Statistical Regression Analysis

September 5, 2013

Chapter 1

Probability Distributions, Estimation, and Testing

1.1 Introduction

Here we introduce probability distributions, and basic estimation/testing methods. **Random variables** are outcomes of an experiment or data-generating process, where the outcome is not known in advance, although the set of possible outcomes is. Random variables can be **discrete** or **continuous**. Discrete random variables can take on only a finite or countably infinite set of possible outcomes. Continuous random variables can take on values along a continuum. In many cases, variables of one type may be treated as or reported as the other type. In general, we will use upper-case letters (such as Y) to represent random variables, and lower-case letters (such as y) to represent specific outcomes. Not all (particularly applied statistics) books follow this convention.

1.1.1 Discrete Random Variables/Probability Distributions

In many applications, the result of the data-generating process is the count of a number of events of some sort. In some cases, a certain number of trials are conducted, and the outcome of each trial is observed as a "Success" or "Failure" (binary outcomes). In these cases, the number of trials ending in Success is observed. Alternatively, a series of trials may be conducted until a pre-selected number of Successes are observed. In other settings, the number of events of interest may be counted in a fixed amount of time or space, without actually breaking the domain into a set of distinct trials.

For discrete random variables, we will use $p(y)$ to represent the probability that the random variable Y takes on the value y. We require that all such probabilities be bounded between 0 and 1 (inclusive), and that they sum to 1:

$$P\left\{Y=y\right\}=p(y) \qquad 0 \leq p(y) \leq 1 \qquad \sum_y p(y)=1$$

The **cumulative distribution function** is the probability that a random variable takes on a value less than or equal to a specific value y^*. It is an increasing function that begins at 0 and increases to 1, and we will denote it as $F\left(y^*\right)$. For discrete random variables it is a step function, taking a step at each point where $p(y) > 0$:

$$F\left(y^*\right) = P\left(Y \leq y^*\right) = \sum_{y \leq y^*} p(y)$$

The **mean** or **Expected Value** (μ) of a random variable is it's long-run average if the experiment was conducted repeatedly ad infinitum. The **variance** $\left(\sigma^2\right)$ is the average squared difference between the random variable and its mean, measuring the dispersion within the distribution. The **standard deviation** (σ) is the positive square root of the variance, and is in the same units as the data.

$$\mu_Y = E\left\{Y\right\} = \sum_y yp(y) \qquad \sigma_Y^2 = V\left\{Y\right\} = E\left\{\left(Y - \mu_Y\right)^2\right\} = \sum_y \left(y - \mu_Y\right)^2 p(y) \qquad \sigma_Y = +\sqrt{\sigma_Y^2}$$

Note that for any function of Y, the expected value and variance of the function is computed as follows:

$$E\left\{g(Y)\right\} = \sum_y g(y)p(y) = \mu_{g(Y)} \qquad V\left\{g(Y)\right\} = E\left\{\left(g(Y) - \mu_{g(Y)}\right)^2\right\} = \sum_y \left(g(y) - \mu_{g(Y)}\right)^2 p(y)$$

For any constants a and b, we have the mean and variance of the linear function $a + bY$:

$$E\left\{a + bY\right\} = \sum_y ap(y) + \sum_y byp(y) = a\sum_y p(y) + b\sum_y yp(y) = a(1) + bE\left\{Y\right\} = a + b\mu_Y$$

$$V\left\{a + bY\right\} = \sum_y \left((a + by) - (a + b\mu_Y)\right)^2 p(y) = b^2 \sum_y \left(y - \mu_Y\right)^2 p(y) = b^2\sigma_Y^2$$

A very useful result in mathematical statistics is the following:

$$\sigma_Y^2 = V\left\{Y\right\} = E\left\{\left(Y - \mu_Y\right)^2\right\} = E\left\{Y^2 - 2\mu_Y Y + \mu_2\right\} = E\left\{Y^2\right\} - 2\mu_Y E\left\{Y\right\} + \mu_Y^2 = E\left\{Y^2\right\} - \mu_Y^2$$

Thus, $E\left\{Y^2\right\} = \sigma_Y^2 + \mu_Y^2$. Also, from this result we obtain: $E\left\{Y(Y-1)\right\} = \sigma_Y^2 + \mu_Y^2 - \mu_Y$. From this, we can obtain $\sigma_Y^2 = E\left\{Y(Y-1)\right\} - \mu_Y^2 + \mu_Y$, which is useful for some discrete probability distributions.

Next, we consider several families of discrete probability distributions: the binomial, poisson, and negative binomial families.

Binomial Distribution

When an experiment consists of n independent trials, each of which can end in one of two outcomes: Success or Failure with constant probability of success, we refer to this as a **binomial experiment**. The random variable Y is the number of Successes in the n trials, and can take on the values $y = 0, 1, \ldots, n$. Note that in some settings, the "Success" can be a negative attribute. We denote the probability of success as π, which lies between 0 and 1. We use the notation: $Y \sim B\left(n, \pi\right)$. The probability distribution, mean and variance of Y depend on the sample size n and probability of success π.

$$p(y) = \binom{n}{y} \pi^y \left(1 - \pi\right)^{n-y} \qquad E\left\{Y\right\} = \mu_Y = n\pi \qquad V\left\{Y\right\} = \sigma_Y^2 = n\pi\left(1 - \pi\right)$$

where $\binom{n}{y} = \frac{n!}{y!(n-y)!}$. In practice, π will be unknown, and estimated from sample data. Note that to obtain the mean and variance, we have:

$$E\{Y\} = \mu_Y = \sum_{y=0}^{n} yp(y) = \sum_{y=0}^{n} y\frac{n!}{y!(n-y)!}\pi^y (1-\pi)^{n-y} = \sum_{y=1}^{n} y\frac{n!}{y!(n-y)!}\pi^y (1-\pi)^{n-y} =$$

$$= n\pi \sum_{y=1}^{n} \frac{(n-1)!}{(y-1)!(n-y)!}\pi^{y-1} (1-\pi)^{n-y} = n\pi \sum_{y^*=0}^{n-1} \binom{n-1}{y^*}\pi^{y^*} (1-\pi)^{n-1-y^*} = n\pi \sum_{y^*} p(y^*) = n\pi \qquad y^* = y-1$$

To obtain the variance, we use the result from above, $\sigma_Y^2 = E\{Y(Y-1)\} - \mu_Y^2 + \mu_Y$:

$$E\{Y(Y-1)\} = \sum_{y=0}^{n} y(y-1)p(y) = \sum_{y=0}^{n} y(y-1)\frac{n!}{y!(n-y)!}\pi^y (1-\pi)^{n-y} = \sum_{y=2}^{n} y(y-1)\frac{n!}{y!(n-y)!}\pi^y (1-\pi)^{n-y} =$$

$$= n(n-1)\pi^2 \sum_{y=2}^{n} \frac{(n-2)!}{(y-2)!(n-y)!}\pi^{y-2} (1-\pi)^{n-y} = n(n-1)\pi^2 \sum_{y^{**}=0}^{n-2} \binom{n-2}{y^{**}}\pi^{y^{**}} (1-\pi)^{n-2-y^{**}}$$

$$n(n-1)\pi^2 \sum_{y^{**}} p(y^{**}) = n(n-1)\pi^2 \qquad y^{**} = y-2$$

$$\Rightarrow \quad \sigma_Y^2 = n(n-1)\pi^2 - n^2\pi^2 + n\pi = n\pi - n\pi^2 = n\pi(1-\pi)$$

Poisson Distribution

In many applications, researchers observe the counts of a random process in some fixed amount of time or space. The random variable Y is a count that can take on any non-negative integer. One important aspect of the Poisson family is that the mean and variance are the same. This is one aspect that does not work for all applications. We use the notation: $Y \sim \text{Poi}(\lambda)$. The probability distribution, mean and variance of Y are:

$$p(y) = \frac{e^{-\lambda}\lambda^y}{y!} \qquad E\{Y\} = \mu_Y = \lambda \qquad V\{Y\} = \sigma_Y^2 = \lambda$$

Note that $\lambda > 0$. The Poisson arises by dividing the time/space into n infinitely small areas, each having either 0 or 1 Success, with Success probability $\pi = \lambda/n$. Then Y is the number of areas having a success.

$$p(y) = \frac{n!}{y!(n-y)!}\left(\frac{\lambda}{n}\right)^y \left(1-\frac{\lambda}{n}\right)^{n-y} = \frac{n(n-1)\cdots(n-y+1)}{y!}\left(\frac{\lambda}{n}\right)^y \left(1-\frac{\lambda}{n}\right)^{n-y} =$$

$$= \frac{1}{y!}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-y+1}{n}\right)\lambda^y \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-y}$$

The limit as n goes to ∞ is:

$$\lim_{n\to\infty} p(y) = \frac{1}{y!}(1)(1)\cdots(1)\lambda^y e^{-\lambda}(1) = p(y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

To obtain the mean of Y for the Poisson distribution, we have:

$$E\{Y\} = \mu_Y = \sum_{y=0}^{\infty} yp(y) = \sum_{y=0}^{\infty} y\frac{e^{-\lambda}\lambda^y}{y!} = \sum_{y=1}^{\infty} y\frac{e^{-\lambda}\lambda^y}{y!} = \sum_{y=1}^{\infty} \frac{e^{-\lambda}\lambda^y}{(y-1)!} =$$

$$\lambda\sum_{y=1}^{\infty}\frac{e^{-\lambda}\lambda^{(y-1)}}{(y-1)!}=\lambda\sum_{y^*=0}^{\infty}\frac{e^{-\lambda}\lambda^{y^*}}{y^*!}=\lambda\sum_{y^*}p\left(y^*\right)=\lambda$$

We use the same result as that for the binomial to obtain the variance for the Poisson distribution:

$$E\left\{Y(Y-1)\right\}=\sum_{y=0}^{\infty}y(y-1)p(y)=\sum_{y=0}^{\infty}y(y-1)\frac{e^{-\lambda}\lambda^{y}}{y!}=\sum_{y=2}^{\infty}y(y-1)\frac{e^{-\lambda}\lambda^{y}}{y!}=\lambda^2\sum_{y=2}^{\infty}\frac{e^{-\lambda}\lambda^{(y-2)}}{(y-2)!}=\lambda^2\sum_{y^{**}=0}^{\infty}\frac{e^{-\lambda}\lambda^{y^{**}}}{y^{**}!}=\lambda^2$$

$$\Rightarrow\quad \sigma_Y^2=\lambda^2-\lambda^2+\lambda=\lambda$$

Negative Binomial Distribution

The negative binomial distribution is used in two quite different contexts. The first is where a binomial type experiment is being conducted, except instead of having a fixed number of trials, the experiment is completed when the r^{th} success occurs. The random variable Y is the number of trials needed until the r^{th} success, and can take on any integer value greater than or equal to r. The probability distribution, its mean and variance are:

$$p(y)=\binom{y-1}{r-1}\pi^r\left(1-\pi\right)^y-r \qquad E\left\{Y\right\}=\mu_Y=\frac{r}{\pi} \qquad V\left\{Y\right\}=\sigma_Y^2=\frac{r\left(1-\pi\right)}{\pi^2}$$

A second use of the negative binomial distribution is as a model for count data. It arises from a mixture of Poisson models. In this setting it has 2 parameters and is more flexible than the Poisson (which has the variance equal to the mean), and can take on any non-negative integer value. In this form, the negative binomial distribution and its mean and variance can be written as (see e.g. Cameron and Trivedi, 2005 or Agresti, 2002):

$$p(y)=\frac{\Gamma\left(\alpha^{-1}+y\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y} \qquad E\left\{Y\right\}=\mu_Y=\mu \qquad V\left\{Y\right\}=\sigma_Y^2=\mu\left(1+\alpha\mu\right)$$

where $\Gamma(\cdot)$ is the gamma integral, and is a built-in function in virtually all computing packages/spreadsheets. If y is an integer, $\Gamma(y)=(y-1)!$.

1.1.2 Continuous Random Variables/Probability Distributions

Random variables that can take on any value along a continuum are continuous. Here, we consider the normal, gamma, t, and F families. Special cases of the gamma family include the exponential and chi-squared distributions. Continuous distributions are density functions, as opposed to probability mass functions. Their density is always non-negative, and integrates to 1. We will use the notation $f(y)$ for density functions. The mean and variance for continuous distributions are obtained in a similar manner as discrete distributions, with integration replacing summation.

$$E\left\{Y\right\}=\mu_Y=\int_{-\infty}^{\infty}yf(y)dy \qquad V\left\{Y\right\}=\sigma_Y^2=\int_{-\infty}^{\infty}\left(y-\mu_Y\right)^2f(y)dy$$

In general, for any function $g(Y)$, we have:

$$E\left\{g(Y)\right\}=\int_{-\infty}^{\infty}g(y)f(y)dy=\mu_{g(Y)} \qquad V\left\{g(Y)\right\}=E\left\{\left(g(Y)-\mu_{g(Y)}\right)^2\right\}=\int_{-\infty}^{\infty}\left(g(y)-\mu_{g(Y)}\right)^2f(y)dy$$

Normal Distribution

The normal distributions, also known as the Gaussian distributions, are a family of symmetric mound-shaped distributions. The distribution has 2 parameters: the mean μ and the variance σ^2, although often it is indexed by its standard deviation σ. We use the notation $Y \sim N\left(\mu, \sigma^2\right)$. The probability density function, the mean and variance are:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \qquad E\{Y\} = \mu_Y = \mu \qquad V\{Y\} = \sigma_Y^2 = \sigma^2$$

The mean μ defines the center (median and mode) of the distribution, and the standard deviation σ is a measure of the spread ($\mu - \sigma$ and $\mu + \sigma$ are the inflection points). Despite the differences in location and spread of the different distributions in the normal family, probabilities with respect to standard deviations from the mean are the same for all normal distributions. For $-\infty < z_1 < z_2 < \infty$, we have:

$$P\left(\mu + z_1\sigma \leq Y \leq \mu + z_2\sigma\right) = \int_{\mu+z_1\sigma}^{\mu+z_2\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy = \int_{z1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \Phi(z_2) - \Phi(z_1)$$

Where Z is **standard normal**, a normal distribution with mean 0, and variance (standard deviation) 1. Here $\Phi(z^*)$ is the cumulative distribution function of the standard normal distribution, up to the point z^*:

$$\Phi(z^*) = \int_{-\infty}^{z^*} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

These probabilities and critical values can be obtained directly or indirectly from standard tables, statistical software, or spreadsheets. Note that:

$$Y \sim N\left(\mu, \sigma^2\right) \qquad \Rightarrow \qquad Z = \frac{Y - \mu}{\sigma} \sim N(0,1)$$

This makes it possible to use the standard normal table for any normal distribution. Plots of three normal distributions are given in Figure 1.1.

Gamma Distribution

The gamma family of distributions are used to model non-negative random variables that are often right-skewed. There are two widely used parameterizations. The first given here is in terms of *shape* and *scale* parameters:

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} \quad y \geq 0, \alpha > 0, \beta > 0 \qquad E\{Y\} = \mu_Y = \alpha\beta \qquad V\{Y\} = \sigma_Y^2 = \alpha\beta^2$$

Here, $\Gamma(\alpha)$ is the gamma function $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$ and is built-in to virtually all statistical packages and spreadsheets. It also has two simple properties:

$$\alpha > 1: \quad \Gamma(\alpha) = (\alpha - 1)\,\Gamma(\alpha - 1) \qquad \Gamma(1/2) = \sqrt{\pi}$$

Thus, if α is an integer, $\Gamma(\alpha) = (\alpha - 1)!$. The second given here is in terms of *shape* and *rate* parameters:

$$f(y) = \frac{\theta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-y\theta} \quad y \geq 0, \alpha > 0, \theta > 0 \qquad E\{Y\} = \mu_Y = \frac{\alpha}{\theta} \qquad V\{Y\} = \sigma_Y^2 = \frac{\alpha}{\theta^2}$$

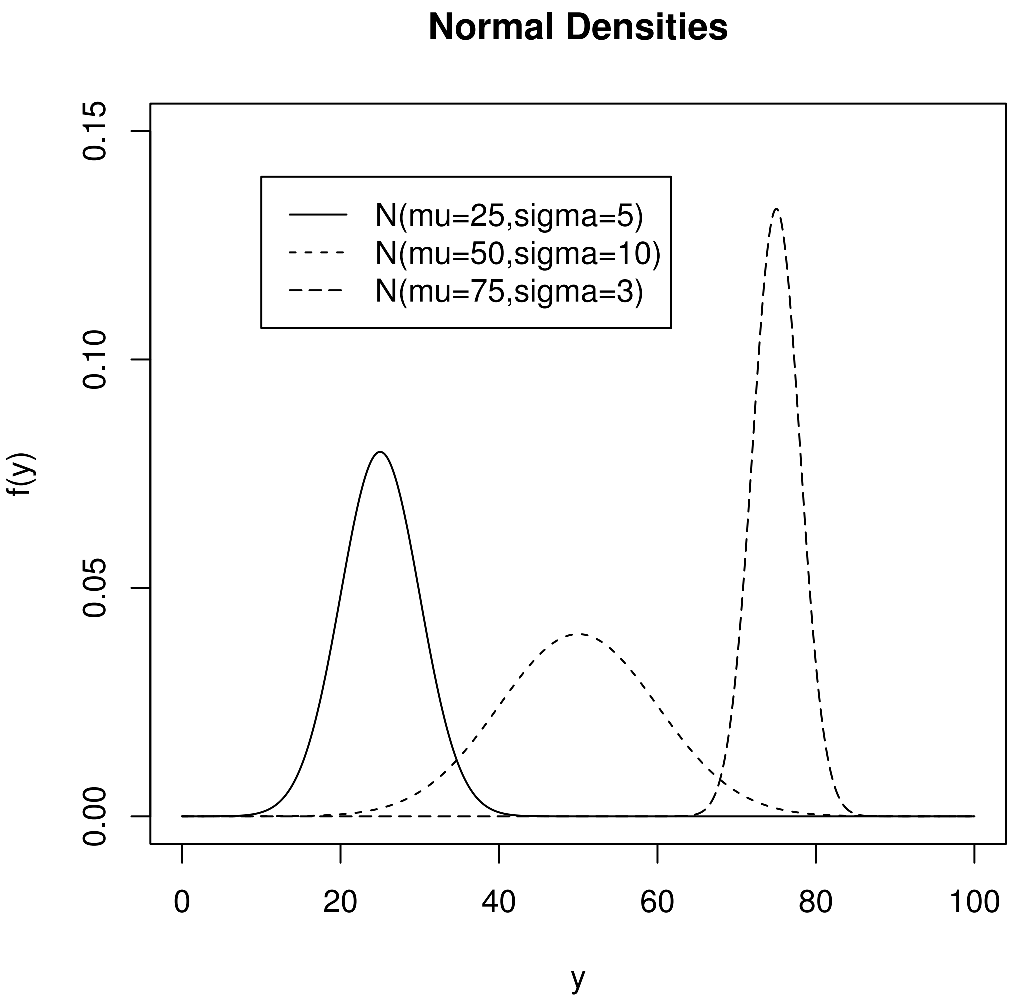

Figure 1.1: Three Normal Densities

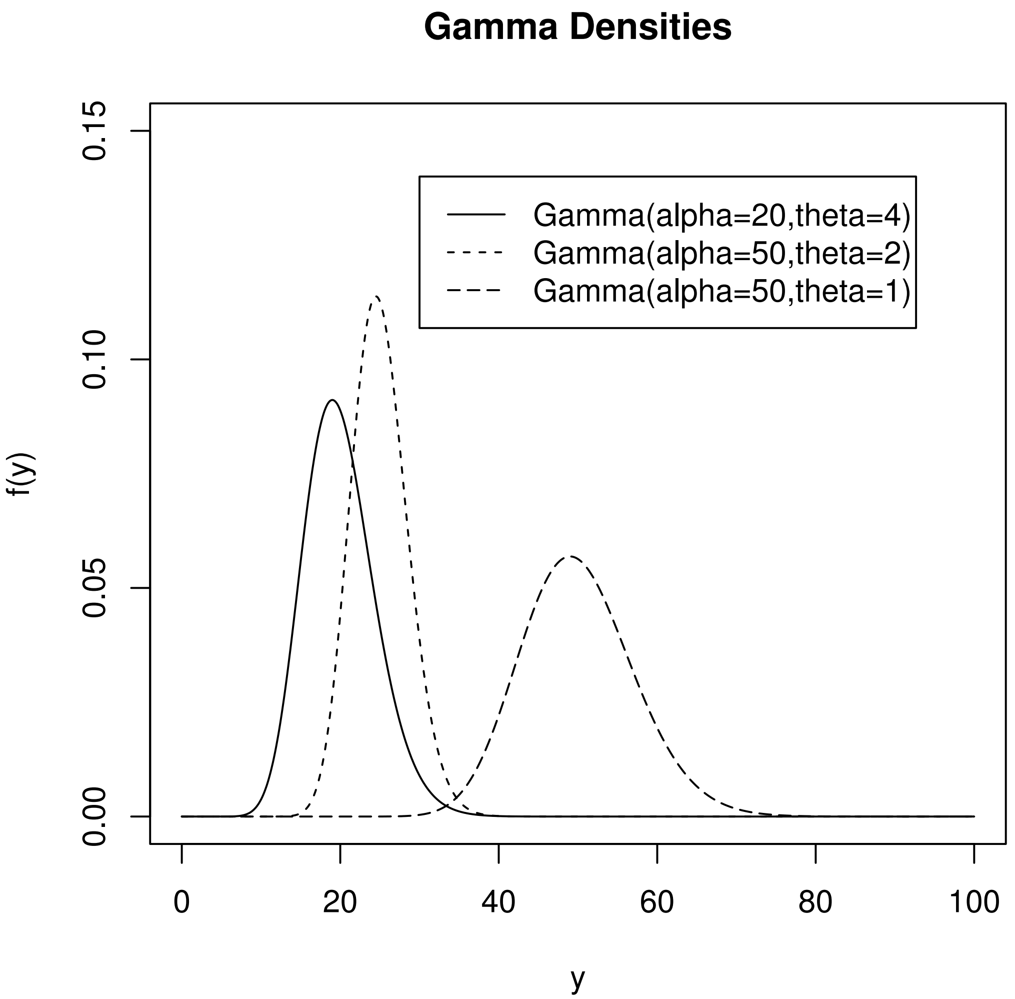

Figure 1.2: Three Gamma Densities

Note that different software packages use different parameterizations in generating samples and giving tail-areas and critical values. For instance, EXCEL uses the first parameterization and R uses the second. Figure reffig:gamma1 displays three gamma densities of various shapes.

Two special cases are the exponential family, where $\alpha = 1$ and the chi-squared family, with $\alpha = \nu/2$ and $\beta = 2$ for integer valued ν. For the exponential family, based on the second parameterization:

$$f(y) = \theta e^{-y\theta} \qquad E\{Y\} = \mu_Y = \frac{1}{\theta} \qquad V\{Y\} = \sigma_Y^2 = \frac{1}{\theta^2}$$

Probabilities for the exponential distribution are trivial to obtain as $F(y^*) = 1 - e^{-y^*\theta}$. Figure 1.3 gives three exponential distributions.

For the chi-squared family, based on the first parameterization:

$$f(y) = \left(\frac{2}{\nu}\right)^{\alpha} ye^{-2y/\nu} \qquad E\{Y\} = \mu_Y = \nu \qquad V\{Y\} = \sigma_Y^2 = 2\nu$$

Here, ν is the **degrees of freedom** and we denote the distribution as: $Y \sim \chi^2(\nu)$. Upper and lower critical values of the chi-squared distribution are available in tabular form, and in statistical packages and spreadsheets. Probabilities can be obtained with statistical packages and spreadsheets. Figure 1.4 gives three Chi-squared distributions.

