Chapter 9

Generalized Linear Models

Review: Linear Regression

- Continuous response y
- Predictors x_j
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi}$
- $Y_i \sim N(\hat{y}_i, \hat{\sigma}^2)$

Review: Logistic Regression

- Response probability limited to 0<=p<=1
- $Y_i \sim Ber(\hat{\pi}_i)$ with $\hat{y}_i = \hat{\pi}_i$
- Or $Y_i \sim Bin(n_i, \hat{y}_i / n_i)$ with $\hat{y}_i / n_i = \hat{\pi}_i$
- $\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots$

Generalized Linear Models

- Extend to other types of responses
- Still have a linear component (we have a linearized model)
- Variance for observations is related to expected value
- Parameter estimates still come from maximum likelihood

Terminology

- **response distribution**: distribution family the responses (the y_i's) are assumed to come from
- **linear predictor**: this is the linear combination we denoted by **η**
- **link function**: function we apply to the response to get a linear combination. We will denote the link function by **g**
- **dispersion parameter**: this is like the error variance in linear regression and is denoted by **ϕ**
- **variance function**: the function **V** such that the distribution's variance is **ϕV(μ)**

Generalization

- Response distribution parametrized by **ϕ** and **V(μ)**
- $Y_i \sim \mathcal{D}(\mu_i, \mathrm{V}(\mu_i), \phi)$
- $\mathrm{g}(\hat{y}_i) = \eta_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots$
- μ_i= $\hat{y}_i$ =$\mathrm{g}^{-1}(\eta_i)$

proc genmod

- For generalized linear models and extensions of generalized linear models
- Similar to procs **reg**, **anova**, **glm**, and **logistic**
- Allows continuous and categorical predictors
- Will need to specify response distribution (**dist** option) and link (**link** option) in **model**

Reference for Details

- **Overview>What is a Generalized Linear Model?**: short description
- **Overview>Examples of Generalized Linear Models**: This gives a few common generalized linear 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**: overview of the whole function; focus on common links and distributions, parameter estimates and type 1 and 3 analyses
- **Details> Generalized Linear Models Theory**: the general theory; note **Log-Likelihood Functions** and **Goodness of Fit** sections

Example: US Crimes Linear Regression

- Linear Regression is special case
- Link is identity, and distribution is normal
- Use predictors **Ex0**, **X**, **Ed**, **Age**, and **U2** and response **R**
- Obtain the parameter estimates, type 1 and type 3 analyses using **proc glm**
- Do same using **proc genmod** and compare results

Example: GHQ Logistic Regression

- Logistic Regression is special case
- Link is logit, and distribution is Bernoulli or binomial
- Use predictors **ghq** and **sex** and response **cases** out of **total**
- Obtain the parameter estimates and type 3 analyses using **proc logistic**
- Do same using **proc genmod** and compare results

Exercise: OZKids data set

- Used ANOVA in Chapter 5
- Counts not continuous; Poisson may be better
- Fit Poisson generalized linear model with log link for **days** as a function of main effects
- Get the goodness of fit statistics, parameter estimates, and type 1 and 3 analyses.
- Comment on which effects are significant.
- Compare with significant effects from ANOVA

Exercise: Overdispersed Poisson Model

- $\phi = 1$ in Poisson model
- Can introduce additional dispersion by estimating **scale** if actual dispersion far from 1
- Model is overdispersed if $\phi > 1$
- Underdispersed if $\phi < 1$

Exercise: Overdispersed Poisson Model

- Use **scale** option to estimate the scale based on the deviance.
- Comment on the goodness of fit statistics, parameter estimates, type 1 and type 3 analyses
- Compare conclusions with those from previous Poisson model
- Compare with the ANOVA model we fit before (e.g. would they select the same terms?)

Exercise: FAP Data Set

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

Note: large counts in response

Exercise: Gamma Model

- For positive continuous response values
- Variance grows as square of expected value
- Counts technically discrete, but can approximate with continuous model

Exercise: Gamma Model

- Fit a gamma model with log link, **resp_n** as the response and other variables as main effects
- Note: male and treat are categorical, but coded so we don't need a **class** statement
- Look at parameter estimates and type 1 and type 3 analyses and comment on significant components in the model

Exercise: Poisson Model

- Now use a Poisson model instead (estimate scale based on deviance if needed)
- Variance will increase as the mean (not the mean squared)
- Comment on the parameter estimates and type 1 and type 3 analyses

Exercise: Residual Checks

- Add **output** statements to **genmod** code for 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
- Any problems with either of these?
- Note: May need to add a **where** statement to restrict to predicted values less than 100 to see the general trends

Exercise: More Residual Checks

- **plots** option to the **proc** statement can plot residuals and other diagnostics vs. either observation number or linear predictor value
- Note there are also analogues of Cook's distance, leverages, etc. for these models