BIOS 6244 Analysis of Categorical Data
November 28, 2005 Lecture

Logit Models for Qualitative Predictors (Sec. 5.4)

Unfortunately, we do not have time this semester to cover this section. It deals with the case in which all of the explanatory variables in the logistic regression (or logit) model are nominal. (Such variables are usually called *factors*.) Agresti demonstrates that logit models with factors are very similar to analysis of variance (ANOVA) models, except that the outcome variable is binary instead of normally distributed. You should read this section and pay close attention to the examples. However, I will not test you on this material.

Multiple Logistic Regression (Sec. 5.5)

This section contains a more general treatment of LR models, in which the explanatory variables can be continuous, ordinal, nominal, and/or binary.

Let $X_1, X_2, \ldots, X_k$ denote the set of explanatory variables and let Y denote the binary response variable. The LR (or logit) model we will be considering in this section is of the form

$$\text{logit}(\pi) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_{k.}$$

Each parameter (or *coefficient*) β_i measures the effect of X_i on the log odds that Y = 1, *controlling for* the other X's. Another way to say this is that β_i measures the change in the log odds corresponding to a 1-unit increase in X_i, *assuming that all other X's remain fixed.* A major advantage of using multiple logistic regression models is that e^{β_i} is the odds ratio for X_i, *adjusted for the effect of all of the other X's*. It measures the multiplicative effect on the odds of Y = 1 corresponding to a 1-unit increase in X_i, assuming that all other X's remain fixed.

Horseshoe Crab Example Using Color & Width as Predictors (Sec. 5.5.1)

Consider a LR model with dependent variable Y defined as before (Y = 1 if the female crab has ≥ 1 satellite, and Y = 0 otherwise) and explanatory variables carapace width and carapace color.

Horseshoe crabs are categorized in terms of color as light, medium light, medium, medium dark and dark. (Color is a surrogate for age, as older crabs tend to be darker.) The data set contained in Table 4.2 contained no light crabs, so only the 4 darker categories will be included in our LR models.

Agresti first treats color as if it were a nominal variable. This requires that 3 *dummy* variables be included in the model:

$$\text{logit}(\pi) = \alpha + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 X \qquad (23)$$

where X = carapace width
C_1 = 1 for medium light color, 0 otherwise
C_2 = 1 for medium color, 0 otherwise
C_3 = 1 for medium dark color, 0 otherwise.

Note that we do not require a dummy variable for dark color since this corresponds to $C_1 = C_2 = C_3 = 0$.

The following SAS code using PROC GENMOD will fit the model in Equation (23) to the horseshoe crab data:

```
data crab;
input  color spine  width  satell  weight;
if satell>0 then y=1; if satell=0 then y=0; n=1;
color = color - 1;
cards;
...
;
proc format;
value colorfmt
      1='med light'
      2='medium'
      3='med dark'
      4='dark';
proc genmod;  class color / order = internal;
     model y/n = color width / dist=bin link=logit;
     format color colorfmt.;
title 'Table 4.2';
title2 'Logistic Regression Model with Width and Color';
title3 'Color Treated as Nominal';
run;
```

The variable COLOR was initially coded using 1 = light, 2 = medium light, etc. Since there were no light-colored crabs in this particular data set, Agresti subtracts 1 from the value of COLOR so that 1 = medium light, 2 = medium, etc. PROC FORMAT is used so that we can label the values of COLOR in the output from PROC GENMOD using the FORMAT statement. The CLASS statement (just as in PROC ANOVA or PROC GLM) tells SAS that the variable COLOR is to be treated as nominal. The ORDER = INTERNAL option tells SAS to order the values of COLOR as originally entered (i.e., 1, 2, 3, 4), and not according to their formatted values (dark, med dark, etc.).

The relevant output produced by the SAS code above is given on the following page:

Table 4.2

Logistic Regression Model with Width and Color
Color Treated as Nominal

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-12.7151	2.7618	-18.1281	-7.3021	21.20	<.0001
color	med light	1	1.3299	0.8525	-0.3410	3.0008	2.43	0.1188
color	medium	1	1.4023	0.5484	0.3274	2.4773	6.54	0.0106
color	med dark	1	1.1061	0.5921	-0.0543	2.2666	3.49	0.0617
color	dark	0	0.0000	0.0000	0.0000	0.0000	.	.
width		1	0.4680	0.1055	0.2611	0.6748	19.66	<.0001

So, for dark crabs, $\log it(\hat{\pi}) = -12.715 + .468x$.
For medium light crabs, $\log it(\hat{\pi}) = (-12.715 + 1.330) + .468x = -11.385 + .468x$.