1.2 Linear Functions of Multiple Random Variables

Suppose we simultaneously observe two random variables: X and Y. Their joint probability distribution can be discrete, continuous, or mixed (one discrete, the other continuous). We consider the **joint distribution** and the **marginal distributions** for the discrete case:

$$p(x,y) = P\{X = x, Y = y\} \qquad p_X(x) = P\{X = x\} = \sum_y p(x,y) \qquad p_Y(y) = P\{Y = y\} = \sum_x p(x,y)$$

For the continuous case, we have the joint and marginal densities and cumulative distribution function:

$$\text{Joint Density when } X = x, Y = y : f(x,y) \qquad f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

$$F(a,b) = P\{X \leq a, Y \leq b\} = \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)dxdy$$

Note that:

$$\text{Discrete: } \sum_x \sum_y p(x,y) = \sum_x p_x(x) = \sum_y p_Y(y) = 1$$

$$\text{Continuous: } \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_{-\infty}^{\infty} f_X(x)dx = \int_{-\infty}^{\infty} f_Y(y)dy = 1$$

The **conditional probability** that $X = x$, given $Y = y$ (or $Y = y$ given $X = x$)is denoted as:

$$p(x|y) = P\{X = x|Y = y\} = \frac{p(x,y)}{p_Y(y)} \qquad p(y|x) = P(Y = y|X = x) = \frac{p(x,y)}{p_X(x)}$$

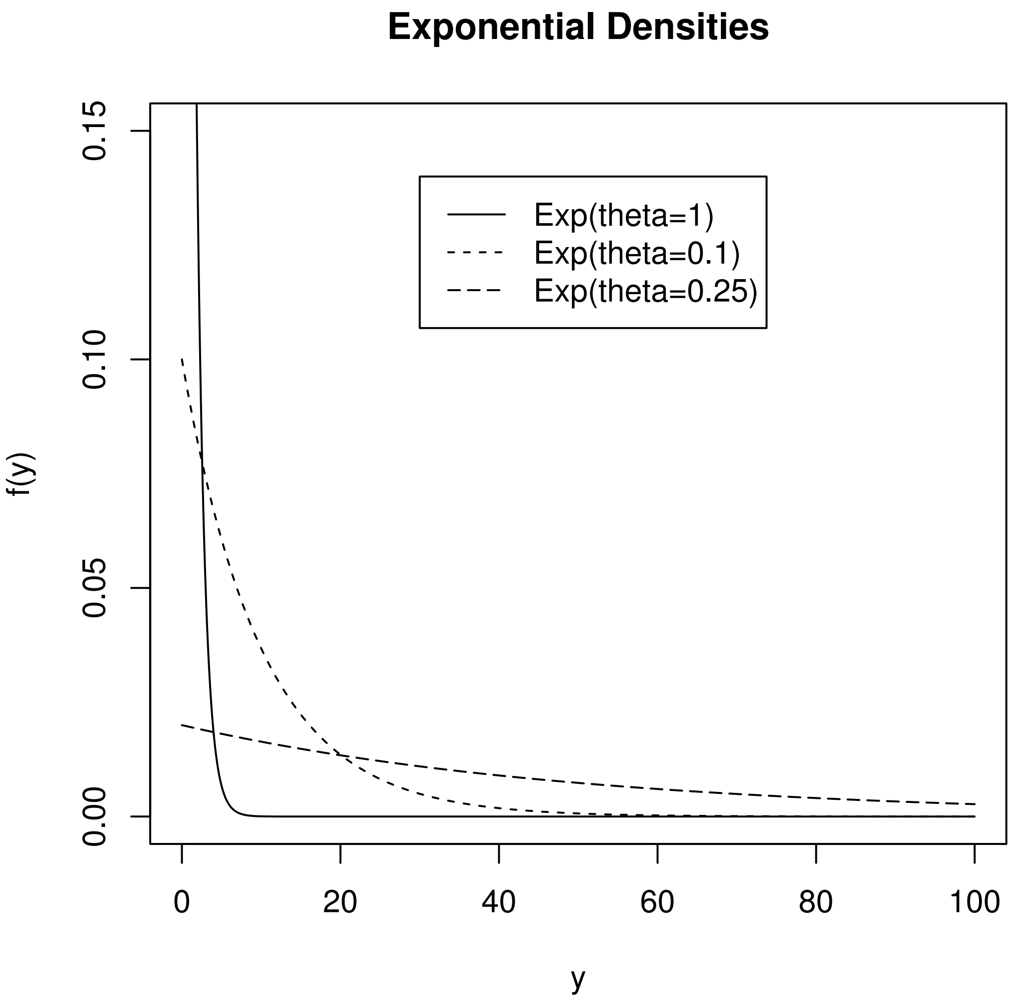

Figure 1.3: Three Exponential Densities

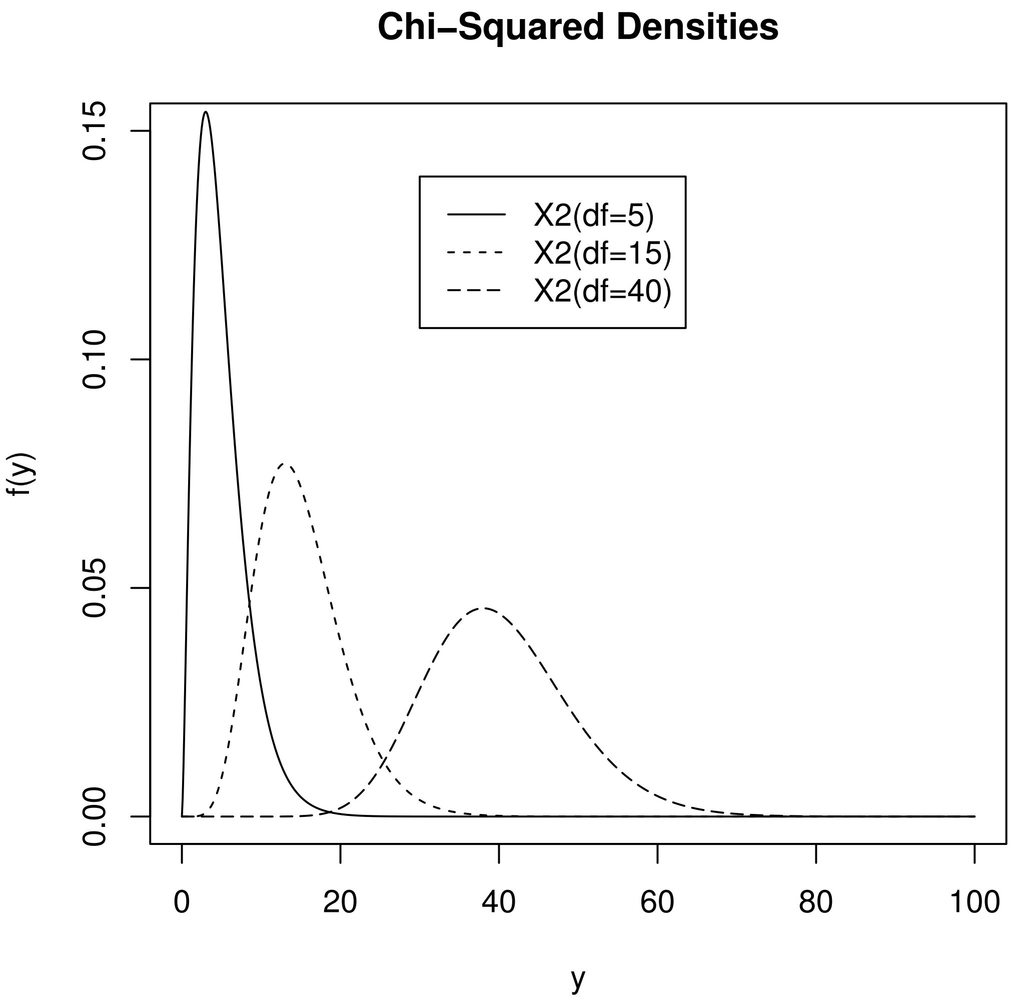

Figure 1.4: Three Chi-Squared Densities

assuming $p_Y(y) > 0$. This simply implies the probability that both occur divided by the probability that $Y = y$. X and Y are said to be independent if $p(x|y) = p(x)$ for all y, and that $p(y|x) = p(y)$ for all x. The conditional densities for continuous random variables are similarly defined based on the joint and marginal densities:

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} \quad f_Y(y) > 0 \qquad\qquad f(y|x) = \frac{f(x,y)}{f_X(x)} \quad f_X(x) > 0$$

The **conditional mean** and **variance** are the mean and variance of the conditional distribution (density), and is often a function of the conditioning variable:

$$\text{Discrete: } E\{Y|X = x\} = \mu_{Y|x} = \sum_y yp(y|x) \qquad \text{Continuous: } E\{Y|X = x\} = \mu_{Y|x} = \int_{-\infty}^{\infty} yf(y|x)dy$$

$$\text{Discrete: } V\{Y|X = x\} = \sigma^2_{Y|x} = \sum_y \left(y - \mu_{Y|x}\right)^2 p(y|x)$$

$$\text{Continuous: } V\{Y|X = x\} = \sigma^2_{Y|x} = \int_{-\infty}^{\infty} \left(y - \mu_{Y|x}\right)^2 f(y|x)dy$$

Next we consider the **variance of the conditional mean** and the **mean of the conditional variance** for the continuous case (with integration being replaced by summation for the discrete case):

$$V_X\{E\{Y|x\}\} = \int_{-\infty}^{\infty} \left(\mu_{Y|x} - \mu_Y\right)^2 f_X(x)dx$$

$$E_X\{V\{Y|x\}\} = \int_{-\infty}^{\infty} \sigma^2_{Y|x} f_X(x)dx$$

Note that we can partition the variance of Y into the sum of the variance of the conditional mean and mean of the conditional variance:

$$V\{Y\} = V_X\{E\{Y|x\}\} + E_X\{V\{Y|x\}\}$$

The **covariance** σ_{XY} between X and Y is the average product of deviations from the mean for X and Y. For the discrete case, we have:

$$\sigma_{XY} = E\{(X - \mu_X)(Y - \mu_Y)\} = \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x,y) = \sum_x \sum_y (xy - x\mu_Y - \mu_X y + \mu_X\mu_Y)p(x,y) =$$

$$= \sum_x \sum_y xyp(x,y) - \mu_Y \sum_x \sum_y xp(x,y) - \mu_Y \sum_x \sum_y xp(x,y) - \mu_X\mu_Y = E\{XY\} - \mu_X\mu_Y$$

For the continuous case, replace summation with integration. If X and Y are independent, $\sigma_{XY} = 0$, but the converse is not typically the case. Covariances can be either positive or negative, depending on the association (if any) between X and Y. The covariance is unbounded, and depends on the scales of measurement for X and Y. The **correlation** ρ_{XY} is a measure that is unit-less, is not affected by linear transformations of X and Y, and bounded between -1 and 1:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

where σ_X and σ_Y are the standard deviations of the maginal distributions of X and Y, respectively.

The mean and variance of any **linear function** of X and Y: $W = aX + bY$ for fixed constants a and b are for the discrete case:

$$E\{W\} = E\{aX + bY\} = \sum_x \sum_y (ax + by)p(x,y) = a\sum_x xp_X(x) + b\sum_y yp_Y(y) = a\mu_X + b\mu_Y$$

$$V\{W\} = V\{aX + bY\} = \sum_x \sum_y \left[(ax + by) - (a\mu_x + b\mu_y)\right]^2 p(x, y) =$$

$$\sum_x \sum_y \left[a^2 (x - \mu_X)^2 + b^2 (y - \mu_Y)^2 + 2ab (x - \mu_X)(y - \mu_Y)\right] p(x, y) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}$$

For the continuous case, replace summation with integration.

In general, if $Y_1, \ldots, Y_n$ are sequence of random variables, and $a_1, \ldots, a_n$ are a sequence of constants:

$$E\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i E\{Y_i\} = \sum_{i=1}^{n} a_i \mu_i$$

$$V\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j \sigma_{ij}$$

Where μ_i is the mean of Y_i, σ_i^2 is the variance of Y_i, and σ_{ij} is the covariance of Y_i and Y_j.

1.3 Functions of Normal Random Variables

First, note that if $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(1)$. Many software packages present Z-tests as (Wald) χ^2-tests. See the section on testing below.

Suppose $Y_1, ..., Y_n$ are independent with $Y_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$. Then the sample mean and sample variance are computed as:

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \qquad S^2 = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}$$

In this case, we obtain the following sampling distributions for the mean and a function of the variance:

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2(n-1) \qquad \overline{Y}, \frac{(n-1)S^2}{\sigma^2} \text{ are independent.}$$

Note that in general, if $Y_1, ..., Y_n$ are normally distributed (and not necessarily with the same mean and/or variance), any linear function of them will be normally distributed, with mean and variance given in the previous section.

Two distributions associated with the normal and chi-squared distributions are **Student's** t and F. Student's t-distribution is similar to the standard normal ($N(0,1)$), except that is indexed by its degrees of freedom and that is has heavier tails than the standard normal. As its degrees of freedom approach infinity, its distribution converges to the standard normal. Let $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$, where Z and W are independent. Then, we get:

$$Y \sim N(\mu, \sigma^2) \quad \Rightarrow \quad Z = \frac{Y - \mu}{\sigma} \sim N(0,1) \qquad T = \frac{Z}{\sqrt{W/\nu}} \sim t(\nu)$$

where the probability density, mean, and variance for Student's t-distribution are:

$$f(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad E\{Y\} = \mu_Y = 0 \qquad V\{Y\} = E\left\{(Y - \mu_Y)^2\right\} = \frac{\nu}{\nu - 2} \quad \nu > 2$$

and we use the notation $Y \sim t(\nu)$. Now consider the sample mean and variance, and the fact they are independent:

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad Z = \frac{\overline{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n}\frac{\overline{Y}-\mu}{\sigma} \sim N(0,1)$$

$$W = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2(n-1) \quad \Rightarrow \quad \sqrt{\frac{W}{\nu}} = \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = \frac{S}{\sigma}$$

$$\Rightarrow \quad T = \frac{Z}{\sqrt{W/\nu}} = \frac{\sqrt{n}\frac{\overline{Y}-\mu}{\sigma}}{\frac{S}{\sigma}} = \sqrt{n}\frac{\overline{Y}-\mu}{S} \sim t(\nu)$$

The F-distribution arises often in Regression and Analysis of Variance applications. If $W_1 \sim \chi^2(\nu_1)$, $W_2 \sim \chi^2(\nu_2)$, and W_1, W_2 are independent, then:

$$F = \frac{\left[\frac{W_1}{\nu_1}\right]}{\left[\frac{W_2}{\nu_2}\right]} \sim F(\nu_1, \nu_2)$$

where the probability density, mean, and variance for the F-distribution are:

$$f(y) = \left[\frac{\Gamma\left(\left(\frac{\nu_1+\nu_2}{2}\right)\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}\right]\left[\frac{y^{\nu_1/2-1}}{(\nu_1 y + \nu_2)^{(\nu_1+\nu_2)/2}}\right]$$

$$E\{Y\} = \mu_Y = \frac{\nu_1}{\nu_2 - 2} \quad \nu_2 > 2 \qquad V\{Y\} = \sigma_Y^2 = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)(\nu_2-4)} \quad \nu_2 > 4$$

Critical values for the t and F-distributions are given in statistical textbooks. Probabilities can be obtained from many statistical packages and spreadsheets. Technically, the t and F distributions described here are **central** t and **central** F distributions.

Inferences Regarding μ and σ^2

We can test hypotheses concerning μ and obtain confidence intervals based on the sample mean and standard deviation when the data are independent $N(\mu, \sigma^2)$. Let $t(\alpha/2, \nu)$ be the value such that if:

$$T \sim t(\nu) \quad \Rightarrow \quad P(T \geq t(\alpha/2, \nu)) = \alpha/2$$

then we get the following probability statement and $(1-\alpha)100\%$ confidence interval for μ:

$$1-\alpha = P\left(-t(\alpha/2, n-1) \leq \sqrt{n}\frac{\overline{Y}-\mu}{S} \leq t(\alpha/2, n-1)\right) = P\left(-t(\alpha/2, n-1)\frac{S}{\sqrt{n}} \leq \overline{Y}-\mu \leq t(\alpha/2, n-1)\frac{S}{\sqrt{n}}\right) =$$

$$= P\left(\overline{Y} - t(\alpha/2, n-1)\frac{S}{\sqrt{n}} \leq \mu \leq \overline{Y} + t(\alpha/2, n-1)\frac{S}{\sqrt{n}}\right)$$

A 2-sided test of whether $\mu = \mu_0$ is set up as follows, where TS is the test statistic, and RR is the rejection region:

$$H_0: \mu = \mu_0 \quad H_A: \mu \neq \mu_0 \qquad TS: t_{obs} = \sqrt{n}\frac{\overline{Y}-\mu_0}{S} \qquad RR: |t_{obs}| \geq t(\alpha/2, \nu)$$

with P-value $= 2P(t(n-1) \geq |t_{obs}|)$.

To make inferences regarding σ^2, we will make use of the following notational convention:

$$W \sim \chi^2(\nu) \quad \Rightarrow \quad P\left(W \geq \chi^2\left(\alpha/2, \nu\right)\right) = \alpha/2$$

Since the chi^2 distribution is not symmetric around 0, as Student's t is, we will have to also obtain $\chi^2\left(1-\alpha/2, \nu\right)$, representing the lower tail of the distribution having area=$\alpha/2$. Then, we can obtain a $(1-\alpha)100\%$ Confidence interval for σ^2, based on the following probability statements:

$$1-\alpha = P\left(\chi^2\left(1-\alpha/2,\nu\right) \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2\left(\alpha/2,\nu\right)\right) = P\left(\frac{(n-1)S^2}{\chi^2\left(\alpha/2,\nu\right)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2\left(1-\alpha/2,\nu\right)}\right)$$

To obtain a $(1-\alpha)100\%$ Confidence interval for σ, take the positive square roots of the end points. To test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_A : \sigma^2 \neq \sigma_0^2$, simply check whether σ_0^2 lies in the confidence interval for σ^2.

1.4 Likelihood Functions and Maximum Likelihood Estimation

Suppose we take a random sample of n items from a probability mass (discrete) or probability density (continuous) function. We can write the marginal probability density (mass) for the each observation (say y_i) as a function of one or more parameters (θ):

$$\text{Discrete: } p\left(y_i|\theta\right) \qquad \text{Continuous: } f\left(y_i|\theta\right)$$

If the data are independent, then we get the joint density (mass) functions as the product of the individual (marginal) functions:

$$\text{Discrete: } p\left(y_1,\ldots,y_n|\theta\right) = \prod_{i=1}^n p\left(y_i|\theta\right) \qquad \text{Continuous: } f\left(y_1,\ldots,y_n|\theta\right) = \prod_{i=1}^n f\left(y_i|\theta\right)$$

Consider the following cases: Binomial, Poisson, Exponential, and Normal. For the binomial case, suppose we consider n individual trials, where each trial can end in Success (with probability π) of Failure (with probability π). Note that each y_i will equal 1 (S) or 0 (F). This is referred to as a **Bernoulli distribution** when each trial is considered individually:

$$p\left(y_i|\pi\right) = \pi^{y_i}\left(1-\pi\right)^{1-y_i} \quad \Rightarrow \quad p\left(y_1,\ldots,y_n|\pi\right) = \prod_{i=1}^n p\left(y_i|\pi\right) = \pi^{\sum y_i}\left(1-\pi\right)^{n-\sum y_i}$$

For the Poisson model, we have:

$$p\left(y_i|\lambda\right) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \quad \Rightarrow \quad p\left(y_1,\ldots,y_n|\lambda\right) = \prod_{i=1}^n p\left(y_i|\lambda\right) = \frac{e^{-n\lambda}\lambda^{\sum y_i}}{\prod y_i!}$$

For the Exponential model, we have:

$$f\left(y_i|\theta\right) = \theta e^{-y_i\theta} \quad \Rightarrow \quad f\left(y_1,\ldots,y_n|\theta\right) = \prod_{i=1}^n f\left(y_i|\theta\right) = \theta^n e^{-\theta\sum y_i}$$

For the normal distribution, we obtain:

$$f\left(y_i|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(y_i-\mu\right)^2}{2\sigma^2}\right) \quad \Rightarrow$$

$$f\left(y_1,\ldots,y_n|\mu,\sigma^2\right)=\prod_{i=1}^{n} f\left(y_i|\mu,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{\sum\left(y_i-\mu\right)^2}{2\sigma^2}\right]$$

Note that in each of these cases (and for other distributions as well), once we have collected the data, the joint distribution can be thought of as a function of unknown parameter(s). This is referred to as the **likelihood function**. Our goal is to choose parameter value(s) that maximize the likelihood function. These are referred to as **maximum likelihood estimators (MLEs)**. For most distributions, it is easier to maximize the log of the likelihood function.

$$\text{Likelihood: } L\left(\theta|y_1,\ldots,y_n\right)=f\left(y_1,\ldots,y_n|\theta\right) \qquad \text{Log-Likelihood: } l=\ln(L)$$

To obtain the MLE(s), we take the derivative of the log-likelihood with respect to the parameter(s) θ, set to zero, and solve for $\hat{\theta}$. Now, we consider the 4 models described above. For the Binomial (series of Bernoulli trials) model, we have:

$$L\left(\pi|y_1,\ldots,y_n\right)=\pi^{\sum y_i}\left(1-\pi\right)^{n-\sum y_i} \quad\Rightarrow\quad l=\ln(L)=\sum y_i\ln(\pi)+\left(n-\sum y_i\right)\ln\left(1-\pi\right)$$

Taking the derivative of l with respect to π, setting to 0, and solving for $\hat{pi}$, we get:

$$\frac{\partial l}{\partial\pi}=\frac{\sum y_i}{\pi}-\frac{n-\sum y_i}{1-\pi}\overset{\text{set}}{=}0 \quad\Rightarrow\quad \hat{\pi}=\frac{\sum y_i}{n}$$

For the Poisson distribution, we have:

$$L\left(\lambda|y_1,\ldots,y_n\right)=\frac{e^{-n\lambda}\lambda^{\sum y_i}}{\prod y_i!} \quad\Rightarrow\quad l=\ln(L)=-n\lambda+\sum y_i\ln(\lambda)-\sum\ln\left(y_i!\right)$$

$$\frac{\partial l}{\partial\lambda}=-n+\frac{\sum y_i}{\lambda}\overset{\text{set}}{=}0 \quad\Rightarrow\quad \hat{\lambda}=\frac{\sum y_i}{n}$$

For the exponential model, we have:

$$L\left(\theta|y_1,\ldots,y_n\right)=\theta^n e^{-\theta\sum y_i} \quad\Rightarrow\quad l=\ln(L)=n\ln(\theta)-\theta\sum y_i$$

$$\frac{\partial l}{\partial\lambda}=\frac{n}{\theta}-\sum y_i\overset{\text{set}}{=}0 \quad\Rightarrow\quad \hat{\theta}=\frac{n}{\sum y_i}$$

For the Normal distribution, we obtain:

$$L\left(\theta|y_1,\ldots,y_n\right)=\left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{\sum\left(y_i-\mu\right)^2}{2\sigma^2}\right] \quad\Rightarrow\quad l=\ln(L)=-\frac{n}{2}\left[\ln(2\pi)+\ln\left(\sigma^2\right)\right]-\frac{\sum\left(y_i-\mu\right)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial\mu}=\frac{\sum\left(y_i-\mu\right)}{\sigma^2}\overset{\text{set}}{=}0 \quad\Rightarrow\quad \hat{\mu}=\frac{\sum y_i}{n}$$

$$\frac{\partial l}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{\sum\left(y_i-\mu\right)^2}{2\sigma^4}\overset{\text{set}}{=}0 \quad\Rightarrow\quad \hat{\sigma}^2=\frac{\sum\left(y_i-\hat{\mu}\right)^2}{n}$$

Under commonly met regularity conditions, the maximum likelihood estimator $\hat{\theta}_{ML}$ is asymptotically normal, with mean equal to the true parameter(s) θ, and variance (or variance-covariance matrix when the number of parameters, $p>1$) equal to:

$$V\left\{\hat{\theta}_{ML}\right\}=-\left(E\left\{\frac{\partial^2 l}{\partial\theta\partial\theta'}\right\}\right)^{-1}$$

where:

$$\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \qquad \frac{\partial^2 l}{\partial\theta\partial\theta'} = \begin{bmatrix} \frac{\partial^2 l}{\partial\theta_1^2} & \cdots & \frac{\partial^2 l}{\partial\theta_1\partial\theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial\theta_p\partial\theta_1} & \cdots & \frac{\partial^2 l}{\partial\theta_p^2} \end{bmatrix}$$

The estimated variance (or variance-covariance matrix) replaces the unknown parameter values θ with their ML estimates $\hat{\theta}_{ML}$. The **standard error** is the standard deviation of its sampling distribution, the square root of the variance.

For the binomial (sequence of Bernoulli trials), we have, where $E\{Y_i\} = \pi$:

$$l = \sum y_i \ln(\pi) + \left(n - \sum y_i\right) \ln(1-\pi) \qquad \frac{\partial l}{\partial \pi} = \frac{\sum y_i}{\pi} - \frac{n - \sum y_i}{1-\pi}$$

$$\Rightarrow \quad \frac{\partial^2 l}{\partial \pi^2} = -\frac{\sum y_i}{\pi^2} - \frac{n - \sum y_i}{(1-\pi)^2} \qquad \Rightarrow \qquad E\left\{\frac{\partial^2 l}{\partial \pi^2}\right\} = -\frac{n\pi}{\pi^2} - \frac{n(1-\pi)}{(1-\pi)^2} = -n\left(\frac{1}{\pi} + \frac{1}{1-\pi}\right) = -\frac{n}{\pi(1-\pi)}$$

$$\Rightarrow \quad V\{\hat{\pi}_{ML}\} = -\left(-\frac{n}{\pi(1-\pi)}\right)^{-1} = \frac{\pi(1-\pi)}{n} \qquad \Rightarrow \quad \hat{V}\{\hat{\pi}_{ML}\} = \frac{\hat{\pi}(1-\hat{\pi})}{n}$$

For the normal model, we have, where $E\{Y_i\} = \mu$ and $E\left\{(Y_i - \mu)^2\right\} = \sigma^2$:

$$l = -\frac{n}{2}\left[\ln(2\pi) + \ln\left(\sigma^2\right)\right] - \frac{\sum (y_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial \mu} = \frac{\sum (y_i - \mu)}{\sigma^2}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum (y_i - \mu)^2}{2\sigma^4}$$

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} \qquad \Rightarrow \qquad E\left\{\frac{\partial^2 l}{\partial \mu^2}\right\} = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial \left(\sigma^2\right)^2} = \frac{n}{2\sigma^4} - \frac{2\sum (y_i - \mu)^2}{2\sigma^6} \qquad \Rightarrow \quad E\left\{\frac{\partial^2 l}{\partial \left(\sigma^2\right)^2}\right\} = \frac{n}{2\sigma^4} - \frac{2n\sigma^2}{2\sigma^6} = -\frac{n}{2\sigma^4}$$

$$\frac{\partial^2 l}{\partial \mu \partial \sigma^2} = -\frac{\sum (y_i - \mu)}{\sigma^4} \qquad \Rightarrow \qquad E\left\{\frac{\partial^2 l}{\partial \mu \partial \sigma^2}\right\} = 0$$

$$\Rightarrow \quad V\left\{\begin{bmatrix} \hat{\mu}_{ML} \\ \hat{\sigma}^2_{ML} \end{bmatrix}\right\} \quad = \quad -\begin{bmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{bmatrix}^{-1} \quad = \quad \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

$$\Rightarrow \quad \hat{V}\left\{\begin{bmatrix} \hat{\mu}_{ML} \\ \hat{\sigma}^2_{ML} \end{bmatrix}\right\} \quad = \quad \begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}$$

$$\Rightarrow \quad V\{\hat{\mu}_{ML}\} = \frac{\sigma^2}{n} \qquad \Rightarrow \qquad \hat{V}\{\hat{\mu}_{ML}\} = \frac{\hat{\sigma}^2}{n}$$

$$\Rightarrow \quad V\left\{\hat{\sigma}^2_{ML}\right\} = \frac{2\sigma^4}{n} \qquad \Rightarrow \qquad \hat{V}\left\{\hat{\sigma}^2_{ML}\right\} = \frac{2\hat{\sigma}^4}{n}$$

Also note that the covariance of $\hat{\mu}$ and $\hat{\sigma}^2$ is zero.

We can construct large-sample Confidence Intervals for the parameter(s) θ, based on the asymptotic normality of the MLEs:

$$\hat{\theta}_{ML} \pm z_{\alpha/2}\sqrt{\hat{V}\left\{\hat{\theta}_{ML}\right\}} \qquad\qquad P\left\{Z \geq z_{\alpha/2}\right\} = \frac{\alpha}{2}$$

1.5 Likelihood Ratio, Wald, and Score (Lagrange Multiplier) Tests

When we wish to test hypotheses regarding value(s) of parameter(s) θ, there are 3 general classes of tests that make use of the likelihood function and MLEs. These are referred to as **Likelihood Ratio**, **Wald**, and **Score (Lagrange Multiplier)** tests. Asymptotically, they are equivalent. In small-samples, their properties can differ. We consider first the case of a single parameter, then the case of multiple parameters.

