SAS Birth Weight 2

```
/* bweight2.sas */
options linesize=79 pagesize=500 noovp formdlim='-';
title 'Low Birth Weight Data';
title2 "Logistic Regression: Just Mom's Weight and Race";

proc format; /* Value labels used in data step below  */
     value lowfmt  0 = '2500 g +'  1 = 'Under 2500 g';
     value racefmt 1 = 'White'
                   2 = 'Black'
                   3 = 'Other';
     value ynfmt   0 = 'No'  1 = 'Yes';

data bigbaby;
     infile 'bweight.data' firstobs=2; /* Skip the first line that R uses */
     input id low age lwt race smoke ptl ht ui ftv bwt;
     if race = 2 then r2 = 1; else r2=0;
     if race = 3 then r3 = 1; else r3=0;
     label id    =  'Identification Code'
           low   =  'Low Birth Weight'
           lwt   =  'Weight at Last Period'
           smoke =  'Smoke during Pregnancy'
           ptl   =  'History of Premature Labour (# of times)'
           ht    =  'History of Hypertension'
           ui    =  'Presence of Uterine Irritability'
           ftv   =  'Visits to Doctor During 1st trimester'
           bwt   =  'Birth Weight in Grams'
           r2    =  'Black vs White'
           r3    =  'Other vs White';
/******  Value labels defined above in proc format  ******/
format low lowfmt.;
format race racefmt.;
format ht ui smoke ynfmt.;

proc logistic;
     title3 'Full model';
     model low (event='Under 2500 g') = lwt r2 r3;
     /* Can also say event=last or event=first. Default is alphabetically
        first, which is backwards if Y=1 means Yes and Y=0 means No.  */

proc logistic order=internal descending;
     title3 'Reduced model for testing race';
     model low (event='Under 2500 g') = lwt  / covb;
     /* Covb option gives estimated asymptotic covariance matrix
        of the beta-hat statistics. */


/* Discuss dangers of full-reduced approach with missing data. */
```

Low Birth Weight Data 1
Logistic Regression: Just Mom's Weight and Race
Full model

The LOGISTIC Procedure

Model Information

Data Set	WORK.BIGBABY	
Response Variable	low	Low Birth Weight
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	189
Number of Observations Used	189

Response Profile

Ordered Value	low	Total Frequency
1	Under 2500 g	59
2	2500 g +	130

Probability modeled is low='Under 2500 g'.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	236.672	231.259
SC	239.914	244.226
-2 Log L	234.672	223.259

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.4129	3	0.0097
Score	10.7572	3	0.0131
Wald	10.1316	3	0.0175

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8057	0.8452	0.9088	0.3404
lwt	1	-0.0152	0.00644	5.5886	0.0181
r2	1	1.0811	0.4881	4.9065	0.0268
r3	1	0.4806	0.3567	1.8156	0.1778

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
lwt	0.985	0.973	0.997
r2	2.948	1.133	7.672
r3	1.617	0.804	3.253

Association of Predicted Probabilities and Observed Responses

Percent Concordant	64.1	Somers' D	0.293
Percent Discordant	34.8	Gamma	0.296
Percent Tied	1.1	Tau-a	0.127
Pairs	7670	c	0.647

Logistic Regression: Just Mom's Weight and Race
Reduced model for testing race

The LOGISTIC Procedure

Model Information

Data Set	WORK.BIGBABY	
Response Variable	low	Low Birth Weight
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	189
Number of Observations Used	189

Response Profile

Ordered Value	low	Total Frequency
1	Under 2500 g	59
2	2500 g +	130

Probability modeled is low='Under 2500 g'.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	236.672	232.691
SC	239.914	239.174
-2 Log L	234.672	228.691

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	5.9813	1	0.0145
Score	5.4382	1	0.0197
Wald	5.1921	1	0.0227

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.9983	0.7853	1.6161	0.2036
lwt	1	-0.0141	0.00617	5.1921	0.0227

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
lwt	0.986	0.974	0.998

Association of Predicted Probabilities and Observed Responses

Percent Concordant	60.1	Somers' D	0.226
Percent Discordant	37.4	Gamma	0.232
Percent Tied	2.5	Tau-a	0.098
Pairs	7670	c	0.613

Estimated Covariance Matrix

Parameter	Intercept	lwt
Intercept	0.616679	-0.00474
lwt	-0.00474	0.000038

Watch out for missing values!

If a case (there are *n* cases) has missing data for a variable used in a calculation, (like calculating a mean or fitting a model), that case is automatically excluded. This is almost always what you want, but suppose you are fitting a full and a reduced model, planning to compare them with a likelihood ratio test.

There are two sets of explanatory variables; call them *A* and *B*. The null hypothesis says all the regression coefficients for set *B* equal zero.

A natural approach is to fit a full model including both *A* and *B*, and a reduced model including just *A*. Then

$$G^2 = \text{Deviance(Reduced)} - \text{Deviance(Full)}$$

Any cases with missing values for at least one variable in set *A* are excluded from both calculations -- no problem.

But if a case has missing values in set *B* but not *A*, it will be excluded from calculation of the full model but not the reduced model. Thus *the reduced model will be based on a larger sample size*. But the deviance is a sum of positive terms, one for each case:

$$-2\ell(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} -2 \log p(y_i; \widehat{\boldsymbol{\beta}})$$

So the deviance of the reduced model is too big! It includes terms from the cases that were excluded from the full model but not the reduced model.

This means G^2 is too big, and the probability of wrongly rejecting the null hypothesis (Type I error rate) is greater than α.

One solution is to create special purpose data sets that contain only cases used to calculate the full model. This can be done in the SAS data step (not the raw data file!), but it's clumsy.

Another solution is to create a new binary response variable that is just a copy of the first one, and then make sure the new Y is missing if the values of *any* explanatory variables in the full model are missing. Use the new Y in fitting both the full and reduced models. This is less clumsy, but still not ideal.