BIOS 6244 Analysis of Categorical Data
November 16, 2005
Computer Lab 3 – Part II

Use of SAS PROC LOGISTIC in Performing Logistic Regression

PROC GENMOD is a general-purpose module that can be used to fit many different GLM's. PROC LOGISTIC is specifically tailored to fitting logistic (or logit) models and is capable of providing more extensive output than PROC GENMOD.

The following SAS code (available on the course website) fits a logistic regression (logit) model to the ungrouped horseshoe crab data:

```
proc logistic desc;
   model y = width / clparm = both  clodds= both  covb  lackfit;
   output out=predict  p=pi_hat  lower=LCL  upper=UCL;
   title 'Table 4.2';
   title2 'Logistic Regression Using PROC LOGISTIC';
   title3 'Model Fitted to Original Data';
   run;

proc print  data=predict;
   title 'Table 4.2';
   title2 'Fitted Probabilities for Width Categories';
   run;
```

Note that we can specify the binary variable "Y" as the outcome (LHS of MODEL statement) without also having to specify the denominator variable as in PROC GENMOD. The DESC option in the PROC LOGISTIC statement (short for "descending") tells SAS to order the values of Y from largest to smallest, rather than smallest to largest. It is necessary to do this so that Y = 1 will be used to indicate a "success" rather than Y = 0.

In the MODEL statement, the option CLPARM tells SAS to calculate confidence limits for the parameters in the logistic regression model, and CLODDS tells SAS to calculate confidence limits for the odds ratio. The option BOTH tells SAS to use both the Wald method and the likelihood ratio method to calculate confidence limits for the parameters. (Use PL instead of BOTH to request only the likelihood-ratio method.) As pointed out in class, the likelihood ratio method is generally preferred, although for most logit models that fit the data well, there will be little difference between the two. The option COVB tells SAS to print the estimated

covariance matrix for the estimated model parameters. The option LACKFIT performs the Hosmer-Lemeshow test based on a partition of the fitted probability values.

We are also interested in the fitted probability values for certain carapace widths. The easiest way to obtain these using PROC LOGISTIC is to tell SAS to calculate them for all values of X in the dataset, output them to a new dataset, and then print the dataset. This is accomplished by using the OUTPUT statement, which creates a new SAS dataset called PREDICT. Contained in this new dataset are variables named "PI-HAT" (fitted probability values), "LCL" (lower confidence limit for π_i), and "UCL" (upper confidence limit for π_i).

The relevant SAS output is given below. (The complete SAS output is available on the course website.) We will divide this output up into several pieces and discuss each one separately.

Table 4.2
Logistic Regression Using PROC LOGISTIC
Model Fitted to Original Data

The LOGISTIC Procedure

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	227.759	198.453
SC	230.912	204.759
-2 Log L	225.759	194.453

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	31.3059	1	<.0001
Score	27.8752	1	<.0001
Wald	23.8872	1	<.0001

On pp. 87-88 of the lecture notes, we described the Wald and likelihood-ratio tests of the null hypothesis H_0: $\beta = 0$. We described how the likelihood-ratio test statistic could be obtained by calculating $-2(L_0 - L_1)$, where L_1 is the the max of the log-likelihood

function under the "full model" when β is unrestricted and L_0 is the max of the log-likelihood function when β is assumed to be 0. SAS provides the value of $-2L_0$ in the "Intercept Only" column and $-2L_1$ in the "Intercept and Covariates" column. Subtracting these two values yields the χ^2 test statistic for the likelihood-ratio test (31.3), as presented in our text and on p. 88 in the lecture notes. The value for the Wald test statistic, 23.9, also agrees with the value presented in our text and on p. 87 in the lecture notes.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.3508	2.6287	22.0749	<.0001
width	1	0.4972	0.1017	23.8872	<.0001

Note that the estimated intercept (-12.351) and slope (.497), highlighted in yellow, agree with the values presented in the text and on p. 84 of the lecture notes. SAS provides the results of the Wald significance test for each of the parameters in the model. The result for the likelihood ratio test for β are given under "Testing Global Null Hypothesis: BETA=0" above

