BIOS 6244 Analysis of Categorical Data
Assignment 5 Solutions

1. Consider Exercise 4.4, p. 98.

   (i) Write the SAS code, including the DATA step, to fit the linear probability model and the logit model to the data in Table 2.7 using the scores indicated in Exercise 4.4.

   Solution

```
data infants_pro;
input alcohol malform total;
cards;
0   48  17114
.5  38  14502
1.5 5    793
4   1    127
7   1     38
;

proc genmod; model malform/total = alcohol  /  dist=bin  link=identity  obstats;
title 'Table 2.7';
title2 'Identity Link';

proc genmod; model malform/total = alcohol   /  dist=bin  link=logit  obstats;
title 'Table 2.7';
title2 'Logit Link';
run;
```

   (ii) Use the following SAS output (pp. 2-3) to answer Part (a) of Exercise 4.4. Be sure to interpret the fitted model in the context of the applied problem. Perform an "eyeball" comparison of the observed and fitted probabilities.

   Solution

   The highlighted portion of the SAS output on the following page (provided with the assignment) is used to answer this question.

Table 2.7
Identity Link

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.0025	0.0003	0.0019	0.0032	58.52	<.0001
alcohol	1	0.0011	0.0007	-0.0003	0.0025	2.24	0.1348
Scale	0	1.0000	0.0000	1.0000	1.0000		

Table 2.7
Identity Link

The GENMOD Procedure

Observation Statistics

Observation	malform	total	alcohol	Pred	Xbeta	Std	HessWgt	Lower	Upper
1	48	17114	0	0.0025476	0.0025476	0.000333	7412655.3	0.0018949	0.0032003
2	38	14502	0.5	0.0030912	0.0030912	0.000356	3991270	0.0023934	0.003789
3	5	793	1.5	0.0041784	0.0041784	0.0009754	287181.35	0.0022667	0.0060901
4	1	127	4	0.0068963	0.0068963	0.0027638	21154.239	0.0014794	0.0123132
5	1	38	7	0.0101578	0.0101578	0.0049381	9729.4498	0.0004793	0.0198363

Fitted linear probability model: $\hat{\pi}(x) = .0025 + .0011x$.

Interpretation: For every increase of 1 drink per day in alcohol consumption, the estimated probability of infant malformation is expected to increase.by .0011.

The sample proportions are compared to the fitted probabilities for the linear probability model in Table 1 on the following page:

Table 1

Alcohol Category Score	Observed Proportion	Fitted Proportion (Linear)	Absolute Residual	Fitted Proportion (Logit)	Absolute Residual
0.0	.0028	.0025	.0003	.0026	.0002
0.5	.0026	.0031	.0005	.0030	.0004
1.5	.0063	.0042	.0021	.0041	.0022
4.0	.0079	.0069	.0010	.0091	.0012
7.0	.0263	.0102	.0161	.0231	.0032

The linear probability model appears to fit the data fairly well, except perhaps for the largest category of alcohol consumption.

(iii) Repeat (ii) above for the logit model.

Solution

The highlighted portion of SAS output on the following page (provided with the assignment) is used to answer this question.

Table 2.7
Logit Link

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-5.9605	0.1154	-6.1867	-5.7342	2666.41	<.0001
alcohol	1	0.3166	0.1254	0.0707	0.5624	6.37	0.0116
Scale	0	1.0000	0.0000	1.0000	1.0000		

Table 2.7
Logit Link

The GENMOD Procedure

Observation Statistics

Observation	malform	total	alcohol	Pred	Xbeta	Std	HessWgt	Lower	Upper
1	48	17114	0	0.0025721	-5.960461	0.1154295	43.905528	0.0020524	0.003223
2	38	14502	0.5	0.0030119	-5.802181	0.1045881	43.54645	0.002455	0.0036946
3	5	793	1.5	0.0041288	-5.48562	0.1725498	3.2606549	0.0029476	0.0057807
4	1	127	4	0.0090651	-4.694219	0.4632045	1.1408289	0.0036766	0.0221752
5	1	38	7	0.0231003	-3.744538	0.8342419	0.8575342	0.0045884	0.1081814

Fitted logit model: $\text{logit}\ \hat{\pi}(x) = -5.96 + .32x$.

Interpretation: For every increase of 1 drink per day in alcohol consumption, the odds of infant malformation are expected to increase by a factor of $e^{.3166} = 1.37$.

The sample proportions are compared to the fitted probabilities for the logit model in Table 1 above. The logit model appears to fit the data fairly well in all categories of alcohol consumption.

(iv) Perform a Wald test of significance of the model coefficients for the linear probability model.