The likelihood ratio test is based on the difference in the log-likelihood function $l(\theta) = \ln L(\theta|y_1, \ldots, y_n)$ at its maximum, evaluated at $\theta = \hat{\theta}$ and when it is evaluated at the null value $\theta = \theta_0$.

The Wald test is based on the difference between the maximized value $\hat{\theta}$ and the null value θ_0 in terms of the estimated standard error (square root of the variance) of $\hat{\theta}$.

The score (Lagrange Multiplier) test is based on a function of the derivative (slope) of the likelihood function evaluated at the null value θ_0. It does not depend on the MLE $\hat{\theta}$, so is often used in complex estimation problems.

1.5.1 Single Parameter Models

For one parameter families (such as the Binomial (Bernoulli), Poisson, and Exponential), the procedures are conducted as follows. Note that a Normal with known variance is also a case, but rare in actual practice.

We wish to test a point null hypothesis $H_0 : \theta = \theta_0$ versus an alternative $H_A : \theta \neq \theta_0$. Note that if θ_0 is at the edge of the parameter space, critical values will need to be adjusted.

The **Likelihood Ratio Test** is conducted as follows:

1. Identify the parameter space Ω, such as $\Omega \equiv \{\theta : 0 < \theta < 1\}$ for Binomial or $\Omega \equiv \{\theta : \theta > 0\}$ for the Poisson.
2. Identify the parameter space under $H_0 : \Omega_0 \equiv \{\theta : \theta = \theta_0\}$
3. Evaluate the maximum log-likelihood (terms not involving θ can be ignored)
4. Evaluate the log-likelihood under H_0 (terms not involving θ can be ignored)
5. Compute $X^2_{LR} = -2\left[l(\theta_0) - l\left(\hat{\theta}\right)\right]$
6. Under the null hypothesis, X^2_{LR} is asymptotically distributed as $\chi^2(1)$, where the 1 degree of freedom refers to the number of restrictions under H_0

7. Reject H_0 for large values of X^2_{LR} $\quad \left(X^2_{LR} \geq \chi^2(1,\alpha)\right)$.

The **Wald Test** makes use of the ML estimate, and its standard error, and asymptotic normality to conduct the test. First, consider the variance of the ML estimator described above (using slightly different notation):

$$V\left\{\hat{\theta}\right\} = \frac{1}{n}I^{-1}(\theta) \qquad\qquad I(\theta) = -\frac{1}{n}E\left\{\frac{\partial^2 l(\theta)}{\partial\theta^2}\right\}$$

where $E\left\{\frac{\partial^2 l(\theta)}{\partial\theta^2}\right\}$ is called the **Fisher Information**. Then we obtain the Wald statistic, which is the square of a large-sample Z-statistic (note the use of the estimated variance):

$$X^2_W = \frac{\left(\hat{\theta}-\theta_0\right)^2}{\hat{V}\left\{\hat{\theta}\right\}} = nI\left(\hat{\theta}\right)\left(\hat{\theta}-\theta_0\right)^2$$

As with the Likelihood Ratio Test, under the null hypothesis, X^2_W is asymptotically $\chi^2(1)$ and we use the same rejection region: $\left(X^2_W \geq \chi^2(1,\alpha)\right)$

The **Score (Lagrange Multiplier) Test** is based on the derivative of the log-likelihood, and actually does not make use of the ML estimate $\hat{\theta}$, which can be an advantage in complex estimation problems.

First, compute the first derivative of the log-likelihood, evaluated at the null value θ_0. Note that this will only equal 0 if $\theta_0 = \hat{\theta}$ (the maximum likelihood estimate). This value is called the **score**:

$$s\left(\theta, y\right) = \frac{\partial l(\theta)}{\partial\theta} \qquad\qquad s\left(\theta_0, y\right) = \left.\frac{\partial l(\theta)}{\partial\theta}\right|_{\theta=\theta_0}$$

Next, multiply the score squared by the variance of the ML estimate, evaluated at the null value θ_0, to obtain the score statistic:

$$\left.V\left\{\hat{\theta}\right\}\right|_{\theta=\theta_0} = \frac{1}{nI\left(\theta_0\right)} \qquad \Rightarrow \qquad X^2_{LM} = \frac{s\left(\theta_0, y\right)^2}{nI\left(\theta_0\right)}$$

As with the Likelihood Ratio and Wald statistics, we reject the null if $X^2_{LM} \geq \chi^2(1,\alpha)$.

In the case of the Exponential distribution, where recall that $\hat{\theta} = \frac{n}{\sum y_i} = \frac{1}{\overline{Y}}$, and $\mu_Y = \frac{1}{\theta}$:

$$L\left(\theta|y_1,\ldots,y_n\right) = \theta^n e^{-\theta\sum y_i} \qquad \Rightarrow \qquad l(\theta) = n\ln(\theta) - \theta\sum y_i$$

$$\frac{\partial l(\theta)}{\partial\theta} = \frac{n}{\theta} - \sum y_i \qquad\qquad \frac{\partial^2 l(\theta)}{\partial\theta^2} = -\frac{n}{\theta^2} \qquad I(\theta) = -\frac{1}{n}E\left\{-\frac{n}{\theta^2}\right\} = \frac{1}{n}\frac{n}{\theta^2} = \frac{1}{\theta^2}$$

For the Likelihood Ratio Test, we obtain:

$$l\left(\hat{\theta}\right) = n\ln\left(\hat{\theta}\right) - \hat{\theta}\sum y_i = n\ln\left(\hat{\theta}\right) - \frac{n}{\sum y_i}\sum y_i = n\ln\left(\hat{\theta}\right) - n$$

$$l\left(\theta_0\right) = n\ln\left(\theta_0\right) - \theta_0\sum y_i = n\ln\left(\theta_0\right) - \theta_0\left(\frac{n}{\hat{\theta}}\right)$$

So that the Likelihood Ratio statistic is:

$$X^2_{LR} = -2\left[l\left(\theta_0\right) - l\left(\hat{\theta}\right)\right] = -2\left[\left(n\ln\left(\theta_0\right) - \theta_0\left(\frac{n}{\hat{\theta}}\right)\right) - n\ln\left(\left(\hat{\theta}\right) - n\right)\right] = -2n\left[\ln\left(\frac{\theta_0}{\hat{\theta}}\right) - \left(\frac{\theta_0}{\hat{\theta}} - 1\right)\right]$$

For the Wald Test, we get the statistic:

$$X^2_W = \frac{\left(\hat{\theta} - \theta_0\right)^2}{\hat{V}\left\{\hat{\theta}\right\}} = nI\left(\hat{\theta}\right)\left(\hat{\theta} - \theta_0\right)^2 = n\frac{\left(\hat{\theta} - \theta_0\right)^2}{\hat{\theta}^2}$$

For the Score (Lagrange Multiplier) Test, we obtain the statistic:

$$s\left(\theta_0, y\right) = \frac{n}{\theta_0} - \sum y_i = \frac{n - \theta_0\sum y_i}{\theta_0} \qquad nI\left(\theta_0\right) = \frac{n}{\theta_0^2}$$

$$\Rightarrow \qquad X^2_{LM} = \frac{s\left(\theta_0, y\right)^2}{nI\left(\theta_0\right)} = \frac{\left(\frac{n - \theta_0\sum y_i}{\theta_0}\right)^2}{\left(\frac{n}{\theta_0^2}\right)} = \frac{\theta_0^2}{n}\left(\frac{n - \theta_0 n\overline{Y}}{\theta_0}\right)^2 = \frac{\theta_0^2 n^2}{n\theta_0^2}\left(1 - \theta_0\overline{Y}\right)^2 = n\left(1 - \theta_0\overline{Y}\right)^2$$

1.5.2 Multiple Parameter Models

For models with multiple parameters, all three tests can be extended to make tests among the parameters (not necessarily all of them). For instance, in a Normal model, we may wish to test $H_0 : \mu = 100, \sigma^2 = 400$ against the alternative that either $\mu \neq 100$ and/or $\sigma^2 \neq 100$. Another possibility is that we may be simultaneously modeling a Poisson model among 3 populations and wish to test $H_0 : \lambda_1 = \lambda_2 = \lambda_3$ versus the alternative that the Poisson parameters are not the same among the populations.

Suppose we have p parameters to be estimated. We have $g \leq p$ **linearly independent** linear hypotheses among the parameters. For instance, we cannot test $H_0 : \mu = 100, \mu = 120$. Note, for an introduction to matrix algebra, see the Regression notes. We can write the null hypothesis as follows:

$$\text{Parameter Vector: } \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \qquad H_0 : R\theta = r \qquad R = \begin{bmatrix} R_{11} & \cdots & R_{1p} \\ \vdots & \ddots & \vdots \\ R_{g1} & \cdots & R_{gp} \end{bmatrix} \qquad r = \begin{bmatrix} r_1 \\ \vdots \\ r_g \end{bmatrix}$$

where R and r are a matrix and vector of constants that define the restrictions from the null hypothesis.

For the Normal model example, we have (with $g = 2$ restrictions):

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad r = \begin{bmatrix} 100 \\ 400 \end{bmatrix}$$

For the Poisson example, there are various ways we could test $H_0 : \lambda_1 = \lambda_2 = \lambda_3$, but keep in mind there are only 2 linearly independent restrictions (we are not testing what value they are, just that they are equal). One possibility is:

$$H_{01} : \lambda_1 = \lambda_2, \lambda_1 = \lambda_3 \qquad \Rightarrow \qquad \lambda_1 - \lambda_2 = 0 \quad \lambda_1 - \lambda_3 = 0$$

Note that with these two statements, we imply that $\lambda_2 = \lambda_3$, and including that would cause a redundancy. A second possibility is:

$$H_{02} : \lambda_1 = \lambda_2, \lambda_2 = \lambda_3 \qquad \Rightarrow \qquad \lambda_1 - \lambda_2 = 0 \quad \lambda_2 - \lambda_3 = 0$$

Again, this implies that $\lambda_1 = \lambda_3$.

For these hypotheses, we have:

$$\theta = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \qquad R_1 = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \qquad R_2 = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \qquad r_1 = r_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Defining $l(\theta_1, \ldots, \theta_p | y) = l(\theta)$ as the log-likelihood, and $n_\bullet$ as the overall sample size (summed across group sizes if comparing several populations), we obtain the following quantities for the three tests:

$$\hat{\theta} \equiv \text{ MLE over entire parameter space} \qquad \tilde{\theta} \equiv \text{ MLE over constraint } H_0$$

$$s_i(\theta, y) = \frac{\partial l(\theta)}{\partial \theta_i} \qquad I_{ij}(\theta) = -\frac{1}{n_\bullet} E\left\{ \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right\}$$

with:

$$s(\theta, y) = \begin{bmatrix} s_1(\theta, y) \\ \vdots \\ s_p(\theta, y) \end{bmatrix} \qquad I(\theta) = \begin{bmatrix} I_{11}(\theta) & \cdots & I_{1p}(\theta) \\ \vdots & \ddots & \vdots \\ I_{p1}(\theta) & \cdots & I_{pp}(\theta) \end{bmatrix}$$

Each of the chi-squared statistics will be asymptotically $\chi^2(g)$, under the null hypothesis, where g is the number of restrictions (rows of R and r). The statistics are obtained as follow:

$$\text{Likelihood Ratio:} \qquad X^2_{LR} = -2\left[l\left(\tilde{\theta}, y\right) - l\left(\hat{\theta}, y\right) \right]$$

$$\text{Wald:} \qquad X^2_W = n_\bullet \left(R\hat{\theta} - r\right)' \left(R\left(I\left(\hat{\theta}\right)\right)^{-1} R' \right)^{-1} \left(R\hat{\theta} - r\right)$$

$$\text{Score (LM):} \qquad X^2_{LM} = \frac{1}{n_\bullet} s\left(\tilde{\theta}, y\right)' \left(I\left(\tilde{\theta}\right)\right)^{-1} s\left(\tilde{\theta}, y\right)$$

1.6 Sampling Distributions and an Introduction to the Bootstrap

Previously we described the sampling distributions of various estimators derived from independent and normally distributed random variables. Also, we considered the large-sample properties of maximum likelihood estimators, that inherently meant we know the underlying distribution of the data.

One useful tool for obtaining the exact distribution of linear functions of random variables (when it even exists) is the **moment-generating function** or mgf. This function serves 2 primary purposes. First, it can be used to obtain the non-central moments of a distribution: $E\left\{Y\right\}, E\left\{Y^2\right\}, E\left\{Y^3\right\}, \ldots$. The moment-generating function (if it exists) for a distribution can be obtained as follows:

$$\text{Discrete Distribution: } M_Y(t) = E\left\{e^{tY}\right\} = \sum_y e^{ty}p(y)$$

$$\text{Continuous Distribution: } M_Y(t) = E\left\{e^{tY}\right\} = \int_{-\infty}^{\infty} e^{ty}f(y)dy$$

Without going through the derivations, we obtain (most involve rules of sums or competing the square or change of variables in integration):

$$\text{Binomial:} \qquad M_Y(t) = \sum_{y=0}^{n} e^{ty}\frac{n!}{y!(n-y)!}\pi^y\left(1-\pi\right)^{n-y} = \sum_{y=0}^{n}\frac{n!}{y!(n-y)!}\left(\pi e^t\right)^y\left(1-\pi\right)^{n-y} = \left(\pi e^t + (1-\pi)\right)^n$$

$$\text{Poisson:} \qquad M_Y(t) = \sum_{y=0}^{\infty} e^{ty}\frac{e^{-\lambda}\lambda^y}{y!} = \sum_{y=0}^{\infty}\frac{e^{-\lambda}\left(\lambda e^t\right)^y}{y!} = e^{\lambda\left(e^t-1\right)}$$

$$\text{Normal: } M_Y(t) = \int_{-\infty}^{\infty} e^{ty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(y-\mu\right)^2}{2\sigma^2}\right]dy = \exp\left[\mu t + \frac{t^2\sigma^2}{2}\right]$$

$$\text{Gamma: } M_Y(t) = \int_{-\infty}^{\infty} e^{ty}\frac{1}{\Gamma(\alpha)\beta^\alpha}y^{\alpha-1}e^{-y/\beta}dy = \left(1-\beta t\right)^{-\alpha}$$

Note that for the other formulation of the Gamma, we would have $M_Y(t) = \left(1-\frac{t}{\theta}\right)^{-\alpha}$. Further, for the Exponential, we would have $M_Y(t) = \left(1-\frac{t}{\theta}\right)^{-1}$ and for the chi-squared, we would have $\left(1-2t\right)^{-\nu/2}$.

The mgf can be used to obtain the non-central moments as follows, based on a series expansion $e^{tY} = \sum_{i=0}^{\infty}\frac{(tY)^i}{i!}$

$$\left.\frac{dM(t)}{dt}\right|_{t=0} = M'(0) = E\left\{Y\right\} \qquad \left.\frac{d^2M(t)}{dt^2}\right|_{t=0} = M''(0) = E\left\{Y^2\right\}$$

and so on such that $M^{(k)}(0) = E\left\{Y^k\right\}$.

More importantly here, if we obtain a sum or linear function of **independent** random of random variables, we can use the uniqueness of the mgf to obtain the distribution of the sum or linear function. Consider $W = Y_1 + \cdots + Y_n$, a sum of independent random variables:

$$M_W(t) = E\left\{e^{tW}\right\} = E\left\{e^{t(Y_1+\cdots+Y_n)}\right\} = \prod_{i=1}^{n} E\left\{e^{tY_i}\right\} = \prod_{i=1}^{n} M_{Y_i}(t)$$

The independence of $Y_1, \ldots, Y_n$ is why we can use this result.

Consider m Binomial random variables, each with success probability π, but with varying sample sizes n_i:

$$Y_i \sim Bin\left(n_i, \pi\right) \quad i = 1, \ldots, m \qquad M_{Y_i}(t) = \left(\pi e^t + (1-\pi)\right)^{n_i}$$

Thus if we let $W = Y_1 + \cdots + Y_m$, we have:

$$M_W(t) = \prod_{i=1}^{m} M_{Y_i}(t) = \prod_{i=1}^{m}\left(\pi e^t + (1-\pi)\right)^{n_i} = \left(\pi e^t + (1-\pi)\right)^{\sum n_i} \qquad \Rightarrow \qquad W \sim \text{Binomial}\left(\sum n_i, \pi\right)$$

Thus, the sum of independent Binomials with common success probability is Binomial with the same success probability, and a sample size equal to the sum of the sample sizes.

Similar results lead to for independent Poissons, where $Y_i \sim \text{Poisson}(\lambda_i)$. Let $W = Y_1 + \cdots + Y_n$:

$$M_W(t) = \prod_{i=1}^{n} M_{Y_i}(t) = \prod_{i=1}^{n} e^{\lambda_i\left(e^t - 1\right)} = exp\left[\left(\sum \lambda_i\right)\left(e^t - 1\right)\right] \qquad \Rightarrow \qquad W \sim \text{Poisson}\left(\sum \lambda_i\right)$$

For a sum of independent Gammas, with common β or θ, that is $Y_i \sim \text{Gamma}(\alpha_i, \beta)$. Let $W = Y_1 + \cdots + Y_n$:

$$M_W(t) = \prod_{i=1}^{n} M_{Y_i}(t) = \prod_{i=1}^{n} (1 - \beta t)^{-\alpha_i} = (1 - \beta t)^{-\sum \alpha_i} \qquad \Rightarrow \qquad W \sim \text{Gamma}\left(\sum \alpha_i, \beta\right)$$

Now consider any linear function $U = a_1 Y_1 + \cdots + a_n Y_n$, for constants $a_1, \ldots, a_n$. This will not work for many distributions:

$$M_U(t) = E\left\{e^{tW}\right\} = E\left\{e^{t(a_1 Y_1 + \cdots + a_n Y_n)}\right\} = \prod_{i=1}^{n} E\left\{e^{t a_i Y_i}\right\} = \prod_{i=1}^{n} M_{Y_i}(a_i t)$$

Now consider independent Normals, with $Y_i \sim N\left(\mu_i, \sigma_i^2\right)$. Let $U = a_1 Y_1 + \cdots + a_n Y_n$:

$$M_U(t) = \prod_{i=1}^{n} M_{Y_i}(a_i t) = \prod_{i=1}^{n} \exp\left[\mu_i a_i t + \frac{a_i^2 t^2 \sigma_i^2}{2}\right] = \exp\left[t\left(\sum a_i \mu_i\right) + \frac{t^2\left(\sum a_i^2 \sigma_i^2\right)}{2}\right] \quad \Rightarrow \quad V \sim N\left(a_i \mu_i, \sum a_i^2 \sigma_i^2\right)$$

So, in special circumstances, when we know the exact distribution of data, we can obtain the exact distribution of some specific estimators. Due to **Central Limit Theorems**, we can also state that sample means of independent observations have sampling distributions that asymptotically converge to the Normal (assuming finite variance). Thus, in many cases:

$$\sqrt{n}\frac{\overline{Y} - \mu}{S} \quad \overset{\text{approx}}{\sim} \quad N(0,1) \qquad \Rightarrow \qquad \overline{Y} \quad \overset{\text{approx}}{\sim} \quad N\left(\mu, \frac{\sigma^2}{n}\right)$$

However, in many settings, estimators either are very complex and no sampling distribution can be derived, or samples are not large enough to rely on large-sample asymptotics. In these settings, the **bootstrap method** is applied to a statistic to obtain a Confidence Interval for the underlying parameter. The method assumes the sample is representative of the underlying population (e.g. no inherent bias). Also, if the sampling plan has any specific patterns to it, such as clusters, the bootstrap should reflect that.

The algorithm works as follows:

1. Obtain a sample of size N from the population of interest.
2. Generate a method (function) to compute the statistic of interest.

3. Generate a random sample with replacement from the original sample, apply the function, and save the result.
4. Repeat this process over many samples.
5. Obtain a $(1-\alpha)100\%$ Confidence Interval for the parameter of interest.

The last step can be conducted various ways, the most common way is to select the cut-off values of the middle $(1-\alpha)100\%$ bootstrap sample results. Other ways, particularly bias-corrected methods are implemented in standard statistical software packages, and make use of the mean and standard deviation (standard error) of the bootstrap estimates. This version is referred to as the **non-parametric bootstrap**, which makes no assumptions on the underlying distribution of the data.

Another possibility when you are confident about the underlying distribution, but unsure of parameter values. Then, the parameters can be estimated (based on methods such as ML in previous sections), and then many samples can be generated using random number generators from the corresponding distribution. The Confidence Interval and mean and standard error of the estimator can be obtained as well. This is referred to as the **non-parametric bootstrap**.

Chapter 2

Simple Linear Regression

2.1 Introduction

Linear regression is used when we have a numeric response variable and numeric (and possibly categorical) predictor (explanatory) variable(s). The mean of the response variable is to be related to the predictor(s) with random error terms assumed to be independent and normally distributed with constant variance. The fitting of linear regression models is very flexible, allowing for fitting curvature and interactions between factors.

When there is a single numeric predictor, we refer to the model as **Simple Regression**. The response variable is denoted as Y and the predictor variable is denoted as X. The assumed model is:

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad\qquad \epsilon \sim N(0, \sigma^2) \text{ independent}$$

Here β_0 is the intercept (mean of Y when X=0) and β_1 is the slope (the change in the mean of Y when X increases by 1 unit). Of primary concern is whether $\beta_1 = 0$, which implies the mean of Y is constant (β_0), and thus Y and X are not associated.

Note that this model assumes:

$$E\{\epsilon\} = 0 \qquad V\{\epsilon\} = E\{\epsilon^2\} = \sigma^2 \qquad COV\{\epsilon_i, \epsilon_j\} = E\{\epsilon_i \epsilon_j\} = 0 \quad i \neq j$$

In practice the variance may not be constant, and the errors may not be independent. These assumptions will be checked after fitting model.

Estimation of Model Parameters

We obtain a sample of pairs $(X_i, Y_i) \quad i = 1, \ldots, n$. Our goal is to choose estimators of β_0 and β_1 that minimize the error sum of squares: $Q = \sum_{i=1}^{n} \epsilon_i^2$. The resulting estimators are (from calculus):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Once we have estimates, we obtain **fitted values** and **residuals** for each observation. The **error sum of squares (SSE)** are obtained as the sum of the squared residuals:

$$\text{Fitted Values: } \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad \text{Residuals: } e_i = Y_i - \hat{Y}_i \qquad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

The (unbiased) estimate of the error variance σ^2 is $s^2 = MSE = \frac{SSE}{n-2}$, where MSE is the **Mean Square Error**. The subtraction of 2 can be thought of as the fact that we have estimated two parameters: β_0 and β_1.

The estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ can be written as linear functions of $Y_1, \ldots, Y_n$:

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i \qquad \hat{\beta}_0 = \sum_{i=1}^{n} b_i Y_i \quad \text{where} \quad a_i = \frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad b_i = \frac{1}{n} + \frac{\overline{X}\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

and thus using the following basic rules of mathematical statistics

$$E\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i E\left\{Y_i\right\} \qquad V\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i^2 V\{Y_i\} + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j \text{COV}\{Y_i, Y_j\}$$

The last term of the variance drops out when the data are independent. Thus, the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$, assuming independent, normal errors with constant variance are:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) \qquad \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]\right)$$

The standard error is the square root of the variance, and the estimated standard error is the standard error with the unknown σ^2 replaced by MSE.

$$SE\{\hat{\beta}_1\} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \qquad SE\{\hat{\beta}_0\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}$$

2.2 Inference Regarding β_1

Primarily of interest is inferences regarding β_1. Note that if $\beta_1 = 0$, Y and X are not associated. We can test hypotheses and construct confidence intervals based on the estimate β_1 and its estimated standard error. The t-test is conducted as follows. Note that the null value β_{10} is almost always 0, and that software packages that report these tests always are treating β_{10} as 0. Here, and in all other tests, TS represents Test Statistic, and RR represents Rejection Region.

$$H_0 : \beta_1 = \beta_{10} \qquad H_A : \beta_1 \neq \beta_{10} \quad TS : t_{obs} = \frac{\hat{\beta}_1 - \beta_{10}}{SE\{\hat{\beta}_1\}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-2} \qquad P\text{-value} : P(t_{n-2} \geq |t_{obs}|)$$

One-sided tests use the same test statistic, but adjusts the Rejection Region and P-value are changed to reflect the alternative hypothesis:

$$H_A^+ : \beta_1 > \beta_{10} \qquad RR : t_{obs} \geq t_{\alpha,n-2} \qquad P\text{-value} : P(t_{n-2} \geq t_{obs})$$

$$H_A^- : \beta_1 < \beta_{10} \qquad RR : t_{obs} \leq -t_{\alpha,n-2} \qquad P\text{-value} : P(t_{n-2} \leq t_{obs})$$

A $(1-\alpha)100\%$ confidence interval for β_1 is obtained as:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} SE\{\hat{\beta}_1\}$$

Note that the confidence interval represents the values of β_{10} that the two-sided test: $H_0 : \beta_1 = \beta_{10} \quad H_A : \beta_1 \neq \beta_{10}$ fails to reject the null hypothesis.

Inferences regarding β_0 are rarely of interest, but can be conducted in analogous manner, using the estimate $\hat{\beta}_0$ and its estimated standard error $SE\{\hat{\beta}_0\}$.

2.3 Estimating a Mean and Predicting a New Observation @ $X = X^*$

We may want to estimate the mean response at a specific level X^*. The parameter of interest is $\mu^* = \beta_0 + \beta_1 X^*$. The point estimate, standard error, and $(1-\alpha)100\%$ Confidence Interval are given below:

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad SE\left\{\hat{Y}^*\right\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]} \qquad (1-\alpha)100\% \text{ CI } : \hat{Y}^* \pm t_{\alpha/2,n-2} SE\left\{\hat{Y}^*\right\}$$

To obtain a $(1-\alpha)100\%$ Confidence Interval for the entire regression line (not just a single point), we use the Working-Hotelling method:

$$\hat{Y}^* \pm \sqrt{2F_{\alpha/2,2,n-2}}SE\left\{\hat{Y}^*\right\}$$

If we are interested in predicting a new observation when $X = X^*$, we have uncertainty with respect to estimating the mean (as seen by the Confidence Interval above), and the random error for the new case (with standard deviation σ). The point prediction is the same as for the mean. The estimate, standard error of prediction, and $(1-\alpha)100\%$ Prediction Interval are given below:

$$\hat{Y}^*_{\text{New}} = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad SE\left\{\hat{Y}^*_{\text{New}}\right\} = \sqrt{MSE\left[1 + \frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]}$$

$$(1-\alpha)100\% \text{ PI } : \hat{Y}^*_{\text{New}} \pm t_{\alpha/2,n-2}SE\left\{\hat{Y}^*_{\text{New}}\right\}$$

Note that the Prediction Interval will tend to be much wider than the Confidence Interval for the mean.