Profile Likelihood Confidence Interval for Parameters

Parameter	Estimate	95% Confidence Limits	
Intercept	-12.3508	-17.8097	-7.4573
width	0.4972	0.3084	0.7090

Wald Confidence Interval for Parameters

Parameter	Estimate	95% Confidence Limits	
Intercept	-12.3508	-17.5030	-7.1986
width	0.4972	0.2978	0.6966

SAS can also provide 95% confidence limits for the intercept and slope parameters. The results for the Wald CI agree with the results presented in our text and on p. 87 of the lecture notes. The confidence limits based on the likelihood-ratio method (called "profile likelihood" by SAS) are generally preferred, but differ very little from the Wald limits for these data.

Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
width	1.0000	1.644	1.361	2.032

Wald Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
width	1.0000	1.644	1.347	2.007

Note that the estimated odds ratio and 95% CI agree with the results presented on p. 87 of the lecture notes. As mentioned above, the "profile likelihood" confidence limits are preferred, but again, they differ very little from the Wald limits for these data.

Estimated Covariance Matrix

Parameter	Intercept	width
Intercept	6.910227	-0.26685
width	-0.26685	0.01035

On pp. 88-89 of the lecture notes, we described how one can find a confidence interval for the true $\pi(x)$ when $X = x$. This method required estimates of $\text{Var}(\hat{\alpha})$, $\text{Var}(\hat{\beta})$, and $\text{Cov}(\hat{\alpha}, \hat{\beta})$. Each of these can be obtained from the "Estimated Covariance Matrix" produced by SAS: 6.910, .01035, and -.2669, respectively.

For a given value of X that is contained in the original data set, SAS can also generate the fitted value and a 95% CI for $\pi(x)$. These values are found in the listing requested by the OUTPUT statement in the SAS code given above. The relevant output for X = 26.5 cm is given below:

Fitted Probabilities for Width Categories

Obs	color	spine	width	satell	weight	y	n	dark	_LEVEL_	pi_hat	LCL	UCL
7	1	1	26.5	0	2.35	0	1	1	1	0.69546	0.61205	0.76775

These confidence limits agree with the values presented on p. 89 of the lecture notes.

For the final analysis in this section, we consider the Hosmer-Lemeshow test for goodness-of-fit discussed on pp. 92-93 of the lecture notes. This test is based on a partition of the fitted probability values into deciles (i.e., 10 groups of approximately equal size). The SAS output for this test is given below. The value of the test statistic, 5.2, does *not* agree with the value in our text book (3.5). This is probably a typo.

Partition for the Hosmer and Lemeshow Test

		y = 1		y = 0	
Group	Total	Observed	Expected	Observed	Expected
1	19	5	5.39	14	13.61
2	18	8	7.62	10	10.38
3	17	11	8.62	6	8.38
4	17	8	9.92	9	7.08
5	16	11	10.10	5	5.90
6	18	11	12.30	7	5.70
7	16	12	12.06	4	3.94
8	16	12	12.90	4	3.10
9	16	13	13.69	3	2.31
10	20	20	18.41	0	1.59

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
5.2465	8	0.7309

Use of SAS PROC LOGISTIC in Performing Logistic Regression for Grouped Data

As stated by Agresti in several places in Chapter 5, examination of goodness-of-fit for logistic regression models can be difficult if there are many settings of the (continuous) explanatory variable for which $n_i = 1$ (as is the case with the horseshoe crab data). It is often advantageous to group the values of the explanatory variable and then fit the logistic regression model to the grouped data, after scoring the categories of the grouped explanatory variable in some meaningful way. This approach is described in more detail on p. 92 of the lecture notes.

In Table 5.3 in our text, Agresti presents the results of a logistic regression analysis applied to the grouped horseshoe crab data, in which he scores the width intervals using the mean width for that interval.

