Introduction to Linear Models

Linear models are **parametric statistical models** that summarize how the probability distribution of a response variable (usually denoted as Y) depends upon one or more explanatory variables (usually denoted with X's: $X_0, X_1, X_2, \ldots, X_k$).

- They are *statistical (or probabilistic)* because they specify a (conditional) probability distribution of a random variable (or at least some aspects of that distribution, like its mean and variance).
- They are *parametric* because the probability distribution is specified up to a finite number of unknown constants, or parameters.
- They are *linear* because the mean of the conditional probability distribution of Y, $\mathrm{E}(Y|X_0, X_1, \ldots, X_k)$, is specified to be a linear function of model parameters.
- This conditional mean, $\mathrm{E}(Y|X_0, X_1, \ldots, X_k)$, is called the regression function for Y on $X_0, \ldots, X_k$.

The classical linear model specifies that

$$Y = X_0\beta_0 + X_1\beta_1 + \cdots X_k\beta_k + e = \mathbf{x}^T\boldsymbol{\beta} + e$$

where

$$\mathbf{x} = \begin{pmatrix} X_0 \\ \vdots \\ X_k \end{pmatrix}_{(k+1)\times 1}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}_{(k+1)\times 1},$$

where e has conditional mean 0, and variance σ^2.

- Notice that this model implies that the regression function is linear in the β's:
$$\mathrm{E}(Y|\mathbf{x}) = X_0\beta_0 + X_1\beta_1 + \cdots X_k\beta_k.$$
- Notice that this model indirectly specifies the first two moments of the conditional distribution of $Y|\mathbf{x}$ by moment assumptions on e:

$$\mathrm{E}(Y|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}, \quad \text{and} \quad \mathrm{var}(Y|\mathbf{x}) = \sigma^2$$

Strictly speaking, the "model" given above is not a true statistical model because it specifies only the first two moments (i.e., the mean and variance) of Y given $\mathbf{x}$ rather than the entire conditional distribution.

Important results concerning the estimation of the regression function, $\mathrm{E}(Y|\mathbf{x})$, are available based only on mean and variance specifications. However, for inference on the model parameters, it is necessary to complete the model specification by assuming further that $e \sim N(0, \sigma^2)$.

- It then follows that $Y|\mathbf{x} \sim N(\mathbf{x}^T\boldsymbol{\beta}, \sigma^2)$.

Typically, we will have a sample of data consisting of observed values of n independent copies of $(Y, X_0, \ldots, X_k)$:

$$(Y_1, X_{10}, X_{11}, \ldots, X_{1k}), \ldots, (Y_n, X_{n0}, X_{n1}, \ldots, X_{nk}).$$

In this case, the classical linear model is supposed to hold for each copy $(Y_i, X_{i0}, X_{i1}, \ldots, X_{ik})$, $i = 1, \ldots, n$.

That is, the model becomes

$$Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n, \quad \text{where} \quad e_1, \ldots, e_n \overset{iid}{\sim} N(0, \sigma^2) \qquad (*)$$

and $\mathbf{x}_i = (X_{i0}, X_{i1}, \ldots, X_{ik})^T$.

- The notation $\overset{iid}{\sim} N(0, \sigma^2)$ means, "are independent, identically distributed random variables each with a normal distribution with mean 0 and variance σ^2."
- Typically, X_{i0} is equal to one for all i in multiple linear regression models, but this need not be so in general.
- In model (*) the parameters are $\beta_0, \beta_1, \ldots, \beta_k, \sigma^2$. The regression parameters are $\beta_0, \beta_1, \ldots, \beta_k$.

More succinctly, we can write model (*) in vector/matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad e_1, \ldots, e_n \overset{iid}{\sim} N(0, \sigma^2).$$

Example - Simple Linear Regression Model

Suppose that for $n = 6$ mother-daughter pairs we have height data: Y_i =height of daughter in i^{th} pair, X_{i1} =height of mother in i^{th} pair; and in addition, we have information on birth order (1 means daughter was first-born daughter).

Pair (i)	X_{i1}	Y_i	Birth Order
1	62.5	64	1
2	67.5	68	3
3	65	63	1
4	65	66	2
5	60	61	2
6	59.5	66	3

It may be of interest to investigate how a woman's height depends upon her mother's height. As part of that investigation we may consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + e_i, \quad i = 1, \ldots, 6,$$

where $e_1, \ldots, e_6 \overset{iid}{\sim} N(0, \sigma^2)$.

In vector notation this model can be written

$$\begin{pmatrix} 64 \\ 68 \\ \vdots \\ 66 \end{pmatrix} = \begin{pmatrix} 1 & 62.5 \\ 1 & 67.5 \\ \vdots & \vdots \\ 1 & 59.5 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_6 \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where $e_1, \ldots, e_6 \overset{iid}{\sim} N(0, \sigma^2)$.

Example - Multiple Linear Regression

Suppose that we observe that daughters' heights don't increase steadily as mothers' heights increase. Instead, daughters' heights level off for large mothers' heights. We may then wish to consider a model which is quadratic in mother's height:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \underbrace{X_{i1}^2}_{\equiv X_{i2}} + e_i, \quad i = 1, \ldots, 6,$$

where e_i's are as before.

- While this model is quadratic in X_{i1} it is still a *linear* model because it is linear in $\beta_0, \beta_1, \beta_2$.

Example - One-way ANOVA

Suppose we knew only the birth order information and not mother's height. Then we might be interested in fitting a model which allowed for different means for each level of birth order.

The **cell-means one-way ANOVA model** does just that. Let Y_{ij} =the height of the j^{th} daughter in the i^{th} birth-order group ($i = 1, 2, 3$, $j = 1, 2$).

The cell-means model is

$$Y_{ij} = \mu_i + e_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2,$$

where $e_{11}, \ldots, e_{32} \overset{iid}{\sim} N(0, \sigma^2)$.

This model can be written in vector notation as

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

An alternative, *but equivalent,* version of this model is the **effects model:**

$$Y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, 3, j = 1, \ldots, 2,$$

where the e_{ij}'s are as before.

The effects model can be written in vector form as

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

- Notice that, like the multiple regression model, this version of the one-way anova model includes an intercept, although here we've called that intercept μ rather than β_0.
- We'll investigate the equivalence between the cell-means and effects forms of ANOVA models later in the course.

Example - An ANCOVA Model

Suppose we had available both birth-order information and mother's height. A model that accounts for dependence of daughter's height on both of these variables is

$$Y_{ij} = \mu + \alpha_i + \beta X_{ij} + e_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2,$$

where X_{ij} =mother's height for the j^{th} pair in the i^{th} birth-order group. The assumptions on e_{ij} are as in the ANOVA models.

In vector notation this model is

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & X_{11} \\ 1 & 1 & 0 & 0 & X_{12} \\ 1 & 0 & 1 & 0 & X_{21} \\ 1 & 0 & 1 & 0 & X_{22} \\ 1 & 0 & 0 & 1 & X_{31} \\ 1 & 0 & 0 & 1 & X_{32} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

In ANOVA models for designed experiments, the explanatory variables (columns of the **design matrix X**) are fixed by design. That is, they are non-random.

In regression models, the X's are often observed simultaneously with the Y's. That is, they are random variables because they are measured (not assigned or selected) characteristics of the randomly selected unit of observation.

In either case, however, the classical linear model treats $\mathbf{X}$ as fixed, by conditioning on the values of the explanatory variables.

That is, the probability distribution of interest is that of $Y|\mathbf{x}$, and all expectations and variances are conditional on $\mathbf{x}$ (e.g., $\mathrm{E}(Y|\mathbf{x})$, $\mathrm{var}(e|\mathbf{x})$, etc.).

- Because this conditioning applies throughout linear models, we will always consider the explanatory variables to be constants and we'll often drop the conditioning notation (the $|\mathbf{x}$ part).

- If we have time, we will consider the case where $\mathbf{X}$ is considered to be random later in the course. See ch. 10 of our text.

Notation:

When dealing with scalar-valued random variables, it is common (and useful) to use upper and lower case to distinguish between a random variable and the value that it takes on in a given realization.

- E.g., Y, Z are random variables with observed values y, and z, respectively. So, we might be concerned with $\Pr(Z = z)$ (if Z is discrete), or we might condition on $Z = z$ and consider the conditional mean $\mathrm{E}(Y|Z = z)$.

However, when working with vectors and matrices I will drop this distinction and instead denote vectors with bold-faced lower case and matrices with bold upper case. E.g., $\mathbf{y}$ and $\mathbf{x}$ are vectors, and $\mathbf{X}$ a matrix.

The distinction between the random vector (or matrix) and its realized value will typically be clear from the context.

Some Concepts from Linear Algebra

Since our topic is the linear model, its not surprising that many of the most useful mathematical tools come from linear algebra.

Matrices, Vectors, and Matrix Algebra

A **matrix** is a rectangular (or square) array of numbers or variables. E.g., we can arrange the mother-daughter height data (p. 4) in a 6×2 matrix

$$\mathbf{A} = \begin{pmatrix} 62.5 & 64 \\ 67.5 & 68 \\ 65 & 63 \\ 65 & 66 \\ 60 & 61 \\ 59.5 & 66 \end{pmatrix}$$

To represent the elements of $\mathbf{A}$ as variables, we use symbols for the elements:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \\ a_{51} & a_{52} \\ a_{61} & a_{62} \end{pmatrix} \equiv (a_{ij}).$$

The size, or **dimension** of $\mathbf{A}$ is its number of rows (r) and columns (c); in this case, we say $\mathbf{A}$ is 6×2, or in general $r \times c$.

A **vector** is simply a matrix with only one column. E.g.,

$$\mathbf{x} = \begin{pmatrix} 64 \\ 68 \\ 63 \\ 66 \\ 61 \\ 66 \end{pmatrix}$$

is the vector formed from the second column of $\mathbf{A}$ above.

- We will typically denote vectors with boldface lower-case letters (e.g., $\mathbf{x}, \mathbf{y}, \mathbf{z}_1$, etc.) and matrices with boldface upper-case letters (e.g., $\mathbf{A}, \mathbf{B}, \mathbf{M}_1, \mathbf{M}_2$, etc.).

- Vectors will always be column vectors. If we need a row vector we will use the transpose of a vector. E.g., $\mathbf{x}^T = (64, 68, 63, 66, 61, 66)$ is the row vector version of $\mathbf{x}$.

A **scalar** is a 1×1 matrix; i.e., a real-valued number or variable. Scalars will be denoted in ordinary (non-bold) typeface.

Matrices of special form:

A **diagonal matrix** is a square matrix with all of its off-diagonal elements equal to 0. We will use the diag$(\cdot)$ function in two ways: if its argument is a square matrix, then diag$(\cdot)$ yields a vector formed from the diagonal of that matrix; if its argument is a vector, then diag$(\cdot)$ yields a diagonal matrix with that vector on the diagonal. E.g.,

$$\text{diag}\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \\ 9 \end{pmatrix} \qquad \text{diag}\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

The $n \times n$ **identity matrix** is a diagonal matrix with 1's along the diagonal. We will denote this as $\mathbf{I}$, or $\mathbf{I}_n$ when we want to make clear what the dimension is. E.g.,

$$\mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- The identity matrix has the property that

$$\mathbf{IA} = \mathbf{A}, \qquad \mathbf{BI} = \mathbf{B},$$

  where $\mathbf{A}, \mathbf{B}, \mathbf{I}$ are assumed conformable to these multiplications.

A vector of 1's is denoted as $\mathbf{j}$, or $\mathbf{j}_n$ when we want to emphasize that the dimension is n. A matrix of 1's is denoted as $\mathbf{J}$, or $\mathbf{J}_{n,m}$ to emphasize the dimension. E.g.,

$$\mathbf{j}_4 = (1, 1, 1, 1)^T, \qquad \mathbf{J}_{2,3} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

A vector or matrix containing all 0's will be denoted by $\mathbf{0}$. Sometimes we will add subscripts to identify the dimension of this quantity.

Lower and upper-triangular matrices have 0's above and below the diagonal, respectively. E.g.,

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 5 & 6 \end{pmatrix} \qquad \mathbf{U} = \begin{pmatrix} 3 & 2 \\ 0 & 1 \end{pmatrix}.$$

I will assume that you know the basic algebra of vectors and matrices. In particular, it is assumed that you are familiar with

- equality/inequality of matrices (two matrices are equal if they have the same dimension and all corresponding elements are equal);
- matrix addition and subtraction (performed elementwise);
- matrix multiplication and conformability (to perform the matrix multiplication $\mathbf{AB}$, it is necessary for $\mathbf{A}$ and $\mathbf{B}$ to be conformable; i.e., the number of columns of $\mathbf{A}$ must equal the number of rows of $\mathbf{B}$);
- scalar multiplication ($c\mathbf{A}$ is the matrix obtained by multiplying each element of $\mathbf{A}$ by the scalar c);
- transpose of a matrix (interchange rows and columns, denoted with a ^{T} superscript);
- the trace of a square matrix (sum of the diagonal elements; the trace of the $n \times n$ matrix $\mathbf{A} = (a_{ij})$ is $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$);
- the determinant of a matrix (a scalar-valued function of a matrix used in computing a matrix inverse; the determinant of $\mathbf{A}$ is denoted $|\mathbf{A}|$);
- the inverse of a square matrix $\mathbf{A}$, say (a matrix, denoted $\mathbf{A}^{-1}$, whose product with $\mathbf{A}$ yields the identity matrix; i.e., $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$).

- Chapter 2 of our text contains a review of basic matrix algebra.

Some Geometric Concepts:

Euclidean Space: A vector $\binom{x}{y}$ of dimension two can be thought of as representing a point on a two-dimensional plane:

and the collection of all such points defines the plane, which we call $\mathcal{R}^2$.

Similarly, a three-dimensional vector $(x, y, z)^T$ can represent a point in 3-dimensional space:

with the collection of all such triples yielding 3 dimensional space, $\mathcal{R}^3$.

More generally, Euclidean n-space, denoted $\mathcal{R}^n$, is given by the collection of all n-tuples (n-dimensional vectors) consisting of real numbers.

- Actually, the proper geometric interpretation of a vector is as a directed line segment extending from the origin (the point $\mathbf{0}$) to the point indicated by the coordinates (elements) of the vector.

Vector Spaces: $\mathcal{R}^n$ (for each possible value of n) is a special case of the more general concept of a vector space:

Let V denote a set of n-dimensional vectors. If, for every pair of vectors in V, $\mathbf{x}_i \in V$ and $\mathbf{x}_j \in V$, it is true that

i. $\mathbf{x}_i + \mathbf{x}_j \in V$, and
ii. $c\mathbf{x}_i \in V$, for all real scalars c,

then V is said to be a vector space of order n.

- Examples: $\mathcal{R}^n$ (Euclidean n-space) is a vector space because it is closed under addition and scalar multiplication. Another example is the set consisting only of $\mathbf{0}$. Moreover, $\mathbf{0}$ belongs to every vector space in $\mathcal{R}^n$.

Spanning Set, Linear Independence, and Basis. The defining characteristics of a vector space ensure that all linear combinations of vectors in a vector space V are also in V. I.e., if $\mathbf{x}_1, \ldots, \mathbf{x}_k \in V$, then $c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_k\mathbf{x}_k \in V$ for any scalars $c_1, \ldots, c_k$.

Suppose that every vector in a vector space V can be expressed as a linear combination of the k vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$. Then the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is said to **span** or **generate** V, and we write $V = \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$ to denote that V is the vector space spanned by $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$.

If the spanning set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ also has the property of linear independence, then $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is called a **basis** of V.

Vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are **linearly independent** if $\sum_{i=1}^{k} c_i\mathbf{x}_i = \mathbf{0}$ implies that $c_1 = 0, c_2 = 0, \ldots, c_k = 0$.

- I.e., If $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are linearly independent (LIN), then there is no redundancy among them in the sense that it is not possible to write $\mathbf{x}_1$ (say) as a linear combination of $\mathbf{x}_2, \ldots, \mathbf{x}_k$.

- Therefore, a basis of V is a spanning set that is LIN.

- It is not hard to prove that every basis of a given vector space V has the same number of elements. That number of elements is called the **dimension** or **rank** of V.

Example:

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}$$

are all in $\mathcal{R}^3$. The space spanned by $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is

$$\begin{aligned} \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) &= \{a\mathbf{x}_1 + b\mathbf{x}_2 + c\mathbf{x}_3 | a, b, c \in R\} \\ &= \left\{ \begin{pmatrix} a+b+3c \\ 2a-b \\ 0 \end{pmatrix} \middle| a, b, c \in R \right\} = \left\{ \begin{pmatrix} d \\ e \\ 0 \end{pmatrix} \middle| d, e \in R \right\} \end{aligned}$$

- Note that $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are not LIN, because it is possible to write any one of the three vectors as a linear combination of the other two. E.g.,

$$\underbrace{\begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}}_{\mathbf{x}_3} = \underbrace{\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}}_{\mathbf{x}_1} + 2 \underbrace{\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}}_{\mathbf{x}_2}.$$

- This linear dependence can be removed by eliminating any one of the three vectors from the set. So, for example, $\mathbf{x}_1, \mathbf{x}_2$ are LIN and span the same set as do $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$. That is, $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \equiv V$, so $\{\mathbf{x}_1, \mathbf{x}_2\}$ and $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ are both spanning sets for V, but only $\{\mathbf{x}_1, \mathbf{x}_2\}$ is a basis for V.

- Bases are not unique. $\{\mathbf{x}_2, \mathbf{x}_3\}$ and $\{\mathbf{x}_1, \mathbf{x}_3\}$ are both bases for V as well in this example.

- Note also that here V is of order 3 and $\mathcal{R}^3$ is of order 3, but $V \neq \mathcal{R}^3$. In general, there are many vector spaces of a given order.

Subspaces. Let V be a vector space and W be a set with $W \subset V$. Then W is a subspace of V if and only if W is also a vector space.

Example: Let V_1, V_2, V_3 be the sets of vectors having the forms $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, respectively, where

$$\mathbf{x} = \begin{pmatrix} \alpha \\ 0 \\ 0 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} 0 \\ 0 \\ \beta \end{pmatrix}, \mathbf{z} = \begin{pmatrix} \gamma \\ 0 \\ \delta \end{pmatrix}, \quad \text{for real } \alpha, \beta, \gamma, \text{ and } \delta.$$

I.e.,

$$V_1 = \mathcal{L}\left(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}\right), V_2 = \mathcal{L}\left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right), V_3 = \mathcal{L}\left(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right).$$

Then V_1, V_2, V_3 each define a vector space of order 3, each of which is a subspace of $\mathcal{R}^3$. In addition, V_1 and V_2 are each subspaces of V_3.

- In this course, we will be concerned with vector spaces that are subspaces of $\mathcal{R}^n$.

Column Space, Rank of a Matrix. The **column space** of a matrix $\mathbf{A}$ is denoted $C(\mathbf{A})$, and defined as the space spanned by the columns of $\mathbf{A}$. I.e., if $\mathbf{A}$ is an $n \times m$ matrix with columns $\mathbf{a}_1, \ldots, \mathbf{a}_m$ so that $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$, then $C(\mathbf{A}) = \mathcal{L}(\mathbf{a}_1, \ldots, \mathbf{a}_m)$.

The **rank** of $\mathbf{A}$ is defined to be the dimension of $C(\mathbf{A})$. I.e., the number of LIN columns of $\mathbf{A}$.

Some properties of the rank of a matrix.

1. For an $m \times n$ matrix $\mathbf{A}$,

$$\text{rank}(\mathbf{A}) \leq \min(m, n).$$

   If $\text{rank}(\mathbf{A}) = \min(m, n)$ then $\mathbf{A}$ is said to be of **full rank**.

2. $\text{rank}(\mathbf{A} + \mathbf{B}) \leq \text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B})$.

3. If the matrices $\mathbf{A}$, $\mathbf{B}$ are conformable to the multiplication $\mathbf{AB}$ then

$$\text{rank}(\mathbf{AB}) \leq \min\{\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\}.$$

4. For any $n \times n$ matrix $\mathbf{A}$, $|\mathbf{A}| = 0$ if and only if $\text{rank}(\mathbf{A}) < n$.

- An $n \times n$ matrix $\mathbf{A}$ has an inverse (i.e., is nonsingular) if and only if $|\mathbf{A}| \neq 0$, so also iff $\text{rank}(\mathbf{A}) = n$.

5. For nonsingular matrices $\mathbf{A}$, $\mathbf{B}$, and any matrix $\mathbf{C}$, then

$$\text{rank}(\mathbf{C}) = \text{rank}(\mathbf{AC}) = \text{rank}(\mathbf{CB}) = \text{rank}(\mathbf{ACB}).$$

6. $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T) = \text{rank}(\mathbf{A}^T\mathbf{A}) = \text{rank}(\mathbf{AA}^T)$.

7. For $\mathbf{A}$ an $m \times n$ matrix and $\mathbf{b}$ an $m \times 1$ vector,

$$\text{rank}([\mathbf{A}, \mathbf{b}]) \geq \text{rank}(\mathbf{A})$$

   (adding a column to $\mathbf{A}$ can't reduce its rank).

And a couple of properties of the column space of a matrix:

8. $C(\mathbf{A}^T\mathbf{A}) = C(\mathbf{A}^T)$.

9. $C(\mathbf{ACB}) = C(\mathbf{AC})$ if $\text{rank}(\mathbf{CB}) = \text{rank}(\mathbf{C})$.

Inner Products, Length, Orthogonality, and Projections.

For two vectors $\mathbf{x}$ and $\mathbf{y}$ in a vector space V of order n, we define $\langle \mathbf{x}, \mathbf{y} \rangle$ to be the **inner product** operation given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x}.$$

- When working in more general spaces where $\mathbf{x}$ and $\mathbf{y}$ may not be vectors, the inner product may still be defined as $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i$ even when the multiplication $\mathbf{x}^T \mathbf{y}$ is not defined. In addition, in some contexts the inner product may be defined differently (e.g., as $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{A} \mathbf{y}$ for some matrix $\mathbf{A}$). Therefore, sometimes it is of use to distinguish $\langle \mathbf{x}, \mathbf{y} \rangle$ from $\mathbf{x}^T \mathbf{y}$. However, in this course these two operations will always be the same. Nevertheless, I will sometimes write $\langle \mathbf{x}, \mathbf{y} \rangle$ for $\mathbf{x}^T \mathbf{y}$ even though they mean the same thing.

- The inner product is sometimes called the dot product. Several notations are commonly used, including $\mathbf{x} \cdot \mathbf{y}$ and $(\mathbf{x}, \mathbf{y})$.

Properties of the inner product:

1. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$
2. $\langle a\mathbf{x}, \mathbf{y} \rangle = a \langle \mathbf{x}, \mathbf{y} \rangle$
3. $\langle \mathbf{x}_1 + \mathbf{x}_2, \mathbf{y} \rangle = \langle \mathbf{x}_1, \mathbf{y} \rangle + \langle \mathbf{x}_2, \mathbf{y} \rangle$

(Euclidean) Length: The Euclidean length of a vector $\mathbf{x} \in \mathcal{R}^n$ is defined to be $\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\mathbf{x}^T \mathbf{x}}$ and is denoted as $||\mathbf{x}||$. That is, $||\mathbf{x}||^2 = \langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^T \mathbf{x}$.

- It is possible to define other types of lengths, but unless otherwise stated, lengths will be assumed to be Euclidean as defined above.

- E.g., for $\mathbf{x} = (3, 4, 12)$, the length of $\mathbf{x}$ is $||\mathbf{x}|| = \sqrt{\sum_i x_i x_i} = \sqrt{9 + 16 + 144} = \sqrt{169} = 13$.

(Euclidean) Distance: The distance between vectors $\mathbf{x}, \mathbf{y} \in \mathcal{R}^n$ is the length of $\mathbf{x} - \mathbf{y}$.

- The inner product between two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{R}^n$ quantifies the angle between them. In particular, if θ is the angle formed between $\mathbf{x}$ and $\mathbf{y}$ then

$$\cos(\theta) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{||\mathbf{x}||||\mathbf{y}||}.$$

Orthogonality: $\mathbf{x}$ and $\mathbf{y}$ are said to be orthogonal (i.e., perpendicular) if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. The orthogonality of $\mathbf{x}$ and $\mathbf{y}$ is denoted with the notation $\mathbf{x} \perp \mathbf{y}$.

- Note that orthogonality is a property of a pair of vectors. When we want to say that each pair in the collection of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is orthogonal, then we say that $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are **mutually orthogonal**.

Example: Consider the model matrix from the ANCOVA example on p. 6:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 62.5 \\ 1 & 1 & 0 & 0 & 67.5 \\ 1 & 0 & 1 & 0 & 65 \\ 1 & 0 & 1 & 0 & 65 \\ 1 & 0 & 0 & 1 & 60 \\ 1 & 0 & 0 & 1 & 59.5 \end{pmatrix}$$

Let $\mathbf{x}_1, \ldots, \mathbf{x}_5$ be the columns of $\mathbf{X}$. Then $\mathbf{x}_2 \perp \mathbf{x}_3, \mathbf{x}_2 \perp \mathbf{x}_4, \mathbf{x}_3 \perp \mathbf{x}_4$ and the other pairs of vectors are not orthogonal. I.e., $\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ are mutually orthogonal.

The length of these vectors are

$$||\mathbf{x}_1|| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{6}, \qquad ||\mathbf{x}_2|| = ||\mathbf{x}_3|| = ||\mathbf{x}_3|| = \sqrt{2},$$

and

$$||\mathbf{x}_5|| = \sqrt{62.5^2 + \cdots + 59.5^2} = 155.09.$$

Pythagorean Theorem: Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be mutually orthogonal vectors in a vector space V. Then

$$\left\| \sum_{i=1}^{k} \mathbf{v}_i \right\|^2 = \sum_{i=1}^{k} ||\mathbf{v}_i||^2$$

Proof:

$$\left\| \sum_{i=1}^{k} \mathbf{v}_i \right\|^2 = \left\langle \sum_{i=1}^{k} \mathbf{v}_i, \sum_{j=1}^{k} \mathbf{v}_j \right\rangle = \sum_{i=1}^{k} \sum_{j=1}^{k} \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \sum_{i=1}^{k} \langle \mathbf{v}_i, \mathbf{v}_i \rangle = \sum_{i=1}^{k} ||\mathbf{v}_i||^2$$

■

Projections: The (orthogonal) **projection** of a vector $\mathbf{y}$ on a vector $\mathbf{x}$ is the vector $\hat{\mathbf{y}}$ such that

1. $\hat{\mathbf{y}} = b\mathbf{x}$ for some constant b; and
2. $(\mathbf{y} - \hat{\mathbf{y}}) \perp \mathbf{x}$ (or, equivalently, $\langle \hat{\mathbf{y}}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$).

- It is possible to define non-orthogonal projections, but by default when we say "projection" we will mean the orthogonal projection as defined above.

- The notation $p(\mathbf{y}|\mathbf{x})$ will denote the projection of $\mathbf{y}$ on $\mathbf{x}$.

Some Pictures:

Vector: A line segment from the origin (**0**, all elements equal to zero) to the point indicated by the coordinates of the (algebraic) vector.

$$\text{E.g.,} \quad \mathbf{x} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$$

Vector Addition:

Scalar Multiplication: Multiplication by a scalar **scales** a vector by shrinking or extending the vector in the same direction (or opposite direction if the scalar is negative).

Projection: The projection of $\mathbf{y}$ on $\mathbf{x}$ is the vector in the direction of $\mathbf{x}$ (part 1 of the definition) whose difference from $\mathbf{y}$ is orthogonal (perpendicular) to $\mathbf{x}$ (part 2 of the definition).

- From the picture above it is clear that $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ on the subspace of all vectors of the form $a\mathbf{x}$, the subspace spanned by $\mathbf{x}$.

It is straight-forward to find $\hat{\mathbf{y}} = p(\mathbf{y}|\mathbf{x})$, the projection of $\mathbf{y}$ on $\mathbf{x}$:

By part 1 of the definition $\hat{\mathbf{y}} = b\mathbf{x}$, and by part 2,

$$\hat{\mathbf{y}}^T\mathbf{x} = \mathbf{y}^T\mathbf{x}.$$

But since

$$\hat{\mathbf{y}}^T\mathbf{x} = (b\mathbf{x})^T\mathbf{x} = b\mathbf{x}^T\mathbf{x} = b\|\mathbf{x}\|^2,$$

the definition implies that

$$b\|\mathbf{x}\|^2 = \mathbf{y}^T\mathbf{x} \quad \Rightarrow \quad b = \frac{\mathbf{y}^T\mathbf{x}}{\|\mathbf{x}\|^2}$$

unless $\mathbf{x} = \mathbf{0}$, in which case b could be any constant.

So, $\hat{\mathbf{y}}$ is given by

$$\hat{\mathbf{y}} = \begin{cases} (\text{any constant})\mathbf{0} = \mathbf{0}, & \text{for } \mathbf{x} = \mathbf{0} \\ \left(\frac{\mathbf{y}^T\mathbf{x}}{\|\mathbf{x}\|^2}\right)\mathbf{x}, & \text{otherwise} \end{cases}$$

Example: In $\mathcal{R}^2$, let $\mathbf{y} = \binom{4}{3}$, $\mathbf{x} = \binom{5}{0}$. Then to find $\hat{\mathbf{y}} = p(\mathbf{y}|\mathbf{x})$ we compute $\mathbf{x}^T\mathbf{y} = 20$, $\|\mathbf{x}\|^2 = 25$, $b = 20/25 = 4/5$ so that $\hat{\mathbf{y}} = b\mathbf{x} = (4/5)\binom{5}{0} = \binom{4}{0}$.
In addition, $\mathbf{y} - \hat{\mathbf{y}} = \binom{0}{3}$ so $\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{x}$.

In this case, the Pythagorean Theorem reduces to its familiar form from high school algebra. The squared length of the hypotenuse ($\|\mathbf{y}\|^2 = 16 + 9 = 25$) is equal to the sum of the squared lengths of the other two sides of a right triangle ($\|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}}\|^2 = 9 + 16 = 25$).

Theorem Among all multiples $a\mathbf{x}$ of $\mathbf{x}$, the projection $\hat{\mathbf{y}} = p(\mathbf{y}|\mathbf{x})$ is the closest vector to $\mathbf{y}$.

Proof: Let $\mathbf{y}^* = c\mathbf{x}$ for some constant c. $\hat{\mathbf{y}}$ is such that

$$(\mathbf{y} - \hat{\mathbf{y}}) \perp a\mathbf{x} \quad \text{for any scalar } a$$

so in particular, for $b = \mathbf{y}^T\mathbf{x}/\|\mathbf{x}\|^2$,

$$(\mathbf{y} - \hat{\mathbf{y}}) \perp \underbrace{(b-c)\mathbf{x}}_{=b\mathbf{x}-c\mathbf{x}=\hat{\mathbf{y}}-\mathbf{y}^*}$$

In addition,

$$\mathbf{y} - \mathbf{y}^* = \underbrace{\mathbf{y} - \hat{\mathbf{y}}} + \underbrace{\hat{\mathbf{y}} - \mathbf{y}^*}$$

so the P.T. implies

$$\begin{aligned} & \|\mathbf{y} - \mathbf{y}^*\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \mathbf{y}^*\|^2 \\ \Rightarrow \quad & \|\mathbf{y} - \mathbf{y}^*\|^2 \geq \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \quad \text{for all } \mathbf{y}^* = c\mathbf{x} \end{aligned}$$

with equality if and only if $\mathbf{y}^* = \hat{\mathbf{y}}$. ∎

The same sort of argument establishes the **Cauchy-Schwartz Inequality:**

Since $\hat{\mathbf{y}} \perp (\mathbf{y} - \hat{\mathbf{y}})$ and $\mathbf{y} = \hat{\mathbf{y}} + (\mathbf{y} - \hat{\mathbf{y}})$, it follows from the P.T. that

$$\begin{aligned} \|\mathbf{y}\|^2 &= \|\hat{\mathbf{y}}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \\ &= b^2\|\mathbf{x}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \\ &= \frac{(\mathbf{y}^T\mathbf{x})^2}{\|\mathbf{x}\|^2} + \underbrace{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}_{\geq 0} \\ \Rightarrow \quad \|\mathbf{y}\|^2 &\geq \frac{(\mathbf{y}^T\mathbf{x})^2}{\|\mathbf{x}\|^2} \end{aligned}$$

or

$$(\mathbf{y}^T\mathbf{x})^2 \leq \|\mathbf{y}\|^2\|\mathbf{x}\|^2$$

with equality if and only if $\|\mathbf{y} - \hat{\mathbf{y}}\| = 0$ (i.e., iff $\mathbf{y}$ is a multiple of $\mathbf{x}$).

Projections onto 0/1 or indicator vectors. Consider a vector space in $\mathcal{R}^n$. Let A be a subset of the indices $1, \ldots, n$. Let $\mathbf{i}_A$ denote the indicator vector for A; that is $\mathbf{i}_A$ is the n-dimensional vector with 1's in the positions given in the set A, and 0's elsewhere.

- E.g., the columns of the model matrix in the cell-means version of the one-way ANOVA model are all indicator variables. Recall the mother-daughter height example that had $n = 6$ and two observations per birth-order group. The model matrix was

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3].$$

  Here, $\mathbf{x}_i$ is an indicator vector for the i^{th} birth-order group. I.e., $\mathbf{x}_i = \mathbf{i}_{A_i}$ where $A_1 = \{1, 2\}, A_2 = \{3, 4\}, A_3 = \{5, 6\}$.

The projection of a vector $\mathbf{y}$ onto an indicator vector is simply the mean of those elements of $\mathbf{y}$ being indicated, times the indicator vector.

I.e.,the projection $\hat{\mathbf{y}}_A$ of $\mathbf{y}$ on $\mathbf{i}_A$ is $b\mathbf{i}_A$ where

$$b = \frac{\mathbf{y}^T \mathbf{i}_A}{\|\mathbf{i}_A\|^2} = \frac{\sum_{i \in A} y_i}{N(A)}$$

where $N(A)$ =the number of indices in A. That is, $b = \bar{y}_A$, the mean of the $y-$values with components in A, so that $\hat{\mathbf{y}}_A = \bar{y}_A \mathbf{i}_A$.

Example: the daughters' height data were given by $\mathbf{y} = (64, 68, 63, 66, 61, 66)^T$. The projection of $\mathbf{y}$ onto $\mathbf{x}_1$ is

$$p(\mathbf{y}|\mathbf{x}_1) = \underbrace{\frac{64+68}{1^2+1^2}}_{=66}\mathbf{x}_1 = \begin{pmatrix} 66 \\ 66 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Similarly,

$$p(\mathbf{y}|\mathbf{x}_2) = \frac{63+66}{2}\mathbf{x}_2 = \begin{pmatrix} 0 \\ 0 \\ 64.5 \\ 64.5 \\ 0 \\ 0 \end{pmatrix}, \qquad p(\mathbf{y}|\mathbf{x}_3) = \frac{61+66}{2}\mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 63.5 \\ 63.5 \end{pmatrix}.$$

Orthogonality to a Subspace: A vector $\mathbf{y}$ is *orthogonal* to a subspace V of $\mathcal{R}^n$ if $\mathbf{y}$ is orthogonal to all vectors in V. If so, we write $\mathbf{y} \perp V$.

Orthogonal Complement: Let $V^\perp = \{\mathbf{v} \in \mathcal{R}^n | \mathbf{v} \perp V\}$. $V^\perp$ is called the orthogonal complement of V.

More generally, if $V \subset W$, then $V^\perp \cap W = \{\mathbf{v} \in W | \mathbf{v} \perp V\}$ is called the orthogonal complement of V with respect to W.

- It can be shown that if W is a vector space and V is a subspace of W, then the orthogonal complement of V with respect to W is a subspace of W and for any $\mathbf{w} \in W$, $\mathbf{w}$ can be written uniquely as $\mathbf{w} = \mathbf{w}_0 + \mathbf{w}_1$, where $\mathbf{w}_0 \in V$ and $\mathbf{w}_1 \in V^\perp \cap W$. The ranks (dimensions) of these subspaces satisfy $\text{rank}(V) + \text{rank}(V^\perp \cap W) = \text{rank}(W)$.

Projection onto a Subspace: The *projection* of a vector $\mathbf{y}$ on a subspace V of $\mathcal{R}^n$ is the vector $\hat{\mathbf{y}} \in V$ such that $(\mathbf{y} - \hat{\mathbf{y}}) \perp V$. The vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ will be called the *residual vector* for $\mathbf{y}$ relative to V.

- Fitting linear models is all about finding projections of a response vector onto a subspace defined as a linear combination of several vectors of explanatory variables. For this reason, the previous definition is central to this course.

- The condition $(\mathbf{y} - \hat{\mathbf{y}}) \perp V$ is equivalent to $(\mathbf{y} - \hat{\mathbf{y}})^T\mathbf{x} = 0$ or $\mathbf{y}^T\mathbf{x} = \hat{\mathbf{y}}^T\mathbf{x}$ for all $\mathbf{x} \in V$. Therefore, the projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto V is the vector which has the same inner product as does $\mathbf{y}$ with each vector in V.

Comment: If vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ span a subspace V then a vector $\mathbf{z} \in V$ equals $p(\mathbf{y}|V)$ if $\langle \mathbf{z}, \mathbf{x}_i \rangle = \langle \mathbf{y}, \mathbf{x}_i \rangle$ for all i.

Why?

Because any vector $\mathbf{x} \in V$ can be written as $\sum_{j=1}^k b_j \mathbf{x}_j$ for some scalars $b_1, \ldots, b_k$, so for any $\mathbf{x} \in V$ if $\langle \mathbf{z}, \mathbf{x}_j \rangle = \langle \mathbf{y}, \mathbf{x}_j \rangle$ for all j, then

$$\langle \mathbf{z}, \mathbf{x} \rangle = \langle \mathbf{z}, \sum_{j=1}^k b_j \mathbf{x}_j \rangle = \sum_{j=1}^k b_j \langle \mathbf{z}, \mathbf{x}_j \rangle = \sum_{j=1}^k b_j \langle \mathbf{y}, \mathbf{x}_j \rangle = \langle \mathbf{y}, \sum_{j=1}^k b_j \mathbf{x}_j \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$$

Since any vector $\mathbf{x}$ in a $k-$dimensional subspace V can be expressed as a linear combination of basis vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$, this suggests that we might be able to compute the projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ on V by summing the projections $\hat{\mathbf{y}}_i = p(\mathbf{y}|\mathbf{x}_i)$.

- We'll see that this works, but only if $\mathbf{x}_1, \ldots, \mathbf{x}_k$ form an **orthogonal basis** for V.

First, does a projection $p(\mathbf{y}|V)$ as we've defined it exist at all, and if so, is it unique?

- We do know that a projection onto a one-dimensional subspace exists and is unique. Let $V = \mathcal{L}(\mathbf{x})$, for $\mathbf{x} \neq \mathbf{0}$. Then we've seen that $\hat{\mathbf{y}} = p(\mathbf{y}|V)$ is given by

$$\hat{\mathbf{y}} = p(\mathbf{y}|\mathbf{x}) = \frac{\mathbf{y}^T\mathbf{x}}{\|\mathbf{x}\|^2}\mathbf{x}.$$

Example Consider $\mathcal{R}^5$, the vector space containing all 5−dimensional vectors of real numbers. Let $A_1 = \{1,3\}, A_2 = \{2,5\}, A_3 = \{4\}$ and $V = \mathcal{L}(\mathbf{i}_{A_1}, \mathbf{i}_{A_2}, \mathbf{i}_{A_3}) = C(\mathbf{X})$, where $\mathbf{X}$ is the 5×3 matrix with columns $\mathbf{i}_{A_1}, \mathbf{i}_{A_2}, \mathbf{i}_{A_3}$.

Let $\mathbf{y} = (6, 10, 4, 3, 2)^T$. It is easy to show that the vector

$$\hat{\mathbf{y}} = \sum_{i=1}^{3} p(\mathbf{y}|\mathbf{i}_{A_i}) = 5\mathbf{i}_{A_1} + 6\mathbf{i}_{A_2} + 3\mathbf{i}_{A_3} = \begin{pmatrix} 5 \\ 6 \\ 5 \\ 3 \\ 6 \end{pmatrix}$$

satisfies the conditions for a projection onto V (need to check that $\hat{\mathbf{y}} \in V$ and $\hat{\mathbf{y}}^T \mathbf{i}_{A_i} = \mathbf{y}^T \mathbf{i}_{A_i}$, $i = 1, 2, 3$).

- The representation of $\hat{\mathbf{y}}$ as the sum of projections on linearly independent vectors spanning V is possible here because $\mathbf{i}_{A_1}, \mathbf{i}_{A_2}, \mathbf{i}_{A_3}$ are mutually orthogonal.

Uniqueness of projection onto a subspace: Suppose $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2$ are two vectors satisfying the definition of $p(\mathbf{y}|V)$. Then $\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 \in V$ and

$$\begin{aligned} & \langle \mathbf{y} - \hat{\mathbf{y}}_1, \mathbf{x} \rangle = 0 = \langle \mathbf{y} - \hat{\mathbf{y}}_2, \mathbf{x} \rangle, \quad \forall \mathbf{x} \in V \\ \Rightarrow \quad & \langle \mathbf{y}, \mathbf{x} \rangle - \langle \hat{\mathbf{y}}_1, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle - \langle \hat{\mathbf{y}}_2, \mathbf{x} \rangle \quad \forall \mathbf{x} \in V \\ & \Rightarrow \quad \langle \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2, \mathbf{x} \rangle = 0 \quad \forall \mathbf{x} \in V \end{aligned}$$

so $\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2$ is orthogonal to all vectors in V including itself, so

$$\begin{aligned} & \|\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2\|^2 = \langle \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2, \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 \rangle = 0 \\ & \Rightarrow \quad \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 = \mathbf{0} \quad \Rightarrow \quad \hat{\mathbf{y}}_1 = \hat{\mathbf{y}}_2 \end{aligned}$$

■

Existence of $p(\mathbf{y}|V)$ is based on showing how to find $p(\mathbf{y}|V)$ if an orthogonal basis for V exists, and then showing that an orthogonal basis always exists.

Theorem: Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be an orthogonal basis for V, a subspace of $\mathcal{R}^k$. Then

$$p(\mathbf{y}|V) = \sum_{i=1}^{k} p(\mathbf{y}|\mathbf{v}_i).$$

Proof: Let $\hat{\mathbf{y}}_i = p(\mathbf{y}|\mathbf{v}_i) = b_i\mathbf{v}_i$ for $b_i = \langle \mathbf{y}, \mathbf{v}_i\rangle / \|\mathbf{v}_i\|^2$. Since $\hat{\mathbf{y}}_i$ is a scalar multiple of $\mathbf{v}_i$, it is orthogonal to $\mathbf{v}_j$, $j \neq i$. From the comment on the top of p.25, we need only show that $\sum_i \hat{\mathbf{y}}_i$ and $\mathbf{y}$ have the same inner product with each $\mathbf{v}_j$. This is true because, for each $\mathbf{v}_j$,

$$\langle \sum_i \hat{\mathbf{y}}_i, \mathbf{v}_j\rangle = \sum_i b_i \langle \mathbf{v}_i, \mathbf{v}_j\rangle = b_j \|\mathbf{v}_j\|^2 = \langle \mathbf{y}, \mathbf{v}_j\rangle.$$

■

Example: Let

$$\mathbf{y} = \begin{pmatrix} 6 \\ 3 \\ 3 \end{pmatrix}, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 3 \\ 0 \\ -3 \end{pmatrix}, \quad V = \mathcal{L}(\mathbf{v}_1, \mathbf{v}_2).$$

Then $\mathbf{v}_1 \perp \mathbf{v}_2$ and

$$\begin{aligned} p(\mathbf{y}|V) = \hat{\mathbf{y}} = p(\mathbf{y}|\mathbf{v}_1) + p(\mathbf{y}|\mathbf{v}_2) &= \left(\frac{12}{3}\right)\mathbf{v}_1 + \left(\frac{9}{18}\right)\mathbf{v}_2 \\ &= \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix} + \begin{pmatrix} 3/2 \\ 0 \\ -3/2 \end{pmatrix} = \begin{pmatrix} 5.5 \\ 4 \\ 2.5 \end{pmatrix} \end{aligned}$$

In addition, $\langle \mathbf{y}, \mathbf{v}_1\rangle = 12$, $\langle \mathbf{y}, \mathbf{v}_2\rangle = 9$ are the same as $\langle \hat{\mathbf{y}}, \mathbf{v}_1\rangle = 12$, and $\langle \hat{\mathbf{y}}, \mathbf{v}_2\rangle = 16.5 - 7.5 = 9$. The residual vector is $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (.5, -1, .5)^T$ which is orthogonal to V.

Note that the Pythagorean Theorem holds:

$$\|\mathbf{y}\|^2 = 54, \quad \|\hat{\mathbf{y}}\|^2 = 52.5, \quad \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = 1.5.$$

- We will see that this result generalizes to become the decomposition of the total sum of squares into model and error sums of squares in a linear model.

Every subspace contains an orthogonal basis (infinitely many, actually). Such a basis can be constructed by the *Gram-Schmidt orthogonalization* method.

Gram-Schmidt Orthogonalization: Let $\mathbf{x}_1, \ldots, \mathbf{x}_k$ be a basis for a $k-$dimensional subspace V of $\mathcal{R}^k$. For $i = 1, \ldots, k$, let $V_i = \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_i)$ so that $V_1 \subset V_2 \subset \cdots \subset V_k$ are nested subspaces. Let

$$\begin{aligned}
&\mathbf{v}_1 = \mathbf{x}_1, \\
&\mathbf{v}_2 = \mathbf{x}_2 - p(\mathbf{x}_2|V_1) = \mathbf{x}_2 - p(\mathbf{x}_2|\mathbf{v}_1), \\
&\mathbf{v}_3 = \mathbf{x}_3 - p(\mathbf{x}_3|V_2) = \mathbf{x}_3 - \{p(\mathbf{x}_3|\mathbf{v}_1) + p(\mathbf{x}_3|\mathbf{v}_2)\}, \\
&\quad \vdots \\
&\mathbf{v}_k = \mathbf{x}_k - p(\mathbf{x}_k|V_{k-1}) = \mathbf{x}_k - \{p(\mathbf{x}_k|\mathbf{v}_1) + \cdots + p(\mathbf{x}_k|\mathbf{v}_{k-1})\}
\end{aligned}$$

- By construction, $\mathbf{v}_1 \perp \mathbf{v}_2$ and $\mathbf{v}_1, \mathbf{v}_2$ span V_2; $\mathbf{v}_3 \perp V_2$ ($\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are mutually orthogonal) and $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ span V_3; etc., etc., so that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are mutually orthogonal spanning V.

If $\mathbf{v}_1, \ldots, \mathbf{v}_k$ form an orthogonal basis for V, then

$$\hat{\mathbf{y}} = p(\mathbf{y}|V) = \sum_{j=1}^{k} p(\mathbf{y}|\mathbf{v}_j) = \sum_{j=1}^{k} b_j \mathbf{v}_j, \quad \text{where } b_j = \langle \mathbf{y}, \mathbf{v}_j \rangle / \|\mathbf{v}_j\|^2$$

so that the Pythagorean Theorem implies

$$\|\hat{\mathbf{y}}\|^2 = \sum_{j=1}^{k} \|b_j \mathbf{v}_j\|^2 = \sum_{j=1}^{k} b_j^2 \|\mathbf{v}_j\|^2 = \sum_{j=1}^{k} \frac{\langle \mathbf{y}, \mathbf{v}_j \rangle^2}{\|\mathbf{v}_j\|^2}.$$

Orthonormal Basis: Two vectors are said to be *orthonormal* if they are orthogonal to one another and each has length one.

- Any vector $\mathbf{v}$ can be rescaled to have length one simply by multiplying that vector by the scalar $1/\|\mathbf{v}\|$ (dividing by its length).

If $\mathbf{v}_1^*, \ldots, \mathbf{v}_k^*$ form an orthonormal basis for V, then the results above simplify to

$$\hat{\mathbf{y}} = \sum_{j}^{k} \langle \mathbf{y}, \mathbf{v}_j^* \rangle \mathbf{v}_j^* \quad \text{and} \quad \|\hat{\mathbf{y}}\|^2 = \sum_{j=1}^{k} \langle \mathbf{y}, \mathbf{v}_j^* \rangle^2.$$

Example: The vectors

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

are a basis for V, the subspace of vectors with the form $(a, b, b)^T$. To orthonormalize this basis, take $\mathbf{v}_1 = \mathbf{x}_1$, then take

$$\begin{aligned} \mathbf{v}_2 = \mathbf{x}_2 - p(\mathbf{x}_2|\mathbf{v}_1) &= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \frac{\langle \mathbf{x}_2, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 \\ &= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \frac{1}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2/3 \\ -1/3 \\ -1/3 \end{pmatrix} \end{aligned}$$

$\mathbf{v}_1, \mathbf{v}_2$ form an orthogonal basis for V, and

$$\mathbf{v}_1^* = \frac{1}{\sqrt{3}} \mathbf{v}_1 = \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix}, \quad \mathbf{v}_2^* = \frac{1}{\sqrt{6/9}} \mathbf{v}_2 = \begin{pmatrix} 2/\sqrt{6} \\ -1/\sqrt{6} \\ -1/\sqrt{6} \end{pmatrix}$$

form an orthonormal basis for V.

- The Gram-Schmidt method provides a method to find the projection of $\mathbf{y}$ onto the space spanned by any collection of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ (i.e., the column space of any matrix).

- Another method is to solve a matrix equation that contains the k simultaneous linear equations known as the *normal equations*. This may necessitate the use of a generalized inverse of a matrix if $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are linearly dependent (i.e., if the matrix with columns $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is not of full rank).

- See homework #1 for how the Gram-Schmidt approach leads to non-matrix formulas for regression coefficients; in what follows we develop the matrix approach.

Consider $V = \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k) = C(\mathbf{X})$, where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k]$. $\hat{\mathbf{y}}$, the projection of $\mathbf{y}$ onto V is a vector in V that forms the same angle as does $\mathbf{y}$ with each of the vectors in the spanning set $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$.

That is, $\hat{\mathbf{y}}$ has the form $\hat{\mathbf{y}} = b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k$ where $\langle \hat{\mathbf{y}}, \mathbf{x}_i \rangle = \langle \mathbf{y}, \mathbf{x}_i \rangle$, for all i.

These requirements can be expressed as a system of equations, called the **normal equations:**

$$\langle \hat{\mathbf{y}}, \mathbf{x}_i \rangle = \langle \mathbf{y}, \mathbf{x}_i \rangle, \quad i = 1, \ldots, k,$$

or, since $\hat{\mathbf{y}} = \sum_j^k b_j \mathbf{x}_j$,

$$\sum_{j=1}^{k} b_j \langle \mathbf{x}_j, \mathbf{x}_i \rangle = \langle \mathbf{y}, \mathbf{x}_i \rangle, \quad i = 1, \ldots, k.$$

More succinctly, these equations can be written as a single matrix equation:

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}, \qquad \textbf{(the normal equation)}$$

where $\mathbf{b} = (b_1, \ldots, b_k)^T$.

