SAS – Logistic – exs. pdf

15	12.08	0.16000	0.1806104	-0.53576	-0.544155
16	14.83	0.93137	0.951464	-0.944241	-0.891258
17	12.33	0.31183	0.2489491	1.40235	1.3675388
18	15.08	0.95902	0.9671908	-0.506855	-0.488396
19	12.56	0.39000	0.3326479	1.2172483	1.2016944
20	15.33	0.96396	0.9779399	-1.0025	-0.919574
21	12.83	0.47222	0.4284346	0.9195778	0.9162826
22	15.58	0.97872	0.9852212	-0.522087	-0.490007
23	13.08	0.47475	0.529902	-1.09953	-1.098226
24	15.83	0.98246	0.9901234	-0.82784	-0.746189
25	17.58	1.00000	0.9994267	0.7756856	1.0968278

Leukemia Data

Feigl and Zelen reported the survival time in weeks and the white cell blood count (WBC) at time of diagnosis for 33 patients who eventually died of acute leukemia. Each person was classified as AG+ or AG-, indicating the presence or absence of a certain morphological characteristic in the white cells. Four variables are given in the **SAS** data set: WBC, a binary **indicator variable** IAG (1 for AG+, 0 for AG-), NTOTAL (the # of patients with the given combination of IAG and WBC), and NRES (the # of NTOTAL that survived at least one year from the time of diagnosis).

The researchers are interested in modeling the probability p of surviving at least one year as a function of WBC and IAG. They believe that WBC should be transformed to a log scale, given the skewness in the WBC values.

As an initial step in the analysis, consider the following model (with multiple effects):

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{LWBC} + \beta_2 \text{IAG},$$

LWBC = quant.
IAG = dummy

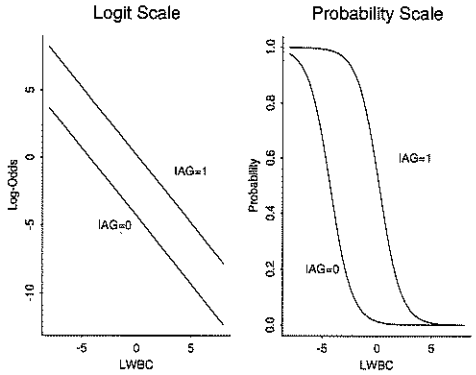
where LWBC = log WBC. The model is best understood by separating the AG+ and AG- cases. For AG- individuals, IAG=0 so the model reduces to

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{LWBC} + \beta_2 * 0 = \beta_0 + \beta_1 \text{LWBC}.$$

For AG+ individuals, IAG=1 and the model implies

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{LWBC} + \beta_2 * 1 = (\beta_0 + \beta_2) + \beta_1 \text{LWBC}.$$

The model without IAG (i.e. $\beta_2 = 0$) is a simple logistic model where the log-odds of surviving one year is linearly related to LWBC, and is independent of AG. The reduced

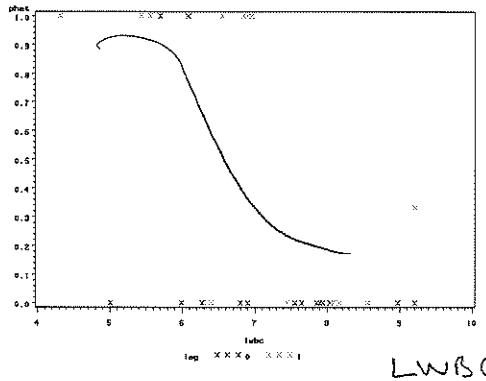


model with $\beta_2 = 0$ implies that there is no effect of the AG level on the survival probability once LWBC has been taken into account.

Including the **binary predictor** IAG in the model implies that there is a linear relationship between the log-odds of surviving one year and LWBC, with a constant slope for the two AG levels. This model includes an effect for the AG morphological factor, but more general models are possible. A natural extension would be to include a product or interaction effect, a point that I will return to momentarily.

The parameters are easily interpreted: β_0 and $\beta_0 + \beta_2$ are intercepts for the population logistic regression lines for AG- and AG+, respectively. The lines have a common slope, β_1. The β_2 coefficient for the IAG indicator is the difference between intercepts for the AG+ and AG- regression lines. A picture of the assumed relationship is given above for $\beta_1 < 0$. The population regression lines are parallel on the logit scale only, but the order between IAG groups is preserved on the probability scale.