Table 5.3 Residuals for Logistic Regression Models Fitted to Grouped Crab Data

Width	Number Cases	Number Yes	Fitted[a] Yes	Pearson[a] Residual	Fitted Yes	Pearson Residual	Adjusted Residual
< 23.25	14	5	8.98	−2.22	3.85	0.69	0.85
23.25–24.25	14	4	8.98	−2.78	5.50	−0.82	−0.93
24.25–25.25	28	17	17.96	−0.38	13.97	1.14	1.35
25.25–26.25	39	21	25.02	−1.34	24.21	−1.06	−1.24
26.25–27.25	22	15	14.12	0.39	15.80	−0.38	−0.42
27.25–28.25	24	20	15.40	1.96	19.16	0.43	0.49
28.25–29.25	18	15	11.55	1.70	15.46	−0.31	−0.36
> 29.25	14	14	8.98	2.80	13.05	1.01	1.14

[a]Independence model, other fitted values and residuals refer to model (5.3.1) with width predictor.

The following SAS code is used to create the dataset for the grouped data and then to carry out a logistic regression analysis:

```
data crabs;
input width cases satell;
cards;
22.69 14  5
23.84 14  4
24.77 28 17
25.84 39 21
26.79 22 15
27.74 24 20
28.67 18 15
30.41 14 14
;
proc logistic desc;
   model satell/cases = width / influence scale = none;
   output out=predict  p=pi_hat  lower=LCL  upper=UCL;
   title 'Table 5.1';
   title2 'Logistic Regression Using PROC LOGISTIC';
   title3 'Model Fitted to Grouped Data';
   run;

proc print  data=predict;
   title 'Table 5.1';
   title2 'Fitted Probabilities for Width Categories';
   run;

proc logistic;
   model satell/cases = / influence  scale = none ;
   title 'Table 5.1';
   title 'Independence Model';
   run;
```

Note that we must use what SAS calls the “events/trials” form of the model statement so that SAS will understand that we have grouped the data. The INFLUENCE option in the model statement tells SAS that we want the influence diagnostics that we discussed in class (Dfbeta, etc.). The SCALE = NONE option tells SAS that we do not want to correct for over-dispersion. A useful by-product of this option is that SAS produces what it calls the “Deviance Table,” which contains the results for the χ^2 and likelihood-ratio GOF tests. The last bit of SAS code fits the “independence model” (i.e., the model in which $\beta = 0$) to the data. The results for the

independence model are needed to reproduce some of the results presented in our text and in the lecture notes; generally, there is no need to fit this model to a set of data, as the diagnostic measures that use it are produced by specifying various options in the MODEL statement used to fit the logistic regression model.

The relevant SAS output is given below. (The complete SAS output is available on the course website.) As in our analysis of the ungrouped horseshoe crab data, we will divide the output up into several pieces and discuss each one separately.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	5.9696	6	0.9949	0.4266
Pearson	5.0289	6	0.8382	0.5401

SAS uses “Deviance” to denote the likelihood ratio GOF test. The values produced by SAS for these 2 GOF tests agree with those given on p. 92 of the lecture notes.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	227.759	201.694
SC	230.912	208.001
-2 Log L	225.759	197.694

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.0644	1	<.0001

As we saw in the analysis of the ungrouped horseshoe crab data described above, the likelihood ratio GOF test statistic for the hypothesis H_0: $\beta = 0$ is found by subtracting the two "-2 Log L" values given under "Model Fit Statistics." As pointed out of p. 94 of the lecture notes, this test statistic can also be found by subtracting the likelihood ratio GOF test statistics (or "deviances") for the logistic regression model and the independence model. As seen in the output on p. 34 of this handout, the deviance for the independence model is 34.034. Subtracting the deviance for the LR model given above (5.9696), we obtain the value (28.06) given in the SAS output above for the likelihood ratio test of H_0: $\beta = 0$. Note that Agresti reports this value as 28.0 in our text.

As we have seen before, the estimated parameters for the LR model are obtained from the following table:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-11.5128	2.5488	20.4031	<.0001
width	1	0.4646	0.0985	22.2312	<.0001

These agree with the values on p. 92 of the lecture notes.