To see this note that $\mathbf{X}^T\mathbf{X}$ has $(i,j)^{\text{th}}$ element $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^T\mathbf{x}_j$, and $\mathbf{X}^T\mathbf{y}$ is the $k \times 1$ vector with i^{th} element $\langle \mathbf{y}, \mathbf{x}_i \rangle = \mathbf{x}_i^T\mathbf{y}$:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_k^T \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k \end{pmatrix} = (\mathbf{x}_i^T\mathbf{x}_j) \quad \text{and} \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_k^T \end{pmatrix} \mathbf{y} = \begin{pmatrix} \mathbf{x}_1^T\mathbf{y} \\ \mathbf{x}_2^T\mathbf{y} \\ \vdots \\ \mathbf{x}_k^T\mathbf{y} \end{pmatrix}$$

- If $\mathbf{X}^T\mathbf{X}$ has an inverse (is **nonsingular**) then the equation is easy to solve:
$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- Assume $\mathbf{X}$ is $n \times k$ with $n \geq k$. From rank property 6 (p. 15) we know that
$$\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X}).$$
Therefore, we conclude that the $k \times k$ matrix $\mathbf{X}^T\mathbf{X}$ has full rank k and thus is nonsingular if and only if $\mathbf{X}_{n\times k}$ has rank k (has *full column rank* or linearly independent columns).

If we write $\hat{\mathbf{y}} = b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k = \mathbf{Xb}$, then for $\mathbf{X}$ of full rank we have

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{Py}, \quad \text{where} \quad \mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

- $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the **(orthogonal) projection matrix** onto $C(\mathbf{X})$ because premultiplying $\mathbf{y}$ by $\mathbf{P}$ produces the projection $\hat{\mathbf{y}} = p(\mathbf{y}|C(\mathbf{X}))$.
- $\mathbf{P}$ is sometimes called the **hat matrix** because it "puts the hat on" $\mathbf{y}$. Our book uses the notation $\mathbf{H}$ instead of $\mathbf{P}$ ('H' for hat, 'P' for projection, but these are just two different terms and symbols for the same thing).

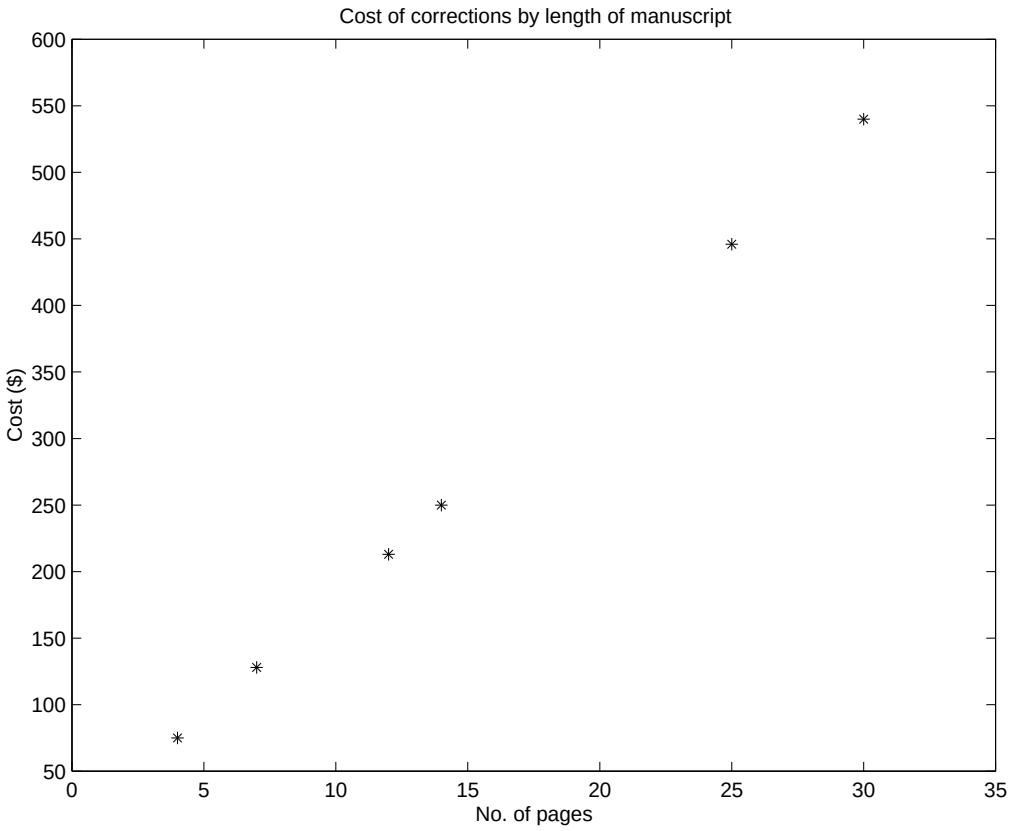
Since $\hat{\mathbf{y}} = p(\mathbf{y}|C(\mathbf{X}))$ is the closest point in $C(\mathbf{X})$ to $\mathbf{y}$, $\mathbf{b}$ has another interpretation: it is the value of $\boldsymbol{\beta}$ that makes the linear combination $\mathbf{X}\boldsymbol{\beta}$ closest to $\mathbf{y}$. I.e., (for $\mathbf{X}$ of full rank) $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ minimizes

$$Q = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 \qquad \textbf{(the least squares criterion)}$$

Example – Typographical Errors: Shown below are the number of galleys for a manuscript (X_1) and the dollar cost of correcting typographical errors (Y) in a random sample of $n = 6$ recent orders handled by a firm specializing in technical manuscripts.

i	X_{i1}	Y_i
1	7	128
2	12	213
3	4	75
4	14	250
5	25	446
6	30	540

A scatterplot of Y versus X_1 appears below.



It is clear from the above scatterplot that the relationship between cost and manuscript length is nearly linear.

Suppose we try to approximate $\mathbf{y} = (128, \ldots, 540)^T$ by a linear function $b_0 + b_1\mathbf{x}_1$ where $\mathbf{x}_1 = (7, \ldots, 30)^T$. The problem is to find the values of (β_0, β_1) in the linear model

$$\mathbf{y} = \underbrace{\begin{pmatrix} 1 & 7 \\ 1 & 12 \\ 1 & 4 \\ 1 & 14 \\ 1 & 25 \\ 1 & 30 \end{pmatrix}}_{=\mathbf{X}} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{pmatrix}$$

$$\mathbf{y} = \beta_0 \underbrace{\mathbf{x}_0}_{=\mathbf{j}_6} + \beta_1\mathbf{x}_1 + \mathbf{e}.$$

Projecting $\mathbf{y}$ onto $\mathcal{L}(\mathbf{x}_0, \mathbf{x}_1) = C(\mathbf{X})$ produces $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ where $\mathbf{b}$ solves the normal equations

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}. \qquad (*)$$

That is, $\mathbf{b}$ minimizes

$$Q = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = ||\mathbf{y} - (\beta_0 + \beta_1\mathbf{x}_1)||^2 = ||\mathbf{e}||^2 = \sum_i^6 e_i^2.$$

- That is, $\mathbf{b}$ minimizes the sum of squared errors. Therefore, Q is called the **least squares criterion** and $\mathbf{b}$ is called the least squares estimator of $\boldsymbol{\beta}$.

$\mathbf{b}$ solves (*), where

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 6 & 92 \\ 92 & 1930 \end{pmatrix} \quad \text{and} \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} 1652 \\ 34602 \end{pmatrix}$$

In this example, $\mathbf{x}_0, \mathbf{x}_1$ are linearly independent so $\mathbf{X}$ has full rank (equal to 2) and $\mathbf{X}^T\mathbf{X}$ is nonsingular with

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} .6194 & -.0295 \\ -.0295 & .0019 \end{pmatrix}$$

so

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 1.5969 \\ 17.8524 \end{pmatrix} \quad \hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \begin{pmatrix} 126.6 \\ 215.8 \\ 73.0 \\ 251.5 \\ 447.9 \\ 537.2 \end{pmatrix}$$

In addition,

$$\|\mathbf{y}\|^2 = 620,394, \quad \|\hat{\mathbf{y}}\|^2 = 620,366, \quad \|\mathbf{e}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = 28.0$$

so the Pythagorean Theorem holds.

Another Example - Gasoline Additives

Suppose that an experiment was conducted to compare the effects on octane for 2 different gasoline additives. For this purpose an investigator obtains 6 one-liter samples of gasoline and randomly divides these samples into 3 groups of 2 samples each. The groups are assigned to receive no additive (treatment C, for control), or 1 cc/liter of additives A, or B, and then octane measurements are made. The resulting data are as follows:

Treatment	Observations	
A	91.7	91.9
B	92.4	91.3
C	91.7	91.2

Let $\mathbf{y}$ be the vector $\mathbf{y} = (y_{11}, y_{12}, y_{21}, y_{22}, y_{31}, y_{32})^T$, where y_{ij} is the response (octane) for the j^{th} sample receiving the i^{th} treatment.

Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ be indicators for the three treatments. That is, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ correspond to the columns of the model matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

The best approximation to $\mathbf{y}$ by a vector in $V = \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = C(\mathbf{X})$ in the least squares sense is

$$\hat{\mathbf{y}} = p(\mathbf{y}|V) = \sum_{i=1}^{3} p(\mathbf{y}|\mathbf{x}_i) = \sum_{i=1}^{3} \bar{y}_i \mathbf{x}_i,$$

where $\bar{y}_i = (1/2)(y_{i1} + y_{i2})$, the mean of the values in the i^{th} treatment.

- The second equality above follows from the orthogonality and linear independence of $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ ($\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ form an orthogonal basis for V).

Easy computations lead to

$$\bar{y}_1 = 91.80, \quad \bar{y}_2 = 91.85, \quad \bar{y}_3 = 91.45,$$

so

$$\hat{\mathbf{y}} = 91.80 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + 91.85 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + 91.45 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 91.80 \\ 91.80 \\ 91.85 \\ 91.85 \\ 91.45 \\ 91.45 \end{pmatrix}$$

It is easily verified that the error sum of squares is $\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = 0.75$ and $\|\hat{\mathbf{y}}\|^2 = 50,453.53$ which sum to $\|\mathbf{y}\|^2 = 50,454.28$.

Projection Matrices.

Definition: $\mathbf{P}$ is a (orthogonal) projection matrix onto V if and only if

i. $\mathbf{v} \in V$ implies $\mathbf{Pv} = \mathbf{v}$ (projection); and
ii. $\mathbf{w} \perp V$ implies $\mathbf{Pw} = \mathbf{0}$ (orthogonality).

- For $\mathbf{X}$ an $n \times k$ matrix of full rank, it is not hard to show that $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ satisfies this definition where $V = C(\mathbf{X})$, and is therefore a projection matrix onto $C(\mathbf{X})$. Perhaps simpler though, is to use the following theorem:

Theorem: $\mathbf{P}$ is a projection matrix onto its column space $C(\mathbf{P}) \subset \mathcal{R}^n$ if and only if

i. $\mathbf{PP} = \mathbf{P}$ (it is idempotent), and
ii. $\mathbf{P} = \mathbf{P}^T$ (it is symmetric).

Proof: First, the $\Rightarrow$ part: Choose any two vectors $\mathbf{w}, \mathbf{z} \in \mathcal{R}^n$. $\mathbf{w}$ can be written $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$ where $\mathbf{w}_1 \in C(\mathbf{P})$ and $\mathbf{w}_2 \perp C(\mathbf{P})$ and $\mathbf{z}$ can be decomposed similarly as $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2$. Note that

$$(\mathbf{I} - \mathbf{P})\mathbf{w} = \mathbf{w} - \mathbf{Pw} = (\mathbf{w}_1 + \mathbf{w}_2) - \underbrace{\mathbf{Pw}_1}_{=\mathbf{w}_1} - \underbrace{\mathbf{Pw}_2}_{=\mathbf{0}} = \mathbf{w}_2$$

and

$$\mathbf{Pz} = \mathbf{Pz}_1 + \mathbf{Pz}_2 = \mathbf{Pz}_1 = \mathbf{z}_1,$$

so

$$\mathbf{0} = \mathbf{z}_1^T\mathbf{w}_2 = (\mathbf{Pz})^T(\mathbf{I} - \mathbf{P})\mathbf{w}.$$

We've shown that $\mathbf{z}^T\mathbf{P}^T(\mathbf{I} - \mathbf{P})\mathbf{w} = \mathbf{0}$ for any $\mathbf{w}, \mathbf{z} \in \mathcal{R}^n$, so it must be true that $\mathbf{P}^T(\mathbf{I} - \mathbf{P}) = \mathbf{0}$ or $\mathbf{P}^T = \mathbf{P}^T\mathbf{P}$. Since $\mathbf{P}^T\mathbf{P}$ is symmetric, $\mathbf{P}^T$ must also be symmetric, and this implies $\mathbf{P} = \mathbf{PP}$.

Second, the $\Leftarrow$ part: Assume $\mathbf{PP} = \mathbf{P}$ and $\mathbf{P} = \mathbf{P}^T$ and let $\mathbf{v} \in C(\mathbf{P})$ and $\mathbf{w} \perp C(\mathbf{P})$. $\mathbf{v} \in C(\mathbf{P})$ means that we must be able to write $\mathbf{v}$ as $\mathbf{Pb}$ for some vector $\mathbf{b}$. So $\mathbf{v} = \mathbf{Pb}$ implies $\mathbf{Pv} = \mathbf{PPb} = \mathbf{Pb} = \mathbf{v}$ (part i. proved). Now $\mathbf{w} \perp C(\mathbf{P})$ means that $\mathbf{P}^T\mathbf{w} = \mathbf{0}$, so $\mathbf{Pw} = \mathbf{P}^T\mathbf{w} = \mathbf{0}$ (proves part ii.). ∎

Now, for $\mathbf{X}$ an $n \times k$ matrix of full rank, $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is a projection matrix onto $C(\mathbf{P})$ because it is symmetric:

$$\mathbf{P}^T = \{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{P}$$

and idempotent:

$$\mathbf{PP} = \{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}\{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{P}.$$

Theorem: Projection matrices are unique.

Proof: Let $\mathbf{P}$ and $\mathbf{M}$ be projection matrices onto some space V. Let $\mathbf{x} \in \mathcal{R}^n$ and write $\mathbf{X} = \mathbf{v} + \mathbf{w}$ where $\mathbf{v} \in V$ and $\mathbf{w} \in V^\perp$. Then $\mathbf{Px} = \mathbf{Pv} + \mathbf{Pw} = \mathbf{v}$ and $\mathbf{Mx} = \mathbf{Mv} + \mathbf{Mw} = \mathbf{v}$. So, $\mathbf{Px} = \mathbf{Mx}$ for any $\mathbf{x} \in \mathcal{R}^n$. Therefore, $\mathbf{P} = \mathbf{M}$. ∎

- Later, we'll see that $C(\mathbf{P}) = C(\mathbf{X})$. This along with the uniqueness of projection matrices means that (in the $\mathbf{X}$ of full rank case) $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the (unique) projection matrix onto $C(\mathbf{X})$.

So, the projection matrix onto $C(\mathbf{X})$ can be obtained from $\mathbf{X}$ as $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Alternatively, the projection matrix onto a subspace can be obtained from an orthonormal basis for the subspace:

Theorem: Let $\mathbf{o}_1, \ldots, \mathbf{o}_k$ be an orthonormal basis for $V \in \mathcal{R}^n$, and let $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_k)$. Then $\mathbf{OO}^T = \sum_{i=1}^k \mathbf{o}_i\mathbf{o}_i^T$ is the projection matrix onto V.

Proof: $\mathbf{OO}^T$ is symmetric and $\mathbf{OO}^T\mathbf{OO}^T = \mathbf{OO}^T$ (idempotent); so, by the previous theorem $\mathbf{OO}^T$ is a projection matrix onto $C(\mathbf{OO}^T) = C(\mathbf{O}) = V$ (using property 8, p.15).∎

Some Examples:

1. The projection (matrix) onto the linear subspace of vectors of the form $(a, 0, b)^T$ (the subspace spanned by $(0, 0, 1)^T$ and $(1, 0, 0)^T$) is

$$\mathbf{P} = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{pmatrix}}_{=\mathbf{O}} \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

2. Let $\mathbf{j}_n$ represent the column vector containing n ones. The projection onto $\mathcal{L}(\mathbf{j}_n)$ is

$$\mathbf{P} = \frac{1}{\sqrt{n}}\mathbf{j}_n \left(\frac{1}{\sqrt{n}}\mathbf{j}_n\right)^T = \frac{1}{n}\mathbf{j}_n\mathbf{j}_n^T = n \times n \text{ matrix with all elements = to } n^{-1}$$

   Note that $\mathbf{P}\mathbf{x} = \bar{x}\mathbf{j}_n$, where $\bar{x} = \langle \mathbf{x}, \mathbf{j}_n \rangle / \|\mathbf{j}_n\|^2 = (\sum_{i=1}^n x_i)/n$.

Projections onto nested subspaces: Let V be a subspace of $\mathcal{R}^n$ and let V_0 be a subspace of V. Let $\mathbf{P}$ and $\mathbf{P}_0$ be the corresponding projection matrices. Then

$$(1) \quad \mathbf{P}\mathbf{P}_0 = \mathbf{P}_0, \quad \text{and} \quad (2) \quad \mathbf{P}_0\mathbf{P} = \mathbf{P}_0.$$

Proof: (1): letting $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y} \in V$ and $\hat{\mathbf{y}}_0 = \mathbf{P}_0\mathbf{y} \in V_0 \subset V$, then for $\mathbf{y} \in \mathcal{R}^n$, $\mathbf{P}(\mathbf{P}_0\mathbf{y}) = \mathbf{P}(\hat{\mathbf{y}}_0) = \hat{\mathbf{y}}_0$. Now (2) follows from the symmetry of projection matrices: Since $\mathbf{P}_0 = \mathbf{P}\mathbf{P}_0$ and $\mathbf{P}_0$ is symmetric it follows that $\mathbf{P}_0 = \mathbf{P}\mathbf{P}_0 = (\mathbf{P}\mathbf{P}_0)^T = \mathbf{P}_0^T\mathbf{P}^T = \mathbf{P}_0\mathbf{P}$.

Projection onto the orthogonal complement:

Theorem: Let $\mathbf{P}$ and $\mathbf{P}_0$ be projection matrices with $C(\mathbf{P}_0) \subset C(\mathbf{P})$. Then (i) $\mathbf{P}-\mathbf{P}_0$ is a projection matrix onto $C(\mathbf{P}-\mathbf{P}_0)$, and (ii) $C(\mathbf{P}-\mathbf{P}_0)$ is the orthogonal complement of $C(\mathbf{P}_0)$ with respect to $C(\mathbf{P})$.

Proof: First prove that $\mathbf{P}-\mathbf{P}_0$ is a projection matrix onto $C(\mathbf{P}-\mathbf{P}_0)$ by checking idempotency:

$$(\mathbf{P}-\mathbf{P}_0)(\mathbf{P}-\mathbf{P}_0) = \mathbf{P}\mathbf{P}-\mathbf{P}\mathbf{P}_0-\mathbf{P}_0\mathbf{P}+\mathbf{P}_0\mathbf{P}_0 = \mathbf{P}-\mathbf{P}_0-\mathbf{P}_0+\mathbf{P}_0 = \mathbf{P}-\mathbf{P}_0$$

and symmetry:

$$(\mathbf{P}-\mathbf{P}_0)^T = \mathbf{P}^T-\mathbf{P}_0^T = \mathbf{P}-\mathbf{P}_0.$$

Now prove that $C(\mathbf{P}-\mathbf{P}_0)$ is the orthogonal complement of $C(\mathbf{P}_0)$ with respect to $C(\mathbf{P})$: $C(\mathbf{P}-\mathbf{P}_0) \perp C(\mathbf{P}_0)$ because $(\mathbf{P}-\mathbf{P}_0)\mathbf{P}_0 = \mathbf{P}\mathbf{P}_0 - \mathbf{P}_0\mathbf{P}_0 = \mathbf{P}_0-\mathbf{P}_0 = \mathbf{0}$. Thus, $C(\mathbf{P}-\mathbf{P}_0)$ is contained in the orthogonal complement of $C(\mathbf{P}_0)$ with respect to $C(\mathbf{P})$. In addition, the orthogonal complement of $C(\mathbf{P}_0)$ with respect to $C(\mathbf{P})$ is contained in $C(\mathbf{P}-\mathbf{P}_0)$ because if $\mathbf{x} \in C(\mathbf{P})$ and $\mathbf{x} \in C(\mathbf{P}_0)^\perp$, then $\mathbf{x} = \mathbf{P}\mathbf{x} = (\mathbf{P}-\mathbf{P}_0)\mathbf{x}+\mathbf{P}_0\mathbf{x} = (\mathbf{P}-\mathbf{P}_0)\mathbf{x}$. ∎

- An important consequence of this theorem is that if $\mathbf{P}_V$ is the projection matrix onto a subspace V, then $\mathbf{I}-\mathbf{P}_V$ is the projection matrix onto the orthogonal complement of V.

Another Example:

3. If $V = \mathcal{L}(\mathbf{j}_n)$, then $V^\perp$ is the set of all $n-$vectors whose elements sum to zero.

   The projection onto $V^\perp$ is $\mathbf{P}_{V^\perp} = \mathbf{I}_n - \mathbf{P}_V = \mathbf{I}_n - \frac{1}{n}\mathbf{j}_n\mathbf{j}_n^T$. Note that $\mathbf{P}_{V^\perp}\mathbf{x}$ is the vector of deviations from the mean:

$$\mathbf{P}_{V^\perp}\mathbf{x} = (\mathbf{I}_n - \frac{1}{n}\mathbf{j}_n\mathbf{j}_n^T)\mathbf{x} = \mathbf{x} - \bar{x}\mathbf{j}_n = (x_1-\bar{x},\ldots,x_n-\bar{x})^T.$$

More on Subspaces:

Direct Sums: In a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, it is useful to decompose the sample space $\mathcal{R}^n$ into two orthogonal subspaces: the model subspace $C(\mathbf{X})$ and the error space $C(\mathbf{X})^{\perp}$. This provides a means of separating sources of variability in the response variable (variability explained by the model versus everything else).

In ANOVA, we go one-step farther and decompose the model space $C(\mathbf{X})$ into mutually orthogonal subspaces of $C(\mathbf{X})$. This provides a means for separating the influences of several explanatory factors on the response, which makes for better understanding.

- To undersand these ideas from the geometric viewpoint, we introduce the ideas of linear independence of subspaces and summation of vector spaces (called the direct sum).

Linear independence of subspaces: Subspaces $V_1, \ldots, V_k$ of $\mathcal{R}^n$ are linearly independent if for $\mathbf{x}_i \in V_i$, $i = 1, \ldots, k$, $\sum_{i=1}^{k} \mathbf{x}_i = \mathbf{0}$ implies that $\mathbf{x}_i = \mathbf{0}$ for all $i = 1, \ldots, k$.

- For a pair of subspaces V_i, V_j, $i \neq j$, the property $V_i \cap V_j = \{\mathbf{0}\}$ is equivalent to linear independence of V_i and V_j.

- However, for several subspaces $V_1, \ldots, V_k$, pairwise linear independence ($V_i \cap V_j = \{\mathbf{0}\}$ for each $i \neq j \in \{1, \ldots, k\}$) does not imply linear independence of the collection $V_1, \ldots, V_k$.

Direct sum: Let $V_1, \ldots, V_k$ be subspaces of $\mathcal{R}^n$. Then

$$V = \left\{ \mathbf{x} \;\middle|\; \mathbf{x} = \sum_{i=1}^{k} \mathbf{x}_i, \mathbf{x}_i \in V_i, i = 1, \ldots, k \right\}$$

is called the **direct sum** of $V_1, \ldots, V_k$, and is denoted by

$$V = V_1 + \cdots + V_k.$$

If these subspaces are linearly independent, then we will write the direct sum as

$$V = V_1 \oplus \cdots \oplus V_k$$

to indicate the linear independence of $V_1, \ldots, V_k$.

- Note that for any subspace $V \in \mathcal{R}^n$, $V^\perp$ is a subspace. In addition, since $V^\perp \cap V = \{\mathbf{0}\}$, $V^\perp$ and V are linearly independent and $\mathcal{R}^n = V \oplus V^\perp$.

Theorem The representation $\mathbf{x} = \sum_{i=1} \mathbf{x}_i$ for $\mathbf{x}_i \in V_i$ of elements $\mathbf{x} \in V = V_1 + \cdots + V_k$ is unique if and only if the subspaces are linearly independent.

Proof: (First the $\Leftarrow$ part:) Suppose that these subspaces are linearly independent and suppose that $\mathbf{x}$ has two such representations: $\mathbf{x} = \sum_{i=1}^k \mathbf{x}_i = \sum_{i=1}^k \mathbf{w}_i$ for $\mathbf{x}_i, \mathbf{w}_i \in V_i$, $i = 1, \ldots, k$. Then $\sum_{i=1}^k (\mathbf{x}_i - \mathbf{w}_i) = \mathbf{0}$, which by linear independence implies $\mathbf{x}_i - \mathbf{w}_i = \mathbf{0}$ or $\mathbf{x}_i = \mathbf{w}_i$ for each i.

(Next the $\Rightarrow$ part:) Suppose the representation is unique, let $\mathbf{v}_i \in V_i$, $i = 1, \ldots, k$, and let $\sum_{i=1}^k \mathbf{v}_k = \mathbf{0}$. Since $\mathbf{0} \in V_i$ for each i, and $\mathbf{0} = \mathbf{0} + \cdots + \mathbf{0}$, it follows that $\mathbf{v}_i = \mathbf{0}$ for each i (uniqueness), implying the linear independence of $V_1, \ldots, V_k$. ∎

Theorem: If $\{\mathbf{v}_{i1}, \ldots, \mathbf{v}_{in_i}\}$ is a basis for V_i, for each $i = 1, \ldots, k$, and $V_1, \ldots, V_k$ are linearly independent, then $\{\mathbf{v}_{11}, \ldots, \mathbf{v}_{1n_1}, \ldots, \mathbf{v}_{k1}, \ldots, \mathbf{v}_{kn_k}\}$ is a basis for $V = V_1 \oplus \cdots \oplus V_k$.

Proof: Omitted.

Corollary: If $V = V_1 \oplus \cdots \oplus V_k$, then

$$\dim(V) = \dim(V_1) + \cdots + \dim(V_k).$$

Proof: Omitted.

Theorem For any subspace $V \subset \mathcal{R}^n$ and any $\mathbf{x} \in \mathcal{R}^n$, there exist unique elements $\mathbf{x}_1, \mathbf{x}_2$ such that $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, where $\mathbf{x}_1 \in V$ and $\mathbf{x}_2 \in V^\perp$.

Proof: For existence take $\mathbf{x}_1 = p(\mathbf{x}|V)$ (which we know exists) and then take $\mathbf{x}_2 = \mathbf{x} - \mathbf{x}_1$. Uniqueness follows from the linear independence of V and $V^\perp$ and the Theorem at the top of the page.

Corollary: $\mathbf{x}_1 = p(\mathbf{x}|V)$ and $\mathbf{x}_2 = p(\mathbf{x}|V^\perp)$.

Proof: Follows from the identities $\mathbf{x} = p(\mathbf{x}|V) + [\mathbf{x} - p(\mathbf{x}|V)]$, $\mathbf{x} = [\mathbf{x} - p(\mathbf{x}|V^\perp)] + p(\mathbf{x}|V^\perp)$ and the uniqueness of the decomposition in the previous Theorem. ∎

Example: Consider $\mathcal{R}^4$, the space of 4-component vectors. Let

$$\mathbf{x}_1 = \mathbf{j}_4 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

and let

$$V_1 = \mathcal{L}(\mathbf{x}_1) = \{ \begin{pmatrix} a \\ a \\ a \\ a \end{pmatrix} | a \in \mathcal{R} \} \quad V_2 = \mathcal{L}(\mathbf{x}_2, \mathbf{x}_3) = \{ \begin{pmatrix} b+c \\ b+c \\ b \\ 0 \end{pmatrix} | b, c \in \mathcal{R} \}$$

Notice that V_1 and V_2 are linearly independent, so that V defined to be

$$V = V_1 \oplus V_2 = \{ \begin{pmatrix} a+b+c \\ a+b+c \\ a+b \\ a \end{pmatrix} | a, b, c \text{ real numbers} \}$$

has dimension $1 + 2 = 3$.

The orthogonal complements are

$$V_1^\perp = \{ \begin{pmatrix} a \\ b \\ c \\ -a-b-c \end{pmatrix} | a, b, c \in \mathcal{R} \} \quad V_2^\perp = \{ \begin{pmatrix} a \\ -a \\ 0 \\ b \end{pmatrix} | a, b \in \mathcal{R} \}$$

$$V^\perp = \{ \begin{pmatrix} a \\ -a \\ 0 \\ 0 \end{pmatrix} | a \in \mathcal{R} \}$$

- In general, for $V = V_1 \oplus V_2$, the relationship $P_V = P_{V_1} + P_{V_2}$ holds only if $V_1 \perp V_2$. They are not orthogonal in this example. Verify that $P_V = P_{V_1} + P_{V_2}$ doesn't hold here by computing $p(\mathbf{y}|V)$, $p(\mathbf{y}|V_1)$ and $p(\mathbf{y}|V_2)$ for some $\mathbf{y}$ (e.g., $\mathbf{y} = (1, 2, 3, 4)^T$).

Earlier, we established that a projection onto a subspace can be accomplished by projecting onto orthogonal basis vectors for that subspace and summing.

In the following theorem we establish that a projection onto a subspace V can be accomplished by summing the projections onto mutually orthogonal subspaces whose sum is V.

Theorem: Let $V_1, \ldots, V_k$ be mutually orthogonal subspaces of $\mathcal{R}^n$. Let $V = V_1 \oplus \cdots \oplus V_k$. Then $p(\mathbf{y}|V) = \sum_{i=1}^{k} p(\mathbf{y}|V_i)$, for all $\mathbf{y} \in \mathcal{R}^n$.

Proof: Let $\hat{\mathbf{y}}_i = p(\mathbf{y}|V_i)$. To show that $\sum_i \hat{\mathbf{y}}_i$ is the projection of $\mathbf{y}$ onto V, we must show that for each $\mathbf{x} \in V$, $\langle \mathbf{y}, \mathbf{x} \rangle = \langle \sum_i \hat{\mathbf{y}}_i, \mathbf{x} \rangle$. Since $\mathbf{x} \in V$, $\mathbf{x}$ can be written as $\mathbf{x} = \sum_{j=1}^{k} \mathbf{x}_j$ for some $\mathbf{x}_1 \in V_1, \mathbf{x}_2 \in V_2, \ldots, \mathbf{x}_k \in V_k$. Thus,

$$\begin{aligned}
\langle \sum_i \hat{\mathbf{y}}_i, \mathbf{x} \rangle &= \langle \sum_i \hat{\mathbf{y}}_i, \sum_j \mathbf{x}_j \rangle = \sum_{i=1}^{k} \sum_{j=1}^{k} \langle \hat{\mathbf{y}}_i, \mathbf{x}_j \rangle \\
&= \sum_i \langle \hat{\mathbf{y}}_i, \mathbf{x}_i \rangle + \underbrace{\sum_{\substack{i,j \\ i \neq j}} \langle \hat{\mathbf{y}}_i, \mathbf{x}_j \rangle}_{=0} \overset{*}{=} \sum_{i=1}^{k} \langle \mathbf{y}, \mathbf{x}_i \rangle \\
&= \langle \mathbf{y}, \sum_{i=1}^{k} \mathbf{x}_i \rangle = \langle \mathbf{y}, \mathbf{x} \rangle
\end{aligned}$$

($*$ since $\hat{\mathbf{y}}_i$ is the projection onto V_i).■

For $V \subset \mathcal{R}^n$, $\mathcal{R}^n$ can always be decomposed into V and $V^\perp$ so that $\mathcal{R}^n = V \oplus V^\perp$.

The previous theorem tells us that any $\mathbf{y} \in \mathcal{R}^n$ can be decomposed as

$$\mathbf{y} = p(\mathbf{y}|V) + p(\mathbf{y}|V^\perp),$$

or since $\mathbf{y} = \mathbf{I}_n\mathbf{y} = \mathbf{P}_{\mathcal{R}^n}\mathbf{y}$,

$$\mathbf{P}_{\mathcal{R}^n}\mathbf{y} = \mathbf{P}_V\mathbf{y} + \mathbf{P}_{V^\perp}\mathbf{y}$$

for all $\mathbf{y}$. Because this is true for all $\mathbf{y}$, it follows that

$$\begin{aligned} \underbrace{\mathbf{P}_{\mathcal{R}^n}}_{=\mathbf{I}} &= \mathbf{P}_V + \mathbf{P}_{V^\perp} \\ \Rightarrow \mathbf{P}_{\mathbf{V}^\perp} &= \mathbf{I} - \mathbf{P}_V \end{aligned}$$

More generally, consider a subspace V_0 that is a proper subset of a subspace V (i.e., $V_0 \subset V$ and is not equal to V, where V is not necessarily $\mathcal{R}^n$). Then V can be decomposed as $V = V_0 \oplus V_1$ where $V_1 = V_0^\perp \cap V$ is the orthogonal complement of V_0 w.r.t. V, and

$$\mathbf{P}_V = \mathbf{P}_{V_0} + \mathbf{P}_{V_1}.$$

Rearranging, this result can be stated as

$$\mathbf{P}_{V_0} = \mathbf{P}_V - \mathbf{P}_{V_0^\perp \cap V}$$

or

$$\mathbf{P}_{V_0^\perp \cap V} = \mathbf{P}_V - \mathbf{P}_{V_0}$$

which are (sometimes) more useful forms of the same relationship.

Example: One-way ANOVA, Effects Version:

Consider the model at the top of p. 6:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Let $\mathbf{x}_1, \ldots, \mathbf{x}_4$ be the columns of $\mathbf{X}$. The sample space $\mathcal{R}^n$ can be decomposed as

$$\mathcal{R}^6 = C(\mathbf{X}) \oplus C(\mathbf{X})^{\perp}.$$

- $C(\mathbf{X})$ is called the model space and $C(\mathbf{X})^{\perp}$ the error space.

The model space $C(\mathbf{X})$ can be decomposed further as

$$C(\mathbf{X}) = \mathcal{L}(\mathbf{x}_1) + \mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$$

- Notice that we use a '+' rather than a '$\oplus$' because $\mathcal{L}(\mathbf{x}_1)$, $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ are not LIN. Why? Because they intersect:

$$\mathcal{L}(\mathbf{x}_1) = \left\{ \begin{pmatrix} a \\ a \\ a \\ a \\ a \\ a \end{pmatrix} \middle| a \in \mathcal{R} \right\}, \quad \mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = \left\{ \begin{pmatrix} b \\ b \\ c \\ c \\ d \\ d \end{pmatrix} \middle| b, c, d \in \mathcal{R} \right\}.$$

  These spaces intersect when $b = c = d$.

- Alternatively, we can check the definition: Does $\mathbf{v} + \mathbf{w} = \mathbf{0}$ imply $\mathbf{v} = \mathbf{w} = \mathbf{0}$ for $\mathbf{v} \in \mathcal{L}(\mathbf{x}_1), \mathbf{w} \in \mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$? No, because we could take $\mathbf{v} = (a, a, a, a, a, a)^T$ and $\mathbf{w} = (-a, -a, -a, -a, -a, -a)^T$ for $a \neq 0$.

A more useful decomposition of the model space is into LIN subspaces. For example,

$$C(\mathbf{X}) = \mathcal{L}(\mathbf{x}_1) \oplus (\mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \cap \mathcal{L}(\mathbf{x}_1)^{\perp}).$$

- Here, $(\mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \cap \mathcal{L}(\mathbf{x}_1)^{\perp})$ is the orthogonal complement of $\mathcal{L}(\mathbf{x}_1)$ with respect to $C(\mathbf{X})$. It can be represented as

$$\left\{ \begin{pmatrix} b \\ b \\ c \\ c \\ -(b+c) \\ -(b+c) \end{pmatrix} \middle| b, c \in \mathcal{R} \right\}.$$

Since $\mathcal{L}(\mathbf{x}_1)$ and $(\mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \cap \mathcal{L}(\mathbf{x}_1)^{\perp})$ are LIN, orthogonal and direct sum to give $C(\mathbf{X})$, the projection of $\mathbf{y}$ onto $C(\mathbf{X})$ (that is, fitting the model) can be done as

$$\hat{\mathbf{y}} = p(\mathbf{y}|C(\mathbf{X})) = p(\mathbf{y}|\mathcal{L}(\mathbf{x}_1)) + p(\mathbf{y}|(\mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \cap \mathcal{L}(\mathbf{x}_1)^{\perp})),$$

where

$$p(\mathbf{y}|\mathcal{L}(\mathbf{x}_1)) = \bar{y}_{..}\mathbf{x}_1 = (\bar{y}_{..}, \bar{y}_{..}, \bar{y}_{..}, \bar{y}_{..}, \bar{y}_{..}, \bar{y}_{..})^T$$

and

$$\begin{aligned} p(\mathbf{y}|(\mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \cap \mathcal{L}(\mathbf{x}_1)^{\perp})) &= \mathbf{P}_{\mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)}\mathbf{y} - \mathbf{P}_{\mathcal{L}(\mathbf{x}_1)}\mathbf{y} \\ &= \bar{y}_{1.}\mathbf{x}_2 + \bar{y}_{2.}\mathbf{x}_3 + \bar{y}_{3.}\mathbf{x}_4 - \bar{y}_{..}\mathbf{x}_1 \\ &= \bar{y}_{1.}\mathbf{x}_2 + \bar{y}_{2.}\mathbf{x}_3 + \bar{y}_{3.}\mathbf{x}_4 - \bar{y}_{..}(\mathbf{x}_2 + \mathbf{x}_3 + \mathbf{x}_4) \\ &= \sum_{i=1}^{3} (\bar{y}_{i.} - \bar{y}_{..})\mathbf{x}_{i+1} \\ &= \begin{pmatrix} \bar{y}_{1.} - \bar{y}_{..} \\ \bar{y}_{1.} - \bar{y}_{..} \\ \bar{y}_{2.} - \bar{y}_{..} \\ \bar{y}_{2.} - \bar{y}_{..} \\ \bar{y}_{3.} - \bar{y}_{..} \\ \bar{y}_{3.} - \bar{y}_{..} \end{pmatrix} \end{aligned}$$

Since $\mathcal{R}^6 = \mathcal{L}(\mathbf{x}_1) \oplus (\mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \cap \mathcal{L}(\mathbf{x}_1)^\perp) \oplus C(\mathbf{X})^\perp$ we have

$$\begin{aligned}
\mathbf{P}_{\mathcal{R}^6} &= \mathbf{P}_{\mathcal{L}(\mathbf{x}_1)} + \mathbf{P}_{\mathcal{L}(\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4)\cap\mathcal{L}(\mathbf{x}_1)^\perp} + \mathbf{P}_{C(\mathbf{X})^\perp} \\
\mathbf{P}_{\mathcal{R}^6}\mathbf{y} &= \mathbf{P}_{\mathcal{L}(\mathbf{x}_1)}\mathbf{y} + \mathbf{P}_{\mathcal{L}(\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4)\cap\mathcal{L}(\mathbf{x}_1)^\perp}\mathbf{y} + \mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{y} \\
\mathbf{y} &= \underbrace{\hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2}_{=\hat{\mathbf{y}}} + \mathbf{e}
\end{aligned}$$

where $\hat{\mathbf{y}}_1 = p(\mathbf{y}|\mathcal{L}(\mathbf{x}_1))$, $\hat{\mathbf{y}}_2 = p(\mathbf{y}|(\mathcal{L}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \cap \mathcal{L}(\mathbf{x}_1)^\perp))$, $\hat{\mathbf{y}} = p(\mathbf{y}|C(\mathbf{X}))$, and $\mathbf{e} = p(\mathbf{y}|C(\mathbf{X})^\perp) = \mathbf{y} - \hat{\mathbf{y}}$.

The Pythagorean Theorem yields

$$||\mathbf{y}||^2 = \underbrace{||\hat{\mathbf{y}}_1||^2 + ||\hat{\mathbf{y}}_2||^2}_{=||\hat{\mathbf{y}}||^2} + ||\mathbf{e}||^2$$

which is usually rearranged to yield the following decomposition of the corrected total sum of squares in the one-way anova:

$$||\mathbf{y}||^2 - ||\hat{\mathbf{y}}_1||^2 = ||\hat{\mathbf{y}}_2||^2 + ||\mathbf{e}||^2 \qquad (*)$$

where

$$\begin{aligned}
||\mathbf{y}||^2 &= \sum_i \sum_j y_{ij}^2 \qquad ||\hat{\mathbf{y}}_1||^2 = \sum_i \sum_j \bar{y}_{\cdot\cdot}^2 \\
||\hat{\mathbf{y}}_2||^2 &= \sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \qquad ||\mathbf{e}||^2 = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2.
\end{aligned}$$

so that (*) becomes

$$\underbrace{\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2}_{=SS_{Total}} = \underbrace{\sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}_{SS_{Trt}} + \underbrace{\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2}_{=SS_E}.$$

Eigenvalues and Eigenvectors

Eigenvalues and Eigenvectors: A square $n \times n$ matrix $\mathbf{A}$ is said to have an *eigenvalue* λ, with a corresponding *eigenvector* $\mathbf{v} \neq \mathbf{0}$ if

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}. \tag{1}$$

Since (1) can be written

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$$

for some nonzero vector $\mathbf{v}$, this implies that the columns of $(\mathbf{A} - \lambda\mathbf{I})$ are linearly dependent $\Leftrightarrow |\mathbf{A} - \lambda\mathbf{I}| = 0$. So, the eigenvalues of $\mathbf{A}$ are the solutions of

$$|\mathbf{A} - \lambda\mathbf{I}| = 0 \qquad \textbf{(characteristic equation)}$$

If we look at how a determinant is calculated it is not difficult to see that $|\mathbf{A} - \lambda\mathbf{I}|$ will be a polynomial in λ of order n so there will be n (not necessarily real, and not necessarily distinct) eigenvalues (solutions).

- if $|\mathbf{A}| = 0$ then $\mathbf{A}\mathbf{v} = \mathbf{0}$ for some nonzero $\mathbf{v}$. That is, if the columns of $\mathbf{A}$ are linearly dependent then $\lambda = 0$ is an eigenvalue of $\mathbf{A}$.
- The eigenvector associated with a particular eigenvalue is unique only up to a scale factor. That is, if $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, then $\mathbf{A}(c\mathbf{v}) = \lambda(c\mathbf{v})$ so $\mathbf{v}$ and $c\mathbf{v}$ are both eigenvectors for $\mathbf{A}$ corresponding to λ. We typically normalize eigenvectors to have length 1 (choose the $\mathbf{v}$ that has the property $\|\mathbf{v}\| = \sqrt{\mathbf{v}^T\mathbf{v}} = 1$).
- We will mostly be concerned with eigen-pairs of symmetric matrices. For $\mathbf{A}_{n\times n}$ symmetric, there exist n not necessarily distinct but real eigenvalues.
- If λ_i and λ_j are distinct eigenvalues of the symmetric matrix $\mathbf{A}$ then their associated eigenvectors $\mathbf{v}_i$, $\mathbf{v}_j$, are orthogonal.
- If there exist exactly k LIN vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ corresponding to the same eigenvalue λ, then λ is said to have *multiplicity* k.
- In this case all vectors in $\mathcal{L}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ are eigenvectors for λ, and k orthogonal vectors from this subspace can be chosen as the eigenvectors of λ.

Computation: By computer typically, but for 2×2 case we have

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \Rightarrow \quad \mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix}$$

$$|\mathbf{A} - \lambda\mathbf{I}| = 0 \quad \Rightarrow$$

$$(a-\lambda)(d-\lambda) - bc = \lambda^2 - (a+d)\lambda + (ad-bc) = 0$$

$$\Rightarrow \quad \lambda = \frac{(a+d) \pm \sqrt{(a+d)^2 - 4(ad-bc)}}{2}$$

To obtain associated eigenvector solve $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. There are infinitely many solutions for $\mathbf{v}$ so choose one by setting $v_1 = 1$, say, and solve for v_2. Normalize to have length one by computing $(1/\|\mathbf{v}\|)\mathbf{v}$.

Orthogonal Matrices: We say that $\mathbf{A}$ is orthogonal if $\mathbf{A}^T = \mathbf{A}^{-1}$ or, equivalently, $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$, so that the columns (and rows) of $\mathbf{A}$ all have length 1 and are mutually orthogonal.

Spectral Decomposition: If $\mathbf{A}_{n\times n}$ is a symmetric matrix then it can be written (decomposed) as follows:

$$\mathbf{A} = \mathbf{U}_{n\times n}\Lambda_{n\times n}\mathbf{U}^T_{n\times n},$$

where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

and $\mathbf{U}$ is an orthogonal matrix with columns $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ the (normalized) eigenvectors corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$.

- We haven't yet talked about how to interpret eigenvalues and eigenvectors, but the spectral decomposition says something about the significance of these quantities: the "information" in $\mathbf{A}$ can be broken apart into its eigenvalues and eigenvectors.

An equivalent representation of the spectral decomposition is

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^T = \underbrace{\begin{pmatrix} \lambda_1\mathbf{u}_1 & \dots & \lambda_n\mathbf{u}_n \end{pmatrix}}_{=\mathbf{U}\Lambda} \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^T.$$

Note that $\mathbf{u}_i\mathbf{u}_i^T = \mathbf{P}_i$ $i = 1, \dots, n$, are projections onto the one-dimensional subspaces $\mathcal{L}(\mathbf{u}_i)$.

So, if the eigenvalues of $\mathbf{A}$ are all distinct, we have

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{P}_i$$

or, if $\mathbf{A}$ has r distinct eigenvalues $\lambda_1, \dots, \lambda_r$ with multiplicities $k_1, \dots, k_r$, then

$$\mathbf{A} = \sum_{i=1}^{r} \lambda_i \mathbf{P}_i^*$$

where $\mathbf{P}_i^*$ is the projection onto the k_j-dimensional subspace spanned by the eigenvectors corresponding to λ_i.

Results:

1. Recall that the *trace* of a matrix is the sum of its diagonal elements. Both the trace and the *determinant* of a matrix give scalar-valued measures of the size of a matrix.

   Using the spectral decomposition, it is easy to show that for a $n \times n$ symmetric matrix $\mathbf{A}$,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad |\mathbf{A}| = \prod_{i=1}^{n} \lambda_i$$

2. If $\mathbf{A}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, then $\mathbf{A}^{-1}$ has the same associated eigenvectors and eigenvalues $1/\lambda_1, \ldots, 1/\lambda_n$ since

$$\mathbf{A}^{-1} = (\mathbf{U}\Lambda\mathbf{U}^T)^{-1} = \mathbf{U}\Lambda^{-1}\mathbf{U}^T.$$

3. Similarly, if $\mathbf{A}$ has the additional property that it is **positive definite** (defined below), then a **square root matrix**, $\mathbf{A}^{1/2}$, can be obtained with the property $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$:

   $\mathbf{A}^{1/2} = \mathbf{U}\Lambda^{1/2}\mathbf{U}^T$, where

$$\Lambda^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{pmatrix}$$

   $\mathbf{A}^{1/2}$ is symmetric, has eigenvalues that are the square roots of the eigenvalues of $\mathbf{A}$, and has the same associated eigenvectors as $\mathbf{A}$.

Quadratic Forms: For a symmetric matrix $\mathbf{A}_{n\times n}$, a quadratic form in $\mathbf{x}_{n\times 1}$ is defined by

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j.$$

($\mathbf{x}^T\mathbf{A}\mathbf{x}$ is a sum in squared (quadratic) terms, x_i^2, and $x_i x_j$ terms.)

- Quadratic forms are going to arise frequently in linear models as squared lengths of projections, or sums of squares.

The spectral decomposition can be used to "diagonalize" (in a certain sense) the matrix in a quadratic form. A quadratic from $\mathbf{x}^T\mathbf{A}\mathbf{x}$ in a symmetric matrix $\mathbf{A}$ can be written

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{U}\Lambda\mathbf{U}^T\mathbf{x} = (\mathbf{U}^T\mathbf{x})^T\Lambda\underbrace{(\mathbf{U}^T\mathbf{x})}_{=\mathbf{y}} = \mathbf{y}^T\underbrace{\Lambda}_{\text{diagonal}}\mathbf{y}.$$

Eigen-pairs of Projection Matrices: Suppose that $\mathbf{P}_V$ is a projection matrix onto a subspace $V \in \mathcal{R}^n$.

Then for any $\mathbf{x} \in V$ and any $\mathbf{y} \in V^\perp$,

$$\mathbf{P}_V\mathbf{x} = \mathbf{x} = (1)\mathbf{x}, \quad \text{and} \quad \mathbf{P}_V\mathbf{y} = \mathbf{0} = (0)\mathbf{y}.$$

Therefore, by definition (see (1) on p.47),

- All vectors in V are eigenvectors of $\mathbf{P}_V$ with eigenvalues 1, and
- All vectors in $V^\perp$ are eigenvectors of $\mathbf{P}_V$ with eigenvalues 0.
- The eigenvalue 1 has multiplicity $= \dim(V)$, and
- The eigenvalue 0 has multiplicity $= \dim(V^\perp) = n - \dim(V)$.
- In addition, since $\text{tr}(\mathbf{A}) = \sum_i \lambda_i$ it follows that $\text{tr}(\mathbf{P}_V) = \dim(V)$ (the trace of a projection matrix is the dimension of the space onto which it projects).
- In addition, $\text{rank}(\mathbf{P}_V) = \text{tr}(\mathbf{P}_V) = \dim(V)$.

Positive Definite Matrices: $\mathbf{A}$ is positive definite (p.d.) if $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$. If $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all nonzero $\mathbf{x}$ then $\mathbf{A}$ is **positive semi-definite** (p.s.d.). If $\mathbf{A}$ is p.d. or p.s.d. then it is said to be **non-negative definite** (n.n.d.).

- If $\mathbf{A}_{n\times n}$ is p.d. then for $i = 1, \ldots, n$, $\mathbf{u}_i^T\mathbf{A}\mathbf{u}_i = \lambda_i > 0$;
- i.e., the eigenvalues of $\mathbf{A}$ are all positive.
- It can also be shown that if the eigenvalues of $\mathbf{A}$ are all positive then $\mathbf{A}$ is p.d.

Example

The matrix

$$\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$$

is positive definite.

Q: Why?

A: Because the associated quadratic form is

$$\begin{aligned}\mathbf{x}^T\mathbf{A}\mathbf{x} &= \begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= 2x_1^2 - 2x_1x_2 + 3x_2^2 = 2(x_1 - \frac{1}{2}x_2)^2 + \frac{5}{2}x_2^2,\end{aligned}$$

which is clearly positive as long as x_1 and x_2 are not both 0.

The matrix

$$\mathbf{B} = \begin{pmatrix} 13 & -2 & -3 \\ -2 & 10 & -6 \\ -3 & -6 & 5 \end{pmatrix}$$

is positive semidefinite because its associated quadratic form is

$$\mathbf{x}^T\mathbf{B}\mathbf{x} = (2x_1 - x_2)^2 + (3x_1 - x_3)^2 + (3x_2 - 2x_3)^2,$$

which is always non-negative, but does equal 0 for $\mathbf{x} = (1, 2, 3)^T$ (or any multiple of $(1, 2, 3)^T$).