Before looking at SAS output for the equal slopes model, note that the data set has 30 distinct IAG and LWBC combinations, or 30 "groups" or samples. Only two samples have more than 1 observation. The majority of the observed proportions surviving at least one year (number surviving $\geq$ 1 year/ group sample size) are 0 (i.e. 0/1) or 1 (i.e. 1/1).



LWBC

IAG = {1 > red, 0 > black}

This sparseness of the data makes it difficult to graphically assess the suitability of the logistic model (see plot above). Although significance tests on the regression coefficients do not require large group sizes, the chi-squared approximation to the deviance and Pearson statistics is suspect in sparse data settings. With small group sizes as we have here most researchers would not interpret p-values for D and X^2 literally. Instead, they would use D/df and X^2/df to informally check the fit of the model. Diagnostics would be used to highlight problems with the model.

```
data d1;
   input ntotal nres Iag wbc @@;
   lwbc  = log(wbc);
   phat = nres/ntotal;
   datalines;
  1 1 1     75   1 1 1    230   1 1 1    430
  1 1 1    260   1 0 1    600   1 1 1   1050
  1 1 1   1000   1 0 1   1700   1 0 1    540
  1 1 1    700   1 1 1    940   1 0 1   3200
  1 0 1   3500   1 0 1   5200   3 1 1  10000
  1 1 0    440   1 1 0    300   1 0 0    400
  1 0 0    150   1 0 0    900   1 0 0    530
  1 0 0   1000   1 0 0   1900   1 0 0   2700
  1 0 0   2800   1 0 0   3100   1 0 0   2600
  1 0 0   2100   1 0 0   7900   2 0 0  10000
 ;
proc sort data = d1 out = d2;
  by lwbc;
proc genmod data=d2;
    model nres/ntotal = Iag lwbc/dist=bin
                                  link=logit noscale obstats type3;
    ods output obstats = r;
data e;
  merge d2 r;
  keep lwbc Iag phat Pred;
symbol1 value=x interpol=none width=2;
symbol2 value=y interpol=join width=2;
proc gplot;
  plot (phat pred)*lwbc = Iag;
run;
```

The GENMOD Procedure

Model Information

Data Set	WORK.D2
Distribution	Binomial
Link Function	Logit
Response Variable (Events)	nres
Response Variable (Trials)	ntotal
Number of Observations Read	30
Number of Observations Used	30
Number of Events	11
Number of Trials	33

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	27	23.0136	0.8524
Scaled Deviance	27	23.0136	0.8524
Pearson Chi-Square	27	19.8094	0.7337
Scaled Pearson X2	27	19.8094	0.7337
Log Likelihood		-13.4164	

Algorithm converged.

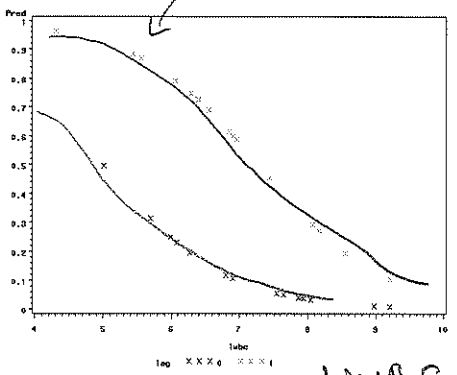
Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.5433	3.0224	-0.3805	11.4672	3.36	0.0666
Iag	1	2.5196	1.0907	0.3819	4.6573	5.34	0.0209
lwbc	1	-1.1088	0.4609	-2.0122	-0.2053	5.79	0.0162
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Iag	1	6.94	0.0084
lwbc	1	8.73	0.0031



The ratios of D and X^2 divided by the df are both less than 1, indicating that there are no gross deficiencies with the model. Recall that the **Analysis Of Parameter Estimates** table gives p-values for testing the hypothesis that the regression coefficients are zero for each predictor in the model. The two predictors are LWBC and IAG, so the small p-values indicate that LWBC and IAG are important predictors of survival in this model. The LR test p-values in the LR table lead to the same conclusion. If either predictor was insignificant, I would consider refitting the model omitting the least significant effect, as in regression.

Given that the model fits reasonably well, a test of $H_0 : \beta_2 = 0$ might be a primary interest here. This checks whether the regression lines are identical for the two AG levels, which is a test for whether AG affects the survival probability, after taking LWBC into account. This test is rejected at any of the usual significance levels, suggesting that the AG level affects the survival probability (assuming a very specific model).

A plot of the predicted survival probabilities as a function of LWBC is given with the **SAS** output, using different colors to identify the two IAG groups. The model indicates that the probability of surviving at least one year from the time of diagnosis is a decreasing function of LWBC. For a given LWBC the survival probability is greater for AG+ patients than for AG- patients. This tendency is consistent with the observed proportions, which show little information about the exact form of the trend.