2.4 Analysis of Variance

When there is no association between Y and X ($\beta_1 = 0$), the best predictor of each observation is $\overline{Y} = \hat{\beta}_0$ (in terms of minimizing sum of squares of prediction errors). In this case, the total variation can be denoted as $TSS = \sum_{i=1}^n (Y_i - \overline{Y})^2$, the **Total Sum of Squares**.

When there is an association between Y and X ($\beta_1 \neq 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ (in terms of minimizing sum of squares of prediction errors). In this case, the error variation can be denoted as $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, the **Error Sum of Squares**.

The difference between TSS and SSE is the variation "explained" by the regression of Y on X (as opposed to having ignored X). It represents the difference between the fitted values and the mean: $SSR = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$ the **Regression Sum of Squares**.

$$TSS = SSE + SSR \qquad \sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$$

Each sum of squares has a **degrees of freedom** associated with it. The **Total Degrees of Freedom** is $df_{\text{Total}} = n - 1$. The **Error Degrees of Freedom** is $df_{\text{Error}} = n - 2$ (for simple regression). The **Regression Degrees of Freedom** is $df_{\text{Regression}} = 1$ (for simple regression).

$$df_{\text{Total}} = df_{\text{Error}} + df_{\text{Regression}} \qquad n - 1 = n - 2 + 1$$

Error and Regression sums of squares have a **Mean Square**, which is the sum of squares divided by its corresponding degrees of freedom: $MSE = SSE/(n-2)$ and $MSR = SSR/1$. It can be shown that

Source	*df*	SS	MS	F_{obs}	P-value
Regression (Model)	1	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$MSR = \frac{SSR}{1}$	$F_{obs} = \frac{MSR}{MSE}$	$P(F_{1,n-2} \geq F_{obs})$
Error (Residual)	$n-2$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-2}$		
Total (Corrected)	$n-1$	$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

Table 2.1: Analysis of Variance Table for Simple Linear Regression

these mean squares have the following **Expected Values**, average values in repeated sampling at the same observed X levels:

$$E\{MSE\} = \sigma^2 \qquad\qquad E\{MSR\} = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

Note that when $\beta_1 = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise $E\{MSR\} > E\{MSE\}$. A second way of testing whether $\beta_1 = 0$ is by the F-test:

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0 \quad TS : F_{obs} = \frac{MSR}{MSE} \qquad RR : F_{obs} \geq F_{\alpha,1,n-2} \qquad P\text{-value} : P(F_{1,n-2} \geq F_{obs})$$

The Analysis of Variance is typically set up in a table as in Table 2.1.

A measure often reported from a regression analysis is the **Coefficient of Determination** or r^2. This represents the variation in Y "explained" by X, divided by the total variation in Y.

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad 0 \leq r^2 \leq 1$$

The interpretation of r^2 is the proportion of variation in Y that is"explained" by X, and is often reported as a percentage ($100r^2$).

2.5 Correlation

The regression coefficient β_1 depends on the units of Y and X. It also depends on which variable is the dependent variable and which is the independent variable. A second widely reported measure is the **Pearson Product Moment Coefficient of Correlation**. It is invariant to linear transformations of Y and X, and does not distinguish which is the dependent and which is the independent variables. This makes it a widely reported measure when researchers are interested in how 2 random variables vary together in a population. The population correlation coefficient is labeled ρ, and the sample correlation is labeled r, and is computed as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \left(\frac{s_X}{s_Y}\right)\hat{\beta}_1$$

where s_X and s_Y are the standard deviations of X and Y, respectively. While $\hat{\beta}_1$ can take on any value, r lies between -1 and +1, taking on the extreme values if all of the points fall on a straight line. The test of whether $\rho = 0$ is mathematically equivalent to the t-test for testig thether $\beta_1 = 0$. The 2-sided test is given below:

$$H_0 : \rho = 0 \qquad H_A : \rho \neq 0 \quad TS : t_{obs} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-2} \qquad P - \text{value} : P(t_{n-2} \geq |t_{obs}|)$$

To construct a large-sample confidence interval, we use **Fisher's z transform** to make r approximately normal. We then construct a confidence interval on the transformed correlation, then "back transform" the end points:

$$z' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad (1-\alpha)100\% \text{ CI for } \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) : z' \pm z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

Labeling the endpoints of the Confidence Interval as (a, b), we obtain:

$$(1-\alpha)100\% \text{ Confidence Interval for } \rho : \left(\frac{e^{2a}-1}{e^{2a}+1}, \frac{e^{2b}-1}{e^{2b}+1}\right)$$

Chapter 3

Matrix Form of Simple Linear Regression

We can write out the regression model in a more concise form using the **matrix form**. This is particularly helpful when we have multiple predictors. We first "string out" the dependent variable (Y), and the predictor variable (X) into **arrays**. In fact, we augment the X^s with a column of 1^s for the intercept:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We can make use of some basic matrix rules to simplify the algebra of regression models. Note that matrices with one row or column are referred to as **vectors**. Matrices with the same number of rows and columns are referred to as **square matrices**. When referring to elements of matrices, the row represents the first subscript, and column is second subscript. Vector elements have one subscript.

The **transpose** of a matrix or vector, is the matrix or vector obtained by interchanging its rows and columns (turning it on its side, counterclockwise). It is typically written with a "prime" or "T" as a superscript.

$$\mathbf{Y}' = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix} \quad \mathbf{X}' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \quad \boldsymbol{\beta}' = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix} \quad \boldsymbol{\varepsilon}' = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_n \end{bmatrix}$$

Matrix Addition/Subtraction: If two matrices are of the same dimension (numbers of rows and columns), then the matrix formed by adding/subtracting each of the elements within the given rows and columns is the addition/subtraction of the two matrices.

$$\mathbf{A} = \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 3 & 1 \\ 8 & 6 \end{bmatrix} \quad \Rightarrow$$

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} 4+3 & 8+1 \\ 2+8 & -4+6 \end{bmatrix} = \begin{bmatrix} 7 & 9 \\ 10 & 2 \end{bmatrix} \qquad \mathbf{A}-\mathbf{B} = \begin{bmatrix} 4-3 & 8-1 \\ 2-8 & -4-6 \end{bmatrix} = \begin{bmatrix} 1 & 7 \\ -6 & -10 \end{bmatrix}$$

Matrix Multiplication: Unlike Addition/Subtraction, Multiplication takes sums of products of matrix elements. The number of **columns** of the **left-hand** matrix must be equal to the number of **rows** of the **right-hand** matrix. The resulting matrix has the same number of rows of the left-hand matrix and the number of columns as the right-hand matrix. Note that multiplication of square matrices of common dimensions will result in a square matrix of the same dimension. The elements of a matrix created by multiplication are the sums of products of elements in the rows of the left-hand matrix with the elements of the columns of the right-hand matrix.

$$\mathbf{AB} = \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 8 & 6 \end{bmatrix} = \begin{bmatrix} 76 & 52 \\ -26 & -22 \end{bmatrix}$$

Note the computation of elements of **AB**:

$$\mathbf{AB}_{11} = 4(3)+8(8) = 12+64 = 76 \qquad \mathbf{AB}_{12} = 4(1)+8(6) = 4+48 = 52$$

$$\mathbf{AB}_{21} = 2(3)+(-4)(8) = 6-32 = -26 \qquad \mathbf{AB}_{22} = 2(1)+(-4)(6) = 2-24 = -22$$

Important matrix multiplications for the simple linear regression model are:

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1\,(\beta_0) + X_1\,(\beta_1) \\ 1\,(\beta_0) + X_2\,(\beta_1) \\ \vdots \\ 1\,(\beta_0) + X_n\,(\beta_1) \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix}$$

The statistical model in matrix form (which easily generalizes to multiple predictors is written as:

$$y_i = \mathbf{x_i}'\boldsymbol{\beta} = \begin{bmatrix} 1 & X_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \epsilon_{\mathbf{i}} \quad \Rightarrow \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{where } \mathbf{X} = \begin{bmatrix} \mathbf{x_1}' \\ \vdots \\ \mathbf{x_n}' \end{bmatrix}$$

Other matrices used in model estimation are:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} 1(1)+\cdots+1(1) & 1(X_1)+\cdots+1(X_n) \\ X_1(1)+\cdots+X_n(1) & X_1^2+\cdots+X_n^2 \end{bmatrix}$$

$$\Rightarrow \quad \mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \qquad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix} \qquad \mathbf{Y}'\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i^2 \end{bmatrix}$$

Identity and Unit Matrices: The **identity** (or **I**) matrix is a square matrix with 1^s on the main diagonal, and 0^s elsewhere. When the identity matrix is multiplied by any multiplication-compatible matrix, it reproduces the multiplied matrix. Thus, it acts like 1 in scalar arithmetic. The **unit** (or **J**) matrix is a matrix of 1^s in all cells. When the unit matrix is multiplied by a multiplication-compatible matrix, it sums the elements of each column (and reproduces the sums for each row).

$$\mathbf{IA} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} = \begin{bmatrix} 1(4)+0(2) & 1(8)+0(-4) \\ 0(4)+1(2) & 0(8)+1(-4) \end{bmatrix} = \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} = \mathbf{A}$$

$$\mathbf{JA} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} = \begin{bmatrix} 1(4)+1(2) & 1(8)+1(-4) \\ 1(4)+1(2) & 1(8)+1(-4) \end{bmatrix} = \begin{bmatrix} 6 & 4 \\ 6 & 4 \end{bmatrix}$$

Matrix Inversion: If a matrix is square and of full rank (no linear functions of a set of columns/rows are equal to another column/row), then an inverse exists. Note that in simple regression, this simply means that the X levels are not all the same among observations. When a square, full rank matrix is multiplied by its inverse, we obtain the identity matrix. This is analogous to the scalar operation: $a(1/a) = 1$, assuming $a \neq 0$. For a 2×2 matrix, the inverse is simple to compute. For larger matrices, we will use computers to obtain them.

$$\mathbf{D} = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \quad \Rightarrow \quad \mathbf{D^{-1}} = \frac{1}{D_{11}D_{22} - D_{12}D_{21}} \begin{bmatrix} D_{22} & -D_{12} \\ -D_{21} & D_{11} \end{bmatrix}$$

Note that if D is not full rank (its columns/rows) are multiples of each other, $D_{11}D_{22} - D_{12}D_{21} = 0$, and its inverse does not exist.

$$\mathbf{A} = \begin{bmatrix} 4 & 8 \\ 2 & -4 \end{bmatrix} \quad \Rightarrow \quad \mathbf{A{-}1} = \frac{1}{4(-4) - 8(2)} \begin{bmatrix} -4 & -8 \\ -2 & 4 \end{bmatrix} = \frac{1}{-32} \begin{bmatrix} -4 & -8 \\ -2 & 4 \end{bmatrix} = \begin{bmatrix} \frac{1}{8} & \frac{1}{4} \\ \frac{1}{16} & -\frac{1}{8} \end{bmatrix}$$

Confirm that $\mathbf{AA^{-1}} = \mathbf{A^{-1}A} = \mathbf{I}$. Serious rounding errors can occur when the division of the determinant $\frac{1}{D_{11}D_{22} - D_{12}D_{21}}$ is rounded down to too few decimal places.

Some very useful results are as follows (assuming matrices are compatible for the operations, which always holds when each is square and of same dimension):

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}' \qquad (\mathbf{AB})^{-1} = \mathbf{B^{-1}A^{-1}} \qquad \mathbf{tr}(\mathbf{AB}) = \mathbf{tr}(\mathbf{BA})$$

The last one is the **trace** of the matrix, which represents the sum of the diagonal elements of a square matrix.

An important application in regression is as follows. The **normal equations** that are obtained from ordinary least squares are: $\mathbf{X'X}\boldsymbol{\beta} = \mathbf{X'Y}$. This is a result from calculus as we try and minimize the error sum of squares:

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\,(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} - \mathbf{2Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Making use of the following calculus results for column vectors **a** and **w**, and symmetric matrix **A**:

$$\frac{\partial \mathbf{a}'\mathbf{w}}{\partial \mathbf{w}} = \mathbf{a} \qquad \frac{\partial \mathbf{w}'\mathbf{A}'\mathbf{w}}{\partial \mathbf{w}} = \mathbf{2Aw}$$

we obtain, the following derivative for Q with respect to $\boldsymbol{\beta}$, set it to 0, and solve for $\hat{\boldsymbol{\beta}}$:

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{Y} + \mathbf{2X}'\mathbf{X}\boldsymbol{\beta} \overset{\text{set}}{=} \mathbf{0} \quad \Rightarrow \quad \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}$$

From this result, we can obtain the vectors of fitted values and residuals, and the sums of squares for the ANOVA from the data matrices and vectors:

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} \quad = \quad \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 X_1 \\ \hat{\beta}_0 + \hat{\beta}_1 X_2 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 X_n \end{bmatrix} \quad = \quad \mathbf{X}\hat{\boldsymbol{\beta}} \quad = \quad \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}$$

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad = \quad \begin{bmatrix} Y_1 - \hat{Y}_1 \\ Y_2 - \hat{Y}_2 \\ \vdots \\ Y_n - \hat{Y}_n \end{bmatrix} \quad = \quad \mathbf{Y} - \hat{\mathbf{Y}} \quad = \quad \left(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\right)\mathbf{Y}$$

$$\overline{\mathbf{Y}} = \begin{bmatrix} \overline{Y} \\ \overline{Y} \\ \vdots \\ \overline{Y} \end{bmatrix} \quad = \quad \frac{1}{n}\begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n Y_i \\ \vdots \\ \sum_{i=1}^n Y_i \end{bmatrix} \quad = \quad \frac{1}{n}\mathbf{JY}$$

The matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$ is very important in regression analysis. It is **symmetric** and **idempotent**. That is, it is equal to its transpose, and when you multiply it by itself (square it), you obtain it again. It is called the **hat** or **projection** matrix, and is often denotes as **H** or **P**. Here, we will use **P** to denote the projection matrix.

The sums of squares can be written in terms of **quadratic forms** of the data vector **Y**. First however note the following results involving matrices used in their construction:

$$\mathbf{PX} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X} = \mathbf{XI} = \mathbf{X} \qquad \mathbf{PP} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' = \mathbf{XI}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' = \mathbf{P}$$

Note that if the model has an intercept (β_0), then the first column of **X**, is a column of 1^s. Then, since $\mathbf{PX} = \mathbf{X}$, that implies $\mathbf{PJ} = \mathbf{J}$, since **J** is a $n \times n$ matrix of 1^s.

$$\mathbf{P}\mathbf{J} = \mathbf{J} \qquad \mathbf{J}\mathbf{J} = n\mathbf{J} \Rightarrow \frac{1}{n}\mathbf{J}\frac{1}{n}\mathbf{J} = \frac{1}{n}\mathbf{J}$$

Now, we re-introduce the sums of squares, and write them in matrix form. The Total (corrected) sum of squares is:

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \left(\mathbf{Y} - \overline{\mathbf{Y}}\right)'\left(\mathbf{Y} - \overline{\mathbf{Y}}\right) \quad = \quad \mathbf{Y}'\left(\mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right)'\left(\mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right)\mathbf{Y} \quad = \quad \mathbf{Y}'\left(\mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right)\mathbf{Y}$$

This is partitioned into the error (SSE) and regression (SSR) sums of squares:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \left(\mathbf{Y} - \hat{\mathbf{Y}}\right)'\left(\mathbf{Y} - \hat{\mathbf{Y}}\right) \quad = \quad \mathbf{Y}'\left(\mathbf{I} - \mathbf{P}\right)'\left(\mathbf{I} - \mathbf{P}\right)\mathbf{Y} \quad = \quad \mathbf{Y}'\left(\mathbf{I} - \mathbf{P}\right)\mathbf{Y}$$

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \left(\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\right)'\left(\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\right) \quad = \quad \mathbf{Y}'\left(\mathbf{P} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right)'\left(\mathbf{P} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right)\mathbf{Y} \quad = \quad \mathbf{Y}'\left(\mathbf{P} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right)\mathbf{Y}$$

Chapter 4

Distributional Results

The model for the observed data (data generating process) can be thought of as Y_i is a random variable with a mean (systematic component) of $\beta_0+\beta_1 X_i$ and a random error term of ϵ_i that reflects all possible sources of variation beyond the predictor X. We assume that the error terms have mean 0, and variance σ_i^2. In general, the error terms may or may not be independent (uncorrelated). The expectation and variance-COVariance matrix of the vector of error terms $\boldsymbol{\varepsilon}$ are:

$$E\{\boldsymbol{\varepsilon}\} = \begin{bmatrix} E\{\epsilon_1\} \\ E\{\epsilon_2\} \\ \vdots \\ E\{\epsilon_n\} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

$$V\{\boldsymbol{\varepsilon}\} = E\left\{(\boldsymbol{\varepsilon}-E\{\boldsymbol{\varepsilon}\})(\boldsymbol{\varepsilon}-E\{\boldsymbol{\varepsilon}\})'\right\} = E\left\{(\boldsymbol{\varepsilon}-\mathbf{0})(\boldsymbol{\varepsilon}-\mathbf{0})'\right\} = E\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right\} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$$

The expectation and variance-covariance matrix of the data vector $\mathbf{Y}$ are:

$$E\{\mathbf{Y}\} = \begin{bmatrix} E\{Y_1\} \\ E\{Y_2\} \\ \vdots \\ E\{Y_n\} \end{bmatrix} = \begin{bmatrix} \beta_0+\beta_1 X_1 \\ \beta_0+\beta_1 X_2 \\ \vdots \\ \beta_0+\beta_1 X_n \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}$$

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = V\{\mathbf{Y}\} = E\left\{(\mathbf{Y}-E\{\mathbf{Y}\})(\mathbf{Y}-E\{\mathbf{Y}\})'\right\} = E\left\{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\right\} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$$

where σ_{ij} is the **covariance** between the i^{th} and j^{th} measurements. When the data are independent, but not necessarily of equal variance (heteroskedastic), we have:

$$V\{\mathbf{Y}\} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

When the data are independent with constant variance (homoskedastic), we have $V\{\mathbf{Y}\} = \sigma^2\mathbf{I}$. This is the common assumption underlying the model, which needs to be checked in practice.

For a random matrix $\mathbf{W}$, and a matrix of fixed constants $\mathbf{A}$ of compatible dimensions for multiplication:

$$E\{\mathbf{AW}\} = \mathbf{A}E\{\mathbf{W}\} \qquad V\{\mathbf{AW}\} = \mathbf{A}V\{\mathbf{W}\}\mathbf{A}'$$

When applied to the least squares estimate $\boldsymbol{\beta}$, we obtain:

$$E\left\{\hat{\boldsymbol{\beta}}\right\} = E\left\{(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}\right\} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'E\{\mathbf{Y}\} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$V\left\{\hat{\boldsymbol{\beta}}\right\} = V\left\{(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}\right\} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'V\{\mathbf{Y}\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$$

When the data are independent, with constant variance, $\boldsymbol{\Sigma}_{\mathbf{Y}} = \sigma^2\mathbf{I}$ then the variance of $\hat{\boldsymbol{\beta}}$ simplifies to:

$$V\left\{\hat{\boldsymbol{\beta}}\right\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} \quad \text{with estimated variance} \quad \hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = s^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} \quad s^2 = MSE = \frac{SSE}{n-2}$$

Further, if $\mathbf{Y}$ is (multivariate) normal, then so is $\hat{\boldsymbol{\beta}}$, and when based on large samples, $\hat{\boldsymbol{\beta}}$ is approximately normal, even when $\mathbf{Y}$ is not, based on Central Limit Theorems.

For **Quadratic forms**, where we have a random column vector, $\mathbf{w}$, and a matrix of constants $\mathbf{A}$ we have the random scalar $\mathbf{w}'\mathbf{A}\mathbf{w}$. If $\mathbf{w}$has mean $\mu_{\mathbf{W}}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{W}} = \sigma^2\mathbf{V}$, then:

$$E\left\{\mathbf{w}'\mathbf{A}\mathbf{w}\right\} = \text{trace}(\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{W}}) + \mu_{\mathbf{W}}'\mathbf{A}\mu_{\mathbf{W}} = \sigma^2\text{trace}((\mathbf{AV}) + \mu_{\mathbf{W}}'\mathbf{A}\mu_{\mathbf{W}}$$

Consider the 3 quantities that make up the Analysis of Variance: $\mathbf{Y}'\mathbf{IY} = \mathbf{Y}'\mathbf{Y}$, $\mathbf{Y}'\mathbf{PY}$, $\mathbf{Y}'\left(\frac{1}{n}\right)\mathbf{JY}$. Here, we have $\mu_{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}} = \sigma^2\mathbf{V}$. Here, we consider the basic case (independent and constant variance, $\mathbf{V} = \mathbf{I}$). Further, recall that trace($\mathbf{AB}$)=trace($\mathbf{BA}$):

$$E\left\{\mathbf{Y}'\mathbf{IY}\right\} = \text{trace}(\mathbf{I}\boldsymbol{\Sigma}_{\mathbf{Y}}) + \mu_{\mathbf{Y}}'\mathbf{I}\mu_{\mathbf{Y}} = \sigma^2\text{trace}\left((\mathbf{II}) + \mu_{\mathbf{Y}}'\mathbf{I}\mu_{\mathbf{Y}}\right) = n\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Recalling that $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$, we can obtain the trace of $\mathbf{P}$ as trace of $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} = \mathbf{I_2}$, which is 2. Further, recall that $\mathbf{PX} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X} = \mathbf{X}$, so that $\boldsymbol{\beta}'\mathbf{X}'\mathbf{PX}\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$:

$$E\left\{\mathbf{Y}'\mathbf{PY}\right\} = \text{trace}(\mathbf{P}\boldsymbol{\Sigma}_{\mathbf{Y}}) + \mu_{\mathbf{Y}}'\mathbf{P}\mu_{\mathbf{Y}} = \sigma^2\text{trace}((\mathbf{PI}) + \mu_{\mathbf{Y}}'\mathbf{P}\mu_{\mathbf{Y}} = 2\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

When we multiply $\mathbf{X}'\mathbf{JX}$, we get:

$$\mathbf{X}'\mathbf{J}\begin{bmatrix} n & n & \cdots & n \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i & \cdots & \sum_{i=1}^n X_i \end{bmatrix} \quad \Rightarrow \quad \mathbf{X}'\mathbf{JX} = \begin{bmatrix} n^2 & n\sum_{i=1}^n X_i \\ n\sum_{i=1}^n X_i & \left(\sum_{i=1}^n X_i\right)^2 \end{bmatrix}$$

$$E\left\{\mathbf{Y}'\left(\frac{1}{n}\right)\mathbf{JY}\right\} = \text{trace}\left(\frac{1}{n}\mathbf{J}\boldsymbol{\Sigma}_{\mathbf{Y}}\right) + \mu_{\mathbf{Y}}'\left(\frac{1}{n}\right)\mathbf{J}\mu_{\mathbf{Y}} = \sigma^2\text{trace}\left(\frac{1}{n}\mathbf{J}\right) + \mu_{\mathbf{Y}}'\left(\frac{1}{n}\right)\mathbf{J}\mu_{\mathbf{Y}} = \sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\frac{1}{n}\mathbf{JX}\boldsymbol{\beta}$$

Now, consider the total, error and regression sums of squares:

$$TSS = \mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} \qquad SSE = \mathbf{Y}'\left(\mathbf{I} - \mathbf{P}\right)\mathbf{Y} \qquad SSR = \mathbf{Y}'\left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$$

Now, consider $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\left(\frac{1}{n}\right)\mathbf{J}\mathbf{X}$:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \quad \mathbf{X}'\left(\frac{1}{n}\right)\mathbf{J}\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \frac{\left(\sum_{i=1}^n X_i\right)^2}{n} \end{bmatrix}$$

$$\Rightarrow \quad \mathbf{X}'\mathbf{X} - \mathbf{X}'\left(\frac{1}{n}\right)\mathbf{J}\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 0 & \sum_{i=1}^n (X_i - \overline{X})^2 \end{bmatrix}$$

Now, we have:

$$E\{TSS\} = \left[n\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] - \left[\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\frac{1}{n}\mathbf{J}\mathbf{X}\boldsymbol{\beta}\right] = (n-1)\sigma^2 + \beta_1^2\sum_{i=1}^n (X_i - \overline{X})^2$$

$$E\{SSE\} = \left[n\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] - \left[2\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] = (n-2)\sigma^2$$

$$E\{SSR\} = \left[2\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] - \left[\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\frac{1}{n}\mathbf{J}\mathbf{X}\boldsymbol{\beta}\right] = \sigma^2 + \beta_1^2\sum_{i=1}^n (X_i - \overline{X})^2$$

Further, if $\mathbf{w}$ is normally distributed, and if $\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{W}}\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{W}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{W}}$ then we have the following results from **Cochran's Theorem**:

$$\frac{\mathbf{w}'\mathbf{A}\mathbf{w}}{\sigma^2} \sim \chi^2\left(df_A, \Omega_A\right) \qquad df_A = \text{ rank}(A) \qquad \Omega_A = \mu_{\mathbf{W}}'\mathbf{A}\mu_{\mathbf{W}}$$

where df_A and Ω_A are the **degrees of freedom** and **non-centrality parameter**, respectively. If $\Omega_A = 0$, then it is the standard (central) chi-square distribution. Two other important results are:

$$\mathbf{w}'\mathbf{A}\mathbf{w} \text{ and } \mathbf{w}'\mathbf{B}\mathbf{w} \text{ are independent if } \mathbf{A}\mathbf{V}\mathbf{B} = \mathbf{0}$$

$$\mathbf{w}'\mathbf{A}\mathbf{w} \text{ and } \mathbf{B}\mathbf{w} \text{ are independent if } \mathbf{B}\mathbf{V}\mathbf{A} = \mathbf{0}$$

Note that with respect to the model with normal, independent errors of constant variance, we have:

$$\mathbf{V} = \mathbf{I} \qquad (\mathbf{I}-\mathbf{P})\mathbf{I}(\mathbf{I}-\mathbf{P})\mathbf{I} = (\mathbf{I}-\mathbf{P})\mathbf{I} \qquad \left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\mathbf{I}\left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\mathbf{I} = \left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\mathbf{I} \qquad \left(\mathbf{P} - \frac{1}{n}\mathbf{J}\right)\mathbf{I}(\mathbf{I}-\mathbf{P}) = \mathbf{0}$$

This leads to the following important results:

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2, 0) \qquad \frac{SSR}{\sigma^2} \sim \chi^2\left(1, \beta_1^2\sum_{i=1}^n (X_i - \overline{X})^2\right)$$

Further, SSE and SSR are independent. Also $\hat{\boldsymbol{\beta}}$ and SSE are independent, since:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad SSE = \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y} \quad \Rightarrow \quad (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}(\mathbf{I}-\mathbf{P}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{0} \quad \mathbf{X}'\mathbf{P} = \mathbf{X}'$$

The ratio of two independent chi-square random variables, each divided by its degrees of freedom, is follows the F-distribution. If the numerator chi-square is non-central, and the denominator is a standard (central) chi-square, it follows a non-central distribution, with the non-centrality parameter of the numerator chi-square. If both are central, the ratio follows a standard F-distribution. Thus, since SSE and SSR are independent:

$$\frac{SSR}{\sigma^2} \sim \chi^2\left(1, \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2\right) \qquad \frac{SSE}{\sigma^2} \sim \chi^2(n-2, 0)$$

$$\Rightarrow \quad F = \frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-2)} = \frac{MSR}{MSE} \sim F\left(1, n-2, \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2\right)$$

When the null hypothesis $H_0 : \beta_1 = 0$, the F-statistic follows the standard (central) F-distribution.

