BIOS 6244 Analysis of Categorical Data
November 30, 2005 Lecture

Model Selection with Several Parameters (Sec.5.5.4)

The horseshoe crab data set in Table 4.2 includes 4 potential predictors of having at least 1 satellite: color, spine condition, weight, and carapace width. Most likely, there were valid scientific reasons for including each of these predictors in the data set, so we want to consider the possibility of including each of them in the "final" logistic regression model.

Variable selection in multiple logistic regression is controversial, just as it is in multiple regression. Several methods have been proposed, but none of them has been established as being "best." The challenges that arise in variable selection in multiple regression also arise in logistic regression. Probably the most important of these is *multicollinearity*, which can occur when 2 or more highly correlated predictors are included in the same model. Multicollinearity is especially problematic when trying to determine the statistical significance of predictors, as it can result in non-significant tests of parameters that should, in fact, be retained in the model.

As an illustration, consider a LR model that includes all 4 potential predictors of having at least 1 satellite, where color and spine condition are treated as nominal (i.e., as *factors*). This model is given by

$$\text{logit}(\pi) = \alpha + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 S_1 + \beta_5 S_2 + \beta_6 X_1 + \beta_7 X_2 \qquad (26)$$

where
X_1 = carapace width
X_2 = weight
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.
S_1 = 1 if both spines are good, 0 otherwise
S_2 = 1 if one spine is worn or broken, 0 otherwise.

Note that spine condition has 3 levels, the 3rd being "both spines are worn or broken."

The following SAS code would be used to fit this model to the data in Table 4.2:

```
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;
...
;
```

(SAS Code continued on next page.)

```
proc format;
value colorfmt
      1='med light'
      2='medium'
      3='med dark'
      4='dark';

proc format;
value spinefmt
      1='both good'
      2='1 worn or broken'
      3='both worn or broken';

proc genmod; class color spine / order = internal;
     model y/n = color spine width weight / dist=bin link=logit;
     format color colorfmt.;
     format spine spinefmt.;
title 'Table 4.2';
title2 'Logistic Regression Model with All Predictors';
title3 'Color and Spine Condition Treated as Nominal';
run;

proc genmod;
   model y/n =  / dist=bin  link=logit ;
title 'Table 4.2';
title2 'Independence Model';
run;
```

The relevant output produced by the SAS code above is as follows:

Table 4.2
Logistic Regression Model with All Predictors
Color and Spine Condition Treated as Nominal

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	165	185.2020	1.1224

To perform the required likelihood-ratio test, we must subtract the deviance for the model in Equation (26) from that of the independence model. We saw earlier that the deviance for the independence model was 225.7585 (df = 172). Thus, the likelihood-ratio test statistic is $-2(L_1 - L_0) = 225.7585 - 185.2020 = 40.56$, df = 172-165 = 7, $p < .0001$. Thus, we conclude that at least 1 of the model coefficients is different from zero.

(SAS output continued on next page.)

Table 4.2

Logistic Regression Model with All Predictors
Color and Spine Condition 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	-9.2734	3.8378	-16.7954	-1.7514	5.84	0.0157
color	med light	1	1.6087	0.9355	-0.2250	3.4423	2.96	0.0855
color	medium	1	1.5058	0.5667	0.3951	2.6164	7.06	0.0079
color	med dark	1	1.1198	0.5933	-0.0430	2.2826	3.56	0.0591
color	dark	0	0.0000	0.0000	0.0000	0.0000	.	.
spine	both good	1	-0.4003	0.5027	-1.3856	0.5850	0.63	0.4259
spine	1 worn or broken	1	-0.4963	0.6292	-1.7294	0.7369	0.62	0.4302
spine	both worn or broken	0	0.0000	0.0000	0.0000	0.0000	.	
width		1	0.2631	0.1953	-0.1197	0.6459	1.82	0.1779
weight		1	0.8258	0.7038	-0.5537	2.2053	1.38	0.2407

Upon examination of the significance test results given above, we see that only “medium color” (yes/no) is statistically significant according to the Wald test. Results such as these are indicative of a multicollinearity problem [i.e., highly significant overall likelihood ratio test ($p < .0001$), but few, if any, significant coefficients]. The significance test result for carapace width is very surprising, since, in Section 5.3, we found it to be a highly significant predictor of having at least 1 satellite ($p < .0001$). As indicated above, after we control for weight, color, and spine condition, carapace width is no longer significant ($p = .178$).

When performing multiple (logistic) regression, it is helpful to examine the correlation matrix for the predictors. The following SAS code can be used to accomplish this:

```
proc corr;
     var y color spine width weight;
title 'Table 4.2';
title2 'Correlation Matrix';
run;
```

The relevant SAS output is given on the following page:

Table 4.2
Correlation Matrix

	y	color	spine	width	weight
y	1.00000	-0.26778	-0.02777	0.40141	0.38719
color	-0.26778	1.00000	0.37850	-0.26439	-0.25078
spine	-0.02777	0.37850	1.00000	-0.12189	-0.16648
width	0.40141	-0.26439	-0.12189	1.00000	0.88687
weight	0.38719	-0.25078	-0.16648	0.88687	1.00000

(Note: Scatterplots between all possible pairs of predictors should also be constructed to help detect possible collinearities.)

