CPH 636 — Spring 2009 — Dr. Charnigo

Lecture 6

The binary logistic regression model

Scenario. We have a single dichotomous (or "binary") response variable Y and multiple explanatory variables $X_1, \ldots, X_k$. The explanatory variables can be any mix of continuous and dichotomous. [[This is not restrictive since a categorical variable with m levels, $m > 2$, can be reduced to a collection of $(m-1)$ dichotomous indicators.]] Ordinarily we assign values of "1" and "0" to Y that reflect the unfavorable event (e.g., developing a disease) and its absence (e.g., not developing a disease), respectively.

Model formulation. Let

$$p_{x_1,\ldots,x_k} := P(Y = 1 | X_1 = x_1, \ldots, X_k = x_k),$$

the probability (or "risk") of the unfavorable event when the explanatory variables have assumed the numerical values $x_1, \ldots, x_k$. For any positive number t, let us define the "logit function" by

$$\text{logit}(t) := \log\left(\frac{t}{1-t}\right).$$

Then

$$\text{logit}(p_{x_1,\ldots,x_k}) = \log\left(\frac{p_{x_1,\ldots,x_k}}{1-p_{x_1,\ldots,x_k}}\right)$$

equals the log odds of the unfavorable event when the explanatory variables have assumed the numerical values $x_1, \ldots, x_k$. While $p_{x_1,\ldots,x_k}$ must lie between 0 and 1, there is no such requirement for $\text{logit}(p_{x_1,\ldots,x_k})$.

The binary logistic regression model

$$\text{logit}(p_{x_1,\ldots,x_k}) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

expresses the log odds of the unfavorable event as a linear combination of the values assumed by the explanatory variables. An alternative formulation of the model is

$$p_{x_1,\ldots,x_k} = \frac{\exp[\alpha + \beta_1 x_1 + \cdots + \beta_k x_k]}{1 + \exp[\alpha + \beta_1 x_1 + \cdots + \beta_k x_k]}.$$

[[We do <u>not</u> write $Y_i = p_{x_{1,i},\ldots,x_{k,i}} + \epsilon_i$ for logistic regression since the ϵ_i would not be readily interpretable. In particular, there would be only two possible values for each ϵ_i, according to whether $Y_i = 1$ or $Y_i = 0$. This contrasts with the ϵ_i in linear regression, which can be interpreted as normally distributed deviations of actual responses from expected responses.]]

Interpreting the coefficients. If having $X_1 = \cdots = X_k = 0$ is possible, then the intercept α represents the log odds of the unfavorable event for an individual with $X_1 = \cdots = X_k = 0$. Alternatively, the probability of the unfavorable event for such an individual is $\exp[\alpha]/(1 + \exp[\alpha])$.

The partial slope coefficient β_1 may be interpreted as follows. Consider two individuals such that the first individual is c units greater on X_1 but the two individuals are the same on $X_2, \ldots, X_k$. Then the odds ratio

$$\frac{\text{Odds of unfavorable event for first individual}}{\text{Odds of unfavorable event for second individual}}$$

equals $\exp[c\beta_1]$. Here I am assuming that there are no mathematical relationships among $X_1, \ldots, X_k$. For instance, I am ruling out relationships like $X_k = X_1^2$ or $X_k = X_1 \times X_2$.

Discussion questions. Suppose that $\beta_1 = -0.05$. If individual A has $X_1 = 10$ and individual B has $X_1 = 0$ but is otherwise similar to individual A, what is the odds ratio? Is the unfavorable event more likely for someone with a larger value of X_1 or a smaller value?

Estimation and prediction. Coefficients in a logistic regression model are estimated by a statistical technique called "maximum likelihood". Roughly speaking, the idea is to produce estimates of the coefficients for which the observed data were most likely to occur. Maximum likelihood estimation in logistic regression must be done by computer; there are no formulas available for pencil and paper.

[[Maximum likelihood estimation in linear regression is equivalent to ordinary least squares. The latter rather than the former is presented to students when they first encounter linear regression since: (i) ordinary least squares has an intuitive and appealing geometric rationale; and, (ii) maximum likelihood is difficult to describe without the notion of a probability density function, which is glossed over in a typical introductory methods course. Maximum likelihood is a general statistical technique and appears in many settings: linear regression and logistic regression are but two.]]