Some Results on p.d. and p.s.d. matrices:

1.
   a. If $\mathbf{A}$ is p.d., then all of its diagonal elements a_{ii} are positive.
   b. If $\mathbf{A}$ is p.s.d., then all $a_{ii} \geq 0$.

2. Let $\mathbf{M}$ be a nonsingular matrix.
   a. If $\mathbf{A}$ is p.d., then $\mathbf{M}^T\mathbf{A}\mathbf{M}$ is p.d.
   b. If $\mathbf{A}$ is p.s.d., then $\mathbf{M}^T\mathbf{A}\mathbf{M}$ is p.s.d.

3. Let $\mathbf{A}$ be a $p \times p$ p.d. matrix and let $\mathbf{B}$ be a $k \times p$ matrix of rank $k \leq p$. Then $\mathbf{B}\mathbf{A}\mathbf{B}^T$ is p.d.

4. Let $\mathbf{A}$ be a $p \times p$ p.d. matrix and let $\mathbf{B}$ be a $k \times p$ matrix. If $k > p$ or if $\text{rank}(\mathbf{B}) < \min(k, p)$ then $\mathbf{B}\mathbf{A}\mathbf{B}^T$ is p.s.d.

5. A symmetric matrix $\mathbf{A}$ is p.d. if and only if there exists a nonsingular matrix $\mathbf{M}$ such that $\mathbf{A} = \mathbf{M}^T\mathbf{M}$.

6. A p.d. matrix is nonsingular.

7. Let $\mathbf{B}$ be an $n \times p$ matrix.
   a. If $\text{rank}(\mathbf{B}) = p$, then $\mathbf{B}^T\mathbf{B}$ is p.d.
   b. If $\text{rank}(\mathbf{B}) < p$, then $\mathbf{B}^T\mathbf{B}$ is p.s.d.

8. If $\mathbf{A}$ is p.d., then $\mathbf{A}^{-1}$ is p.d.

9. If $\mathbf{A}$ is p.d. and is partitioned in the form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

   where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are square, then $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are both p.d.

Inverses of Partitioned Matrices: A very useful result is the following: Let $\mathbf{A}$ be a symmetric matrix that can be partitioned as above in result 9 (note this implies $\mathbf{A}_{21} = \mathbf{A}_{12}^T$). Then

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T & -\mathbf{C}\mathbf{B}^{-1} \\ -\mathbf{B}^{-1}\mathbf{C}^T & \mathbf{B}^{-1} \end{pmatrix}$$

where $\mathbf{B} = \mathbf{A}_{22} - \mathbf{A}_{12}^T\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$, and $\mathbf{C} = \mathbf{A}_{11}^{-1}\mathbf{A}_{12}$. More general results are available for $\mathbf{A}$ nonsymmetric (see, e.g., Ravishanker & Dey, 2002, *A First Course in Linear Model Theory*, p. 37).

Systems of Equations

The system of n (linear) equations in p unknowns,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p &= c_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p &= c_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{np}x_p &= c_n \end{aligned}$$

can be written in matrix form as

$$\mathbf{A}\mathbf{x} = \mathbf{c} \qquad (*)$$

where $\mathbf{A}$ is $n \times p$, $\mathbf{x}$ is $p \times 1$ and $\mathbf{c}$ is $n \times 1$.

- If $n \neq p$ then $\mathbf{x}$ and $\mathbf{c}$ are of different sizes.
- If $n = p$ and $\mathbf{A}$ is nonsingular, then there exists a unique solution vector $\mathbf{x}$ given by $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$.
- If $n > p$ so that $\mathbf{A}$ has more rows than columns, then (*) usually (but not always) does not have a solution.
- If $n < p$, so that $\mathbf{A}$ has fewer rows than columns, (*) usually (but not always) has an infinite number of solutions.

Consistency: If (*) has one or more solution vectors then it is said to be consistent. Systems without any solutions are said to be inconsistent.

We will most often be concerned with systems where $\mathbf{A}$ is square. Suppose $\mathbf{A}$ is $p \times p$ with $\text{rank}(\mathbf{A}) = r < p$.

Recall $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T)$ so that the rank of $\mathbf{A}$ (the number of linearly independent columns in $\mathbf{A}$) = the number of linearly independent columns of $\mathbf{A}^T$ = the number of linearly independent rows of $\mathbf{A}$.

Therefore, there are $r < p$ linearly independent rows of $\mathbf{A}$ which implies there exists a $\mathbf{b} \neq \mathbf{0}$ so that

$$\mathbf{A}^T\mathbf{b} = \mathbf{0}, \quad \text{or, equivalently,} \quad \mathbf{b}^T\mathbf{A} = \mathbf{0}^T.$$

Multiplying both sides of (*) by $\mathbf{b}^T$ we have

$$\begin{aligned} & \mathbf{b}^T\mathbf{A}\mathbf{x} = \mathbf{b}^T\mathbf{c} \\ \Rightarrow \quad & \mathbf{0}^T\mathbf{x} = \mathbf{b}^T\mathbf{c} \\ \Rightarrow \quad & \mathbf{b}^T\mathbf{c} = 0 \end{aligned}$$

Otherwise, if $\mathbf{b}^T\mathbf{c} \neq \mathbf{0}$, there is no $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \mathbf{c}$.

- Hence, in order for $\mathbf{A}\mathbf{x} = \mathbf{c}$ to be consistent, the same linear relationships, if any, that exist among the rows of $\mathbf{A}$ must exist among the rows (elements) of $\mathbf{c}$.
  - This idea is formalized by comparing the rank of $\mathbf{A}$ with the rank of the *augmented matrix* $[\mathbf{A}, \mathbf{c}]$.

Theorem: The system of equations $\mathbf{A}\mathbf{x} = \mathbf{c}$ has at least one solution vector (is consistent) if and only if $\text{rank}(\mathbf{A}) = \text{rank}([\mathbf{A}, \mathbf{c}])$.

Proof: Suppose $\text{rank}(\mathbf{A}) = \text{rank}([\mathbf{A}, \mathbf{c}])$. Then $\mathbf{c}$ is a linear combination of the columns of $\mathbf{A}$; that is, there exists some $\mathbf{x}$ so that we can write

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_p\mathbf{a}_p = \mathbf{c},$$

or, equivalently, $\mathbf{A}\mathbf{x} = \mathbf{c}$.

Now suppose there exists a solution vector $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \mathbf{c}$. In general, $\text{rank}(\mathbf{A}) \leq \text{rank}([\mathbf{A}, \mathbf{c}])$ (result 7, p.15). But since there exists an $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \mathbf{c}$, we have

$$\begin{aligned} \text{rank}([\mathbf{A}, \mathbf{c}]) &= \text{rank}([\mathbf{A}, \mathbf{A}\mathbf{x}]) = \text{rank}(\mathbf{A}[\mathbf{I}, \mathbf{x}]) \\ &\leq \text{rank}(\mathbf{A}) \quad \text{(by result 3, p.15)} \end{aligned}$$

and we have

$$\text{rank}(\mathbf{A}) \leq \text{rank}([\mathbf{A}, \mathbf{c}]) \leq \text{rank}(\mathbf{A})$$

so that $\text{rank}(\mathbf{A}) = \text{rank}([\mathbf{A}, \mathbf{c}])$. ∎

Generalized Inverses

Generalized Inverse: A *generalized inverse* of an $n \times k$ matrik $\mathbf{X}$ is defined to be any $k \times n$ matrix $\mathbf{X}^-$ that satisfies the condition

(1) $\mathbf{X}\mathbf{X}^-\mathbf{X} = \mathbf{X}$.

- Such a matrix always exists.
- Such a matrix is not unique.
- If $\mathbf{X}$ is nonsingular, then the generalized inverse of $\mathbf{X}$ is unique and is equal to $\mathbf{X}^{-1}$.

Example:

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 3 \\ 1 & 0 & 1 \\ 3 & 2 & 4 \end{pmatrix}.$$

The first two rows of $\mathbf{A}$ are linearly independent, but the third row is equal to the sum of the others, so $\text{rank}(\mathbf{A}) = 2$.

The matrices

$$\mathbf{A}_1^- = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{A}_2^- = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -\frac{3}{2} & \frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix}$$

are each generalized inverses of $\mathbf{A}$ since straight-forward matrix multiplication verifies that $\mathbf{A}\mathbf{A}_1^-\mathbf{A} = \mathbf{A}$ and $\mathbf{A}\mathbf{A}_2^-\mathbf{A} = \mathbf{A}$.

A matrix need not be square to have a generalized inverse. For example,

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

has generalized inverses $\mathbf{x}_1^- = (1, 0, 0, 0)$, $\mathbf{x}_2^- = (0, 1/2, 0, 0)$, $\mathbf{x}_3^- = (0, 0, 1/3, 0)$ and $\mathbf{x}_4^- = (0, 0, 0, 1/4)$. In each case, it is easily verified that

$$\mathbf{x}\mathbf{x}_i^-\mathbf{x} = \mathbf{x}, \qquad i = 1, \ldots, 4.$$

Let $\mathbf{X}$ be $n \times k$ of rank r, let $\mathbf{X}^-$ be a generalized inverse of $\mathbf{X}$, and let $\mathbf{G}$ and $\mathbf{H}$ be any two generalized inverse of $\mathbf{X}^T\mathbf{X}$. Then we have the following results concerning generalized inverses:

1. $\text{rank}(\mathbf{X}^-\mathbf{X}) = \text{rank}(\mathbf{X}\mathbf{X}^-) = \text{rank}(\mathbf{X}) = r$.

2. $(\mathbf{X}^-)^T$ is a generalized inverse of $\mathbf{X}^T$. Furthermore, if $\mathbf{X}$ is symmetric, then $(\mathbf{X}^-)^T$ is a generalized inverse of $\mathbf{X}$.

3. $\mathbf{X} = \mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{X}$.

4. $\mathbf{G}\mathbf{X}^T$ is a generalized inverse of $\mathbf{X}$.

5. $\mathbf{X}\mathbf{G}\mathbf{X}^T$ is symmetric, idempotent, and is invariant to the choice of $\mathbf{G}$; that is, $\mathbf{X}\mathbf{G}\mathbf{X}^T = \mathbf{X}\mathbf{H}\mathbf{X}^T$.

Proof:

1. By result 3 on rank (p.15), $\text{rank}(\mathbf{X}^-\mathbf{X}) \leq \min\{\text{rank}(\mathbf{X}^-), \text{rank}(\mathbf{X})\} \leq \text{rank}(\mathbf{X}) = r$. In addition, because $\mathbf{X}\mathbf{X}^-\mathbf{X} = \mathbf{X}$, we have $r = \text{rank}(\mathbf{X}) \leq \min\{\text{rank}(\mathbf{X}), \text{rank}(\mathbf{X}^-\mathbf{X})\} \leq \text{rank}(\mathbf{X}^-\mathbf{X})$. Putting these together we have $r \leq \text{rank}(\mathbf{X}^-\mathbf{X}) \leq r \Rightarrow \text{rank}(\mathbf{X}^-\mathbf{X}) = r$. We can show $\text{rank}(\mathbf{X}\mathbf{X}^-) = r$ similarly.

2. This follows immediately upon transposing both sides of the equation $\mathbf{X}\mathbf{X}^-\mathbf{X} = \mathbf{X}$.

3. For $\mathbf{v} \in \mathcal{R}^n$, let $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ where $\mathbf{v}_1 \in C(\mathbf{X})$ and $\mathbf{v}_2 \perp C(\mathbf{X})$. Let $\mathbf{v}_1 = \mathbf{X}\mathbf{b}$ for some vector $\mathbf{b} \in \mathcal{R}^n$. Then for any such $\mathbf{v}$,

   $$\mathbf{v}^T\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{v}_1^T\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{b}^T(\mathbf{X}^T\mathbf{X})\mathbf{G}(\mathbf{X}^T\mathbf{X}) = \mathbf{b}^T(\mathbf{X}^T\mathbf{X}) = \mathbf{v}^T\mathbf{X}.$$

   Since this is true for all $\mathbf{v} \in \mathcal{R}^n$, it follows that $\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{X}$, and since $\mathbf{G}$ is arbitrary and can be replaced by another generalized inverse $\mathbf{H}$, it follows that $\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{X} = \mathbf{X}$ as well.

4. Follows immediately from result 3.

5. Invariance: observe that for the arbitrary vector $\mathbf{v}$, above,

$$\mathbf{XGX}^T\mathbf{v} = \mathbf{XGX}^T\mathbf{Xb} = \mathbf{XHX}^T\mathbf{Xb} = \mathbf{XHX}^T\mathbf{v}.$$

Since this holds for any $\mathbf{v}$, it follows that $\mathbf{XGX}^T = \mathbf{XHX}^T$.
Symmetry: $(\mathbf{XGX}^T)^T = \mathbf{XG}^T\mathbf{X}^T$, but since $\mathbf{G}^T$ is a generalized inverse for $\mathbf{X}^T\mathbf{X}$, the invariance property implies $\mathbf{XG}^T\mathbf{X}^T = \mathbf{XGX}^T$.
Idempotency: $\mathbf{XGX}^T\mathbf{XGX}^T = \mathbf{XGX}^T$ from result 3. ∎

Result 5 says that $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ is symmetric and idempotent for any generalized inverse $(\mathbf{X}^T\mathbf{X})^-$. Therefore, $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ is the unique projection matrix onto $C(\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T)$.

In addition, $C(\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T) = C(\mathbf{X})$, because $\mathbf{v} \in C(\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T)$ $\Rightarrow \mathbf{v} \in C(\mathbf{X})$, and if $\mathbf{v} \in C(\mathbf{X})$ then there exists a $\mathbf{b} \in \mathcal{R}^n$ so that $\mathbf{v} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{Xb} \Rightarrow \mathbf{v} \in C(\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T)$.

We've just proved the following theorem:

Theorem: $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ is *the* projection matrix onto $C(\mathbf{X})$ (projection matrices are uniqe).

- Although a generalized inverse is not unique, this does not pose any particular problem in the theory of linear models, because we're mainly interested in using a generalized inverse to obtain the projection matrix $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ onto $C(\mathbf{X})$, which is unique.

A generalized inverse $\mathbf{X}^-$ of $\mathbf{X}$ which satisfies (1) and has the additional properties

(2) $\mathbf{X}^-\mathbf{XX}^- = \mathbf{X}^-$,
(3) $\mathbf{X}^-\mathbf{X}$ is symmetric, and
(4) $\mathbf{XX}^-$ is symmetric,

is unique, and is known as the **Moore-Penrose Inverse**, but we have little use for the Moore-Penrose inverse in this course.

- A generalized inverse of a symmetric matrix is not necessarily symmetric. However, it is true that a symmetric generalized inverse can always be found for a symmetric matrix. In this course, we'll generally assume that generalized inverses of symmetric matrices are symmetric.

Generalized Inverses and Systems of Equations:

A solution to a consistent system of equations can be expressed in terms of a generalized inverse.

Theorem: If the system of equations $\mathbf{Ax} = \mathbf{c}$ is consistent and if $\mathbf{A}^-$ is any generalized inverse of $\mathbf{A}$, then $\mathbf{x} = \mathbf{A}^-\mathbf{c}$ is a solution.

Proof: Since $\mathbf{AA}^-\mathbf{A} = \mathbf{A}$, we have

$$\mathbf{AA}^-\mathbf{Ax} = \mathbf{Ax}.$$

Substituting $\mathbf{Ax} = \mathbf{c}$ on both sides, we obtain

$$\mathbf{AA}^-\mathbf{c} = \mathbf{c}.$$

Writing this in the form $\mathbf{A}(\mathbf{A}^-\mathbf{c}) = \mathbf{c}$, we see that $\mathbf{A}^-\mathbf{c}$ is a solution to $\mathbf{Ax} = \mathbf{c}$. ∎

For consistent systems of equations with > 1 solution, different choices of $\mathbf{A}^-$ will yield different solutions of $\mathbf{Ax} = \mathbf{c}$.

Theorem: If the system of equations $\mathbf{Ax} = \mathbf{c}$ is consistent, then all possible solutions can be obtained in either of the following two ways:

i. Use a specific $\mathbf{A}^-$ in the equation $\mathbf{x} = \mathbf{A}^-\mathbf{c} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{h}$, combined with all possible values of the arbitrary vector $\mathbf{h}$.

ii. Use all possible values of $\mathbf{A}^-$ in the equation $\mathbf{x} = \mathbf{A}^-\mathbf{c}$.

Proof: See Searle (1982, *Matrix Algebra Useful for Statistics*, p.238).

An alternative necessary and sufficient condition for $\mathbf{Ax} = \mathbf{c}$ to be consistent instead of the $\text{rank}(\mathbf{A}) = \text{rank}([\mathbf{A}, \mathbf{c}])$ condition given before is

Theorem: The system of equations $\mathbf{Ax} = \mathbf{c}$ has a solution if and only if for *any* generalized inverse $\mathbf{A}^-$ of $\mathbf{A}$ it is true that

$$\mathbf{AA}^-\mathbf{c} = \mathbf{c}.$$

Proof: Suppose $\mathbf{Ax} = \mathbf{c}$ is consistent. Then $\mathbf{x} = \mathbf{A}^-\mathbf{c}$ is a solution. Therefore, we can multiply $\mathbf{c} = \mathbf{Ax}$ by $\mathbf{AA}^-$ to get

$$\mathbf{AA}^-\mathbf{c} = \mathbf{AA}^-\mathbf{Ax} = \mathbf{Ax} = \mathbf{c}.$$

Now suppose $\mathbf{AA}^-\mathbf{c} = \mathbf{c}$. Then we can multiply $\mathbf{x} = \mathbf{A}^-\mathbf{c}$ by $\mathbf{A}$ to obtain

$$\mathbf{Ax} = \mathbf{AA}^-\mathbf{c} = \mathbf{c}.$$

Hence a solution exists, namely $\mathbf{x} = \mathbf{A}^-\mathbf{c}$. ∎

The Cholesky Decomposition: Let $\mathbf{A}$ be a symmetric positive semi-definite matrix. There exist an infinite number of "square-root matrices"; that is, $n \times n$ matrices $\mathbf{B}$ such that

$$\mathbf{A} = \mathbf{B}^T\mathbf{B}.$$

- The matrix square root $\mathbf{A}^{1/2} = \mathbf{U}\Lambda^{1/2}\mathbf{U}^T$ obtained earlier based on spectral decomposition $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^T$ is the unique *symmetric* square root, but there are many other non-symmetric square roots.

In particular, there exists a unique *upper-triangular* matrix $\mathbf{B}$ so that this decomposition holds. This choice of $\mathbf{B}$ is called the *Cholesky factor* and the decomposition $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ for $\mathbf{B}$ upper-triangular is called the **Cholesky decomposition.**

Random Vectors and Matrices

Definitions:

Random Vector: A vector whose elements are random variables. E.g.,

$$\mathbf{x}_{k\times 1} = \begin{pmatrix} x_1 & x_2 & \cdots & x_k \end{pmatrix}^T,$$

where $x_1, \ldots, x_k$ are each random variables.

Random Matrix: A matrix whose elements are random variables. E.g., $\mathbf{X}_{n\times k} = (x_{ij})$, where $x_{11}, x_{12}, \ldots, x_{nk}$ are each random variables.

Expected Value: The expected value (population mean) of a random matrix (vector) is the matrix (vector) of expected values. For $\mathbf{X}_{n\times k}$,

$$\mathrm{E}(\mathbf{X}) = \begin{pmatrix} \mathrm{E}(x_{11}) & \mathrm{E}(x_{12}) & \cdots & \mathrm{E}(x_{1k}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(x_{n1}) & \mathrm{E}(x_{n2}) & \cdots & \mathrm{E}(x_{nk}) \end{pmatrix}.$$

- $\mathrm{E}(\mathbf{X})$ will often be denoted $\boldsymbol{\mu}_{\mathbf{X}}$ or just $\boldsymbol{\mu}$ when the random matrix (vector) for which $\boldsymbol{\mu}$ is the mean is clear from the context.
- Recall, for a univariate random variable X,

$$\mathrm{E}(X) = \begin{cases} \int_{-\infty}^{\infty} x f_X(x) dx & \text{if } X \text{ is continuous;} \\ \sum_{\text{all } x} x f_X(x) & \text{if } X \text{ is discrete.} \end{cases}$$

  Here, $f_X(x)$ is the probability density function of X in the continuous case, $f_X(x)$ is the probability function of X in the discrete case.

(Population) Variance-Covariance Matrix: For a random vector $\mathbf{x}_{k\times 1} = (x_1, x_2, \ldots, x_k)^T$, the matrix

$$\begin{pmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) & \cdots & \text{cov}(x_1, x_k) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) & \cdots & \text{cov}(x_2, x_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(x_k, x_1) & \text{cov}(x_k, x_2) & \cdots & \text{var}(x_k) \end{pmatrix}$$

$$\equiv \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{pmatrix}$$

is called the variance-covariance matrix of $\mathbf{x}$ and is denoted $\text{var}(\mathbf{x})$ or $\Sigma_{\mathbf{x}}$ or sometimes Σ when it is clear which random vector is being referred to.

- Note that the $\text{var}(\cdot)$ function takes a single argument which is a vector or scalar.
- The book uses the notation $\text{cov}(\mathbf{x})$ for the var-cov matrix of $\mathbf{x}$. This is not unusual, but I like to use $\text{var}(\cdot)$ when there's one argument, and $\text{cov}(\cdot, \cdot)$ when there are two.
- Recall: for a univariate random variable x_i with expected value μ_i,

$$\sigma_{ii} = \text{var}(x_i) = \text{E}[(x_i - \mu_i)^2]$$

- Recall: for univariate random variables x_i and x_j,

$$\sigma_{ij} = \text{cov}(x_i, x_j) = \text{E}[(x_i - \mu_i)(x_j - \mu_j)]$$

- $\text{var}(\mathbf{x})$ is symmetric because $\sigma_{ij} = \sigma_{ji}$.
- In terms of vector/matrix algebra, $\text{var}(\mathbf{x})$ has formula

$$\text{var}(\mathbf{x}) = \text{E}[(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T].$$

- If the random variables $x_1, \ldots, x_k$ in $\mathbf{x}$ are mutually independent, then $\text{cov}(x_i, x_j) = 0$, when $i \neq j$, and $\text{var}(\mathbf{x})$ is **diagonal** with $(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{kk})^T$ along the diagonal and zeros elsewhere.

(Population) Covariance Matrix: For random vectors $\mathbf{x}_{k\times 1} = (x_1, \ldots, x_k)^T$, and $\mathbf{y}_{n\times 1} = (y_1, \ldots, y_n)^T$ let $\sigma_{ij} = \text{cov}(x_i, y_j)$, $i = 1, \ldots, k$, $j = 1, \ldots, n$. The matrix

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kn} \end{pmatrix} = \begin{pmatrix} \text{cov}(x_1, y_1) & \cdots & \text{cov}(x_1, y_n) \\ \vdots & \ddots & \vdots \\ \text{cov}(x_k, y_1) & \cdots & \text{cov}(x_k, y_n) \end{pmatrix}$$

is the **covariance matrix** of $\mathbf{x}$ and $\mathbf{y}$ and is denoted $\text{cov}(\mathbf{x}, \mathbf{y})$, or sometimes $\Sigma_{\mathbf{x},\mathbf{y}}$.

- Notice that the $\text{cov}(\cdot, \cdot)$ function takes two arguments, each of which can be a scalar or a vector.
- In terms of vector/matrix algebra, $\text{cov}(\mathbf{x}, \mathbf{y})$ has formula

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \text{E}[(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})^T].$$

- Note that $\text{var}(\mathbf{x}) = \text{cov}(\mathbf{x}, \mathbf{x})$.

(Population) Correlation Matrix: For a random variable $\mathbf{x}_{k\times 1}$, the population correlation matrix is the matrix of correlations among the elements of $\mathbf{x}$:

$$\text{corr}(\mathbf{x}) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix},$$

where $\rho_{ij} = \text{corr}(x_i, x_j)$.

- Recall: for random variables x_i and x_j,

$$\rho_{ij} = \text{corr}(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$$

measures the amount of linear association between x_i and x_j.

- For any $\mathbf{x}$, $\text{corr}(\mathbf{x})$ is symmetric.

- Sometimes we will use the corr function with two arguments, $\text{corr}(\mathbf{x}_{k\times 1}, \mathbf{y}_{n\times 1})$ to mean the $k \times n$ matrix of correlations between the elements of $\mathbf{x}$ and $\mathbf{y}$:

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} \text{corr}(x_1, y_1) & \cdots & \text{corr}(x_1, y_n) \\ \vdots & \ddots & \vdots \\ \text{corr}(x_k, y_1) & \cdots & \text{corr}(x_k, y_n) \end{pmatrix}.$$

- Notice that $\text{corr}(\mathbf{x}) = \text{corr}(\mathbf{x}, \mathbf{x})$.

- For random vectors $\mathbf{x}_{k\times 1}$ and $\mathbf{y}_{n\times 1}$, let

$$\boldsymbol{\rho}_{\mathbf{x}} = \text{corr}(\mathbf{x}), \quad \Sigma_{\mathbf{x}} = \text{var}(\mathbf{x}), \quad \boldsymbol{\rho}_{\mathbf{x},\mathbf{y}} = \text{corr}(\mathbf{x}, \mathbf{y}), \quad \Sigma_{\mathbf{x},\mathbf{y}} = \text{cov}(\mathbf{x}, \mathbf{y}),$$
$$\mathbf{V}_{\mathbf{x}} = \text{diag}(\text{var}(x_1), \ldots, \text{var}(x_k)), \quad \text{and} \quad \mathbf{V}_{\mathbf{y}} = \text{diag}(\text{var}(y_1), \ldots, \text{var}(y_n)).$$

  The relationship between $\boldsymbol{\rho}_{\mathbf{x}}$ and $\Sigma_{\mathbf{x}}$ is

$$\Sigma_{\mathbf{x}} = \mathbf{V}_{\mathbf{x}}^{1/2} \boldsymbol{\rho}_{\mathbf{x}} \mathbf{V}_{\mathbf{x}}^{1/2}$$
$$\boldsymbol{\rho}_{\mathbf{x}} = (\mathbf{V}_{\mathbf{x}}^{1/2})^{-1} \Sigma_{\mathbf{x}} (\mathbf{V}_{\mathbf{x}}^{1/2})^{-1}$$

  and the relationship between the covariance and correlation matrices of $\mathbf{x}$ and $\mathbf{y}$ is

$$\Sigma_{\mathbf{x},\mathbf{y}} = \mathbf{V}_{\mathbf{x}}^{1/2} \boldsymbol{\rho}_{\mathbf{x},\mathbf{y}} \mathbf{V}_{\mathbf{y}}^{1/2}$$
$$\boldsymbol{\rho}_{\mathbf{x},\mathbf{y}} = \mathbf{V}_{\mathbf{x}}^{-1/2} \Sigma_{\mathbf{x},\mathbf{y}} \mathbf{V}_{\mathbf{y}}^{-1/2}$$

Properties:

Let $\mathbf{X}$, $\mathbf{Y}$ be random matrices of the same dimension and $\mathbf{A}$, $\mathbf{B}$ be matrices of constants such that $\mathbf{AXB}$ is defined

1. $\mathrm{E}(\mathbf{X}+\mathbf{Y}) = \mathrm{E}(\mathbf{X}) + \mathrm{E}(\mathbf{Y})$.
2. $\mathrm{E}(\mathbf{AXB}) = \mathbf{A}\mathrm{E}(\mathbf{X})\mathbf{B}$.
   - In particular, $\mathrm{E}(\mathbf{AX}) = \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}}$.

Now let $\mathbf{x}_{k\times 1}$, $\mathbf{y}_{n\times 1}$ be random vectors and let $\mathbf{c}_{k\times 1}$ and $\mathbf{d}_{n\times 1}$ be vectors of constants. Let $\mathbf{A}$, $\mathbf{B}$ be matrices conformable to the products $\mathbf{Ax}$ and $\mathbf{By}$.

3. $\mathrm{cov}(\mathbf{x},\mathbf{y}) = \mathrm{cov}(\mathbf{y},\mathbf{x})^T$.
4. $\mathrm{cov}(\mathbf{x}+\mathbf{c},\mathbf{y}+\mathbf{d}) = \mathrm{cov}(\mathbf{x},\mathbf{y})$.
5. $\mathrm{cov}(\mathbf{Ax},\mathbf{By}) = \mathbf{A}\mathrm{cov}(\mathbf{x},\mathbf{y})\mathbf{B}^T$.

Let $\mathbf{x}_1, \mathbf{x}_2$ be two $k \times 1$ random vectors and $\mathbf{y}_1, \mathbf{y}_2$ be two $n \times 1$ random vectors. Then

6. $\mathrm{cov}(\mathbf{x}_1+\mathbf{x}_2,\mathbf{y}_1) = \mathrm{cov}(\mathbf{x}_1,\mathbf{y}_1) + \mathrm{cov}(\mathbf{x}_2,\mathbf{y}_1)$ and $\mathrm{cov}(\mathbf{x}_1,\mathbf{y}_1+\mathbf{y}_2) = \mathrm{cov}(\mathbf{x}_1,\mathbf{y}_1) + \mathrm{cov}(\mathbf{x}_1,\mathbf{y}_2)$.

   - Taken together, properties 5 and 6 say that $\mathrm{cov}(\cdot,\cdot)$ is linear in both arguments (that is, it is *bilinear*).

Several properties of $\mathrm{var}(\cdot)$ follow directly from the properties of $\mathrm{cov}(\cdot,\cdot)$ since $\mathrm{var}(\mathbf{x}) = \mathrm{cov}(\mathbf{x},\mathbf{x})$:

7. $\mathrm{var}(\mathbf{x}_1+\mathbf{c}) = \mathrm{cov}(\mathbf{x}_1+\mathbf{c},\mathbf{x}_1+\mathbf{c}) = \mathrm{cov}(\mathbf{x}_1,\mathbf{x}_1) = \mathrm{var}(\mathbf{x}_1)$.
8. $\mathrm{var}(\mathbf{Ax}) = \mathbf{A}\mathrm{var}(\mathbf{x})\mathbf{A}^T$.
9. $\mathrm{var}(\mathbf{x}_1+\mathbf{x}_2) = \mathrm{cov}(\mathbf{x}_1+\mathbf{x}_2,\mathbf{x}_1+\mathbf{x}_2) = \mathrm{var}(\mathbf{x}_1) + \mathrm{cov}(\mathbf{x}_1,\mathbf{x}_2) + \mathrm{cov}(\mathbf{x}_2,\mathbf{x}_1) + \mathrm{var}(\mathbf{x}_2)$.

If $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent, then $\mathrm{cov}(\mathbf{x}_1,\mathbf{x}_2) = \mathbf{0}$, so property 9 implies $\mathrm{var}(\mathbf{x}_1+\mathbf{x}_2) = \mathrm{var}(\mathbf{x}_1) + \mathrm{var}(\mathbf{x}_2)$. This result extends easily to a sum of n independent $\mathbf{x}_i$'s so that

$$\mathrm{var}\Big(\sum_{i=1}^{n}\mathbf{x}_i\Big) = \sum_{i=1}^{n}\mathrm{var}(\mathbf{x}_i), \quad \text{if } \mathbf{x}_1,\ldots,\mathbf{x}_n \text{ are independent.}$$

In addition, if $\text{var}(\mathbf{x}_1) = \cdots = \text{var}(\mathbf{x}_n) = \Sigma_{\mathbf{x}}$ then $\text{var}(\sum_{i=1}^n \mathbf{x}_i) = n\Sigma_{\mathbf{x}}$. The formula for the variance of a sample mean vector follows easily:

$$\text{var}(\bar{\mathbf{x}}) = \text{var}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i) = \left(\frac{1}{n}\right)^2 \text{var}(\sum_{i=1}^n \mathbf{x}_i) = \left(\frac{1}{n^2}\right) n\Sigma_{\mathbf{x}} = \frac{1}{n}\Sigma_{\mathbf{x}}.$$

- Notice this generalizes the familiar result from the univariate case.

In linear models, quadratic forms $\mathbf{x}^T\mathbf{A}\mathbf{x}$ in some random vector $\mathbf{x}$ and symmetric matrix $\mathbf{A}$ often arise, and its useful to have a general result about how to take the expected value of such quantities.

- Note that for $\mathbf{A}$ not symmetric, $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is still a quadratic form, because it is possible to write $\mathbf{x}^T\mathbf{A}\mathbf{x}$ as $\mathbf{x}^T\mathbf{B}\mathbf{x}$ for the symmetric matrix $\mathbf{B} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T)$. That is, $Q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ can be written

$$\begin{aligned} Q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} &= \frac{1}{2}(\mathbf{x}^T\mathbf{A}\mathbf{x} + \underbrace{\mathbf{x}^T\mathbf{A}\mathbf{x}}_{=(\mathbf{x}^T\mathbf{A}^T\mathbf{x})^T}) = \frac{1}{2}(\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{A}^T\mathbf{x}) \\ &= \mathbf{x}^T \underbrace{\left\{\frac{1}{2}(\mathbf{A} + \mathbf{A}^T)\right\}}_{=\mathbf{B} \text{ symmetric}} \mathbf{x} \end{aligned}$$

That quadratic forms are common and important in linear models is familiar once we realize that any quadratic form can be written as a weighted sum of squares, and vice versa.

Let $\mathbf{A}$ be an $n \times n$ symmetric matrix with spectral decomposition $\mathbf{A} = \sum_{i=1}^n \lambda_i \mathbf{u}_i\mathbf{u}_i^T$. Then

$$Q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^n \lambda_i \mathbf{x}^T\mathbf{u}_i\mathbf{u}_i^T\mathbf{x} = \sum_{i=1}^n \lambda_i \underbrace{(\mathbf{u}_i^T\mathbf{x})}_{\equiv w_i}(\mathbf{u}_i^T\mathbf{x}) = \sum_{i=1}^n \lambda_i w_i^2.$$

The expected value of a quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$ follows immediately from a more general result concerning the expected value of a **bilinear form**, $Q(\mathbf{x},\mathbf{y}) = (\mathbf{x}_{k\times 1})^T\mathbf{A}_{k\times n}\mathbf{y}_{n\times 1}$.

Theorem: (E.V. of a bilinear form) Let $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}_{\mathbf{x}}$ and $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}_{\mathbf{y}}$, $\mathrm{cov}(\mathbf{x},\mathbf{y}) = \Sigma_{\mathbf{x},\mathbf{y}} = (\sigma_{ij})$ and $\mathbf{A} = (a_{ij})$. Then,

$$\mathrm{E}(\mathbf{x}^T\mathbf{A}\mathbf{y}) = \sum_i\sum_j a_{ij}\sigma_{ij} + \boldsymbol{\mu}_{\mathbf{x}}^T\mathbf{A}\boldsymbol{\mu}_{\mathbf{y}} = \mathrm{tr}(\mathbf{A}\Sigma_{\mathbf{x},\mathbf{y}}^T) + \boldsymbol{\mu}_{\mathbf{x}}^T\mathbf{A}\boldsymbol{\mu}_{\mathbf{y}}.$$

Proof: Writing the bilinear form in summation notation we have $\mathbf{x}^T\mathbf{A}\mathbf{y} = \sum_i\sum_j a_{ij}x_iy_j$. In addition, $\mathrm{E}(x_iy_j) = \mathrm{cov}(x_i,y_j) + \mu_{\mathbf{x},i}\mu_{\mathbf{y},j} = \sigma_{ij} + \mu_{\mathbf{x},i}\mu_{\mathbf{y},j}$, so

$$\begin{aligned}\mathrm{E}(\mathbf{x}^T\mathbf{A}\mathbf{y}) &= \sum_{i=1}^{k}\underbrace{\sum_{j=1}^{n}a_{ij}\sigma_{ij}} + \sum_{i=1}^{k}\sum_{j=1}^{n}a_{ij}\mu_{\mathbf{x},i}\mu_{\mathbf{y},j}\\ &\qquad\qquad = (i,i)^{\text{th}}\text{ term of }\mathbf{A}\Sigma_{\mathbf{x},\mathbf{y}}^T\\ &= \mathrm{tr}(\mathbf{A}\Sigma_{\mathbf{x},\mathbf{y}}^T) + \boldsymbol{\mu}_{\mathbf{x}}^T\mathbf{A}\boldsymbol{\mu}_{\mathbf{y}}\end{aligned}$$

■

Letting $\mathbf{y} = \mathbf{x}$ we obtain

Theorem: (E.V. of a quadratic form) Let $Q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$, $\mathrm{var}(\mathbf{x}) = \Sigma$, $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$ and $\mathbf{A} = (a_{ij})$. Then,

$$\mathrm{E}\left\{Q(\mathbf{x})\right\} = \sum_i\sum_j a_{ij}\mathrm{cov}(x_i,x_j) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} = \mathrm{tr}(\mathbf{A}\Sigma) + Q(\boldsymbol{\mu}).$$

Example: Let $x_1, \ldots, x_n$ be independent random variables each with mean μ and variance σ^2. Then for $\mathbf{x} = (x_1, \ldots, x_n)^T$, $\mathrm{E}(\mathbf{x}) = \mu\mathbf{j}_n$, $\mathrm{var}(\mathbf{x}) = \sigma^2\mathbf{I}_n$.

Consider the quadratic form

$$Q(\mathbf{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \|\mathbf{P}_{V^\perp}\mathbf{x}\|^2 = \mathbf{x}^T(\mathbf{I}_n - \mathbf{P}_V)\mathbf{x},$$

where $V = \mathcal{L}(\mathbf{j}_n)$, and $\mathbf{P}_V = (1/n)\mathbf{J}_{n,n}$. To obtain $\mathrm{E}\{Q(\mathbf{x})\}$ we note the matrix in the quadratic form is $\mathbf{A} = \mathbf{I}_n - \mathbf{P}_V$ and $\Sigma_{\mathbf{x}} = \sigma^2\mathbf{I}_n$. Then $\mathbf{A}\Sigma_{\mathbf{x}} = \sigma^2(\mathbf{I}_n - \mathbf{P}_V)$, and $\mathrm{tr}(\mathbf{A}\Sigma) = \sigma^2(n-1)$ (trace of a projection matrix equals the dimension onto which it projects). Thus

$$\mathrm{E}\{Q(\mathbf{x})\} = \sigma^2(n-1) + \underbrace{Q(\mu\mathbf{j}_n)}_{=0}$$

- An immediate consequence of this is the unbiasedness of the sample variance, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

An alternative method of obtaining this result is to define $\mathbf{y} = (x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})^T$, and apply the preceding theorem to $Q(\mathbf{x}) = \sum_i (x_i - \bar{x})^2 = \mathbf{y}^T\mathbf{I}_n\mathbf{y}$.

Since $\mathbf{y} = \mathbf{P}_{V^\perp}\mathbf{x}$, $\mathbf{y}$ has var-cov matrix $\mathbf{P}_{V^\perp}(\sigma^2\mathbf{I}_n)\mathbf{P}_{V^\perp}^T = \sigma^2\mathbf{P}_{V^\perp}$ (because $\mathbf{P}_{V^\perp}$ is idempotent and symmetric) and mean $\mathbf{0}$.

So,

$$\mathrm{E}\{Q(\mathbf{x})\} = \mathrm{tr}\{\mathbf{I}_n(\sigma^2\mathbf{P}_{V^\perp})\} = \mathrm{tr}\{\sigma^2(\mathbf{I}_n - \mathbf{P}_V)\} = \sigma^2(n-1),$$

as before.

The Multivariate Normal Distribution

Multivariate Normal Distribution: A random vector $\mathbf{y}_{n\times 1}$ is said to have a multivariate normal distribution if $\mathbf{y}$ has the same distribution as

$$\mathbf{A}_{n\times p}\mathbf{z}_{p\times 1} + \boldsymbol{\mu}_{n\times 1} \equiv \mathbf{x}$$

where, for some p, $\mathbf{z}$ is a vector of independent $N(0,1)$ random variables, $\mathbf{A}$ is a matrix of constants, and $\boldsymbol{\mu}$ is a vector of constants.

- The type of transformation used in going from $\mathbf{z}$ to $\mathbf{x}$ above is called an *affine transformation.*

By using the form of $\mathbf{x}$ and the fact that the elements of $\mathbf{z}$ are independent standard normal, we can determine the density function of $\mathbf{x}$ and hence of $\mathbf{y}$.

- This is only possible in the case that $n = p$ and $\text{rank}(\mathbf{A}) = p$. We will restrict attention to this case.

Define $g(\mathbf{z}) = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$ to be the transformation from $\mathbf{z}$ to $\mathbf{x}$. For $\mathbf{A}$ a $p \times p$ full rank matrix, $g(\mathbf{z})$ is a 1–1 function from $\mathcal{R}^p$ to $\mathcal{R}^p$ so that we can use the following change of variable formula for the density of $\mathbf{x}$:

$$f_{\mathbf{x}}(\mathbf{x}) = f_{\mathbf{z}}\{g^{-1}(\mathbf{x})\}\text{abs}\left(\left|\frac{\partial g^{-1}(\mathbf{x})}{\partial \mathbf{x}^T}\right|\right) = f_{\mathbf{z}}\{\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}\text{abs}(|\mathbf{A}^{-1}|).$$

(Here, $\text{abs}(\cdot)$ denotes the absolute value and $|\cdot|$ the determinant.)

Since the elements of $\mathbf{z}$ are independent standard normals, the density of $\mathbf{z}$ is

$$f_{\mathbf{z}}(\mathbf{z}) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = (2\pi)^{-p/2} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right).$$

Plugging into our change of variable formula we get

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-p/2} \underbrace{\text{abs}(|\mathbf{A}|)}_{=|\mathbf{A}|}{}^{-1} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Note that $\Sigma = \text{var}(\mathbf{x})$ is equal to $\text{var}(\mathbf{Az}+\boldsymbol{\mu}) = \text{var}(\mathbf{Az}) = \mathbf{AIA}^T = \mathbf{AA}^T$, so

$$|\Sigma| = |\mathbf{AA}^T| = |\mathbf{A}|^2 \quad \Rightarrow \quad |\mathbf{A}| = |\Sigma|^{1/2}.$$

In addition, $\text{E}(\mathbf{x}) = \text{E}(\mathbf{Az} + \boldsymbol{\mu}) = \boldsymbol{\mu}$.

So, a multivariate normal random vector of dimension p with mean $\boldsymbol{\mu}$ and p.d. var-cov matrix Σ has density

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\Sigma)^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \quad \text{for all } \mathbf{x} \in \mathcal{R}^p.$$

- In the case where $\mathbf{A}$ is $n \times p$ with $\text{rank}(\mathbf{A}) \neq n$, $\mathbf{x}$ is still multivariate normal, but its density does not exist. Such cases correspond to multivariate normal distributions with non p.d. var-cov matrices, which arise rarely and which we won't consider in this course. Such distributions do not have p.d.f.'s but can be characterized using the characteristic function, which always exists.

Recall that the probability density function is just one way to characterize (fully describe) the distribution of a random variable (or vector). Another function that can be used for this purpose is the moment generating function m.g.f.

The m.g.f. of a random vector $\mathbf{x}$ is $m_{\mathbf{x}}(\mathbf{t}) = \text{E}(e^{\mathbf{t}^T\mathbf{x}})$. So, for $\mathbf{x} = \mathbf{Az}+\boldsymbol{\mu} \sim N_n(\boldsymbol{\mu}, \mathbf{AA}^T = \Sigma)$, the m.g.f. of $\mathbf{x}$ is

$$m_{\mathbf{x}}(\mathbf{t}) = \text{E}[\exp\{\mathbf{t}^T(\mathbf{Az}+\boldsymbol{\mu})\}] = e^{\mathbf{t}^T\boldsymbol{\mu}}\text{E}(e^{\mathbf{t}^T\mathbf{Az}}) = e^{\mathbf{t}^T\boldsymbol{\mu}}m_{\mathbf{z}}(\mathbf{A}^T\mathbf{t}). \qquad (*)$$

The m.g.f. of a standard normal r.v. z_i is $m_{z_i}(u) = e^{u^2/2}$, so the m.g.f. of $\mathbf{z}$ is

$$m_{\mathbf{z}}(\mathbf{u}) = \prod_{i=1}^{p} \exp(u_i^2/2) = e^{\mathbf{u}^T\mathbf{u}/2}.$$

Substituting into (*) we get

$$m_{\mathbf{x}}(\mathbf{t}) = e^{\mathbf{t}^T\boldsymbol{\mu}} \exp\{\frac{1}{2}(\mathbf{A}^T\mathbf{t})^T(\mathbf{A}^T\mathbf{t})\} = e^{\mathbf{t}^T\boldsymbol{\mu}} \exp(\frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}).$$

- So, if $m_{\mathbf{x}}(\mathbf{t})$ completely characterizes the distribution of $\mathbf{x}$ and $m_{\mathbf{x}}(\mathbf{t})$ depends only upon $\mathbf{x}$'s mean and variance $\boldsymbol{\mu}$ and Σ, then that says that a multivariate normal distribution is completely specified by these two parameters.

- I.e., for $\mathbf{x}_1 \sim N_n(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathbf{x}_2 \sim N_n(\boldsymbol{\mu}_2, \Sigma_2)$, $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same distribution if and only if $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\Sigma_1 = \Sigma_2$.

Theorem: Let $\boldsymbol{\mu}$ be an element of $\mathcal{R}^n$ and Σ an $n \times n$ symmetric p.s.d. matrix. Then there exists a multivariate normal distribution with mean $\boldsymbol{\mu}$ and var-cov matrix Σ.

Proof: Since Σ is symmetric and p.s.d., there exists a $\mathbf{B}$ so that $\Sigma = \mathbf{B}\mathbf{B}^T$ (e.g., the Cholesky decomposition). Let $\mathbf{z}$ be an $n \times 1$ vector of independent standard normals. Then $\mathbf{x} = \mathbf{B}\mathbf{z} + \boldsymbol{\mu} \sim N_n(\boldsymbol{\mu}, \Sigma)$. ∎

- This result suggests that we can always generate a multivariate normal random vector with given mean $\boldsymbol{\mu}$ and given var-cov matrix Σ by generating a vector of independent standard normals $\mathbf{z}$ and then pre-multiplying $\mathbf{z}$ by the lower-triangular Cholesky factor $\mathbf{B}$ and then adding on the mean vector $\boldsymbol{\mu}$.

Theorem: Let $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$ where Σ is p.d. Let $\mathbf{y}_{r\times 1} = \mathbf{C}_{r\times n}\mathbf{x} + \mathbf{d}$ for $\mathbf{C}$ and $\mathbf{d}$ containing constants. Then $\mathbf{y} \sim N_r(\mathbf{C}\boldsymbol{\mu} + \mathbf{d}, \mathbf{C}\Sigma\mathbf{C}^T)$.

Proof: By definition, $\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$ for some $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^T = \Sigma$, and $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$. Then

$$\begin{aligned}\mathbf{y} &= \mathbf{C}\mathbf{x} + \mathbf{d} = \mathbf{C}(\mathbf{A}\mathbf{z} + \boldsymbol{\mu}) + \mathbf{d} = \mathbf{C}\mathbf{A}\mathbf{z} + \mathbf{C}\boldsymbol{\mu} + \mathbf{d} \\ &= (\mathbf{C}\mathbf{A})\mathbf{z} + (\mathbf{C}\boldsymbol{\mu} + \mathbf{d}).\end{aligned}$$

So, by definition, $\mathbf{y}$ has a multivariate normal distribution with mean $\mathbf{C}\boldsymbol{\mu} + \mathbf{d}$ and var-cov matrix $(\mathbf{C}\mathbf{A})(\mathbf{C}\mathbf{A})^T = \mathbf{C}\Sigma\mathbf{C}^T$.∎

Simple corollaries of this theorem are that if $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$, then

i. any subvector of $\mathbf{x}$ is multivariate normal too, with mean and variance given by the corresponding subvector of $\boldsymbol{\mu}$ and submatrix of Σ, respectively, and

ii. any linear combination $\mathbf{a}^T\mathbf{x} \sim N(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\Sigma\mathbf{a})$ (univariate normal) for $\mathbf{a}$ a vector of constants.

Theorem: Let $\mathbf{y}_{n\times 1}$ have a multivariate normal distribution, and partition $\mathbf{y}$ as

$$\mathbf{y} = \begin{pmatrix} \underset{p\times 1}{\mathbf{y}_1} \\ \cdots \\ \underset{(n-p)\times 1}{\mathbf{y}_2} \end{pmatrix}.$$

Then $\mathbf{y}_1$ and $\mathbf{y}_2$ are independent if and only if $\text{cov}(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{0}$.

Proof: 1st, independence implies 0 covariance: Suppose $\mathbf{y}_1, \mathbf{y}_2$ are independent with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$. Then

$$\text{cov}(\mathbf{y}_1, \mathbf{y}_2) = \text{E}\{(\mathbf{y}_1 - \boldsymbol{\mu}_1)(\mathbf{y}_2 - \boldsymbol{\mu}_2)^T\} = \text{E}\{(\mathbf{y}_1 - \boldsymbol{\mu}_1)\}\text{E}\{(\mathbf{y}_2 - \boldsymbol{\mu}_2)^T\} = \mathbf{0}(\mathbf{0}^T) = \mathbf{0}.$$

2nd, 0 covariance and normality imply independence: To do this we use the fact that two random vectors are independent if and only if their joint m.g.f. is the product of their marginal m.g.f.'s. Suppose $\text{cov}(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{0}$. Let $\mathbf{t}_{n\times 1}$ be partitioned as $\mathbf{t} = (\mathbf{t}_1^T, \mathbf{t}_2^T)^T$ where $\mathbf{t}_1$ is $p \times 1$. Then $\mathbf{y}$ has m.g.f.

$$m_{\mathbf{y}}(\mathbf{t}) = \exp(\underbrace{\mathbf{t}^T\boldsymbol{\mu}}_{=\mathbf{t}_1^T\boldsymbol{\mu}_1 + \mathbf{t}_2^T\boldsymbol{\mu}_2}) \exp(\frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}),$$

where

$$\Sigma = \text{var}(\mathbf{y}) = \begin{pmatrix} \Sigma_{11} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \text{var}(\mathbf{y}_1) & \mathbf{0} \\ \mathbf{0} & \text{var}(\mathbf{y}_2) \end{pmatrix}$$

Because of the form of Σ, $\mathbf{t}^T\Sigma\mathbf{t} = \mathbf{t}_1^T\Sigma_{11}\mathbf{t}_1 + \mathbf{t}_2^T\Sigma_{22}\mathbf{t}_2$, so

$$\begin{aligned} m_{\mathbf{y}}(\mathbf{t}) &= \exp(\mathbf{t}_1^T\boldsymbol{\mu}_1 + \frac{1}{2}\mathbf{t}_1^T\Sigma_{11}\mathbf{t}_1 + \mathbf{t}_2^T\boldsymbol{\mu}_2 + \frac{1}{2}\mathbf{t}_2^T\Sigma_{22}\mathbf{t}_2) \\ &= \exp(\mathbf{t}_1^T\boldsymbol{\mu}_1 + \frac{1}{2}\mathbf{t}_1^T\Sigma_{11}\mathbf{t}_1) \exp(\mathbf{t}_2^T\boldsymbol{\mu}_2 + \frac{1}{2}\mathbf{t}_2^T\Sigma_{22}\mathbf{t}_2) = m_{\mathbf{y}_1}(\mathbf{t}_1)m_{\mathbf{y}_2}(\mathbf{t}_2). \end{aligned}$$

■

Lemma: Let $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \Sigma)$ where we have the partitioning

$$\mathbf{y} = \begin{pmatrix} \underset{p\times 1}{\mathbf{y}_1} \\ \cdots \\ \underset{(n-p)\times 1}{\mathbf{y}_2} \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{21} = \Sigma_{12}^T$. Let $\mathbf{y}_{2|1} = \mathbf{y}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1$. Then $\mathbf{y}_1$ and $\mathbf{y}_{2|1}$ are independent with

$$\mathbf{y}_1 \sim N_p(\boldsymbol{\mu}_1, \Sigma_{11}), \quad \mathbf{y}_{2|1} \sim N_{n-p}(\boldsymbol{\mu}_{2|1}, \Sigma_{22|1}),$$

where

$$\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\mu}_1, \quad \text{and} \quad \Sigma_{22|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Proof: We can write $\mathbf{y}_1 = \mathbf{C}_1\mathbf{y}$ where $\mathbf{C}_1 = (\mathbf{I}, \mathbf{0})$ and we can write $\mathbf{y}_{2|1} = \mathbf{C}_2\mathbf{y}$ where $\mathbf{C}_2 = (-\Sigma_{21}\Sigma_{11}^{-1}, \mathbf{I})$, so by the theorem on the bottom of p. 72, both $\mathbf{y}_1$ and $\mathbf{y}_{2|1}$ are normal. Their mean and variance-covariances are $\mathbf{C}_1\boldsymbol{\mu} = \boldsymbol{\mu}_1$ and $\mathbf{C}_1\Sigma\mathbf{C}_1^T = \Sigma_{11}$ for $\mathbf{y}_1$, and $\mathbf{C}_2\boldsymbol{\mu} = \boldsymbol{\mu}_{2|1}$ and $\mathbf{C}_2\Sigma\mathbf{C}_2^T = \Sigma_{22|1}$ for $\mathbf{y}_{2|1}$. Independence follows from the fact that these two random vectors have covariance matrix $\text{cov}(\mathbf{y}_1, \mathbf{y}_{2|1}) = \text{cov}(\mathbf{C}_1\mathbf{y}, \mathbf{C}_2\mathbf{y}) = \mathbf{C}_1\Sigma\mathbf{C}_2^T = \mathbf{0}$.

