Chapters 8 Notes

Logistic Regression

In our discussions of linear regression, we considered models of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \varepsilon_i$$

where the ε_i were iid mean 0 normally distributed residuals. If $\hat{y}_i$ is the expected value of Y_i, then we could equivalently write

$$Y_i \sim N(\hat{y}_i, \hat{\sigma}^2)$$

so each of the responses is assumed to be from a normal distribution with mean equal to the expected value and a common variance estimated by $\hat{\sigma}^2$. In the linear regression model, the response is continuous and can potentially take on any real value.

In many situations the assumptions of the linear regression model will not be appropriate. One such case is in modeling a dichotomous response (a response that can only take on two possible values). In such cases we would instead prefer to model the probability of observing one of the possible outcomes. We can code our dichotomous response variable as 0's and 1's and model the probability π of being in case 1 (or case 0 if we prefer) given the predictors. Our binary response y_i is then distributed according to a Bernoulli distribution

$$Y_i \sim Ber(\hat{\pi}_i)$$

and the expected value of Y_i is $\hat{y}_i = \hat{\pi}_i$.

The **odds** of an event is the probability that an event happens divided by the probability that event doesn't happen. In the logistic regression model, we model the **log odds** as a linear function

$$\log(\hat{\pi}_i/(1-\hat{\pi}_i)) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots$$

The expected odds would then be given by

$$\hat{\pi}_i/(1-\hat{\pi}_i) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots)$$

and the estimated probability is

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots)}$$

where exp is the exponential function.

We should note several things:

- we no longer have an assumption of normality
- we can no longer separate the error term from the rest of model
- estimates for the $\hat{\beta}_j$'s will come from maximum likelihood estimation assuming Bernoulli (or binomial) distributed responses
- the $\hat{\pi}_i$ values will fall between 0 and 1
- holding all other predictor variables constant, a change of 1 in the j^{th} predictor will result in a change of $\exp(\hat{\beta}_j)$ in the odds
- the odds ratio between two sets of conditions (values for the predictors variables) will tell us how much greater (or less) the odds of an event is under one set of conditions than under another (e.g. how many times greater is the odds for females to encounter a specific event than for males, or how much greater is the odds for an event to happen if a continuous predictor increases by 1)
- confidence intervals for the $\hat{\beta}_j$'s will be symmetric, but confidence intervals for the odds and odds ratios will not be symmetric because they are based on exponentials of the symmetric intervals

Note that under the assumption of normality, least squares estimates are equivalent to maximum likelihood estimates for the parameters since the likelihood will be maximized (or equivalently the log likelihood will be maximized) when the sum of squares is minimized. This will no longer be the case for logistic regression and results and diagnostics will need to be based on likelihoods rather than sums of squares.

If we have counts of events out of a possible number of trial (e.g. if we count the number of heads out of *n* coin tosses), our counts would follow binomial distributions instead of Bernoulli (recall that Bernoulli is a binomial with n=1), and we would now have

$$Y_i \sim Bin(n_i, \hat{\pi}_i)$$

and the expected value of Y_i is $\hat{y}_i = n_i\hat{\pi}_i$.

The logistic procedure in SAS

The procedure specific to logistic regression is **proc logistic**. For our usage, the syntax will be largely similar to that for **proc reg**. There are a couple differences we will note in the input specification. In the **model** statement specification, we can now have event counts and trial counts in the specification of the response

model events/trials = *<insert model terms as before>*

In the events/trials specification we are looking at **events** being observations from a binomial with n=**trials**, and we are still interested in the estimates of the probability p=$\hat{\pi}$.

We can also have both continuous explanatory variables (like in **proc reg**) and categorical predictors (specified with a **class** statement like in **proc anova** and **proc glm**). This will be very useful because often when predicting probabilities of outcomes, the probabilities depend on categorical factors.

For instance with the General Health Questionnaire (GHQ) data set in Chapter 8 of the text , there is a **sex** variable. It is not unreasonable that some illnesses would be more likely to affect men than women or vice versa. It may also be the case that different age groups might have different impacts on a binary response (for instance, if we wanted to predict probability of being in favor of a particular political issue, the age group from 25 to 35 may be different from the 75+ age group).

We've previously mentioned that categorical predictors can be incorporated in a linear model. In **proc reg** we would have to code these ourselves and treat the coded variables as continuous predictors. In **proc glm** we could specify them and let the procedure worry about the coding (this is what was happening in the unbalanced ANOVA examples via **proc glm**). We can have a look at the **"Regression Models and Models with Classification Effects"** section of the **"Introduction to Statistical Modeling with SAS/STAT Software"** portion of the SAS documentation to get see how these terms get incorporated.

Coding of categorical variables

Binary predictor variables are typically coded as:

- 0/1 (this is **reference cell coding**) or
- -1/1 (**deviations from means coding**)

For categorical predictor variables with *n* possible values, we would typically create *n-1* reference-coded variables.