Let $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k$ denote the estimates of $\alpha, \beta_1, \ldots, \beta_k$. With generic values of the explanatory variables, we refer to

$$\mathrm{logit}[\hat{p}_{x_1,\ldots,x_k}] = \hat{\alpha} + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

or

$$\hat{p}_{x_1,\ldots,x_k} = \frac{\exp[\hat{\alpha} + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k]}{1 + \exp[\hat{\alpha} + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k]}$$

as a "fitted model". With values of the explanatory variables corresponding to subject i specifically, we refer to

$$\hat{p}_{x_{1,i},\ldots,x_{k,i}} = \frac{\exp[\hat{\alpha} + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_k x_{k,i}]}{1 + \exp[\hat{\alpha} + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_k x_{k,i}]}$$

as a "fitted probability".

Note that a fitted probability of 0.30, for example, does not mean that someone with the same values of the explanatory variables as subject i would be projected to have $Y = 0.30$. Rather, such a person has an estimated 30% chance of having $Y = 1$ and an estimated 70% chance of having $Y = 0$.

If we wish, we can make "predictions" by saying that anyone with a fitted probability of at least 0.50 would be projected to have $Y = 1$ while anyone with a fitted probability less than 0.50 would be projected to have $Y = 0$. However, we are not required to choose 0.50 as the threshold. In particular, if the unfavorable event is rare yet very serious, then a threshold lower than 0.50 may be appropriate.

Quantifying predictive capabilities. Several measures have been proposed to quantify the predictive capabilities of a logistic regression model. I describe a few of them here.

• Sensitivity: This is the fraction of individuals experiencing events for whom the fitted probabilities are above the threshold for projecting that $Y = 1$. We can compute sensitivity for the training subset, validation subset, or test subset. If computing sensitivity for the validation subset or test subset, we still use the coefficient estimates from the training subset to obtain the fitted probabilities.

• Specificity: This is the fraction of individuals not experiencing events for whom the fitted probabilities are below the threshold for projecting that $Y = 1$.

• False positive rate: Among the individuals with fitted probabilities above the threshold for projecting that $Y = 1$, this is the fraction who did <u>not</u> experience events.

• False negative rate: Among the individuals with fitted probabilities below the threshold for projecting that $Y = 1$, this is the fraction who experienced events.

• Correct classification rate: This is the fraction of individuals for whom the actual experiences (i.e., event or non-event) were the same as the projected experiences. The correct classification rate is a number between 0 and 1 that constitutes an overall summary of the model's predictive capabilities. The cutoff for a "worthless" model is not 0 but rather 0.5. [[**Discussion question**: Why?]]

• Area under the Receiver Operator Curve (or "C"): If we vary the threshold, then sensitivity and specificity will change. A plot of sensitivity against 1 minus specificity is called a receiver operator curve. The area under the receiver operator curve is a number between 0 and 1 that constitutes an overall summary of the model's predictive capabilities. Again, the cutoff for a "worthless" model is not 0 but rather 0.5.

Illustrating binary logistic regression

Introduction. We have already encountered the South African heart disease data set {saheart.sas7bdat} in Lecture 3. Now I will describe some of the output that I obtained when I applied the LOGISTIC macro to a training subset. The output is available in {RCLIB.SATRAIN224.rtf}.

Page 13. We see ("Model Information") that the response variable was CHD, that the training subset had 231 observations, and that ("Response Profile") 86 out of the 231 subjects in the training subset had $CHD = 1$.

We also see ("Descriptive Statistics for Continuous Variables") what the explanatory variables were and what their means and standard deviations were, both overall and within each of the two strata determined by the response variable.

Page 14. Results for three tests of $H_0 : \beta_1 = \cdots = \beta_k = 0$ are presented in the "Testing Global Null Hypothesis" box. With very large samples, the three tests are almost equivalent. With small to medium samples, there may be noticeable differences in the results; some statisticians favor the first test ("Likelihood Ratio") over the other two.

Page 15. The "Analysis of Maximum Likelihood Estimates" box displays estimates of $\alpha, \beta_1, \ldots, \beta_k$ ("Estimate"), standard errors ("Standard Error"), and the results for "Wald" tests of $H_0 : \alpha = 0$ against $H_1 : \alpha \neq 0$, $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, and so forth ("Wald Chi-Square" and "Pr > ChiSq"). [[A Wald test divides the parameter estimate by the standard error, squares the quotient, and then compares the result to an upper quantile of the chi-square distribution on one degree of freedom.]]