SAS can also produce several measures that are useful in diagnosing the goodness of fit of the fitted model, including the Pearson residuals, Dfbeta, and a statistic *c* based on joint confidence interval estimation of α and β.

Regression Diagnostics

Case Number	Covariates width	Pearson Residual Value	Pearson Residual (1 unit = 0.14) -8 -4 0 2 4 6 8	Deviance Residual Value	Deviance Residual (1 unit = 0.18) -8 -4 0 2 4 6 8	Hat Matrix Diagonal Value	Hat Matrix Diagonal (1 unit = 0.02) 0 2 4 6 8 12 16
1	22.6900	0.6901	| | * |	0.6719	| | * |	0.3458	| *|
2	23.8400	-0.8196	| * | |	-0.8370	| * | |	0.2244	| * |
3	24.7700	1.1443	| | *|	1.1487	| | * |	0.2807	| * |
4	25.8400	-1.0606	| * | |	-1.0485	| * | |	0.2726	| * |
5	26.7900	-0.3772	| * | |	-0.3727	| * | |	0.1741	| * |
6	27.7400	0.4272	| | * |	0.4372	| | * |	0.2551	| * |
7	28.6700	-0.3146	| * | |	-0.3072	| * | |	0.2382	| * |
8	30.4100	1.0113	| | * |	1.4051	| | *|	0.2092	| * |

Case Number	Intercept DfBeta Value	(1 unit = 0.07) -8 -4 0 2 4 6 8	width DfBeta Value	(1 unit = 0.07) -8 -4 0 2 4 6 8	Confidence Interval Displacement C Value	(1 unit = 0.04) 0 2 4 6 8 12 16
1	0.5603	| * |	-0.5410	|* |	0.3848	| * |
2	-0.3956	| * |	0.3740	| * |	0.2505	| * |
3	0.4775	| * |	-0.4295	| * |	0.7102	| * |
4	-0.0365	| * |	-0.0150	| * |	0.5794	| * |
5	0.0821	| * |	-0.0935	| * |	0.0363	| * |
6	-0.2012	| * |	0.2149	| * |	0.0839	| * |
7	0.1646	| * |	-0.1721	| * |	0.0406	| * |
8	-0.5316	|* |	0.5469	| * |	0.3422	| * |

These values agree with those presented in Tables 5.3 and 5.4 in our text. None of the small graphs presented with the diagnostics indicate that any one width category has more influence or appears to be discordant from the others.

In the SAS code given above, we also requested that SAS provide the fitted probability values for each of the carapace width intervals. The output from the PROC PRINT statement is as follows:

Table 5.1
Fitted Probabilities for Width Categories

Obs	width	cases	satell	pi_hat	LCL	UCL
1	22.69	14	5	0.27481	0.15971	0.43036
2	23.84	14	4	0.39268	0.28007	0.51800
3	24.77	28	17	0.49901	0.40218	0.59592
4	25.84	39	21	0.62086	0.53879	0.69655
5	26.79	22	15	0.71801	0.63346	0.78953
6	27.74	24	20	0.79835	0.70525	0.86756
7	28.67	18	15	0.85913	0.76133	0.92102
8	30.41	14	14	0.93192	0.84095	0.97256

We see that if we multiply the fitted probability values (labelled PI_HAT in the output) times the # of crabs in each interval (labelled as CASES in the output), we obtain the fitted frequencies presented in Table 5.3.

We also included some SAS code to fit the independence model to the grouped data. Generally, this is not required when fitting a logistic regression model, unless one needs the results of the χ^2 or likelihood-ratio GOF tests for that model:

Independence Model

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	34.0340	7	4.8620	<.0001
Pearson	29.2766	7	4.1824	0.0001

What SAS calls the "deviance" statistic (same as the likelihood ratio GOF test) is used to perform the likelihood ratio test of H_0: $\beta = 0$, as described on p. 32 of this handout.