Chapter 5

Model Diagnostics and Influence Measures

The inferences regarding the simple linear regression model (tests and confidence intervals) are based on the following assumptions:

- Relation between Y and X is linear
- Errors are normally distributed
- Errors have constant variance
- Errors are independent

These assumptions can be checked graphically, as well as by statistical tests.

5.1 Checking Linearity

A plot of the residuals versus X should be a random cloud of points centered at 0 (they sum to 0). A "U-shaped" or "inverted U-shaped" pattern is inconsistent with linearity.

A test for linearity can be conducted when there are repeat observations at certain X-levels (methods have also been developed to "group X values). Suppose we have c distinct X-levels, with n_j observations at the j^{th} level. The data need to be re-labeled as Y_{ij} where j represents the X group, and i represents the individual case within the group ($i = 1, \ldots, n_j$). We compute the following quantities:

$$\overline{Y}_j = \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j} \qquad \hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$$

We then decompose the Error Sum of Squares into **Pure Error** and **Lack of Fit**:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n_j}\sum_{j=1}^{c}\left(Y_{ij} - \overline{Y}_j\right)^2 + \sum_{j=1}^{c} n_j\left(\overline{Y}_j - \hat{Y}_j\right) \qquad SSE = SSPE + SSLF$$

We then partition the error degrees of freedom $(n-2)$ into Pure Error $(n-c)$ and Lack of Fit $(c-2)$. This leads to an F-test for testing H_0: Relation is Linear versus H_A: Relation is not Linear:

$$TS : F_{obs} = \frac{[SSLF/(c-2)]}{[SSPE/(n-c)]} = \frac{MSLF}{MSPE} \qquad RR : F_{obs} \geq F_{\alpha,c-2,n-c} \qquad P\text{-Value } : P\left(F_{c-2,n-c} \geq F_{obs}\right)$$

If the relationship is not linear, we can add polynomial terms to allow for ”bends” in the relationship between Y and X using multiple regression.

5.2 Checking Normality

A normal probability plot of the ordered residuals versus their predicted values should fall approximately on a straight line. A histogram should be mound-shaped. Neither of these methods work well with small samples (even data generated from a normal distribution will not necessarily look like it is normal).

Various tests are computed directly by statistical computing packages. The Shapiro-Wilk and Kolmogorov-Smirnov tests are commonly reported, reporting P-values for testing H_0: Errors are normally distributed.

When data are not normally distributed, the **Box-Cox transformation** is often applied to the data. This involves fitting regression models for various power transformations of Y on X, where:

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda}-1}{\lambda(\dot{Y})^{(\lambda-1)}} & \lambda \neq 0 \\ \dot{Y}\ln(Y_i) & \lambda = 0 \end{cases}$$

Here $\dot{Y}$ is the geometric mean of $Y_1, \ldots, Y_n$ are all strictly positive (a constant can be added to all observations to assure this).

$$\dot{Y} = \left(\prod_{i=1}^{n} Y_i\right)^{1/n} \quad = \quad \exp\left\{\frac{\sum_{i=1}^{n}\ln(Y_i)}{n}\right\}$$

Values of λ between -2 and 2 by 0.1 are typically run, and the value of λ that has the smallest Error Sum of Squares (equivalently Maximum Likelihood) is identified. Software packages will present a confidence interval for λ.

5.3 Checking Equal Variance

A plot of the residuals versus the fitted values should be a random cloud of points centered at 0. When the variances are unequal, the variance tends to increase with the mean, and we observe a funnel-type shape.

Two tests for equal variance are the Brown-Forsyth test and the Breusch-Pagan (aka Cook-Weisberg) test.

Brown-Forsyth Test - Splits data into two groups of approximately equal sample sizes based on their fitted values (any cases with the same fitted values should be in the same group). Then labeling the residuals $e_{11}, \ldots, e_{1n_1}$ and $e_{21}, \ldots, e_{2n_2}$, obtain the median residual for each group: $\tilde{e}_1$ and $\sim e_2$, respectively. Then compute the following:

$$d_{ij} = |e_i j - \tilde{e}_i| \quad i = 1, 2; j = 1, \ldots, n_i \qquad \overline{d}_i = \frac{\sum_{i=1}^{n_i} d_{ij}}{n_i} \qquad s_i^2 = \frac{\sum_{i=1}^{n_i} \left(d_{ij} - \overline{d}_i\right)^2}{n_i - 1} \qquad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Then, a 2-sample t-test is conducted to test H_0: Equal Variances in the 2 groups:

$$TS : t_{obs} = \frac{\overline{d}_1 - \overline{d}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad RR : |t_{obs}| \geq t_{\alpha/2, n-2} \qquad P\text{-value} = P\left(t_{n-2} \geq |t_{obs}|\right)$$

Breusch-Pagan Test(aka Cook-Weisberg Test) - Fits a regression of the squared residuals on X and tests whether the (natural) log of the variance is linearly related to X. When the regression of the squared residuals is fit, we obtain SSR_{e^2}, the regression sum of squares. The test is conducted as follows, where SSE is the Error Sum of Squares for the original regression of Y on X:

$$TS : X_{obs}^2 = \frac{(SSR_{e^2}/2)}{(SSE/n)^2} \qquad RR : X_{obs}^2 \geq \chi_{\alpha,1}^2 \qquad P\text{-value: } P\left(\chi_1^2 \geq X_{obs}^2\right)$$

When the variance is not constant, we can transform Y (often can use the Box-Cox transformation to obtain constant variance).

We can also use **Estimated Weighted Least Squares** by relating the standard deviation (or variance) of the errors to the mean. This is an iterative process, where the weights are re-weighted each iteration. The weights are the reciprocal of the estimated variance (as a function of the mean). Iteration continues until the regression coefficient estimates stabilize.

Another, simpler method is to obtain robust standard errors of the OLS estimators based on the residuals from the linear regression (using the squared residuals as estimates of the variances for the individual cases). This method was originally proposed by White (1980). The estimated variance-COVariance matrix (with resulting **robust to heteroskedasticity standard errors** for $\hat{\boldsymbol{\beta}}$ is:

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = (\mathbf{X'X})^{-1}\mathbf{X'\hat{E}_2X}(\mathbf{X'X})^{-1} \qquad \mathbf{\hat{E}_2} = \begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix}$$

When the distribution of Y is a from a known family (e.g. Binomial, Poisson, Gamma), we can fit a **Generalized Linear Model**.

5.4 Checking Independence

When the data are a time (or spatial) series, the errors can be correlated over time (or space), referred to as **autocorrelated**. A plot of residuals versus time should be random, not displaying a trending pattern (linear or cyclical). If it does show these patterns, autocorrelation may be present.

The Durbin-Watson test is used to test for serial autocorrelation in the errors, where the null hypothesis is that the errors are uncorrelated. Unfortunately, the formal test can end in one of 3 possible outcomes: reject H_0, accept H_0, or inconclusive. Statistical software packages can report an approximate P-value. The test is obtained as follows:

$$TS: DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad \text{Decision Rule: } DW < d_L \text{Reject } H_0 \quad DW > d_U \text{Accept } H_0 \quad \text{Otherwise Inconclusive}$$

where tables of d_L and d_U are in standard regression texts and posted on the internet. These values are indexed by the number of predictor variables (1, in the case of simple regression) and the sample size (n).

When errors are not independent, estimated standard errors of estimates tend to be too small, making t-statistics artificially large and confidence intervals artificially narrow.

The **Cochrane-Orcutt** method transforms the Y and X variables, and fits the model based on the transformed responses. Another approach is to use **Estimated Generalized Least Squares (EGLS)**. This uses the estimated COVariance structure of the observations to obtain estimates of the regression coefficients and their estimated standard errors.

5.5 Detecting Outliers and Influential Observations

These measures are widely used in multiple regression, as well, when there are p predictors, and $p' = p + 1$ parameters (including intercept, β_0). Many of the "rules of thumb" are based on p', which is 1+1=2 for simple regression. Most of these methods involve matrix algebra, but are obtained from statistical software packages. Their matrix forms are not given here (see references).

Also, many of these methods make use of the estimated variance when the i^{th} case was removed (to remove its effect if it is an outlier):

$$MSE_{(i)} = \frac{SSE_{(i)}}{n-p'-1} = \frac{SSE - e_i^2}{n-p'-1} \quad \text{for simple regression } p' = 2$$

Studentized Residuals - Residuals divided by their estimated standard error, with their contribution to SSE having been removed (see above). Since residuals have mean 0, the studentized residuals are like t-statistics. Since we are simultaneously checking whether n of these are outliers, we conclude any cases are outliers if the absolute value of their studentized residuals exceed $t_{\alpha/2n,n-p'-1}$, where p' is the number of independent variables plus one (for simple regression, p'=2).

Leverage Values (Hat Values) - These measure each case's potential to influence the regression due to its X levels. Cases with high leverage values (often denoted v_{ii} or h_{ii}) have X levels "away" from the center of the distribution. The leverage values sum to p' (2 for simple regression), and cases with leverage values greater than $2p'/n$ (twice the average) are considered to be potentially influential due to their X-levels.

DFFITS - These measure how much an individual case's fitted value shifts when it is included in the regression fit, and when it is excluded. The shift is divided by its standard error, so we are measuring how many standard errors a fitted value shifts, due to its being included in the regression model. Cases with the DFFITS values greater than $2\sqrt{p'/n}$ in absolute value are considered influential on their own fitted values.

DFBETAS - One of these is computed for each case, for each regression coefficient (including the intercept). DFBETAS measures how much the estimated regression coefficient shifts when that case is included and excluded from the model, in units of standard errors. Cases with DFBETAS values larger than $2/\sqrt{n}$ in absolute value are considered to be influential on the estimated regression coefficient.

Cook's D - Is a measure that represents each case's aggregate influence on all regression coefficients, and all cases' fitted values. Cases with Cook's D larger than $F_{.50,p',n-p'}$ are considered influential.

COVRATIO - This measures each case's influence on the estimated standard errors of the regression coefficients (inflating or deflating them). Cases with COVRATIO outside of $1 \pm 3p'/n$ are considered influential.

Chapter 6

Multiple Linear Regression

When there are more than one predictor variables, the model generalizes to multiple linear regression. The calculations become more complex, but conceptually, the ideas remain the same. We will use the notation of p as the number of predictors, and $p' = p + 1$ as the number of parameters in the model (including the intercept). The model can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad \epsilon \sim N(0, \sigma^2) \text{ independent}$$

We then obtain least squares (and maximum likelihood) estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize the error sum of squares. The fitted values, residuals, and error sum of squares are obtained as:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots \hat{\beta}_p X_{ip} \qquad e_i = Y_i - \hat{Y}_i \qquad SSE = \sum_{i=1}^{n} e_i^2$$

The degrees of freedom for error are now $n - p' = n - (p + 1)$, as we have now estimated $p' = p + 1$ parameters.

In the multiple linear regression model, β_j represents the change in $E\{Y\}$ when X_j increases by 1 unit, with all other predictor variables being held constant. It is thus often referred to as the **partial regression coefficient**.

6.1 Testing and Estimation for Partial Regression Coefficients

Once we fit the model, obtaining the estimated regression coefficients, we also obtain standard errors for each coefficient (actually, we obtain an estimated variance-COVariance matrix for the coefficients).

If we wish to test whether Y is associated with X_j, after controlling for the remaining $p-1$ predictors, we are testing whether $\beta_j = 0$. This is equivalent to the t-test from simple regression (in general, we can test whether a regression coefficient is any specific number, although software packages are testing whether it is 0):

$$H_0 : \beta_j = \beta_{j0} \qquad H_A : \beta_j \neq \beta_{j0} \quad TS : t_{obs} = \frac{\hat{\beta}_j - \beta_{j0}}{SE\{\hat{\beta}_j\}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-p'} \qquad P\text{-value} : P(t_{n-p'} \geq |t_{obs}|)$$

One-sided tests make the same adjustments as in simple linear regression:

$$H_A^+ : \beta_j > \beta_{j0} \qquad RR : t_{obs} \geq t_{\alpha,n-p'} \qquad P\text{-value} : P(t_{n-p'} \geq t_{obs})$$

$$H_A^- : \beta_j < \beta_{j0} \qquad RR : t_{obs} \leq -t_{\alpha,n-p'} \qquad P\text{-value} : P(t_{n-p'} \leq t_{obs})$$

A $(1-\alpha)100\%$ Confidence Interval for β_j is obtained as:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p'} SE\{\hat{\beta}_j\}$$

Note that the Confidence Interval represents the values of β_{j0} that the two-sided test: $H_0 : \beta_j = \beta_{j0} \quad H_A : \beta_j \neq \beta_{j0}$ fails to reject the null hypothesis.

6.2 Analysis of Variance

When there is no association between Y and $X_1, \ldots, X_p$ ($\beta_1 = \cdots = \beta_p = 0$), the best predictor of each observation is $\overline{Y} = \hat{\beta}_0$ (in terms of minimizing sum of squares of prediction errors). In this case, the total variation can be denoted as $TSS = \sum_{i=1}^n (Y_i - \overline{Y})^2$, the **Total Sum of Squares**, just as with simple regression.

When there is an association between Y and at least one of $X_1, \ldots, X_p$ (not all $\beta_i = 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$ (in terms of minimizing sum of squares of prediction errors). In this case, the error variation can be denoted as $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, the **Error Sum of Squares**.

The difference between TSS and SSE is the variation "explained" by the regression of Y on $X_1, \ldots, X_p$ (as opposed to having ignored $X_1, \ldots, X_p$). It represents the difference between the fitted values and the mean: $SSR = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$ the **Regression Sum of Squares**.

$$TSS = SSE + SSR \qquad \sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$$

Source	*df*	SS	MS	F_{obs}	P-value
Regression (Model)	p	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$MSR = \frac{SSR}{p}$	$F_{obs} = \frac{MSR}{MSE}$	$P(F_{p,n-p'} \geq F_{obs})$
Error (Residual)	$n-p'$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-p'}$		
Total (Corrected)	$n-1$	$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

Table 6.1: Analysis of Variance Table for Multiple Linear Regression

Each sum of squares has a **degrees of freedom** associated with it. The **Total Degrees of Freedom** is $df_{\text{Total}} = n-1$. The **Error Degrees of Freedom** is $df_{\text{Error}} = n-p'$. The **Regression Degrees of Freedom** is $df_{\text{Regression}} = p$. Note that when we have $p = 1$ predictor, this generalizes to simple regression.

$$df_{\text{Total}} = df_{\text{Error}} + df_{\text{Regression}} \qquad n-1 = n-p'+p$$

Error and Regression sums of squares have a **Mean Square**, which is the sum of squares divided by its corresponding degrees of freedom: $MSE = SSE/(n-p')$ and $MSR = SSR/p$. It can be shown that these mean squares have the following **Expected Values**, average values in repeated sampling at the same observed X levels:

$$E\{MSE\} = \sigma^2 \qquad\qquad E\{MSR\} \geq \sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\left(\mathbf{I}-\left(\frac{1}{n}\right)\mathbf{J}\right)\mathbf{X}\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ and $\mathbf{X}$ are matrix/vector extensions of the simple linear regression model (see below). Note that when $\beta_1 = \cdots \beta_p = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise $E\{MSR\} > E\{MSE\}$. A way of testing whether $\beta_1 = \cdots \beta_p = 0$ is by the F-test:

$$H_0 : \beta_1 = \cdots \beta_p = 0 \qquad H_A : \text{ Not all } \beta_j = 0$$

$$TS : F_{obs} = \frac{MSR}{MSE} \qquad RR : F_{obs} \geq F_{\alpha,p,n-p'} \qquad P\text{-value} : P(F_{p,n-p'} \geq F_{obs})$$

The Analysis of Variance is typically set up in a table as in Table **??**.

A measure often reported from a regression analysis is the **Coefficient of Determination** or R^2. This represents the variation in Y ”explained” by $X_1, \ldots, X_p$, divided by the total variation in Y.

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad 0 \leq R^2 \leq 1$$

The interpretation of R^2 is the proportion of variation in Y that is ”explained” by $X_1, \ldots, X_p$, and is often reported as a percentage ($100R^2$).

6.3 Testing a Subset of $\beta^s = 0$

The F-test from the Analysis of Variance and the t-tests represent extremes as far as model testing (all variables simultaneously versus one-at-a-time). Often we wish to test whether a group of predictors do not improve prediction, after controlling for the remaining predictors.

Suppose that after controlling for g predictors, we wish to test whether the remaining $p-g$ predictors are associated with Y. That is, we wish to test:

$$H_0 : \beta_{g+1} = \cdots \beta_p = 0 \qquad H_A : \text{ Not all of } \beta_{g+1}, \dots, \beta_p = 0$$

Note that, the t-tests control for all other predictors, while here, we want to control for only $X_1, \dots, X_g$. To do this, we fit two models: the **Complete** or **Full Model** with all p predictors, and the **Reduced Model** with only the g "control" variables. For each model, we obtain the Regression and Error sums of squares, as well as R^2. This leads to the test statistic and rejection region:

$$TS : F_obs = \frac{\left[\frac{SSE(R)-SSE(F)}{(n-g')-(n-p')}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{SSR(F)-SSR(R)}{p-g}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{R_F^2-R_R^2}{p-g}\right]}{\left[\frac{1-R_F^2}{n-p'}\right]}$$

$$RR : F_{obs} \geq F_{\alpha,p-g,n-p'} \qquad P\text{-value} : P(F_{p-g,n-p'} \geq F_{obs})$$

6.4 Matrix Form of Multiple Regression Model

The matrix form is virtually identical (at least symbolically) for multiple regression as simple regression. The primary difference is the dimension of the various matrices and vectors. Now, $\mathbf{X}$ still has n rows, but it hat $p+1$ columns (one for the intercept, and one each for the p predictors). The vectors $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ each have $p' = p+1$ rows.

We still have that the estimated variance of $\hat{\boldsymbol{\beta}} = s^2(\mathbf{X'X})^{-1}$ which is how the estimated standard errors for the partial regression coefficients used in t-tests and confidence intervals are obtained, in the case of the model with independent errors with constant variance.

The **general linear test** can be used to test any set of up to $p+1$ linear hypotheses among the β^s, that are linearly independent. The tests described above are special cases. Here we wish to test:

$$H0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m} \quad \Rightarrow \quad \mathbf{K}'\boldsymbol{\beta} - \mathbf{m} = \mathbf{0}$$

where $\mathbf{K}'$ is a $q \times (p+1)$ matrix of constants defining the the hypotheses among the $\boldsymbol{\beta}$ elements and $\mathbf{m}$ is the $q \times 1$ vector of hypothesized values for the q linear functions. Some special cases are given below, assuming $p = 3$ (three predictor variables):

$$H_{01} : \beta_1 = \beta_2 = \beta_3 = 0 \qquad \mathbf{K_1'} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{m_1} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$H_{02}: \beta_1 = \beta_2 = \beta_3 \qquad \mathbf{K_2'} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \qquad \mathbf{m_2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$H_{03}: \beta_0 = 100, \beta_1 = 10, \beta_2 = \beta_3 \qquad \mathbf{K_3'} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \qquad \mathbf{m_3} = \begin{bmatrix} 100 \\ 10 \\ 0 \end{bmatrix}$$

The estimator $\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}$ has an estimated variance-COVariance matrix of $s^2\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}$ which is $q \times q$. Then, we can form the F-statistic (based on assuming normal and independent errors with constant variance):

$$F_{obs} = \frac{\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)' \left(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}\right)^{-1} \left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right)}{qs^2}$$

which under the null hypothesis is distributed $F_{q,n-p'}$.

Note that even if the data are not normally distributed, the quantity qF_{obs} is asymptotically distributed as χ^2_q, so the test can be conducted in this manner in large samples. Note that in this large-sample case, the tests are identical as $F_{\alpha,q,\infty} = \frac{\chi^2_{\alpha,q}}{q}$.

6.5 Models With Categorical (Qualitative) Predictors

Often, one or more categorical variables are included in a model. If we have a categorical variable with m levels, we will need to create $m-1$ **dummy** or **indicator variables**. The variable will take on 1 if the i^{th} observation is in that level of the variable, 0 otherwise. Note that one level of the variable will have 0^s for all $m-1$ dummy variables, making it the reference group. The β^s for the other groups (levels of the qualitative variable) reflect the difference in the mean for that group with the reference group, controlling for all other predictors.

Note that if the qualitative variable has 2 levels, there will be a single dummy variable, and we can test for differences in the effects of the 2 levels with a t-test, controlling for all other predictors. If there are $m-1$ ¿ 2 dummy variables, we can use the F-test to test whether all $m-1$ β^s are 0, controlling for all other predictors.

6.6 Models With Interaction Terms

When the effect of one predictor depends on the level of another predictor (and vice versa), the predictors are said to **interact**. The way we can model interaction(s) is to create a new variable that is the product of the 2 predictors. Suppose we have Y, and 2 numeric predictors: X_1 and X_2. We create a new predictor $X_3 = X_1X_2$. Now, consider the model:

$$E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 = \beta_0 + \beta_2X_2 + (\beta_1 + \beta_3X_2)\,X_1$$

Thus, the slope with respect to X_1 depends on the level of X_2, unless $\beta_3 = 0$, which we can test with a t-test. This logic extends to qualitative variables as well. We create cross-product terms between numeric (or other categorical) predictors with the $m-1$ dummy variables representing the qualitative predictor. Then t-test ($m-1=1$) or F-test ($m-1>2$) can be conducted to test for interactions among predictors.

6.7 Models With Curvature

When a plot of Y versus one or more of the predictors displays curvature, we can include polynomial terms to "bend" the regression line. Often, to avoid multicollinearity, we work with centered predictor(s), by subtracting off their mean(s). If the data show k bends, we will include $k+1$ polynomial terms. Suppose we have a single predictor variable, with 2 "bends" appearing in a scatterplot. Then, we will include terms up to the a third order term. Note that even if lower order terms are not significant, when a higher order term is significant, we keep the lower order terms (unless there is some physical reason not to). We can now fit the model:

$$E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

If we wish to test whether the fit is linear, as opposed to "not linear," we could test $H_0 : \beta_2 = \beta_3 = 0$. In many instances it is preferable to center the data (subtract off the mean) or to center and scale the data (divide centered values by a scale constant) for ease of interpretation and reduce collinearity among the predictors.

Response surfaces are often fit when we have 2 or more predictors, and include "linear effects," "quadratic effects," and "interaction effects". In the case of 3 predictors, a full model would be of the form:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3$$

We typically wish to simplify the model, to make it more parsimonious, when possible.

6.8 Model Building

When we have many predictors, we may wish to use an algorithm to determine which variables to include in the model. These variables can be main effects, interactions, and polynomial terms. Note that there are two common approaches. One method involves testing variables based on t-tests, or equivalently F-tests for partial regression coefficients. An alternative method involves comparing models based on model based measures, such as Akaike Information Criterion (AIC), or Schwartz Bayesian Information criterion (BIC or SBC). These measures can be written as follows (note that different software packages print different versions, as some parts are constant for all potential models). The goal is to minimize the measures.

$$AIC(\text{Model}) = nln(SSE(\text{Model})) + 2p' - nln(n) \qquad BIC(\text{Model}) = nln(SSE(\text{Model})) + [ln(n)]p' - nln(n)$$

Note that SSE(Model) depends on the variables included in the current model. The measures put a penalty on excess predictor variables, with BIC placing a higher penalty when $ln(n) > 2$. Note that p' is the number of parameters in the model (including the intercept), and n is the sample size.

6.8.1 Backward Elimination

This is a "top-down" method, which begins with a "Complete" Model, with all potential predictors. The analyst then chooses a significance level to stay in the model (SLS). The model is fit, and the predictor with the lowest t-statistic in absolute value (largest P-value) is identified. If the P-value is larger than SLS, the variable is dropped from the model. Then the model is re-fit with all other predictors (this will change all regression coefficients, standard errors, and P-values). The process continues until all variables have P-values below SLS.

The model based approach fits the full model, with all predictors and computes AIC (or BIC). Then, each variable is dropped one-at-a-time, and AIC (or BIC) is obtained for each model. If none of the models with one dropped variable has AIC (or BIC) below that for the full model, the full model is kept, otherwise the model with the lowest AIC (or BIC) is kept as the new full model. The process continues until no variables should be dropped (none of the "drop one variable models" has a lower AIC (or BIC) than the "full model."

6.8.2 Forward Selection

This is a "bottom-up, which begins with all "Simple" Models, each with one predictor. The analyst then chooses a significance level to enter into the model (SLE). Each model is fit, and the predictor with the highest t-statistic in absolute value (smallest P-value) is identified. If the P-value is smaller than SLE, the variable is entered into the model. Then all two variable models including the best predictor in the first round, with each of the other predictors. The best second variable is identified, and its P-value is compared with SLE. If its P-value is below SLE, the variable is added to the model. The process continues until no potential added variables have P-values below SLE.

The model based approach fits each simple model, with one predictor and computes AIC (or BIC). The best variable is identified (assuming its AIC (or BIC) is smaller than that for the null model, with no predictors). Then, each potential variable is added one-at-a-time, and AIC (or BIC) is obtained for each model. If none of the models with one added variable has AIC (or BIC) below that for the best simple model, the simple model is kept, otherwise the model with the lowest AIC (or BIC) is kept as the new full model. The process continues until no variables should be added (none of the "add one variable models" has a lower AIC (or BIC) than the "reduced model."

