11. Generalized Linear Models: An Introduction

1. Introduction

A synthesis due to Nelder and Wedderburn, generalized linear models (GLMs) extend the range of application of linear statistical models by accommodating response variables with non-normal conditional distributions.

Except for the error, the right-hand side of a generalized linear model is essentially the same as for a linear model.

2. Goals

- To introduce the format and structure of generalized linear models
- To show how the familiar linear, logit, and probit models fit into the GLM framework.
- To introduce Poisson generalized linear models for count data.
- To describe diagnostics for generalized linear models.

3. The Structure of Generalized Linear Models

A generalized linear model consists of three components:

1. A random component, specifying the conditional distribution of the response variable, $Y_i$, given the predictors.
   - Traditionally, the random component is a member of an "exponential family" — the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian families of distributions — but generalized linear models have been extended beyond the exponential families.
   - The Gaussian and binomial distributions are familiar.
   - Poisson distributions are often used in modeling count data. Poisson random variables take on non-negative integer values, $0, 1, 2, \ldots$.
     Some examples are shown in Figure 1.
The gamma and inverse-Gaussian distributions are for positive continuous data; some examples are given in Figure 2.

2. A linear function of the regressors, called the **linear predictor**, 
\[ \eta_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \]
on which the expected value \( \mu_i \) of \( Y_i \) depends.

- The \( X \)'s may include quantitative predictors, but they may also include transformations of predictors, polynomial terms, contrasts generated from factors, interaction regressors, etc.

3. An invertible **link function** \( g(\eta_i) = \eta_i \), which transforms the expectation of the response to the linear predictor.

- The inverse of the link function is sometimes called the **mean function**: 
\[ g^{-1}(\eta_i) = \mu_i. \]

- Standard link functions and their inverses are shown in the following table:

<table>
<thead>
<tr>
<th>Link</th>
<th>( \eta_i = g(\mu_i) )</th>
<th>( \mu_i = g^{-1}(\eta_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>identity</td>
<td>( \mu_i )</td>
<td>( \eta_i )</td>
</tr>
<tr>
<td>log</td>
<td>( \log(\mu_i) )</td>
<td>( e^{\eta_i} )</td>
</tr>
<tr>
<td>inverse</td>
<td>( \mu_i^{-1} )</td>
<td>( \eta_i^{-1} )</td>
</tr>
<tr>
<td>inverse-square</td>
<td>( \mu_i^{-2} )</td>
<td>( \eta_i^{-1/2} )</td>
</tr>
<tr>
<td>square-root</td>
<td>( \sqrt{\mu_i} )</td>
<td>( \eta_i^{1/2} )</td>
</tr>
<tr>
<td>logit</td>
<td>( \log \left( \frac{1}{\mu_i} \right) )</td>
<td>( \frac{1}{1 + e^{-\eta_i}} )</td>
</tr>
<tr>
<td>probit</td>
<td>( \Phi^{-1}(\mu_i) )</td>
<td>( \Phi(\eta_i) )</td>
</tr>
<tr>
<td>log-log</td>
<td>( \log \left( \frac{1}{\mu_i} \right) )</td>
<td>( \exp(-\exp(-\eta_i)) )</td>
</tr>
<tr>
<td>complementary log-log</td>
<td>( \log \left( \frac{1}{\mu_i} \right) )</td>
<td>( 1 - \exp(-\exp(-\eta_i)) )</td>
</tr>
</tbody>
</table>

- The logit, probit, and complementary-log-log links are for **binomial data**, where \( Y_i \) represents the observed proportion and \( \mu_i \) the expected proportion of “successes” in \( n_i \) binomial trials — that is, \( \mu_i \) is the probability of a success.
– For the probit link, \( \Phi \) is the standard-normal cumulative distribution function, and \( \Phi^{-1} \) is the standard-normal quantile function.

– An important special case is binary data, where all of the binomial trials are 1, and therefore all of the observed proportions \( Y_i \) are either 0 or 1. This is the case that we examined the previous lecture.

For distributions in the exponential families, the conditional variance of \( Y_i \) is a function of the mean \( \mu_i \) together with a dispersion parameter \( \phi \) (as shown in the table below).

