Chapter 9 Notes

Generalized Linear Models

In our discussions of linear and logistic regression, we built models around a **linear predictor** of the form

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

We will denote this linear predictor by η, so we can write

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

In the linear regression case, we predict a **continuous** response y by this linear predictor, so we have

$$\hat{y}_i = \mu_i = \eta_i$$

where μ_i is the expected value of Y_i given the x values from the i^{th} data point.

In the logistic case, we model the **probability** of a binary response either from data with a binary response (the Bernoulli case) or from counts of events out of total observations (the binomial case). Here, we model the expected value of a single trial (the probability of an event happening in a single trial), so we have that the expected probability is

$$\hat{\pi}_i = \mu_i$$

and

$$\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \eta_i$$

so the probability of an event given the explanatory variables from the i^{th} data point can be predicted by

$$\mu_i = \hat{\pi}_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

because the logit function is invertible for $0 \leq \mu \leq 1$ and can be inverted as showed above.

The changes from the linear regression case are in the assumed distribution of the response (binary outcome for Bernoulli case, and proportion of "events" observed from a binomial with a known number of trials) and the thing we equate to the linear predictor.

Here we have a **linearized** model for proportions. The proportions themselves are not modeled by a linear combination of explanatory variables, but a function of the response is. Further, that function of the linear combination is invertible for proportions, so we can predict proportions by a function of the linear predictor.

This brings us to the generalized linear model generalization of linear regression, which will allow us to model other types of responses via linearized models as well. To define this class of models, we will introduce the following terms:

- **response distribution**: the distribution family the responses are assumed to come from (e.g. normal for linear regression).
- **linear predictor**: this is the linear combination we denoted by η.
- **link function**: the function we apply to the response to get a linear combination. We will denote the link function by $\mathbf{g}$. Then we have $\mathbf{g}(\mu) = \boldsymbol{\eta}$, and we can get predictions from $\mathbf{g}^{-1}(\boldsymbol{\eta})$.
- **dispersion parameter**: this is like the error variance in linear regression (and in fact is the error variance in linear regression), and is denoted by $\boldsymbol{\varphi}$.
- **variance function**: the function **V** such that we can write the distribution’s variance as $\boldsymbol{\varphi}\mathbf{V}(\boldsymbol{\mu})$, so the variance changes with the expected value, but only depends on the dispersion and the expected value.

Rather than transcribe all the theory (this is an applied class, anyway, right?), we will point to some relevant documentation in SAS and then switch to discussion of the practical aspects and some examples. The function of interest in SAS is **proc genmod**. In the GENMOD Procedure documentation, we should note the following entries:

- **Overview>What is a Generalized Linear Model?**: This gives a short description of the general class of models and why they may be useful.
- **Overview>Examples of Generalized Linear Models**: This gives a few common generalized models with the classification that will be of most interest to us (the type of response variable, the distribution, and the link function).
- **Overview>The GENMOD Procedure**: This is the overview of the whole function, and covers more than we will need. For us, it will more than suffice to note the common link functions and for which kinds of response variables they will be appropriate, the most common distributions (normal, Poisson, binomial, gamma, inverse Gaussian) and how the variance is related to the mean for these distributions, that parameter estimates will be based on maximum likelihood and that deviances will be the analog of sums of squares (from linear regression) in type I and type III analyses of the model terms.
- **Details> Generalized Linear Models Theory**: This goes into far more theory than we need and is mentioned here mostly for completeness and as reference for those interested in a deeper understanding of the theory. Information on **Log-Likelihood Functions** and **Goodness of Fit** may be of general interest, but don’t let the theory scare you; the software will take care of the theoretical aspects for us.

The Wikipedia entry at http://en.wikipedia.org/wiki/Generalized_linear_model also gives a nice description for those wanting another perspective. Keep in mind that the Wikipedia entry also includes details on extensions such as multinomial logistic regression and negative binomial regression.

Practical issues

For our purposes, specification of the model will be fairly similar to specifying a linear or logistic model with **proc reg** or **proc logistic** or **proc glm**. Our **class** statement will look the same. The model will be specified in the same way as in these other functions, but we will need to give a **dist** option to define the response distribution and a **link** option to define the link function. This leaves us to choose an appropriate distribution for our responses, and an appropriate link function (we will see an example where we need to use a non-default scale parameter, but often we can just let **proc genmod** take care of that).

The parameter estimates will still be the β's from the linear predictor, and a statistically significant parameter will be significantly different from 0. Significance will be based on asymptotic normality of the parameter estimates, so we will have Wald statistics like in the logistic case (profile likelihood is also a possible method of estimating the confidence intervals, but we will stick with the default Wald values).

The variance function is critical to this class of models and results based on them, but for named distributions it is defined by the distribution. (Note: As long as a distributional structure can be defined such that the variance is equal to some constant times a function of only the distribution's mean, we could fit that distributional structure into the generalized linear model framework. It's worth knowing that **proc genmod** allows for this extension, but it is not something we will use in class.)

Quick Note on Asymptotic Normality

We have seen Wald statistics and likelihood ratio statistics based on normal approximations already. We have noted that these are based on large sample asymptotics. Asymptotic normality means that as the sample size increases, the distribution of a statistic tends to a normal distribution. For least squares estimates with the assumption of normality (e.g. the β estimates in the linear regression case), we get t statistics with degrees of freedom based on the sample size. The t statistic is derived from the assumption of normal errors.

