Lecture 16: Graphical Methods for Binomial Data and Introduction to Poisson GLMs
(Text Sections 9.1-9.2)

Summary of Graphical Methods for Binomial Data

As discussed in the last lecture, formal goodness-of-fit tests of binomial GLMs rely on the assumption that we have a large number of replicates (i.e. the \( m_i \)'s are large). Graphical methods for choosing a model and assessing its fit also work best when the \( m_i \)'s are large. However, they can give us some insight into the fit even if the asymptotic results of the formal tests cannot be trusted.

Types of plots:

1. To choose a link function: For each continuous predictor variable, plot the observed proportion of successes for each value of this predictor vs. the predictor variable. The shape of the resulting curve can provide guidance in terms of appropriate link functions. If there are few or no replicates, we might bin the values of the predictor and plot the observed proportion of successes vs. the average predictor value in each bin. (For instance, consider problem of plotting the proportion of companies who repay their loans vs. company size in the loan data example.)

2. To assess the fit of the model:
   
   (a) Plot the observed vs. fitted values (i.e. \( y_i \) vs. \( m_i \hat{\pi}_i \)) and look for a straight line.
   
   (b) Plot the Pearson or deviance residuals vs. each continuous predictor variable and look for random scatter and no outlying observations.

3. To assess (partially) independence: If the observations are ordered (e.g. by time), plot Pearson or deviance residuals vs. their order and check for serial correlation.

In addition, if there is a large number of replicates and formal tests show that the saturated model fits “significantly” better than the proposed model, it is possible that these graphs will show that the “lack of fit” is too small to be concerned about in practice.

Introduction to Poisson GLMs

Recall that if \( Y \sim \text{Poisson}(\mu) \), then the distribution of \( Y \) is given by

\[
f_Y(y) = \frac{e^{-\mu} \mu^y}{y!}
\]

and \( \text{E}[Y] = \text{Var}[Y] \). The Poisson distribution is almost always used to model count data where the assumption that \( \text{E}[Y] = \text{Var}[Y] \) seems reasonable, and where the observed data (often the number of 0’s, in particular) is reasonably well described by this distribution.
Count data arise in numerous common settings. Sometimes the counts themselves are of interest. At other times, a continuous response is actually of interest, but is difficult to measure. In these cases, counts may be a surrogate for the response of interest.

Example 1: Health services

Health economics research is concerned with the link between health service utilization and economic variables such as income and price. Ideally, one would measure utilization by expenditures. But, if the data come from surveys of individuals, it is more common to have information on the number of times health services are used (e.g. number of visits to a doctor, number of days in the hospital in the last year). Excess 0’s are common.

Example 2: Accident insurance

In insurance, the frequency of accidents is often a variable of interest because of their impact on insurance premiums. Demographic information (e.g. age, number of years accident-free) may provide useful predictors of accident frequency.

Example 3: Manufacturing defects

The number of defects on a surface may be of interest (as a function of variables in the manufacturing process). Excess 0’s are common.

We need to be careful about how we model $\mu$. When we have multiple observations, it often makes more sense to model a rate rather than the mean. In this case, we need to take the exposure (e.g. number of years for which the individual is reporting doctor visits
in Example 1, area of the surface in question in Example 3) into account in our model. In particular, in Example 1, we might assume that $Y_i \sim \text{Poisson}(t_i \theta)$, where $Y_i$ is the number of doctor visits reported by patient $i$ over the last $t_i$ years. Here, $\theta$ would be the the mean number of doctor visits per year. In Example 3, we might assume that $Y_i \sim \text{Poisson}(a_i \theta)$, where $Y_i$ is the number of defects in a part with surface area $a_i$. Here $\theta$ is the mean number of defects per unit area.

In other cases, the exposure is constant, in which case we do not need to include it explicitly in the model. In Example 2, insurance agencies may specifically collect data on the number of accidents each customer has in a given year. If $Y_i$ is the number of accidents for customer $i$ in a given year, then we might use the model $Y_i \sim \text{Poisson}(\mu)$, where $\mu$ is the mean number of accidents per customer (per year).

Modelling $\mu$

Assume $Y_1, \ldots, Y_n$ are independent with $Y_i \sim \text{Poisson}(\mu_i)$. If we have covariates $x_1, \ldots, x_n$, we might want to use the model

$$
\mu_i = g^{-1} \left( \sum_{j=1}^{p} x_{ij} \beta_j \right).
$$

Recall that

$$
 f_Y(y_i) = \exp\{y_i \log \mu_i - \mu_i - \log y_i! \}
$$

so that $g(\mu_i) = \log(\mu_i)$ is the canonical link.

The log link is by far the most commonly used link function. This link function has the property that

$$
\mu_i = \exp \left( \sum_{j=1}^{p} x_{ij} \beta_j \right)
\begin{align*}
&= e^{x_{i1} \beta_1} \cdots e^{x_{ip} \beta_p} \\
&= g^{-1}(x_{i1} \beta_1) \cdots g^{-1}(x_{ip} \beta_p).
\end{align*}
$$

Therefore, with this model, (functions of) the covariates have a multiplicative rather than additive effect on the mean.

Two other link functions provided by S-PLUS are

- The identity link: $g(\mu) = \mu$.
- The square-root link: $g(\mu) = \sqrt{\mu}$. 

3
However, these link functions can be problematic for predictions of $\mu_i$, since

$$\hat{\eta}_i = \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$$

could be negative. In this case, if $g$ were the square-root link, $\sqrt{\hat{\mu}_i} = \hat{\eta}_i$ (and hence $\hat{\mu}_i$) would be undefined. Likewise, if $g$ were the identity link so that $\hat{\mu}_i = \hat{\eta}_i$, we would have $\hat{\mu}_i < 0$. Since we know $\mu_i \geq 0$, such predictions would not be reasonable.

A Note on Offsets

We will see in the next lecture how to extend the Poisson GLM to the case where we model a rate rather than a mean. Fitting the model in this case will require the use of an offset term. (Recall the use of an offset in the dilution series example in Lecture 14.) Offsets are used when we know the value of one of the regression coefficients.

In a linear regression model, an offset can be absorbed directly into the response. For example, if the coefficient of $x_1$ is a known value $c$, then

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$$

$$Y_i = cx_{i1} + \sum_{j=2}^{p} x_{ij}\beta_j + \epsilon_i$$

$$Y_i - cx_{i1} = \sum_{j=2}^{p} x_{ij}\beta_j + \epsilon_i$$

Defining $Y'_i = Y_i - cx_{i1}$ as our new response variable, we can estimate $\beta_2, \ldots, \beta_p$ in the usual way by regressing $Y'_i$ on $x_2, \ldots, x_p$.

In a GLM, however, this is not possible because

1. The form of the model (i.e. the modelling of the relationship between the covariates and the mean of the observations rather than the relationship between the covariates and the observations themselves) means that we can’t subtract (or do some other simple operation) the offset from the observations.

2. If we were to adjust (e.g. by subtraction of the offset) the responses and then try to build a model for the adjusted responses, it would be difficult to determine an appropriate distribution.