The estimated (ML) survival probabilities satisfy

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 5.54 - 1.11\text{LWBC} + 2.52\text{IAG}.$$

For AG- individuals with IAG=0, this reduces to

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 5.54 - 1.11\text{LWBC},$$

or equivalently,

$$\hat{p} = \frac{\exp(5.54 - 1.11\text{LWBC})}{1 + \exp(5.54 - 1.11\text{LWBC})}.$$

For AG+ individuals with IAG=1,

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 5.54 - 1.11\text{LWBC} + 2.52 * (1) = 8.06 - 1.11\text{LWBC},$$

or

$$\hat{p} = \frac{\exp(8.06 - 1.11\text{LWBC})}{1 + \exp(8.06 - 1.11\text{LWBC})}.$$

The equivalence between logit and probability scales is that $\log(p/(1-p)) = z$ implies $p = \exp(z)/\{1 + \exp(z)\}$, regardless of how many effects define z.

Using the **logit scale**, the difference between AG+ and AG- individuals in the estimated log-odds of surviving at least one year, at a fixed but arbitrary LWBC, is the estimated IAG regression coefficient:

$$(8.06 - 1.11\text{LWBC}) - (5.54 - 1.11\text{LWBC}) = 2.52.$$

Using properties of exponential functions, the odds that an AG+ patient lives at least one year is $\exp(2.52) = 12.42$ times larger than the odds that an AG- patient lives at least one year, regardless of LWBC. This statistical summary is called the **adjusted odds ratio** for the IAG effect. A CI for the adjusted odds ratio is obtained by exponentiating the CI for the corresponding regression effect given in the parameter estimates table: $\exp(.38) = 1.47$ to $\exp(4.66) = 105.35$.

odds ↓ 5. decreases by 67% for each unit increase in LWBC.

$\pi^{+1} = e^{\frac{1}{3}} \pi^{\odot}$

Similarly, the estimated odds ratio of .33 = exp(−1.11) for LWBC implies that the odds of surviving at least one year is reduced by a factor of 3 for each unit increase of LWBC.

Although the equal slopes model appears to fit well, a more general model might fit better. A natural generalization here would be to add an **interaction**, or product term, IAG * LWBC to the model. The logistic model with an IAG effect and the IAG * LWBC interaction is equivalent to fitting separate logistic regression lines to the two AG groups. This interaction model provides an easy way to test whether the slopes are equal across AG levels. I will note that the interaction term is not needed here.

Data from a Budworm Experiment

The purpose of this experiment was to assess the resistance of tobacco budworm larvae to cypermethrin. Batches of 20 pyrethroid-resistant moths (i.e. moths that have grown resistant to the pesticide pyrethroid) of each sex were exposed to a range of doses of cypermethrin two days after emergence from pupation. The number of moths which were either knocked down (uncoordinated movements) or dead was recorded 72 hours after treatment. We will model the probability of uncoordinated movement, or death, as a function of sex and dose.

p = prob(death)

The **SAS** program below inputs the number of responders for each combination of sex and dose. The sample size of 20 for each sex-dose combination is defined directly in the **data** step, rather than entered with the data. A plot of the observed proportion of responders against the log(dose) (log-dose is traditionally used in bioassay settings such as this) shows that increased doses are more effective, and that males tend to be more affected than females.

The first **genmod** proc fits a logistic model with main effects for **sex** (a factor) and log(dose) (a predictor) and a sex-by-log(dose) interaction. The **model** statement specifies the response in the number of events divided by sample size format as used in the menarche data. The **class** statement identifies sex as a factor.

You should probably read the **ANCOVA** (analysis of covariance) notes from my Data Analysis II website if you are not familiar with how **factors** are treated in regression models. Factors are typically used with qualitative effects such as sex, ethnicity and so on. The (main) effect for a factor with c levels (sex has $c = 2$ levels) is captured through c indicator (binary,

or 0-1) variables, one for each level of the factor. However, perfect collinearity results from including all c binary predictors in a model with an intercept. Without loss of generality, one of the binary predictors can be removed from the model. The group corresponding to the omitted binary predictor is called the **baseline** group. Interactions between a factor and a predictor (or another factor) corresponds to adding one *df* effect for each product of the binary variables with the other feature in the interaction. If a factor has $c > 2$ levels the SAS output will include estimates of the regression coefficients for the $c - 1$ single *df* effects while the LR table will include a test of significance for the entire effect (i.e. a single $c - 1$ df test on the importance of the effect).