Note that the p-value for the likelihood ratio GOF test indicates that the independence model does not fit the data well. This is also indicated by the plot of the Pearson residuals:

Independence Model
The LOGISTIC Procedure

Regression Diagnostics

Case Number	Pearson Residual Value	(1 unit = 0.35) -8 -4 0 2 4 6 8	Deviance Residual Value	(1 unit = 0.44) -8 -4 0 2 4 6 8	Hat Matrix Diagonal Value	(1 unit = 0.01) 0 2 4 6 8 12 16
1	-2.2197	| * | |	-2.1585	| * | |	0.0809	| * |
2	-2.7771	|* | |	-2.7059	| * | |	0.0809	| * |
3	-0.3804	| *| |	-0.3779	| *| |	0.1618	| * |
4	-1.3434	| * | |	-1.3210	| * | |	0.2254	| *|
5	0.3932	| |* |	0.3968	| |* |	0.1272	| * |
6	1.9586	| | * |	2.0815	| | * |	0.1387	| * |
7	1.6962	| | * |	1.8027	| | * |	0.1040	| * |
8	2.7964	| | *|	3.5250	| | *|	0.0809	| * |

NOTE: SAS Output continued on nexst page.

Regression Diagnostics

Case Number	Intercept DfBeta Value	(1 unit = 0.11) -8 -4 0 2 4 6 8	Confidence Interval Displacement C Value	(1 unit = 0.05) 0 2 4 6 8 12 16	Confidence Interval Displacement CBar Value	(1 unit = 0.04) 0 2 4 6 8 12 16
1	-0.6870	| * | |	0.4720	| * |	0.4338	| * |
2	-0.8596	|* | |	0.7388	| *|	0.6791	| *|
3	-0.1826	| * | |	0.0333	| * |	0.0279	| * |
4	-0.8235	|* | |	0.6782	| * |	0.5253	| * |
5	0.1606	| |* |	0.0258	| * |	0.0225	| * |
6	0.8470	| | *|	0.7174	| * |	0.6179	| * |
7	0.6107	| | * |	0.3729	| * |	0.3341	| * |
8	0.8655	| | *|	0.7492	| *|	0.6885	| *|

Case Number	Delta Deviance Value	(1 unit = 0.82) 0 2 4 6 8 12 16	Delta Chi-Square Value	(1 unit = 0.53) 0 2 4 6 8 12 16
1	5.0931	| * |	5.3610	| * |
2	8.0007	| * |	8.3911	| *|
3	0.1708	|* |	0.1727	|* |
4	2.2704	| * |	2.3301	| * |
5	0.1799	|* |	0.1771	|* |
6	4.9507	| * |	4.4541	| * |
7	3.5837	| * |	3.2113	| * |
8	13.1139	| *|	8.5084	| *|

We can also use SAS to perform exact logistic regression when the maximum likelihood estimation procedure fails to converge. The SAS code for accomplishing this for the grouped horseshoe crab data is as follows:

```
proc logistic desc;
   model satell/cases = width;
   exact 'E1' intercept width / estimate;
   title 'Table 5.1';
   title2 'Exact Logistic Regression Using PROC LOGISTIC';
   title3 'Model Fitted to Grouped Data';
   run;
```

Be prepared to wait (perhaps as long as 1 or 2 minutes) for SAS to execute this code. Note that the results are very similar to what we obtained previously for the maximum likelihood estimates (-11.78 vs. -11.51 for $\hat{\alpha}$, .462 vs .465 for $\hat{\beta}$):

Table 5.1
Exact Logistic Regression Using PROC LOGISTIC
Model Fitted to Grouped Data

The LOGISTIC Procedure

Exact Conditional Analysis

Conditional Exact Tests for 'E1'

Effect	Test	Statistic	--- p-Value --- Exact	Mid
Intercept	Score	22.4375	<.0001	<.0001
	Probability	1.556E-6	<.0001	<.0001
width	Score	25.5344	<.0001	<.0001
	Probability	1.56E-10	<.0001	<.0001

Exact Parameter Estimates for 'E1'

Parameter	Estimate	95% Confidence Limits		p-Value
Intercept	-11.7842	-18.7616	-5.1867	<.0001
width	0.4618	0.2766	0.6626	<.0001