Initial Example

Looking at the **ghq** example we can see the difference between modeling the case probability as a function of questionnaire score using a linear model and a logistic model. The tendency to 0 and 1 in the logistic model as the test score gets small and as the test score gets large is clear, as is the fact that the linear model will overshoot 0 and 1 when test scores get too large or too small.

Now let's fit the model where case probability (given as cases divided by total of cases and non-cases) is predicted by both **sex** and **ghq** score. Our results and interpretation will be similar in concept to those for linear regression. A few notable differences are:

- the goodness of fit measures are computed from log-likelihood (actually the R^2 values are based on log-likelihoods, too…) and will follow approximate chi-square distributions; we can see more information about these measures under the **Details>Model Fitting Information** tab of the **proc logistic** docs

- we are interested in the odds and the odds ratio being significantly different from 1 (the odds is 1 when $\pi = 1 - \pi$), so confidence intervals for odds or odds ratios which contain 1 will be insignificant and those not containing 1 will be significant (the β values will still be significant when they are significantly different from 0)
- elimination criteria for selection methods will also be based on likelihoods, and hence be chi-squared rather than F distributed
- residuals and results based on residuals will need to be modified or generalized to maintain a useful interpretation

After modeling probability of being a case based just on test score, we may want to consider the impact of gender on the probability of being a cases as well.

- Plot probability of being a case vs. **ghq** score with the data grouped by **sex**. Does it seem reasonable that there may be a difference between males and females?
- Add the **sex** variable as a categorical predictor in the model and interpret the results? Is there a difference between male and female? Is the test score still significant? Etc.

Exercises

ESR and Plasma data set

This data contains **fibrinogen** and **gamma** globulin plasma levels and an **esr** variable. **esr** is an indicator variable for whether an individual has an unhealthy erythrocyte sedimentation rate(ESR). The variable is coded as 0 for healthy and 1 for unhealthy in this data set.

For 0|1 responses, SAS will treat the 0 case as the case of interest in **proc logistic** by default, so it will treat **esr**=0 as the event. To model unhealthy esr, we will want to use the **desc** option to reverse the order of response values.

- Use a logistic regression model with the main effects **fibrinogen** and **gamma** as the explanatory variables (no interaction in the model). What do the global tests, beta estimates and odds ratio test us about the significance of the impact of **fibrinogen** and **gamma** on **esr**?
- If either or both of the main effects is insignificant, remove the insignificant ones, refit and comment on the significance of the results again.
- We could also include the interaction between **fibrinogen** and **gamma** as a product term. Fit the logistic regression model with both main effects and the interaction. Comment on the significance of the main and interaction effects.
- If any terms are insignificant, use backward selection (starting with both main effects and the interaction) to obtain a best model and comment on how that model compares with the model obtained when just picking the significant main effects.

Danish Do-It-Yourself data set

An interesting feature of this data set is that it contains classification variables with more than 2 levels, so the classification variables will need to be coded as multiple 0|1 variables. We will want to use the

param option to specify that we want to use reference coding and the **ref** option to tell **proc logistic** that we want the **first** value to be the reference value.

The explanatory variables are **work**, **tenure**, **type** of home, and **agegrp**, and the response of interest is the probability of responding yes to a question about do-it-yourself home improvement.

- Look at the actual data using **proc print**. Look at the cross tabulation of the response probability given in the text (and in the initial code file for the chapter). Are there any obvious trends in the cross-tabulation?
- Fit the model based on "yes"-response probability based only on **work** type. One of the three **work** categories will be rolled into the intercept term. That category will give a baseline for comparison, so estimates for the other two **work** categories will be comparing to that baseline category. Comment on any significant differences.
- Now use all four main effects and backward selection to obtain the best (based on the backward selection criteria…) model and comment on the impacts of the remaining predictors on "yes"-response probability. Interpret the results.

Residuals and Pointwise Diagnostics

As mentioned above, the concepts of residuals and other diagnostics (such as leverage-based diagnostics) from linear regression must be modified for logistic (and other generalized linear models). We will talk about these a bit more when we discuss generalized linear models, but will point out the ideas now.

In linear regression, residuals have many useful properties and can be interpreted in terms of distance (sums of squares), log-likelihoods, and deviances. In the case of independent identically distributed normal responses, these all boil down to sums of squares and raw residuals (observed-expected values).

For logistic and generalized linear models, **Pearson residuals** extend the sum of squares interpretation, **deviance residuals** extend the deviance interpretation, and **likelihood residuals** extend the likelihood interpretation. Each takes into account the distributional assumptions in some way. In particular, the Pearson residuals are the individual terms in the Pearson chi-square statistic (like in categorical data analysis before) and deviance residuals are the individual contributions from the deviance (a likelihood ratio-based measure).

Likewise, leverage diagnostics can be obtained by using the formulas as before, but incorporating a one-step linear approximation adjustment based on the distributional assumptions. Single step linear approximations are used because the true values would require an iterative estimation for each point, which would be very time intensive, and the linear approximations should be reasonably close for diagnostic purposes.

We can find some additional information under the under the **Details>Regression Diagnostics** tab of the **proc logistic** docs.