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

Use of SAS in Fitting Generalized Linear Models
Reading Assignment: Stokes et al., pp.182-203, 235-239, 349-362

Consider the data in Table 4.1 in our text:

Table 4.1 Relationship Between Snoring and Heart Disease

Snoring	Heart Disease: Yes	Heart Disease: No	Proportion Yes	Linear Fit[a]	Logit Fit	Probit Fit
Never	24	1355	.017	.017	.021	.020
Occasional	35	603	.055	.057	.044	.046
Nearly every night	21	192	.099	.096	.093	.095
Every night	30	224	.118	.116	.132	.131

[a]Model fits refer to proportion of yes responses.
Source: P. G. Norton and E .V. Dunn, *Brit. Med. J., 291*: 630–632 (1985), published by BMJ Publishing Group. See also *Small Data Sets*, D. J. Hand et al., ed. (London: Chapman and Hall, 1994).

In the lecture on October 31, we considered fitting a GLM with binomial random component and identity link to these data. The following SAS code (available on the course website) creates the SAS dataset for this example.

```
data glm;
input snoring disease total;
cards;
0  24  1379
2  35   638
4  21   213
5  30   254
;
```

The following SAS code (available on the course website) fits the GLM with binomial random component and identity link:

```
proc genmod; model disease/total = snoring  /  dist=bin  link=identity  obstats;
title 'Table 4.1';
title2 'Identity Link';
```

The DIST option in the MODEL statement specifies the random component of the GLM and the LINK option specifies the link. The OBSTATS option tells SAS to produce predicted values, residuals, etc. Note that whenever a binomial link is specified, the LHS of the model specified in the MODEL command must be of the form {number of successes}/ {sample size}. In these data, the variable DISEASE contains the number of heart disease cases and the variable TOTAL contains the sample size at each level of the explanatory variable SNORING. Therefore, the LHS of the model is correctly specified by "DISEASE/TOTAL".

The relevant SAS output is given below. (The complete SAS output is available on the course website.)

Table 4.1
Identity Link

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.0172	0.0034	0.0105	0.0240	25.18	<.0001
snoring	1	0.0198	0.0028	0.0143	0.0253	49.97	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

(The SAS output is continued on the following page.)

Table 4.1
Identity Link

The GENMOD Procedure

Observation Statistics

Observation	disease	total	snoring	Pred	Xbeta	Std	HessWgt	Lower	Upper
1	24	1379	0	0.0172466	0.0172466	0.0034369	82089.692	0.0105104	0.0239829
2	35	638	2	0.0568023	0.0568023	0.0053524	11525.472	0.0463117	0.0672929
3	21	213	4	0.096358	0.096358	0.0103975	2496.876	0.0759793	0.1167367
4	30	254	5	0.1161358	0.1161358	0.0130889	2511.0091	0.090482	0.1417897

Note that the estimated intercept (.0172) and slope (.0198), highlighted in yellow, and the fitted (predicted) values at each of the values of the explanatory variable (labelled "snoring") , highlighted in green, agree with the values presented in the text and on p. 71 of the lecture notes.

In the lecture on October 31, we also considered fitting a GLM with binomial random component and logit link. The following SAS code (available on the course website) fits this GLM.

```
proc genmod; model disease/total = snoring  /  dist=bin  link=logit  obstats ;
title 'Table 4.1';
title2 'Logit Link';
run;
```

The relevant SAS output is given below. (The complete SAS output is available on the course website.)

Table 4.1
Logit Link

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-3.8662	0.1662	-4.1920	-3.5405	541.06	<.0001
snoring	1	0.3973	0.0500	0.2993	0.4954	63.12	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Table 4.1
Logit Link

The GENMOD Procedure

Observation Statistics

Observation	disease	total	snoring	Pred	Xbeta	Std	HessWgt	Lower	Upper
1	24	1379	0	0.0205074	-3.866248	0.1662144	27.699783	0.0148906	0.0281823
2	35	638	2	0.0442951	-3.071575	0.104568	27.008489	0.0363854	0.0538283
3	21	213	4	0.0930541	-2.276902	0.1193745	17.976144	0.0750996	0.1147685
4	30	254	5	0.1324388	-1.879565	0.1530077	29.184295	0.1016107	0.1708415

Note that the estimated intercept (-3.87) and slope (.40), highlighted in yellow, and the fitted (predicted) values at each of the values of the explanatory variable (labelled "snoring") , highlighted in green, agree with the values presented in the text. However, there was a typo in the lecture notes on p. 74 where these values were presented that has since been corrected. (The slope was mistakenly presented as -.387.)

Use of SAS in Performing Poisson Regression

In the lecture on November 2, we considered Poisson regression models for the relationship between # of satellites and carapace width in nesting female horseshoe crabs. The following SAS DATA step was used to create the SAS dataset for the horseshoe crab data, which is available in its entirety on the course website at http://www.bios6244.com/.

```
data crab;
input  color spine  width  satell  weight;
```

The following SAS code (available on the course website) fits a Poisson regression model with log link (a *loglinear model*) to these data:

```
proc genmod; model satell = width / dist=poi link=log;
title 'Table 4.2';
title2 'Poisson Regression';
title3 'Log Link';
```

Note that when a Poisson link is used, the outcome variable is specified on the LHS of the model in the MODEL statement, unlike when a binomial link is used.