Based on the coefficient estimates, the estimated log odds of developing coronary heart disease is

$$-6.9699 + 0.0169SBP + 0.1255TO + 0.1399LDL + 0.0479AD +$$

$$0.9254FAM + 0.0383TA - 0.0584OB - 0.00531AL + 0.0163AGE.$$

[[For a retrospective study, the estimated intercept may overstate α because, by design, there are disproportionately many "cases" (i.e., subjects who experienced the unfavorable event) in the data set. However, the estimated partial slope coefficients do not have any analogous problem. Thus, although the fitted probabilities may be systematically too high, their ordering is meaningful. We may reasonably regard a person with a fitted probability of 0.60 as being at greater risk than a person with a fitted probability of 0.40, even if the 0.60 and 0.40 themselves are too high.]]

The "Odds Ratio Estimates" box displays estimates and 95% confidence intervals for the odds ratios $\exp[\beta_1], \ldots, \exp[\beta_9]$. The confidence intervals are computed by the "Wald" method. [[The lower confidence limit is the exponential of {the estimate minus 1.96 times the standard error}, while the upper confidence limit is the exponential of {the estimate plus 1.96 times the standard error}.]] For example, a person with a family history of heart disease is estimated to have 2.523 times the odds of developing coronary heart disease compared to an otherwise similar person without a family history; however, the data are not inconsistent with the odds ratio being as low as 1.337 or as high as 4.762.

Several measures of the model's predictive capabilities are presented in the "Association of Predicted Probabilities and Observed Responses" box. I emphasize only one, the area under the receiver operator curve ("C").

Page 16. The "Profile Likelihood Confidence Interval for Adjusted Odds Ratios" box presents alternative 95% confidence intervals computed by the "profile likelihood" method. With very large samples, these confidence intervals are almost the same as those produced by the Wald method. With small to medium samples, there may be noticeable differences; some statisticians favor the profile likelihoood method in such instances.

The next two boxes pertain to the Hosmer-Lemeshow Goodness of Fit Test. I do not use this test, but some statisticians are strong advocates of it. Roughly speaking, a p-value <u>greater</u> than 0.05 suggests that the model provides a "good fit" to the data.

Page 17. Suppose that the threshold for projecting $Y = 1$ is set to 0.50. We can find the corresponding sensitivity, specificity, false positive rate, and false negative rate (in the training subset) by looking in row 0.50 of the "Classification Table".

There were 46 subjects who experienced events and had fitted probabilities greater than 0.50, 120 subjects who did not experience events and had fitted probabilities less than 0.50, 25 subjects who had fitted probabilities greater than 0.50 but did not experience events, and 40 subjects who had fitted probabilities less than 0.50 but did experience events.

The correct classification rate is $(46+120)/(46+120+25+40) = 71.9\%$, the sensitivity is $46/(46+40) = 53.5\%$, the specificity is $120/(120+25) = 82.8\%$, the false positive rate is $25/(25+46) = 35.2\%$, and the false negative rate is $40/(40+120) = 25.0\%$.

Pages 18 and 19. Here we have the 10% of subjects in the training subset whose fitted probabilities were lowest and the 10% whose fitted probabilities were highest.

Pages 24 and 25. On page 24 are listed observations that have been flagged as "outliers" or "influential". The outliers are the observations for which the projected experience based on the logistic regression model (i.e., event or non-event) was near definitive yet different from the actual experience. In other words, the outliers are observations that do not "fit the pattern" established by the bulk of the data, with respect to the associations of the response variable with the explanatory variables. The requirement to be flagged as an outlier is a squared "deviance residual" exceeding 4, although

4 is a pretty low cutoff. [[Actually, the requirement is a squared deviance residual exceeding {4 minus the "confidence interval displacement"}, but the confidence interval displacement is usually small.]]

The influential observations are those for which the configuration of explanatory variable values was atypical, irrespective of the response variable value. These are the observations most capable of strongly influencing the estimation of $\alpha, \beta_1, \ldots, \beta_k$. The requirement to be flagged as an influential observation is a hat matrix element more than twice the average hat matrix element. [[The hat matrix, already defined in Lecture 5, does not play the same role in logistic regression that it does in linear regression. In particular, it does not convert a vector of actual responses into a vector of fitted probabilities. Even so, the hat matrix can be used for diagnostic purposes in logistic regression.]]