- For the binomial and Poisson distributions, the dispersion parameter is fixed to 1.
- For the Gaussian distribution, the dispersion parameter is the usual error variance, which we previously symbolized by \( \sigma^2 \) (and which doesn’t depend on \( \mu \)).

\[ \begin{array}{|c|c|c|}
\hline
\text{Family} & \text{Canonical Link} & \text{Range of } Y_i \\
\hline
\text{Gaussian} & \text{identity} & (-\infty, +\infty) \\
\text{binomial} & \text{logit} & 0, 1, \ldots, n_i \\
\text{Poisson} & \text{log} & 0, 1, 2, \ldots \\
\text{gamma} & \text{inverse} & (0, \infty) \\
\text{inverse-Gaussian} & \text{inverse-square} & (0, \infty) \\
\hline
\end{array} \]

The canonical link for each family is not only the one most commonly used, but also arises naturally from the general formula for distributions in the exponential families.

- Other links may be more appropriate for the specific problem at hand.
- One of the strengths of the GLM paradigm — in contrast, for example, to transformation of the response variable in a linear model — is the separation of the link function from the conditional distribution of the response.

GLMs are typically fit to data by the method of maximum likelihood.

- Denote the maximum-likelihood estimates of the regression parameters as \( \hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k \).
- These imply an estimate of the mean of the response, \( \hat{\mu}_i = g^{-1}(\hat{\alpha} + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}) \).

The log-likelihood for the model, maximized over the regression coefficients, is

\[ \log L_0 = \sum_{i=1}^{n} \log p(\hat{\mu}_i; \phi; y_i) \]

where \( p(\cdot) \) is the probability or probability-density function corresponding to the family employed.

- A “saturated” model, which dedicates one parameter to each observation, and hence fits the data perfectly, has log-likelihood

\[ \log L_1 = \sum_{i=1}^{n} \log p(y_i; \phi; y_i) \]

- Twice the difference between these log-likelihoods defines the residual deviance under the model, a generalization of the residual sum of squares for linear models:

\[ D(y; \hat{\mu}) = 2(\log L_1 - \log L_0) \]
• Dividing the deviance by the estimated dispersion produces the scaled deviance: $D(y; \hat{\mu})/\hat{\phi}$.

• Likelihood-ratio tests can be formulated by taking differences in the residual deviance for nested models.

• For models with an estimated dispersion parameter, one can alternatively use incremental $I^2$-tests.

• Wald tests for individual coefficients are formulated using the estimated asymptotic standard errors of the coefficients.

Some familiar examples:

• Combining the identity link with the Gaussian family produces the normal linear model.
  – The maximum-likelihood estimates for this model are the ordinary least-squares estimates.

• Combining the logit link with the binomial family produces the logistic-regression model (linear-logit model).

Combining the probit link with the binomial family produces the linear probit model.

- Although the logit and probit links are familiar, the log-log and complementary log-log links for binomial data are not.
- These links are compared in Figure 3.
- The log-log or complementary log-log link may be appropriate when the probability of the response as a function of the linear predictor approaches 0 and 1 asymmetrically.

Some familiar examples:

• Combining the identity link with the Gaussian family produces the normal linear model.

4. Poisson GLMs for Count Data

Poisson generalized linear models arise in two common formally identical but substantively distinguishable contexts:

1. when the response variable in a regression model takes on non-negative integer values, such as a count;

2. to analyze associations among categorical variables in a contingency table of counts (an application that I won’t take up here).

- The canonical link for the Poisson family is the log link.

Figure 3. Comparison of logit, probit, and complementary log-log links. The probit link is rescaled to match the variance of the logistic distribution, $\pi^2/3$. 
4.1 Poisson Regression

Recall Ornstein's data on interlocking director and top-executive positions among 248 major Canadian firms. Ornstein performed a least-squares regression of the number of interlocks maintained by each firm on the firm's assets, and dummy variables for the firm's nation of control and sector of operation. I found that a square-root transformation of the response variable tends to stabilize residual variance and make the distribution of the residuals more symmetric.