Solution

The highlighted portion of the following SAS output provided with the assignment is used to answer this question.

Table 2.7
Identity Link

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.0025	0.0003	0.0019	0.0032	58.52	<.0001
alcohol	1	0.0011	0.0007	-0.0003	0.0025	2.24	0.1348
Scale	0	1.0000	0.0000	1.0000	1.0000		

The results for the Wald tests of the model coefficients are summarized in the following table:

Table 2

Model	**Parameter**	**Estimate**	**X^2**	**df**	**p-value**
Linear	α	.0025	58.52	1	<.0001
	β	.0011	2.24	1	.135
Logit	α	-5.96	2666.41	1	< .0001
	β	.32	6.37	1	.012

For the linear probability model, the test for the intercept coefficient α is significant ($p < .0001$), but the test for the slope coefficient β is not ($p = .135$). We conclude that there is no significant association between alcohol consumption and infant malformation.

(v) Repeat (iv) above for the logit model.

Solution

The following SAS output provided with the assignment is used to answer this question.

Table 2.7
Logit Link

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-5.9605	0.1154	-6.1867	-5.7342	2666.41	<.0001
alcohol	1	0.3166	0.1254	0.0707	0.5624	6.37	0.0116
Scale	0	1.0000	0.0000	1.0000	1.0000		

The results for the Wald tests of the model coefficients are summarized in Table 2 above.

For the logit model, the test for both coefficients are significant: $p < .0001$ for α and $p = .012$ for β. We conclude that there is a significant association between alcohol consumption and infant malformation.

(vi) Using the results of (ii) – (v) above, compare the fits of the linear probability and logit models. Which model fits the data better? Give a reason for your answer.

Solution

In terms of fitted probabilities, an eyeball comparison indicates that the linear and logit models fit the data equally well in the 1st four categories of alcohol consumption (the linear model has smaller absolute residuals in 2 categories and the logit model has smaller absolute residuals in the other 2). However, in the highest category, the logit model fits much better in terms of the unadjusted absolute residual (.0032 vs. .0161).

In terms of significance tests of the individual model coefficients, the logit model is preferred since both coefficients are statistically significant. In the linear model, only the intercept parameter is significant.

(vii) Based on your answer to (vi) above, find an approximate 95% CI for the true probability of an infant malformation among mothers who drink ≥ 6 drinks per day, on average.

Solution

The highlighted portion of the following SAS output provided with the assignment is used to answer this question.

Table 2.7
Logit Link

The GENMOD Procedure

Observation Statistics

Observation	malform	total	alcohol	Pred	Xbeta	Std	HessWgt	Lower	Upper
1	48	17114	0	0.0025721	-5.960461	0.1154295	43.905528	0.0020524	0.003223
2	38	14502	0.5	0.0030119	-5.802181	0.1045881	43.54645	0.002455	0.0036946
3	5	793	1.5	0.0041288	-5.48562	0.1725498	3.2606549	0.0029476	0.0057807
4	1	127	4	0.0090651	-4.694219	0.4632045	1.1408289	0.0036766	0.0221752
5	1	38	7	0.0231003	-3.744538	0.8342419	0.8575342	0.0045884	0.1081814

Thus, an approximate 95% CI[$\pi(7)$] is given by (.005, .108).

2. Consider Exercise 4.9, p. 99.

(i) Write the SAS code, including the DATA step, to answer Parts (a) – (c) of this Exercise. You need not reproduce all of the data lines, but the INPUT statement and all other necessary statements in the DATA step are required. (Note that in the horseshoe data set given on the course website, weight is recorded in grams, rather than kilograms.)

Solution

```
data crab;
input  color spine  width  satell  weight;
weight = weight/1000;
cards;
3  3  28.3  8  3050
...
;
```

(SAS Code continued on following page.)

```
proc genmod; model satell = weight / dist=poi link=log  obstats;
title 'Table 4.2';
title2 'Poisson Regression';
title3 'Log Link';
title4 '# of satellites vs. weight';
run;
```

(ii) Use the SAS output below to answer parts (a) – (c) of this Exercise. In part (a), only give a point estimate for the mean # of satellites. In Part (b), be sure to interpret $\hat{\beta}$ in the context of the applied problem. For the confidence interval requested in Part (b), use the Wald interval.

Solution

(a) The highlighted portion of the following SAS output provided with the assignment is used to answer this question.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.4284	0.1789	-0.7791	-0.0777	5.73	0.0167
weight	1	0.5893	0.0650	0.4619	0.7167	82.15	<.0001

$\log \hat{\mu}(x) = -.43 + .59x$

$\log \hat{\mu}(2.44) = -.4284 + .5893(2.44) = 1.0095$. So, we estimate that there will be $\hat{\mu}(2.44) = e^{1.0095} = 2.74 \rightarrow 3$ satellites for a female horseshoe crab weighing 2.44 kg.