The high correlation between carapace width and weight of the crab (r = .89) is the most likely explanation for the unexpected result for the test of the coefficient for width. Upon further reflection, we see that it makes little sense to consider a LR model that examines the effect of carapace width while controlling for crab weight since weight naturally increases along with width.

Agresti chooses to eliminate weight from further consideration as a predictor. This is reasonable, given its high correlation with width and its slightly lower correlation with the binary outcome Y (.387 vs. .401). He proceeds to consider possible LR models that include main effects for width (W), color (C), and spine condition (S). (C and S are treated as nominal variables in these analyses.) He denotes the various models he considers by using the highest order term in the model; for example, he uses C + S * W to denote a model having an interaction between spine condition and width, but no interactions involving color. The models he considers are listed in Table 5.8:

Table 5.8 Results of Fitting Several Logistic Regression Models to Horseshoe Crab Data

Model	Predictors	Deviance	*DF*	Models Compared	Difference	$(Y, \hat{\pi})$ Correlation
(1)	$C * S * W$	170.44	152	—	—	0.526
(2)	$C * S + C * W + S * W$	173.68	155	(2)-(1)	3.2 ($df = 3$)	
(3a)	$C * S + S * W$	177.34	158	(3a)-(2)	3.7 ($df = 3$)	
(3b)	$C * W + S * W$	181.56	161	(3b)-(2)	7.9 ($df = 6$)	
(3c)	$C * S + C * W$	173.69	157	(3c)-(2)	0.0 ($df = 2$)	
(4a)	$S + C * W$	181.64	163	(4a)-(3c)	8.0 ($df = 6$)	
(4b)	$W + C * S$	177.61	160	(4b)-(3c)	3.9 ($df = 3$)	
(5)	$C + S + W$	186.61	166	(5)-(4b)	9.0 ($df = 6$)	
(6a)	$C + S$	208.83	167	(6a)-(5)	22.2 ($df = 1$)	
(6b)	$S + W$	194.42	169	(6b)-(5)	7.8 ($df = 3$)	
(6c)	$C + W$	187.46	168	(6c)-(5)	0.8 ($df = 2$)	0.452
(7a)	C	212.06	169	(7a)-(6c)	24.5 ($df = 1$)	0.285
(7b)	W	194.45	171	(7b)-(6c)	7.0 ($df = 3$)	0.402
(8)	C=dark + W	187.96	170	(8)-(6c)	0.5 ($df = 2$)	0.447
(9)	None	225.76	172	(9)-(8)	37.8 ($df = 2$)	0.000

Note: C = color, S = spine condition, W = width

Backward Elimination of Predictors (Sec. 5.5.5)

Table 5.8 also summarizes the results of fitting and comparing the models considered by Agresti. As we have discussed before (Sec. 5.3.2), the *deviance* of a model is the likelihood-ratio test statistic G^2 for comparing the model being examined to the *saturated* model that has a separate parameter for each study subject. The difference in deviances for any 2 models is the likelihood-ratio test statistic $-2(L_1 - L_0)$ for comparing the 2 models. Agresti recommends using a *backward elimination* procedure, which starts with the model containing all possible terms that could be included and then successively removes terms. (His approach differs from the backward elimination procedure implemented in SAS.)

The most complex model based on the 3 predictors width, color, and spine condition contains the 3 main effects (C, S, W), all possible 2-way interactions among these variables (C*S, C*W, S*W), and the 3-way interaction (C*S*W). At each stage of the backward elimination procedure, the term that has the largest p-value when we test the significance of its coefficient(s) using the likelihood-ratio test is eliminated. *Any term that has a significant p-value should be retained.* At each stage, we test only the highest-order term for each variable since it is inappropriate to remove a main effect term if an interaction involving that main effect has been retained in the model. (For example, you would not remove color from the model if the interaction of color & width had been found to be statistically significant even if the test for the color parameter was not significant.)