By default, SAS uses the category of the **class** variable with the highest alphanumeric value as the baseline group. Thus, in the budworm analysis, males (M) are the baseline group, and the sex effect corresponds to a single binary predictor with levels 1 for females and 0 for males. The interaction between **ld** and **sex** corresponds to a predictor that is the product of ld and the binary sex effect.

ld = log(dose)

Letting p_j be the success probability (i.e. probability of death or uncoordinated movement) for batch j, and defining $sex_j = 1$ if batch j consists of females and 0 if batch j consists of males (baseline), the fitted model is equivalent to:

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 sex_j + \beta_2 ld_j + \beta_3 sex_j * ld_j$$

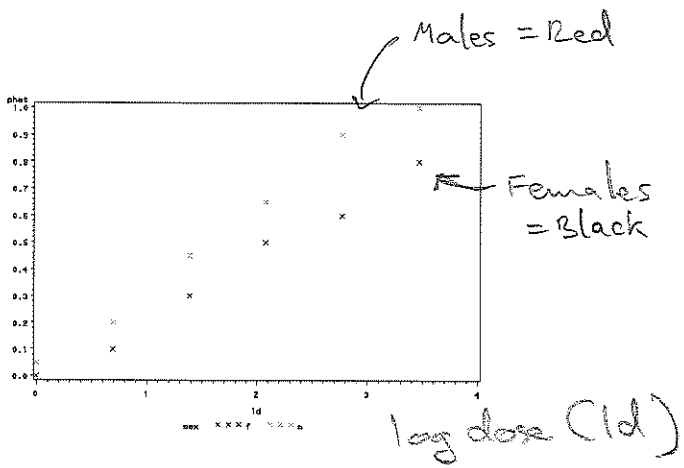
where ld_j is the log-dose given to batch j. The model implies that the log-odds of a response is linearly related to log(dose), but that males and females have distinct lines. To see this, note that the model reduces to

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_2 ld_j \qquad \text{sex} = 0$$

for males (sex = 0) and

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 + (\beta_2 + \beta_3) ld_j \qquad \text{sex} = 1$$

for females. Thus, β_0 and β_2 give the intercept and slope for the baseline group, males, and β_1 and β_3 give the **difference** in intercepts and slopes between females and males.



It is important to note that the choice of a baseline group for a factor impacts the interpretation of the parameters, but does not impact the interpretation of the model - provided the model follows the hierarchy principal. For example, how would the parameters and model change if we ran the same program but had defined sex to have 2 levels: M for males and W for females?

We could have treated IAG as a factor in the leukemia analysis (even though it was coded numerically), but SAS would have treated IAG=1 as the baseline group. Treating IAG as a predictor (not a class variable) essentially leads to IAG=0 as the baseline group or level. You should read the online help for the **class** statement in **genmod** to learn how to change the default choice for the baseline group.

```
options ls=79 nodate;
  data d1;
    input sex $ dose nresp @@;
    ntot  = 20;
    phat  = nresp/ntot;
    ld    = log(dose);
    cards;
    m  1  1 m  2  4 m  4  9 m  8  13
    m 16 18 m 32 20
    f  1  0 f  2  2 f  4  6 f  8  10
    f 16 12 f 32 16
    ;
  symbol1 value=x interpol=none width=2;
  symbol2 value=y interpol=join width=2;
```



```
proc gplot;
  plot phat*ld=sex;

run;
```

Model with unequal slopes

```
proc genmod data=d1;
  classes sex;
  model nresp/ntot = sex ld sex*ld/dist=bin
                     link=logit noscale type3;

run;
```

Model Information

Data Set	WORK.D1
Distribution	Binomial
Link Function	Logit
Response Variable (Events)	nresp
Response Variable (Trials)	ntot
Number of Observations Read	12
Number of Observations Used	12
Number of Events	111
Number of Trials	240

Class Level Information

Class	Levels	Values
sex	2	f m

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	8	4.9937	0.6242
Scaled Deviance	8	4.9937	0.6242
Pearson Chi-Square	8	3.5047	0.4381
Scaled Pearson X2	8	3.5047	0.4381
Log Likelihood		-105.7388	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-2.8186	0.5480	-3.8926	-1.7445	26.46	<.0001
sex	f	1	-0.1750	0.7783	-1.7004	1.3505	0.05	0.8221
sex	m	0	0.0000	0.0000	0.0000	0.0000	.	.
ld		1	1.8163	0.3059	1.2166	2.4159	35.24	<.0001

ld*sex	f	1	-0.5091	0.3895	-1.2726	0.2543	1.71	0.1912
ld*sex	m	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
sex	1	0.05	0.8221
ld	1	112.73	<.0001
ld*sex	1	1.76	0.1842