(b) The highlighted portion of the following SAS output provided with the assignment is used to answer this question.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.4284	0.1789	-0.7791	-0.0777	5.73	0.0167
weight	1	0.5893	0.0650	0.4619	0.7167	82.15	<.0001

$\hat{\beta} = .5893 \Rightarrow$ for each 1 kg increase in weight, the predicted # of satellites will increase by a factor of $e^{.5893} = 1.80$.

From the SAS output above, an approximate 95% CI(β) = (.4619, .7167), so for every 1 kg increase in weight, we are 95% sure that the # of satellites will increase by a factor somewhere between $(e^{.4619}, e^{.7167}) = (1.59, 2.05)$.

(c) The highlighted portion of the following SAS output provided with the assignment is used to answer this question.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.4284	0.1789	-0.7791	-0.0777	5.73	0.0167
weight	1	0.5893	0.0650	0.4619	0.7167	82.15	<.0001

Thus, $X^2 = 82.15$, df = 1, $p < .0001$. So, we reject H_0: $\beta = 0$ and conclude that the # of satellites is not independent of weight.

3. Consider Exercise 5.1, p. 135. Use the SAS output below to answer parts (a) – (c) of this Exercise. Note that in Part (b), Agresti is asking you to use extrapolation to answer the question concerning thermal distress at 31^o. In Part (c), compare the results of the Wald and likelihood-ratio tests and comment.

Solution

(a) The highlighted portion of the SAS output below (provided with the assignment) is used to answer this question.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	15.0429	7.3786	4.1563	0.0415
temp	1	-0.2322	0.1082	4.6008	0.0320

Fitted logistic regression model: $\text{logit}\ \hat{\pi}(x) = 15.04 - .23x$

Interpretation: For every 1^o F increase in temperature at the time of the flight, the odds of thermal distress in an O ring.are expected to decrease by a factor of $e^{-.2322} = .79$.

(b) From the coefficients of the fitted model (highlighted in above SAS output), we see that

$\text{logit}\ \hat{\pi}(31) = 15.0429 - .2322(31) = 7.8447$

$\Rightarrow\ \hat{\pi}(31) = \dfrac{e^{.78447}}{1+e^{.78447}} = .9996$. Therefore, the predicted probability of thermal distress at 31^o F is .9996.

$\hat{\pi}(x) = .5$ at $x = -\dfrac{\hat{\alpha}}{\hat{\beta}} = \dfrac{15.0429}{.2322} = 65^o$ F. Therefore, the predicted probability = .5 at $x = 65^o$ F.

A linear approximation for the change in $\hat{\pi}(x)$ per 1^o F increase in temperature at $x = 65^o$ F is given by $\hat{\beta}\hat{\pi}(65)[1-\hat{\pi}(65)] = (-.2322)(.5)(1-.5) = -.058$. Therefore, the predicted probability of thermal distress is expected to decrease by .058 for each 1^o F increase in temperature for temperatures around 65^o F.

(c) The highlighted portions of the SAS output below (provided with the assignment) are used to answer this question.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
temp	0.793	0.641	0.980

$\widehat{OR}$ =.79 $\Rightarrow$ that for every 1^o F increase in temperature, the odds of thermal distress in at least 1 of the O-rings decrease by a factor of .79 (or, for every 1^o F *decrease* in temperature, the odds of thermal distress in at least 1 of the O-rings increase by a factor of 1/.79 = 1.27.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.9520	1	0.0048
Wald	4.6008	1	0.0320

The results for Wald and likelihood-ratio tests of H_0: $\beta = 0$ are obtained from the highlighted portion of the SAS output above (provided with the assignment) and are summarized in the following table:

Table 3

Test	**X^2**	**df**	**p-value**
Wald	4.60	1	.032
Likelihood-ratio	7.95	1	.005

The results for the likelihood-ratio test are much more significant (p = .005 vs. p = .032). This illustrates the benefits of using the likelihood-ratio test.

4. Consider Exercise 5.7, pp. 136-137.

(i) Write the SAS code, including the DATA step, to fit a logistic regression model to these data and produce the SAS output required to answer the questions in Part (ii) below.

Solution

```
data smoking;
input cigs cases at_risk;
cards;
0    90  436
7.5  57  148
19.5 65  113
30   40  58
;
proc logistic desc;
   model cases/at_risk = cigs / clparm = pl  influence  scale = none;
   title 'Table 5.11';
   title2 'Logistic Regression Using PROC LOGISTIC';
   run;
```

(ii) Use the SAS output on the following pages to answer the following questions:

(a) Perform the likelihood-ratio goodness-of-fit test for the logistic regression model.