As the sample size gets large, the t statistic approaches a z statistic (standard normal distribution) and for a sufficiently large sample, we can approximate a t statistic with a z statistic. That's why t tables stop with some finite number of degrees of freedom, like 30 or 100, and then include a row for ∞ degrees of freedom. The ∞ case is the normal approximation (the asymptotic result), and for large samples that approximation will be good enough.

In the case of maximum likelihood estimates, we still have that the distribution of the β estimates will tend to a normal distribution as the sample size gets large (under some conditions that will be met in the generalized linear models case), but we don't have an analogue of the t statistic for small samples. So Wald statistics (based on the asymptotic normality of maximum likelihood estimates) will only be approximate and the approximation will tend to be better for larger sample sizes.

In short, when we see Wald statistics and their p-values, these are approximations based on large samples. We will use the results like we have for t and F statistics in linear models, but they are approximations which will become more accurate for larger samples. For linear regression models, the t and F statistics (given by **proc reg** and **proc glm**) will be more accurate than the approximate results returned by **proc genmod** (which is able to assume less because it covers a broader set of distributions).

Examples & Exercises

US Crimes data set (linear regression as a special case)

In chapter 7, we created a linear regression model for crime rate as a function of several predictors. In class we eventually chose the model based on the variables **Ex0**, **X**, **Ed**, **Age**, and **U2** as our best model.

- Obtain the parameter estimates, type 1 and type 3 analyses for this model using **proc glm**.
- Do the same for the model via **proc genmod**.
- Compare the results. What differences do you notice in these results?

GHQ data set (logistic regression as a special case)

For the ghq data set:

- Use **proc logistic** to obtain the parameter estimates and type 3 analysis using **ghq** and **sex** (a classification variable) as explanatory variables for case probability.
- Do the same using **proc genmod**.
- How do the results compare? What differences do you notice in these results?

OZKids data set (log-linear model for count data)

In Chapter 5, we modeled **days** absent for Australian kids based on several categorical predictor variables. We used an ANOVA model, which is really a linear regression model based on classification explanatory variables. Days absent is really a count variable so a Poisson model would be more appropriate.

- Use a Poisson generalized linear model with log link to model days as a function of the main effects.
- Get the goodness of fit statistics, parameter estimates, and type 1 and 3 analyses.
- Comment on which effects are significant.

The scale parameter in the standard Poisson model is taken to be 1 (we can see this from the fact that the variance for Poisson is the mean μ, so the variance is equal to 1μ). If the scaled deviance and scaled Pearson X^2 are not close to 1 that is an indication that we may need to incorporate a scale other than 1 (this is called an overdispersed model if the scale is greater than 1 and underdispersed if it is less than 1).

- In the previous model, if the scaled deviance and Pearson X^2 values are not roughly 1, use a **scale** option to use an estimate of the scale based on the deviance.

- Comment on the goodness of fit statistics, parameter estimates, type 1 and type 3 analyses. (Note that we have F statistics as well as chi-squares. If the scale is taken to be known, we have asymptotic chi-squares. If the scale is estimated from the deviance or Pearson X^2, we get asymptotic F statistics as described in the **Details> F Statistics** tab in the documentation).
- How do the conclusions of the overdispersed Poisson count model (the one where we estimate the scale) compare with those for the regular Poisson model (the one where we just use the default scale of 1)?
- How do the conclusions of the overdispersed model compare with the ANOVA model we fit before (e.g. would they select the same terms)?

FAP data set (positive data with expected increasing variation)

The FAP data set contains the following variables:

- **male** – 1 for male, 0 for female
- **treat** – 1 for active drug, 0 for placebo
- **base_n** – number of polyps before treatment
- **age** – patient's age in years
- **resp_n** – number of polyps 3 months after treatment

While gender and treatment type would be classification variables, the **male** and **treatment** variables are already coded so we do not need to treat them as classification variables.

- Print the data and notice there are some very large count values.

If we approximate the counts after treatment as continuous rather than discrete and assume the error variance of the response counts increases as the square of the expected value (the mean), we could use a gamma distribution. The most common link for a gamma model is the log link.

- Fit a gamma model with **resp_n** as the response and the other variables as main effects.
- Look at the parameter estimates and type 1 and type 3 analyses and comment on significant components in the model.

Since we are dealing with counts, we may want to try a Poisson model.

- Change the distribution to Poisson and refit the model. (Now the variance will increase with the mean rather than the square of the mean.)
- Again comment on the parameter estimates and type 1 and type 3 analyses.

For diagnostics, we will need to modify the residuals (and other diagnostics) to account for the change in assumptions of the model. Standardized Pearson residuals will account for the fact that the error variance should change with the mean and standardized deviance residuals will take into account a log-likelihood based interpretation of residuals, so we can look at plots of the standardized Pearson and deviance residuals and plot those instead.

- Add output statements to the **genmod** code for the gamma and Poisson models to save the predicted values and standardized Pearson and deviance residuals to an output data set.
- Create scatter plots of these residuals against predicted values. We should be able to diagnose problems with the residuals from these plots like we do with standardized residual plots in linear regression.
  (Note that we can use the **plots** option to the **proc** statement to plot residuals and other diagnostics vs. either observation number or linear predictor value)
- Do we see problems with either of these? (You may need to add a **where** statement to restrict to predicted values less than 100 to see the general trends.)