Note that only the coefficients for medium color and width are statistically significant in the fitted model. This poses a problem since if we delete C_1 and C_3 from the model, all crabs except those of medium color will be lumped together. Alternative models that allow us to get around this problem will be considered later in this section.

The model in Equation (23) that we just fit to the data assumes that there is no *interaction* between color and width in terms of their effect on the probability of having at least 1 satelllite. In other words, width is assumed to have the same effect ($\hat{\beta}_4 = .468$) regardless of color, so the curves relating π to width for each color separately are all parallel (Fig. 5.4).

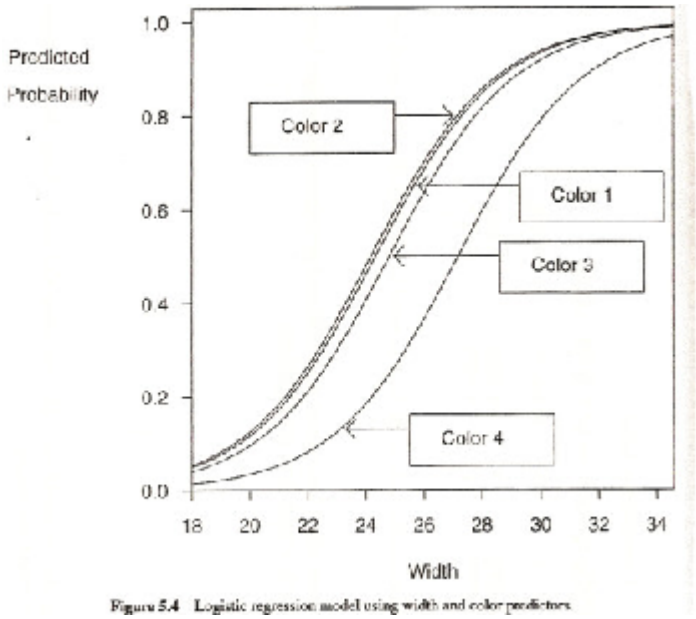

Figure 5.4 Logistic regression model using width and color predictors

For each color, a 1 cm increase in carapace width has a multiplicative effect of $e^{.468} = 1.60$ on the odds that Y = 1.

The positive association between width and the probability of having at least 1 satellite seen in Fig. 5.4 is similar to what we saw in Fig 5.3 when width was the only explanatory variable (p. 96 of these lecture notes).

We can calculate fitted probabilities using multiple logistic regression model equations just as we did for a simple LR model (i.e., one containing a single explanatory variable).

For example, for a medium light crab of average width (26.3 cm), the fitted probability is

$$\hat{\pi} = \frac{e^{-11.385+.468(26.3)}}{1+e^{-11.385+.468(26.3)}} = .715.$$

By comparison, a medium dark crab of average width has fitted probability

$$\hat{\pi} = \frac{e^{-11.609+.468(26.3)}}{1+e^{-11.609+.468(26.3)}} = .669.$$

An additional advantage of using multiple logistic regression models is that the exponentiated difference between 2 color parameter estimates is the odds ratio comparing these colors.

For example, for comparing medium light to medium dark crabs, the parameter difference is

$$\hat{\beta}_1 - \hat{\beta}_3 = 1.330 - 1.106 = .224.$$

At any given width, the estimated odds that a medium light crab will have a satellite are $e^{.224} = 1.25$ times the estimated odds for a medium dark crab. For example, using the fitted probabilities we just calculated for a crab of average width, the odds are $\frac{.715}{.285} = 2.51$ for a medium light crab and $\frac{.669}{.331} = 2.02$ for a medium dark crab, yielding an OR of 1.24, which agrees to within roundoff error with the value of OR = 1.25 given above. Thus, medium dark crabs are less likely than medium light crabs to have satellites.

<u>Model Comparisons (Sec. 5.5.2)</u>

One can use the likelihood-ratio method to test hypotheses about parameters in multiple logistic regression models by taking differences of deviances. For example, to test whether color makes a significant contribution to the model in Equation (23), we test

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0. \qquad (24)$$

This null hypothesis is properly interpreted as "controlling for width, the probability of having at least 1 satellite is independent of color."