On page 25 is a graphical display that identifies the outliers and influential observations. We would be most concerned about an observation in the upper right quadrant because such an observation, being both an outlier and influential, might have pushed the estimates of $\alpha, \beta_1, \ldots, \beta_k$ away from values that would have better described the pattern established by the bulk of the data, with respect to the associations of the response variable with the explanatory variables.

Pages 26 and 27. On page 26 is a graphical display of the receiver operator curve. On page 27 is a graphical display of the correct classification rate, false positive rate, and false negative rate as functions of the threshold for projecting that $Y = 1$.

Model selection from the training subset

Pages 6, 7, and 8. On page 6 of {RCLIB.SATRAIN224.rtf} is a list of AIC ("Akaike Information Criterion") and SC ("Schwarz Criterion") values for nine candidate logistic regression models. The nine candidate models correspond to adding one explanatory variable at a time in an order determined by a forward selection algorithm. We are interested in models with small values of the AIC and SC. Note that the detailed results on pages 13 to 27 of {RCLIB.SATRAIN224.rtf} pertain only to the ninth candidate logistic regression model, which has nine explanatory variables. [[The AIC and SC are as defined in Lecture 5; the SC and SBC are the same.]]

Pages 7 and 8 provide graphical displays of the AIC and SC values, with long green bars identifying models that are judged unfavorably, short green bars identifying models that are judged only slightly less favorably than the preferred model, and no green bar identifying the preferred model.

Model selection and evaluation from the validation and test subsets

Model selection. If we have a validation subset, we do not have to make a final decision about which model to choose based on the AIC and SC values computed from the training subset. Much as we computed R^2_{valid} for each candidate model in linear regression, we can compute the correct classification rate on the validation subset for each candidate model in logistic regression. I have provided SAS code for this task in {SASMacroInstr5.txt}.

Our final decision about which model to choose is that we will select the model producing the highest correct classification rate on the validation subset. [[A limitation of this approach is that, if the event is rare, one may be able to surpass the correct classification rate of even a very good model by blindly "predicting" that nobody will experience the event.

Hence, another option is to look at the area under the receiver operator curve on the validation subset rather than at the correct classification rate. This is more work but can be carried out with the code in {http://web.as.uky.edu/statistics/users/rjchar2/CPH930F06/SASroc.txt} along with that in {SASMacroInstr5.txt}. **Discussion question**: Why not just "predict" that nobody will experience the event?]]

Of course, the correct classification rate depends on the threshold for projecting $Y = 1$, so we either need to fix the threshold for all candidate models (perhaps at 0.50) or agree to use the "best" threshold for each candidate model, in the sense of maximizing the correct classification rate on the validation subset for each candidate model.

Model evaluation. In the end, we formally evaluate whichever model we have selected by computing the correct classification rate on the test subset for that model.

Diagnostic plots

Logit and partial delta logit plot. This is a logistic regression analogue to the augmented partial residual plot in linear regression. If interested, you may refer to page 168 of the textbook for details; see pages 2 through 10 of {RCLIB.SATRAIN222.rtf} for illustrations. Two key ideas are as follows.

• The configuration of L points represents the unadjusted linear relationship between the log odds of the event and the explanatory variable. The configuration of P points and the parabola through them represent the adjusted quadratic relationship between the log odds of the event and the explanatory variable. Thus, a pronounced curvature in the configuration of P points suggests that inclusion of a quadratic term (e.g., a term of the form X_1^2) may be warranted, while a large discrepancy between the configurations of L points and P points suggests that some of the other explanatory variables may be confounding the relationship between the log odds of the event and the present explanatory variable.

• For a continuous explanatory variable, a severe compression of P points is indicative of collinearity. The variance inflation factor, calculated in the same manner as for linear regression, is also reported at the top of the plot.

Interaction plots. These plots, illustrated on pages 11 through 17 of {RCLIB.SATRAIN222.rtf}, are like the analogous plots in linear regression. However, for some reason that is not clear to me, the present output has considerably fewer plots (seven) than the number of possible interactions when there are nine explanatory variables (thirty-six).