Theorem: For $\mathbf{y}$ defined as in the previous theorem, the conditional distribution of $\mathbf{y}_2$ given $\mathbf{y}_1$ is

$$\mathbf{y}_2|\mathbf{y}_1 \sim N_{n-p}(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1), \Sigma_{22|1}).$$

Proof: Since $\mathbf{y}_{2|1}$ is independent of $\mathbf{y}_1$, its conditional distribution for a given value of $\mathbf{y}_1$ is the same as its marginal distribution, $\mathbf{y}_{2|1} \sim N_{n-p}(\boldsymbol{\mu}_{2|1}, \Sigma_{22|1})$. Notice that $\mathbf{y}_2 = \mathbf{y}_{2|1} + \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1$. Conditional on the value of $\mathbf{y}_1$, $\Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1$ is constant, so the conditional distribution of $\mathbf{y}_2$ is that of $\mathbf{y}_{2|1}$ plus a constant, or $(n-p)-$variate normal, with mean

$$\boldsymbol{\mu}_{2|1} + \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1 = \boldsymbol{\mu}_2 - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\mu}_1 + \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1 = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1),$$

and var-cov matrix $\Sigma_{22|1}$. ∎

Partial and Multiple Correlation

Example — Height and Reading Ability:

Suppose that an educational psychologist studied the relationship between height y_1 and reading ability y_2 of children based on scores on a standardized test. For 200 children in grades 3, 4, and 5 he measured y_1 and y_2 and found that the sample correlation between these variables was .56.

Is there a linear association between height and reading ability?

Well, yes, but only because we've ignored one or more "lurking" variables. There is likely no direct effect of height on reading ability. Instead, older children with more years of schooling tend to be better readers and tend to be taller. The effects of age on both y_1 and y_2 have been ignored by just examining the simple correlation between y_1 and y_2.

The partial correlation coefficient is a measure of linear relationship between two variables, with the linear effects of one or more other variables (in this case, age) removed.

Partial Correlation: Suppose $\mathbf{v} \sim N_{p+q}(\boldsymbol{\mu}, \Sigma)$ and let $\mathbf{v}$, $\boldsymbol{\mu}$ and Σ be partitioned as

$$\mathbf{v} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{y}} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix},$$

where $\mathbf{x} = (v_1, \ldots, v_p)^T$ is $p \times 1$ and $\mathbf{y} = (v_{p+1}, \ldots, v_{p+q})^T$ is $q \times 1$.

Recall that the conditional var-cov matrix of $\mathbf{y}$ given $\mathbf{x}$ is

$$\text{var}(\mathbf{y}|\mathbf{x}) = \Sigma_{\mathbf{yy}} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}} \equiv \Sigma_{\mathbf{y}|\mathbf{x}}.$$

Let $\sigma_{ij|1,\ldots,p}$ denote the $(i,j)^{\text{th}}$ element of $\Sigma_{\mathbf{y}|\mathbf{x}}$.

Then the **partial correlation coefficient** of y_i and y_j given $\mathbf{x} = \mathbf{c}$ is defined by

$$\rho_{ij|1,\ldots,p} = \frac{\sigma_{ij|1,\ldots,p}}{[\sigma_{ii|1,\ldots,p}\sigma_{jj|1,\ldots,p}]^{1/2}}$$

(provided the denominator is non-zero, in which case $\rho_{ij|1,\ldots,p}$ is undefined).

- Like the ordinary correlation coefficient, the partial correlation satisfies
$$-1 \leq \rho_{ij|1,\ldots,p} \leq 1.$$

- Interpretation: the partial correlation $\rho_{ij|1,\ldots,p}$ measures the linear association between y_i and y_j after accounting for (or removing) the linear association between y_i and $\mathbf{x}$ and between y_j and $\mathbf{x}$.

  - E.g., if $v_1 =$ age, $v_2 =$height, $v_3 =$reading ability, $\mathbf{x} = v_1$ and $\mathbf{y} = (v_2, v_3)^T$, then $\rho_{23|1} =$ the correlation between height and reading ability after removing the effects of age on each of these variables. I would expect $\rho_{23|1} \approx 0$.

The partial correlation measures the linear association between two variables after removing the effects of several others. The multiple correlation coefficient measures the linear association between one variable and a group of several others.

Multiple Correlation: Suppose $\mathbf{v} \sim N_{p+1}(\boldsymbol{\mu}, \Sigma)$ and let $\mathbf{v}$, $\boldsymbol{\mu}$ and Σ be partitioned as

$$\mathbf{v} = \begin{pmatrix} \mathbf{x} \\ y \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y} \\ \boldsymbol{\sigma}_{y\mathbf{x}} & \sigma_{yy} \end{pmatrix},$$

where $\mathbf{x} = (v_1, \ldots, v_p)^T$ is $p \times 1$, and $y = v_{p+1}$ is a scalar random variable.

Recall that the conditional mean of y given $\mathbf{x}$ is

$$\mu_y + \boldsymbol{\sigma}_{y\mathbf{x}} \Sigma_{\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) \equiv \mu_{y|\mathbf{x}}.$$

Then the **squared multiple correlation coefficient** between y and $\mathbf{x}$ is defined as

$$\rho^2_{y,\mathbf{x}} = \frac{\mathrm{cov}(\mu_{y|\mathbf{x}}, y)}{[\mathrm{var}(\mu_{y|\mathbf{x}})\mathrm{var}(y)]^{1/2}}.$$

- A computationally simple formula is given by

$$\rho^2_{y,\mathbf{x}} = \left\{ \frac{\boldsymbol{\sigma}_{y\mathbf{x}} \Sigma_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{\mathbf{x}y}}{\sigma_{yy}} \right\}^{1/2}.$$

- My notation for this quantity is $\rho^2_{y,\mathbf{x}}$ rather than $\rho_{y,\mathbf{x}}$ because it behaves like the square of a correlation coefficient. It is bounded between zero and one:

  $$0 \leq \rho^2_{y,\mathbf{x}} \leq 1,$$

  and quantifies the strength of the linear association, but not the direction.

- The sample squared multiple correlation coefficient is called the **coefficient of determination** and usually denoted R^2.

- We'll talk about sample versions of the partial and multiple correlation coefficients later, when we get to fitting linear models.

Distributions of Quadratic Forms: The χ^2, F, and t Distributions

- All three of these distributions arise as the distributions of certain functions of normally distributed random variables.

- Their central (ordinary) versions are probably familiar as the distributions of normal-theory test statistics under the usual null hypotheses and as the basis for confidence intervals.

- The non-central versions of these distributions arise as the distributions of normal-theory test statistics under the alternatives to the usual null hypotheses. Thus, they are important in, for example, determining the power of tests.

Chi-square Distribution: Let $x_1, \ldots, x_n$ be independent normal random variables with means $\mu_1, \ldots, \mu_n$ and common variance 1. Then

$$y = x_1^2 + \cdots + x_n^2 = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2, \quad \text{where } \mathbf{x} = (x_1, \ldots, x_n)^T$$

is said to have a **noncentral chi-square distribution** with n **degrees of freedom** and **noncentrality parameter** $\lambda = \frac{1}{2}\sum_{i=1}^n \mu_i^2$. We denote this as $y \sim \chi^2(n, \lambda)$.

- As in the t and F distributions, the central chi-square distribution occurs when the noncentrality parameter λ equals 0.

- The central χ^2 with n degrees of freedom will typically be denoted $\chi^2(n) \equiv \chi^2(n, 0)$.

  - In particular, for $z_1, \ldots, z_n \overset{iid}{\sim} N(0, 1)$, $z_1^2 + \cdots + z_n^2 \sim \chi^2(n)$.

- The non-central $\chi^2(n, \lambda)$ distribution depends only on its parameters, n and λ.

A random variable $Y \sim \chi^2(n)$ (a central χ^2 with n d.f.) has p.d.f.

$$f_Y(y; n) = \frac{y^{n/2-1} e^{-y/2}}{\Gamma(n/2) 2^{n/2}}, \quad \text{for } y > 0.$$

- This is a special case of a gamma density with power parameter $n/2$ and scale parametr $1/2$.

The non-central χ^2 density is a Poisson mixture of central χ^2's. If $Z \sim \chi^2(n, \lambda)$ then Z has p.d.f.

$$f_Z(z; n, \lambda) = \sum_{k=0}^{\infty} p(k; \lambda) f_Y(z; n + 2k), \quad \text{for } z > 0,$$

where $p(k; \lambda) = \{e^{-\lambda}(\lambda)^k\}/k!$ is the Poisson probability function with rate (mean) λ.

- I.e., the noncentral χ^2 is a weighted sum of central χ^2's with Poisson weights.

Theorem: Let $Y \sim \chi^2(n, \lambda)$. Then

i. $\mathrm{E}(Y) = n + 2\lambda$;
ii. $\mathrm{var}(Y) = 2n + 8\lambda$; and
iii. the m.g.f. of Y is

$$m_Y(t) = \frac{\exp[-\lambda\{1 - 1/(1 - 2t)\}]}{(1 - 2t)^{n/2}}.$$

Proof: The proof of (i) and (ii) we leave as a homework problem. The proof of (iii) simply involves using the definition of expectation to evaluate

$$m_Y(t) = \mathrm{E}(e^{tY}) = \mathrm{E}\{e^{t(\mathbf{x}^T\mathbf{x})}\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t(\mathbf{x}^T\mathbf{x})} \underbrace{f_{\mathbf{x}}(\mathbf{x})}_{\text{a } N(\boldsymbol{\mu}, \mathbf{I}) \text{ density}} dx_1 \ldots dx_n,$$

where $\mathbf{x}$ is an $n \times 1$ vector of independent normals with mean $\boldsymbol{\mu}$ and constant variance 1. See Graybill (1976, p. 126) for details. ∎

The χ^2 distribution has the convenient property that sums of independent χ^2's are χ^2 too:

Theorem: If $v_1, \ldots, v_k$ are independent random variables distributed as $\chi^2(n_i, \lambda_i)$, $i = 1, \ldots, k$, respectively, then

$$\sum_{i=1}^{k} v_i \sim \chi^2(\sum_{i=1}^{k} n_i, \sum_{i=1}^{k} \lambda_i).$$

Proof: follows easily from the definition.

Distribution of a Quadratic Form:

From the definition of a central chi-square, it is immediate that if $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \mathbf{I}_n)$ then

$$\|\mathbf{y} - \boldsymbol{\mu}\|^2 = (\mathbf{y} - \boldsymbol{\mu})^T(\mathbf{y} - \boldsymbol{\mu}) \sim \chi^2(n).$$

For $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \Sigma)$ where Σ is p.d., we can extend this to

$$(\mathbf{y} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu}) \sim \chi^2(n)$$

by noting that

$$\begin{aligned}(\mathbf{y} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu}) &= (\mathbf{y} - \boldsymbol{\mu})^T\Sigma^{-1/2}\Sigma^{-1/2}(\mathbf{y} - \boldsymbol{\mu}) \\ &= \{\Sigma^{-1/2}(\mathbf{y} - \boldsymbol{\mu})\}^T \underbrace{\{\Sigma^{-1/2}(\mathbf{y} - \boldsymbol{\mu})\}}_{\equiv \mathbf{z}} \\ &= \mathbf{z}^T\mathbf{z},\end{aligned}$$

where $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$, and $\Sigma^{1/2}$ is the unique symmetric square root of Σ (see p. 51 of these notes). Since $\mathbf{z} = \Sigma^{-1/2}(\mathbf{y} - \boldsymbol{\mu}) \sim N_n(\mathbf{0}, \mathbf{I}_n)$, we have that $\mathbf{z}^T\mathbf{z} = (\mathbf{y} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu}) \sim \chi^2(n)$.

We'd like to generalize these results on the distribution of quadratic forms to obtain the the distribution of $\mathbf{y}^T\mathbf{A}\mathbf{y}$ for $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \Sigma)$ for $\mathbf{A}$ a matrix of constants. We can do this, but first we need a couple of results.

Theorem: If $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \Sigma)$ then the m.g.f. of $\mathbf{y}^T\mathbf{A}\mathbf{y}$ is

$$m_{\mathbf{y}^T\mathbf{A}\mathbf{y}}(t) = |\mathbf{I}_n - 2t\mathbf{A}\Sigma|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{\mu}^T\{\mathbf{I}_n - (\mathbf{I}_n - 2t\mathbf{A}\Sigma)^{-1}\}\Sigma^{-1}\boldsymbol{\mu}\right].$$

Proof: Again, the proof of this result "simply" involves evaluating the expectation,

$$\mathrm{E}(e^{t\mathbf{y}^T\mathbf{A}\mathbf{y}}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t\mathbf{y}^T\mathbf{A}\mathbf{y}} \underbrace{f_y(\mathbf{y})}_{\text{a multi'normal density}} dy_1 \ldots dy_n.$$

See Searle, 1971, p.55, for details. ∎

We also need a couple of eigenvalue results that we probably should have stated earlier:

Result: If λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ is the corresponding eigenvector of $\mathbf{A}$, then for scalars c and k, $(c\lambda + k, \mathbf{x})$ is an eigenvalue-eigenvector pair of the matrix $c\mathbf{A} + k\mathbf{I}$.

Proof: Because $(\lambda, \mathbf{x})$ is an eigen-pair for $\mathbf{A}$, they satisfy $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ which implies

$$c\mathbf{A}\mathbf{x} = c\lambda\mathbf{x}.$$

Adding $k\mathbf{x}$ to both sides of this equation we have

$$\begin{aligned} & c\mathbf{A}\mathbf{x} + k\mathbf{x} = c\lambda\mathbf{x} + k\mathbf{x} \\ \Rightarrow \qquad & (c\mathbf{A} + k\mathbf{I})\mathbf{x} = (c\lambda + k)\mathbf{x}, \end{aligned}$$

so that $(c\lambda + k, \mathbf{x})$ is an eigenvalue-eigenvector pair of the matrix $c\mathbf{A} + k\mathbf{I}$. ∎

Result: If all of the eigenvalues of $\mathbf{A}$ satisfy $-1 < \lambda < 1$, then

$$(\mathbf{I} - \mathbf{A})^{-1} = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots = \mathbf{I} + \sum_{k=1}^{\infty} \mathbf{A}^k. \qquad (*)$$

Proof: This can be verified by multiplying $\mathbf{I} - \mathbf{A}$ times $(\mathbf{I} + \sum_{k=1}^{\infty} \mathbf{A}^k)$ to obtain the identity matrix. Note that $-1 < \lambda < 1$ for all eigenvalues of $\mathbf{A}$ ensures $\lim_{k\to\infty} \mathbf{A}^k \to \mathbf{0}$ so that $\sum_{k=1}^{\infty} \mathbf{A}^k$ converges. ∎

Previously, we established that a projection matrix $\mathbf{P}_V$ onto a subspace $V \in \mathcal{R}^n$ where $\dim(V) = k$ has k eigenvalues equal to 1, and $n - k$ eigenvalues equal to 0.

Recall that a projection matrix is symmetric and idempotent. More generally, this result can be extended to all idempotent matrices.

Theorem: If $\mathbf{A}$ is an $n \times n$ idempotent matrix of rank r, then $\mathbf{A}$ has r eigenvalues equal to 1, and $n - r$ eigenvalues equal to 0.

Proof: In general, if λ is an eigenvalue for $\mathbf{A}$, then λ^2 is an eigenvalue for $\mathbf{A}^2$ since

$$\mathbf{A}^2\mathbf{x} = \mathbf{A}(\mathbf{A}\mathbf{x}) = \mathbf{A}\lambda\mathbf{x} = \lambda\mathbf{A}\mathbf{x} = \lambda\lambda\mathbf{x} = \lambda^2\mathbf{x}.$$

Since $\mathbf{A}^2 = \mathbf{A}$, we have $\mathbf{A}^2\mathbf{x} = \mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. Equating the right sides of $\mathbf{A}^2\mathbf{x} = \lambda^2\mathbf{x}$ and $\mathbf{A}^2\mathbf{x} = \lambda\mathbf{x}$, we have

$$\lambda\mathbf{x} = \lambda^2\mathbf{x}, \quad \text{or} \quad (\lambda - \lambda^2)\mathbf{x} = \mathbf{0}.$$

But $\mathbf{x} \neq \mathbf{0}$, so $\lambda - \lambda^2 = 0$, from which λ must be either 0 or 1. Since $\mathbf{A}$ is idempotent, it must be p.s.d., so the number of nonzero eigenvalues is equal to $\text{rank}(\mathbf{A}) = r$ and therefore, r eigenvalues are 1 and $n - r$ are 0. ∎

Now we are ready to state our main result:

Theorem: Let $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \Sigma)$ and $\mathbf{A}$ be a $n \times n$ symmetric matrix of constants with $\text{rank}(\mathbf{A}) = r$. Let $\lambda = \frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$. Then

$$\mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2(r, \lambda) \quad \text{if and only if } \mathbf{A}\Sigma \text{ is idempotent.}$$

Proof: From the theorem on p. 81, the moment generating function of $\mathbf{y}^T\mathbf{A}\mathbf{y}$ is

$$m_{\mathbf{y}^T\mathbf{A}\mathbf{y}}(t) = |\mathbf{I}_n - 2t\mathbf{A}\Sigma|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{\mu}^T\{\mathbf{I}_n - (\mathbf{I}_n - 2t\mathbf{A}\Sigma)^{-1}\}\Sigma^{-1}\boldsymbol{\mu}\right].$$

By the result on p. 81, the eigenvalues of $\mathbf{I}_n - 2t\mathbf{A}\Sigma$ are $1 - 2t\lambda_i$, $i = 1, \ldots, n$, where λ_i is an eigenvalue of $\mathbf{A}\Sigma$. Since the determinant equals the product of the eigenvalues, we have $|\mathbf{I}_n - 2t\mathbf{A}\Sigma| = \prod_{i=1}^n (1 - 2t\lambda_i)$. In addition, by (*) we have $(\mathbf{I}_n - 2t\mathbf{A}\Sigma)^{-1} = \mathbf{I}_n + \sum_{k=1}^\infty (2t)^k(\mathbf{A}\Sigma)^k$ provided that $-1 < 2t\lambda_i < 1$ for all i. Thus $m_{\mathbf{y}^T\mathbf{A}\mathbf{y}}(t)$ can be written as

$$m_{\mathbf{y}^T\mathbf{A}\mathbf{y}}(t) = \left(\prod_{i=1}^n (1 - 2t\lambda_i)^{-1/2}\right) \exp\left[-\frac{1}{2}\boldsymbol{\mu}^T\{-\sum_{k=1}^\infty (2t)^k(\mathbf{A}\Sigma)^k\}\Sigma^{-1}\boldsymbol{\mu}\right].$$

Now suppose $\mathbf{A}\Sigma$ is idempotent of rank $r = \text{rank}(\mathbf{A})$. Then $(\mathbf{A}\Sigma)^k = \mathbf{A}\Sigma$ and r of the λ_i's are equal to 1, and $n - r$ of the λ_i's are equal to 0. Therefore,

$$\begin{aligned} m_{\mathbf{y}^T\mathbf{A}\mathbf{y}}(t) &= \left(\prod_{i=1}^r (1 - 2t)^{-1/2}\right) \exp[-\frac{1}{2}\boldsymbol{\mu}^T\{-\sum_{k=1}^\infty (2t)^k\}\mathbf{A}\Sigma\Sigma^{-1}\boldsymbol{\mu}] \\ &= (1 - 2t)^{-r/2} \exp[-\frac{1}{2}\boldsymbol{\mu}^T\{1 - (1 - 2t)^{-1}\}\mathbf{A}\boldsymbol{\mu}], \end{aligned}$$

provided that $-1 < 2t < 1$ or $-\frac{1}{2} < t < \frac{1}{2}$, which is compatible with the requirement that the m.g.f. exist for t in a neighborhood of 0. (Here we have used the series expansion $1/(1 - x) = 1 + \sum_{k=1}^\infty x^k$ for $-1 < x < 1$.) Thus,

$$m_{\mathbf{y}^T\mathbf{A}\mathbf{y}}(t) = \frac{\exp[-\frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}\{1 - 1/(1 - 2t)\}]}{(1 - 2t)^{r/2}},$$

which is the m.g.f. of a $\chi^2(r, \frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu})$ random variable.

For a proof of the converse (that $\mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2(r, \lambda)$ implies $\mathbf{A}\Sigma$ idempotent), see Searle (1971, pp. 57–58). ∎

Several useful results follow easily from the previous theorem as corollaries:

Corollary 1: If $\mathbf{y} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, then $\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2(r)$ if and only if $\mathbf{A}$ is idempotent of rank r.

Corollary 2: If $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, then $\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2(r, \frac{1}{2\sigma^2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu})$ if and only if $\mathbf{A}$ is idempotent of rank r.

Corollary 3: Suppose $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$ and let $\mathbf{P}_V$ be the projection matrix onto a subspace $V \in \mathcal{R}^n$ of dimension $r \leq n$. Then

$$\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{P}_V\mathbf{y} = \frac{1}{\sigma^2}\|p(\mathbf{y}|V)\|^2 \sim \chi^2(r, \frac{1}{2\sigma^2}\boldsymbol{\mu}^T\mathbf{P}_V\boldsymbol{\mu}) = \chi^2(r, \frac{1}{2\sigma^2}\|p(\boldsymbol{\mu}|V)\|^2).$$

Corollary 4: Suppose $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$ and let $\mathbf{c}$ be an $n \times 1$ vector of constants. Then

$$(\mathbf{y} - \mathbf{c})^T\Sigma^{-1}(\mathbf{y} - \mathbf{c}) \sim \chi^2(n, \lambda) \quad \text{for} \quad \lambda = \frac{1}{2}(\boldsymbol{\mu} - \mathbf{c})^T\Sigma^{-1}(\boldsymbol{\mu} - \mathbf{c}).$$

The classical linear model has the form $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$, where $\boldsymbol{\mu}$ is assumed to lie in a subspace $V = \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k) = C(\mathbf{X})$. That is, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathcal{R}^k$.

Therefore, we'll be interested in statistical properties of $\hat{\mathbf{y}}_V = p(\mathbf{y}|V)$, the projection of $\mathbf{y}$ onto V, and of functions of $\hat{\mathbf{y}}_V$ and the residual vector $\mathbf{y} - \hat{\mathbf{y}}_V = p(\mathbf{y}|V^\perp)$.

- The distributional form (normal, chi-square, etc.) of functions of $\mathbf{y}$ (e.g., $\hat{\mathbf{y}}_V$) are determined by the distributional form of $\mathbf{y}$ (usually assumed normal).
- The expectation of linear functions of $\mathbf{y}$ is determined solely by $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$.
- The expectation of quadratic functions of $\mathbf{y}$ (e.g., $\|\hat{\mathbf{y}}_V\|^2$) is determined by $\boldsymbol{\mu}$ and $\mathrm{var}(\mathbf{y})$.

In particular, we have the following results

Theorem: Let V be a $k-$dimensional subspace of $\mathcal{R}^n$, and let $\mathbf{y}$ be a random vector in $\mathcal{R}^n$ with mean $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$. Then

1. $\mathrm{E}\{p(\mathbf{y}|V)\} = p(\boldsymbol{\mu}|V)$;

2. if $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}_n$ then

$$\mathrm{var}\{p(\mathbf{y}|V)\} = \sigma^2\mathbf{P}_V \quad \text{and} \quad \mathrm{E}\left\{\|p(\mathbf{y}|V)\|^2\right\} = \sigma^2 k + \|p(\boldsymbol{\mu}|V)\|^2;$$

   and

3. if we assume additionally that $\mathbf{y}$ is m'variate normal i.e., $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$, then

$$p(\mathbf{y}|V) \sim N_n(p(\boldsymbol{\mu}|V), \sigma^2\mathbf{P}_V),$$

   and

$$\frac{1}{\sigma^2}\|p(\mathbf{y}|V)\|^2 = \frac{1}{\sigma^2}\mathbf{y}^T\mathbf{P}_V\mathbf{y} \sim \chi^2(k, \frac{1}{2\sigma^2}\underbrace{\boldsymbol{\mu}^T\mathbf{P}_V\boldsymbol{\mu}}_{=\|p(\boldsymbol{\mu}|V)\|^2}).$$

Proof:

1. Since the projection operation is linear, $\mathrm{E}\{p(\mathbf{y}|V)\} = p(\mathrm{E}(\mathbf{y})|V) = p(\boldsymbol{\mu}|V)$.

2. $p(\mathbf{y}|V) = \mathbf{P}_V\mathbf{y}$ so $\mathrm{var}\{p(\mathbf{y}|V)\} = \mathrm{var}(\mathbf{P}_V\mathbf{y}) = \mathbf{P}_V\sigma^2\mathbf{I}_n\mathbf{P}_V^T = \sigma^2\mathbf{P}_V$. In addition, $\|p(\mathbf{y}|V)\|^2 = p(\mathbf{y}|V)^T p(\mathbf{y}|V) = (\mathbf{P}_V\mathbf{y})^T\mathbf{P}_V\mathbf{y} = \mathbf{y}^T\mathbf{P}_V\mathbf{y}$. So, $\mathrm{E}(\|p(\mathbf{y}|V)\|^2) = \mathrm{E}(\mathbf{y}^T\mathbf{P}_V\mathbf{y})$ is the expectation of a quadratic form and therefore equals

$$\begin{aligned}\mathrm{E}(\|p(\mathbf{y}|V)\|^2) = \mathrm{tr}(\sigma^2\mathbf{P}_V) + \boldsymbol{\mu}^T\mathbf{P}_V\boldsymbol{\mu} &= \sigma^2\mathrm{tr}(\mathbf{P}_V) + \boldsymbol{\mu}^T\mathbf{P}_V^T\mathbf{P}_V\boldsymbol{\mu} \\ &= \sigma^2 k + \|p(\boldsymbol{\mu}|V)\|^2.\end{aligned}$$

3. Result 3 is just a restatement of corollary 3 above, and follows immediately from the Theorem on the bottom of p. 82. ∎

So, we have distributional results for a projection and for the squared length of that projection. In linear models we typically decompose the sample space by projecting onto the model space V to form $\hat{\mathbf{y}}_V = p(\mathbf{y}|V)$ and onto its orthogonal complement $V^\perp$ to form the residual $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}_V$. In some cases we go further and decompose the model space by projecting onto subspaces within the model space.

What's the joint distribution of such projections?

Well, it turns out that if the subspaces are orthogonal and if the conditions of the classical linear model (normality, independence, and homoscedasticity) hold, then the projections onto these subspaces are independent normal random vectors, and their squared lengths (the sums of squares in an ANOVA) are independent chi-square random variables.

So, if we understand the geometry underlying our linear model (e.g., underlying an ANOVA we'd like to do for a particular linear model), then we can use the following result:

Theorem: Let $V_1, \ldots, V_k$ be mutually orthogonal subspaces of $\mathcal{R}^n$ with dimensions $d_1, \ldots, d_k$, respectively, and let $\mathbf{y}$ be a random vector taking values in $\mathcal{R}^n$ which has mean $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$. Let $\mathbf{P}_i$ be the projection matrix onto V_i so that $\hat{\mathbf{y}}_i = p(\mathbf{y}|V_i) = \mathbf{P}_i\mathbf{y}$ and let $\boldsymbol{\mu}_i = \mathbf{P}_i\boldsymbol{\mu}$, $i = 1, \ldots, n$. Then

1. if $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}_n$ then $\mathrm{cov}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j) = \mathbf{0}$, for $i \neq j$; and

2. if $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$ then $\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_k$ are independent, with

$$\hat{\mathbf{y}}_i \sim N(\boldsymbol{\mu}_i, \sigma^2\mathbf{P}_i);$$

   and

3. if $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$ then $\|\hat{\mathbf{y}}_1\|^2, \ldots, \|\hat{\mathbf{y}}_k\|^2$ are independent, with

$$\frac{1}{\sigma^2}\|\hat{\mathbf{y}}_i\|^2 \sim \chi^2(d_i, \frac{1}{2\sigma^2}\|\boldsymbol{\mu}_i\|^2).$$

Proof: Part 1: For $i \neq j$, $\text{cov}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j) = \text{cov}(\mathbf{P}_i\mathbf{y}, \mathbf{P}_j\mathbf{y}) = \mathbf{P}_i\text{cov}(\mathbf{y}, \mathbf{y})\mathbf{P}_j = \mathbf{P}_i\sigma^2\mathbf{I}\mathbf{P}_j = \sigma^2\mathbf{P}_i\mathbf{P}_j = \mathbf{0}$. (For any $\mathbf{z} \in \mathcal{R}^n$, $\mathbf{P}_i\mathbf{P}_j\mathbf{z} = \mathbf{0} \Rightarrow \mathbf{P}_i\mathbf{P}_j = \mathbf{0}$.)

Part 2: If $\mathbf{y}$ is m'variate normal then $\hat{\mathbf{y}}_i = \mathbf{P}_i\mathbf{y}$, $i = 1, \ldots, k$, are jointly multivariate normal and are therefore independent if and only if $\text{cov}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j) = \mathbf{0}$, $i \neq j$. The mean and variance-covariance of $\hat{\mathbf{y}}_i$ are $\text{E}(\hat{\mathbf{y}}_i) = \text{E}(\mathbf{P}_i\mathbf{y}) = \mathbf{P}_i\boldsymbol{\mu} = \boldsymbol{\mu}_i$ and $\text{var}(\hat{\mathbf{y}}_i) = \mathbf{P}_i\sigma^2\mathbf{I}\mathbf{P}_i^T = \sigma^2\mathbf{P}_i$.

Part 3: If $\hat{\mathbf{y}}_i = \mathbf{P}_i\mathbf{y}$, $i = 1, \ldots, k$, are mutually independent, then any (measurable*) functions $f_i(\hat{\mathbf{y}}_i)$, $i = 1, \ldots, k$, are mutually independent. Thus $\|\hat{\mathbf{y}}_i\|^2$, $i = 1, \ldots, k$, are mutually independent. That $\sigma^{-2}\|\hat{\mathbf{y}}_i\|^2 \sim \chi^2(d_i, \frac{1}{2\sigma^2}\|\boldsymbol{\mu}_i\|^2)$ follows from part 3 of the previous theorem. ∎

Alternatively, we can take an algebraic approach to determining whether projections and their squared lengths (in, for example, an ANOVA) are independent. The geometric approach is easier, perhaps, but only if you understand the geometry. But we will describe the algebraic approach as well (next).

Independence of Linear and Quadratic Forms:

Here we consider the statistical independence of:

1. a linear form and a quadratic form (e.g., consider whether $\bar{y}$ and s^2 in a one sample problem are independent; or consider the independence of $\hat{\boldsymbol{\beta}}$ and s^2 in a regression setting);

2. two quadratic forms (e.g., consider the independence of the sum of squares due to regression and the error sum of squares in a regression problem); and

3. several quadratic forms (e.g., consider the joint distribution of SS_A, SS_B, SS_{AB}, SS_E in a two-way layout analysis of variance).

* All continuous functions, and most "well-behaved" functions are measurable

Before proceeding, we need a Lemma and its corollary:

Lemma: If $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \Sigma)$, then

$$\text{cov}(\mathbf{y}, \mathbf{y}^T\mathbf{A}\mathbf{y}) = 2\Sigma\mathbf{A}\boldsymbol{\mu}.$$

Proof: Using the definition of covariance and the expectation of a quadratic form, we have

$$\begin{aligned}\text{cov}(\mathbf{y}, \mathbf{y}^T\mathbf{A}\mathbf{y}) &= \text{E}[(\mathbf{y} - \boldsymbol{\mu})\{\mathbf{y}^T\mathbf{A}\mathbf{y} - \text{E}(\mathbf{y}^T\mathbf{A}\mathbf{y})\}] \\ &= \text{E}[(\mathbf{y} - \boldsymbol{\mu})\{\mathbf{y}^T\mathbf{A}\mathbf{y} - \text{tr}(\mathbf{A}\Sigma) - \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}\}].\end{aligned}$$

Now using the algebraic identity $\mathbf{y}^T\mathbf{A}\mathbf{y} - \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} = (\mathbf{y} - \boldsymbol{\mu})^T\mathbf{A}(\mathbf{y} - \boldsymbol{\mu}) + 2(\mathbf{y} - \boldsymbol{\mu})^T\mathbf{A}\boldsymbol{\mu}$, we obtain

$$\begin{aligned}\text{cov}(\mathbf{y}, \mathbf{y}^T\mathbf{A}\mathbf{y}) &= \text{E}[(\mathbf{y} - \boldsymbol{\mu})\{(\mathbf{y} - \boldsymbol{\mu})^T\mathbf{A}(\mathbf{y} - \boldsymbol{\mu}) + 2(\mathbf{y} - \boldsymbol{\mu})^T\mathbf{A}\boldsymbol{\mu} - \text{tr}(\mathbf{A}\Sigma)\}] \\ &= \text{E}\{(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T\mathbf{A}(\mathbf{y} - \boldsymbol{\mu})\} + 2\text{E}\{(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T\mathbf{A}\boldsymbol{\mu}\} \\ &\quad - \text{E}\{(\mathbf{y} - \boldsymbol{\mu})\text{tr}(\mathbf{A}\Sigma)\} \\ &= \mathbf{0} + 2\Sigma\mathbf{A}\boldsymbol{\mu} - \mathbf{0}.\end{aligned}$$

The first term equals $\mathbf{0}$ here because all third central moments of the multivariate normal distribution are 0. ∎

Corollary: Let $\mathbf{B}$ be a $k \times n$ matrix of constants and $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \Sigma)$. Then

$$\text{cov}(\mathbf{B}\mathbf{y}, \mathbf{y}^T\mathbf{A}\mathbf{y}) = 2\mathbf{B}\Sigma\mathbf{A}\boldsymbol{\mu}.$$

∎

Now we are ready to consider (1.):

Theorem: Suppose $\mathbf{B}$ is a $k \times n$ matrix of constants, $\mathbf{A}$ a $n \times n$ symmetric matrix of constants, and $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \Sigma)$. Then $\mathbf{B}\mathbf{y}$ and $\mathbf{y}^T\mathbf{A}\mathbf{y}$ are independent if and only if $\mathbf{B}\Sigma\mathbf{A} = \mathbf{0}_{k\times n}$.

Proof: We prove this theorem under the additional assumption that $\mathbf{A}$ is a projection matrix (that is, assuming $\mathbf{A}$ is idempotent as well as symmetric), which is the situation in which we're most interested in this course. See Searle (1971, p.59) for the complete proof.

Assuming $\mathbf{A}$ is symmetric and idempotent, then we have

$$\mathbf{y}^T\mathbf{A}\mathbf{y} = \mathbf{y}^T\mathbf{A}^T\mathbf{A}\mathbf{y} = \|\mathbf{A}\mathbf{y}\|^2.$$

Now suppose $\mathbf{B}\Sigma\mathbf{A} = \mathbf{0}$. Then $\mathbf{B}\mathbf{y}$ and $\mathbf{A}\mathbf{y}$ are each normal, with

$$\text{cov}(\mathbf{B}\mathbf{y}, \mathbf{A}\mathbf{y}) = \mathbf{B}\Sigma\mathbf{A} = \mathbf{0}.$$

Therefore, $\mathbf{B}\mathbf{y}$ and $\mathbf{A}\mathbf{y}$ are independent of one another. Furthermore, $\mathbf{B}\mathbf{y}$ is independent of any (measureable) function of $\mathbf{A}\mathbf{y}$, so that $\mathbf{B}\mathbf{y}$ is independent of $\|\mathbf{A}\mathbf{y}\|^2 = \mathbf{y}^T\mathbf{A}\mathbf{y}$.
Now for the converse. Suppose $\mathbf{B}\mathbf{y}$ and $\mathbf{y}^T\mathbf{A}\mathbf{y}$ are independent. Then $\text{cov}(\mathbf{B}\mathbf{y}, \mathbf{y}^T\mathbf{A}\mathbf{y}) = \mathbf{0}$ so

$$\mathbf{0} = \text{cov}(\mathbf{B}\mathbf{y}, \mathbf{y}^T\mathbf{A}\mathbf{y}) = 2\mathbf{B}\Sigma\mathbf{A}\boldsymbol{\mu},$$

by the corollary to the lemma above. Since this holds for all possible $\boldsymbol{\mu}$, it follows that $\mathbf{B}\Sigma\mathbf{A} = \mathbf{0}$. ∎

Corollary: If $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, then $\mathbf{B}\mathbf{y}$ and $\mathbf{y}^T\mathbf{A}\mathbf{y}$ are independent if and only if $\mathbf{B}\mathbf{A} = \mathbf{0}$.

Example: Suppose we have a random sample $\mathbf{y} = (y_1, \ldots, y_n)^T \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$, and consider

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\mathbf{j}_n^T\mathbf{y}$$

and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\|\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)^\perp}\mathbf{y}\|^2$$

where $\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)^\perp} = \mathbf{I}_n - \frac{1}{n}\mathbf{j}_n\mathbf{j}_n^T$. By the above corollary, $\bar{y}$ and s^2 are independent because

$$\frac{1}{n}\mathbf{j}_n^T\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)^\perp} = \frac{1}{n}(\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)^\perp}^T\mathbf{j}_n)^T = (\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)^\perp}\mathbf{j}_n)^T = \frac{1}{n}(\mathbf{0})^T = \mathbf{0}^T.$$

Theorem: Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric matrices of constants. If $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \Sigma)$, then $\mathbf{y}^T\mathbf{A}\mathbf{y}$ and $\mathbf{y}^T\mathbf{B}\mathbf{y}$ are independent if and only if $\mathbf{A}\Sigma\mathbf{B} = \mathbf{0}$.

Proof: Again, we consider the special case when $\mathbf{A}$ and $\mathbf{B}$ are symmetric and idempotent (projection matrices), and we only present the "if" part of the proof (see Searle, 1971, pp. 59–60, for complete proof).

Suppose $\mathbf{A}$ and $\mathbf{B}$ are symmetric and idempotent and that $\mathbf{A}\Sigma\mathbf{B} = \mathbf{0}$. Then $\mathbf{y}^T\mathbf{A}\mathbf{y} = \|\mathbf{A}\mathbf{y}\|^2$, $\mathbf{y}^T\mathbf{B}\mathbf{y} = \|\mathbf{B}\mathbf{y}\|^2$. If $\mathbf{A}\Sigma\mathbf{B} = \mathbf{0}$ then

$$\text{cov}(\mathbf{A}\mathbf{y}, \mathbf{B}\mathbf{y}) = \mathbf{A}\Sigma\mathbf{B} = \mathbf{0}.$$

Each of $\mathbf{A}\mathbf{y}$ and $\mathbf{B}\mathbf{y}$ are multivariate normal and they have covariance $\mathbf{0}$, hence they're independent. Furthermore, any (measurable) functions of $\mathbf{A}\mathbf{y}$ and $\mathbf{B}\mathbf{y}$ are independent, so that $\|\mathbf{A}\mathbf{y}\|^2$ and $\|\mathbf{B}\mathbf{y}\|^2$ are independent. ∎

Corollary: If $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ then $\mathbf{y}^T\mathbf{A}\mathbf{y}$ and $\mathbf{y}^T\mathbf{B}\mathbf{y}$ are independent if and only if $\mathbf{A}\mathbf{B} = \mathbf{0}$. ∎

Finally, we have a theorem and corollary concerning the mutual independence of several quadratic forms in normal random vectors:

Theorem: Let $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$, let $\mathbf{A}_i$ be symmetric of rank r_i, $i = 1, \ldots, k$, and let $\mathbf{A} = \sum_{i=1}^k \mathbf{A}_i$ with rank r so that $\mathbf{y}^T\mathbf{A}\mathbf{y} = \sum_{i=1}^k \mathbf{y}^T\mathbf{A}_i\mathbf{y}$. Then

1. $\mathbf{y}^T\mathbf{A}_i\mathbf{y}/\sigma^2 \sim \chi^2(r_i, \boldsymbol{\mu}^T\mathbf{A}_i\boldsymbol{\mu}/\{2\sigma^2\})$, $i = 1, \ldots, k$; and
2. $\mathbf{y}^T\mathbf{A}_i\mathbf{y}$ and $\mathbf{y}^T\mathbf{A}_j\mathbf{y}$ are independent for all $i \neq j$; and
3. $\mathbf{y}^T\mathbf{A}\mathbf{y}/\sigma^2 \sim \chi^2(r, \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}/\{2\sigma^2\})$;

if and only if any two of the following statements are true:

a. each $\mathbf{A}_i$ is idempotent;
b. $\mathbf{A}_i\mathbf{A}_j = \mathbf{0}$ for all $i \neq j$;
c. $\mathbf{A}$ is idempotent;

or if and only if (c) and (d) are true where (d) is as follows:

d. $r = \sum_{i=1}^k r_i$.

Proof: See Searle (1971, pp. 61–64). ∎

- Note that any two of (a), (b), and (c) implies the third.

- The previous theorem concerned the partitioning of a weighted sum of squares $\mathbf{y}^T\mathbf{A}\mathbf{y}$ into several components. We now state a corollary that treats the special case where $\mathbf{A} = \mathbf{I}$ and the total (unweighted sum of squares) $\mathbf{y}^T\mathbf{y}$ is decomposed into a sum of quadratic forms.

Corollary: Let $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, let $\mathbf{A}_i$ be symmetric of rank r_i, $i = 1, \ldots, k$, and suppose that $\mathbf{y}^T\mathbf{y} = \sum_{i=1}^k \mathbf{y}^T\mathbf{A}_i\mathbf{y}$ (i.e., $\sum_{i=1}^k \mathbf{A}_i = \mathbf{I}$). Then

1. each $\mathbf{y}^T\mathbf{A}_i\mathbf{y} \sim \chi^2(r_i, \boldsymbol{\mu}^T\mathbf{A}_i\boldsymbol{\mu}/\{2\sigma^2\})$; and
2. the $\mathbf{y}^T\mathbf{A}_i\mathbf{y}$'s are mutually independent;

if and only if any one of the following statements holds:

a. each $\mathbf{A}_i$ is idempotent;
b. $\mathbf{A}_i\mathbf{A}_j = \mathbf{0}$ for all $i \neq j$;
c. $n = \sum_{i=1}^k r_i$.

∎

- This theorem subsumes "Cochran's Theorem" which is often cited as the justification for independence of sums of squares in a decomposition of the total sum of squares in an analysis of variance.

F-Distribution: Let

$$U_1 \sim \chi^2(n_1, \lambda), \qquad U_2 \sim \chi^2(n_2) \quad \text{(central)}$$

be independent. Then

$$V = \frac{U_1/n_1}{U_2/n_2}$$

is said to have a noncentral F distribution with noncentrality parameter λ, and n_1 and n_2 degrees of freedom.

- For $\lambda = 0$, V is said to have a central F distribution with degrees of freedom n_1 and n_2.
- The noncentral F has three parameters: λ, the noncentrality parameter, n_1, the numerator degrees of freedom, and n_2, the denominator degrees of freedom. The central F has two parameters, the numerator and denominator degrees of freedom, n_1 and n_2.
- We'll denote the noncentral F by $F(n_1, n_2, \lambda)$ and the central F by $F(n_1, n_2)$. The $100\gamma^{\text{th}}$ percentile of the $F(n_1, n_2)$ distribution will be denoted $F_\gamma(n_1, n_2)$.
- The p.d.f. of the noncentral F is an ugly looking thing not worth reproducing here. It has the form of a Poisson mixture of central F's.
- The mean of the noncentral $F(n_1, n_2, \lambda)$ is $\frac{n_2}{n_2-2}(1 + 2\lambda/n_1)$. Its variance is much more complicated (see Stapleton, p.67, if you're interested).

t-Distribution: Let

$$W \sim N(\mu, 1), \qquad Y \sim \chi^2(m)$$

be independent random variables. Then

$$T = \frac{W}{\sqrt{Y/m}}$$

is said to have a (Student's) t distribution with noncentrality parameter μ and m degrees of freedom.

- We'll denote this distribution as $t(m, \mu)$.
- If the numerator random variable W has distribution $N(\mu, \sigma^2)$ then the noncentrality parameter becomes μ/σ since $W/\sigma \sim (\mu/\sigma, 1)$.
- Again, when the noncentrality parameter μ is zero, we get the central t distribution, $t(m) \equiv t(m, 0)$. The $100\gamma^{\text{th}}$ percentile of the central t will be denoted $t_\gamma(m)$.
- The p.d.f. of the noncentral t is too complicated to be worth becoming familiar with. It may be found in Stapleton, p.68.
- The mean and variance of the central $t(m)$ distribution are 0, and $m/(m-2)$, respectively. For the noncentral t, these moments are not so simple.
- We can think of the central $t(m)$ has a more dispersed version of the standard normal distribution with fatter tails. As $m \to \infty$, $m/(m-2) \to 1$, so the $t(m)$ converges to the $N(0, 1)$.

Notice the **important relationship** that the square of a r.v. with a $t(m)$ distribution has an $F(1, m)$ distribution:

If

$$T = \frac{W}{\sqrt{Y/m}} \sim t(m),$$

then

$$T^2 = \frac{W^2/1}{Y/m} \sim F(1, m).$$

Example: Let $Y_1, \ldots, Y_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. The following results should be familiar to you from earlier coursework (we will prove them in the following theorem): For $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$,

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1), \quad \text{and } \bar{Y}, S^2 \text{ are independent.}$$

Let μ_0 be a constant, and define $\theta = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$. Then

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim \begin{cases} t(n-1), & \text{if } \mu_0 = \mu; \\ t(n-1, \theta), & \text{otherwise.} \end{cases}$$

Here's the theorem establishing these results, and providing the basis for the one-sample $t-$test.

Theorem: Let $Y_1, \ldots, Y_n$ be a random sample (i.i.d. r.v.'s) from a $N(\mu, \sigma^2)$ distribution, and let $\bar{Y}, S^2$, and T be defined as above. Then

1. $\bar{Y} \sim N(\mu, \sigma^2/n)$;
2. $V = \frac{S^2(n-1)}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \chi^2(n-1)$;
3. $\bar{Y}$ and S^2 are independent; and
4. $T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t(n-1, \lambda)$ where $\lambda = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$ for any constant μ_0.

Proof: Part 1: By assumption, $\mathbf{y} = (Y_1, \ldots, Y_n)^T \sim N_n(\mu \mathbf{j}_n, \sigma^2 \mathbf{I}_n)$. Let $V = \mathcal{L}(\mathbf{j}_n)$, a 1-dimensional subspace of $\mathcal{R}^n$. Then $p(\mathbf{y}|V) = \bar{Y}\mathbf{j}_n$. $\bar{Y} = n^{-1}\mathbf{j}_n^T \mathbf{y}$ is an affine transformation of $\mathbf{y}$ so it is normal, with mean $n^{-1}\mathbf{j}_n^T \mathrm{E}(\mathbf{y}) = \mu$ and variance $n^{-2}\mathbf{j}_n^T \sigma^2 \mathbf{I}_n \mathbf{j}_n = \sigma^2/n$.

Part 3: By part 2 of the theorem on p. 86, $p(\mathbf{y}|V^\perp) = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})^T$ is independent of $p(\mathbf{y}|V) = \bar{Y}\mathbf{j}_n$, and hence independent of $\bar{Y}$. Thus, S^2, a function of $p(\mathbf{y}|V^\perp)$, and $\bar{Y}$ are independent. Alternatively, we could use the corollary on p. 89 giving the necessary and sufficient condition for independence of a linear form ($\bar{Y}$) and a quadratic form (S^2). See the example on p. 89.

Part 2: Here we use part 3 of the theorem on p. 85. That result implies that $\frac{1}{\sigma^2}S^2(n-1) = \frac{1}{\sigma^2}\|p(\mathbf{y}|V^\perp)\|^2 \sim \chi^2(\dim(V^\perp), \lambda) = \chi^2(n-1, \lambda)$ and

$$\lambda = \frac{1}{2\sigma^2}\|p(\mathrm{E}(\mathbf{y})|V^\perp)\|^2 = \frac{1}{2\sigma^2}\| \underbrace{p(\underbrace{\mu\mathbf{j}_n}_{\in V}|V^\perp)}_{=\mathbf{0}} \|^2 = 0.$$

Part 4: Let $U = \frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}$. Then $U \sim N\left(\frac{\mu-\mu_0}{\sigma/\sqrt{n}}, 1\right)$. From part 3 of this theorem, U and V are independent. Note that T can be written as

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} = \frac{\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} = \frac{U}{\sqrt{V/(n-1)}},$$

so by the definition of the noncentral t distribution, $T \sim t\left(n-1, \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right)$.
■

Finally, the following theorem puts together some of our results on independence of squared lengths of projections (quadratic forms) and the definition of the F distribution to give the distribution of F-tests in the ANOVA:

Theorem: Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{X}$ is $n \times k$ with rank $r \leq k$. and $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Let V_1 be a subspace of $C(\mathbf{X})$ with $\dim(V_1) = r_1$ which is smaller than $\dim(C(\mathbf{X})) = r$. Let $\hat{\mathbf{y}} = p(\mathbf{y}|C(\mathbf{X}))$. Then the random variable

$$F = \frac{\|p(\mathbf{y}|V_1)\|^2/r_1}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-r)} \sim F(r_1, n-r, \|p(\mathbf{X}\boldsymbol{\beta}|V_1)\|^2/\{2\sigma^2\}).$$

Proof: Since $\mathbf{y} - \hat{\mathbf{y}} = p(\mathbf{y}|C(\mathbf{X})^\perp)$, it follows from part 3 of the theorem on p. 86 that $Q_1 \equiv \|p(\mathbf{y}|V_1)\|^2$ and $Q_2 \equiv \|p(\mathbf{y}|C(\mathbf{X})^\perp)\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$ are independent random variables and $Q_1/\sigma^2 \sim \chi^2(r_1, \|p(\mathbf{X}\boldsymbol{\beta}|V_1)\|^2/\{2\sigma^2\})$ and $Q_2/\sigma^2 \sim \chi^2(n-r)$. Thus, by the definition of the non-central F distribution, the result follows. ∎

The Linear Model

The Full Rank Case: Multiple Linear Regression

Suppose that on a random sample of n units (patients, animals, trees, etc.) we observe a response variable Y and explanatory variables $X_1, \ldots, X_k$.