Solution

The highlighted portion of the SAS output below (provided with the assignment) is used to answer this question.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	2.9953	2	1.4976	0.2237

Thus, $X^2 = 3.00$, df = 2, p = .224. Since .224 > .05, we conclude that the logistic regression model fits the data well.

(b) Give the parameter estimates for the logistic regression model fit to these data.

Solution

The highlighted portion of the SAS output below (provided with the assignment) is used to answer this question.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.2627	0.1050	144.5010	<.0001
cigs	1	0.0771	0.00849	82.3915	<.0001

Thus, the fitted model is given by logit $\hat{\pi}(x) = -1.26 + .077x$

(c) Use the likelihood ratio method to test the null hypothesis H_0: $\beta = 0$ and find a 95% CI(β).

Solution

The highlighted portions of the SAS output below (provided with the assignment) are used to answer this question.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	91.3777	1	<.0001

Thus, $X^2 = 91.38$, df = 1, $p < .0001$. Since $p < .05$, conclude that MI is not independent of average # of cigarettes smoked per day.

Profile Likelihood Confidence Interval for Parameters

Parameter	Estimate	95% Confidence Limits	
Intercept	-1.2627	-1.4723	-1.0602
cigs	0.0771	0.0607	0.0941

Thus, an approximate 95% CI(β) is given by (.06, .09).

(d) Obtain the Pearson residuals and other diagnostic measures (Dfbeta, etc.) that we discussed in class (pp. 95-98 of the lecture notes). Do any of these measures indicate lack of fit of the model? Give a reason for your answer.

Solution

The highlighted portions of the SAS output below (provided with the assignment) are used to answer this question.

Regression Diagnostics

Case Number	Covariates cigs	Pearson Residual Value	Pearson Residual (1 unit = 0.16) -8 -4 0 2 4 6 8	Deviance Residual Value	Deviance Residual (1 unit = 0.16) -8 -4 0 2 4 6 8	Hat Matrix Diagonal Value	Hat Matrix Diagonal (1 unit = 0.05) 0 2 4 6 8 12 16
1	0	-0.7093	| * | |	-0.7149	| * | |	0.8269	| *|
2	7.5000	1.2850	| | *|	1.2708	| | *|	0.2246	| * |
3	19.5000	0.3278	| | * |	0.3283	| | * |	0.4715	| * |
4	30.0000	-0.8898	| * | |	-0.8726	| * | |	0.4770	| * |

(SAS Output continued on next page.)

Regression Diagnostics

Case Number	Intercept DfBeta Value	(1 unit = 0.47) -8 -4 0 2 4 6 8	cigs DfBeta Value	(1 unit = 0.29) -8 -4 0 2 4 6 8	Confidence Interval Displacement C Value	(1 unit = 0.87) 0 2 4 6 8 12 16
1	-3.7267	|* | |	2.3066	| | *|	13.8884	| *|
2	0.6245	| |* |	-0.0125	| * |	0.6167	| * |
3	0.00825	| * |	0.3294	| |* |	0.1814	|* |
4	0.2991	| |* |	-1.0778	| * | |	1.3810	| * |

Regression Diagnostics

Case Number	Confidence Interval Displacement CBar Value	(1 unit = 0.15) 0 2 4 6 8 12 16	Delta Deviance Value	(1 unit = 0.18) 0 2 4 6 8 12 16	Delta Chi-Square Value	(1 unit = 0.18) 0 2 4 6 8 12 16
1	2.4038	| *|	2.9149	| *|	2.9069	| *|
2	0.4782	| * |	2.0931	| * |	2.1295	| * |
3	0.0959	| * |	0.2036	| * |	0.2033	| * |
4	0.7222	| * |	1.4837	| * |	1.5141	| * |

Diagnostic	Indication
Pearson residuals	No apparent lack of fit
Dfbeta	The model does not appear to fit the data very well for the 0 cigs/day category (Dfbeta = 2.3).
c	Same as for Dfbeta, but even more severe: *c* = 13.9 for the 0 cigs/day category, whereas all of the other c values are in the range .2 – 1.4.

Thus, 2 of the 3 diagnostic measures that we examined here indicate possible lack of fit. Perhaps other GLM's should be considered – a plot of the fitted probabilities vs. the scores for the smoking categories might suggest such a model. Alternatively, a dummy variable for the 0 cigs/day category could be incorporated into the logit model in an attempt to accommodate what appears to be a influential observation.