Looking at the output, we see

1. from the data plot that the pesticide is more effective as the dose increases
2. the effect of dose is statistically significant ($p < .0001$)
3. the sex-by-log(dose) interaction is not significant (p = .18 on LR test) so an equal slopes model (i.e. the original model with no interaction) might be considered
4. the sex effect (which estimates the difference in intercepts) is not significant in the **presence** of the interaction (but is significant if interaction is excluded)
5. The Deviance and Pearson statistics are smaller than the degrees of freedom, which indicate that the model fits the data. A chi-squared approximation is appropriate here, and the p-values would be large.

The fitted model is

$$\log\left(\frac{\tilde{p}_j}{1-\tilde{p}_j}\right) = -2.82 - .18sex_j + 1.82ld_j - .51sex_j * ld_j$$

which reduces to

$$\log\left(\frac{\tilde{p}_j}{1-\tilde{p}_j}\right) = -2.82 + 1.82ld_j$$

for males and

$$\log\left(\frac{\tilde{p}_j}{1-\tilde{p}_j}\right) = -2.82 - .18 + (1.82 - .51)ld_j = -3.00 + 1.31ld_j$$

for females.

Recalling that $\log(p/(1-p)) = z$ implies $p = \exp(z)/\{1+\exp(z)\}$, we can write the fitted model as

$$\tilde{p}_j = \frac{\exp(-2.82 + 1.82ld_j)}{1+\exp(-2.82+1.82ld_j)}$$

for males and

$$\tilde{p}_j = \frac{\exp(-3.00 + 1.31ld_j)}{1+\exp(-3.00+1.31ld_j)}$$

for females.

A reasonable next step would be to fit a model with equal slopes (i.e. no sex-by-ld interaction), which is done below. I will leave to you to decide what would be the best final model and how to summarize the analysis.