Our data are then $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, or, in vector/matrix form $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_k$ where $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$; or $\mathbf{y}, \mathbf{X}$ where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$.

Either by design or by conditioning on their observed values, $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are regarded as vectors of known constants.

The linear model in its classical form makes the following assumptions:

A1. (additive error) $\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}$ where $\mathbf{e} = (e_1, \ldots, e_n)^T$ is an unobserved random vector with $\mathrm{E}(\mathbf{e}) = \mathbf{0}$. This implies that $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y})$ is the unknown mean of $\mathbf{y}$.

A2. (linearity) $\boldsymbol{\mu} = \beta_1\mathbf{x}_1 + \cdots + \beta_k\mathbf{x}_k = \mathbf{X}\boldsymbol{\beta}$ where $\beta_1, \ldots, \beta_k$ are unknown parameters. This assumption says that $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} \in \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k) = C(\mathbf{X})$ lies in the column space of $\mathbf{X}$; i.e., it is a linear combination of explanatory vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ with coefficients the unknown parameters in $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$.

- Linear in $\beta_1, \ldots, \beta_k$ not in the x's.

A3. (independence) $e_1, \ldots, e_n$ are independent random variables (and therefore so are $y_1, \ldots, y_n$).

A4. (homoscedasticity) $e_1, \ldots, e_n$ all have the same variance σ^2; that is, $\mathrm{var}(e_1) = \cdots = \mathrm{var}(e_n) = \sigma^2$ which implies $\mathrm{var}(y_1) = \cdots = \mathrm{var}(y_n) = \sigma^2$.

A5. (normality) $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

- Taken together, assumptions (A3) and (A4) say $\text{var}(\mathbf{y}) = \text{var}(\mathbf{e}) = \sigma^2\mathbf{I}_n$. We say that $\mathbf{y}$ (and $\mathbf{e}$) has a **spherical** variance-covariance matrix.

- Note that assumption (A5) subsumes (A3) and (A4), but we can obtain many useful results without invoking normality, so its useful to separate the assumptions of independence, homoscedasticity, and normality.

- Taken together, all five assumptions can be stated more succinctly as $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

- The unknown parameters of the model are $\beta_1, \ldots, \beta_k, \sigma^2$.

- In multiple linear regression models, there is typically an intercept term in the model. That is, one of the explanatory variables is equal to 1, for all i.

- This could be accommodated by just setting $\mathbf{x}_1 = \mathbf{j}_n$ and letting β_1 represent the intercept. However, we'll follow our book's convention and include the intercept as an additional term. Our model then becomes

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \quad i = 1, \ldots, n,$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{=\boldsymbol{\beta}} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

or

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

- In multiple regression, it is typical for the explanatory vectors $\mathbf{j}_n, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ to be LIN. At times this is violated, but usually because of the data we happen to observe rather than because of the structure of the model. That is, in contrast to ANOVA models, there is usually no structural dependence among the explanatory vectors in multiple regression. **Therefore, we will (for now) assume that X is of full rank.**

- Note that our model assumes that we can re-express $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} \in \mathcal{R}^n$ as $\mathbf{X}\boldsymbol{\beta} \in C(\mathbf{X})$ where $\dim(C(\mathbf{X})) = k+1$. Therefore, as long as $n > k+1$, our model involves some reduction or summarization by assuming that the n-element vector $\boldsymbol{\mu}$ falls in a $k+1$ dimensional subspace of $\mathcal{R}^n$.

- We will assume $n > k+1$ throughout our discussion of the linear model. If $n = k+1$ (when there are as many parameters in $\boldsymbol{\beta}$ as there are data points), then the model involves no data reduction at all, only a data transformation. A linear model with $k+1 > n$ parameters is useless.

Interpretation of the β_j's: The elements of $\boldsymbol{\beta}$ in a multiple linear regression model are usually called simply regression coefficients, but are more properly termed **partial regression coefficients**.

In the model

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}_{=\mu} + e,$$

note that $\frac{\partial \mu}{\partial x_k} = \beta_j$. That is, β_j represents the change in $\mathrm{E}(y) = \mu$ associated with a unit change in x_j, assuming all of the other x_k's are held constant.

In addition, this effect depends upon what other explanatory variables are present in the model. For example, β_0 and β_1 in the model

$$\mathbf{y} = \beta_0 \mathbf{j}_n + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \mathbf{e}$$

will typically be different than β_0^* and β_1^* in the model

$$\mathbf{y} = \beta_0^* \mathbf{j}_n + \beta_1^* \mathbf{x}_1 + \mathbf{e}^*.$$

- For example, consider again our reading ability and height example. If I regressed reading ability (r) on height (h) for children from grades 1–5,

$$r_i = \beta_0 + \beta_1 h_i + e_i \quad i = 1, \ldots, n,$$

I would expect to obtain a large positive (and significant) estimate $\hat{\beta}_1$. However, if we were to add age (a) to our model:

$$r_i = \beta_0^* + \beta_1^* h_i + \beta_2^* a_i + e_i^* \quad i = 1, \ldots, n,$$

I would expect $\hat{\beta}_1^* \approx 0$.

- The issue here is very similar to the distinction between correlation and partial correlation. However, regression coefficients quantify association between a random y and a fixed x_j. Correlation coefficients regard both y and x_j as random.

Estimation in the Full-Rank Linear Model

Recall in the classical linear model (CLM) we assume

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}, \quad \text{for} \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \quad \text{for some } \boldsymbol{\beta} \in \mathcal{R}^{k+1}, \quad \text{and} \quad \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

- Sometimes we are interested only in estimating $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$ (e.g., when focused on prediction).

- More often, however, the parameters $\beta_0, \beta_1, \ldots, \beta_k$, or some subset or function of the parameters are of interest in and of themselves as interpretable, meaningful quantities to be estimated. Therefore, we will focus on estimation of $\boldsymbol{\beta}$ rather than $\boldsymbol{\mu}$.

- However, the relationship between the mean parameter $\boldsymbol{\mu}$ and the regression parameter $\boldsymbol{\beta}$ is $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, so we can always estimate $\boldsymbol{\mu}$ once we estimate $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

- In addition, in the $\mathbf{X}$ of full rank case, we can write $\boldsymbol{\beta}$ in terms of $\boldsymbol{\mu}$: $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\mu}$, so we could just as well focus on estimation of $\boldsymbol{\mu}$ and then obtain $\hat{\boldsymbol{\beta}}$ as $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\mu}}$.

Least-Squares Estimation:

Recall that the projection of $\mathbf{y}$ onto $C(\mathbf{X})$, the set of all vectors of the form $\mathbf{Xb}$ for $\mathbf{b} \in \mathcal{R}^{k+1}$, yields the closest point in $C(\mathbf{X})$ to $\mathbf{y}$. That is, $p(\mathbf{y}|C(\mathbf{X}))$ yields the minimizer of

$$Q(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad \textbf{(the least squares criterion)}$$

This leads to the estimator $\hat{\boldsymbol{\beta}}$ given by the solution of

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \qquad \textbf{(the normal equations)}$$

or

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

All of this has already been established back when we studied projections (see pp. 30–31). Alternatively, we could use calculus:

To find a stationary point (maximum, minimum, or saddle point) of $Q(\boldsymbol{\beta})$, we set the partial derivative of $Q(\boldsymbol{\beta})$ equal to zero and solve:

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}}Q(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= \frac{\partial}{\partial\boldsymbol{\beta}}(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}) \\ &= \mathbf{0} - 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Here we've used the vector differentiation formulas $\frac{\partial}{\partial\mathbf{z}}\mathbf{c}^T\mathbf{z} = \mathbf{c}$ and $\frac{\partial}{\partial\mathbf{z}}\mathbf{z}^T\mathbf{A}\mathbf{z} = 2\mathbf{A}\mathbf{z}$ (see §2.14 of our text).

Setting this result equal to zero, we obtain the normal equations, which has solution $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. That this is a minimum rather than a max, or saddle point can be verified by checking the second derivative matrix of $Q(\boldsymbol{\beta})$:

$$\frac{\partial^2 Q(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 2\mathbf{X}^T\mathbf{X}$$

which is positive definite (result 7, p. 54), therefore $\hat{\boldsymbol{\beta}}$ is a minimum.

Example — Simple Linear Regression

Consider the case $k = 1$:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n$$

where $e_1, \ldots, e_n$ are i.i.d. each with mean 0 and variance σ^2. Then the model equation becomes

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{=\boldsymbol{\beta}} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

It follows that

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}, \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}$$
$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}.$$

Therefore, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ yields

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{pmatrix} (\sum_i x_i^2)(\sum_i y_i) - (\sum_i x_i)(\sum_i x_i y_i) \\ -(\sum_i x_i)(\sum_i y_i) + n\sum_i x_i y_i \end{pmatrix}.$$

After a bit of algebra, these estimators simplify to

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$
$$\text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

In the case that $\mathbf{X}$ is of full rank, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \quad \hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{P}_{C(\mathbf{X})}\mathbf{y}.$$

- Notice that both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ are linear functions of $\mathbf{y}$. That is, in each case the estimator is given by some matrix times $\mathbf{y}$.

Note also that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}.$$

From this representation several important properties of the least squares estimator $\hat{\boldsymbol{\beta}}$ follow easily:

1. (unbiasedness):

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathrm{E}(\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}) = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \underbrace{\mathrm{E}(\mathbf{e})}_{=\mathbf{0}} = \boldsymbol{\beta}.$$

2. (var-cov matrix)

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}) &= \mathrm{var}(\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \underbrace{\mathrm{var}(\mathbf{e})}_{=\sigma^2\mathbf{I}} \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\end{aligned}$$

3. (normality) $\hat{\boldsymbol{\beta}} \sim N_k(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ (if $\mathbf{e}$ is assumed normal).

- These three properties require increasingly strong assumptions. Property (1) holds under assumptions A1 and A2 (additive error and linearity).
- Property (2) requires, in addition, the assumption of sphericity.
- Property (3) requires assumption A5 (normality). However, later we will present a central limit theorem-like result that establishes the *asymptotic* normality of $\hat{\boldsymbol{\beta}}$ under certain conditions even when $\mathbf{e}$ is not normal.

Example — Simple Linear Regression (Continued)

Result 2 on the previous page says for $\text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$, $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Therefore, in the simple linear regression case,

$$\begin{aligned}
\text{var}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \frac{\sigma^2}{n\sum_i x_i^2 - (\sum_i x_i)^2}\begin{pmatrix}\sum_i x_i^2 & -\sum_i x_i\\ -\sum_i x_i & n\end{pmatrix} \\
&= \frac{\sigma^2}{\sum_i (x_i-\bar{x})^2}\begin{pmatrix}n^{-1}\sum_i x_i^2 & -\bar{x}\\ -\bar{x} & 1\end{pmatrix}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\text{var}(\hat{\beta}_0) &= \frac{\sigma^2\sum_i x_i^2/n}{\sum_i(x_i-\bar{x})^2} = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i(x_i-\bar{x})^2}\right], \\
\text{var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_i(x_i-\bar{x})^2}, \\
\text{and}\quad \text{cov}(\hat{\beta}_0,\hat{\beta}_1) &= \frac{-\sigma^2\bar{x}}{\sum_i(x_i-\bar{x})^2}
\end{aligned}$$

- Note that if $\bar{x} > 0$, then $\text{cov}(\hat{\beta}_0,\hat{\beta}_1)$ is negative, meaning that the slope and intercept are inversely related. That is, over repeated samples from the same model, the intercept will tend to decrease when the slope increases.

Gauss-Markov Theorem:

We have seen that in the spherical errors, full-rank linear model, the least-squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is unbiased and it is a linear estimator.

The following theorem states that in the class of linear and unbiased estimators, the least-squares estimator is optimal (or best) in the sense that it has minimum variance among all estimators in this class.

Gauss-Markov Theorem: Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{e}$ where $\mathbf{X}$ is $n \times (k+1)$ of rank $k+1$, where $n > k+1$, $\mathrm{E}(\mathbf{e}) = \mathbf{0}$, and $\mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{I}$. The least-squares estimators $\hat{\beta}_j,\ j = 0, 1, \ldots, k$ (the elements of $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ have minimum variance among all linear unbiased estimators.

Proof: Write $\hat{\beta}_j$ as $\hat{\beta}_j = \mathbf{c}^T\hat{\boldsymbol{\beta}}$ where $\mathbf{c}$ is the indicator vector containing a 1 in the $(j+1)$st position and 0's elsewhere. Then $\hat{\beta}_j = \mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{a}^T\mathbf{y}$ where $\mathbf{a} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}$. The quantity being estimated is $\beta_j = \mathbf{c}^T\boldsymbol{\beta} = \mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\mu} = \mathbf{a}^T\boldsymbol{\mu}$ where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

Consider an arbitrary linear estimator $\tilde{\beta}_j = \mathbf{d}^T\mathbf{y}$ of β_j. For such an estimator to be unbiased, it must satisfy $\mathrm{E}(\tilde{\beta}_j) = \mathrm{E}(\mathbf{d}^T\mathbf{y}) = \mathbf{d}^T\boldsymbol{\mu} = \mathbf{a}^T\boldsymbol{\mu}$ for any $\boldsymbol{\mu} \in C(\mathbf{X})$. I.e.,

$$\mathbf{d}^T\boldsymbol{\mu} - \mathbf{a}^T\boldsymbol{\mu} = \mathbf{0} \Rightarrow (\mathbf{d}-\mathbf{a})^T\boldsymbol{\mu} = \mathbf{0} \quad \text{for all } \boldsymbol{\mu} \in C(\mathbf{X}),$$

or $(\mathbf{d}-\mathbf{a}) \perp C(\mathbf{X})$. Then

$$\tilde{\beta}_j = \mathbf{d}^T\mathbf{y} = \mathbf{a}^T\mathbf{y} + (\mathbf{d}-\mathbf{a})^T\mathbf{y} = \hat{\beta}_j + (\mathbf{d}-\mathbf{a})^T\mathbf{y}.$$

The random variables on the right-hand side, $\hat{\beta}_j$ and $(\mathbf{d}-\mathbf{a})^T\mathbf{y}$, have covariance

$$\mathrm{cov}(\mathbf{a}^T\mathbf{y}, (\mathbf{d}-\mathbf{a})^T\mathbf{y}) = \mathbf{a}^T\mathrm{var}(\mathbf{y})(\mathbf{d}-\mathbf{a}) = \sigma^2\mathbf{a}^T(\mathbf{d}-\mathbf{a}) = \sigma^2(\mathbf{d}^T\mathbf{a} - \mathbf{a}^T\mathbf{a}).$$

Since $\mathbf{d}^T\boldsymbol{\mu} = \mathbf{a}^T\boldsymbol{\mu}$ for any $\boldsymbol{\mu} \in C(\mathbf{X})$ and $\mathbf{a} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c} \in C(\mathbf{X})$, it follows that $\mathbf{d}^T\mathbf{a} = \mathbf{a}^T\mathbf{a}$ so that

$$\mathrm{cov}(\mathbf{a}^T\mathbf{y}, (\mathbf{d}-\mathbf{a})^T\mathbf{y}) = \sigma^2(\mathbf{d}^T\mathbf{a} - \mathbf{a}^T\mathbf{a}) = \sigma^2(\mathbf{a}^T\mathbf{a} - \mathbf{a}^T\mathbf{a}) = \mathbf{0}.$$

It follows that

$$\text{var}(\tilde{\beta}_j) = \text{var}(\hat{\beta}_j) + \text{var}((\mathbf{d}-\mathbf{a})^T\mathbf{y}) = \text{var}(\hat{\beta}_j) + \sigma^2||\mathbf{d}-\mathbf{a}||^2.$$

Therefore, $\text{var}(\tilde{\beta}_j) \geq \text{var}(\hat{\beta}_j)$ with equality if and only if $\mathbf{d} = \mathbf{a}$, or equivalently, if and only if $\tilde{\beta}_j = \hat{\beta}_j$. ∎

Comments:

1. Notice that nowhere in this proof did we make use of the specific form of $\mathbf{c}$ as an indicator for one of the elements of $\boldsymbol{\beta}$. That is, we have proved a slightly more general result than that given in the statement of the theorem. We have proved that $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ is the minimum variance estimator in the class of linear unbiased estimators of $\mathbf{c}^T\boldsymbol{\beta}$ *for any* vector of constant $\mathbf{c}$.

2. The least-squares estimator $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is often called the B.L.U.E. (best linear unbiased estimator) of $\mathbf{c}^T\boldsymbol{\beta}$. Sometimes, it is called the Gauss-Markov estimator.

3. The variance of the BLUE is

   $$\text{var}(\mathbf{c}^T\hat{\boldsymbol{\beta}}) = \sigma^2\|\mathbf{a}\|^2 = \sigma^2\left[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}\right]^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c} = \sigma^2\left[\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}\right].$$

   Note that this variance formula depends upon $\mathbf{X}$ through $(\mathbf{X}^T\mathbf{X})^{-1}$. Two implications of this observation are:

   - If the columns of the $\mathbf{X}$ matrix are mutually orthogonal, then $(\mathbf{X}^T\mathbf{X})^{-1}$ will be diagonal, so that the elements of $\hat{\boldsymbol{\beta}}$ are uncorrelated.

   - Even for a given set of explanatory variables, the values at which the explanatory variable are observed will affect the variance (precision) of the resulting parameter estimators.

4. What is remarkable about the Gauss-Markov Theorem is its distributional generality. It does not require normality! It says that $\hat{\boldsymbol{\beta}}$ is BLUE regardless of the distribution of $\mathbf{e}$ (or $\mathbf{y}$) as long as we have mean zero, spherical errors.

An additional property of least-squares estimation is that the estimated mean $\hat{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is invariant to (doesn't change as a result of) linear changes of scale in the explanatory variables.

That is, consider the linear models

$$\mathbf{y} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{=\mathbf{X}} \boldsymbol{\beta} + \mathbf{e}$$

and

$$\mathbf{y} = \underbrace{\begin{pmatrix} 1 & c_1x_{11} & c_2x_{12} & \cdots & c_kx_{1k} \\ 1 & c_1x_{21} & c_2x_{22} & \cdots & c_kx_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & c_1x_{n1} & c_2x_{n2} & \cdots & c_kx_{nk} \end{pmatrix}}_{=\mathbf{Z}} \boldsymbol{\beta}^* + \mathbf{e}$$

Then, $\hat{\boldsymbol{\mu}}$, the least squares estimator of $\mathrm{E}(\mathbf{y})$, is the same in both of these two models. This follows from a more general theorem:

Theorem: In the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and $\mathbf{X}$ is of full rank, $\hat{\boldsymbol{\mu}}$, the least-squares estimator of $\mathrm{E}(\mathbf{y})$ is invariant to a full rank linear transformation of $\mathbf{X}$.

Proof: A full rank linear transformation of $\mathbf{X}$ is given by

$$\mathbf{Z} = \mathbf{XH}$$

where $\mathbf{H}$ is square and of full rank. In the original (untransformed) linear model $\hat{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{P}_{C(\mathbf{X})}\mathbf{y}$. In the transformed model $\mathbf{y} = \mathbf{Z}\boldsymbol{\beta}^* + \mathbf{e}$, $\hat{\boldsymbol{\mu}} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} = \mathbf{P}_{C(\mathbf{Z})}\mathbf{y} = \mathbf{P}_{C(\mathbf{XH})}\mathbf{y}$. So, it suffices to show that $\mathbf{P}_{C(\mathbf{X})} = \mathbf{P}_{C(\mathbf{XH})}$. This is true because if $\mathbf{x} \in C(\mathbf{XH})$ then $\mathbf{x} = \mathbf{XHb}$ for some $\mathbf{b}$, $\Rightarrow \mathbf{x} = \mathbf{Xc}$ where $\mathbf{c} = \mathbf{Hb} \Rightarrow \mathbf{x} \in C(\mathbf{X}) \Rightarrow C(\mathbf{XH}) \subset C(\mathbf{X})$. In addition, if $\mathbf{x} \in C(\mathbf{X})$ then $\mathbf{x} = \mathbf{Xd}$ for some $\mathbf{d} \Rightarrow \mathbf{x} = \mathbf{XHH}^{-1}\mathbf{d} = \mathbf{XHa}$ where $\mathbf{a} = \mathbf{H}^{-1}\mathbf{d} \Rightarrow \mathbf{x} \in C(\mathbf{XH}) \Rightarrow C(\mathbf{X}) \subset C(\mathbf{XH})$. Therefore, $C(\mathbf{X}) = C(\mathbf{XH})$. ∎

- The simple case described above where each of the x_j's is rescaled by a constant c_j occurs when $\mathbf{H} = \mathrm{diag}(1, c_1, c_2, \ldots, c_k)$.

Maximum Likelihood Estimation:

Least-squares provides a simple, intuitively reasonable criterion for estimation. If we want to estimate a parameter describing $\boldsymbol{\mu}$, the mean of $\mathbf{y}$, then choose the parameter value that minimizes the squared distance between $\mathbf{y}$ and $\boldsymbol{\mu}$. If $\text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$, then the resulting estimator is BLUE (optimal, in some sense).

- Least-squares is based only on assumptions concerning the mean and variance-covaraince matrix (the first two moments) of $\mathbf{y}$.

- Least-squares tells us how to estimate parameters associated with the mean (e.g., $\boldsymbol{\beta}$) but nothing about how to estimate parameters describing the variance (e.g., σ^2) or other aspects of the distribution of $\mathbf{y}$.

An alternative method of estimation is maximum likelihood estimation.

- Maximum likelihood requires the specification of the entire distribution of $\mathbf{y}$ (up to some unknown parameters), rather than just the mean and variance of that distribution.

- ML estimation provides a criterion of estimation for any parameter describing the distribution of $\mathbf{y}$, including parameters describing the mean (e.g., $\boldsymbol{\beta}$), variance (σ^2), or any other aspect of the distribution.

- Thus, ML estimation is simultaneously more general and less general than least squares in certain senses. It can provide estimators of all sorts of parameters in a broad array of model types, including models much more complex than those for which least-squares is appropriate; but it requires stronger assumptions than least-squares.

ML Estimation:

Suppose we have a discrete random variable Y (possibly a vector) with observed value y. Suppose Y has probability mass function

$$f(y;\boldsymbol{\gamma}) = \Pr(Y = y;\boldsymbol{\gamma})$$

which depends upon an unknown $p \times 1$ parameter vector $\boldsymbol{\gamma}$ taking values in a parameter space Γ.

The **likelihood function**, $L(\boldsymbol{\gamma}; y)$ is defined to equal the probability mass function but viewed as a function of $\boldsymbol{\gamma}$, not y:

$$L(\boldsymbol{\gamma}; y) = f(y;\boldsymbol{\gamma})$$

Therefore, the likelihood at $\boldsymbol{\gamma}_0$, say, has the interpretation

$$\begin{aligned} L(\boldsymbol{\gamma}_0; y) &= \Pr(Y = y \quad \text{when} \quad \boldsymbol{\gamma} = \boldsymbol{\gamma}_0) \\ &= \Pr(\text{observing the obtained data when } \boldsymbol{\gamma} = \boldsymbol{\gamma}_0) \end{aligned}$$

Logic of ML: choose the value of $\boldsymbol{\gamma}$ that makes this probability largest $\Rightarrow$ $\hat{\boldsymbol{\gamma}}$, the Maximum Likelihood Estimator or MLE.

We use the same procedure when Y is continuous, except in this context Y has a probability density function $f(y;\boldsymbol{\gamma})$, rather than a p.m.f.. Nevertheless, the likelihood is defined the same way, as $L(\boldsymbol{\gamma}; y) = f(y;\boldsymbol{\gamma})$, and we choose $\boldsymbol{\gamma}$ to maximize L.

Often, our data come from a random sample so that we observe $\mathbf{y}$ corresponding to $\mathbf{Y}_{n\times 1}$, a random vector. In this case, we either

(i) specify a multivariate distribution for $\mathbf{Y}$ directly and then the likelihood is equal to that probability density function (e.g. we assume $\mathbf{Y}$ is multivariate normal and then the likelihood would be equal to a multivariate normal density), or

(ii) we use an assumption of independence among the components of $\mathbf{Y}$ to obtain the joint density of $\mathbf{Y}$ as the product of the marginal densities of its components (the Y_i's).

Under independence,

$$L(\boldsymbol{\gamma}; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\gamma})$$

Since its easier to work with sums than products its useful to note that in general

$$\arg\max_{\boldsymbol{\gamma}} L(\boldsymbol{\gamma}; y) = \arg\max_{\boldsymbol{\gamma}} \underbrace{\log L(\boldsymbol{\gamma}; y)}_{\equiv \ell(\boldsymbol{\gamma}; y)}$$

Therefore, we define a MLE of $\boldsymbol{\gamma}$ as a $\hat{\boldsymbol{\gamma}}$ so that

$$\ell(\hat{\boldsymbol{\gamma}}, y) \geq \ell(\boldsymbol{\gamma}; y) \quad \text{for all } \boldsymbol{\gamma} \in \Gamma$$

If Γ is an open set, then $\hat{\boldsymbol{\gamma}}$ must satisfy (if it exists)

$$\frac{\partial \ell(\hat{\boldsymbol{\gamma}})}{\partial \gamma_j} = 0, \quad j = 1, \ldots, p$$

or in vector form

$$\frac{\partial \ell(\hat{\boldsymbol{\gamma}}; y)}{\partial \boldsymbol{\gamma}} = \begin{pmatrix} \frac{\partial \ell(\hat{\boldsymbol{\gamma}})}{\partial \gamma_1} \\ \vdots \\ \frac{\partial \ell(\hat{\boldsymbol{\gamma}})}{\partial \gamma_p} \end{pmatrix} = \mathbf{0}, \quad \text{(the } \textbf{likelihood equation}\text{, a.k.a. } \textbf{score equation}\text{)}$$

In the classical linear model, the unknown parameters of the model are $\boldsymbol{\beta}$ and σ^2, so the pair $(\boldsymbol{\beta}, \sigma^2)$ plays the role of $\boldsymbol{\gamma}$.

Under the assumption A5 that $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, it follows that $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, so the likelihood function is given by

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right\}$$

for $\boldsymbol{\beta} \in \mathcal{R}^{k+1}$ and $\sigma^2 > 0$.

The log-likelihood is a bit easier to work with, and has the same maximizers. It is given by

$$\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

We can maximize this function with respect to $\boldsymbol{\beta}$ and σ^2 in two steps: First maximize with respect to $\boldsymbol{\beta}$ treating σ^2 as fixed, then second plug that estimator back into the loglikelihood function and maximize with respect to σ^2.

For fixed σ^2, maximizing $\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ is equivalent to maximizing the third term $-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ or, equivalently, minimizing $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$. This is just what we do in least-squares, and leads to the estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Next we plug this estimator back into the loglikelihood (this gives what's known as the *profile loglikelihood* for σ^2):

$$\ell(\hat{\boldsymbol{\beta}}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$$

and maximize with respect to σ^2.

Since the 2 exponent in σ^2 can be a little confusing when taking derivatives, let's change symbols from σ^2 to ϕ. Then taking derivatives and setting equal to zero we get the (profile) likelihood equation

$$\frac{\partial \ell}{\partial \phi} = \frac{-n/2}{\phi} + \frac{(1/2)\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{\phi^2} = 0,$$

which has solution

$$\hat{\phi} = \hat{\sigma}^2 = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2,$$

where $\mathbf{x}_i^T$ is the i^{th} row of $\mathbf{X}$.

- Note that to be sure that the solution to this equation is a maximum (rather than a minimum or saddle-point) we must check that $\frac{\partial^2 \ell}{\partial \phi^2}$ is negative. I leave it as an exercise for you to check that this is indeed the case.

Therefore, the MLE of $(\boldsymbol{\beta}, \sigma^2)$ in the classical linear model is $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

and

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \\ &= \frac{1}{n}\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2,\end{aligned}$$

where $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

- Note that

$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{y} - p(\mathbf{y}|C(\mathbf{X}))\|^2 = \frac{1}{n}\|p(\mathbf{y}|C(\mathbf{X})^{\perp})\|^2.$$

Estimation of σ^2:

- Maximum likelihood estimation provides a unified approach to estimating *all* parameters in the model, $\boldsymbol{\beta}$ *and* σ^2.
- In contrast, least squares estimation only provides an estimator of $\boldsymbol{\beta}$.

We've seen that the LS and ML estimators of $\boldsymbol{\beta}$ coincide. However, the MLE of σ^2 is not the usually preferred estimator of σ^2 and is not the estimator of σ^2 that is typically combined with LS estimation of $\boldsymbol{\beta}$.

Why not?

Because $\hat{\sigma}^2$ is biased.

That $\mathrm{E}(\hat{\sigma}^2) \neq \sigma^2$ can easily be established using our results for taking expected values of quadratic forms:

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2) = \mathrm{E}\left(\frac{1}{n}\|\mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{y}\|^2\right) &= \frac{1}{n}\mathrm{E}\left\{(\mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{y})^T\mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{y}\right\} = \frac{1}{n}\mathrm{E}(\mathbf{y}^T\mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{y}) \\
&= \frac{1}{n}\left\{\begin{array}{r}\sigma^2\dim(C(\mathbf{X})^\perp) + \|\underbrace{\mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{X}\boldsymbol{\beta}}_{=\mathbf{0},\text{ because }\mathbf{X}\boldsymbol{\beta}\in C(\mathbf{X})}\|^2\end{array}\right\} \\
&= \frac{\sigma^2}{n}\dim(C(\mathbf{X})^\perp) \\
&= \frac{\sigma^2}{n}\{n - \underbrace{\dim(C(\mathbf{X}))}_{=\mathrm{rank}(\mathbf{X})}\} = \frac{\sigma^2}{n}\{n-(k+1)\}
\end{aligned}$$

Therefore, the MLE $\hat{\sigma}^2$ is biased by a multiplicative factor of $\{n-k-1\}/n$ and an alternative unbiased estimator of σ^2 can easily be constructed as

$$s^2 \equiv \frac{n}{n-k-1}\hat{\sigma}^2 = \frac{1}{n-k-1}\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2,$$

or more generally (that is, for $\mathbf{X}$ not necessarily of full rank),

$$s^2 = \frac{1}{n-\mathrm{rank}(\mathbf{X})}\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2.$$

- s^2 rather than $\hat{\sigma}^2$ is generally the preferred estimator of σ^2. In fact, it can be shown that in the spherical errors linear model, s^2 is the best (minimum variance) estimator of σ^2 in the class of quadratic (in $\mathbf{y}$) unbiased estimators.

- *In the special case* that $\mathbf{X}\boldsymbol{\beta} = \mu\mathbf{j}_n$ (i.e., the model contains only an intercept, or constant term), so that $C(\mathbf{X}) = \mathcal{L}(\mathbf{j}_n)$, we get $\hat{\boldsymbol{\beta}} = \hat{\mu} = \bar{y}$, and $\text{rank}(\mathbf{X}) = 1$. Therefore, s^2 becomes the usual sample variance from the one-sample problem:

$$s^2 = \frac{1}{n-1}\|\mathbf{y} - \bar{y}\mathbf{j}_n\|^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

If $\mathbf{e}$ has a normal distribution, then by part 3 of the theorem on p. 85,

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/\sigma^2 \sim \chi^2(n - \text{rank}(\mathbf{X}))$$

and, since the central $\chi^2(m)$ has mean m and variance $2m$,

$$s^2 = \frac{\sigma^2}{n - \text{rank}(\mathbf{X})}\underbrace{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/\sigma^2}_{\sim\chi^2(n-\text{rank}(\mathbf{X}))}$$

implies

$$\text{E}(s^2) = \frac{\sigma^2}{n - \text{rank}(\mathbf{X})}\{n - \text{rank}(\mathbf{X})\} = \sigma^2,$$

and

$$\text{var}(s^2) = \frac{\sigma^4}{\{n - \text{rank}(\mathbf{X})\}^2}2\{n - \text{rank}(\mathbf{X})\} = \frac{2\sigma^4}{n - \text{rank}(\mathbf{X})}.$$

Properties of $\hat{\boldsymbol{\beta}}$ and s^2 — Summary:

Theorem: Under assumptions A1–A5 of the classical linear model,

i. $\hat{\boldsymbol{\beta}} \sim N_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$,
ii. $(n-k-1)s^2/\sigma^2 \sim \chi^2(n-k-1)$, and
iii. $\hat{\boldsymbol{\beta}}$ and s^2 are independent.

Proof: We've already shown (i.) and (ii.). Result (iii.) follows from the fact that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\mu}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y}$ and $s^2 = (n-k-1)^{-1}||\mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{y}||^2$ are functions of projections onto mutually orthogonal subspaces $C(\mathbf{X})$ and $C(\mathbf{X})^\perp$. ∎

Minimum Variance Unbiased Estimation:

- The Gauss-Markov Theorem establishes that the least-squares estimator $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ for $\mathbf{c}^T\boldsymbol{\beta}$ in the linear model with spherical, but not-necessarily-normal, errors is the minimum variance *linear* unbiased estimator.

- If, in addition, we add the assumption of normal errors, then the least-squares estimator has minimum variance among *all* unbiased estimators.

- The general theory of minimum variance unbiased estimation is beyond the scope of this course, but we will present the background material we need without proof or detailed discussion. Our main goal is just to establish that $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ and s^2 are minimum variance unbiased.

Our model is the classical linear model with normal errors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

We first need the concept of a *complete sufficient statistic:*

Sufficiency: Let $\mathbf{y}$ be random vector with p.d.f. $f(\mathbf{y};\boldsymbol{\theta})$ depending on an unknown $k \times 1$ parameter $\boldsymbol{\theta}$. Let $\mathbf{T}(\mathbf{y})$ be an $r \times 1$ vector-valued statistic that is a function of $\mathbf{y}$. Then $\mathbf{T}(\mathbf{y})$ is said to be a **sufficient statistic for** $\boldsymbol{\theta}$ if and only if the conditional distribution of $\mathbf{y}$ given the value of $\mathbf{T}(\mathbf{y})$ does not depend upon $\boldsymbol{\theta}$.

- If $\mathbf{T}$ is sufficient for $\boldsymbol{\theta}$ then, loosely, $\mathbf{T}$ summarizes all of the information in the data $\mathbf{y}$ relevant to $\boldsymbol{\theta}$. Once we know $\mathbf{T}$, there's no more information in $\mathbf{y}$ about $\boldsymbol{\theta}$.

The property of completeness is needed as well, but it is somewhat technical. Briefly, it ensures that if a function of the sufficient statistic exists that is unbiased for the quantity being estimated, then it is unique.

Completeness: A vector-valued sufficient statistic $\mathbf{T}(\mathbf{y})$ is said to be complete if and only if $\mathrm{E}\{h(\mathbf{T}(\mathbf{y}))\} = 0$ for all $\boldsymbol{\theta}$ implies $\Pr\{h(\mathbf{T}(\mathbf{y})) = 0\} = 1$ for all $\boldsymbol{\theta}$.

Theorem: If $\mathbf{T}(\mathbf{y})$ is a complete sufficient statistic, then $f(\mathbf{T}(\mathbf{y}))$ is a minimum variance unbiased estimator of $\mathrm{E}\{f(\mathbf{T}(\mathbf{y}))\}$.

Proof: This theorem is known as the Lehmann-Scheffé Theorem and its proof follows easily from the Rao-Blackwell Theorem. See, e.g., Bickel and Doksum, p. 122, or Casella and Berger, p. 320.

In the linear model, the p.d.f. of $\mathbf{y}$ depends upon $\boldsymbol{\beta}$ and σ^2, so the pair $(\boldsymbol{\beta}, \sigma^2)$ plays the role of $\boldsymbol{\theta}$.

Is there a complete sufficient statistic for $(\boldsymbol{\beta}, \sigma^2)$ in the classical linear model?

Yes, by the following result:

Theorem: Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^T$ and let $\mathbf{y}$ be a random vector with probability density function

$$f(\mathbf{y}) = c(\boldsymbol{\theta}) \exp\left\{ \sum_{i=1}^{r} \theta_i T_i(\mathbf{y}) \right\} h(\mathbf{y}).$$

Then $\mathbf{T}(\mathbf{y}) = (T_1(\mathbf{y}), \ldots, T_r(\mathbf{y}))^T$ is a complete sufficient statistic provided that neither $\boldsymbol{\theta}$ nor $\mathbf{T}(\mathbf{y})$ satisfy any linear constraints.

- The density function in the above theorem describes the *exponential family of distributions*. For this family, which includes the normal distribution, then it is easy to find a complete sufficient statistic.

Consider the classical linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

The density of $\mathbf{y}$ can be written as

$$\begin{aligned} f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\{-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/(2\sigma^2)\} \\ &= c_1(\sigma^2) \exp\{-(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta})/(2\sigma^2) \\ &= c_2(\boldsymbol{\beta}, \sigma^2) \exp[\{(-1/(2\sigma^2))\mathbf{y}^T\mathbf{y} + (\sigma^{-2}\boldsymbol{\beta}^T)(\mathbf{X}^T\mathbf{y})\} \end{aligned}$$

If we reparameterize in terms of $\boldsymbol{\theta}$ where

$$\theta_1 = -\frac{1}{2\sigma^2}, \quad \begin{pmatrix} \theta_2 \\ \vdots \\ \theta_{k+2} \end{pmatrix} = \frac{1}{\sigma^2}\boldsymbol{\beta},$$

then this density can be seen to be of the exponential form, with vector-valued complete sufficient statistic $\begin{pmatrix} \mathbf{y}^T\mathbf{y} \\ \mathbf{X}^T\mathbf{y} \end{pmatrix}$.

So, since $\mathbf{c}^T\hat{\boldsymbol{\beta}} = \mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, is a function of $\mathbf{X}^T\mathbf{y}$ and is an unbiased estimator of $\mathbf{c}^T\boldsymbol{\beta}$, it must be minimum variance among all unbiased estimators.

In addition, $s^2 = \frac{1}{n-k-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ is an unbiased estimator of σ^2 and can be written as a function of the complete sufficient statistic as well:

$$\begin{aligned} s^2 &= \frac{1}{n-k-1}[(\mathbf{I}_n - \mathbf{P}_{C(\mathbf{X})})\mathbf{y}]^T[(\mathbf{I}_n - \mathbf{P}_{C(\mathbf{X})})\mathbf{y}] \\ &= \frac{1}{n-k-1}\mathbf{y}^T(\mathbf{I}_n - \mathbf{P}_{C(\mathbf{X})})\mathbf{y} = \frac{1}{n-k-1}\{\mathbf{y}^T\mathbf{y} - (\mathbf{y}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})\}. \end{aligned}$$

Therefore, s^2 is a minimum variance unbiased estimator as well.

Taken together, these results prove the following theorem:

Theorem: For the full rank, classical linear model with $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, $\mathbf{e} \sim N_n(0, \sigma^2\mathbf{I}_n)$, s^2 is a minimum variance unbiased estimator of σ^2, and $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ is a minimum variance unbiased estimator of $\mathbf{c}^T\boldsymbol{\beta}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the least squares estimator (MLE) of $\boldsymbol{\beta}$.

Generalized Least Squares

Up to now, we have assumed $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$ in our linear model. There are two aspects to this assumption: (i) uncorrelatedness ($\text{var}(\mathbf{e})$ is diagonal), and (ii) homoscedasticity (the diagonal elements of $\text{var}(\mathbf{e})$ are all the same.

Now we relax these assumptions simultaneously by considering a more general variance-covaraince structure. We now consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \text{where} \quad \text{E}(\mathbf{e}) = \mathbf{0}, \text{var}(\mathbf{e}) = \sigma^2\mathbf{V},$$

where $\mathbf{X}$ is full rank as before, and where $\mathbf{V}$ is a **known** positive definite matrix.

- Note that we assume $\mathbf{V}$ is known, so there still is only one variance-covariance parameter to be estimated, σ^2.

- In the context of least-squares, allowing $\mathbf{V}$ to be unknown complicates things substantially, so we postpone discussion of this case. $\mathbf{V}$ unknown can be handled via ML estimation and we'll talk about that later. Of course, $\mathbf{V}$ unknown is the typical scenario in practice, but there are cases when $\mathbf{V}$ would be known.

- A good example of such a situation is the simple linear regression model with uncorrelated, but heteroscedastic errors:

  $$y_i = \beta_0 + \beta_1 x_i + e_i,$$

  where the e_i's are independent, each with mean 0, and $\text{var}(e_i) = \sigma^2 x_i$. In this case, $\text{var}(\mathbf{e}) = \sigma^2\mathbf{V}$ where $\mathbf{V} = \text{diag}(x_1, \ldots, x_n)$, a known matrix of constants.

Estimation of $\boldsymbol{\beta}$ and σ^2 when $\text{var}(\mathbf{e}) = \sigma^2\mathbf{V}$:

A nice feature of the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad \text{where} \quad \text{var}(\mathbf{e}) = \sigma^2\mathbf{V} \tag{1}$$

is that, although it is not a Gauss-Markov (spherical errors) model, it is simple to transform this model into a Gauss-Markov model. This allows us to apply what we've learned about the spherical errors case to obtain methods and results for the non-spherical case.

Since $\mathbf{V}$ is known and positive definite, it is possible to find a matrix $\mathbf{Q}$ such that $\mathbf{V} = \mathbf{Q}\mathbf{Q}^T$ (e.g., $\mathbf{Q}^T$ could be the Cholesky factor of $\mathbf{V}$).

Multiplying on both sides of the model equation in (1) by the known matrix $\mathbf{Q}^{-1}$, it follows that the following transformed model holds as well:

$$\begin{aligned} &\mathbf{Q}^{-1}\mathbf{y} = \mathbf{Q}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{Q}^{-1}\mathbf{e} \\ &\text{or} \quad \tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{e}} \quad \text{where} \quad \text{var}(\tilde{\mathbf{e}}) = \sigma^2\mathbf{I} \end{aligned} \tag{2}$$

where $\tilde{\mathbf{y}} = \mathbf{Q}^{-1}\mathbf{y}$, $\tilde{\mathbf{X}} = \mathbf{Q}^{-1}\mathbf{X}$ and $\tilde{\mathbf{e}} = \mathbf{Q}^{-1}\mathbf{e}$.

- Notice that model (2) is a Gauss-Markov model because

$$\text{E}(\tilde{\mathbf{e}}) = \mathbf{Q}^{-1}\text{E}(\mathbf{e}) = \mathbf{Q}^{-1}\mathbf{0} = \mathbf{0}$$

and

$$\begin{aligned} \text{var}(\tilde{\mathbf{e}}) &= \mathbf{Q}^{-1}\text{var}(\mathbf{e})(\mathbf{Q}^{-1})^T = \sigma^2\mathbf{Q}^{-1}\mathbf{V}(\mathbf{Q}^{-1})^T \\ &= \sigma^2\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^T(\mathbf{Q}^{-1})^T = \sigma^2\mathbf{I} \end{aligned}$$

The least-squares estimator based on the transformed model minimizes

$$\begin{aligned}\tilde{\mathbf{e}}^T\tilde{\mathbf{e}} = \mathbf{e}^T(\mathbf{Q}^{-1})^T\mathbf{Q}^{-1}\mathbf{e} &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Q}\mathbf{Q}^T)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \quad \text{(The GLS Criterion)}\end{aligned}$$

- So the generalized least squares estimates of $\boldsymbol{\beta}$ from model (1) minimize a squared statistical (rather than Euclidean) distance between $\mathbf{y}$ and $\mathbf{X}\boldsymbol{\beta}$ that takes into account the differing variances among the y_i's and the covariances (correlations) among the y_i's.

- There is some variability in terminology here. Most authors refer to this approach as *generalized least-squares* when $\mathbf{V}$ is an arbitrary, known, positive definite matrix and use the term *weighted least-squares* for the case in which $\mathbf{V}$ is diagonal. Others use the terms interchangeably.

Since GLS estimators for model (1) are just ordinary least squares estimators from model (2), many properties of GLS estimators follow easily from the properties of ordinary least squares.

Properties of GLS Estimators:

1. The best linear unbiased estimator of $\boldsymbol{\beta}$ in model (1) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

Proof: Since model (2) is a Gauss-Markov model, we know that $(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{y}}$ is the BLUE of $\boldsymbol{\beta}$. But this estimator simplifies to

$$\begin{aligned}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{y}} &= [(\mathbf{Q}^{-1}\mathbf{X})^T(\mathbf{Q}^{-1}\mathbf{X})]^{-1}(\mathbf{Q}^{-1}\mathbf{X})^T(\mathbf{Q}^{-1}\mathbf{y}) \\ &= [\mathbf{X}^T(\mathbf{Q}^{-1})^T\mathbf{Q}^{-1}\mathbf{X}]^{-1}\mathbf{X}^T(\mathbf{Q}^{-1})^T\mathbf{Q}^{-1}\mathbf{y} \\ &= [\mathbf{X}^T(\mathbf{Q}\mathbf{Q}^T)^{-1}\mathbf{X}]^{-1}\mathbf{X}^T(\mathbf{Q}\mathbf{Q}^T)^{-1}\mathbf{y} \\ &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.\end{aligned}$$

■

2. Since $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ in model (1), the estimated mean of $\mathbf{y}$ is

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

In going from the $\mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{I}$ case to the $\mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{V}$ case, we've changed our estimate of the mean from

$$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{P}_{C(\mathbf{X})}\mathbf{y}$$

to

$$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

Geometrically, we've changed from using the Euclidean (or orthogonal) projection matrix $\mathbf{P}_{C(\mathbf{X})}$ to using a non-Euclidean (or oblique) projection matrix $\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$. The latter accounts for correlation and heteroscedasticity among the elements of $\mathbf{y}$ when projecting onto $C(\mathbf{X})$

3. The var-cov matrix of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}.$$

Proof:

$$\begin{aligned}
\mathrm{var}(\hat{\boldsymbol{\beta}}) &= \mathrm{var}\{(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}\} \\
&= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\underbrace{\mathrm{var}(\mathbf{y})}_{=\sigma^2\mathbf{V}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}.
\end{aligned}$$

■

4. An unbiased estimator of σ^2 is

$$s^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-k-1} = \frac{\mathbf{y}^T [\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}]\mathbf{y}}{n-k-1},$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$.

Proof: Homework.

5. If $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$, then the MLEs of $\boldsymbol{\beta}$ and σ^2 are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$
$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Proof: We already know that $\hat{\boldsymbol{\beta}}$ is the OLS estimator in model (2) and that the OLS estimator and MLE in such a Gauss-Markov model coincide, so $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$. In addition, the MLE of σ^2 is the MLE of this quantity in model (2), which is

$$\hat{\sigma}^2 = \frac{1}{n}||(\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T)\tilde{\mathbf{y}}||^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

after plugging in $\tilde{\mathbf{X}} = \mathbf{Q}^{-1}\mathbf{X}$, $\tilde{\mathbf{y}} = \mathbf{Q}^{-1}\mathbf{y}$ and some algebra. ∎

Misspecification of the Error Structure:

Q: What happens if we use OLS when GLS is appropriate?

A: The OLS estimator is still linear and unbiased, but no longer best. In addition, we need to be careful to compute the var-cov matrix of our estimator correctly.

Suppose the true model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathrm{E}(\mathbf{e}) = \mathbf{0}, \mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{V}.$$

The BLUE of $\boldsymbol{\beta}$ here is the GLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$, with var-cov matrix $\sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$.

However, suppose we use OLS here instead of GLS. That is, suppose we use the estimator

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Obviously, this estimator is still linear, and it is unbiased because

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}^*) &= \mathrm{E}\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{E}(\mathbf{y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.\end{aligned}$$

However, the variance formula $\mathrm{var}(\hat{\boldsymbol{\beta}}^*) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ is no longer correct, because this was derived under the assumption that $\mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{I}$ (see p. 103). Instead, the correct var-cov of the OLS estimator here is

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}^*) &= \mathrm{var}\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underbrace{\mathrm{var}(\mathbf{y})}_{=\sigma^2\mathbf{V}}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}. \qquad (*)\end{aligned}$$

In contrast, if we had used the GLS estimator (the BLUE), the var-cov matrix of our estimator would have been

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}. \qquad (**)$$

Since $\hat{\boldsymbol{\beta}}$ is the BLUE, we know that the variances from (*) will be $\geq$ the variances from (**), which means that the OLS estimator here is a less efficient (precise), but not necessarily *much* less efficient, estimator under the GLS model.

Misspecification of E($\mathbf{y}$):

Suppose that the true model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{e}$ where we return to the spherical errors case: var($\mathbf{e}$) = $\sigma^2\mathbf{I}$. We want to consider what happens when we omit some explanatory variable is $\mathbf{X}$ and when we include too many x's. So, let's partition our model as

$$\begin{aligned}\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} &= (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \mathbf{e} \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}. \qquad (\dagger)\end{aligned}$$

- If we leave out $\mathbf{X}_2\boldsymbol{\beta}_2$ when it should be included (when $\boldsymbol{\beta}_2 \neq \mathbf{0}$) then we are **underfitting**.
- If we include $\mathbf{X}_2\boldsymbol{\beta}_2$ when it doesn't belong in the true model (when $\boldsymbol{\beta}_2 = \mathbf{0}$) then we are **overfitting**.
- We will consider the effects of both overfitting and underfitting on the bias and variance of $\hat{\boldsymbol{\beta}}$. The book also consider effects on predicted values and on the MSE s^2.

Underfitting:

Suppose model (†) holds, but we fit the model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \mathbf{e}^*, \quad \text{var}(\mathbf{e}^*) = \sigma^2\mathbf{I}. \qquad (\clubsuit)$$

The following theorem gives the bias and var-cov matrix of $\hat{\boldsymbol{\beta}}_1^*$ the OLS estimator from ♣.

Theorem: If we fit model ♣ when model (†) is the true model, then the mean and var-cov matrix of the OLS estimator $\hat{\boldsymbol{\beta}}_1^* = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}$ are as follows:

(i) $\text{E}(\hat{\boldsymbol{\beta}}_1^*) = \boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2$, where $\mathbf{A} = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2$.

(ii) $\text{var}(\hat{\boldsymbol{\beta}}_1^*) = \sigma^2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}$.