To test H_0 in Equation (24), we compare the maximized log likelihood L_1 for the full model in Equation (23) with the maximized log likelihood L_0 for the simpler model in which $\beta_1 = \beta_2 = \beta_3 = 0$. This yields a test statistic of $-2(L_1 - L_0) = 7.00$, df = 3, p = .072. This is not statistically significant, but provides weak evidence of a color effect. (Sometimes we say that the data are *suggestive* of a color effect.) This is not surprising since 2 of the 3 coefficients for color were not statistically significant, as we saw in the output from PROC GENMOD. The following output from PROC GENMOD can be used to perform the likelihood-ratio test that we just described. It was generated by running PROC GENMOD with both WIDTH & COLOR as explanatory variables, and then running PROC GENMOD again with WIDTH as the only explanatory variable.

Table 4.2
Logistic Regression Model with Width and Color
Color Treated as Nominal

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	168	187.4570	1.1158

Table 4.2
Logistic Regression Model with Width Only
Color Treated as Nominal

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	171	194.4527	1.1372

Thus, the test statistic for the likelihood-ratio test just described is $-2(L_1 - L_0) = 194.4527 - 187.4570 = 6.9957$ (rounded to 7.00) and the df are 171 – 168 = 3, as given above.

More generally, one can compare maximized log-likelihoods for any pair of models in which one model is a special case of the other. Such a comparison can be used to determine whether the logistic regression model requires interaction terms, e.g., one can test whether adding the interaction between width and color to the model in Equation (23) will result in a better-fitting model. This more complex model would allow for a different width effect for each carapace color. It would have 3 additional terms, one for the cross-product of width with each dummy variable for color. Fitting this model is equivalent to fitting a logistic regression model with width as the only explanatory variable separately for the crabs in each of the 4 color categories. For each color, there would then be a different S-shaped curve relating carapace width to the probability of having ≥ 1 satellite (and these curves would be likely

to cross). Thus, a comparison of 2 colors of horseshoe crabs would vary according to the width of the carapace. The following SAS code would be used to fit this more complex model:

```
proc genmod; class color;
     model y/n = color width color * width / dist=bin link=logit;
title 'Table 4.2';
title2 'Logistic Regression Model with Width and Color';
title3 'Color Treated as Nominal';
title4 'Interaction Term Included';
run;
```

Note that the interaction term is denoted by COLOR * WIDTH in the MODEL statement.

The relevant SAS output for performing the likelihood ratio test for testing the need for the interaction term is as follows:

Table 4.2
Logistic Regression Model with Width and Color
Color Treated as Nominal
Interaction Term Included

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	165	183.0806	1.1096

Thus, the test statistic for the likelihood-ratio test just described is $-2(L_1 - L_0) = 187.4570 - 183.0806 = 4.3764$ (rounded to 4.38) and the df are 168 - 165 = 3, p = .223. Thus, we conclude that there is no need for the interaction terms. This makes sense when we examine the significance tests for each of the interaction coefficients; none of them are significant, as indicated below:

Table 4.2
Logistic Regression Model with Width and Color
Color Treated as Nominal
Interaction Term Included

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-5.8538	6.6939	-18.9737	7.2660	0.76	0.3818
color	med light	1	4.1012	13.2753	-21.9179	30.1204	0.10	0.7574
color	medium	1	-4.1861	7.5809	-19.0445	10.6722	0.30	0.5808
color	med dark	1	-15.6642	9.5606	-34.4027	3.0743	2.68	0.1013
color	dark	0	0.0000	0.0000	0.0000	0.0000	.	.
width		1	0.2004	0.2617	-0.3124	0.7133	0.59	0.4437

(SAS Output continued on next page.)

width*color	med light	1	-0.0944	0.5004	-1.0752	0.8864	0.04	0.8503
width*color	medium	1	0.2184	0.2952	-0.3602	0.7971	0.55	0.4594
width*color	med dark	1	0.6579	0.3753	-0.0777	1.3936	3.07	0.0796
width*color	dark	0	0.0000	0.0000	0.0000	0.0000	.	.

Quantitative Treatment of Ordinal Predictors (Sec. 5.5.3)

Color for female horseshoe crabs as defined above is actually an ordinal variable (medium light < medium < medium dark < dark). As we have seen previously in this course, one should always make use of the fact that an explanatory variable is ordinal whenever possible. This will generally result in a more parsimonious model (i.e., one with fewer terms) and tests of an ordinal predictor generally have greater power than tests based on the corresponding nominal predictor.