From the results in Table 5.8, we see that if we remove the 3-factor interaction C*S*W, the likelihood-ratio test statistic for comparing Model (2) vs. Model (1) is $-2(L_1 - L_0) = 3.24$, df = 3, p = .356 by SimCalc. This indicates that the 3-factor interaction term is not needed in the model. ("Thank goodness," as Agresti says, since no one really understands how to interpret 3-way and higher order interactions.) At the next stage, we consider removing one of the 2-way interactions from Model (2). Comparing the 3 models with one 2-way interaction removed (Models 3a, 3b, 3c), we see that the deviance for Models (2) and (3c) are almost identical (173.68 vs. 173.69). Therefore, we can safely remove the S*W interaction. Now, what about dropping one of the 2 remaining 2-way interactions? To do this, we compare Models (4a) & (4b) with (3c). Using the results in the "Difference" column in Table 5.8, along with SimCalc, we see that the p-value for retaining the C*S interaction (4a vs. 3c) is .242 and for the C*W interaction (4b vs. 3c) it is .270. We drop the term that has the largest p-value (C*W), and that leaves us with model (4b). Comparing this model with the one in which the C*S interaction has been dropped (Model 5), we have a p-value of .174, so we can safely drop the C*S interaction. We have now eliminated the 3-way interaction and all 2-way interactions. Note that we could have examined the effect of removing *all* 2-way interactions at once by comparing Models (2) and (5). The results here are $-2(L_1 - L_0) = 186.61-173.68 = 12.93$, df = 166-155 = 11, p = .298. Therefore, we are justified in dropping all of the 2-way interactions at once.

Our working model is now Model (5), which contains only the main effect terms. The next stage of the backward elimination procedure involves possibly dropping one of them. Using SimCalc to calculate the p-values for the "Difference" values in Table 5.8, we see that removing S has almost no effect (Model 5 vs. 6c, p = .670). However, we should *not* remove W (Model 5 vs. 6a, p < .0001). Agresti recommends that we also keep C (Model 5 vs. 6b, p = .0503). After removing S, we are now at Model (6c). Should we remove either W or C? The answer is definitely "no" for W (Model 6c vs. 7a, p < .0001). Again, Agresti recommends keeping C even though the test does not quite reach statistical significance (Model 6c vs. 7b, p = .072). (Note that this is the same result that we presented on p. 103 of

these lecture notes.) As we saw earlier, Agresti also considers the effect of dichotomizing color into "dark" and "lighter." If we call this variable "D," then we see that we do have statistical significance when we compare Model W (7b) with Model D + W (8): $-2(L_1 - L_0) = 6.49$, df = 1, p = .011. Therefore, we should retain D. Removing W is certainly not justified [$-2(L_1 - L_0) = 25.10$, df = 1, $p < .0001$), so we stop the variable elimination process.

We should now use the methods of Section 5.3 to check the overall fit of the model D + W.

As Agresti points out, use of "automated" variable selection procedures is to be discouraged. Simulation studies have shown that explanatory variables can be retained by such procedures even if they are completely independent of the outcome. This is less likely to occur using the "deviance comparison" process we have outlined here. Agresti's approach is one way of looking at "all possible regressions," which is a method of variable selection recommended by many authors for use in multiple regression analysis if the number of explanatory variables is not too large.

Generally, we want all of the explanatory variables that we retain in our LR model to be statistically significant. However, in some Epi studies, we may wish to demonstrate that a potential risk factor for a certain disease is no longer significant after adjusting for one or more other risk factors that may be more important. For example, suppose that alcohol consumption is being considered as a possible risk factor for a certain type of cancer. When included as the lone predictor in a LR model, alcohol is significant. However, after smoking is added to the model, alcohol is no longer significant, but smoking is. (This type of confounding is likely to occur since heavy drinkers are more likely to smoke than non-drinkers.) In terms of the "final" LR model that we would include in a journal submission or other report, we would "force" alcohol into the LR model and retain it as a predictor even though it is not significant so that we can demonstrate the significant confounding effect of smoking on the association between alcohol and the particular cancer we are considering.

A Correlation Summary of Predictive Power (Sec. 5.5.6)

Note that there is another column in Table 5.8 with the heading "(Y, $\hat{\pi}$) Correlation." This gives the Pearson correlation coefficient R_{LR} between the observed binary responses Y and the fitted (or predicted) values $\hat{\pi}$ obtained from the model. R_{LR} is a crude measure of predictive power, and it does not have some of the desirable properties that a similar measure , R^2, has for multiple regression models. For example, R^2 always increases as you add terms to the model; R_{LR} does not. In addition, the values of R_{LR} tend to be small, whereas for good fitting multiple regression models, you hope to see $R^2 \geq .75$. From Table 5.8, we see that the R_{LR} value for the complex model C*S*W is .526; this model contains all possible terms involving 3 predictors. The next largest R_{LR} value (.452) is for the model C + W, but this model has a non-significant coefficient for C. The R_{LR} value for the model D + W, which used a dummy variable for "dark" color rather than the original color variable C has an R_{LR} value of .447. This represents a drop of only .005 in the R_{LR} value in going from the C + W model to D + W. In addition, if we divide the R_{LR} value for D + W by the R_{LR} value for the complex model C*S*W, we obtain $\frac{.447}{.526} = .85$, which is more like the R^2 values we see in multiple regression. Since both coefficients in D + W are statistically significant, my choice would be to go with this model.