6.8.3 Stepwise Regression

This approach is a hybrid of forward selection and backward elimination. It begins like forward selection, but then applies backward elimination at each step. In forward selection, once a variable is entered, it stays in the model. In stepwise regression, once a new variable is entered, all previously entered variables are tested, to confirm they should stay in the model, after controlling for the new entrant, as well as the other previous entrant.

6.8.4 All Possible Regressions

We can fit all possible regression models, and use model based measures to choose the "best" model. Commonly used measures are: Adjusted-R^2 (equivalently MSE), Mallow's C_p statistic, AIC, and BIC. The formulas, and decision criteria are given below (where p' is the number of parameters in the "current" model being fit:

Adjusted-R^2 - $1-\left(\frac{n-1}{n-p'}\right)\frac{SSE}{TSS}$ - Goal is to maximize

Mallow's C_p - $C_p=\frac{SSE(\text{Model})}{MSE(\text{Complete})}+2p'-n$ - Goal is to have $C_p \leq p'$

Akaike Information Criterion - $AIC(\text{Model})=nln(SSE(\text{Model}))+2p'-nln(n)$ - Goal is to minimize

Bayesian Information Criterion - $BIC(\text{Model})=nln(SSE(\text{Model}))+[ln(n)]p'-nln(n)$ - Goal is to minimize

6.9 Issues of Collinearity

When the predictor variables are highly correlated among themselves, the regression coefficients become unstable, with increased standard errors. This leads to smaller t-statistics for tests regarding the partial regression coefficients and wider confidence intervals. At its most extreme case, the sign of a regression coefficient can change when a new predictor variable is included. One widely reported measure of collinearity is the **Variance Inflation Factor (VIF)**. This is computed for each predictor variable, by regressing it on the remaining $p-2$ predictors. Then $VIF_J=\frac{1}{1-R_j^2}$ where R_j^2 is the coefficent of dettermination of the regression of X_j on the remaining predictors. Values of VIF_j greater than 10 are considered problematic, but if results are significant, it should not be problematic.

Various remedies exist. One is determining which variable(s) make the most sense theoretically for the model, and removing other variables, which are correlated with the other more meaningful predictors. A second method involves generating uncorrelated predictor variables from the original set of predictors. While this method based on **principal components** removes the collinearity problem, the new variables may lose their meaning, thus making it harder to describe the process. A third method **ridge regression** introduces a bias factor into the regression, that reduces the inflated variance due to collinearity, and through that reduces the Mean Square Error of the regression coefficients. Unfortunately, there is no simple rule on choosing the bias factor.

6.9.1 Principal Components Regression

For **principal components regression**, if we have p predictors: $X_1,\ldots,X_p$, we can generate p linearly independent predictors that are linear functions of $X_1,\ldots,X_p$. When the new variables with small eigenvalues are removed, the estimate of $\boldsymbol{\beta}$ obtained from the new regression is biased. The amount of bias depends on the relative size of the eigenvalues of the removed principal components, however the collinearity problem will be removed and the variance of the estimator will have been reduced. The process is conducted as follows (e.g. Rawlings, Pantula, and Dickey (1998), Section 13.2.2):

1. Create $Z_1, \ldots Z_p$ from the original variables $X_1, \ldots, X_p$ by subtracting the mean and dividing by a multiple of the standard deviation.

$$Z_{ij} = \frac{X_{ij} - \overline{X}_j}{\sqrt{n-1}s_j} \quad i = 1, ..., n; j = 1, ..., p \qquad \overline{X}_j = \frac{\sum_{i=1}^n X_{ij}}{n} \qquad s_j = \sqrt{\frac{\sum_{i=1}^n \left(X_{ij} - \overline{X}_j\right)^2}{n-1}}$$

2. Obtain the eigenvalues $\lambda_1, \ldots, \lambda_p$ (and place in a diagonal matrix $\mathbf{L}$) and eigenvectors (as columns in matrix $\mathbf{V}$) of the $p \times p$ matrix $\mathbf{R} = \mathbf{Z}'\mathbf{Z}$, where $\mathbf{R}$ is the correlation matrix among the predictor variables $X_1, \ldots, X_p$. These can be obtained in any matrix computer package ($\mathbf{Z}$ does not contain a column for an intercept).

$$\mathbf{R} = \mathbf{Z}'\mathbf{Z} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix} \qquad \mathbf{Z}'\mathbf{Z} = \mathbf{VLV}' = \sum_{i=1}^{n} \lambda_i \left(\mathbf{v_i}\mathbf{v_i'}\right)$$

3. Create the matrix of principal components $\mathbf{W} = \mathbf{ZV}$.

4. Fit the regression $\mathbf{Y} = \mathbf{W}\boldsymbol{\gamma}$ and obtain $SSR\left(\hat{\gamma}_j\right)$, the partial sum of squares for each generated predictor variable (principal component):

$$\hat{\boldsymbol{\gamma}} = (\mathbf{W}'\mathbf{W})^{-\mathbf{1}}\mathbf{W}'\mathbf{Y} \qquad \hat{V}\{\hat{\boldsymbol{\gamma}}\} = s^2(\mathbf{W}'\mathbf{W})^{-\mathbf{1}}$$

5. For each generated predictor, test $H_0 : \gamma_j = 0$, based on the t-test or F-test. Eliminate any principal components with high VIF and do not have significant coefficients.

6. Let $\hat{\boldsymbol{\gamma}}_{(g)}$ be the vector of retained coefficients from previous part. Then $SSR_{PC} = \sum SSR\left(\hat{\gamma}_j\right)$, with g degrees of freedom (the number of retained principal components (generated predictors)).

7. Scaling back to the original variables (in their standardized (mean=0, standard deviation=1) format), we get: $\hat{\boldsymbol{\beta}}_g^{PC} = \mathbf{V_{(g)}}\hat{\boldsymbol{\gamma}}_{(g)}$where $\mathbf{V_{(g)}}$ is the $p \times g$ portion of the eigenvector matrix (columns) corresponding to the retained principal components.

8. The estimated variance-COVariance matrix of $\hat{\boldsymbol{\beta}}_g^{PC}$ is:

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}_{\mathbf{g}}^{\mathbf{PC}}\right\} = s^2\mathbf{V_{(g)}}\mathbf{L}_{\mathbf{(g)}}^{-\mathbf{1}}\mathbf{V}'_{\mathbf{(g)}} \qquad \mathbf{L_{(g)}} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_g \end{bmatrix}$$

9. The fitted regression equation can be written with respect to the original (standardized) variables, or the principal components:

$$\hat{\mathbf{Y}}_{(g)} = \overline{\mathbf{Y}} + \mathbf{Z}\hat{\boldsymbol{\beta}}_g^{PC} \qquad \hat{\mathbf{Y}}_{(g)} = \overline{\mathbf{Y}} + \mathbf{W}\hat{\boldsymbol{\gamma}}_{(g)}$$

Note that the bias in the estimation comes when we reduce the principal components regression from p to g components. If the model reflects no need to remove any principal components, the estimator is unbiased. The bias comes from the fact that we have:

$$\hat{\boldsymbol{\beta}}_g^{PC} = \mathbf{V}_{(\mathbf{g})}\mathbf{V}'_{(\mathbf{g})}\hat{\boldsymbol{\beta}} \quad \Rightarrow \quad E\left\{\hat{\boldsymbol{\beta}}_g^{PC}\right\} = \mathbf{V}_{(\mathbf{g})}\mathbf{V}'_{(\mathbf{g})}\boldsymbol{\beta} \qquad \text{Note: } \mathbf{V}_{(\mathbf{p})}\mathbf{V}'_{(\mathbf{p})} = \mathbf{I}$$

6.9.2 Ridge Regression

In **Ridge Regression**, a biased estimator is directly induced that reduces its variance and mean square error (variance + squared bias). Unfortunately, the bias-inducing constant varies among applications, so it must be selected comparing results over various possible levels. We begin with a **standardized regression model** with no bias, based on the $p \times p$ correlation matrix among the predictors $\mathbf{R_{XX}}$ and the $p \times 1$ vector of correlations $\mathbf{R_{YX}}$ between the predictors and response variable.

$$\mathbf{R_{XX}} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix} \qquad \mathbf{R_{YX}} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Yp} \end{bmatrix}$$

The estimated standardized regression coefficients are obtained as:

$$\mathbf{R_{XX}}\hat{\boldsymbol{\beta}}^* = \mathbf{R_{XY}} \qquad \Rightarrow \qquad \hat{\boldsymbol{\beta}}^* = \mathbf{R_{XX}^{-1}}\mathbf{R_{YX}}$$

The standardized regression coefficients $\hat{\boldsymbol{\beta}}^*$ measure the change in Y in standard deviation units as each predictor increases by 1 standard deviation, thus removing the effects of scaling each predictor. It can also be obtained by transforming each X and Y by the following transformations:

$$X_{ij}^* = \frac{X_{ij} - \overline{X}_j}{\sqrt{n-1}s_j} \qquad Y_i^* = \frac{Y_i - \overline{Y}}{\sqrt{n-1}s_Y}$$

In matrix form, we have:

$$\mathbf{X}^* = \begin{bmatrix} X_{11}^* & X_{12}^* & \cdots & X_{1p}^* \\ X_{21}^* & X_{22}^* & \cdots & X_{2p}^* \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1}^* & X_{n2}^* & \cdots & X_{np}^* \end{bmatrix} \qquad \mathbf{Y}^* = \begin{bmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_n^* \end{bmatrix} \qquad \hat{\boldsymbol{\beta}}^* = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{\mathbf{-1}}\mathbf{X}^{*'}\mathbf{Y}^*$$

Note that $\mathbf{R_{XX}} = \mathbf{X}^{*'}\mathbf{X}^*$ and $\mathbf{R_{YX}} = \mathbf{X}^{*'}\mathbf{Y}^*$ The standardized ridge estimator is obtained as follows (e.g. Kutner, Nachtsheim, Neter, and Li (2005), Section 11.2):

$$\hat{\boldsymbol{\beta}}^{RR} = (\mathbf{R_{XX}}+c\mathbf{I})^{-1}\mathbf{R_{YX}} = \left(\mathbf{X}^{*'}\mathbf{X}^* + c\mathbf{I}\right)^{-1}\mathbf{X}^{*'}\mathbf{Y}^*$$

A **ridge trace plot** of the regression coefficients (vertical axis) versus c (horizontal axis) leads to a choice of c, where the coefficients stabilize or "flatten out." The fitted regression equation in transformed scale is:

$$\hat{\mathbf{Y}}^* = \mathbf{X}^*\hat{\boldsymbol{\beta}}^{RR} \qquad \Rightarrow \qquad \hat{Y}_i^* = \hat{\beta}_1^{RR}X_{i1}^* + \cdots + \hat{\beta}_p^{RR}X_{ip}^*$$

In terms of the originally scaled response, we have $\hat{\beta}_j = (s_Y/s_j)\hat{\beta}_j^{RR}$ and $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}_1 - \cdots - \hat{\beta}_p\overline{X}_p$.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_{i1} + \cdots + \hat{\beta}_pX_{ip}$$

6.10 Models with Unequal Variances (Heteroskedasticity)

When the data are independent, but with unequal variances, we can use (estimated) **Weighted Least Squares**. In rare occasions, the variances are known, and they will be used directly. One setting where this occurs in practice is when the "data" are averages among a group of units with common X levels. If each individual unit is independent with constant variance σ^2, the average of the m_i units (Y_i in this setting) has variance $V\left\{Y_i\right\} = \sigma^2/m_i$. In this case, we would use the reciprocal of the variance as the weight for each case (observations based on larger sample sizes have smaller variances and larger weights).

$$\mathbf{W} = \mathbf{\Sigma_Y^{-1}} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix} = \begin{bmatrix} \frac{m_1}{\sigma^2} & 0 & \cdots & 0 \\ 0 & \frac{m_2}{\sigma^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{m_n}{\sigma^2} \end{bmatrix} \qquad \hat{\boldsymbol{\beta}}^W = \left(\mathbf{X'WX}\right)^{\mathbf{-1}}\mathbf{X'WY}$$

The variance of the least squares estimator is obtained as follows:

$$V\left\{\hat{\boldsymbol{\beta}}^W\right\} = \left(\mathbf{X'WX}\right)^{\mathbf{-1}}\mathbf{X'W\Sigma_Y WX}\left(\mathbf{X'WX}\right)^{\mathbf{-1}} = \left(\mathbf{X'WX}\right)^{\mathbf{-1}}$$

In the case with $V\left\{Y_i\right\} = \sigma^2/m_i$, we can estimate σ^2 based on weighted mean square error:

$$MSE_W = \frac{\sum_{i=1}^n m_i\left(Y_i - \hat{Y}_i\right)^2}{n-p'} \qquad \hat{Y}_i = \hat{\beta}_0^W + \hat{\beta}_1^W X_{i1} + \cdots + \hat{\beta}_p^W X_{ip}$$

In this case (where data are averages):

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}^W\right\}\left(\mathbf{X'\hat{W}X}\right)^{-1} \qquad \mathbf{\hat{W}} = \begin{bmatrix} \frac{m_1}{MSE_W} & 0 & \cdots & 0 \\ 0 & \frac{m_2}{MSE_W} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{m_n}{MSE_W} \end{bmatrix}$$

Note that weighted least squares can be conducted as ordinary least squares on transformed bfX and bfY, which makes it possible to conduct using EXCEL, and non-statistical computing packages:

$$\mathbf{X^*} = \mathbf{W^{1/2}X} \qquad \mathbf{Y^*} = \mathbf{W^{1/2}Y} \qquad \hat{\boldsymbol{\beta}}^W = \left(\mathbf{X^{*\prime}X^*}\right)^{\mathbf{-1}}\mathbf{X^{*\prime}Y^*}$$

where $\mathbf{W^{1/2}}$ is the (diagonal) matrix with elements equal to the square roots of the elements of $\mathbf{W}$.

In most cases, the variances are unknown, and must be estimated. In this case, the squared residuals (variance) or absolute residuals (standard deviation) are regressed against one or more of the predictor variables or the mean (fitted values). The process is iterative. We begin by fitting ordinary least squares, obtaining the residuals, then regressing the squared or absolute residuals on the the predictor variables or fitted values, leading to (assuming all p predictors are used in the residual regression):

Variance Function Case: $\hat{v}_i = \hat{\delta}_0 + \hat{\delta}_1 X_{i1} + \cdots \hat{\delta}_p X_{ip}$ Standard Deviation Case: $\hat{s}_i = \hat{\delta}_0 + \hat{\delta}_1 X_{i1} + \cdots \hat{\delta}_p X_{ip}$

Once the estimated variance (standard deviation) is obtained for each case, we get the estimated weights:

$$\text{Variance Function Case: } \hat{w}_i = \frac{1}{\hat{v}_i} \qquad \text{Standard Deviation Case: } \hat{w}_i = \frac{1}{\hat{s}_i^2}$$

Then we compute the estimated weighted least squares estimator as:

$$\hat{\boldsymbol{\beta}}^{\hat{W}} = \left(\mathbf{X'\hat{W}X}\right)^{\mathbf{-1}} \mathbf{X'\hat{W}Y} \qquad \mathbf{\hat{W}} = \begin{bmatrix} \hat{w}_1 & 0 & \cdots & 0 \\ 0 & \hat{w}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{w}_n \end{bmatrix}$$

The process is continued until the estimated regression coefficients are stable from iteration to iteration. The estimated variance is:

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}^{\hat{W}}\right\} = \left(\mathbf{X'\hat{W}X}\right)^{\mathbf{-1}}$$

When the data collection process is based on a well designed controlled experiment, with multiple cases for each set of X levels, the variance of the errors can be estimated within each distinct group, and used in the estimated weighted least squares equation.

Another, simpler method is to obtain robust standard errors of the ordinary least squares (OLS) estimators based on the residuals from the linear regression (using the squared residuals as estimates of the variances for the individual cases). This method was originally proposed by White (1980). The estimated variance-covariance matrix with resulting **robust to heteroskedasticity standard errors** for $\hat{\boldsymbol{\beta}}$ is:

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = \mathbf{(X'X)^{-1}X'\hat{E}_2X(X'X)^{-1}} \qquad \mathbf{\hat{E}_2} = \begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix}$$

6.10.1 Bootstrap Methods When Distribution of Errors is Unknown

The bootstrap is widely used in many applications when the distribution of the data is unknown, or when the distribution of the estimator of is unknown. In regression applications, there are various ways of bootstrapping (see e.g. Cameron and Trivedi (2009, Chapter 13) and Kutner, et al (2005, Section 11.5)). All sampling is done with replacement (except the parametric bootstrap).

One possibility is to bootstrap the individual cases from the dataset, and repeatedly re-fit the regression, and saving the regression coefficients, obtaining their standard error. Then we can construct Confidence Intervals for the regression coefficients by taking the original estimate and adding and subtracting 2 standard errors for approximate 95% Confidence Intervals. This method is widely used when the X levels are random (not set up by the experimenter), and when the errors may not have constant variance. Also, the cut-off values for the middle $(1-\alpha)100\%$ of the bootstrap estimates can be used.

Another possibility that is useful is to retain the fitted values from the original regression $\hat{Y}_1, \ldots, \hat{Y}_n$ and bootstrap the residuals $e_1, \ldots, e_n$. Then the bootstrapped residuals are added to the original fitted values and the regression coefficients are obtained, and their standard error is computed and used as above.

The reflection method (see e.g. Kutner, et al (2005, Section 11.5)). In this case, we obtain the lower $\alpha/2$ percentile of the bootstrapped regression coefficients $\left(\hat{\beta}_j^*(\alpha/2)\right)$ and the upper $1-\alpha/2$ percentile of the regression coefficients $\left(\hat{\beta}_j^*(1-\alpha/2)\right)$ and obtain the interval:

$$\hat{\beta}_j - \hat{\beta}_j^*(\alpha/2) \leq \beta_j \leq \hat{\beta}_j^*(1-\alpha/2) - \hat{\beta}_j \quad j = 0,1,\ldots,p$$

In the parametric bootstrap, the residuals are sampled from a specified distribution with parameter(s) estimated from the original sample.

There are various bias-corrected methods applied as well.

6.11 Generalized Least Squares for Correlated Errors

Typically when data are collected over time and/or space, the errors are correlated, with correlation tending to be higher among observations that are closer in time or space. In this case, the variance-covariance matrix of the error terms is written generally:

$$V\{\boldsymbol{\varepsilon}\} = \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$$

For the case where the observations are equally spaced in time, and the error terms form an autoregressive process of order 1, we have:

$$\epsilon_t = \nu_t + \rho\epsilon_{t-1} \quad -1 < \rho < 1 \qquad \nu_t \sim iid\left(0,\sigma^2\right) \qquad \{\epsilon\} \perp \{\nu\}$$

Note that this autoregressive process can extend back to q time points, but the covariance structure gets more complicated. If we start with the definition that $E\{\epsilon_1\} = 0$ and $V\{\epsilon_1\} = \frac{\sigma^2}{1-\rho^2}$, we obtain:

$$E\{\epsilon_2\} = E\{\nu_2 + \rho\epsilon_1\} = 0$$

$$V\{\epsilon_2\} = V\{\nu_2 + \rho\epsilon_1\} = V\{\nu_2\} + V\{\rho\epsilon_1\} + 2\text{COV}\{\nu_2, \rho\epsilon_1\} = \sigma^2 + \frac{\rho^2\sigma^2}{1-\rho^2} + 2(0) = \frac{\sigma^2}{1-\rho^2}$$

$$\text{COV}\{\epsilon_1,\epsilon_2\} = \text{COV}\{\epsilon_1, \nu_2 + \rho\epsilon_1\} = \frac{\rho\sigma^2}{1-\rho^2}$$

In general, this extends to the following general results:

$$V\{\epsilon_t\} = \frac{\sigma^2}{1-\rho^2} = \gamma(0) \qquad \text{COV}\{\epsilon_t,\epsilon_{t+k}\} = \frac{\rho^{|k|}\sigma^2}{1-\rho^2} = \gamma(k) \qquad \rho = \frac{\text{COV}\{\epsilon_t,\epsilon_{t+1}\}}{V\{\epsilon_t\}} = \frac{\gamma(1)}{\gamma(0)}$$

$$V\{\boldsymbol{\varepsilon}\} = \boldsymbol{\Sigma} = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix}$$

If ρ were known, then we could use **Generalized Least Squares** to estimate $\boldsymbol{\beta}$. Let $\boldsymbol{\Sigma} = \sigma^2\mathbf{W}$. Then we would have:

$$\hat{\boldsymbol{\beta}}^{\mathbf{GLS}} = \left(\mathbf{X'W^{-1}X}\right)^{-1}\mathbf{X'W^{-1}Y} \qquad s^2 = \frac{1}{n-p'}\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathbf{GLS}}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathbf{GLS}}\right) \qquad \hat{V}\left\{\hat{\boldsymbol{\beta}}^{\mathbf{GLS}}\right\} = s^2\left(\mathbf{X'W^{-1}X}\right)^{-1}$$

In practice, ρ will be unknown, and can be estimated from the data. Further, if we make the following transformations for $AR(1)$ errors, the transformed response will have uncorrelated errors:

$$\mathbf{Y}^* = \mathbf{T^{-1}Y} \qquad \mathbf{X}^* = \mathbf{T^{-1}X} \qquad \mathbf{T^{-1}} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}$$

For this model, the transformed $\mathbf{Y}^*$, has the following model and variance-covariance structure:

$$\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{T^{-1}}\boldsymbol{\varepsilon} \qquad V\{\mathbf{Y}^*\} = \mathbf{T^{-1}}\boldsymbol{\sigma^2}\mathbf{W}\mathbf{T^{-1}}' = \boldsymbol{\sigma^2}\mathbf{I}$$

Then for **Estimated Generalized Least Squares (EGLS)** also known as **Feasible Generalized Least Squares (FGLS)**, we obtain estimates of ρ and σ^2 based on the residuals from the OLS regression, then re-fit the EGLS model.

$$\hat{\gamma}(0) = \frac{\sum_{t=1}^{n} e_t^2}{n} \qquad \hat{\gamma}(1) = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{n} \qquad \hat{\rho} = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} \qquad \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\rho}\hat{\gamma}(1)$$

Next, we obtain the "estimated" transformation matrix, and the transformed $\mathbf{Y}^*$ and $\mathbf{X}^*$:

$$\mathbf{Y}^* = \hat{\mathbf{T}}^{\mathbf{-1}}\mathbf{Y} \qquad \mathbf{X}^* = \hat{\mathbf{T}}^{\mathbf{-1}}\mathbf{X} \qquad \hat{\mathbf{T}}^{\mathbf{-1}} = \begin{bmatrix} \sqrt{1-\hat{\rho}^2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\hat{\rho} & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\hat{\rho} & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\hat{\rho} & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\hat{\rho} & 1 \end{bmatrix}$$

Note that the transformed responses have the following pattern (the predictors (and intercept) will have a similar structure):

$$Y_1^* = \sqrt{1-\hat{\rho}^2}Y_1 \qquad Y_t^* = Y_t - \hat{\rho}Y_{t-1} \quad t = 2,\ldots,n$$

The EGLS estimator, its variance-covariance matrix, and the estimator for σ^2 are obtained as follow:

$$\hat{\boldsymbol{\beta}}^{\mathbf{EGLS}} = \left(\mathbf{X}'\hat{\mathbf{T}}^{-1'}\hat{\mathbf{T}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{T}}^{-1'}\hat{\mathbf{T}}^{-1}\mathbf{Y} = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{Y}^*$$

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}^{\mathbf{EGLS}}\right\} = \hat{\sigma}_e^2\left(\mathbf{X}'\hat{\mathbf{T}}^{-1'}\hat{\mathbf{T}}^{-1}\mathbf{X}\right)^{-1} = \hat{\sigma}_e^2\left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}$$

$$\hat{\sigma}_e^2 = \frac{\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathbf{EGLS}}\right)'\hat{\mathbf{T}}^{-1'}\hat{\mathbf{T}}^{-1}\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathbf{EGLS}}\right)}{n-p'-1} = \frac{\left(\mathbf{Y}^*-\mathbf{X}^*\hat{\boldsymbol{\beta}}^{\mathbf{EGLS}}\right)'\left(\mathbf{Y}^*-\mathbf{X}^*\hat{\boldsymbol{\beta}}^{\mathbf{EGLS}}\right)}{n-p'-1}$$

Tests and Confidence Intervals for regression coefficients are obtained as follow:

$$H_0 : \beta_j = 0 \quad H_A : \beta_j \neq 0 \qquad TS : t_{obs} = \frac{\hat{\beta}_j^{EGLS}}{SE\left\{\hat{\beta}_j^{EGLS}\right\}} \qquad RR : |t_{obs}| \geq t\left(n - p' - 1, \alpha/2\right)$$

$$(1 - \alpha)100\% \text{ Confidence Interval: } \hat{\beta}_j^{EGLS} \pm t\left(n - p' - 1, \alpha/2\right) SE\left\{\hat{\beta}_j^{EGLS}\right\}$$

A test for the autoregressive parameter ρ is obtained as follows:

$$s^2 = \frac{\hat{\gamma}(0) - \hat{\rho}\hat{\gamma}(1)}{n - p' - 1} \qquad SE\left\{\hat{\rho}\right\} = \sqrt{\frac{s^2}{\hat{\gamma}(0)}} \qquad t_{obs} = \frac{\hat{\rho}}{SE\left\{\hat{\rho}\right\}}$$

Chapter 7

Nonlinear Regression

Often theory leads to a relationship between the response and the predictor variable(s) to be nonlinear, based on differential equations. While polynomial regression models allow for bends, these models are nonlinear with respect to the parameters. Many models with multiplicative errors can be transformed to be linear models. For instance:

$$Y = \beta_0 X^{\beta_1}\epsilon \quad E\{\epsilon\} = 1 \quad \Rightarrow \quad \ln(Y) = \ln(\beta_0) + \beta_1 \ln(X) + \ln(\epsilon) \quad \Rightarrow \quad Y^* = \beta_0^* + \beta_1 X^* + \epsilon^*$$

If the error term had been additive (with mean=0), the linearizing transformation would not have been possible, and a nonlinear regression would need to have been fitted. Consider the relationship:

$$Y_i = g\left(\mathbf{x_i'}, \boldsymbol{\beta}\right) + \epsilon_i \qquad \epsilon_i \sim NID\left(0, \sigma^2\right)$$

for some nonlinear function g (noting that linear regression ends up being a special case of this method). In matrix form, we have:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{g}\left(\boldsymbol{\beta}\right) = \begin{bmatrix} g\left(\mathbf{x_1'}, \boldsymbol{\beta}\right) \\ g\left(\mathbf{x_2'}, \boldsymbol{\beta}\right) \\ \vdots \\ g\left(\mathbf{x_n'}, \boldsymbol{\beta}\right) \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \qquad \mathbf{Y} = \mathbf{g}\left(\boldsymbol{\beta}\right) + \epsilon$$

Then by nonlinear least squares (NLS), we wish to estimate $\boldsymbol{\beta}$.

$$Q = \sum_{i=1}^{n} \left[Y_i - g\left(\mathbf{x_i'}, \boldsymbol{\beta}\right)\right]^2 = \left(\mathbf{Y} - \mathbf{g}\left(\boldsymbol{\beta}\right)\right)'\left(\mathbf{Y} - \mathbf{g}\left(\boldsymbol{\beta}\right)\right) \qquad \frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\sum_{i=1}^{n} \left[Y_i - g\left(\mathbf{x_i'}, \beta\right)\right]\left(\frac{\partial g\left(\mathbf{x_i'}, \beta\right)}{\partial \boldsymbol{\beta}}\right)$$

Note that for linear regression, $\frac{\partial g\left(\mathbf{x_i'}, \beta\right)}{\partial \boldsymbol{\beta}} = \mathbf{x_i'}$. Defining the matrix $\mathbf{G}\left(\boldsymbol{\beta}\right)$ as follows, we can obtain the NLS iterative algorithm.

$$\mathbf{G}\left(\boldsymbol{\beta}\right) = \begin{bmatrix} \frac{\partial g\left(\mathbf{x_1'}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}_1} & \frac{\partial g\left(\mathbf{x_1'}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial g\left(\mathbf{x_1'}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}_p} \\ \frac{\partial g\left(\mathbf{x_2'}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}_1} & \frac{\partial g\left(\mathbf{x_2'}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial g\left(\mathbf{x_2'}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g\left(\mathbf{x_n'}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}_1} & \frac{\partial g\left(\mathbf{x_n'}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial g\left(\mathbf{x_n'}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}_p} \end{bmatrix} \qquad \text{where} \qquad \mathbf{x_i'} = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

The Gauss-Newton algorithm can be used to obtain the values $\hat{\boldsymbol{\beta}}$ that minimize Q by setting $\frac{\partial Q}{\partial \boldsymbol{\beta}} = \mathbf{0}$:

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\sum_{i=1}^{n}\left[Y_i - g\left(\mathbf{x_i'}, \beta\right)\right]\left(\frac{\partial g\left(\mathbf{x_i'}, \beta\right)}{\partial \boldsymbol{\beta}}\right) = -2\left[\mathbf{Y} - \mathbf{g}\left(\boldsymbol{\beta}\right)\right]' \mathbf{G}\left(\boldsymbol{\beta}\right) = \left[\begin{array}{cccc} 0 & 0 & \cdots & 0 \end{array}\right]$$

The algorithm begins with setting starting values $\boldsymbol{\beta}^0$, and iterating to convergence (which can be difficult with poor starting values):

$$\hat{\boldsymbol{\beta}}^{(1)} = \boldsymbol{\beta}^0 + \left(\mathbf{G}\left(\boldsymbol{\beta^0}\right)' \mathbf{G}\left(\boldsymbol{\beta^0}\right)\right)^{-1} \mathbf{G}\left(\boldsymbol{\beta^0}\right)' \left[\mathbf{Y} - \mathbf{g}\left(\boldsymbol{\beta^0}\right)\right]$$

At the second round $\boldsymbol{\beta}^0$ is replaced by $\hat{\boldsymbol{\beta}}^{(1)}$, and we obtain $\hat{\boldsymbol{\beta}}^{(2)}$. Then iterate to convergence (hopefully).