Equal slopes model

```
proc genmod data=d1;
  class sex;
  model nresp/ntot = sex ld/dist=bin link=logit noscale type3;
run;
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	9	6.7571	0.7508
Pearson Chi-Square	9	5.3060	0.5896

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-2.3724	0.3855	-3.1280	-1.6168	37.87	<.0001

sex	f	1	-1.1007	0.3558	-1.7982	-0.4033	9.57	0.0020
sex	m	0	0.0000	0.0000	0.0000	0.0000	.	.
ld		1	1.5353	0.1891	1.1647	1.9060	65.92	<.0001

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
sex	1	10.23	0.0014
ld	1	112.04	<.0001

Compare prob of failure between M & F

$$\log\left(\frac{p}{1-p}\right)_M = \beta_0 + \beta_2 ld \quad , \quad \beta_0 = -2.37 \;,\; \beta_2 = 1.54$$

$$\log\left(\frac{p}{1-p}\right)_F = \beta_0 + \beta_1 + \beta_2 ld \quad , \quad \beta_1 = -1.00$$

$$\Rightarrow \log \pi_M - \log \pi_F = \log\left(\frac{\pi_M}{\pi_F}\right) = \beta_1$$

$$\left\{ \text{where } \pi_M = \frac{p_M}{1-p_M} \;,\; \pi_F = \frac{p_F}{1-p_F} \right\}$$

$$\Rightarrow \frac{\pi_M}{\pi_F} = e^{\beta_1} \approx 2.72$$

$$\Rightarrow \pi_M = 2.7\, \pi_F$$

Odds of failure for M are 2.7 times the odds of failure for F

Using PROC LOGISTIC

This handout illustrates the **logistic** procedure in SAS - the document is old, and has not been updated with nicer graphics and newer features in SAS.

The UNM Trauma Data (Model survival probability)

The data to be analyzed here were collected on 3132 patients admitted to The University of New Mexico Trauma Center between the years 1991 and 1994. For each patient, the attending physician recorded their age, their revised trauma score (RTS), their injury severity score (ISS), whether their injuries were blunt (i.e. the result of a car crash: BP=0) or penetrating (i.e. gunshot wounds: BP=1), and whether they eventually survived their injuries (SURV=0 if not, SURV=1 if survived). Approximately 10% of patients admitted to the UNM Trauma Center eventually die from their injuries.

The ISS is an overall index of a patient's injuries, based on the approximately 1300 injuries cataloged in the Abbreviated Injury Scale. The ISS can take on values from 0 for a patient with no injuries to 75 for a patient with 3 or more life threatening injuries. The ISS is the standard injury index used by trauma centers throughout the U.S. The RTS is an index of physiologic injury, and is constructed as a weighted average of an incoming patient's systolic blood pressure, respiratory rate, and Glasgow Coma Scale. The RTS can take on values from 0 for a patient with no vital signs to 7.84 for a patient with normal vital signs.

Champion et al. (1981) proposed a logistic regression model to estimate the probability of a patient's survival as a function of RTS, the injury severity score ISS, and the patient's age, which is used as a surrogate for physiologic reserve. Subsequent survival models included the binary effect BP as a means to differentiate between blunt and penetrating injuries.

We will develop a logistic model for predicting survival from ISS, AGE, BP, and RTS. Data on the number of severe injuries in each of the nine body regions is also included in the database, so we will also assess whether these features have any predictive power. The following labels were used to identify the number of severe injuries in the nine regions: AS

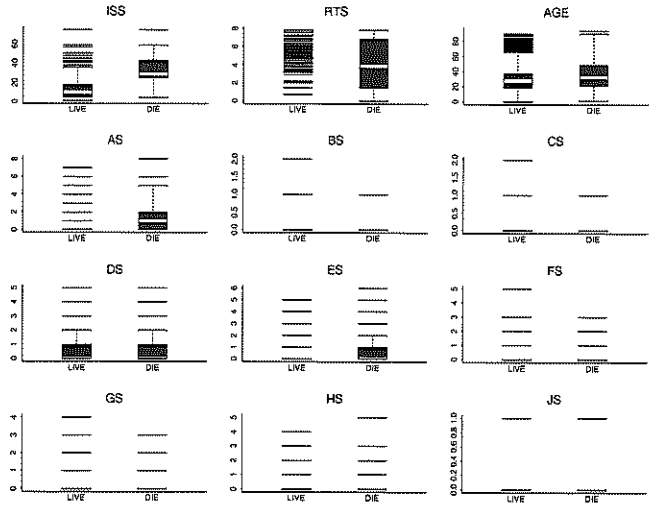
= HEAD, BS = FACE, CS = NECK, DS = THORAX, ES = ABDOMEN, FS = SPINE, GS = UPPER EXTREMITIES, HS = LOWER EXTREMITIES, and JS = SKIN.

I made side-by-side boxplots of the distributions of ISS, AGE, RTS, and AS through JS for the survivors and non-survivors. In several body regions the number of injuries is limited, so these boxplots are not very enlightening. Survivors tend to have lower ISS scores, tend to be slightly younger, tend to have higher RTS scores, and tend to have fewer severe head (AS) and abdomen injuries (ES) than non-survivors. The importance of the effects individually towards predicting survival is directly related to the separation between the survivors and non-survivors scores.

Selecting Predictors in the Trauma Data

The logistic procedure has automated methods for model building, including **backward elimination**, forward selection, and stepwise methods, among others. A **SAS** proc is given below for performing a backward elimination, starting with a full model having 13 effects: ISS, BP, RTS, AGE, and AS-JS. The elimination process begins by identifying the least significant effect in the full model, based on Wald test p-values. If the least significant effect is not important, the effect is omitted from the model and the resulting model is fitted. The process is repeated until the least significant effect in the model is too important to exclude. This gives the final model. The default significance level for testing effects in backward elimination is .05. The test level was changed to .10 using the **slstay=.10** option on the **model** statement.

In our previous logistic regression analyses, the cases in the data set were pre-aggregated into groups of observations having identical levels of the predictor variables. The numbers of cases in the success category and the group sample sizes were specified in the model statement, along with the names of the predictors. The trauma data set, which is not reproduced here, is **raw data** consisting of one record per patient (i.e. 3132 lines). The logistic model is fitted to data on individual cases by specifying the binary response variable



(SURV) and the predictors on the **model** statement. Some care is needed because **SAS** will define the logistic model in such a way that the **success** category corresponds to the **lower** of the two levels of response variable. Here SURV=0 corresponds to patients that died whereas SURV=1 corresponds to patients that lived, so the default model is for the log-odds of **dying** as a function of the predictor variables. To reverse the role of the response levels and model the log-odds of **surviving**, I used the **descending** option on the **model** statement (see the printed note from the log window contained with the **SAS** output). This switch has no impact on the model building process. In particular, you can convert a model for the log-odds of surviving to a model for the log-odds of dying by simply changing the signs of each regression coefficient in the model.