Proof:

(i)

$$\begin{aligned}\text{E}(\hat{\boldsymbol{\beta}}_1^*) &= \text{E}[(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}] = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\text{E}(\mathbf{y}) \\ &= (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) \\ &= \boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2.\end{aligned}$$

(ii)

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}_1^*) &= \text{var}[(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}] \\ &= (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T(\sigma^2\mathbf{I})\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1} \\ &= \sigma^2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}.\end{aligned}$$

■

- This result says that when underfitting, $\hat{\boldsymbol{\beta}}_1^*$ is biased by an amount that depends upon both the omitted and included explanatory variables.

Corollary If $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$, i.e.. if the columns of $\mathbf{X}_1$ are orthogonal to the columns of $\mathbf{X}_2$, then $\hat{\boldsymbol{\beta}}_1^*$ is unbiased.

Note that in the above theorem the var-cov matrix of $\hat{\boldsymbol{\beta}}_1^*$, $\sigma^2(\mathbf{X}_1^T\mathbf{X}_1)^{-1}$ is not the same as the var-cov matrix of $\hat{\boldsymbol{\beta}}_1$, the corresponding portion of the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ from the full model. How these var-cov matrices differ is established in the following theorem:

Theorem: Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ from the full model (†) be partitioned as

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$$

and let $\hat{\boldsymbol{\beta}}_1^* = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}$ be the estimator from the reduced model ♣. Then

$$\text{var}(\hat{\boldsymbol{\beta}}_1) - \text{var}(\hat{\boldsymbol{\beta}}_1^*) = \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^T$$

a n.n.d. matrix. Here, $\mathbf{A} = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2$ and $\mathbf{B} = \mathbf{X}_2^T\mathbf{X}_2 - \mathbf{X}_2^T\mathbf{X}_1\mathbf{A}$.

- Thus $\text{var}(\hat{\beta}_j) \geq \text{var}(\hat{\beta}_j^*)$, meaning that underfitting results in smaller variances of the $\hat{\beta}_j$'s and overfitting results in larger variances of the $\hat{\beta}_j$'s.

Proof: Partitioning $\mathbf{X}^T\mathbf{X}$ to conform to the partitioning of $\mathbf{X}$ and $\boldsymbol{\beta}$, we have

$$\begin{aligned} \text{var}(\hat{\boldsymbol{\beta}}) = \text{var}\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2 \\ \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_2 \end{pmatrix}^{-1} \\ = \sigma^2\begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{pmatrix}^{-1} = \sigma^2\begin{pmatrix} \mathbf{H}^{11} & \mathbf{H}^{12} \\ \mathbf{H}^{21} & \mathbf{H}^{22} \end{pmatrix}, \end{aligned}$$

where $\mathbf{H}_{ij} = \mathbf{X}_i^T\mathbf{X}_j$ and $\mathbf{H}^{ij}$ is the corresponding block of the inverse matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ (see p. 54).

So, $\text{var}(\hat{\boldsymbol{\beta}}_1) = \sigma^2\mathbf{H}^{11}$. Using the formulas for inverses of partitioned matrices,

$$\mathbf{H}^{11} = \mathbf{H}_{11}^{-1} + \mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{B}^{-1}\mathbf{H}_{21}\mathbf{H}_{11}^{-1},$$

where

$$\mathbf{B} = \mathbf{H}_{22} - \mathbf{H}_{21}\mathbf{H}_{11}^{-1}\mathbf{H}_{12}.$$

In the previous theorem, we showed that $\text{var}(\hat{\boldsymbol{\beta}}_1^*) = \sigma^2(\mathbf{X}_1^T\mathbf{X}_1)^{-1} = \sigma^2\mathbf{H}_{11}^{-1}$. Hence,

$$\begin{aligned}
\text{var}(\hat{\boldsymbol{\beta}}_1) - \text{var}(\hat{\boldsymbol{\beta}}_1^*) &= \sigma^2(\mathbf{H}^{11} - \mathbf{H}_{11}^{-1}) \\
&= \sigma^2(\mathbf{H}_{11}^{-1} + \mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{B}^{-1}\mathbf{H}_{21}\mathbf{H}_{11}^{-1} - \mathbf{H}_{11}^{-1}) \\
&= \sigma^2(\mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{B}^{-1}\mathbf{H}_{21}\mathbf{H}_{11}^{-1}) \\
&= \sigma^2[(\mathbf{X}_1^T\mathbf{X}_1)^{-1}(\mathbf{X}_1^T\mathbf{X}_2)\mathbf{B}^{-1}(\mathbf{X}_2^T\mathbf{X}_1)(\mathbf{X}_1^T\mathbf{X}_1)^{-1}] \\
&= \sigma^2\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^T.
\end{aligned}$$

We leave it as homework for you to show that $\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^T$ is n.n.d. ∎

- To summarize, we've seen that underfitting reduces the variances of regression parameter estimators, but introduces bias. On the other hand, overfitting produces unbiased estimators with increased variances. Thus it is the task of a regression model builder to find an optimum set of explanatory variables to balance between a biased model and one with large variances.

The Model in Centered Form

For some purposes it is useful to write the regression model in centered form; that is, in terms of the centered explanatory variables (the explanatory variables minus their means).

The regression model can be written

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \\ &= \alpha + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \cdots + \beta_k(x_{ik} - \bar{x}_k) + e_i, \end{aligned}$$

for $i = 1, \ldots, n$, where

$$\alpha = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \cdots + \beta_k \bar{x}_k, \qquad (\heartsuit)$$

and where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$.

In matrix form, the equivalence between the original model and centered model that we've written above becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = (\mathbf{j}_n, \mathbf{X}_c)\begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \mathbf{e},$$

where $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_k)^T$, and

$$\mathbf{X}_c = \underbrace{(\mathbf{I} - \frac{1}{n}\mathbf{J}_{n,n})}_{=\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)^\perp}} \mathbf{X}_1 = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{pmatrix},$$

and $\mathbf{X}_1$ is the matrix consisting of all but the first columns of $\mathbf{X}$, the original model matrix.

- $\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)^\perp} = (\mathbf{I} - \frac{1}{n}\mathbf{J}_{n,n})$ is sometimes called the *centering matrix*.

Based on the centered model, the least squares estimators become:

$$\begin{aligned} \begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} &= [(\mathbf{j}_n, \mathbf{X}_c)^T(\mathbf{j}_n, \mathbf{X}_c)]^{-1}(\mathbf{j}_n, \mathbf{X}_c)^T\mathbf{y} = \begin{pmatrix} n & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_c^T\mathbf{X}_c \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{j}_n^T \\ \mathbf{X}_c^T \end{pmatrix} \mathbf{y} \\ &= \begin{pmatrix} n^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}_c^T\mathbf{X}_c)^{-1} \end{pmatrix} \begin{pmatrix} n\bar{y} \\ \mathbf{X}_c^T\mathbf{y} \end{pmatrix} = \begin{pmatrix} \bar{y} \\ (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y} \end{pmatrix}, \end{aligned}$$

or

$$\hat{\alpha} = \bar{y}, \quad \text{and}$$
$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}.$$

$\hat{\boldsymbol{\beta}}_1$ here is the same as the usual least-squares estimator. That is, it is the same as $\hat{\beta}_1, \ldots, \hat{\beta}_k$ from $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. However, the intercept $\hat{\alpha}$ differs from $\hat{\beta}_0$. The relationship between $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ is just what you'd expect from the reparameterization (see (♡)):

$$\hat{\alpha} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + \cdots + \hat{\beta}_k\bar{x}_k.$$

From the expression for the estimated mean based on the centered model:

$$\widehat{\mathrm{E}(y_i)} = \hat{\alpha} + \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \cdots + \hat{\beta}_k(x_{ik} - \bar{x}_k)$$

it is clear that the fitted regression plane passes through the point of averages: $(\bar{y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$.

In general, we can write SSE, the error sum of squares, as

$$\begin{aligned}
\mathrm{SSE} &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{y} - \mathbf{P}_{C(\mathbf{X})}\mathbf{y})^T(\mathbf{y} - \mathbf{P}_{C(\mathbf{X})}\mathbf{y}) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} - \mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} + \mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}.
\end{aligned}$$

From the centered model we see that $\widehat{\mathrm{E}(\mathbf{y})} = \mathbf{X}\hat{\boldsymbol{\beta}} = [\mathbf{j}_n, \mathbf{X}_c]\begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix}$, so SSE can also be written as

$$\begin{aligned}
\mathrm{SSE} &= \mathbf{y}^T\mathbf{y} - (\hat{\alpha}, \hat{\boldsymbol{\beta}}_1^T)\begin{pmatrix} \mathbf{j}_n^T \\ \mathbf{X}_c^T \end{pmatrix}\mathbf{y} \\
&= \mathbf{y}^T\mathbf{y} - \bar{y}\,\mathbf{j}_n^T\mathbf{y} - \hat{\boldsymbol{\beta}}_1^T\mathbf{X}_c^T\mathbf{y} \\
&= (\mathbf{y} - \bar{y}\,\mathbf{j}_n)^T\mathbf{y} - \hat{\boldsymbol{\beta}}_1^T\mathbf{X}_c^T\mathbf{y} \\
&= (\mathbf{y} - \bar{y}\,\mathbf{j}_n)^T(\mathbf{y} - \bar{y}\,\mathbf{j}_n) - \hat{\boldsymbol{\beta}}_1^T\mathbf{X}_c^T\mathbf{y} \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\boldsymbol{\beta}}_1^T\mathbf{X}_c^T\mathbf{y} \qquad (*)
\end{aligned}$$

R^2, the Estimated Coefficient of Determination

Rearranging (*), we obtain a decomposition of the total variability in the data:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}_1^T \mathbf{X}_c^T \mathbf{y} + \text{SSE}$$

$$\text{or} \qquad \text{SST} = \text{SSR} + \text{SSE}$$

- Here SST is the (corrected) total sum of squares. The term "corrected" here indicates that we've taken the sum of the squared y's after correcting, or adjusting, them for the mean. The uncorrected sum of squares would be $\sum_{i=1}^{n} y_i^2$, but this quantity arises less frequently, and by "SST" or "total sum of squares" we will generally mean the corrected quantity unless stated otherwise.

- Note that SST quantifies the total variability in the data (if we added a $\frac{1}{n-1}$ multiplier in front, SST would become the sample variance).

- The first term on the right-hand side is called the regression sum of squares. It represents the variability in the data (the portion of SST) that can be explained by the regression terms $\beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k$.

- This interpretation can be seen by writing SSR as

$$\text{SSR} = \hat{\boldsymbol{\beta}}_1^T \mathbf{X}_c^T \mathbf{y} = \hat{\boldsymbol{\beta}}_1^T \mathbf{X}_c^T \mathbf{X}_c (\mathbf{X}_c^T \mathbf{X}_c)^{-1} \mathbf{X}_c^T \mathbf{y} = (\mathbf{X}_c \hat{\boldsymbol{\beta}}_1)^T (\mathbf{X}_c \hat{\boldsymbol{\beta}}_1).$$

The proportion of the total sum of squares that is due to regression is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\hat{\boldsymbol{\beta}}_1^T \mathbf{X}_c^T \mathbf{X}_c \hat{\boldsymbol{\beta}}_1}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y} - n\bar{y}^2}{\mathbf{y}^T \mathbf{y} - n\bar{y}^2}.$$

- This quantity is called the **coefficient of determination**, and it is usually denoted as R^2. It is the sample estimate of the squared multiple correlation coefficient we discussed earlier (see p. 77).

Facts about R^2:

1. The range of R^2 is $0 \leq R^2 \leq 1$, with 0 corresponding to the explanatory variables $x_1, \ldots, x_k$ explaining none of the variability in $\mathbf{y}$ and 1 corresponding to $x_1, \ldots, x_k$ explaining all of the variability in $\mathbf{y}$.

2. R, the multiple correlation coefficient or positive square root of R^2, is equal to the sample correlation coefficient between the observed y_i's and their fitted values, the $\hat{y}_i$'s. (Here the fitted value is just the estimated mean: $\hat{y}_i = \widehat{\mathrm{E}(y_i)} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$.)

3. R^2 will always stay the same or (typically) increase if an explanatory variable x_{k+1} is added to the model.

4. If $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, then

$$\mathrm{E}(R^2) = \frac{k}{n-1}.$$

- From properties 3 and 4, we see that R^2 tends to be higher for a model with many predictors than for a model with few predictors, even if those models have the same explanatory power. That is, as a measure of goodness of fit, R^2 rewards complexity and penalizes parsimony, which is certainly not what we would like to do.

- Therefore, a version of R^2 that penalizes for model complexity was developed, known as R_a^2 or **adjusted** R^2:

$$R_a^2 = \frac{\left(R^2 - \frac{k}{n-1}\right)(n-1)}{n-k-1} = \frac{(n-1)R^2 - k}{n-k-1}.$$

5. Unless the x_j's $j = 1, \ldots, k$ are mutually orthogonal, R^2 cannot be written as a sum of k components uniquely attributable to $x_1, \ldots, x_k$. (R^2 represents the joint explanatory power of the x_j's not the sum of the explanatory powers of each of the individual x_j's.)

6. R^2 is invariant to a full-rank linear transformation of $\mathbf{X}$ and to a scale change on $\mathbf{y}$ (but not invariant to a joint linear transformation on $[\mathbf{y}, \mathbf{X}]$).

7. Geometrically, R, the multiple correlation coefficient, is equal to $R = \cos(\theta)$ where θ is the angle between $\mathbf{y}$ and $\hat{\mathbf{y}}$ corrected for their means, $\bar{y}\,\mathbf{j}_n$. This is depicted in the picture below.

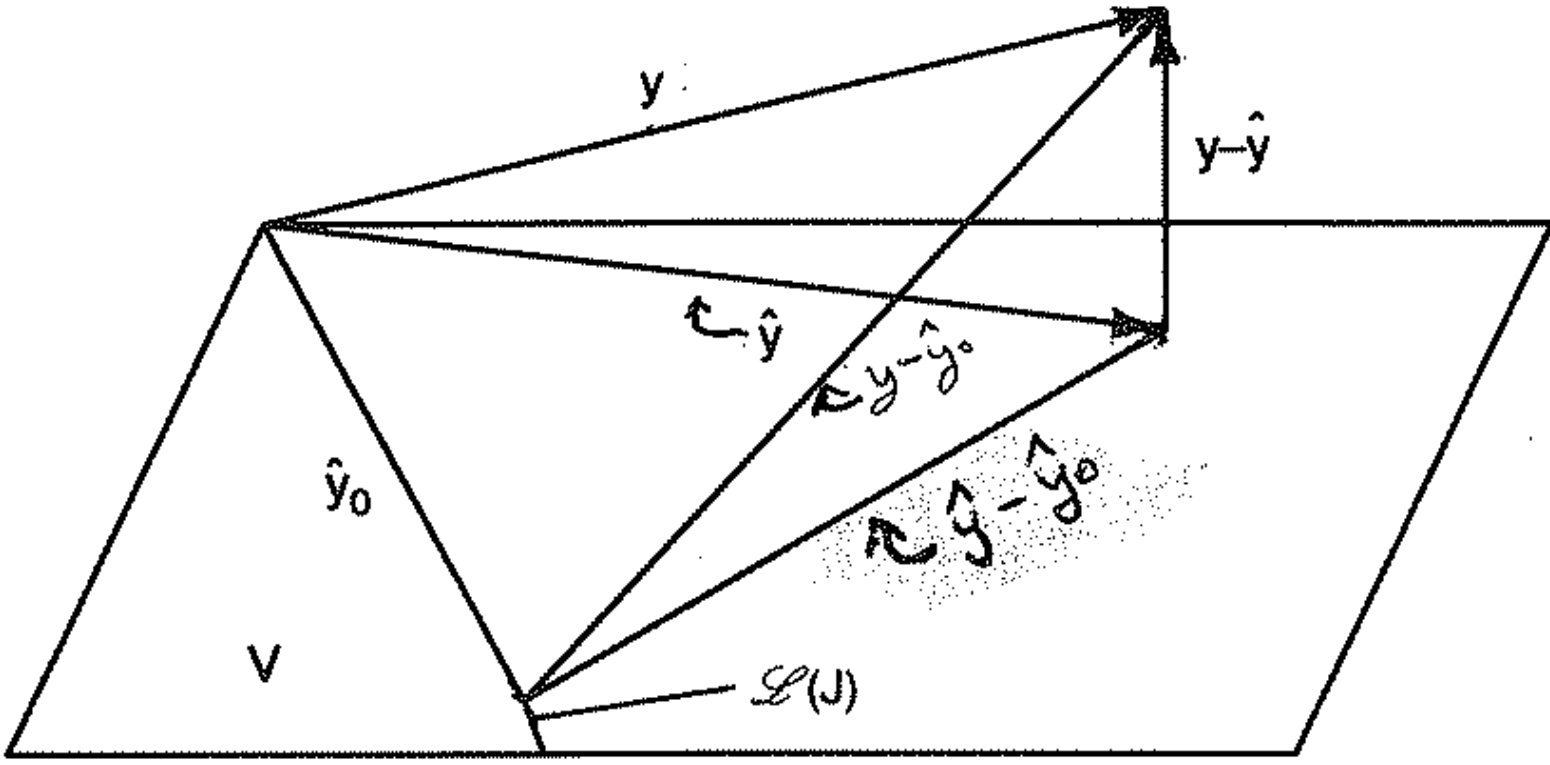


FIGURE 3.5 The multiple regression coefficient R, where $R^2 = \|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2 / \|\mathbf{y} - \hat{\mathbf{y}}_0\|^2$

Inference in the Multiple Regression Model

Testing a Subset of $\boldsymbol{\beta}$: Testing Nested Models

All testing of linear hypotheses (nonlinear hypotheses are rarely encountered in practice) in linear models reduces essentially to putting linear constraints on the model space. The test amounts to comparing the resulting constrained model against the original unconstrained model.

We start with a model we know (assume, really) to be valid:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}, \quad \text{where} \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \in C(\mathbf{X}) \equiv V, \quad \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

and then ask the question of whether or not a simpler model holds corresponding to $\boldsymbol{\mu} \in V_0$ where V_0 is a proper subset of V. (E.g., $V_0 = C(\mathbf{X}_0)$ where $\mathbf{X}_0$ is a matrix consisting of a subset of the columns of $\mathbf{X}$.)

For example, consider the second order response surface model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2} + e_i, \quad i = 1, \ldots, n. \quad (\dagger)$$

This model says that $\mathrm{E}(y)$ is a quadratic function of x_1 and x_2.

A hypothesis we might be interested in here is that the second-order terms are unnecessary; i.e., we might be interested in $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$, under which the model is linear in x_1 and x_2:

$$y_i = \beta_0^* + \beta_1^* x_{i1} + \beta_2^* x_{i2} + e_i^*, \quad i = 1, \ldots, n. \quad (\ddagger)$$

- Testing $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ is equivalent to testing H_0 :model $(\ddagger)$ holds versus H_1 :model $(\dagger)$ holds but $(\ddagger)$ does not.
- I.e., we test $H_0 : \boldsymbol{\mu} \in C([\mathbf{j}_n, \mathbf{x}_1, \mathbf{x}_2])$ versus

  $H_1 : \boldsymbol{\mu} \in C([\mathbf{j}_n, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_1 * \mathbf{x}_1, \mathbf{x}_2 * \mathbf{x}_2, \mathbf{x}_1 * \mathbf{x}_2])$ and $\boldsymbol{\mu} \notin C([\mathbf{j}_n, \mathbf{x}_1, \mathbf{x}_2])$.

  Here $*$ denotes the element-wise product and $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y})$.

Without loss of generality, we can always arrange the linear model so the terms we want to test appear last in the linear predictor. So, we write our model as

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \mathbf{e} \\ &= \underbrace{\mathbf{X}_1}_{n\times(k+1-h)}\boldsymbol{\beta}_1 + \underbrace{\mathbf{X}_2}_{n\times h}\boldsymbol{\beta}_2 + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \qquad \text{(FM)}\end{aligned}$$

where we are interested in the hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$.

Under $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ the model becomes

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \mathbf{e}^*, \quad \mathbf{e}^* \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \qquad \text{(RM)}$$

The problem is to test

$$H_0 : \boldsymbol{\mu} \in C(\mathbf{X}_1) \quad \text{(RM)} \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \notin C(\mathbf{X}_1)$$

under the *maintained hypothesis* that $\boldsymbol{\mu} \in C(\mathbf{X}) = C([\mathbf{X}_1, \mathbf{X}_2])$ (FM).

We'd like to find a test statistic whose size measures the strength of the evidence against H_0. If that evidence is overwhelming (the test statistic is large enough) then we reject H_0.

The test statistic should be large, but large relative to what?

Large relative to its distribution under the null hypothesis.

How large?

That's up to the user, but an $\alpha-$level test rejects H_0 if, assuming H_0 is true, the probability of getting a test statistic at least as far from expected as the one obtained (the $p-$value) is less than α.

- E.g., suppose we compute a test statistic and obtain a $p-$value of $p = 0.02$. This says that assuming H_0 is true, the results that we obtained were very unlikely (results this extreme should happen only 2% of the time). If these results are so unlikely assuming H_0 is true, perhaps H_0 is not true. The cut-off for how unlikely our results must be before we're willing to reject H_0 is the significance level α. (We reject if $p < \alpha$.)

So, we want a test statistic that measures the strength of the evidence against $H_0 : \boldsymbol{\mu} \in C(\mathbf{X}_1)$ (i.e., one that is small for $\boldsymbol{\mu} \in C(\mathbf{X}_1)$ and large for $\boldsymbol{\mu} \notin C(\mathbf{X}_1)$) whose distribution is available.

- This will lead to an F test which is equivalent to the likelihood ratio test, and which has some optimality properties.

Note that under RM, $\boldsymbol{\mu} \in C(\mathbf{X}_1) \subset C(\mathbf{X}) = C([\mathbf{X}_1, \mathbf{X}_2])$. Therefore, if RM is true, then FM must be true as well. So, if RM is true, then the least squares estimates of the mean $\boldsymbol{\mu}$: $\mathbf{P}_{C(\mathbf{X}_1)}\mathbf{y}$ and $\mathbf{P}_{C(\mathbf{X})}\mathbf{y}$ are estimates of the same thing.

This suggests that the difference between the two estimates

$$\mathbf{P}_{C(\mathbf{X})}\mathbf{y} - \mathbf{P}_{C(\mathbf{X}_1)}\mathbf{y} = (\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}$$

should be small under $H_0 : \boldsymbol{\mu} \in C(\mathbf{X}_1)$.

- Note that $\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)}$ is the projection matrix onto $C(\mathbf{X}_1)^\perp \cap C(\mathbf{X})$, the orthogonal complement of $C(\mathbf{X}_1)$ with respect to $C(\mathbf{X})$, and $C(\mathbf{X}_1) \oplus [C(\mathbf{X}_1)^\perp \cap C(\mathbf{X})] = C(\mathbf{X})$. (See bottom of p. 43 of these notes.)

So, under H_0, $(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}$ should be "small". A measure of the "smallness" of this vector is its squared length:

$$\|(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}\|^2 = \mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}.$$

By our result on expected values of quadratic forms,

$$\begin{aligned}\mathrm{E}[\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}] &= \sigma^2 \dim[C(\mathbf{X}_1)^\perp \cap C(\mathbf{X})] + \boldsymbol{\mu}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\boldsymbol{\mu} \\ &= \sigma^2 h + [(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\boldsymbol{\mu}]^T[(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\boldsymbol{\mu}] \\ &= \sigma^2 h + (\mathbf{P}_{C(\mathbf{X})}\boldsymbol{\mu} - \mathbf{P}_{C(\mathbf{X}_1)}\boldsymbol{\mu})^T(\mathbf{P}_{C(\mathbf{X})}\boldsymbol{\mu} - \mathbf{P}_{C(\mathbf{X}_1)}\boldsymbol{\mu})\end{aligned}$$

Under H_0, $\boldsymbol{\mu} \in C(\mathbf{X}_1)$ and $\boldsymbol{\mu} \in C(\mathbf{X})$, so

$$(\mathbf{P}_{C(\mathbf{X})}\boldsymbol{\mu} - \mathbf{P}_{C(\mathbf{X}_1)}\boldsymbol{\mu}) = \boldsymbol{\mu} - \boldsymbol{\mu} = \mathbf{0}.$$

Under H_1,

$$\mathbf{P}_{C(\mathbf{X})}\boldsymbol{\mu} = \boldsymbol{\mu}, \quad \text{but} \quad \mathbf{P}_{C(\mathbf{X}_1)}\boldsymbol{\mu} \neq \boldsymbol{\mu}.$$

I.e., letting $\boldsymbol{\mu}_0$ denote $p(\boldsymbol{\mu}|C(\mathbf{X}_1))$,

$$\mathrm{E}[\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}] = \begin{cases} \sigma^2 h, & \text{under } H_0; \\ \sigma^2 h + \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2, & \text{under } H_1. \end{cases}$$

- That is, under H_0 we expect the squared length of

  $$\mathbf{P}_{C(\mathbf{X})}\mathbf{y} - \mathbf{P}_{C(\mathbf{X}_1)}\mathbf{y} \equiv \hat{\mathbf{y}} - \hat{\mathbf{y}}_0$$

  to be small, on the order of $\sigma^2 h$. If H_0 is not true, then the squared length of $\hat{\mathbf{y}} - \hat{\mathbf{y}}_0$ will be larger, with expected value $\sigma^2 h + \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2$.

Therefore, if σ^2 is known

$$\frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2}{\sigma^2 h} = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/h}{\sigma^2} \begin{cases} \approx 1, & \text{under } H_0 \\ > 1, & \text{under } H_1 \end{cases}$$

is an appropriate test statistic for testing H_0.

Typically, σ^2 will not be known, so it must be estimated. The appropriate estimator is $s^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-k-1)$, the mean squared error from FM, the model which is valid under H_0 and under H_1. Our test statistic then becomes

$$F = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/h}{s^2} = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/h}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-k-1)} \begin{cases} \approx 1, & \text{under } H_0 \\ > 1, & \text{under } H_1. \end{cases}$$

By the theorems on pp. 84–85, the following results on the numerator and denominator of F hold:

Theorem: Suppose $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ where $\mathbf{X}$ is $n \times (k+1)$ of full rank where $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$, and $\mathbf{X}_2$ is $n \times h$. Let $\hat{\mathbf{y}} = p(\mathbf{y}|C(\mathbf{X})) = \mathbf{P}_{C(\mathbf{X})}\mathbf{y}$, $\hat{\mathbf{y}}_0 = p(\mathbf{y}|C(\mathbf{X}_1)) = \mathbf{P}_{C(\mathbf{X}_1)}\mathbf{y}$, and $\boldsymbol{\mu}_0 = p(\boldsymbol{\mu}|C(\mathbf{X}_1)) = \mathbf{P}_{C(\mathbf{X}_1)}\boldsymbol{\mu}$. Then

(i) $\frac{1}{\sigma^2}\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \frac{1}{\sigma^2}\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathbf{y} \sim \chi^2(n-k-1)$;

(ii) $\frac{1}{\sigma^2}\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2 = \frac{1}{\sigma^2}\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y} \sim \chi^2(h, \lambda_1)$, where

$$\lambda_1 = \frac{1}{2\sigma^2}\|(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\boldsymbol{\mu}\|^2 = \frac{1}{2\sigma^2}\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2;$$

and

(iii) $\frac{1}{\sigma^2}\|\mathbf{y} - \hat{\mathbf{y}}\|^2$ and $\frac{1}{\sigma^2}\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2$ are independent.

Proof: Parts (i) and (ii) folllow immediately from part (3) of the theorem on p. 84. Part (iii) follows because

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = ||p(\mathbf{y}|C(\mathbf{X})^\perp)||^2$$

and

$$\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2 = ||p(\mathbf{y}|\underbrace{C(\mathbf{X}_1)^\perp \cap C(\mathbf{X})}_{\subset C(\mathbf{X})})||^2$$

are squared lengths of projections onto orthogonal subspaces, so they are independent according to the theorem on p. 85. ∎

From this result, the distribution of our test statistic F follows easily:

Theorem: Under the conditions of the previous theorem,

$$F = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/h}{s^2} = \frac{\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}/h}{\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathbf{y}/(n-k-1)}$$
$$\sim \begin{cases} F(h, n-k-1), & \text{under } H_0; \text{ and} \\ F(h, n-k-1, \lambda_1), & \text{under } H_1, \end{cases}$$

where λ_1 is as given in the previous theorem.

Proof: Follows the previous theorem and the definition of the F distribution. ∎

Therefore, the $\alpha-$level $F-$test for $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$ (equivalently, of RM vs. FM) is:

$$\text{reject } H_0 \text{ if } \quad F > F_{1-\alpha}(h, n-k-1).$$

- It is worth noting that the numerator of this F test can be obtained as the difference in the SSE's under FM and RM divided by the difference in the dfE (degrees of freedom for error) for the two models. This is so because the Pythagorean Theorem yields

  $$\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 - \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \text{SSE(RM)} - \text{SSE(FM)}.$$

  The difference in the dfE's is $(n-h-k-1) - (n-k-1) = h$. Therefore,

  $$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/[\text{dfE(RM)} - \text{dfE(FM)}]}{\text{SSE(FM)}/\text{dfE(FM)}}.$$

- In addition, because SSE = SST − SSR,

  $$\begin{aligned} \|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2 &= \text{SSE(RM)} - \text{SSE(FM)} \\ &= \text{SST} - \text{SSR(RM)} - [\text{SST} - \text{SSR(FM)}] \\ &= \text{SSR(FM)} - \text{SSR(RM)} \equiv \text{SS}(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) \end{aligned}$$

  which we denote as $\text{SS}(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$, and which is known as the *"extra" regression sum of squares due to $\boldsymbol{\beta}_2$ after accounting for $\boldsymbol{\beta}_1$.*

The results leading to the F-test for $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ that we have just developed can be summarized in an ANOVA table:

Source of Variation	Sum of Squares	df	Mean Squares	F
Due to $\boldsymbol{\beta}_2$ adjusted for $\boldsymbol{\beta}_1$	$SS(\boldsymbol{\beta}_2\|\boldsymbol{\beta}_1) = \mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}$	h	$\frac{SS(\boldsymbol{\beta}_2\|\boldsymbol{\beta}_1)}{h}$	$\frac{MS(\boldsymbol{\beta}_2\|\boldsymbol{\beta}_1)}{MSE}$
Error	$SSE = \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathbf{y}$	$n-k-1$	$\frac{SSE}{n-k-1}$	
Total (Corr.)	$SST = \mathbf{y}^T\mathbf{y} - n\bar{y}^2$			

An additional column is sometimes added to the ANOVA table for $\mathrm{E}(MS)$, or expected mean squares. The expected mean squares here are

$$\begin{aligned}\mathrm{E}\{MS(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)\} &= \frac{1}{h}\mathrm{E}\{SS(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)\} = \frac{\sigma^2}{h}\mathrm{E}\{SS(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/\sigma^2\} \\ &= \frac{\sigma^2}{h}\{h + 2\lambda_1\} = \sigma^2 + \frac{1}{h}||\boldsymbol{\mu} - \boldsymbol{\mu}_0||^2\end{aligned}$$

and

$$\mathrm{E}(MSE) = \frac{1}{n-k-1}\mathrm{E}(SSE) = \frac{1}{n-k-1}(n-k-1)\sigma^2 = \sigma^2.$$

These expected mean squares give additional insight into why F is an appropriate test of $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. Any mean square can be thought of as an estimate of its expectation. Therefore, MSE estimates σ^2 (always), and $MS(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$ estimates σ^2 under H_0, and estimates σ^2 plus a positive quantity under H_1. Therefore, our test statistic F will behave as

$$F \begin{cases} \approx 1, & \text{under } H_0 \\ > 1, & \text{under } H_1 \end{cases}$$

where how much larger F is than 1 depends upon "how false" H_0 is.

Overall Regression Test:

An important special case of the test of $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ that we have just developed is when we partition $\boldsymbol{\beta}$ so that $\boldsymbol{\beta}_1$ contains just the intercept and when $\boldsymbol{\beta}_2$ contains all of the regression coefficients. That is, if we write the model as

$$\begin{aligned}\mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e} \\ &= \beta_0\mathbf{j}_n + \underbrace{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{=\mathbf{x}_2} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}}_{=\boldsymbol{\beta}_2} + \mathbf{e}\end{aligned}$$

then our hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ is equivalent to

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

which says that the collection of explanatory variables $x_1, \ldots, x_k$ have no linear effect on (do not predict) y.

The test of this hypothesis is called the **overall regression test** and occurs as a special case of the test of $\boldsymbol{\beta}_2 = \mathbf{0}$ that we've developed. Under H_0,

$$\hat{\mathbf{y}}_0 = p(\mathbf{y}|C(\mathbf{X}_1)) = p(\mathbf{y}|\mathcal{L}(\mathbf{j}_n)) = \bar{y}\,\mathbf{j}_n$$

and $h = k$, so the numerator of our F-test statistic becomes

$$\begin{aligned}\frac{1}{k}\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{\mathcal{L}(\mathbf{j}_n)})\mathbf{y} &= \frac{1}{k}(\mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} - \mathbf{y}^T\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)}\mathbf{y}) \\ &= \frac{1}{k}\{(\mathbf{P}_{C(\mathbf{X})}\mathbf{y})^T\mathbf{y} - \mathbf{y}^T\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)}^T \underbrace{\mathbf{P}_{\mathcal{L}(\mathbf{j}_n)}\mathbf{y}}_{=\bar{y}\mathbf{j}_n}\} \\ &= \frac{1}{k}(\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} - n\bar{y}^2) = SSR/k \equiv MSR\end{aligned}$$

Thus, the test statistic of overall regression is given by

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$$
$$\sim \begin{cases} F(k, n-k-1), & \text{under } H_0 : \beta_1 = \cdots = \beta_k = 0 \\ F(k, n-k-1, \frac{1}{2\sigma^2}\boldsymbol{\beta}_2^T \mathbf{X}_2^T \mathbf{P}_{\mathcal{L}(\mathbf{j}_n)^\perp} \mathbf{X}_2 \boldsymbol{\beta}_2), & \text{otherwise.} \end{cases}$$

The ANOVA table for this test is given below. This ANOVA table is typically part of the output of regression software (e.g., PROC REG in SAS).

Source of Variation	Sum of Squares	df	Mean Squares	F
Regression	SSR $= \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y} - n\bar{y}^2$	k	$\frac{SSR}{k}$	$\frac{MSR}{MSE}$
Error	SSE $= \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathbf{y}$	$n-k-1$	$\frac{SSE}{n-k-1}$	
Total (Corr.)	SST $= \mathbf{y}^T\mathbf{y} - n\bar{y}^2$			

F test in terms of R^2:

The F test statistics we have just developed can be written in terms of R^2, the coefficient of determination. This relationship is given by the following theorem.

Theorem: The F statistic for testing $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ in the full rank model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$ (top of p. 138) can be written in terms of R^2 as

$$F = \frac{(R^2_{FM} - R^2_{RM})/h}{(1 - R^2_{FM})/(n-k-1)},$$

where R^2_{FM} corresponds to the full model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$, and R^2_{RM} corresponds to the reduced model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}^*_1 + \mathbf{e}^*$.

Proof: Homework.∎

Corollary: The F statistic for overall regression (for testing $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$) in the full rank model, $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i$, $i = 1, \ldots, n$, $e_1, \ldots, e_n \overset{iid}{\sim} N(0, \sigma^2)$ can be written in terms of R^2, the coefficient of determination from this model as follows:

$$F = \frac{R^2/k}{(1 - R^2)/(n-k-1)}.$$

Proof: For this hypothesis h, the dimension of the regression parameter being tested, is k. In addition, the reduced model here is

$$\mathbf{y} = \mathbf{j}_n\beta_0 + \mathbf{e},$$

so $(\mathbf{X}\hat{\boldsymbol{\beta}})_{RM}$, the estimated mean of $\mathbf{y}$, under the reduced model is $(\mathbf{X}\hat{\boldsymbol{\beta}})_{RM} = \mathbf{j}_n\bar{y}$. So, R^2_{RM} in the previous theorem is (cf. p. 131):

$$R^2_{RM} = \frac{[(\mathbf{X}\hat{\boldsymbol{\beta}})^T_{RM}\mathbf{y} - n\bar{y}^2]}{\mathbf{y}^T\mathbf{y} - n\bar{y}^2}$$
$$= \frac{[\bar{y}\overbrace{\mathbf{j}_n^T\mathbf{y}}^{=n\bar{y}} - n\bar{y}^2]}{\mathbf{y}^T\mathbf{y} - n\bar{y}^2} = 0.$$

The result now follows from the previous theorem. ∎

The General Linear Hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$

The hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ is called the general linear hypothesis. Here $\mathbf{C}$ is a $q \times (k+1)$ matrix of (known) coefficients with $\text{rank}(\mathbf{C}) = q$. We will consider the slightly simpler case $H : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ (i.e., $\mathbf{t} = \mathbf{0}$) first.

Most of the questions that are typically asked about the coefficients of a linear model can be formulated as hypotheses that can be written in the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, for some $\mathbf{C}$. For example, the hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ in the model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

can be written as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = (\underbrace{\mathbf{0}}_{h\times(k+1-h)}, \mathbf{I}_h) \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \boldsymbol{\beta}_2 = \mathbf{0}.$$

The test of overall regression can be written as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = (\underbrace{\mathbf{0}}_{k\times 1}, \mathbf{I}_k) \begin{pmatrix} \beta_0 \\ \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \mathbf{0}.$$

Hypotheses encompassed by $H_: \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ are not limitted to ones in which certain regression coefficients are set equal to zero. Another example that can be handled is the hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$. For example, suppose $k = 4$, then this hypothesis can be written as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \\ \beta_3 - \beta_4 \end{pmatrix} = \mathbf{0}.$$

Another equally good choice for $\mathbf{C}$ in this example is

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix}$$

The test statistic for $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is based on comparing $\mathbf{C}\hat{\boldsymbol{\beta}}$ to its null value $\mathbf{0}$, using a squared statistical distance (quadratic form) of the form

$$\begin{aligned} Q &= \{\mathbf{C}\hat{\boldsymbol{\beta}} - \underbrace{\mathrm{E}_0(\mathbf{C}\hat{\boldsymbol{\beta}})}_{=\mathbf{0}}\}^T \{\hat{\mathrm{var}}_0(\mathbf{C}\hat{\boldsymbol{\beta}})\}^{-1} \{\mathbf{C}\hat{\boldsymbol{\beta}} - \mathrm{E}_0(\mathbf{C}\hat{\boldsymbol{\beta}})\} \\ &= (\mathbf{C}\hat{\boldsymbol{\beta}})^T \{\hat{\mathrm{var}}_0(\mathbf{C}\hat{\boldsymbol{\beta}})\}^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}}). \end{aligned}$$

- Here, the 0 subscript is there to indicate that the expected value and variance are computed under H_0.

Recall that $\hat{\boldsymbol{\beta}} \sim N_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. Therefore,

$$\mathbf{C}\hat{\boldsymbol{\beta}} \sim N_q(\mathbf{C}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T).$$

We estimate σ^2 using $s^2 = \mathrm{MSE} = \mathrm{SSE}/(n-k-1)$, so

$$\hat{\mathrm{var}}_0(\mathbf{C}\hat{\boldsymbol{\beta}}) = s^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T$$

and Q becomes

$$\begin{aligned} Q &= (\mathbf{C}\hat{\boldsymbol{\beta}})^T \{s^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}\hat{\boldsymbol{\beta}} \\ &= \frac{(\mathbf{C}\hat{\boldsymbol{\beta}})^T \{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}}{\mathrm{SSE}/(n-k-1)} \end{aligned}$$

To use Q to form a test statistic, we need its distribution, which is given by the following theorem:

Theorem: If $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ where $\mathbf{X}$ is $n \times (k+1)$ of full rank and $\mathbf{C}$ is $q \times (k+1)$ of rank $q \leq k+1$, then

(i) $\mathbf{C}\hat{\boldsymbol{\beta}} \sim N_q[\mathbf{C}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]$;
(ii) $(\mathbf{C}\hat{\boldsymbol{\beta}})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}/\sigma^2 \sim \chi^2(q, \lambda)$, where

$$\lambda = (\mathbf{C}\boldsymbol{\beta})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}\boldsymbol{\beta}/(2\sigma^2);$$

(iii) $\text{SSE}/\sigma^2 \sim \chi^2(n-k-1)$; and
(iv) $(\mathbf{C}\hat{\boldsymbol{\beta}})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}$ and SSE are independent.

Proof: Part (i) follows from the normality of $\hat{\boldsymbol{\beta}}$ and that $\mathbf{C}\hat{\boldsymbol{\beta}}$ is an affine transformation of a normal. Part (iii) has been proved previously (p. 138).

(ii) Recall the theorem on the bottom of p. 82 (thm 5.5A in our text). This theorem said that if $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \Sigma)$ and $\mathbf{A}$ was $n \times n$ of rank r, then $\mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2(r, \frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu})$ iff $\mathbf{A}\Sigma$ is idempotent. Here $\mathbf{C}\hat{\boldsymbol{\beta}}$ plays the role of $\mathbf{y}$, $\mathbf{C}\boldsymbol{\beta}$ plays the role of $\boldsymbol{\mu}$, $\sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T$ plays the role of Σ, and $\{\sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}$ plays the role of $\mathbf{A}$. Then the result follows because $\mathbf{A}\Sigma = \{\sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T = \mathbf{I}$ is obviously idempotent.
(iv) Since $\hat{\boldsymbol{\beta}}$ and SSE are independent (p. 115) then $(\mathbf{C}\hat{\boldsymbol{\beta}})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}$ (a function of $\hat{\boldsymbol{\beta}}$) and SSE must be independent. ∎

Therefore,

$$F = Q/q = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}})^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}/q}{\text{SSE}/(n-k-1)} = \frac{\text{SSH}/q}{\text{SSE}/(n-k-1)}$$

has the form of a ratio of independent χ^2's each divided by its d.f.

- Here, SSH denotes $(\mathbf{C}\hat{\boldsymbol{\beta}})^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}$, the sum of squares due to the Hypothesis H_0.

Theorem: If $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ where $\mathbf{X}$ is $n \times (k+1)$ of full rank and $\mathbf{C}$ is $q \times (k+1)$ of rank $q \leq k+1$, then

$$\begin{aligned} F &= \frac{(\mathbf{C}\hat{\boldsymbol{\beta}})^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}/q}{\text{SSE}/(n-k-1)} \\ &= \frac{\text{SSH}/q}{\text{SSE}/(n-k-1)} \\ &\sim \begin{cases} F(q, n-k-1), & \text{if } H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \text{ is true;} \\ F(q, n-k-1, \lambda), & \text{if } H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \text{ is false,} \end{cases} \end{aligned}$$

where λ is as in the previous theorem.

Proof: Follows from the previous theorem and the definition of the F distribution. ∎

So, to conduct a hypothesis test of $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, we compute F and reject at level α if $F > F_{1-\alpha}(q, n-k-1)$ ($F_{1-\alpha}$ denotes the $(1-\alpha)^{\text{th}}$ quantile, or *upper* α^{th} quantile of the F distribution).

The general linear hypothesis as a test of nested models:

We have seen that the test of $\boldsymbol{\beta}_2 = \mathbf{0}$ in the model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$ can be formulated as a test of $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$. Therefore, special cases of the general linear hypothesis correspond to tests of nested (full and reduced) models. In fact, all F tests of the general linear hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ can be formulated as tests of nested models.

Theorem: The F test for the general linear hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is a full-and-reduced-model test.

Proof: The book, in combination with a homework problem, provides a proof based on Lagrange multipliers. Here we offer a different proof based on geometry.

Under H_0,

$$\begin{aligned} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad & \text{and} \qquad \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \\ \Rightarrow \quad & \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \\ \Rightarrow \quad & \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\mu} = \mathbf{0} \\ \Rightarrow \quad & \mathbf{T}^T\boldsymbol{\mu} = \mathbf{0} \quad \text{where} \quad \mathbf{T} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T. \end{aligned}$$

That is, under H_0, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \in C(\mathbf{X}) = V$ and $\boldsymbol{\mu} \perp C(\mathbf{T})$, or

$$\boldsymbol{\mu} \in [C(\mathbf{T})^\perp \cap C(\mathbf{X})] = V_0$$

where $V_0 = C(\mathbf{T})^\perp \cap C(\mathbf{X})$ is the orthogonal complement of $C(\mathbf{T})$ with respect to $C(\mathbf{X})$.

- Thus, under $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{\mu} \in V_0 \subset V = C(\mathbf{X})$, and under $H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}$, $\boldsymbol{\mu} \in V$ but $\boldsymbol{\mu} \notin V_0$. That is, these hypotheses correspond to nested models. It just remains to establish that the F test for these nested models is the F test for the general linear hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ given on p. 147.

The F test statistic for nested models given on p. 139 is

$$F = \frac{\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}/h}{\text{SSE}/(n-k-1)}$$

Here, we replace $\mathbf{P}_{C(\mathbf{X}_1)}$ by the projection matrix onto V_0:

$$\mathbf{P}_{V_0} = \mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{T})}$$

and replace h with $\dim(V) - \dim(V_0)$, the reduction in dimension of the model space when we go from the full to the reduced model.

Since V_0 is the orthogonal complement of $C(\mathbf{T})$ with respect to $C(\mathbf{X})$, $\dim(V_0)$ is given by

$$\dim(V_0) = \dim(C(\mathbf{X})) - \dim(C(\mathbf{T})) = \text{rank}(\mathbf{X}) - \text{rank}(\mathbf{T}) = k + 1 - q$$

Here, $\text{rank}(\mathbf{T}) = q$ by the following argument:

$$\text{rank}(\mathbf{T}) = \text{rank}(\mathbf{T}^T) \geq \text{rank}(\mathbf{T}^T\mathbf{X}) = \text{rank}(\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{C}) = q$$

and

$$\begin{aligned}\text{rank}(\mathbf{T}) = \text{rank}(\mathbf{T}^T\mathbf{T}) &= \text{rank}(\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T) \\ &= \text{rank}(\underbrace{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T}_{q\times q}) \leq q.\end{aligned}$$

Therefore, $q \leq \text{rank}(\mathbf{T}) \leq q \Rightarrow \text{rank}(\mathbf{T}) = q$.

So,

$$h = \dim(V) - \dim(V_0) = (k+1) - [(k+1) - q] = q.$$

Thus the full vs. reduced model F statistic becomes

$$F = \frac{\mathbf{y}^T[\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{V_0}]\mathbf{y}/q}{\text{SSE}/(n-k-1)} = \frac{\mathbf{y}^T[\mathbf{P}_{C(\mathbf{X})} - (\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{T})})]\mathbf{y}/q}{\text{SSE}/(n-k-1)}$$
$$= \frac{\mathbf{y}^T\mathbf{P}_{C(\mathbf{T})}\mathbf{y}/q}{\text{SSE}/(n-k-1)}$$

where

$$\begin{aligned}
\mathbf{y}^T\mathbf{P}_{C(\mathbf{T})}\mathbf{y} &= \mathbf{y}^T\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y} \\
&= \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \underbrace{\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}}_{=\hat{\boldsymbol{\beta}}^T}\mathbf{C}^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}_{=\hat{\boldsymbol{\beta}}} \\
&= \hat{\boldsymbol{\beta}}^T\mathbf{C}^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}
\end{aligned}$$

which is our test statistic for the general linear hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ from p. 147. ∎

The case $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ where $\mathbf{t} \neq \mathbf{0}$:

Extension to this case is straightforward. The only requirement is that the system of equations $\mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ be consistent, which is ensured by $\mathbf{C}$ having full row rank q.

Then the F test statistic for $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ is given by

$$F = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{t})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{t})/q}{\text{SSE}/(n-k-1)} \sim \begin{cases} F(q, n-k-1), & \text{under } H_0 \\ F(q, n-k-1, \lambda), & \text{otherwise,} \end{cases}$$

where $\lambda = (\mathbf{C}\boldsymbol{\beta} - \mathbf{t})^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{t})/(2\sigma^2)$.

Tests on β_j and on $\mathbf{a}^T\boldsymbol{\beta}$:

Tests of $H_0 : \beta_j = 0$ or $H_0 : \mathbf{a}^T\boldsymbol{\beta} = 0$ occur as special cases of the tests we have already considered. To test $H_0 : \mathbf{a}^T\boldsymbol{\beta} = 0$, we use $\mathbf{a}^T$ in place of $\mathbf{C}$ in our test of the general linear hypothesis $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$. In this case $q = 1$ and the test statistic becomes

$$F = \frac{(\mathbf{a}^T\hat{\boldsymbol{\beta}})^T[\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}]^{-1}\mathbf{a}^T\hat{\boldsymbol{\beta}}}{\text{SSE}/(n-k-1)} = \frac{(\mathbf{a}^T\hat{\boldsymbol{\beta}})^2}{s^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}$$
$$\sim F(1, n-k-1) \quad \text{under } H_0 : \mathbf{a}^T\boldsymbol{\beta} = 0.$$

- Note that since $t^2(\nu) = F(1, \nu)$, an equivalent test of $H_0 : \mathbf{a}^T\boldsymbol{\beta} = 0$ is given by the t-test with test statistic

$$t = \frac{\mathbf{a}^T\hat{\boldsymbol{\beta}}}{s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}} \sim t(n-k-1) \quad \text{under } H_0.$$

An important special case of the hypothesis $H_0 : \mathbf{a}^T\boldsymbol{\beta} = 0$ occurs when $\mathbf{a} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ where the 1 appears in the $j+1$th position. This is the hypothesis $H_0 : \beta_j = 0$, and it says that the j^{th} explanatory variable x_j has no partial regression effect on y (no effect above and beyond the effects of the other explanatory variables in the model).

The test statistic for this hypothesis simplifies from that given above to yield

$$F = \frac{\hat{\beta}_j^2}{s^2 g_{jj}} \sim F(1, n-k-1) \quad \text{under } H_0 : \beta_j = 0,$$

where g_{jj} is the j^{th} diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. Equivalently, we could use the t test statistic

$$t = \frac{\hat{\beta}_j}{s\sqrt{g_{jj}}} = \frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)} \sim t(n-k-1) \quad \text{under } H_0 : \beta_j = 0.$$

Confidence and Prediction Intervals

Hypothesis tests and confidence regions (e.g., intervals) are really two different ways to look at the same problem.

- For an α-level test of a hypothesis of the form $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$ is given by all those values of $\boldsymbol{\theta}_0$ such that the hypothesis would not be rejected. That is, the *acceptance region* of the α-level test is the $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$.

- Conversely, $\boldsymbol{\theta}_0$ falls outside of a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$ iff an α level test of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is rejected.

- That is, we can *invert* the statistical tests that we have derived to obtain confidence regions for parameters of the linear model.

Confidence Region for $\boldsymbol{\beta}$:

If we set $\mathbf{C} = \mathbf{I}_{k+1}$ and $\mathbf{t} = \boldsymbol{\beta}$ in the F statistic on the bottom of p. 150, we obtain

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/(k+1)}{s^2} \sim F(k+1, n-k-1)$$

From this distributional result, we can make the probability statement,

$$\Pr\left\{ \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{s^2(k+1)} \leq F_{1-\alpha}(k+1, n-k-1) \right\} = 1 - \alpha.$$

Therefore, the set of all vectors $\boldsymbol{\beta}$ that satisfy

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (k+1) s^2 F_{1-\alpha}(k+1, n-k-1)$$

forms a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\beta}$.

- Such a region is an ellipse, and is only easy to draw and make easy interpretation of for $k = 1$ (e.g., simple linear regression).