All of the distributional arguments given below are based on large sample asymtotics, however simulation results have shown that in small samples, tests generally work well. For more information on nonlinear regression models, see e.g. (Gallant (1987), and Rawlings, Pantula, Dickey (2001, Chapter 15)). When the errors are independent and normally distributed with equal variances $\left(\sigma^2\right)$, the estimator $\hat{\boldsymbol{\beta}}$ is approximately Normal, with:

$$E\left\{\hat{\boldsymbol{\beta}}\right\} = \boldsymbol{\beta} \qquad V\left\{\hat{\boldsymbol{\beta}}\right\} = \sigma^2(\mathbf{G'G})^{-\mathbf{1}} \qquad \hat{\boldsymbol{\beta}} \overset{\text{approx}}{\sim} N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{G'G})^{-\mathbf{1}}\right)$$

The estimated variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is:

$$\hat{V}\left\{\hat{\boldsymbol{\beta}}\right\} = s^2\left(\mathbf{\hat{G}'\hat{G}}\right)^{-\mathbf{1}} = \mathbf{\hat{S}}\hat{\rho}\mathbf{\hat{S}} \qquad \hat{G} = G\left(\hat{\boldsymbol{\beta}}\right) \qquad s^2 = \frac{\left(\mathbf{Y} - \mathbf{g}\left(\hat{\boldsymbol{\beta}}\right)\right)'\left(\mathbf{Y} - \mathbf{g}\left(\hat{\boldsymbol{\beta}}\right)\right)}{n-p}$$

where $\mathbf{\hat{S}}$ is the diagonal matrix of estimated standard errors of the elements of $\hat{\boldsymbol{\beta}}$, and $\hat{\rho}$ is the matrix of correlations of the elements of $\hat{\boldsymbol{\beta}}$, which are printed out in various software packages. Estimates (Confidence Intervals) and tests for the regression coefficients can be conducted (approximately) based on the t-distribution as in linear regression.

$$(1-\alpha)100\% \text{ CI for } \beta_j : \hat{\beta}_j \pm t\left(\alpha/2, n-p\right) SE\left\{\hat{\beta}_j\right\} \qquad \frac{\hat{\beta}_j - \beta_{j0}}{SE\left\{\hat{\beta}_j\right\}} \overset{\text{approx}}{\sim} t(n-p) \text{ Under } H_0 : \beta_j = \beta_{j0}$$

For **linear** functions of $\boldsymbol{\beta}$ of the form $\mathbf{K'}\boldsymbol{\beta}$, we then have approximate normality of the estimator $\mathbf{K'}\hat{\boldsymbol{\beta}}$:

$$\mathbf{K'}\hat{\boldsymbol{\beta}} \overset{\text{approx}}{\sim} N\left(\mathbf{K'}\boldsymbol{\beta}, \sigma^2\mathbf{K'}\left(\mathbf{G'G}\right)^{-\mathbf{1}}\mathbf{K}\right)$$

Thus, to test $H_0 : \mathbf{K'}\boldsymbol{\beta} = \mathbf{m}$, where $\mathbf{K'}$ has q linearly independent rows (restrictions on the regression coefficients), we have the following test statistic and rejection region:

$$TS : F_{obs} = \frac{\left(\mathbf{K'}\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{K'}\left(\mathbf{\hat{G}'\hat{G}}\right)^{-\mathbf{1}}\mathbf{K}\right]^{-\mathbf{1}}\left(\mathbf{K'}\hat{\boldsymbol{\beta}} - \mathbf{m}\right)}{qs^2} \qquad RR : F_{obs} \geq F(q, n-p)$$

with a P-value as the area above the test statistic F_{obs} for the $F(q, n-p)$ distribution.

By the nature of nonlinear models, we may also be interested in **nonlinear** functions of the parameters, say $h\left(\boldsymbol{\beta}\right)$, which cannot be written in the form $\mathbf{K'}\hat{\boldsymbol{\beta}}$. In this case, the estimator $h\left(\hat{\boldsymbol{\beta}}\right)$ is approximately normally distributed:

$$h\left(\hat{\boldsymbol{\beta}}\right) \overset{\text{approx}}{\sim} N\left(h\left(\boldsymbol{\beta}\right), \sigma^2\left(\mathbf{H}\left(\mathbf{G'G}\right)^{-\mathbf{1}}\mathbf{H'}\right)\right)$$

where

$$\mathbf{H}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \end{bmatrix}$$

The estimated variance of $h\left(\hat{\boldsymbol{\beta}}\right)$ replaces both $\mathbf{H}$ and $\mathbf{G}$ with their estimates, replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$. Estimates (Confidence Intervals) and tests concerning $h(\boldsymbol{\beta})$ can be obtained as follow:

$$(1-\alpha)100\% \text{ CI for } h(\boldsymbol{\beta}) : h\left(\hat{\boldsymbol{\beta}}\right) \pm t(\alpha/2, n-p)\, SE\left\{h\left(\hat{\boldsymbol{\beta}}\right)\right\} \qquad \frac{h\left(\hat{\boldsymbol{\beta}}\right) - h_0}{SE\left\{h\left(\hat{\boldsymbol{\beta}}\right)\right\}} \overset{\text{approx}}{\sim} t(n-p) \text{ Under } H_0 : h(\boldsymbol{\beta}) = h_0$$

where:

$$SE\left\{h\left(\hat{\boldsymbol{\beta}}\right)\right\} = \sqrt{s^2 \hat{\mathbf{H}} \left(\hat{\mathbf{G}}'\hat{\mathbf{G}}\right)^{-\mathbf{1}} \hat{\mathbf{H}}'}$$

When there are several (q) nonlinear functions, an approximate Wald test of $\mathbf{h}(\boldsymbol{\beta}) = \mathbf{h_0}$ is:

$$TS : F_{obs} = \frac{\left(\mathbf{h}\left(\hat{\boldsymbol{\beta}}\right) - \mathbf{h_0}\right)' \left[\hat{\mathbf{H}} \left(\hat{\mathbf{G}}'\hat{\mathbf{G}}\right)^{-\mathbf{1}} \hat{\mathbf{H}}'\right]^{-\mathbf{1}} \left(\mathbf{h}\left(\hat{\boldsymbol{\beta}}\right) - \mathbf{h_0}\right)}{qs^2} \qquad RR : W \geq F(q, n-p)$$

with P-value being the upper-tail area above F_{obs} for the $F(q, n-p)$ distribution. Here, we define:

$$\mathbf{h}(\boldsymbol{\beta}) = \begin{bmatrix} h_1(\boldsymbol{\beta}) \\ h_2(\boldsymbol{\beta}) \\ \vdots \\ h_q(\boldsymbol{\beta}) \end{bmatrix} \qquad \mathbf{H}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial h_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial h_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial h_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \\ \frac{\partial h_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial h_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial h_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} & \frac{\partial h_q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} & \cdots & \frac{\partial h_q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} \end{bmatrix}$$

When the error variance is not constant, we can fit estimated weighted NLS. The weights would be the inverse of the estimated variances, as in the case of Linear Regression described previously. The variances may be related to the mean and/or the levels of one or more predictor variables. This will necessarily be an iterative process. The function we want to minimize is:

$$Q_W = \sum_{i=1}^{n} \left[v\left(g\left(\mathbf{x_i'}; \boldsymbol{\beta}\right)\right)\right]^{-1} \left[Y_i - g\left(\mathbf{x_i'}; \boldsymbol{\beta}\right)\right]^2 \qquad \text{where} \qquad \sigma_i^2 = v\left(g\left(\mathbf{x_i'}; \boldsymbol{\beta}\right)\right)$$

If the errors are correlated with a known correlation structure, such as AR(1), the autoregressive parameter(s) can be estimated and plugged into the variance-covariance matrix, and we can fit estimated generalized NLS. Here we want to minimize:

$$\left[Y_i - g\left(\mathbf{x_i'}; \boldsymbol{\beta}\right)\right]' \mathbf{V^{-1}} \left[Y_i - g\left(\mathbf{x_i'}; \boldsymbol{\beta}\right)\right]$$

where the elements of $\mathbf{V}$ are functions of unknown parameters which are estimated from the residuals. See the AR(1) description for Linear Regression.

Chapter 8

Generalized Linear Models

8.1 Introduction

When data come from exponential families (e.g. Binomial, Poisson, Negative Binomial (with known α), Gamma, Normal, among others), the linear models that are applied to normal data, such as regression and ANOVA models can be generalized. The exponential family can be written in terms of its probability density (mass) function as (Venables and Ripley (1997)):

$$f\left(y_i;\theta_i,\varphi\right) = \exp\left[\frac{A_i\{y_i\theta_i - \gamma(\theta_i)\}}{\varphi} + \tau\left(y_i,\frac{\varphi}{A_i}\right)\right] \qquad \Rightarrow \qquad \log\left(f(y)\right) = \frac{A\{y\theta - \gamma(\theta)\}}{\varphi} + \tau\left(y,\frac{\varphi}{A}\right)$$

where φ is a scale parameter (known for some settings), A_i is a known weight, and θ_i is a function of the linear predictor.

The generalized linear model has three primary components, and can be applied to any distribution in the exponential family:

- A response variable y, believed to come from a probability distribution in the exponential family
- A set of predictor variables that are believed to be related to y through a linear predictor: $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
- The mean of y is an invertible function of the linear predictor: $\eta = g(\mu) \qquad \mu = g^{-1}(\eta)$

8.1.1 Normal (Gaussian) Distribution

For the Normal Distribution, with mean μ and variance σ^2, we have the following results:

$$f\left(y;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \quad \Rightarrow \quad \log\left(f(y)\right)=\frac{y\mu-0.5y^2-0.5\mu^2}{\sigma^2}-0.5\log\left(2\pi\sigma^2\right)$$

Thus, for the normal distribution, $A_i = 1$, $\theta = \mu$, $\varphi = \sigma^2$, and $\gamma(\theta) = \mu^2/2$.

8.1.2 Binomial Distribution

For the Binomial Distribution, with n trials and probability of Success π, we have the following results, where the "data" are the sample proportions y/n:

$$f\left(y;n,\pi\right)=\binom{n}{y}\pi^y(1-\pi)^{n-y}=\binom{n}{y}\left(\frac{\pi}{1-\pi}\right)^y(1-\pi)^n \quad \Rightarrow$$

$$\log\left(f(y)\right)=n\left[\frac{y}{n}\log\left(\frac{\pi}{1-\pi}\right)+\log(1-\pi)\right]+\log\binom{n}{y}$$

So, for the Binomial distribution, $A_i = n_i$, $\theta = \log\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi)$, $\gamma(\theta) = -\log(1-\pi) = \log\left(1+e^\theta\right)$, and $\varphi = 1$.

8.1.3 Poisson Distribution

For the Poisson Distribution, with mean of λ, we have the following results:

$$f\left(y;\lambda\right)=\frac{e^{-\lambda}\lambda^y}{y!} \qquad \Rightarrow \qquad \log\left(f(y)\right)=-\lambda+y\log(\lambda)-\log(y!)$$

So, for the Poisson Distribution, $A_i = 1$, $\theta = \log(\lambda)$, $\gamma(\theta) = \lambda = e^\theta$, and $\varphi = 1$.

8.1.4 Gamma Distribution

For the Gamma Distribution, with shape parameter α and scale parameter β (with rate parameter $1/\beta$), we have the following results (note that the mean is $\alpha\beta$ and the variance is $\alpha\beta^2$ for the Gamma Distribution):

$$f\left(y;\alpha,\beta\right)=\frac{1}{\beta^{\alpha}\Gamma(\alpha)}y^{\alpha-1}\exp\left[-\frac{y}{\beta}\right]\quad\Rightarrow$$
$$\log\left(f(y)\right)=(\alpha-1)\log(y)-\frac{y}{\beta}-\alpha\log(\beta)-\log\left(\Gamma(\alpha)\right)$$

Let: $\theta=-\frac{1}{\alpha\beta}\quad\Rightarrow\quad\beta=-\frac{1}{\alpha\theta}\quad\Rightarrow\quad\log\left(f(y)\right)=y\alpha\theta-\alpha\log\left(-\frac{1}{\alpha\theta}\right)+(\alpha-1)\log(y)-\log\left(\Gamma(\alpha)\right)$

Now subtract and add $\alpha\log(\alpha)$, and let $\varphi=1/\alpha$:

$$\Rightarrow\quad\log\left(f(y)\right)=\left\{y\alpha\theta-\alpha\log\left(-\frac{1}{\alpha\theta}\right)-\alpha\log(\alpha)\right\}+\left\{\alpha\log(\alpha)+(\alpha-1)\log(y)-\log\left(\Gamma(\alpha)\right)\right\}=$$
$$=\left\{y\alpha\theta-\alpha\log\left(-\frac{\alpha}{\alpha\theta}\right)\right\}+\left\{\alpha\log(\alpha)+(\alpha-1)\log(y)-\log\left(\Gamma(\alpha)\right)\right\}=$$
$$=\left\{\frac{y\theta-\log(-1/\theta)}{\varphi}\right\}+\left\{(1/\varphi)\log(1/\varphi)+((1/\varphi)-1)\log(y)-\log\left(\Gamma(1/\varphi)\right)\right\}$$

For the Gamma Distribution: $A_i=1$, $\theta=-\frac{1}{\alpha\beta}$, $\gamma(\theta)=\log(-1/\theta)$, and $\varphi=1/\alpha$.

8.2 Log-Likelihood Function

Once the data have been collected, we form the log-likelihood function:

$$l(\theta,\varphi;y)=\sum_{i=1}^{n}\left[\frac{A_i\left(y_i\theta_i-\gamma(\theta_i)\right)}{\varphi}+\tau\left(y_i,\frac{\varphi}{A_i}\right)\right]=\sum_{i=1}^{n}l_i(\theta_i,\varphi;y_i)$$

As stated in McCullaugh and Nelder (1989, Section 2.2):

$$E\left\{\frac{\partial l_i}{\partial\theta_i}\right\}=0\qquad E\left\{\frac{\partial^2 l_i}{\partial\theta_i^2}\right\}+E\left\{\left(\frac{\partial l_i}{\partial\theta_i}\right)^2\right\}=0$$

Taking the partial derivatives and expectations for each observation, we get:

$$\frac{\partial l_i}{\partial\theta_i}=\frac{A_i\left(y_i-\frac{\partial\gamma(\theta_i)}{\partial\theta_i}\right)}{\varphi}=\frac{A_i\left(y_i-\gamma'(\theta_i)\right)}{\varphi}\qquad\frac{\partial^2 l_i}{\partial\theta_i^2}=\frac{-A_i\gamma''(\theta_i)}{\varphi}$$

$$E\left\{\frac{\partial l_i}{\partial \theta_i}\right\} = 0 = E\left\{\frac{A_i\left(y_i - \gamma'(\theta_i)\right)}{\varphi}\right\} \quad \Rightarrow \quad E\{y_i\} = \mu_i = \gamma'(\theta_i)$$

$$E\left\{\frac{\partial^2 l_i}{\partial \theta_i^2}\right\} + E\left\{\left(\frac{\partial l_i}{\partial \theta_i}\right)^2\right\} = 0 = E\left\{\frac{-A_i\gamma''(\theta_i)}{\varphi}\right\} + E\left\{\left(\frac{A_i\left(y_i - \gamma'(\theta_i)\right)}{\varphi}\right)^2\right\} = \frac{-A_i\gamma''(\theta_i)}{\varphi} + \frac{A_i^2 V\{y_i\}}{\varphi^2}$$

$$\Rightarrow \quad V\{y_i\} = \frac{\varphi\gamma''(\theta_i)}{A_i}$$

$V\{y_i\}$ is called the **Variance Function**.

8.3 Mean and Variance for Normal, Binomial, Poisson and Gamma Families

Normal Distribution: $A_i = 1$, $\theta = \mu$, $\varphi = \sigma^2$, and $\gamma(\theta) = \mu^2/2$

$$\Rightarrow \quad E\{y_i\} = \mu_i = \gamma'(\theta_i) = \frac{\partial\gamma(\mu_i)}{\partial\mu_i} = \frac{2\mu_i}{2} = \mu_i \qquad V\{y_i\} = \frac{\varphi\gamma''(\theta_i)}{A_i} = \frac{\sigma^2(1)}{1} = \sigma^2$$

Binomial distribution: $A_i = n_i$, $\theta = \log\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi)$, $\gamma(\theta) = -\log(1-\pi) = \log\left(1+e^{\theta}\right)$, and $\varphi = 1$.

$$\gamma(\theta_i) = \log\left(1+e_i^{\theta}\right) \quad \Rightarrow \quad \gamma'(\theta_i) = \left(1+e^{\theta_i}\right)^{-1} e^{\theta_i}$$

$$\Rightarrow \quad \gamma''(\theta_i) = \frac{\left(1+e^{\theta_i}\right)e^{\theta_i} - e^{\theta_i}e^{\theta_i}}{\left(1+e^{\theta_i}\right)^2} = \frac{e^{\theta_i}}{\left(1+e^{\theta_i}\right)^2}$$

$$\Rightarrow \quad E\left\{\frac{y_i}{n_i}\right\} = \mu_i = \gamma'(\theta_i) = \frac{e^{\theta_i}}{1+e^{\theta_i}} \qquad V\left\{\frac{y_i}{n_i}\right\} = \frac{e^{\theta_i}}{n_i\left(1+e^{\theta_i}\right)^2} = \frac{\pi_i(1-\pi_i)}{n_i}$$

Poisson Distribution, $A_i = 1$, $\theta = \log(\lambda)$, $\gamma(\theta) = \lambda = e^{\theta}$, and $\varphi = 1$

$$\gamma(\theta_i) = e^{\theta_i} \quad \Rightarrow \quad \gamma'(\theta_i) = \gamma''(\theta_i) = e^{\theta_i}$$

$$\Rightarrow \quad E\{y_i\} = \mu_i = \gamma'(\theta_i) = e^{\theta_i} \qquad V\{y_i\} = e^{\theta_i}$$

Gamma Distribution: $A_i = 1$, $\theta = -\frac{1}{\alpha\beta}$, $\gamma(\theta) = \log(-1/\theta)$, and $\varphi = 1/\alpha$.

$$\gamma(\theta_i) = \log(-1/\theta) \quad \Rightarrow \quad \gamma'(\theta_i) = \left(\frac{1}{-1/\theta}\right)\left(\frac{1}{\theta^2}\right) = -\frac{1}{\theta} \quad \Rightarrow \quad \gamma''(\theta_i) = \frac{1}{\theta^2}$$

$$\Rightarrow \quad E\{y_i\} = \mu_i = \gamma'(\theta_i) = -\frac{1}{\theta_i} = \alpha_i\beta_i \qquad V\{y_i\} = \left(\frac{1}{\alpha_i}\right)\left(\frac{1}{\theta^2}\right) = \frac{\alpha_i^2\beta_i^2}{\alpha_i} = \alpha_i\beta_i^2$$

8.4 Canonical Link Functions

A link function for a probability distribution is the function of the mean $\eta = g(\mu)$ that is linearly related to the independent variables: $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. The parameter θ for the exponential family is called the canonical parameter (McCullaugh and Nelder (1989, Section 2.2)). The inverse function of $\gamma'(\theta)$ is the **canonical link** (Venables and Ripley (1997)). The canonical links for the Normal, Binomial, Poisson, and Gamma are given below.

Normal Distribution:

$$\theta = \mu \quad \text{(identity link)}$$

Binomial Distribution:

$$\theta = \log\left(\frac{\mu}{1-\mu}\right) \quad \text{(logit link)}$$

Poisson Distribution:

$$\theta = \log(\mu) \quad \text{(log link)}$$

Gamma Distribution:

$$\theta = -\frac{1}{\alpha\beta} \quad \text{(inverse link)}$$

8.5 Maximum Likelihood Estimation

Once we have chosen the distribution for the model, we use the method of maximum likelihood to estimate model parameters, and obtain the variances and covariances among the estimators. Unlike for linear models, where we have "close-formed" solutions, for generalized linear models, we must use iterative methods. Typically, we are maximizing the log of the likelihood function, as opposed to the likelihood, as the process is easier in this form. Both functions will be maximized at the same parameter values.

Once we have the systematic component selected, including all predictors, interactions, and polynomial terms of interest, we can write the mean of the linear predictor for the i^{th} case as:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x_i'}\boldsymbol{\beta} = \begin{bmatrix} 1 & x_{i1} & \cdots & x_{ip} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Then, we take derivatives of the log Likelihood function with respect to each of the regression parameters (in the form of a $(p+1) \times 1$ vector):

$$l(\theta, \varphi; y) = \sum_{i=1}^{n} \left[\frac{A_i \left(y_i \theta_i - \gamma(\theta_i) \right)}{\varphi} + \tau \left(y_i, \frac{\varphi}{A_i} \right) \right] = \sum_{i=1}^{n} l_i(\theta_i, \varphi; y_i)$$

$$g(\boldsymbol{\beta}) \quad = \quad \frac{\partial l}{\partial \boldsymbol{\beta}} \quad = \quad \sum_{i=1}^{n} \left[\frac{A_i \left(y_i \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} - \frac{\partial \gamma(\theta_i)}{\partial \boldsymbol{\beta}} \right)}{\varphi} \right]$$

Taking the second derivative with respect to $\boldsymbol{\beta}'$, we obtain the $(p+1) \times (p+1)$ matrix:

$$G(\boldsymbol{\beta}) \quad = \quad \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \quad = \quad \sum_{i=1}^{n} \frac{A_i}{\varphi} \left[y_i \frac{\partial^2 \theta_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} - \frac{\partial^2 \gamma(\theta_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right]$$

Then, we iterate until convergence with the following Newton-Raphson algorithm:

$$\hat{\boldsymbol{\beta}}_{\text{New}} = \hat{\boldsymbol{\beta}}_{\text{Old}} - \left[G\left(\hat{\boldsymbol{\beta}}_{\text{Old}} \right) \right]^{-1} g\left(\hat{\boldsymbol{\beta}}_{\text{Old}} \right)$$

The estimated Variance-Covariance matrix for $\hat{\beta}$ is $-G\left(\hat{\boldsymbol{\beta}}\right)^{-1}$.