```
data d1;
   infile 'C:\Documents and Settings\ed\My Documents\LONGITUDINALcourse\albuq.bed';
   input id surv a1 a2 a3 a4 a5 a6 b1 b2 b3 b4 b5 b6 c1 c2 c3 c4 c5 c6
                 d1 d2 d3 d4 d5 d6 e1 e2 e3 e4 e5 e6 f1 f2 f3 f4 f5 f6
                 g1 g2 g3 g4 g5 g6 h1 h2 h3 h4 h5 h6 j1 j2 j3 j4 j5 j6
                 iss iciss bp rts age prob;
    as = a3+a4+a5+a6;
    bs = b3+b4+b5+b6;
    cs = c3+c4+c5+c6;
    ds = d3+d4+d5+d6;
    es = e3+e4+e5+e6;
    fs = f3+f4+f5+f6;
    gs = g3+g4+g5+g6;
    hs = h3+h4+h5+h6;
    js = j3+j4+j5+j6;

proc logistic descending;
   model surv = as bs cs ds es fs gs hs js
            iss rts age bp/
            selection=backward scale=none lackfit slstay=.10
            ctable pprob=(.10 to 1 by .10);
run;
```

```
FROM LOG WINDOW: NOTE: PROC LOGISTIC is modeling the probability that surv=1.
```

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	as	1	12	0.0489	0.8250
2	bs	1	11	0.1196	0.7294
3	fs	1	10	0.1468	0.7016
4	cs	1	9	0.6680	0.4137
5	ds	1	8	0.8078	0.3688
6	gs	1	7	1.1229	0.2893
7	hs	1	6	1.1977	0.2738
8	js	1	5	1.7047	0.1917

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	2573	866.0687	0.3366	1.0000
Pearson	2573	2534.2677	0.9849	0.7031

Number of unique profiles: 2579

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.3558	0.4429	0.6454	0.4218
es	1	-0.4613	0.1101	17.5567	<.0001
iss	1	-0.0569	0.00741	58.9887	<.0001
rts	1	0.8431	0.0553	232.1313	<.0001
age	1	-0.0497	0.00529	88.2479	<.0001
bp	1	-0.6351	0.2496	6.4751	0.0109

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
es	0.630	0.508	0.782
iss	0.945	0.931	0.958
rts	2.324	2.085	2.590
age	0.952	0.942	0.961
bp	0.530	0.325	0.864

I only included the summary table from the backward elimination, and information on the fit of the selected model. The final model includes effects for ES (number of severe abdominal injuries), ISS, RTS, AGE, and BP. All of the effects in the selected model are significant at the 5% level.

Letting p be the probability of survival, the estimated survival probability is given by

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = .3558 - .4613\text{ES} - .6351\text{BP} - .0569\text{ISS} + .8431\text{RTS} - .0497\text{AGE}.$$

The p-values for D and X^2 are large, indicating no gross deficiencies with the selected model.

Let us interpret the sign of the coefficients, and the odds ratios, in terms of the impact that individual predictors have on the survival probability.

Checking Predictions for the Trauma Model

To assess the ability of the selected model to accurately predict survival, assume that the two types of errors (a false positive prediction of survival, and a false negative prediction) have equal costs. Under this assumption, the optimal classification rule is to predict survival for an individual with an estimated survival probability of .50 or larger.

$\hat{p} > 0.5$

In the **SAS** program above, the **ctable pprob=** option gives classifications for cutoff probabilities $.1, .2, ..., 1$ based on the selected model. This summary uses the **jackknife** method for assessing classification accuracy. To implement the jackknife method, each observation is temporarily held out and the selected model is fitted to the remaining (3131) cases. This leads to an estimated survival probability for the case, which is compared to the actual result. The results are summarized in terms of total number of correct and incorrect classifications for each possible outcome. In the table below, the columns labeled **Event** (a survival) and **Non-Event** (a death) refer to the classification of observations, and not to the actual outcomes. The columns labeled **Correct** and **Incorrect** identify whether the classifications are accurate.