In order to be able to treat color as an ordinal variable in the logistic regression analysis, we must score the color categories. As we saw previously, Agresti uses 1 = medium light, 2 = medium, 3 = medium dark, and 4 = dark. He fits the following logistic regression model, where C is the color variable with values 1, 2, 3, 4:

$$\text{logit}(\pi) = \alpha + \beta_1 C + \beta_2 X. \qquad (25)$$

The following SAS code fits this model to the horseshoe crab data:

```
proc genmod;
     model y/n = color width / dist=bin link=logit;
title 'Table 4.2';
title2 'Logistic Regression Model with Width and Color';
title3 'Color Treated as Ordinal';
run;
```

Note that the lack of a CLASS statement tells SAS to treat COLOR as numeric.

The relevant SAS output is as follows:

Table 4.2
Logistic Regression Model with Width and Color
Color Treated as Ordinal

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	170	189.1212	1.1125

(SAS output continued on next page.)

Table 4.2
Logistic Regression Model with Width and Color
Color Treated as Ordinal

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-10.0708	2.8069	-15.5722	-4.5695	12.87	0.0003
color	1	-0.5090	0.2237	-0.9475	-0.0706	5.18	0.0229
width	1	0.4583	0.1040	0.2544	0.6622	19.41	<.0001

Thus, the fitted model is $\text{logit}(\hat{\pi}) = -10.071 - .509c + .458x$. All coefficients in this model are statistically significant. At a given width, for every 1-category increase in color darkness, the estimated odds of having at least 1 satellite are multiplied by $e^{-.509} = .60$. For example, the estimated odds for a medium-colored crab to have at least 1 satellite are 60% of those for a medium-light crab.

One can compare the model in Equation (25) with the more complex one in Equation (23) using the likelihood-ratio test. The test statistic is -2($L_1 - L_0$) = 189.1212 - 187.4570 = 1.66, df = 170-168 = 2, p = .436. Thus, the simpler model in Equation (25) is preferred. This model comparison is equivalent to testing the hypothesis that the color parameters {β_1, β_2, β_3, 0} in the model in Equation (23}, when plotted against the scores {1, 2, 3, 4}, follow a linear trend.

As indicated in the SAS output on p. 101 of these lecture notes, the estimates of the color parameters in the model in Equation (23} are {1.33, 1.40, 1.11, 0}. The above likelihood-ratio test indicates that these values do not depart significantly from a linear trend; however, the 1st 3 values are very similar when compared with the 4th. This suggests another possible set of scores for the color categories: {1, 1, 1, 0}. In other words, we could redefine the color variable in the model in Equation (25) as C = 0 for dark-colored crabs and C= 1 otherwise.

The SAS code for fitting this new model is given below:

```
data crab;
input  color spine  width  satell  weight;
if satell>0 then y=1; if satell=0 then y=0; n=1;
color = color - 1;
if color=4 then lighter=0; if color < 4 then lighter=1;
cards;
...
proc genmod;
     model y/n = lighter width / dist=bin link=logit;
title 'Table 4.2';
title2 'Logistic Regression Model with Width and Color';
title3 'Color Treated as Binary – Lighter vs. Dark';
run;
```

The relevant SAS output for performing the likelihood-ratio test is as follows:

Table 4.2
Logistic Regression Model with Width and Color
Color Treated as Binary - Lighter vs. Dark

The GENMOD Procedure

Criterion	DF	Value	Value/DF
Deviance	170	187.9579	1.1056

The likelihood-ratio test statistic for comparing this simplified model with that in Equation (25) is 187.9579 - 187.4570 = .5009, df = 170 - 168 = 2, p = .779, indicating that the simpler model with a binary color variable provides an adequate fit.

The fitted model produced by SAS is as follows:

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-12.9795	2.7272	-18.3248	-7.6342	22.65	<.0001
lighter	1	1.3005	0.5259	0.2698	2.3312	6.12	0.0134
width	1	0.4782	0.1041	0.2741	0.6823	21.08	<.0001

Thus, at any given width, the estimated odds that a lighter-colored crab has a satellite are $e^{1.301} = 3.67$ times the estimated odds for a dark crab.

Note that, in all of the logistic regression models that included some variable representing color along with carapace width, the estimated coefficient for width did not change very much, as indicated in the following table:

Color Variable	$\hat{\beta}$	**95% CI(β) [Wald method]**
None	0.50	(0.30, 0.70)
Nominal	0.47	(0.26, 0.67)
Ordinal	0.46	(0.25, 0.66)
Binary	0.48	(0.27, 0.68)

In Epi studies, we would say that the *independent effect* of carapace width on the probability of having at least one satellite is an odds ratio of approximately $e^{.50} = 1.65$. In other words, the OR for carapace width varies very little when color of the crab is taken into account. Thus, carapace color would not be a *clinically significant* confounder of the association between carapace width and probability of having at least 1 satellite.