- If one can't plot the region and then plot a point to see whether its in or out of the region (i.e., for $k > 1$) then this region isn't any more informative than the test of $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. To decide whether $\boldsymbol{\beta}_0$ is in the region, we essentially have to perform the test!

- More useful are confidence intervals for the individual β_j's and for linear combinations of the form $\mathbf{a}^T\boldsymbol{\beta}$.

Confidence Interval for $\mathbf{a}^T\boldsymbol{\beta}$:

If we set $\mathbf{C} = \mathbf{a}^T$ and $\mathbf{t} = \mathbf{a}^T\boldsymbol{\beta}$ in the F statistic on the bottom of p. 150, we obtain

$$\frac{(\mathbf{a}^T\hat{\boldsymbol{\beta}} - \mathbf{a}^T\boldsymbol{\beta})^2}{s^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}} \sim F(1, n-k-1)$$

which implies

$$\frac{(\mathbf{a}^T\hat{\boldsymbol{\beta}} - \mathbf{a}^T\boldsymbol{\beta})}{s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}} \sim t(n-k-1).$$

From this distributional result, we can make the probability statement,

$$\Pr\left\{\underbrace{t_{\alpha/2}(n-k-1)}_{-t_{1-\alpha/2}(n-k-1)} \leq \frac{(\mathbf{a}^T\hat{\boldsymbol{\beta}} - \mathbf{a}^T\boldsymbol{\beta})}{s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}} \leq t_{1-\alpha/2}(n-k-1)\right\} = 1-\alpha.$$

Rearranging this inequality so that $\mathbf{a}^T\boldsymbol{\beta}$ falls in the middle, we get

$$\Pr\Big\{\mathbf{a}^T\hat{\boldsymbol{\beta}} - t_{1-\alpha/2}(n-k-1)s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}} \leq \mathbf{a}^T\boldsymbol{\beta} \\ \leq \mathbf{a}^T\hat{\boldsymbol{\beta}} + t_{1-\alpha/2}(n-k-1)s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}\Big\} = 1-\alpha.$$

Therefore, a $100(1-\alpha)\%$ CI for $\mathbf{a}^T\boldsymbol{\beta}$ is given by

$$\mathbf{a}^T\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}(n-k-1)s\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}.$$

Confidence Interval for β_j:

A special case of this interval occurs when $\mathbf{a} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$, where the 1 is in the $j+1$th position. In this case $\mathbf{a}^T\boldsymbol{\beta} = \beta_j$, $\mathbf{a}^T\hat{\boldsymbol{\beta}} = \hat{\beta}_j$, and $\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} = \{(\mathbf{X}^T\mathbf{X})^{-1}\}_{jj} \equiv g_{jj}$. The confidence interval for β_j is then given by

$$\hat{\beta}_j \pm t_{1-\alpha/2}(n-k-1)s\sqrt{g_{jj}}.$$

Confidence Interval for $\mathrm{E}(y)$:

Let $\mathbf{x}_0 = (1, x_{01}, x_{02}, \ldots, x_{0k})^T$ denote a particular choice of the vector of explanatory variables $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)^T$ and let y_0 denote the corresponding response.

We assume that the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ applies to $(y_0, \mathbf{x}_0)$ as well. This may be because $(y_0, \mathbf{x}_0)$ were in the original sample to which the model was fit (i.e., $\mathbf{x}_0^T$ is a row of $\mathbf{X}$), or because we believe that $(y_0, \mathbf{x}_0)$ will behave similarly to the data $(\mathbf{y}, \mathbf{X})$ in the sample. Then

$$y_0 = \mathbf{x}_0^T\boldsymbol{\beta} + e_0, \quad e_0 \sim N(0, \sigma^2)$$

where $\boldsymbol{\beta}$ and σ^2 are the same parameters in the fitted model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$.

Suppose we wish to find a CI for

$$\mathrm{E}(y_0) = \mathbf{x}_0^T\boldsymbol{\beta}.$$

This quantity is of the form $\mathbf{a}^T\boldsymbol{\beta}$ where $\mathbf{a} = \mathbf{x}_0$, so the BLUE of $\mathrm{E}(y_0)$ is $\mathbf{x}_0^T\hat{\boldsymbol{\beta}}$ and a $100(1-\alpha)\%$ CI for $\mathrm{E}(y_0)$ is given by

$$\mathbf{x}_0^T\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}(n-k-1)s\sqrt{\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

- This confidence interval holds for a particular value $\mathbf{x}_0^T\boldsymbol{\beta}$. Sometimes, it is of interest to form simultaneous confidence intervals around each and every point $\mathbf{x}_0^T\boldsymbol{\beta}$ for all $\mathbf{x}_0$ in the range of $\mathbf{x}$. That is, we sometimes desire a simultaneous confidence band for the entire regression line (or plane, for $k > 1$). The confidence interval given above, if plotted for each value of $\mathbf{x}_0$, does not give such a simultaneous band; instead it gives a "point-wise" band. For discussion of simultaneous intervals see §8.6.7 of our text.

- The confidence interval given above is for $\mathrm{E}(y_0)$, **not** for y_0 itself. $\mathrm{E}(y_0)$ is a parameter, y_0 is a random variable. Therefore, we can't *estimate* y_0 or form a confidence interval for it. However, we can predict its value, and an interval around that prediction that quantifies the uncertainty associated with that prediction is called a prediction interval.

Prediction Interval for an Unobserved y-value:

For an unobserved value y_0 with known explanatory vector $\mathbf{x}_0$ assumed to follow our linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, we predict y_0 by

$$\hat{y}_0 = \mathbf{x}_0^T\hat{\boldsymbol{\beta}}.$$

- Note that this predictor of y_0 coincides with our estimator of $\mathrm{E}(y_0)$. However, the uncertainty associated with the quantity $\mathbf{x}_0^T\hat{\boldsymbol{\beta}}$ as a predictor of y_0 is different from (greater than) its uncertainty as an estimator of $\mathrm{E}(y_0)$. *Why?* Because observations (e.g., y_0) are more variable than their means (e.g., $\mathrm{E}(y_0)$).

To form a CI for the estimator $\mathbf{x}_0^T\hat{\boldsymbol{\beta}}$ of $\mathrm{E}(y_0)$ we examine the variance of the error of estimation:

$$\mathrm{var}\{\mathrm{E}(y_0) - \mathbf{x}_0^T\hat{\boldsymbol{\beta}}\} = \mathrm{var}(\mathbf{x}_0^T\hat{\boldsymbol{\beta}}).$$

In contrast, to form a PI for the predictor $\mathbf{x}_0^T\hat{\boldsymbol{\beta}}$ of y_0, we examine the variance of the error of prediction:

$$\begin{aligned}\mathrm{var}(y_0 - \mathbf{x}_0^T\hat{\boldsymbol{\beta}}) &= \mathrm{var}(y_0) + \mathrm{var}(\mathbf{x}_0^T\hat{\boldsymbol{\beta}}) - 2\underbrace{\mathrm{cov}(y_0, \mathbf{x}_0^T\hat{\boldsymbol{\beta}})}_{0} \\ &= \mathrm{var}(\mathbf{x}_0^T\boldsymbol{\beta} + e_0) + \mathrm{var}(\mathbf{x}_0^T\hat{\boldsymbol{\beta}}) \\ &= \mathrm{var}(e_0) + \mathrm{var}(\mathbf{x}_0^T\hat{\boldsymbol{\beta}}) = \sigma^2 + \sigma^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0.\end{aligned}$$

Since σ^2 is unknown, we must estimate this quantity with s^2, yielding

$$\hat{\mathrm{var}}(y_0 - \hat{y}_0) = s^2\{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0\}.$$

It's not hard to show that

$$\frac{y_0 - \hat{y}_0}{s\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}} \sim t(n-k-1),$$

therefore

$$\Pr\left\{-t_{1-\alpha/2}(n-k-1) \le \frac{y_0 - \hat{y}_0}{s\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}} \le t_{1-\alpha/2}(n-k-1)\right\} = 1-\alpha.$$

Rearranging,

$$\begin{aligned}\Pr\Big\{&\hat{y}_0 - t_{1-\alpha/2}(n-k-1)s\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0} \le y_0 \\ &\le \hat{y}_0 + t_{1-\alpha/2}(n-k-1)s\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}\Big\} = 1-\alpha.\end{aligned}$$

Therefore, a $100(1-\alpha)\%$ prediction interval for y_0 is given by

$$\hat{y}_0 \pm t_{1-\alpha/2}(n-k-1)s\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

- Once again, this is a point-wise interval. Simultaneous prediction intervals for predicting multiple y-values with given coverage probability are discussed in §8.6.7.

Equivalence of the $F-$test and Likelihood Ratio Test:

Recall that for the classical linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, with normal, homoscedastic errors, the likelihood function is given by

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\{-\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2/(2\sigma^2)\},$$

or expressing the likelihood as a function of $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ instead of $\boldsymbol{\beta}$:

$$L(\boldsymbol{\mu}, \sigma^2; \mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\{-\|\mathbf{y} - \boldsymbol{\mu}\|^2/(2\sigma^2)\}.$$

- $L(\boldsymbol{\mu}, \sigma^2; \mathbf{y})$ gives the probability of observing $\mathbf{y}$ for specified values of the parameters $\boldsymbol{\mu}$ and σ^2 (to be more precise, the probability that the response vector is "close to" the observed value $\mathbf{y}$).
  - or, roughly, it measures how likely the data are for given values of the parameters.

The idea behind a likelihood ratio test (LRT) for some hypothesis H_0 is to compare the likelihood function maximized over the parameters subject to the restriction imposed by H_0 (the constrained maximum likelihood) with the likelihood function maximized over the parameters without assuming H_0 is true (the unconstrained maximum likelihood).

- That is, we compare how probable the data are under the most favorable values of the parameters subject to H_0 (the constrained MLEs), with how probable the data are under the most favorable values of the parameters under the maintained hypothesis (the unconstrained MLEs).

- If assuming H_0 makes the data substantially less probable than not assuming H_0, then we reject H_0.

Consider testing $H_0 : \boldsymbol{\mu} \in V_0$ versus $H_1 : \boldsymbol{\mu} \notin V_0$ under the maintained hypothesis that $\boldsymbol{\mu}$ is in V. Here $V_0 \subset V$ and $\dim(V_0) = k+1-h \leq k+1 = \dim(V)$.

Let $\hat{\mathbf{y}} = p(\mathbf{y}|V)$ and $\hat{\mathbf{y}}_0 = p(\mathbf{y}|V_0)$. Then the unconstrained MLEs of $(\boldsymbol{\mu}, \sigma^2)$ are $\hat{\boldsymbol{\mu}} = \hat{\mathbf{y}}$ and $\hat{\sigma}^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2/n$ and the constrained MLEs are $\hat{\boldsymbol{\mu}}_0 = \hat{\mathbf{y}}_0$ and $\hat{\sigma}_0^2 = \|\mathbf{y} - \hat{\mathbf{y}}_0\|^2/n$.

Therefore, the likelihood ratio statistic is

$$\begin{aligned} LR &= \frac{\sup_{\boldsymbol{\mu} \in V_0} L(\boldsymbol{\mu}, \boldsymbol{\sigma}^2; \mathbf{y})}{\sup_{\boldsymbol{\mu} \in V} L(\boldsymbol{\mu}, \boldsymbol{\sigma}^2; \mathbf{y})} = \frac{L(\hat{\mathbf{y}}_0, \hat{\sigma}_0^2)}{L(\hat{\mathbf{y}}, \hat{\sigma}^2)} \\ &= \frac{(2\pi\hat{\sigma}_0^2)^{-n/2} \exp\{-\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2/(2\hat{\sigma}_0^2)\}}{(2\pi\hat{\sigma}^2)^{-n/2} \exp\{-\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(2\hat{\sigma}^2)\}} \\ &= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{-n/2} \frac{\exp(-n/2)}{\exp(-n/2)} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{-n/2} \end{aligned}$$

- We reject for small values of LR. Typically in LRTs, we work with $\lambda = -2\log(LR)$ so that we can reject for large values of λ. In this case, $\lambda = n \log\left(\hat{\sigma}_0^2/\hat{\sigma}^2\right)$.

- Equivalently we reject for large values of $\left(\hat{\sigma}_0^2/\hat{\sigma}^2\right)$ where

$$\begin{aligned} \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right) &= \frac{\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2} = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2} \\ &= 1 + \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2} = 1 + \left(\frac{h}{n-k-1}\right) F \end{aligned}$$

  a monotone function of F (cf. the F statistic on the top of p. 138).

Therefore, large values of λ correspond to large values of F and the decision rules based on LR and on F are the same.

- Therefore, the LRT and the $F-$test are equivalent.

Analysis of Variance Models: The Non-Full Rank Linear Model

- To this point, we have focused exclusively on the case when the model matrix $\mathbf{X}$ of the linear model is of full rank. We now consider the case when $\mathbf{X}$ is $n \times p$ with $\text{rank}(\mathbf{X}) = k < p$.

- The basic ideas behind estimation and inference in this case are the same as in the full rank case, but the fact that $(\mathbf{X}^T\mathbf{X})^{-1}$ doesn't exist and therefore the normal equations have no unique solution causes a number of technical complications.

- We wouldn't bother to dwell on these technicalities if it weren't for the fact that the non-full rank case does arise frequently in applications in the form of analysis of variance models .

The One-way Model:

Consider the balanced one-way layout model for y_{ij} a response on the j^{th} unit in the i^{th} treatment group. Suppose that there are a treatments and n units in the i^{th} treatment group. The **cell-means** model for this situation is

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, \ldots, a, j = 1, \ldots, n,$$

where the e_{ij}'s are i.i.d. $N(0, \sigma^2)$.

An alternative, but equivalent, linear model is the **effects model** for the one-way layout:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, a, j = 1, \ldots, n,$$

with the same assumptions on the errors.

The cell means model can be written in vector notation as

$$\mathbf{y} = \mu_1\mathbf{x}_1 + \mu_2\mathbf{x}_2 + \cdots + \mu_a\mathbf{x}_a + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

and the effects model can be written as

$$\mathbf{y} = \mu\mathbf{j}_N + \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_a\mathbf{x}_a + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

where $\mathbf{x}_i$ is an indicator for treatment i, and $N = an$ is the total sample size.

- That is, the effects model has the same model matrix as the cell-means model, but with one extra column, a column of ones, in the first position.
- Notice that $\sum_i \mathbf{x}_i = \mathbf{j}_N$. Therefore, the columns of the model matrix for the effects model are linearly dependent.

Let $\mathbf{X}_1$ denote the model matrix in the cell-means model, $\mathbf{X}_2 = (\mathbf{j}_N, \mathbf{X}_1)$ denote the model matrix in the effects model.

- Note that $C(\mathbf{X}_1) = C(\mathbf{X}_2)$.

In general, two linear models $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e}_1$, $\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}_2$ with the same assumptions on $\mathbf{e}_1$ and $\mathbf{e}_2$ are equivalent linear models if $C(\mathbf{X}_1) = C(\mathbf{X}_2)$.

Why?

Because the mean vectors $\boldsymbol{\mu}_1 = \mathbf{X}_1\boldsymbol{\beta}_1$ and $\boldsymbol{\mu}_2 = \mathbf{X}_2\boldsymbol{\beta}_2$ in the two cases are both restricted to fall in the same subspace $C(\mathbf{X}_1) = C(\mathbf{X}_2)$.

In addition,

$$\hat{\boldsymbol{\mu}}_1 = p(\mathbf{y}|C(\mathbf{X}_1)) = p(\mathbf{y}|C(\mathbf{X}_2)) = \hat{\boldsymbol{\mu}}_2$$

is the same in both models, and

$$S^2 = \frac{1}{n - \dim(C(\mathbf{X}_1))}\|\mathbf{y} - \hat{\boldsymbol{\mu}}_1\|^2 = \frac{1}{n - \dim(C(\mathbf{X}_2))}\|\mathbf{y} - \hat{\boldsymbol{\mu}}_2\|^2$$

is the same in both models.

- The cell-means and effects models are simply reparameterizations of one-another. The relationship between the parameters in this case is very simple: $\mu_i = \mu + \alpha_i$, $i = 1, \ldots, a$.

- Let $V = C(\mathbf{X}_1) = C(\mathbf{X}_2)$. In the case of the cell-means model, $\text{rank}(\mathbf{X}_1) = a = \dim(V)$ and $\boldsymbol{\beta}_1$ is $a \times 1$. In the case of the effects model, $\text{rank}(\mathbf{X}_2) = a = \dim(V)$ but $\boldsymbol{\beta}_2$ is $(a+1) \times 1$. The effects model is **overparameterized**.

To understand overparameterization, consider the model

$$y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \ldots, a, j = 1, \ldots, n.$$

This model says that $\text{E}(y_{ij}) = \mu + \alpha_i = \mu_i$, or

$$\begin{aligned} \text{E}(y_{1j}) &= \mu + \alpha_1 = \mu_1 \quad j = 1, \ldots, n, \\ \text{E}(y_{2j}) &= \mu + \alpha_2 = \mu_2 \quad j = 1, \ldots, n, \\ &\vdots \end{aligned}$$

Suppose the true treatment means are $\mu_1 = 10$ and $\mu_2 = 8$. In terms of the parameters of the effects model, μ and the α_i's, these means can be represented in an infinity of possible ways,

$$\begin{aligned} \text{E}(y_{1j}) &= 10 + 0 \quad j = 1, \ldots, n, \\ \text{E}(y_{2j}) &= 10 + (-2) \quad j = 1, \ldots, n, \end{aligned}$$

($\mu = 10$, $\alpha_1 = 0$, and $\alpha_2 = -2$), or

$$\begin{aligned} \text{E}(y_{1j}) &= 8 + 2 \quad j = 1, \ldots, n, \\ \text{E}(y_{2j}) &= 8 + 0 \quad j = 1, \ldots, n, \end{aligned}$$

($\mu = 8$, $\alpha_1 = 2$, and $\alpha_2 = 0$), or

$$\begin{aligned} \text{E}(y_{1j}) &= 1 + 9 \quad j = 1, \ldots, n, \\ \text{E}(y_{2j}) &= 1 + 7 \quad j = 1, \ldots, n, \end{aligned}$$

($\mu = 1$, $\alpha_1 = 9$, and $\alpha_2 = 7$), etc.

Why would we want to consider an overparameterized model like the effects model?

In a simple case like the one-way layout, I would argue that we wouldn't.

The most important criterion for choice of parameterization of a model is interpretability. Without imposing any constraints, the parameters of the effects model do not have clear interpretations.

However, **subject to the constraint** $\sum_i \alpha_i = 0$, the parameters of the effects model have the following interpretations:

> μ =grand mean response across all treatments
> α_i =deviation from the grand mean placing μ_i (the i^{th} treatment mean) up or down from the grand mean; i.e., the effect of the i^{th} treatment.

Without the constraint, though, μ is not constrained to fall in the center of the μ_i's. μ is in no sense the grand mean, it is just an arbitrary baseline value.

In addition, adding the constraint $\sum_i \alpha_i = 0$ has essentially the effect of reparameterizing from the overparameterized (non-full rank) effects model to a just-parameterized (full rank) model that is equivalent (in the sense of having the same model space) as the cell means model.

To see this consider the one-way effects model with $a = 3$, $n = 2$. Then $\sum_{i=1}^{a} \alpha_i = 0$ implies $\alpha_1 + \alpha_2 + \alpha_3 = 0$ or $\alpha_3 = -(\alpha_1 + \alpha_2)$. Subject to the constraint, the effects model is

$$\mathbf{y} = \mu\mathbf{j}_N + \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3 + \mathbf{e}, \quad \text{where } \alpha_3 = -(\alpha_1 + \alpha_2),$$

or

$$\begin{aligned}\mathbf{y} &= \mu\mathbf{j}_N + \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + (-\alpha_1 - \alpha_2)\mathbf{x}_3 + \mathbf{e} \\ &= \mu\mathbf{j}_N + \alpha_1(\mathbf{x}_1 - \mathbf{x}_3) + \alpha_2(\mathbf{x}_2 - \mathbf{x}_3) + \mathbf{e} \\ &= \mu \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \alpha_1 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ -1 \\ -1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} + \mathbf{e},\end{aligned}$$

which has the same model space as the cell-means model.

Thus, when faced with a non-full rank model like the one-way effects model, we have three ways to proceed:

(1) Reparameterize to a full rank model.

- E.g., switch from the effects model to the cell-means model.

(2) Add constraints to the model parameters to remove the overparameterization.

- E.g., add a constraint such as $\sum_{i=1}^{a} \alpha_i = 0$ to the one-way effects model.
- Such constraints are usually called **side-conditions**.
- Adding side conditions essentially accomplishes a reparameterization to a full rank model as in (1).

(3) Analyze the model as a non-full rank model, but limit estimation and inference to those functions of the (overparameterized) parameters that can be uniquely estimated.

- Such functions of the parameters are called **estimable**.
- It is only in this case that we are actually using an overparameterized model, for which some new theory is necessary. (In cases (1) and (2) we remove the overparameterization somehow.)

Why would we choose option (3)?

Three reasons:

i. We can. Although we may lose nice parameter interpretations in using an unconstrained effects model or other unconstrained, non-full-rank model, there is no theoretical or methodological reason to avoid them (they can be handled with a little extra trouble).

ii. It is often easier to formulate an appropriate (and possibly overparameterized) model without worrying about whether or not its of full rank than to specify that model's "full-rank version" or to identify and impose the appropriate constraints on the model to make it full rank. This is especially true in modelling complex experimental data that are not balanced.

iii. Non-full rank model matrices may arise for reasons other than the structure of the model that's been specified. E.g., in an observational study, several explanatory variables may be colinear.

So, let's consider the overparameterized (non-full-rank) case.

- In the non-full-rank case, it is not possible to obtain linear unbiased estimators of all of the components of $\boldsymbol{\beta}$.

To illustrate this consider the effects version of the one-way layout model with no parameter constraints.

Can we find an unbiased linear estimator of α_1?

To be linear, such an estimator (call it T) would be of the form $T = \sum_i \sum_j d_{ij} y_{ij}$ for some coefficients $\{d_{ij}\}$. For T to be unbiased we require $\mathrm{E}(T) = \alpha_1$. However,

$$\mathrm{E}(T) = \mathrm{E}(\sum_i \sum_j d_{ij} y_{ij}) = \sum_i \sum_j d_{ij}(\mu + \alpha_i) = \mu d_{\cdot\cdot} + \sum_i d_{i\cdot}\alpha_i$$

Thus, the unbiasedness requirement $\mathrm{E}(T) = \alpha_1$ implies $d_{\cdot\cdot} = 0$, $d_{1\cdot} = 1$, $d_{2\cdot} = \cdots = d_{a\cdot} = 0$. This is impossible!

- So, α_1 is non-estimable. In fact, all of the parameters of the unconstrained one-way effects model are non-estimable. More generally, in any non-full rank linear model, at least one of the individual parameters of the model is not estimable.

If the parameters of a non-full rank linear model are non-estimable, what does least-squares yield?

Even if $\mathbf{X}$ is not of full rank, the least-squares criterion is still a reasonable one for estimation, and it still leads to the normal equation:

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}. \qquad (\clubsuit)$$

Theorem: For $\mathbf{X}$ and $n \times p$ matrix of rank $k < p \leq n$, ($\clubsuit$) is a consistent system of equations.

Proof: By the Theorem on p. 60 of these notes, ($\clubsuit$) is consistent iff

$$\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y}.$$

But this equation holds by result 3, on p. 57. ∎

So ($\clubsuit$) is consistent, and therefore has a *non-unique* (for $\mathbf{X}$ not of full rank) solution given

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y},$$

where $(\mathbf{X}^T\mathbf{X})^-$ is some (any) generalized inverse of $\mathbf{X}^T\mathbf{X}$.

What does $\hat{\boldsymbol{\beta}}$ estimate in the non-full rank case?

Well, in general a statistic estimates its expectation, so for a particular generalized inverse $(\mathbf{X}^T\mathbf{X})^-$, $\hat{\boldsymbol{\beta}}$ estimates

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathrm{E}\{(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}\} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathrm{E}(\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}.$$

- That is, in the non-full rank case, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$ is not unbiased for $\boldsymbol{\beta}$. This is not surprising given that we said earlier that $\boldsymbol{\beta}$ is not estimable.

- Note that $\mathrm{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$ depends upon which (of many possible) generalized inverses $(\mathbf{X}^T\mathbf{X})^-$ is used in $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$. That is, $\hat{\boldsymbol{\beta}}$, a solution of the normal equations, is not unique, and each possible choice estimates something different.

- This is all to reiterate that $\boldsymbol{\beta}$ is not estimable, and $\hat{\boldsymbol{\beta}}$ is not an estimator of $\boldsymbol{\beta}$ in the not-full rank model. However, certain linear combinations of $\boldsymbol{\beta}$ are estimable, and we will see that the corresponding linear combinations of $\hat{\boldsymbol{\beta}}$ are BLUEs of these estimable quantities.

Estimability: Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T$ be a vector of constants. The parameter $\boldsymbol{\lambda}^T\boldsymbol{\beta} = \sum_j \lambda_j\beta_j$ is said to be **estimable** if there exists a vector $\mathbf{a}$ in $\mathcal{R}^n$ such that

$$\mathrm{E}(\mathbf{a}^T\mathbf{y}) = \boldsymbol{\lambda}^T\boldsymbol{\beta}, \quad \text{for all } \boldsymbol{\beta} \in \mathcal{R}^p. \qquad (\dagger)$$

Since (†) is equivalent to $\mathbf{a}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, it follows that $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable if and only if there exists $\mathbf{a}$ such that $\mathbf{X}^T\mathbf{a} = \boldsymbol{\lambda}$ (i.e., iff $\boldsymbol{\lambda}$ lies in the row space of $\mathbf{X}$).

This and two other necessary and sufficient conditions for estimability of $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ are given in the following theorem:

Theorem: In the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$, the linear function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable if and only if any one of the following conditions hold:

(i) $\boldsymbol{\lambda}$ lies in the row space of $\mathbf{X}$. I.e., $\boldsymbol{\lambda} \in \mathbf{C}(\mathbf{X}^T)$, or, equivalently, if there exists a vector $\mathbf{a}$ such that

$$\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}.$$

(ii) $\boldsymbol{\lambda} \in C(\mathbf{X}^T\mathbf{X})$. I.e., if there exists a vector $\mathbf{r}$ such that

$$\boldsymbol{\lambda} = (\mathbf{X}^T\mathbf{X})\mathbf{r}.$$

(iii) $\boldsymbol{\lambda}$ satisfies

$$\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda} = \boldsymbol{\lambda},$$

where $(\mathbf{X}^T\mathbf{X})^-$ is any symmetric generalized inverse of $\mathbf{X}^T\mathbf{X}$.

Proof: Part (i) follows from the comment directly following the definition of estimability. That is, (†), the definition of estimability, is equivalent to $\mathbf{a}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, which happens iff $\mathbf{a}^T\mathbf{X} = \boldsymbol{\lambda}^T \Leftrightarrow \boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$, $\Leftrightarrow$ $\boldsymbol{\lambda} \in C(\mathbf{X}^T)$.

Now, condition (iii) is equivalent to condition (i) because (i) implies (iii): (i) implies

$$\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X} = \mathbf{a}^T\underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}}_{=\mathbf{X}} = \mathbf{a}^T\mathbf{X} = \boldsymbol{\lambda}^T$$

which, taking transposes of both sides, implies (iii); and (iii) implies (i): (iii) says

$$\boldsymbol{\lambda} = \mathbf{X}^T\underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}}_{=\mathbf{a}}$$

which is of the form $\mathbf{X}^T\mathbf{a}$ for $\mathbf{a} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}$.

Finally, condition (ii) is equivalent to condition (iii) because (ii) implies (iii): (ii) implies

$$\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}\mathbf{r} = \mathbf{X}^T\mathbf{X}\mathbf{r} = \boldsymbol{\lambda};$$

and (iii) implies (ii): (iii) says

$$\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{X}\underbrace{(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}}_{=\mathbf{a}}$$

which is of the form $\mathbf{X}^T\mathbf{X}\mathbf{a}$ for $\mathbf{a} = (\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}$. ∎

Example: Let $\mathbf{x}_1 = (1,1,1,1)^T$, $\mathbf{x}_2 = (1,1,1,0)^T$ and $\mathbf{x}_3 = 3\mathbf{x}_1 - 2\mathbf{x}_2 = (1,1,1,3)^T$. Then $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ is 4×3 but has rank of only 2.

Consider the linear combination $\eta = 5\beta_1 + 3\beta_2 + 9\beta_3 = \boldsymbol{\lambda}^T\boldsymbol{\beta}$, where $\boldsymbol{\lambda} = (5,3,9)^T$. η is estimable because $\boldsymbol{\lambda}$ is in the row space of $\mathbf{X}$:

$$\boldsymbol{\lambda} = \begin{pmatrix} 5 \\ 3 \\ 9 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 3 \end{pmatrix}}_{=\mathbf{X}^T} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \end{pmatrix}$$

The parameters β_1 and $\beta_1 - \beta_2$ are not estimable. Why? Because for $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ to be estimable, there must exist an $\mathbf{a}$ so that

$$\mathbf{X}^T\mathbf{a} = \boldsymbol{\lambda}$$

or

$$\begin{aligned} \mathbf{x}_1^T\mathbf{a} &= \lambda_1 \\ \mathbf{x}_2^T\mathbf{a} &= \lambda_2 \quad \text{and} \\ 3\mathbf{x}_1^T\mathbf{a} - 2\mathbf{x}_2^T\mathbf{a} &= \lambda_3 \end{aligned}$$

which implies $3\lambda_1 - 2\lambda_2 = \lambda_3$ must hold for $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ to be estimable. This equality does not hold for $\beta_1 = 1\beta_1 + 0\beta_2 + 0\beta_3$ or for $\beta_1 - \beta_2 = 1\beta_1 + (-1)\beta_2 + 0\beta_3$. It does hold for $\boldsymbol{\lambda} = (5,3,9)^T$ because $3(5) - 2(3) = 9$.

Theorem: In the non-full rank linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, the number of linearly independent estimable functions of $\boldsymbol{\beta}$ is the rank of $\mathbf{X}$.

Proof: This follows from the fact that estimable functins $\boldsymbol{\lambda}\boldsymbol{\beta}$ must satisfy $\boldsymbol{\lambda} \in C(\mathbf{X}^T)$ and $\dim\{C(\mathbf{X}^T)\} = \text{rank}(\mathbf{X}^T) = \text{rank}(\mathbf{X})$. ∎

- Let $\mathbf{x}_i^T$ be the i^{th} row of $\mathbf{X}$. Since each $\mathbf{x}_i$ is in the row space of $\mathbf{X}$, it follows that every $\mathbf{x}_i^T\boldsymbol{\beta}$ (every element of $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$) is estimable, $i = 1, \ldots, n$.

- Similarly, from the theorem on p. 165, every row (element) of $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$ is estimable, and therefore $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$ itself is estimable.

- In fact, all estimable functions can be obtained from $\mathbf{X}\boldsymbol{\beta}$ or $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$.

Theorem: In the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\text{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$, any estimable function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ can be obtained by taking a linear combination of the elements of $\mathbf{X}\boldsymbol{\beta}$ or of the elements of $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$.

Proof: Follows directly from the theorem on p. 166. ∎

Example: The one-way layout model (effects version).

Consider again the effects version of the (balanced) one way layout model:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, a, j = 1, \ldots, n.$$

Suppose that $a = 3$ and $n = 2$. Then, in matrix notation, this model is

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \mathbf{e}.$$

The previous theorem says that any estimable function of $\boldsymbol{\beta}$ can be obtained as a linear combination of the elements of $\mathbf{X}\boldsymbol{\beta}$. In addition, by the theorem on p. 166, vice versa (any linear combination of the elements of $\mathbf{X}\boldsymbol{\beta}$ is estimable).

So, any linear combination $\mathbf{a}^T\mathbf{X}\boldsymbol{\beta}$ for some $\mathbf{a}$ is estimable.

Examples:

$$\begin{aligned}
\mathbf{a}^T = (1,0,-1,0,0,0) \quad &\Rightarrow \quad \mathbf{a}^T\mathbf{X}\boldsymbol{\beta} = (0,1,-1,0)\boldsymbol{\beta} \\
&\qquad\qquad = \alpha_1 - \alpha_2 \\
\mathbf{a}^T = (0,0,1,0,-1,0) \quad &\Rightarrow \quad \mathbf{a}^T\mathbf{X}\boldsymbol{\beta} = (0,0,1,-1)\boldsymbol{\beta} \\
&\qquad\qquad = \alpha_2 - \alpha_3 \\
\mathbf{a}^T = (1,0,0,0,-1,0) \quad &\Rightarrow \quad \mathbf{a}^T\mathbf{X}\boldsymbol{\beta} = (0,1,0,-1)\boldsymbol{\beta} \\
&\qquad\qquad = \alpha_1 - \alpha_3 \\
\mathbf{a}^T = (1,0,0,0,0,0) \quad &\Rightarrow \quad \mathbf{a}^T\mathbf{X}\boldsymbol{\beta} = (1,1,0,0)\boldsymbol{\beta} \\
&\qquad\qquad = \mu + \alpha_1 \\
\mathbf{a}^T = (0,0,1,0,0,0) \quad &\Rightarrow \quad \mathbf{a}^T\mathbf{X}\boldsymbol{\beta} = (1,0,1,0)\boldsymbol{\beta} \\
&\qquad\qquad = \mu + \alpha_2 \\
\mathbf{a}^T = (0,0,0,0,1,0) \quad &\Rightarrow \quad \mathbf{a}^T\mathbf{X}\boldsymbol{\beta} = (1,0,0,1)\boldsymbol{\beta} \\
&\qquad\qquad = \mu + \alpha_3
\end{aligned}$$

- So, all treatment means (quantities of the form $\mu+\alpha_i$) are estimable, and all pairwise differences in the treatment effects (quantities of the form $\alpha_i - \alpha_j$ are estimable in the one-way layout model. Actually, any **contrast** in the treatment effects is estimable. A contrast is a linear combination whose coefficients sum to zero.

- Thus, even though the individual parameters $(\mu, \alpha_1, \alpha_2, \ldots)$ of the one-way layout model are non-estimable, it is still useful, because all of the quantities of interest in the model (treatment means and contrasts) *are* estimable.

Estimation in the non-full rank linear model:

A natural candidate for an estimator of an estimable function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is a solution of the least squares normal equation $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ (that is, where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$ for some generalized inverse $(\mathbf{X}^T\mathbf{X})^-$).

The following theorem shows that this estimator is unbiased, and even though $\hat{\boldsymbol{\beta}}$ is not unique, $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ is.

Theorem: Let $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ be an estimable function of $\boldsymbol{\beta}$ in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$. Let $\hat{\boldsymbol{\beta}}$ be any solution of the normal equation $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$. Then the estimator $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ has the following properties:

(i) (unbiasedness) $\mathrm{E}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{\lambda}^T\boldsymbol{\beta}$; and

(ii) (uniqueness) $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ is invariant to the choice of $\hat{\boldsymbol{\beta}}$ (to the choice of generalized inverse $(\mathbf{X}^T\mathbf{X})^-$ in the formula $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$.

Proof: Part (i):

$$\mathrm{E}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{\lambda}^T\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}^T\boldsymbol{\beta}$$

where the last equality follows from part (iii) of the theorem on p. 165.

Part (ii): Because $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable, $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$ for some $\mathbf{a}$. Therefore,

$$\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} = \mathbf{a}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} = \mathbf{a}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y}.$$

The result now follows from the fact that projection matrices are unique (see pp. 57–58). ∎

- Note that $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ can be written as $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} = \mathbf{r}^T\mathbf{X}^T\mathbf{y}$ for $\mathbf{r}$ a solution of $\mathbf{X}^T\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}$. (This fact is used quite a bit in our book).

Theorem: Under the conditions of the previous theorem, and where $\text{var}(\mathbf{e}) = \text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$, the variance of $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ is unique, and is given by

$$\text{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda},$$

where $(\mathbf{X}^T\mathbf{X})^-$ is any generalized inverse of $\mathbf{X}^T\mathbf{X}$.

Proof:

$$\begin{aligned}
\text{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}) &= \boldsymbol{\lambda}^T\text{var}((\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y})\boldsymbol{\lambda} \\
&= \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}\{(\mathbf{X}^T\mathbf{X})^-\}^T\boldsymbol{\lambda} \\
&= \sigma^2\underbrace{\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}}_{=\boldsymbol{\lambda}^T}\{(\mathbf{X}^T\mathbf{X})^-\}^T\boldsymbol{\lambda} \\
&= \sigma^2\boldsymbol{\lambda}^T\{(\mathbf{X}^T\mathbf{X})^-\}^T\boldsymbol{\lambda} \\
&= \sigma^2\mathbf{a}^T\mathbf{X}\{(\mathbf{X}^T\mathbf{X})^-\}^T\mathbf{X}^T\mathbf{a} \quad \text{(for some } \mathbf{a}\text{)} \\
&= \sigma^2\mathbf{a}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{a} = \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}.
\end{aligned}$$

Uniqueness: since $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$ for some $\mathbf{a}$. Therefore,

$$\begin{aligned}
\text{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}) &= \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda} \\
&= \sigma^2\mathbf{a}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{a} = \sigma^2\mathbf{a}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{a}
\end{aligned}$$

Again, the result follows from the fact that projection matrices are unique. ∎

Theorem: Let $\boldsymbol{\lambda}_1^T\boldsymbol{\beta}$ and $\boldsymbol{\lambda}_2^T\boldsymbol{\beta}$ be two estimable function in the model considered in the previous theorem (the spherical errors, non-full-rank linear model). Then the covariance of the least-squares estimators of these quantities is

$$\text{cov}(\boldsymbol{\lambda}_1^T\hat{\boldsymbol{\beta}}, \boldsymbol{\lambda}_2^T\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{\lambda}_1^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}_2.$$

Proof: Homework. ∎

In the full rank linear model, the Gauss-Markov theorem established that $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ was the BLUE of its mean $\boldsymbol{\lambda}^T\boldsymbol{\beta}$. This result holds in the non-full rnak linear model as well, as long as $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable.

Theorem: (Gauss-Markov in the non-full rank case) If $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable in the spherical errors non-full rank linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, then $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ is its BLUE.

Proof: Since $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is estimable, $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$ for some $\mathbf{a}$. $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} = \mathbf{a}^T\mathbf{X}\hat{\boldsymbol{\beta}}$ is a linear estimator because it is of the form

$$\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} = \mathbf{a}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} = \mathbf{a}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} = \mathbf{c}^T\mathbf{y}$$

where $\mathbf{c} = \mathbf{P}_{C(\mathbf{X})}\mathbf{a}$. We have already seen that $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ is unbiased. Consider any other linear estimator $\mathbf{d}^T\mathbf{y}$ of $\boldsymbol{\lambda}^T\boldsymbol{\beta}$. For $\mathbf{d}^T\mathbf{y}$ to be unbiased, the mean of $\mathbf{d}^T\mathbf{y}$, which is $\mathrm{E}(\mathbf{d}^T\mathbf{y}) = \mathbf{d}^T\mathbf{X}\boldsymbol{\beta}$, must satisfy $\mathrm{E}(\mathbf{d}^T\mathbf{y}) = \boldsymbol{\lambda}^T\boldsymbol{\beta}$, for all $\boldsymbol{\beta}$, or equivalently, it must satisfy $\mathbf{d}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}^T\boldsymbol{\beta}$, for all $\boldsymbol{\beta}$, which implies

$$\mathbf{d}^T\mathbf{X} = \boldsymbol{\lambda}^T.$$

The covariance between $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ and $\mathbf{d}^T\mathbf{y}$ is

$$\begin{aligned}\mathrm{cov}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}, \mathbf{d}^T\mathbf{y}) &= \mathrm{cov}(\mathbf{c}^T\mathbf{y}, \mathbf{d}^T\mathbf{y}) = \sigma^2\mathbf{c}^T\mathbf{d} \\ &= \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{d} = \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}.\end{aligned}$$

Now

$$\begin{aligned}0 &\leq \mathrm{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} - \mathbf{d}^T\mathbf{y}) = \mathrm{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}) + \mathrm{var}(\mathbf{d}^T\mathbf{y}) - 2\mathrm{cov}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}, \mathbf{d}^T\mathbf{y}) \\ &= \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda} + \mathrm{var}(\mathbf{d}^T\mathbf{y}) - 2\sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda} \\ &= \mathrm{var}(\mathbf{d}^T\mathbf{y}) - \underbrace{\sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}}_{=\mathrm{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}})}\end{aligned}$$

Therefore,

$$\mathrm{var}(\mathbf{d}^T\mathbf{y}) \geq \mathrm{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}})$$

with equality holding iff $\mathbf{d}^T\mathbf{y} = \boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$. I.e., an arbitrary linear unbiased estimator $\mathbf{d}^T\mathbf{y}$ has variance $\geq$ to that of the least squares estimator with equality iff the arbitrary estimator is the least squares estimator. ∎

ML Estimation:

In the not-necessarily-full-rank model with normal errors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n), \qquad (*)$$

where $\mathbf{X}$ is $n \times p$ with rank $k \leq p \leq n$, the ML estimators of $\boldsymbol{\beta}$, σ^2 change as expected from their values in the full rank case. That is, we replace inverses with generalized inverses in the formulas for the MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, and the MLE of $\boldsymbol{\beta}$ coincides with the OLS estimator, which is BLUE.

Theorem: In model (*) MLEs of $\boldsymbol{\beta}$ and σ^2 are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y},$$
$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2.$$

Proof: As in the full rank case, the loglikelihood is

$$\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

(cf. p. 110). By inspection, it is clear that the maximum of $\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ with respect to $\boldsymbol{\beta}$ is the same as the minimizer of $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, which is the least-squares criterion. Differentiating the LS criterion w.r.t. $\boldsymbol{\beta}$ gives the normal equations, which we know has solution $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$. Plugging $\hat{\boldsymbol{\beta}}$ back into $\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ gives the profile loglikelihood for σ^2, which we then maximize w.r.t. σ^2. These steps follow exactly as in the full rank case, leading to $\hat{\sigma}^2 = \frac{1}{n}||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2$. ∎

- Note that $\hat{\boldsymbol{\beta}}$ is not *the* (unique) MLE, but is an MLE of $\boldsymbol{\beta}$ corresponding to one particular choice of generalized inverse $(\mathbf{X}^T\mathbf{X})^-$.

- $\hat{\sigma}^2$ is the unique MLE of σ^2, though, because $\hat{\sigma}^2$ is a function of $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} = p(\mathbf{y}|C(\mathbf{X}))$, which is invariant to the choice of $(\mathbf{X}^T\mathbf{X})^-$.

$s^2 = \text{MSE}$ is an unbiased estimator of σ^2:

As in the full rank case, the MLE $\hat{\sigma}^2$ is biased as an estimator of σ^2, and is therefore not the preferred estimator. The bias of $\hat{\sigma}^2$ can be seen as follows:

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2) &= \frac{1}{n}\mathrm{E}\{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\} \\
&= \frac{1}{n}\mathrm{E}\{[(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T)\mathbf{y}]^T \underbrace{(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T)}_{\mathbf{P}_{C(\mathbf{X})^\perp}}\mathbf{y}\} \\
&= \frac{1}{n}\mathrm{E}\{\mathbf{y}^T\mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{y}\} \\
&= \frac{1}{n}\{\sigma^2 \dim[C(\mathbf{X})^\perp] + (\mathbf{X}\boldsymbol{\beta})\mathbf{P}_{C(\mathbf{X})^\perp}\underbrace{(\mathbf{X}\boldsymbol{\beta})}_{\in C(\mathbf{X})} \\
&= \frac{1}{n}\sigma^2(n - \dim[C(\mathbf{X})]) + 0 = \frac{1}{n}\sigma^2(n - \mathrm{rank}(\mathbf{X})) = \sigma^2\frac{n-k}{n}.
\end{aligned}$$

Therefore, an unbiased estimator of σ^2 is

$$s^2 = \frac{n}{n-k}\hat{\sigma}^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{\text{SSE}}{\text{dfE}} = \text{MSE}.$$

Theorem: In the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, $\mathrm{E}(\mathbf{e}) = \mathbf{0}$, $\mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{I}$, and where $\mathbf{X}$ is $n \times p$ of rank $k \leq p \leq n$, we have the following properties of s^2:

(i) (unbiasedness) $\mathrm{E}(s^2) = \sigma^2$.
(ii) (uniqueness) s^2 is invariant to the choice of $\hat{\boldsymbol{\beta}}$ (i.e., to the choice of generalized inverse $(\mathbf{X}^T\mathbf{X})^-$).

Proof: (i) follows from the construction of s^2 as $n\hat{\sigma}^2/(n-k)$ and the bias of $\hat{\sigma}^2$. (ii) follows from the uniqueness (invariance) of $\hat{\sigma}^2$. ∎

Distributions of $\hat{\boldsymbol{\beta}}$ and s^2:

In the normal-errors, not-necessarily full rank model (*), the distribution of $\hat{\boldsymbol{\beta}}$ and s^2 can be obtained. These distributional results are essentially the same as in the full rank case, except for the mean and variance of $\hat{\boldsymbol{\beta}}$:

Theorem: In model (*),

(i) For any given choice of $(\mathbf{X}^T\mathbf{X})^-$,

$$\hat{\boldsymbol{\beta}} \sim N_p[(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}\{(\mathbf{X}^T\mathbf{X})^-\}^T],$$

(ii) $(n-k)s^2/\sigma^2 \sim \chi^2(n-k)$, and

(iii) For any given choice of $(\mathbf{X}^T\mathbf{X})^-$, $\hat{\boldsymbol{\beta}}$ and s^2 are independent.

Proof: Homework. Proof is essentially the same as in the full rank case. Adapt the proof on p. 115. ∎

- In the full rank case we saw that with normal with spherical var-cov structure, $\hat{\boldsymbol{\beta}}$ and s^2 were minimimum variance unbiased estimators. This result continues to hold in the not-full-rank case.

Reparameterization:

The idea in reparameterization is to transform from the vector of non-estimable parameters $\boldsymbol{\beta}$ in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{X}$ is $n \times p$ with rank $k < p \leq n$, to a vector of linearly independent estimable functions of $\boldsymbol{\beta}$:

$$\begin{pmatrix} \mathbf{u}_1^T\boldsymbol{\beta} \\ \mathbf{u}_2^T\boldsymbol{\beta} \\ \vdots \\ \mathbf{u}_k^T\boldsymbol{\beta} \end{pmatrix} = \mathbf{U}\boldsymbol{\beta} \equiv \boldsymbol{\gamma}.$$

Here $\mathbf{U}$ is the $k \times p$ matrix with rows $\mathbf{u}_1^T, \ldots, \mathbf{u}_k^T$, so that the elements of $\boldsymbol{\gamma} = \mathbf{U}\boldsymbol{\beta}$ are a "full set" of linearly independent estimable functions of $\boldsymbol{\beta}$.

The new full-rank model is

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}, \qquad (*)$$

where $\mathbf{Z}$ is $n \times k$ of full rank, and $\mathbf{Z}\boldsymbol{\gamma} = \mathbf{X}\boldsymbol{\beta}$ (the mean under the non-full rank model is the same as under the full rank model, we've just changed the parameterization; i.e., switched from $\boldsymbol{\beta}$ to $\boldsymbol{\gamma}$.)

To find the new (full rank) model matrix $\mathbf{Z}$, note that $\mathbf{Z}\boldsymbol{\gamma} = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\gamma} = \mathbf{U}\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$ imply

$$\begin{aligned} \mathbf{ZU}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}, \quad \text{for all } \boldsymbol{\beta}, \quad &\Rightarrow \quad \mathbf{ZU} = \mathbf{X} \\ &\Rightarrow \quad \mathbf{ZUU}^T = \mathbf{XU}^T \\ &\Rightarrow \quad \mathbf{Z} = \mathbf{XU}^T(\mathbf{UU}^T)^{-1}. \end{aligned}$$

- Note that $\mathbf{U}$ is of full rank, so $(\mathbf{UU}^T)^{-1}$ exists.
- Note also that we have constructed $\mathbf{Z}$ to be of full rank:

  $$\text{rank}(\mathbf{Z}) \geq \text{rank}(\mathbf{ZU}) = \text{rank}(\mathbf{X}) = k$$

  but

  $$\text{rank}(\mathbf{Z}) \leq k, \quad \text{because } \mathbf{Z} \text{ is } n \times k.$$

  Therefore, $\text{rank}(\mathbf{Z}) = k$.

Thus the reparameterized model (*) is a full rank model, and we can obtain the BLUE of E($\mathbf{y}$) as

$$\hat{\boldsymbol{\mu}} = \mathbf{P}_{C(\mathbf{Z})}\mathbf{y} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y},$$

and the BLUE of $\boldsymbol{\gamma}$ as

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}.$$

If we are interested in any other estimable function of the original parameter $\boldsymbol{\beta}$ than those given by $\boldsymbol{\gamma} = \mathbf{U}\boldsymbol{\beta}$, such quantities are easily estimated from $\hat{\boldsymbol{\gamma}}$. Any estimable function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ must satisfy $\boldsymbol{\lambda}^T = \mathbf{a}^T\mathbf{X}$ for some $\mathbf{a}$. So

$$\boldsymbol{\lambda}^T\boldsymbol{\beta} = \mathbf{a}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{a}^T\mathbf{Z}\boldsymbol{\gamma} = \mathbf{b}^T\boldsymbol{\gamma}$$

for $\mathbf{b} = \mathbf{Z}^T\mathbf{a}$. Therefore, any estimable function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ can be written as $\mathbf{b}^T\boldsymbol{\gamma}$ for some $\mathbf{b}$ and the BLUE of $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\lambda}^T\boldsymbol{\beta}} = \mathbf{b}^T\hat{\boldsymbol{\gamma}}.$$

- Note that the choice of a "full set" of linearly independent estimable functions $\mathbf{U}\boldsymbol{\beta} = \boldsymbol{\gamma}$ is not unique. We could choose another set of LIN estimable functions $\mathbf{V}\boldsymbol{\beta} = \boldsymbol{\delta}$, and then reparameterize to a different full rank linear model $\mathbf{y} = \mathbf{W}\boldsymbol{\delta} + \mathbf{e}$ where $\mathbf{W}\boldsymbol{\delta} = \mathbf{Z}\boldsymbol{\gamma} = \mathbf{X}\boldsymbol{\beta}$. Any reparameterization leads to the same estimator of $\boldsymbol{\lambda}^T\boldsymbol{\beta}$.

Example: The Two-way ANOVA Model

In a two-way layout, observations are taken at all combinations of the levels of two treatment factors. Suppose factor A has a levels and factor B has b levels, then in a balanced two-way layout n observations are obtained in each of the ab treatments (combinations of A and B).

Let y_{ijk} = the k^{th} observation at the i^{th} level of A combined with the j^{th} level of B.