The data set has 2865 survivors and 267 people that died. Using a .50 cutoff, 2825 of the survivors would be correctly identified, and 40 misclassified. Similarly, 138 of the patients that died would be correctly classified and 129 would not. The overall percentage of cases correctly classified is $100 * (2825 + 138)/3132 = 94.6$. The **false positive rate**, which is the % of those predicted to survive that did not, is 4.4% or 129/(129+2825). The **false negative rate**, which is the % of those predicted to die that did not is $40/(40 + 138) = 22.5\%$. The **sensitivity** is the percentage of survivors that are correctly classified, $2825/(2825 + 40) =$

98.6%. The **specificity** is the percentage of patients that died that are correctly classified, $138/(138 + 129) = 51.7\%$.

The misclassification rate seems small, but you should remember that approximately 10% of patients admitted to UNM eventually die from their injuries. Given this historical information only, you could achieve a 10% misclassification rate by completely ignoring the data and classifying each admitted patient as a survivor. Using the data reduces the misclassification rate by about 50%, which is an important reduction in this problem.

Classification Table

	Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
	0.100	2860	77	190	5	93.8	99.8	28.8	6.2	6.1
	0.200	2853	105	162	12	94.4	99.6	39.3	5.4	10.3
	0.300	2848	117	150	17	94.7	99.4	43.8	5.0	12.7
	0.400	2836	125	142	29	94.5	99.0	46.8	4.8	18.8
**	0.500	2825	138	129	40	94.6	98.6	51.7	4.4	22.5
	0.600	2801	156	111	64	94.4	97.8	58.4	3.8	29.1
	0.700	2773	171	96	92	94.0	96.8	64.0	3.3	35.0
	0.800	2726	195	72	139	93.3	95.1	73.0	2.6	41.6
	0.900	2626	227	40	239	91.1	91.7	85.0	1.5	51.3
	1.000	0	267	0	2865	8.5	0.0	100.0	.	91.5

Historical Example: O-Ring Data

The table below presents data from Dalal, Fowlkes, and Hoadley (1989) on field O-ring failures in the 23 pre-*Challenger* space shuttle launches. Temperature at lift-off and O-ring joint pressure are predictors. The binary version of the data views each flight as an independent trial. The result of a trial is a 1 if at least one field O-ring failure occurred on the flight and a 0 if all six O-rings functioned properly. A regression for the binary data would model the probability that **any** O-ring failed as a function of temperature and pressure. The binomial version of these data is also presented. It views the six field O-rings on each flight as independent trials. A regression for the binomial data would model the probability that

Summary of Logistic Regression classif.

Data

Model		S	D	
	S	2825	129	2954
	D	40	138	178
		2865	267	3132

$$\text{Correctly Classified} = \frac{2825+138}{3132} = 0.946$$

$$\text{False Positive Rate} = \frac{\text{Model} = S \text{ but Data} = D}{\text{Model} = S} = \frac{129}{2954} = 0.044$$

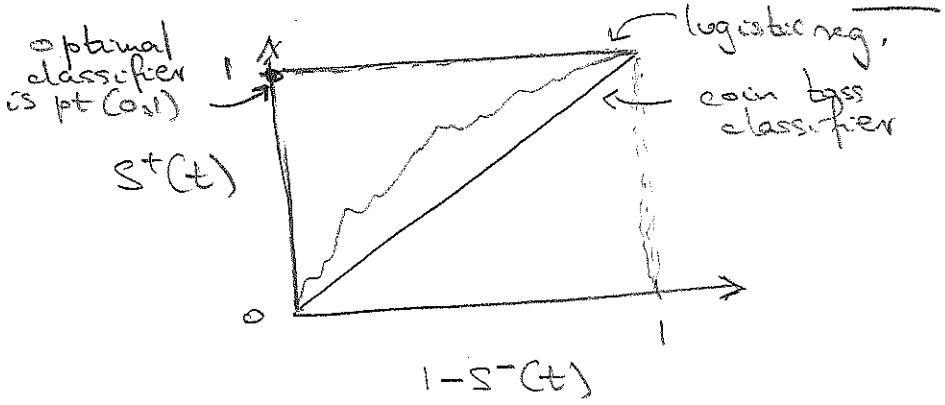
$$\text{False Negative Rate} = \frac{\text{Model} = D \text{ but Data} = S}{\text{Model} = D} = \frac{40}{178} = 0.225$$

$$\underline{\text{Sensitivity}} = \text{proportion of survivors correctly classified} = \frac{2825}{2825+40} = 0.986$$

$$\underline{\text{Specificity}} = \text{prop. of non-survivors correctly classified} = \frac{138}{138+129} = 0.517$$

Goals of Classification

- Simultaneously maximize sensitivity & specificity.
- Sensitivity $= S^+(t)$
  Specificity $= S^-(t)$
  t = threshold cutoff for classifying (e.g. $p > 0.5$)
- Plot $1 - S^-(t)$ vs. $S^+(t)$ — RoC curve



- Summarize RoC curve with AUC = area under curve

$0.5 \le AUC \le 1.$

↑ coin toss classifier

↑ optimal classifier