8.5.1 Binomial Distribution

For the Binomial distribution, with "data" y_i/n_i for m distinct cases, we have $A_i = n_i$, $\varphi = 1$ and:

$$l = \sum_{i=1}^{m} \log\left(f(y_i)\right) = \sum_{i=1}^{m} \left\{ n_i \left[\frac{y_i}{n_i} \log\left(\frac{\pi_i}{1-\pi_i} \right) + \log(1-\pi_i) \right] + \log \binom{n_i}{y_i} \right\}$$

with:

$$\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x_i'}\boldsymbol{\beta} \qquad \gamma\left(\theta_i\right) = \log\left(1+e^{\theta_i}\right)$$

$$\Rightarrow \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \mathbf{x_i} \qquad\qquad \frac{\partial \gamma(\theta_i)}{\partial \boldsymbol{\beta}} = \frac{\partial \log\left(1+e^{\mathbf{x_i'}\boldsymbol{\beta}}\right)}{\partial \boldsymbol{\beta}} = \frac{\mathbf{x_i} e^{\mathbf{x_i'}\boldsymbol{\beta}}}{1+e^{\mathbf{x_i'}\boldsymbol{\beta}}}$$

$$\Rightarrow \frac{\partial^2 \theta_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = 0 \qquad\qquad \frac{\partial \gamma(\theta_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \left(\frac{e^{\mathbf{x_i'}\boldsymbol{\beta}}}{1+e^{x'_i\beta}}\right) \mathbf{x_i}\mathbf{x_i'}$$

$$\Rightarrow g\left(\beta\right) \quad = \quad \frac{\partial l}{\partial \beta} \quad = \quad \sum_{i=1}^{m} \left[n_i \left(\frac{y_i}{n_i}\mathbf{x_i} - \frac{e^{\mathbf{x_i'}\boldsymbol{\beta}}}{1+e^{\mathbf{x_i'}\boldsymbol{\beta}}}\mathbf{x_i} \right) \right] = \sum_{i=1}^{m} \left[n_i \left(\frac{y_i}{n_i} - \frac{e^{\mathbf{x_i'}\boldsymbol{\beta}}}{1+e^{\mathbf{x_i'}\boldsymbol{\beta}}} \right) \right] \mathbf{x_i}$$

$$\Rightarrow \qquad G\left(\boldsymbol{\beta}\right) \quad = \quad \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \quad = \quad -\sum_{i=1}^{m} \left[n_i \left(\frac{e^{\mathbf{x_i'}\boldsymbol{\beta}}}{1+e^{\mathbf{x_i'}\boldsymbol{\beta}}} \right) \mathbf{x_i}\mathbf{x_i'} \right] \qquad -G\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{X}'\hat{\mathbf{W}}\mathbf{X}$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1'} \\ \mathbf{x_2'} \\ \vdots \\ \mathbf{x_m'} \end{bmatrix} \qquad \hat{\mathbf{W}} = \text{diag}\left[n_i \hat{\pi}_i \left(1-\hat{\pi}_i\right)\right] \qquad \hat{\pi}_i = \frac{e^{\mathbf{x_i'}\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x_i'}\hat{\boldsymbol{\beta}}}}$$

For starting values of $\hat{\boldsymbol{\beta}}$, set:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \hat{\pi} = \frac{\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} n_i}$$

8.5.2 Poisson Distribution

For the Poisson model, we have $A_i = \varphi = 1$ and:

$$l = \sum_{i=1}^{n} \log\left(f(y_i)\right) = \sum_{i=1}^{n} \left[-\lambda_i + y_i \log(\lambda_i) - \log(y_i!)\right]$$

with:

$$\theta_i = \log\left(\lambda_i\right) = \mathbf{x_i'}\boldsymbol{\beta} \qquad \gamma\left(\theta_i\right) = \lambda_i = e_i^{\theta} = e^{\mathbf{x_i'}\boldsymbol{\beta}}$$

$$\Rightarrow \frac{\partial\theta_i}{\partial\boldsymbol{\beta}} = \mathbf{x_i} \qquad \frac{\partial\gamma(\theta_i)}{\partial\boldsymbol{\beta}} = \frac{\partial e^{\mathbf{x}_i'\boldsymbol{\beta}}}{\partial\boldsymbol{\beta}} = e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x_i}$$

$$\Rightarrow \frac{\partial^2\theta_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = 0 \qquad \frac{\partial\gamma(\theta_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x_i}\mathbf{x_i'}$$

$$\Rightarrow g\left(\boldsymbol{\beta}\right) \quad = \quad \frac{\partial l}{\partial\boldsymbol{\beta}} \quad = \quad \sum_{i=1}^{n}\left[y_i\mathbf{x_i} - e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x_i}\right] = \sum_{i=1}^{n}\left[y_i - e^{\mathbf{x}_i'\boldsymbol{\beta}}\right]\mathbf{x_i}$$

$$\Rightarrow \qquad G\left(\boldsymbol{\beta}\right) \quad = \quad \frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} \quad = \quad -\sum_{i=1}^{n}\left[e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x_i}\mathbf{x_i'}\right] \qquad -G\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{X'\hat{W}X}$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1'} \\ \mathbf{x_2'} \\ \vdots \\ \mathbf{x_n'} \end{bmatrix} \qquad \hat{\mathbf{W}} = \text{diag}\left[\hat{\lambda}_i\right] \qquad \hat{\lambda}_i = e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}$$

For starting values of $\hat{\beta}$, set:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \log\left(\overline{Y}\right) \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

8.5.3 Gamma Distribution

For the Gamma distribution, with parameters α and β, we have $A_i = 1$ and $\varphi = 1/\alpha$ and we have:

$$l = \sum_{i=1}^{n}\log\left(f(y_i)\right) = \sum_{i=1}^{n}\left[-\log\left(\Gamma(\alpha)\right) - \alpha\log(\beta) + (\alpha-1)\log\left(y_i\right) + -\left(\frac{y}{\beta}\right)\right]$$

where we treat α as a fixed known constant while estimating $\boldsymbol{\beta}$, with:

$$\theta_i = -\frac{1}{\alpha\beta_i} = \mathbf{x_i'}\boldsymbol{\beta} \qquad \gamma\left(\theta_i\right) = \log\left(-\frac{1}{\theta_i}\right) = \log\left(\frac{1}{-\mathbf{x_i'}\boldsymbol{\beta}}\right) = -\log\left(-\mathbf{x_i'}\boldsymbol{\beta}\right)$$

$$\Rightarrow \frac{\partial\theta_i}{\partial\boldsymbol{\beta}} = \mathbf{x_i} \qquad \frac{\partial\gamma(\theta_i)}{\partial\boldsymbol{\beta}} = \frac{\partial\left\{-\log\left(-\mathbf{x_i'}\boldsymbol{\beta}\right)\right\}}{\partial\boldsymbol{\beta}} = -\frac{1}{\mathbf{x_i'}\boldsymbol{\beta}}\mathbf{x_i}$$

$$\Rightarrow \frac{\partial^2 \theta_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = 0 \qquad \frac{\partial \gamma(\theta_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \left(\frac{1}{\mathbf{x_i'} \boldsymbol{\beta}} \right)^2 \mathbf{x_i x_i'}$$

$$\Rightarrow g(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} = \alpha \sum_{i=1}^{n} \left[y_i \mathbf{x_i} + \frac{1}{\mathbf{x_i'}\boldsymbol{\beta}} \mathbf{x_i} \right] = \alpha \sum_{i=1}^{n} \left[y_i + \frac{1}{\mathbf{x_i'}\boldsymbol{\beta}} \right] \mathbf{x_i}$$

$$\Rightarrow \quad G(\boldsymbol{\beta}) = \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\alpha \sum_{i=1}^{n} \left[\left(\frac{1}{\mathbf{x_i'}\boldsymbol{\beta}} \right)^2 \mathbf{x_i x_i'} \right] \qquad -G\left(\hat{\boldsymbol{\beta}}\right) = \alpha \mathbf{X}' \hat{\mathbf{W}} \mathbf{X}$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1'} \\ \mathbf{x_2'} \\ \vdots \\ \mathbf{x_n'} \end{bmatrix} \qquad \hat{\mathbf{W}} = \text{diag} \left[\left(\frac{1}{\mathbf{x_i'} \hat{\boldsymbol{\beta}}} \right)^2 \right]$$

For starting values of $\hat{\beta}$, set:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} -\frac{1}{\overline{Y}} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

After estimating $\boldsymbol{\beta}$, we can obtain a moment based estimate of α, the inverse of the square of the coefficient of variation as follows (see McCullaugh and Nelder, 1987, p. 296):

$$\tilde{\alpha}^{-1} = \frac{1}{n - p'} \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)^2$$

Then the variance-covariance matrix is estimated as:

$$V\left\{\hat{\boldsymbol{\beta}}\right\} = \tilde{\alpha}^{-1} \left(\mathbf{X}' \hat{\mathbf{W}} \mathbf{X} \right)^{-1}$$

8.6 Assessing Model Fit

The **scaled deviance** is a measurement that describes how well a model fits the data, similar to the Error Sum of Squares in Linear Regression models. It measures the difference in -2 times the difference in the log-likelihood when the model is fit, and the log-likelihood when the data are the fitted values (saturated model). If the model is a good fit the scaled deviance divided by $n - p'$, should be around 1, where the scaled deviance is the deviance divided by ϕ.

For the Binomial (logit) model, we fit the model with a set of predictors, and obtain $\hat{\boldsymbol{\beta}}$, then obtain the fitted values $\hat{\mu}_i = \hat{\pi}_i = \frac{e^{\mathbf{x_i'}\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x_i'}\hat{\boldsymbol{\beta}}}}$. For the saturated model, we use the observed y_i as the fitted values:

$$l(\mu) = \sum_{i=1}^{m} \left[n_i \left(\frac{y_i}{n_i} \log \left(\frac{\mu_i}{1 - \mu_i} \right) + \log(1 - \mu_i) \right) + \log \binom{n_i}{y_i} \right] \qquad \Rightarrow$$

$$l\left(\hat{\mu}\right) = \sum_{i=1}^{m}\left[n_i\left(\frac{y_i}{n_i}\log\left(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right) + \log\left(1-\hat{\mu}_i\right)\right) + \log\left(\binom{n_i}{y_i}\right)\right]$$

$$l\left(y\right) = \sum_{i=1}^{m}\left[n_i\left(\frac{y_i}{n_i}\log\left(\frac{y_i/n_i}{\left(n_i-y_i\right)/n_i}\right) + \log\left(\frac{n_i-y_i}{n_i}\right)\right) + \log\binom{n_i}{y_i}\right]$$

And the deviance is $D\left(y,\hat{\mu}\right) = -2\left[l\left(\hat{\mu}\right) - l\left(y\right)\right]$:

$$D\left(y,\hat{\mu}\right) = -2\sum_{i=1}^{m}\left\{n_i\left(\frac{y_i}{n_i}\left[\log\left(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right) - \log\left(\frac{y_i/n_i}{\left(n_i-y_i\right)/n_i}\right)\right] + \left[\log\left(1-\hat{\mu}_i\right) - \log\left(\frac{n_i-y_i}{n_i}\right)\right]\right)\right\}$$

$$\Rightarrow \qquad D\left(y,\hat{\mu}\right) = 2\sum_{i=1}^{m}\left\{y_i\log\left(\frac{y_i}{n_i\hat{\mu}_i}\right) + \left(n_i-y_i\right)\log\left(\frac{n_i-y_i}{n_i\left(1-\hat{\mu}_i\right)}\right)\right\}$$

For the Poisson model, we fit the model with a set of predictors, and obtain $\hat{\boldsymbol{\beta}}$, then obtain the fitted values $\hat{\mu}_i = e^{\mathbf{x_i}'\hat{\boldsymbol{\beta}}}$. For the saturated model, we use the observed y_i as the fitted values:

$$l\left(\mu\right) = \sum_{i=1}^{n}\left[-\mu_i + y_i\log\left(\mu_i\right)\log\left(y_i!\right)\right] \qquad \Rightarrow$$

$$l\left(\hat{\mu}\right) = \sum_{i=1}^{n}\left[-\hat{\mu}_i + y_i\log\left(\hat{\mu}_i\right)\log\left(y_i!\right)\right] \qquad l\left(y\right) = \sum_{i=1}^{n}\left[-y_i + y_i\log\left(y_i\right)\log\left(y_i!\right)\right]$$

And the deviance is $D\left(y,\hat{\mu}\right) = -2\left[l\left(\hat{\mu}\right) - l\left(y\right)\right]$:

$$D\left(y,\hat{\mu}\right) = -2\sum_{i=1}^{n}\left[\left(-\hat{\mu}_i + y_i\log\left(\hat{\mu}_i\right)\log\left(y_i!\right)\right) - \left(-y_i + y_i\log\left(y_i\right)\log\left(y_i!\right)\right)\right] = 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i-\hat{\mu}_i\right)\right]$$

For the Gamma model, the **scaled deviance** is -2 times the difference between the log-likelihood for the fitted model and the saturated model. It is $D^*\left(y,\hat{\mu}\right) = D\left(y,\hat{\mu}\right)/\phi = \alpha D\left(y,\hat{\mu}\right)$. For this model, $\hat{m}u_i = -1/\exp\left(\mathbf{x_i}'\hat{\boldsymbol{\beta}}\right)$. The log-likelihood can be written as (ignoring terms that do not change between $l\left(\hat{\mu}\right)$ and $l\left(y\right)$):

$$l\left(\mu\right) = -\alpha\sum_{i=1}^{n}\left[\log\left(\mu_i\right) + \log\left(y_i\right) + \frac{y_i}{\mu_i}\right]$$

$$l\left(\hat{\mu}\right) = -\alpha\sum_{i=1}^{n}\left[\log\left(\hat{\mu}_i\right) + \log\left(y_i\right) + \frac{y_i}{\hat{\mu}_i}\right] \qquad l\left(y\right) = -\alpha\sum_{i=1}^{n}\left[\log\left(y_i\right) + \log\left(y_i\right) + \frac{y_i}{y_i}\right]$$

$$\Rightarrow \qquad D^*\left(y,\hat{\mu}\right) = 2\alpha\sum_{i=1}^{n}\left[-\log\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}\right]$$

The deviance is $D = D^*/\alpha$.

A second measure commonly computed is the Pearson chi-square statistic:

$$X^2 = \sum_{i=1}^{n}\frac{\left(y_i-\hat{\mu}_i\right)^2}{V\left\{\hat{\mu}_i\right\}}$$

The model is rejected if $X^2 \geq \chi^2\left(\alpha, n-p'\right)$.

8.7 Residuals

For each observation, we can obtain various types of residuals. For GLM's, these are typically scaled. The **Pearson residual** scales the raw residual by the standard deviation of $\hat{\mu}_i$:

$$r_P = \frac{y_i - \hat{\mu}_i}{\sqrt{V\left\{\hat{\mu}_i\right\}}} \qquad\qquad X^2 = \sum_{i=1}^{n} r_P^2$$

The **Deviance residual** takes the square root of the contribution of each observation towards the deviance, and multiplies it by the sign of $y_i - \hat{\mu}_i$. Note that the contribution depends on the distribution.

For the Binomial, we have:

$$r_D = \text{sign}\left(y_i - \hat{\mu}_i\right)\sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + \left(n_i - y_i\right)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right]}$$

For the Poisson, the Deviance residuals are:

$$r_D = \text{sign}\left(y_i - \hat{\mu}_i\right)\sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i - \hat{\mu}_i\right)\right]}$$

For the Gamma, we have:

$$r_D = \text{sign}\left(y_i - \hat{\mu}_i\right)\sqrt{2\left[-\log\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right]}$$

Note that for each case $D\left(y, \hat{\mu}\right) = \sum_{i=1}^{n} r_D^2$.

8.8 Interpreting Regression Coefficients

Here we consider the interpretation of regression coefficients for the conjugate link functions.

In the case of binomial data and the logistic regression model, we have:

$$\mu_i = \pi_i \qquad \log\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x_i'}\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

Back transforming, we get:

$$\mu_i = \frac{e^{\mathbf{x_i'}\boldsymbol{\beta}}}{1 + e^{\mathbf{x_i'}\boldsymbol{\beta}}} \qquad \Rightarrow \qquad \frac{\mu_i}{1-\mu_i} = \frac{\left[\frac{e^{\mathbf{x_i'}\boldsymbol{\beta}}}{1+e^{\mathbf{x_i'}\boldsymbol{\beta}}}\right]}{\left[\frac{1}{1+e^{\mathbf{x_i'}\boldsymbol{\beta}}}\right]} = e^{\mathbf{x_i'}\boldsymbol{\beta}} = e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}$$

where $\frac{\mu_i}{1-\mu_i}$ is the **odds** of the event occurring (the number of times it occurs per each non-occurrence). The effect of increasing X_j by 1, while holding the other predictors constant is measured by the **Odds Ratio**:

$$OR_j = \frac{\text{odds}\left(X_1, \ldots, X_{j-1}, X_j + 1, X_{j+1}, \ldots, X_p\right)}{\text{odds}\left(X_1, \ldots, X_{j-1}, X_j, X_{j+1}, \ldots, X_p\right)} = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1} X_{j-1} + \beta_j \left(X_j + 1\right) + \beta_{j+1} X_{j+1} + \cdots + \beta_p X_p}}{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1} X_{j-1} + \beta_j X_j + \beta_{j+1} X_{j+1} + \cdots + \beta_p X_p}} = e^{\beta_j}$$

Thus, the multiplicative change in the odds is e^{β_j} as X_j increases 1 unit, holding all other predictors constant. If β_j is positive, the probability of success increases with X_j, if it is negative, it decreases as X_j increases, and if it is 0, the probability of success is not related to X_j, controlling for all other predictors.

For the Poisson regression model, we have:

$$\mu_i = e^{\mathbf{x_i'}\boldsymbol{\beta}} = e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}$$

Here, if we increase X_j by 1 unit, we get a multiplicative change in the mean of Y of e^{β_j}. The interpretations are similar to those of the odds for the logistic regression model.

For the Gamma regression model, we have:

$$\mu_i = -\frac{1}{\mathbf{x_i'}\boldsymbol{\beta}} = -\frac{1}{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}$$

Here, if we change X_j by 1, the ratio change in μ_i is:

$$\frac{\mu\left(X_j\right)}{\mu\left(X_j+1\right)} = \frac{\beta_0+\beta_1 X_1+\cdots+\beta_{j-1}X_{j-1}+\beta_j\left(X_j+1\right)+\beta_{j+1}X_{j+1}+\cdots+\beta_p X_p}{\beta_0+\beta_1 X_1+\cdots+\beta_{j-1}X_{j-1}+\beta_j X_j+\beta_{j+1}X_{j+1}+\cdots+\beta_p X_p} = 1+\frac{\beta_j}{\mu\left(X_j\right)}$$

Note that if β_j is positive, μ decreases as X_j increases; if β_j is negative, μ increases as X_j increases; and if $\beta_j = 0$, μ is not related to X_j, controlling for all other predictors.

8.9 Inferences Regarding Regression Coefficients

Tests regarding regression coefficients can be conducted as **Likelihood-Ratio** and **Wald** tests. Also, large-sample Confidence can be formed based on asymptotic normality results.

For Likelihood-Ratio tests, the log-likelihood is computed under the null hypothesis, which we will denote as l_0, and under the alternative hypothesis, which we will denote as l_A. Then we compute the likelihood-ratio test statistic, and and define the rejection region:

$$TS: X^2_{LR} = -2\left(l_0 - l_A\right) \qquad RR: X^2_{LR} \geq \chi^2\left(\alpha, q\right)$$

where q is the number of restrictions under the null hypothesis.

For Wald tests of the form $\mathbf{K'}\boldsymbol{\beta} - \mathbf{m} = \mathbf{0}$, where the number of linear independent rows in $\mathbf{K'}$ is q, the number of restrictions under the null hypothesis, we have the following test statistic and rejection region:

$$TS: X^2_W = \frac{\left(\mathbf{K'}\hat{\boldsymbol{\beta}} - \mathbf{m}\right)'\left[\mathbf{K'}\hat{V}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{K}\right]^{-1}\left(\mathbf{K'}\hat{\boldsymbol{\beta}} - \mathbf{m}\right)}{q} \qquad RR: X^2_W \geq \chi^2\left(\alpha, q\right)$$

The special case of testing for individual regression coefficients has the form (some software packages including R, use the Z-statistic (not squared):

$$TS: X^2_W = \left(\frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)}\right)^2 \qquad RR: X^2_W \geq \chi^2\left(\alpha, 1\right)$$

A large-sample $(1-\alpha)100\%$ for β_j can be computed as:

$$\hat{\beta}_j \pm z\left(\alpha/2\right) SE\left(\hat{\beta}_j\right)$$

8.10 Negative Binomial Regression

In many instances with count data the variance is larger than the mean, making the Poisson model inappropriate. The Negative Binomial model allows for the variance to be larger than the mean. The probability distribution/likelihood for the i^{th} observation is of the form:

$$\begin{aligned}
L_i = p\left(y_i\right) &= \frac{\Gamma\left(\alpha^{-1}+y_i\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y_i+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \\
&= \frac{\left(y_i+\alpha^{-1}-1\right)\cdots\left(\alpha^{-1}\right)\Gamma\left(\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y_i+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \\
&= \frac{\left(y_i+\alpha^{-1}-1\right)\cdots\left(\alpha^{-1}\right)}{y_i!}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \\
&= \frac{\left(y_i+e^{-\alpha^*}-1\right)\cdots\left(e^{-\alpha^*}\right)}{y_i!}\left(\frac{e^{-\alpha^*}}{e^{-\alpha^*}+\mu_i}\right)^{e^{-\alpha^*}}\left(\frac{\mu_i}{e^{-\alpha^*}+\mu_i}\right)^{y_i}
\end{aligned}$$

where $-\alpha^* = \ln\left(\alpha^{-1}\right)$ and $\mu_i = \mathbf{x_i'}\boldsymbol{\beta}$. The log-likelihood function for the i^{th} observation is:

$$l_i = \ln\left(L_i\right) = \sum_{j=1}^{y_i-1}\ln\left(e^{-\alpha^*}+j\right) - \ln\left(y_i!\right) + e^{-\alpha^*}\ln\left(e^{-\alpha^*}\right) + y_i\ln\left(\mu_i\right) - \left(e^{-\alpha^*}+y_i\right)\ln\left(\mu_i+e^{-\alpha^*}\right)$$

The first and second partial derivatives with respect to $-\alpha^*$ and $\boldsymbol{\beta}$ are:

$$\frac{\partial l_i}{\partial\left(-\alpha^*\right)} = e^{-\alpha^*}\left\{\sum_{j=1}^{y_i-1}\frac{1}{e^{-\alpha^*}+j} + 1 + \ln\left(e^{-\alpha^*}\right) - \frac{e^{-\alpha^*}+y_i}{e^{-\alpha^*}+\mu_i} - \ln\left(e^{-\alpha^*}+\mu_i\right)\right\}$$

$$\begin{aligned}
\frac{\partial^2 l_i}{\partial\left(-\alpha^*\right)^2} = e^{-\alpha^*}&\left\{\sum_{j=1}^{y_i-1}\frac{1}{e^{-\alpha^*}+j} + 1 + \ln\left(e^{-\alpha^*}\right) - \frac{e^{-\alpha^*}+y_i}{e^{-\alpha^*}+\mu_i} - \ln\left(e^{-\alpha^*}+\mu_i\right)\right. \\
&\left. -e^{-\alpha^*}\sum_{j=1}^{y_i-1}\frac{1}{\left(e^{-\alpha^*}+j\right)^2} + 1 - e^{-\alpha^*}\left(\frac{\mu_i-y_i}{\left(\mu_i+e^{-\alpha^*}\right)^2}\right) - \frac{e^{-\alpha^*}}{\mu_i+e^{-\alpha^*}}\right\}
\end{aligned}$$

$$\frac{\partial^2 l_i}{\partial\left(-\alpha^*\right)\partial\boldsymbol{\beta}} = e^{-\alpha^*}\left\{\frac{y_i-\mu_i}{\left(e^{-\alpha^*}\right)^2}\right\}\mathbf{x_i}$$

$$\frac{\partial l_i}{\partial\boldsymbol{\beta}} = e^{-\alpha^*}\left\{\frac{y_i-\mu_i}{\mu_i+e^{-\alpha^*}}\right\}\mathbf{x_i}$$

$$\frac{\partial^2 l_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = -e^{-\alpha^*}\mu_i\left\{\frac{e^{-\alpha^*}+y_i}{\left(\mu_i+e^{-\alpha^*}\right)^2}\right\}\mathbf{x_i}\mathbf{x_i'}$$

The Newton-Raphson algorithm then is conducted as follows:

$$g_{-\alpha^*} = \sum_{i=1}^{n}\frac{\partial l_i}{\partial\left(-\alpha^*\right)} \qquad G_{-\alpha^*} = \sum_{i=1}^{n}\frac{\partial^2 l_i}{\partial\left(-\alpha^*\right)^2}$$

$$g_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \boldsymbol{\beta}} \qquad G_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$$

$$g_{-\alpha^* \boldsymbol{\beta}} = \begin{bmatrix} g_{-\alpha^*} \\ g_{\boldsymbol{\beta}} \end{bmatrix} \qquad G_{-\alpha^* \boldsymbol{\beta}} = \begin{bmatrix} G_{-\alpha^*} & \left(\sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial(-\alpha^*)\partial \boldsymbol{\beta}}\right)' \\ \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial(-\alpha^*)\partial \boldsymbol{\beta}} & G_{\boldsymbol{\beta}} \end{bmatrix}$$

To complete the algorithm, the following steps can be fit.

First, set $-\alpha^* = 0$, which is equivalent to setting $\alpha = \alpha^{-1} = 1$, and obtain an estimate of $\boldsymbol{\beta}$ by iterating to convergence:

$$\hat{\boldsymbol{\beta}}^{(i)} = \hat{\boldsymbol{\beta}}^{(i-1)} - \left[G_{\boldsymbol{\beta}}\right]^{-1} g_{\boldsymbol{\beta}}$$

Second, set $\boldsymbol{\beta}' = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$ and obtain an estimate of $-\alpha^*$ by iterating to convergence:

$$\left(-\hat{\alpha}^*\right)^{(i)} = \left(-\hat{\alpha}^*\right)^{(i-1)} - \left[G_{-\alpha^*}\right]^{-1} g_{-\alpha^*}$$

Third, use the results from the first two steps to obtain an estimate of θ:

$$\hat{\theta}^{(i)} = \hat{\theta}^{(i-1)} - \left[G_{-\alpha^* \boldsymbol{\beta}}\right]^{-1} g_{-\alpha^* \boldsymbol{\beta}} \qquad \theta = \begin{bmatrix} -\alpha^* \\ \boldsymbol{\beta} \end{bmatrix}$$

Fourth, back-transform to get estimate of $\alpha^{-1} = e^{-\alpha^*}$.

Note that for the Negative Binomial distribution: $E\{Y_i\} = \mu_i$ and $V\{Y_i\} = \mu_i(1 + \alpha\mu_i)$.