One way to think about the analysis of a two-way layout, is that if we ignore factors A and B, then what we have is really just a one-way experiment with ab treatments. Therefore, a one-way layout-type model, with a mean for each treatment can be used. This leads to the cell-means model for the two-way layout, which is a full-rank model:

$$y_{ijk} = \mu_{ij} + e_{ijk}, \quad i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n.$$

Often, an effects model is used instead for the two-way layout. In the effects model, the $(i, j)^{\text{th}}$ treatment mean is decomposed into a constant term plus additive effects for factor A, factor B, and factor A combined with factor B:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

This leads to the effects model for the two-way layout:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \quad i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n.$$

Suppose $a = 2$, $b = 3$ and $n = 2$. Then the effects model is

$$\begin{pmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{131} \\ y_{132} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{231} \\ y_{232} \end{pmatrix} = \mu \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \alpha_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \beta_3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

$$+ (\mathbf{I}_6 \otimes \begin{pmatrix} 1 \\ 1 \end{pmatrix}) \begin{pmatrix} (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{13} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22} \\ (\alpha\beta)_{23} \end{pmatrix} + \mathbf{e},$$

or

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{13} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22} \\ (\alpha\beta)_{23} \end{pmatrix}.$$

Obviously, the effects model is overparameterized and $\mathbf{X}$ is not of full rank. In fact, $\text{rank}(\mathbf{X}) = ab = 6$.

One way to reparameterize the effects model is to choose $\boldsymbol{\gamma}$ to be the vector of treatment means. That is, take

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \\ \gamma_6 \end{pmatrix} = \begin{pmatrix} \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11} \\ \mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12} \\ \mu + \alpha_1 + \beta_3 + (\alpha\beta)_{13} \\ \mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21} \\ \mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22} \\ \mu + \alpha_2 + \beta_3 + (\alpha\beta)_{23} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{13} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22} \\ (\alpha\beta)_{23} \end{pmatrix} = \mathbf{U}\boldsymbol{\beta}.$$

To reparameterize in terms of $\boldsymbol{\gamma}$, we can use

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \mathbf{I}_6 \otimes \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

I leave it to you to verify that $\mathbf{Z}\boldsymbol{\gamma} = \mathbf{X}\boldsymbol{\beta}$, and that $\mathbf{ZU} = \mathbf{X}$.

- Note that this choice of $\boldsymbol{\gamma}$ and $\mathbf{Z}$ amounts to a reparameterization from the effects model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ to the cell-means model $\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}$. That is, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_6)^T$ is just a relabelling of $(\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23})^T$.

- Any estimable function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ can be obtained as $\mathbf{b}^T\boldsymbol{\gamma}$ for some $\mathbf{b}$. For example, the main effect of A corresponds to

$$\begin{aligned}\boldsymbol{\lambda}^T\boldsymbol{\beta} &= \{\alpha_1 + \frac{1}{3}[(\alpha\beta)_{11} + (\alpha\beta)_{12} + (\alpha\beta)_{13}]\} - \{\alpha_2 + \frac{1}{3}[(\alpha\beta)_{21} + (\alpha\beta)_{22} + (\alpha\beta)_{23}]\} \\ &= \{\alpha_1 + (\bar{\alpha\beta})_{1.}\} - \{\alpha_2 + (\bar{\alpha\beta})_{2.}\}\end{aligned}$$

  (that is, $\boldsymbol{\lambda} = (0, 1, -1, 0, 0, 0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})^T$),which can be written as $\mathbf{b}^T\boldsymbol{\gamma}$ for $\mathbf{b} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})^T$.

Side Conditions:

Another approach for removing the rank deficiency of $\mathbf{X}$ in the non-full rank case is to impose linear constraints on the parameters, called side conditions. We have already seen one example (pp. 162–163): the one-way effects model with effects that sum to zero

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad \sum_j \alpha_j = 0,$$

for $i = 1, \ldots, a$, $j = 1, \ldots, n$.

Consider the case $a = 3$ and $n = 2$. Then the model can be written as

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}}_{=\mathbf{X}} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \mathbf{e}, \quad \text{where} \quad \alpha_1 + \alpha_2 + \alpha_3 = 0.$$

Imposing the constraint $\alpha_3 = -(\alpha_1 + \alpha_2)$ on the model equation, we can rewrite it as

$$\mathbf{y} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix}}_{=\tilde{\mathbf{X}}} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \mathbf{e},$$

where now we are back in the full rank case with the same model space, since $C(\mathbf{X}) = C(\tilde{\mathbf{X}})$.

Another Example - The Two-way Layout Model w/o Interaction:

We have seen that there are two equivalent models for the two-way layout with interaction: the cell-means model,

$$y_{ijk} = \mu_{ij} + e_{ijk}, \quad \begin{matrix} i = 1, \ldots, a \\ j = 1, \ldots, b \\ k = 1, \ldots, n_{ij} \end{matrix} \qquad (*)$$

and the effects model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}. \qquad (**)$$

- Model (**) is overparameterized and has a non-full rank model matrix. Model (*) is just-parameterized and has a full rank model matrix. However, they both have the same model space, and are therefore equivalent.

- We've now developed theory that allows us to use model (**) "as is" by restricting attention to estimable functions, using generalized inverses, etc.

- We've also seen that we can reparameterize from model (**) to model (*) by identifying the μ_{ij}'s where $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ as a full set of LIN estimable functions of the parameters in (**).

Another way to get around the overparameterization of (**) is to impose side conditions. Side conditions are not unique; there are lots of valid choices. But in this model, the set of side conditions that is most comonly used is

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0,$$

$$\sum_i \gamma_{ij} = 0 \quad \text{for each } j, \text{ and} \quad \sum_j \gamma_{ij} = 0 \quad \text{for each } i.$$

As in the one-way effects model example, substituting these constraints into the model equation leads to an equivalent full rank model.

- These side conditions, and those considered in the one-way model, are often called the "sum-to-zero constraints", or the "usual constraints", or sometimes "the anova constraints".

- The sum-to-zero constraints remove the rank deficiency in the model matrix, but they also give the parameters nice interpretations. E.g., in the on-way layout model, the constraint $\sum_i \alpha_i = 0$ forces μ to fall in the middle of all of the $\mu + \alpha_i$'s, rather than being some arbitrary constant. Thus, μ is the overall mean response averaged over all of the treatment groups, and the α_i's are deviations from this "grand mean" associated with the i^{th} treatment.

In both the one-way model and the two-way model with interaction, there's an obvious alternative (the cell-means model) to reparameterize to, so perhaps these aren't the best examples to motivate the use of side conditions.

A better example is the two-way model with no interaction. That is, suppose we want to assume that there is no interaction between the two treatment factors. That is, we want to assume that the difference between any two levels of factor A is the same across levels of factor B.

How could we formulate such a model?

The easiest way is just to set the interaction effects γ_{ij} in the effects model (**) to 0, yielding the (still overparameterized) model

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \qquad (***)$$

- In contrast, it is not so easy to see how a no-interaction, full-rank version of the cell-means model can be formed. And, therefore, reparameterization from (***) is not an easy option, since its not so obvious what the model is that we would like to reparameterize to.

Side conditions are a much easier option to remove the overparameterization in (***). Again, the "sum-to-zero" constraints are convenient because they remove the rank deficiency and give the parameters nice interpretations.

Under the sum-to-zero constraints, model (***) becomes

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \quad \begin{array}{l} \sum_i \alpha_i = 0 \\ \sum_j \beta_j = 0 \end{array} \qquad (\dagger)$$

In model (†) we can substitute

$$\alpha_a = -\sum_{i=1}^{a-1} \alpha_i \quad \text{and} \quad \beta_b = -\sum_{j=1}^{b-1} \beta_j$$

into the model equation

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \quad \begin{array}{l} i = 1, \ldots, a \\ j = 1, \ldots, b \\ k = 1, \ldots, n_{ij} \end{array}$$

For example, consider the following data from an unbalanced two-way layout in which rats were fed diets that differed in factor A, protein level (high and low), and factor B, food type (beef, cereal, pork). The response is weight gain.

High Protein			Low Protein		
Beef	Cereal	Pork	Beef	Cereal	Pork
73	98	94	90	107	49
102	74	79	76	95	82
	56		90		

Letting α_i represent the effect of the i^{th} level of protein and β_j be the effect of the j^{th} food type, the model matrix for model (***) (the unconstrained version of model (†)) based on these data is

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

If we add the constraints so that we use the full-rank model (†) instead of (***) we can substitute

$$\alpha_2 = -\alpha_1, \quad \text{and} \quad \beta_3 = -\beta_1 - \beta_2$$

so that the model mean becomes

$$\mathbf{X}\begin{pmatrix} \mu \\ \alpha_1 \\ -\alpha_1 \\ \beta_1 \\ \beta_2 \\ -\beta_1 - \beta_2 \end{pmatrix} = \mathbf{X}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 \end{pmatrix}\begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Therefore, the constrained model (†) is equivalent to a model with unconstrained parameter vector $(\mu, \alpha_1, \beta_1, \beta_2)^T$ and full rank model matrix

$$\tilde{\mathbf{X}} = \mathbf{X}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{pmatrix}$$

- Another valid set of side conditions in this model is

  $$\alpha_2 = 0, \quad \beta_3 = 0.$$

  I leave it to you to derive the reduced model matrix (the model matrix under the side conditions) for these conditions, and to convince yourself that these side consitions, like the sum-to-zero constraints, leave the model space unchanged.

- In the previous example, note that the two sets of side conditions were

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \alpha_1 + \alpha_2 \\ \beta_1 + \beta_2 + \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \alpha_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

  In each case, the side condition was of the form $\mathbf{T}\boldsymbol{\beta} = \mathbf{0}$ where $\mathbf{T}\boldsymbol{\beta}$ was a vector of non-estimable functions of $\boldsymbol{\beta}$.

This result is general. Side conditions must be restrictions on non-estimable functions of $\boldsymbol{\beta}$. If constraints are placed on estimable functions of $\boldsymbol{\beta}$ then this actually changes the model space, which is not our goal.

- Note also that in the example the rank deficiency of $\mathbf{X}$ was 2 ($\mathbf{X}$ had 6 columns but rank equal to 4). Therefore, two side conditions were necessary to remove this rank deficiency (the number of elements of $\mathbf{T}\boldsymbol{\beta}$ was 2).

In general, for the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathrm{E}(\mathbf{e}) = \mathbf{0}, \quad \mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{I}_n,$$

where $\mathbf{X}$ is $n \times p$ with $\mathrm{rank}(\mathbf{X}) = k < p \leq n$, we define **side conditions** to be a set of constraints of the form $\mathbf{T}\boldsymbol{\beta} = \mathbf{0}$ where $\mathbf{T}$ has rank q where $q = p - k$ (q =the rank deficiency), and

i. $\mathrm{rank}\begin{pmatrix}\mathbf{X}\\ \mathbf{T}\end{pmatrix} = p$, and

ii. $\mathrm{rank}\begin{pmatrix}\mathbf{X}\\ \mathbf{T}\end{pmatrix} = \mathrm{rank}(\mathbf{T}) + \mathrm{rank}(\mathbf{X})$.

- Note that (i) and (ii) imply $\mathbf{T}\boldsymbol{\beta}$ is nonestimable.

Theorem: In the spherical errors linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{X}$ is $n \times p$ of rank $k < p$, the unique least-squares estimator of $\boldsymbol{\beta}$ under the side condition $\mathbf{T}\boldsymbol{\beta} = \mathbf{0}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \mathbf{T}^T\mathbf{T})^{-1}\mathbf{X}^T\mathbf{y}.$$

Proof: As in problem 8.19 from your homework, we can use the method of Lagrange multipliers. Introducing a Lagrange multiplier $\boldsymbol{\lambda}$, the constrained least squares estimator of $\boldsymbol{\beta}$ minimizes

$$u = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\lambda}^T(\mathbf{T}\boldsymbol{\beta} - \mathbf{0})$$

Differentiating u with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ leads to the equations

$$\begin{aligned}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \frac{1}{2}\mathbf{T}^T\boldsymbol{\lambda} &= \mathbf{X}^T\mathbf{y}\\ \mathbf{T}\boldsymbol{\beta} &= \mathbf{0},\end{aligned}$$

which can be written as the single equation

$$\begin{pmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{T}^T\\ \mathbf{T} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}\\ \frac{1}{2}\boldsymbol{\lambda}\end{pmatrix} = \begin{pmatrix}\mathbf{X}^T\mathbf{y}\\ \mathbf{0}\end{pmatrix}.$$

Under the conditions for $\mathbf{T}\boldsymbol{\beta} = \mathbf{0}$ to be a side condition, it can be shown that $\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{T}^T \\ \mathbf{T} & \mathbf{0} \end{pmatrix}$ is nonsingular with inverse given by

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{T}^T \\ \mathbf{T} & \mathbf{0} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{H}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{H}^{-1} & \mathbf{H}^{-1}\mathbf{T}^T \\ \mathbf{T}\mathbf{H}^{-1} & \mathbf{0} \end{pmatrix},$$

where $\mathbf{H} = \mathbf{X}^T\mathbf{X} + \mathbf{T}^T\mathbf{T}$. (See Wang and Chow, *Advanced Linear Models*, §5.2, for details.)

Therefore, the constrained least-squares equations have a unique solution given by

$$\begin{aligned} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \frac{1}{2}\hat{\boldsymbol{\lambda}} \end{pmatrix} &= \begin{pmatrix} \mathbf{H}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{H}^{-1} & \mathbf{H}^{-1}\mathbf{T}^T \\ \mathbf{T}\mathbf{H}^{-1} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{H}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{H}^{-1}\mathbf{X}^T\mathbf{y} \\ \mathbf{T}\mathbf{H}^{-1}\mathbf{X}^T\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{H}^{-1}\mathbf{X}^T\mathbf{y} \\ \mathbf{0} \end{pmatrix} \end{aligned}$$

Here, the last equality follows because

i. $\mathbf{X}^T\mathbf{X}\mathbf{H}^{-1}\mathbf{X}^T = \mathbf{X}^T$ and
ii. $\mathbf{T}\mathbf{H}^{-1}\mathbf{X}^T = \mathbf{0}$.

To show (i) and (ii) let $\mathbf{X}_1$ be a $k \times p$ matrix containing the linearly independent rows of $\mathbf{X}$ and let $\mathbf{L} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{T} \end{pmatrix}$. Then $\mathbf{L}$ is a $p \times p$ nonsingular matrix.

There exists an $n \times k$ matrix $\mathbf{C}$ such that $\mathbf{X} = \mathbf{C}\mathbf{X}_1 = (\mathbf{C}, \mathbf{0})\mathbf{L}$. In addition, we can write $\mathbf{T} = \mathbf{0}\mathbf{X}_1 + \mathbf{T} = (\mathbf{0}, \mathbf{I}_q)\mathbf{L}$.

Therefore,

$$\mathbf{X}^T\mathbf{X} = \mathbf{L}^T \begin{pmatrix} \mathbf{C}^T\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{L}, \quad \text{and} \quad \mathbf{T}^T\mathbf{T} = \mathbf{L}^T \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \mathbf{L}.$$

Note that $\mathbf{C}^T\mathbf{C}$ is nonsingular (this follows from result 3 on rank, p. 15), so direct calculation gives

$$\begin{aligned}
\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{T}^T\mathbf{T})^{-1}\mathbf{X}^T &= \mathbf{L}^T \begin{pmatrix} \mathbf{C}^T\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{L} \left\{ \mathbf{L}^T \begin{pmatrix} \mathbf{C}^T\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \mathbf{L} \right\}^{-1} \mathbf{X}^T \\
&= \mathbf{L}^T \begin{pmatrix} \mathbf{C}^T\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{L} \left\{ \mathbf{L}^{-1} \begin{pmatrix} (\mathbf{C}^T\mathbf{C})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} (\mathbf{L}^T)^{-1} \right\} \mathbf{X}^T \\
&= \mathbf{L}^T \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} (\mathbf{L}^T)^{-1}\mathbf{X}^T = \mathbf{L}^T \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} (\mathbf{L}^T)^{-1}\mathbf{L}^T \begin{pmatrix} \mathbf{C}^T \\ \mathbf{0} \end{pmatrix} \\
&= \mathbf{L}^T \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{C}^T \\ \mathbf{0} \end{pmatrix} = \mathbf{L}^T \begin{pmatrix} \mathbf{C}^T \\ \mathbf{0} \end{pmatrix} = \mathbf{X}^T
\end{aligned}$$

establishing (i) and

$$\begin{aligned}
&\mathbf{T}(\mathbf{X}^T\mathbf{X} + \mathbf{T}^T\mathbf{T})^{-1}\mathbf{X}^T \\
&\quad = (\mathbf{0}, \mathbf{I}_q)\mathbf{L} \left\{ \mathbf{L}^{-1} \begin{pmatrix} (\mathbf{C}^T\mathbf{C})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} (\mathbf{L}^T)^{-1} \right\} \mathbf{L}^T \begin{pmatrix} \mathbf{C}^T \\ \mathbf{0} \end{pmatrix} = \mathbf{0}
\end{aligned}$$

which establishes (ii). ∎

Example - Weight Gain in Rats (Continued):

Returning to the data of p. 186, we now have two equivalent methods of obtaining the unique least-squares parameter estimates of the constrained model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \quad \begin{matrix} \sum_i \alpha_i = 0 \\ \sum_j \beta_j = 0 \end{matrix} \qquad (\dagger)$$

First, we can solve the constraints to yield

$$\alpha_2 = -\alpha_1, \quad \text{and} \quad \beta_3 = -\beta_1 - \beta_2.$$

Substituting into the model equation gives the full rank model matrix $\tilde{\mathbf{X}}$ given on p. 187. Thus the least-squares estimate for the unconstrained parameter vector $\boldsymbol{\delta} = (\mu, \alpha_1, \beta_1, \beta_2)^T$ based on model (†) is given by

$$\hat{\boldsymbol{\delta}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{y} = \begin{pmatrix} 82.733 \\ -0.941 \\ 3.278 \\ 3.455 \end{pmatrix}$$

Alternatively, we can use the method of Lagrange multipliers to obtain the least-squares estimate of the original parameter vector $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3)^T$ subject to the constraints. This estimator is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \mathbf{T}^T\mathbf{T})^{-1}\mathbf{X}^T\mathbf{y} \qquad (\ddagger)$$

where $\mathbf{T}$ is the constraint matrix given on the top of p. 187:

$$\mathbf{T} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Substituting the original model matrix $\mathbf{X}$ from p. 186 and $\mathbf{T}$ into (‡) we obtain

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 82.733 \\ -0.941 \\ 0.941 \\ 3.278 \\ 3.455 \\ -6.733 \end{pmatrix}.$$

- Note that the solution $\hat{\boldsymbol{\beta}}$ is one (of infinitely many) valid solutions to the unconstrained model (***). It corresponds to one possible choice of generalized inverses for $\mathbf{X}^T\mathbf{X}$.

- In particular it corresponds to choosing $(\mathbf{X}^T\mathbf{X} + \mathbf{T}^T\mathbf{T})^{-1}$ as the generalized inverse of $\mathbf{X}^T\mathbf{X}$. That $(\mathbf{X}^T\mathbf{X} + \mathbf{T}^T\mathbf{T})^{-1}$ is a generalized inverse of $\mathbf{X}^T\mathbf{X}$ follows from result (i) on p.190. By (i),

$$\underbrace{\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{T}^T\mathbf{T})^{-1}\mathbf{X}^T}_{=\mathbf{X}^T}\mathbf{X} = \mathbf{X}^T\mathbf{X}$$

  which is the defining property of a generalized inverse.

- Whichever approach we take, we obtain the same estimate of $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and of any estimable functions of $\boldsymbol{\beta}$ (for example, $\alpha_1 - \alpha_2$, the difference in treatment effects for high and low protein) in our original overparameterized model. See ratsexamp.txt.

Hypothesis Testing:

For inference, we need distributional assumptions, so throughout this section we assume the non-full rank linear model with normal errors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n), \qquad (\spadesuit)$$

where $\mathbf{X}$ is $n \times p$ with rank $k < p \leq n$.

Testable hypotheses:

Suppose we are interested in testing a hypothesis of the form

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}.$$

- Not all such hypothesis are testable. E.g., if $\mathbf{C}\boldsymbol{\beta}$ is nonestimable (is a vector with non-estimable elements) then H_0 cannot be tested.

- This should come as no surprise. E.g., in the one-way layout model $y_{ij} = \mu + \alpha_i + e_{ij}$ without any constraints, we cannot test $\mu = \mu_0$ because μ is not identifiable. μ could be any one of an infinite number of values without changing the model (as long as we change the α_i's accordingly), therefore how could we test whether its equal to a given null value?

A hypothesis is said to be testable when we can calculate an $F-$statistic that is suitable for testing it. There are three conditions that $\mathbf{C}$ must satisfy for H_0 to be testable:

1. $\mathbf{C}\boldsymbol{\beta}$ must be estimable (must have estimable elements).
   $\Rightarrow \mathbf{C}^T$ must lie in the row space of $\mathbf{X}$.
   $\Rightarrow$ there must exist a $\mathbf{A}$ so that $\mathbf{C}^T = \mathbf{X}^T\mathbf{A}$ or, equivalently, $\mathbf{C} = \mathbf{A}^T\mathbf{X}$ for some $\mathbf{A}$.
   (It would be meaningless to test hypotheses on nonestimable hypotheses anyway.)

2. $\mathbf{C}$ must have full row rank. I.e., for $\mathbf{C}$ $m \times p$, we require $\text{rank}(\mathbf{C}) = m$.
   $\Rightarrow$ this ensures that the hypothesis contains no redundant statements.

   – E.g., suppose $\boldsymbol{\beta}$ is 3×1 and we wish to test the hypothesis that $\beta_1 = \beta_2 = \beta_3$. We can express this hypothesis in the form $H_0 : \mathbf{C}\boldsymbol{\beta}$ for (infinitely) many choices of $\mathbf{C}$. A valid choice of $\mathbf{C}$ is

   $$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$$

   Notice that this 2×3 matrix has row rank 2.

   An invalid choice of $\mathbf{C}$ is

   $$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \end{pmatrix}$$

   Notice that the last row is redundant (given that $\beta_1 = \beta_2$ and $\beta_2 = \beta_3$ its redundant to require that $\beta_1 = \beta_3$), and $\text{rank}(\mathbf{C}) = 2$.

3. $\mathbf{C}$ must have no more than $\text{rank}(\mathbf{X}) = k$ rows.
   This is because, in general, one can only construct up to $\text{rank}(\mathbf{X})$ linearly independent estimable functions.

Subject to these conditions, H_0 is testable. As in the full rank case, there are two equivalent approaches to testing H_0:

1. Formulate a test statistic based on a comparison between $\mathbf{C}\hat{\boldsymbol{\beta}}$ and its null value $\mathbf{0}$.

2. Recast the null hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ in an equivalent form $H_0 : \boldsymbol{\mu} \in V_0$ for $\boldsymbol{\mu} = \text{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and an appropriate subspace $V_0 \subset V = C(\mathbf{X})$. That is, translate testing H_0 into a full vs. reduced model testing problem.

- In the full rank case we described both of these approaches. Because of time constraints, we'll only describe approach 1 in the non-full rank case, but approach 2 follows in much the same way as before.

The General Linear Hypothesis:

As in the full rank case, our F test is based on the quadratic form

$$\begin{aligned}&\{\mathbf{C}\hat{\boldsymbol{\beta}} - \mathrm{E}_0(\mathbf{C}\hat{\boldsymbol{\beta}})\}^T\{\hat{\mathrm{var}}_0(\mathbf{C}\hat{\boldsymbol{\beta}})\}^{-1}\{\mathbf{C}\hat{\boldsymbol{\beta}} - \mathrm{E}_0(\mathbf{C}\hat{\boldsymbol{\beta}})\}\\ &\quad = (\mathbf{C}\hat{\boldsymbol{\beta}})^T\{\hat{\mathrm{var}}_0(\mathbf{C}\hat{\boldsymbol{\beta}})\}^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}})\end{aligned}$$

(Here the 0 subscript indicates that the expected value and variance are computed under $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$.)

The theorem leading to the F statistic is as follows:

Theorem: In model (♠), if $\mathbf{C}$ is $m \times p$ of rank $m \leq k = \mathrm{rank}(\mathbf{X})$ such that $\mathbf{C}\boldsymbol{\beta}$ is a set of m LIN estimable functions, and if $\hat{\boldsymbol{\beta}} = \mathbf{G}\mathbf{X}^T\mathbf{y}$, for some generalized inverse $\mathbf{G}$ of $\mathbf{X}^T\mathbf{X}$, then

(i) $\mathbf{C}\mathbf{G}\mathbf{C}^T$ is nonsingular and invariant to the choice of $\mathbf{G}$;

(ii) $\mathbf{C}\hat{\boldsymbol{\beta}} \sim N_m(\mathbf{C}\boldsymbol{\beta}, \sigma^2\mathbf{C}\mathbf{G}\mathbf{C}^T)$;

(iii) $\mathrm{SSH}/\sigma^2 = (\mathbf{C}\hat{\boldsymbol{\beta}})^T(\mathbf{C}\mathbf{G}\mathbf{C}^T)^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}/\sigma^2 \sim \chi^2(m, \lambda)$, where

$$\lambda = (\mathbf{C}\boldsymbol{\beta})^T(\mathbf{C}\mathbf{G}\mathbf{C}^T)^{-1}\mathbf{C}\boldsymbol{\beta}/(2\sigma^2);$$

(iv) $\mathrm{SSE}/\sigma^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}^T)\mathbf{y}/\sigma^2 \sim \chi^2(n-k)$; and

(v) SSH and SSE are independent.

Proof:

(i) Since $\mathbf{C}\boldsymbol{\beta}$ is a vector of estimable function, there must exist an $\mathbf{A}$ so that $\mathbf{C} = \mathbf{A}^T\mathbf{X}$. Therefore,

$$\mathbf{C}\mathbf{G}\mathbf{C}^T = \mathbf{A}^T\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{A} = \mathbf{A}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{A},$$

which is unique. To show $\mathbf{C}\mathbf{G}\mathbf{C}^T$ is nonsingular we show that it is of full rank.

In general, we have the following two results about rank: (i) for any matrix $\mathbf{M}$, $\text{rank}(\mathbf{M}^T\mathbf{M}) = \text{rank}(\mathbf{M}^T)$; and (ii) for any matrices $\mathbf{M}$ and $\mathbf{N}$, $\text{rank}(\mathbf{MN}) \leq \text{rank}(\mathbf{M})$.

In addition, $\mathbf{G}$ can be chosen to be a symmetric generalized inverse of $\mathbf{X}^T\mathbf{X}$ (this is always possible), so $\mathbf{G}$ can be written $\mathbf{G} = \mathbf{L}^T\mathbf{L}$ for some $\mathbf{L}$.

Therefore,

$$\begin{aligned}\text{rank}(\mathbf{CGC}^T) &= \text{rank}(\mathbf{CL}^T\mathbf{LC}^T) = \text{rank}(\mathbf{CL}^T) \\ &\geq \text{rank}(\mathbf{CL}^T\mathbf{L}) = \text{rank}(\mathbf{CG}) = \text{rank}(\mathbf{A}^T\mathbf{XG}) \\ &\geq \text{rank}(\mathbf{A}^T\mathbf{XGX}^T\mathbf{X}) = \text{rank}(\mathbf{A}^T\mathbf{X}) = \text{rank}(\mathbf{C}) = m\end{aligned}$$

So we've established that $\text{rank}(\mathbf{CGC}^T) \geq m$. But, since $\mathbf{CGC}^T$ is $m \times m$ it follows that $\text{rank}(\mathbf{CGC}^T) \leq m$. Together, these results imply

$$\text{rank}(\mathbf{CGC}^T) = m.$$

(ii) By the theorem on p. 176,

$$\hat{\boldsymbol{\beta}} \sim N_p[\mathbf{GX}^T\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{GX}^T\mathbf{XG}].$$

Therefore, Since $\mathbf{C}\hat{\boldsymbol{\beta}}$ is an affine transformation of a normal, and $\mathbf{C} = \mathbf{A}^T\mathbf{X}$ for some $\mathbf{A}$,

$$\begin{aligned}\mathbf{C}\hat{\boldsymbol{\beta}} &\sim N_m[\mathbf{CGX}^T\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{CGX}^T\mathbf{XGC}^T] \\ &= N_m[\mathbf{A}^T\underbrace{\mathbf{XGX}^T\mathbf{X}}_{=\mathbf{X}}\boldsymbol{\beta}, \sigma^2\mathbf{A}^T\underbrace{\mathbf{XGX}^T\mathbf{X}}_{=\mathbf{X}}\mathbf{GX}^T\mathbf{A}] \\ &= N_m(\mathbf{C}\boldsymbol{\beta}, \sigma^2\mathbf{CGC}^T].\end{aligned}$$

(iii) By part (ii), $\text{var}(\mathbf{C}\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{C}\mathbf{G}\mathbf{C}^T$. Therefore, SSH/σ^2 is a quadratic form in a normal random vector. Since,

$$\sigma^2[\mathbf{C}\mathbf{G}\mathbf{C}^T]^{-1}\mathbf{C}\mathbf{G}\mathbf{C}^T/\sigma^2 = \mathbf{I},$$

the result follows from the theorem on the bottom of p. 81.

(iv) This was established in part (ii) of the theorem on p. 176.

(v) Homework. ∎

Putting the results of this theorem together, we obtain the F statistic for testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ for H_0 a testable hypothesis:

Theorem: In the setup of the previous theorem, then the F statistic for testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is as follows:

$$\begin{aligned} F &= \frac{\text{SSH}/m}{\text{SSE}/(n-k)} \\ &= \frac{(\mathbf{C}\hat{\boldsymbol{\beta}})^T[\mathbf{C}\mathbf{G}\mathbf{C}^T]^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}/m}{\text{SSE}/(n-k)} \\ &\sim F(m, n-k, \lambda), \end{aligned}$$

where $\mathbf{G}$ is a generalized inverse of $\mathbf{X}^T\mathbf{X}$ and

$$\lambda = \begin{cases} \frac{1}{2\sigma^2}(\mathbf{C}\boldsymbol{\beta})^T(\mathbf{C}\mathbf{G}\mathbf{C}^T)^{-1}(\mathbf{C}\boldsymbol{\beta}) & \text{in general} \\ 0 & \text{under } H_0. \end{cases}$$

Proof: Follows directly from the previous theorem and the definition of the noncentral F distribution. ∎

- As in the full rank case, this F statistic can be extended to test a hypothesis of the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ for $\mathbf{t}$ a vector of constants, $\mathbf{C}\boldsymbol{\beta}$ estimable, and $\mathbf{C}$ of full row rank. The resulting test statistic is identical to F above, but with $\mathbf{C}\hat{\boldsymbol{\beta}}$ replaced by $\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{t}$.

Breaking up Sums of Squares:

Consider again the problem of testing a full versus reduced model. That is suppose we are interested in testing the hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ in the model

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \mathbf{e} \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \qquad \text{(FM)}\end{aligned}$$

Under $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ the model becomes

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \mathbf{e}^*, \quad \mathbf{e}^* \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \qquad \text{(RM)}$$

The problem is to test

$$H_0 : \boldsymbol{\mu} \in C(\mathbf{X}_1) \quad \text{(RM)} \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \notin C(\mathbf{X}_1)$$

under the *maintained hypothesis* that $\boldsymbol{\mu} \in C(\mathbf{X}) = C([\mathbf{X}_1, \mathbf{X}_2])$ (FM).

When discussing the full rank CLM, we saw that the appropriate test statistic for this problem was

$$\begin{aligned}F &= \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/h}{s^2} = \frac{\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}/h}{\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathbf{y}/(n-p)} \\ &\sim \begin{cases} F(h, n-p), & \text{under } H_0; \text{ and} \\ F(h, n-p, \lambda_1), & \text{under } H_1, \end{cases}\end{aligned}$$

where

$$\lambda_1 = \frac{1}{2\sigma^2}\|(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\boldsymbol{\mu}\|^2 = \frac{1}{2\sigma^2}\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2,$$

$h = \dim(\boldsymbol{\beta}_2) = \text{rank}(\mathbf{X}_2) = \text{rank}(\mathbf{X}) - \text{rank}(\mathbf{X}_1)$, and $p = \dim(\boldsymbol{\beta}) = \text{rank}(\mathbf{X}) =$ the number of columns in $\mathbf{X}$ (which we called $k+1$, previously).

More generally, in the not-necessarily full rank CLM where $\mathbf{X}$ is $n \times p$ with $\text{rank}(\mathbf{X}) = k \leq p < n$, this result generalizes:

$$\begin{aligned} F &= \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/m}{s^2} \\ &= \frac{\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}/m}{\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathbf{y}/(n-k)} \\ &\sim \begin{cases} F(m, n-k), & \text{under } H_0; \text{ and} \\ F(m, n-k, \lambda_1), & \text{under } H_1, \end{cases} \end{aligned}$$

where λ_1 is as before and $m = \{\text{rank}(\mathbf{X}) - \text{rank}(\mathbf{X}_1)\}$.

Recall that the squared projection length $\mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}$ in the numerator of this F statistic is equal to

$$SS(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) \equiv SSR(FM) - SSR(RM) = SSE(RM) - SSE(FM)$$

so that $F = \frac{SS(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/m}{MSE}$.

The quantity $SS(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$ goes by several different names: *the extra regression sum of squares* or *the reduction in error sum of squares* due to $\boldsymbol{\beta}_2$ after fitting $\boldsymbol{\beta}_1$, and several different notations: $R(\boldsymbol{\beta}|\boldsymbol{\beta}_1)$, $SSR(X_2|X_1)$, etc.

Regardless of the notation or terminology, $SS(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$ is a *sum of squares* that quantifies the amount of variability in $\mathbf{y}$ accounted for by adding $\mathbf{X}_2\boldsymbol{\beta}_2$ to the regresion model that includes only $\mathbf{X}_1\boldsymbol{\beta}_1$.

That is, it quantifies the contribution to the regression model of the explanatory variables in $\mathbf{X}_2$ above and beyond the explanatory variables in $\mathbf{X}_1$.

The contributions of distinct sets of explanatory variables to the model are typically captured by breaking up the overall regression (or model) sum of squares into distinct components.

This is useful quite generally in linear models, but especially in ANOVA models where the response is modeled in terms of one or more class variables or factors. In such cases, the model sum of squares is decomposed into sums of squares for each of the distinct sets of dummy, or indicator, variables necessary to capture each of the factors in the model.

For example, the following model is appropriate for a randomized complete block design (RCBD)

$$y_{ij} = \mu + \beta_j + \alpha_i + e_{ij}$$

where y_{ij} is the response from the ith treatment in the jth block, and β_j and α_i are block and treatment effects, respectively. This model can also be written as

$$\mathbf{y} = \mu\mathbf{j}_n + \beta_1\mathbf{b}_1 + \cdots + \beta_b\mathbf{b}_b + \alpha_1\mathbf{t}_1 + \cdots + \alpha_a\mathbf{t}_a + \mathbf{e} \qquad (*)$$

In this context, the notation $SS(\alpha|\beta,\mu)$ denotes *the extra regression sum of squares due to fitting the α_is after fitting μ and the β_js* and is given by

$$SS(\alpha|\beta,\mu) = \mathbf{y}^T(\mathbf{P}_{C(\mathbf{X})} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}$$

where $\mathbf{X}_1 = (\mathbf{j}_n, \mathbf{b}_1, \ldots, \mathbf{b}_b)$ and $\mathbf{X} = (\mathbf{X}_1, \mathbf{t}_1, \ldots, \mathbf{t}_a)$.

- Sums of squares like this one that can be computed by fitting successively more complex models and taking the difference in regression/model sum of squares at each step are called *sequential sums of squares.*
- They represent the contribution of each successive group of explanatory variables above and beyond those explanatory variables already in the model.

Any model that can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{X}_3\boldsymbol{\beta}_3 + \cdots + \mathbf{e}$$

has a sequential sum of squares decomposition. That is, the regression or model sum of squares $SS_{\text{Model}} = \mathbf{y}^T\mathbf{P}_{C(\mathbf{X})}\mathbf{y} = \|\mathbf{P}_{C(\mathbf{X})}\mathbf{y}\|^2$ can always be decomposed as follows:

$$\begin{aligned} SS_{\text{Model}} &= \|\mathbf{P}_{C(\mathbf{X})}\mathbf{y}\|^2 \\ &= \|\mathbf{P}_{C(\mathbf{X}_1)}\mathbf{y}\|^2 + \|(\mathbf{P}_{C(\mathbf{X}_1,\mathbf{X}_2)} - \mathbf{P}_{C(\mathbf{X}_1)})\mathbf{y}\|^2 \\ &\qquad + \|(\mathbf{P}_{C(\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3)} - \mathbf{P}_{C(\mathbf{X}_1,\mathbf{X}_2)})\mathbf{y}\|^2 + \cdots \\ \text{or} \quad SS_{\text{Model}} &= SS(\boldsymbol{\beta}_1) + SS(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) + SS(\boldsymbol{\beta}_3|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) + \cdots \end{aligned}$$

- Note that by construction, the projections and squared lengths of projections in such a decomposition are independent because the spaces onto which we are projecting are mutually orthogonal.
- Such a decomposition can be extended to any number of terms.

Consider the RCBD model (*). This model can be written as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{X}_3\boldsymbol{\beta}_3 + \mathbf{e}$$

where

$$\mathbf{X}_1 = \mathbf{j}_N, \quad \mathbf{X}_2 = (\mathbf{b}_1, \ldots, \mathbf{b}_b), \quad \mathbf{X}_3 = (\mathbf{t}_1, \ldots, \mathbf{t}_a)$$

and $\boldsymbol{\beta}_1 = \mu$, $\boldsymbol{\beta}_2 = (\beta_1, \ldots, \beta_b)^T$, and $\boldsymbol{\beta}_3 = (\alpha_1, \ldots, \alpha_a)^T$.

The sequential break-down of the model sum of squares here is

$$SS_{\text{Model}} = SS(\mu) + SS(\beta|\mu) + SS(\alpha|\beta, \mu) \qquad (**)$$

Consider the null hypothesis $H_0 : \alpha_1 = \cdots = \alpha_a = 0$. The null model corresponding to this hypothesis is $y_{ij} = \mu + \beta_j + e_{ij}$.

Fitting just the null model we have

$$SS_{\text{Model0}} = SS(\mu) + SS(\beta|\mu).$$

Note that $\text{SSE} = SS_T - SS_{\text{Model}}$, where $SS_T = \|\mathbf{y}\|^2$ is the total (uncorrected) sum of squares. Therefore, the difference in error sums of squares between the null model and the maintained model is

$$\begin{aligned} \text{SSE}_0 - \text{SSE} &= (SS_T - SS_{\text{Model0}}) - (SS_T - SS_{\text{Model}}) \\ &= SS_{\text{Model}} - SS_{\text{Model0}} = SS(\alpha|\beta, \mu). \end{aligned}$$

- That is, $SS(\alpha|\beta, \mu)$ is an appropriate sum of squares for the numerator of the $F-$ test for testing $y_{ij} = \mu + \beta_j + e_{ij}$ versus $y_{ij} = \mu + \beta_j + \alpha_i + e_{ij}$.
- Similarly, we can test $y_{ij} = \mu + e_{ij}$ versus $y_{ij} = \mu + \beta_j + e_{ij}$ using $SS(\beta|\mu)$ as the numerator sum of squares.
- Finally, we can test $y_{ij} = \mu + e_{ij}$ versus $y_{ij} = \mu + \beta_j + \alpha_i + e_{ij}$ using $SS(\alpha, \beta|\mu) \equiv SS(\alpha|\beta, \mu) + SS(\beta|\mu)$.

This last test is a test for significance of the entire model (other than the constant term), or the overall regression test we have already encountered.

The sequential sums of squares used in decomposition (**) are known as **Type I sums of squares.** This terminology is from SAS, but it has taken hold more generally.

Notice that with Type I (sequential) sums of squares we can decompose SS_{Model} either as

$$SS_{\text{Model}} = SS(\mu) + SS(\beta|\mu) + SS(\alpha|\beta,\mu)$$

or as

$$SS_{\text{Model}} = SS(\mu) + SS(\alpha|\mu) + SS(\beta|\alpha,\mu)$$

- That is, if we happen to add the block effects to the model first, then the appropriate test statistic is based on $SS(\alpha|\beta,\mu)$. If we happen to add the treatment effects first then the test is based on $SS(\alpha|\mu)$.
- In addition, although $SS(\alpha|\beta,\mu) = SS(\alpha|\mu)$ for <u>some</u> models (e.g., balanced ANOVA models), such a result is not true, in general.
  - Clearly, there's something dissatisfying about obtaining different tests based on the order of the terms in the model.
  - There's an asymmetry in the way that the α's and β's are treated in the Type I SSs.
- In contrast we might choose to always test the α's based on $SS(\alpha|\beta,\mu)$ and to test the β's based on $SS(\beta|\alpha,\mu)$.
- Sums of squares like $SS(\alpha|\beta,\mu)$ and $SS(\beta|\alpha,\mu)$ are called **Type II sums of squares** in SAS.

- Type II SS's correct the order-dependence of Type I SS's. In the RCBD mode, for example, Type II SS's treat main effects for blocks ($SS(\beta|\alpha,\mu)$) and treatments ($SS(\alpha|\beta,\mu)$) in a symmetric way.

- For a full definition of Type II SS's we need to understand what a hierarchical model is.

Hierarchical models: Hierarchical models are models in which the inclusion of any interaction effect necessarily implies the inclusion of all lower-level interactions and main effects involving the factors of the original interaction.

- E.g., in the context of a two-way layout, the usual model has main effects α_i and β_j for the levels of each of the two factors A and B, and interaction effects $(\alpha\beta)_{ij}$ corresponding to $A * B$. However, there is nothing to prevent us from considering simpler models.

- The model
$$y_{ijk} = \mu + \alpha_i + (\alpha\beta)_{ij} + e_{ijk}$$
is <u>not</u> a hierachical model, because we have included an $A * B$ interaction, but no main effect for factor B. In a hierarchical model, the inclusion of $(\alpha\beta)_{ij}$ requires the inclusion of both α_i and β_j.

- Similarly, suppose we have a three-way layout. The full hierarchical model is
$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijkl}$$
Here, γ_k is an effect for the k^{th} level of factor C, $(\alpha\gamma)_{ik}$ and $(\beta\gamma)_{jk}$ are two way interactions for $A * C$ and $B * C$, and $(\alpha\beta\gamma)_{ijk}$ is the three-way interaction $A * B * C$. Two examples of non-hierarchical three-factor models are
$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijkl}$$
and $$y_{ijkl} = \mu + \alpha_i + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\beta\gamma)_{ijk} + e_{ijkl}$$

- I believe (and most statisticians would agree) that, in general, it is best to restrict attention to hierarchical models unless there is a compelling reason that in a particular application the omission of a lower-order term makes sense (e.g., is suggested by some theory or known fact from the context of the problem).

  - This principle is similar to the notion that in a polynomial regression model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_q x_i^q + e_i$ one should not consider any model in which a term $\beta_k x_i^k$ is included, but where any of the terms $\beta_0, \beta_1 x_i, \ldots, \beta_{k-1} x_i^{k-1}$ are excluded.

Type II SS's computes the SS for a factor U, say, as the reduction in SS's obtained by adding a term for factor U to the model that is the largest hierarchical model that does not contain U.

- E.g., in the two-way layout, the Type II SS's are

$$SS_A = SS(\alpha|\beta,\mu), \quad SS_B = SS(\beta|\alpha,\mu), \quad SS_{AB} = SS((\alpha\beta)|\alpha,\beta,\mu)$$

  - Notice that there is no longer an order effect. Factor B is adjusted for A and factor A is adjusted for B.

- Another example: in the three-way layout, the Type II SS's are

$$SS_A = SS(\alpha|\mu,\beta,\gamma,(\beta\gamma)), \quad SS_B = SS(\beta|\mu,\alpha,\gamma,(\alpha\gamma)), \quad SS_C = SS(\gamma|\mu,\alpha,\beta,(\alpha\beta))$$
$$SS_{AB} = SS((\alpha\beta)|\mu,\alpha,\beta,\gamma,(\alpha\gamma),(\beta\gamma)) \quad SS_{AC} = SS((\alpha\gamma)|\mu,\alpha,\beta,\gamma,(\alpha\beta),(\beta\gamma))$$
$$SS_{BC} = SS((\beta\gamma)|\mu,\alpha,\beta,\gamma,(\alpha\beta),(\alpha\gamma)) \quad SS_{ABC} = SS((\alpha\beta\gamma)|\mu,\alpha,\beta,\gamma,(\alpha\beta),(\alpha\gamma),(\beta\gamma))$$

Type III SS's:

When using ANOVA models for the analysis of experimental data, the scientific focus is often on comparing mean responses across different levels of the treatment factors (e.g., low, medium high doses of a drug; presence vs. abscence of fertilizer,; etc.).

In an expperiment with only one treatment factor, the levels of that factor are the treatments, and means across these treatments are typically of interest. However, in a factorial design, the treatments are the experimental conditions corresponding to combinations of the levels of two or more factors.

- E.g., in a two-way layout with two factors, A and B, with a and b levels, respectively, we may be interested in comparing means across the treatments, where the treatment means correspond to the cells of the following table

Levels of Factor A	Levels of Factor B: 1	2	$\cdots$	b	
1	μ_{11}	μ_{12}	$\cdots$	μ_{1b}	$\bar{\mu}_{1\cdot}$
2	μ_{21}	μ_{22}	$\cdots$	μ_{2b}	$\bar{\mu}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a	μ_{a1}	μ_{a2}	$\cdots$	μ_{ab}	$\bar{\mu}_{a\cdot}$
	$\bar{\mu}_{\cdot 1}$	$\bar{\mu}_{\cdot 2}$	$\cdots$	$\bar{\mu}_{\cdot b}$	

Here, μ_{ij} is the population mean for the i, jth treatment, or the treatment corresponding to the ith level of factor A combinaed with the jth level of factor B. These μ_{ij} are the parameters of the full-rank cell means model:

$$y_{ijk} = \mu_{ij} + e_{ijk}, \quad i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n_{ij}$$

or, in terms of the overparameterized effects model,

$$y_{ijk} = \underbrace{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}}_{=\mu_{ij}} + e_{ijk}$$

While it is often of interest to compare treatment means (e.g., to test $H_0 : \mu_{ij} = \mu_{i'j'}$), it also often of interest to compare the mean response across levels of factor A *marginally*; that is, after averaging across factor B (or vice versa).

We define the marginal mean at the ith level of factor A to be $\bar{\mu}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}$ or, in terms of the effects model,

$$\bar{\mu}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}(\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}) = \mu + \alpha_i + \bar{\beta}_{\cdot} + \bar{(\alpha\beta)}_{i\cdot}.$$

- Marginal means for each of the level of factor B are defined similarly.

If we are interested in making comparisons among the marginal means, it would be nice if our sum of squares for factor A, and for factor B led to an F test, $F = \frac{SS_A/\mathrm{df}_A}{MSE}$, which tested a simple hypothesis of interest like $H_0 : \bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \cdots = \bar{\mu}_{a\cdot}$.

- It turns out that both the Type I and Type II approach to calculating SS_A do not test such a hypothesis. Type III SS's, also know as **marginal sums of squares**, do.

It can be shown that, in terms of the marginal means, the hypotheses tested by Type I and Type II SS's are difficult to interpret, and not at all the sorts of hypotheses that one would typically be interested in if the focus of the analysis was to compare treatment means (which it usually is).

- For example, in the two-way layout, the hypotheses tested by $F_A = \frac{SS_A/\mathrm{df}_A}{MS_E}$ for Type I and II versions of SS_A are:

$$\text{Type I:} \qquad H_0 : \sum_{j=1}^{b} \frac{n_{1j}\mu_{1j}}{n_{1\cdot}} = \cdots = \sum_{j=1}^{b} \frac{n_{aj}\mu_{aj}}{n_{a\cdot}}$$

$$\text{Type II:} \qquad H_0 : \sum_{j=1}^{b} n_{1j}\mu_{1j} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{n_{1j}n_{ij}\mu_{ij}}{n_{\cdot j}}, \ldots, \sum_{j=1}^{b} n_{aj}\mu_{aj} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{n_{aj}n_{ij}\mu_{ij}}{n_{\cdot j}}$$

- These hypotheses (especially those from Type II) are strange. They correspond to testing whether certain weighted marginal averages of the treatment means are equal. Testing such hypotheses is seldom of interest. If one is interested in comparing means across the levels of factor A, these SS's are definitely not what one would want to use.

Type III SS's are designed to always test simple hypotheses on (unweighted) marginal population means. In particular, for the Type III version of SS_A, F_A tests the hypothesis

$$\text{Type III:} \qquad H_0 : \bar{\mu}_{1\cdot} = \cdots = \bar{\mu}_{a\cdot}$$

Similarly, the Type III version of SS_B leads to a test of

$$\text{Type III:} \qquad H_0 : \bar{\mu}_{\cdot 1} = \cdots = \bar{\mu}_{\cdot b}$$

- All three types of SS's lead to the same (reasonable and appropriate) hypothesis for $F_{AB} = MS_{AB}/MS_E$. Namely,

$$H_0 : (\mu_{ij} - \mu_{ij'}) - (\mu_{i'j} - \mu_{i'j'}) = 0, \quad \text{for all } i, i', j, j'$$

Type III SS's also have an interpretation in terms of reduction in SS's. For the two way layout model **with sum-to-zero restrictions on the parameters** the Type III SS's are:

$$SS_A = SS(\alpha|\mu, \beta, (\alpha\beta)), \quad SS_B = SS(\beta|\mu, \alpha, (\alpha\beta)), \quad SS_{AB} = SS((\alpha\beta)|\mu, \alpha, \beta).$$

- Note that this interpretation only applies to the sum-to-zero restricted version of the two-way layout model. For other restrictions, the interpretation would be different. A much better way to understand Type III SS's is in terms of the hypotheses tested on the marginal means, as described above.

Type IV SS's:

The fourth type of SS's is useful when there are certain treatment combinations for which $n_{ij} = 0$.

- My recommendation is to avoid the use of Type IV SSs. If factorial designs are encountered with missing treatments, I suggest instead the use of a one-way anova model, treating the treatments with data as levels of a single treatment factor. Interactions and main effects can be investigated by testing hypotheses on the treatment means.

- If you're really interested, you can refer to Milliken & Johnson (1992) or Littell, Freund, & Spector (*SAS System for Linear Models, Third Edition, 1991*) for information on Type IV SS's.

Relationships Among the Types and Recommendations:

In certain situations, the following equalities among the types of SS's hold:

$$\begin{aligned} I = II = III = IV & \qquad \text{for balanced data} \\ II = III = IV & \qquad \text{for no-interaction models} \\ III = IV & \qquad \text{for all-cells-filled data} \end{aligned}$$

If one is interested in model-building (finding a parsimonious well-fitting model for the data) then

i. use Type I for choosing between models of sequentially increasing complexity; and

ii. use Type II for choosing between hierarchical models.

If one is interested in testing hypotheses that compare means across the levels of experimentally controlled factors

iii. use Type III.

- Note that Type I SS's are the only type of sum of squares that, in general, lead decompose the total sum of squares. E.g., in the two-way anova model, they are the only type of SS that guarantee that $SS_T = SS_A + SS_B + SS_{AB} + SS_E$ holds, in general.

- However, all four types yield sums of squares that are independent of SS_E and all lead to valid F tests (just of different hypotheses).

- Independence of these SSs from the SSE is guaranteed because all four types of SSs are squared lengths of projections onto some subspace of $C(\mathbf{X})$, whereas SSE is the squared length of a projection onto $C(\mathbf{X})^{\perp}$.