The relevant SAS output is given below. (The complete SAS output is available on the course website.)

Table 4.2
Poisson Regression
Log Link

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-3.3048	0.5422	-4.3675	-2.2420	37.14	<.0001
width	1	0.1640	0.0200	0.1249	0.2032	67.51	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Table 4.2
Poisson Regression
Log Link

The GENMOD Procedure

Observation Statistics

Observation	satell	width	Pred	Xbeta	Std	HessWgt	Lower	Upper
107	1	26.3	2.7445814	1.0096286	0.047064	2.7445814	2.5027367	3.009796

Note that the estimated intercept (-3.305) and slope (.164), highlighted in yellow, agree with the values presented in the text. However, there was a typo in the lecture notes on p. 77 where these values were presented that has since been corrected. (The slope was mistakenly presented as -3.035.). Note that SAS provides a Wald confidence interval and significance test for each of the parameters in the model.

On p. 78 in the lecture notes, we considered the fitted (predicted) # of satellites for a horseshoe crab with carapace width 26.3 cm. If we specify the OBSTATS option in the MODEL statement of PROC GENMOD, SAS will generate fitted values (and 95% confidence limits) at each value of the explanatory variable (labelled "width" in the above SAS output). The results for the single horseshoe crab with width = 26.3 (obs #107) are highlighted in green in the SAS output above, and agree with the fitted value of 2.74 presented in the text and in the lecture notes. The 95% CI for the true # of satellites for a crab with this width is also highlighted in green. Notice that the observed # of satellites for this crab was 1, which is not contained in the 95% CI. This provides evidence of a poor fit of the Poisson regression model, as we noted from the plot in Figure 4.5 in our text.

We also considered a Poisson regression model in which an identity link was used. The following SAS code (available on the course website) fits this Poisson regression model to these data:

```
proc genmod; model satell = width / dist=poi link=identity;
title 'Table 4.2';
title2 'Poisson Regression';
title3 'Identity Link';
```

This code generates output similar to that presented above for the Poisson regression model with log link and will not be reproduced here. It is available on the course website.

Use of SAS PROC GENMOD in Performing Logistic Regression

In the lecture on November 7, we considered logistic regression models for the relationship between the probability of having at least 1 satellite and carapace width using the data on nesting female horseshoe crabs. A few statements must be added to the SAS DATA step used earlier to create the SAS dataset for the horseshoe crab data in order to create the binary dependent variable:

```
data crab;
input  color spine  width  satell  weight;
if satell>0 then y=1; if satell=0 then y=0; n=1;
```

The SAS statement "n=1;" is used to create the denominator to be used in the MODEL statement in PROC GENMOD.

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

```
proc genmod;
   model y/n = width / dist=bin  link=logit  obstats ;
   title 'Table 4.2';
   title2 'Logistic Regression Using PROC GENMOD';
```

Recall that any time a binomial link is used, the LHS of the model in the MODEL statement must be of the form {number of successes}/ {sample size}. In the DATA step above, we created a new binary variable "Y" that has the value 0 or 1, depending on whether or not the female crab had any satellites, and a new variable "N" that has the value 1. Thus, the LHS of the model in the MODEL statement above will have the values 0 or 1, as required.

The relevant SAS output is given below. (The complete SAS output is available on the course website.)

Table 4.2
Logistic Regression Using PROC GENMOD

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-12.3508	2.6287	-17.5030	-7.1986	22.07	<.0001
width	1	0.4972	0.1017	0.2978	0.6966	23.89	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Table 4.2
Logistic Regression Using PROC GENMOD

The GENMOD Procedure

Observation Statistics

Observation	y	n	width	Pred	Xbeta	Std	HessWgt	Lower	Upper
14	0	1	21	0.129096	-1.908975	0.5166846	0.1124302	0.0510932	0.289813
141	1	1	33.5	0.9866974	4.3064069	0.8042372	0.0131256	0.9387814	0.9972205

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. Note that SAS provides a Wald confidence interval and significance test for each of the parameters in the model. (We will discuss how to obtain likelihood ratio test results later.)

On p. 84 in the lecture notes, we also considered the fitted (predicted) probability of having at least 1 satellite for a horseshoe crab with the min and max carapace widths in the data set, 21.0 and 33.5 cm, respectively. As we saw earlier, if we specify the OBSTATS option in the MODEL statement of PROC GENMOD, SAS will generate fitted values at each value of the explanatory variable (labelled “width” in the above SAS output). The fitted probabilities for the horseshoe crabs with the min and max widths are highlighted in green in the SAS output above, and agree with the fitted values (.129 & .987) presented in the text and in the lecture notes. The wide range of fitted probabilities for the min and max values of width suggests that the LR model provides a reasonably good fit to the horseshoe crab data.