Because the response variable is a count, a Poisson linear model might be preferable. The marginal distribution of number of interlocks, in Figure 4, shows many zero counts and an extreme positive skew. Fitting a Poisson GLM with log link to Ornstein's data produces the following results:

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.32</td>
</tr>
<tr>
<td>Assets</td>
<td>0.0000209</td>
</tr>
<tr>
<td>Nation of Control (baseline: Canada)</td>
<td>0.0000012</td>
</tr>
<tr>
<td>Other</td>
<td>-0.163</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>-0.577</td>
</tr>
<tr>
<td>United States</td>
<td>-0.826</td>
</tr>
<tr>
<td>Sector (baseline: Agriculture and Food)</td>
<td>0.049</td>
</tr>
<tr>
<td>Banking</td>
<td>-0.409</td>
</tr>
<tr>
<td>Construction</td>
<td>-0.620</td>
</tr>
<tr>
<td>Finance</td>
<td>0.677</td>
</tr>
<tr>
<td>Holding Company</td>
<td>0.208</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>0.0527</td>
</tr>
<tr>
<td>Merchandizing</td>
<td>0.178</td>
</tr>
<tr>
<td>Mining</td>
<td>0.621</td>
</tr>
<tr>
<td>Transportation</td>
<td>0.678</td>
</tr>
<tr>
<td>Wood and Forest Products</td>
<td>0.712</td>
</tr>
</tbody>
</table>

An analysis of deviance table for the model shows that all three explanatory variables have highly statistically significant effects:

<table>
<thead>
<tr>
<th>Source</th>
<th>G^2</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assets</td>
<td>390.90</td>
<td>1</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Nation of Control</td>
<td>328.94</td>
<td>3</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Sector</td>
<td>361.46</td>
<td>9</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

The deviance for the null model (with only a constant) is 3737.0, and 1887.4 for the full model; thus an analog of the squared multiple correlation is

\[ R^2 = 1 - \frac{1887.4}{3737.0} = .495 \]

Effect displays for the model are shown in Figure 5.
4.2 Over-Dispersed Binomial and Poisson Models

- The binomial and Poisson GLMs fix the dispersion parameter $\phi$ to 1.
- It is possible to fit versions of these models in which the dispersion is a free parameter, to be estimated along with the coefficients of the linear predictor.
  - The resulting error distribution is not an exponential family.
- The regression coefficients are unaffected by allowing dispersion different from 1, but the coefficient standard errors are multiplied by the square-root of $\phi$.
  - Because the estimated dispersion typically exceeds 1, this inflates the standard errors.
  - That is, failing to account for “over-dispersion” produces misleadingly small standard errors.
- So-called over-dispersed binomial and Poisson models arise in several different circumstances.

- For example, in modeling proportions, it is possible that
  - the probability of success $\mu$ varies for different individuals who share identical values of the predictors (this is called “unmodeled heterogeneity”);
  - or the individual successes and failures for a “binomial” observation are not independent, as required by the binomial distribution.

5. Diagnostics for GLMs

- Most regression diagnostics extend straightforwardly to generalized linear models.
- These extensions typically take advantage of the computation of maximum-likelihood estimates for generalized linear models by iterated weighted least squares (the procedure typically used to fit GLMs).
5.1 Outlier, Leverage, and Influence Diagnostics

5.1.1 Hat-Values

Hat-values for a generalized linear model can be taken directly from the final iteration of the IWLS procedure.

They have the usual interpretation — except that the hat-values in a GLM depend on $Y$ as well as on the configuration of the $X$'s.

5.1.2 Residuals

Several kinds of residuals can be defined for generalized linear models:

- **Response residuals** are simply the differences between the observed response and its estimated expected value: $Y_i - \hat{\mu}_i$.

- **Working residuals** are the residuals from the final WLS fit. These may be used to define partial residuals for component-plus-residual plots (see below).

- **Pearson residuals** are case-wise components of the Pearson goodness-of-fit statistic for the model:
  \[
  \phi^{1/2}(Y_i - \hat{\mu}_i) \\
  \sqrt{V(Y_i|\eta_i)}
  \]
  where $\phi$ is the dispersion parameter for the model and $V(Y_i|\eta_i)$ is the variance of the response given the linear predictor.

- **Standardized Pearson residuals** correct for the conditional response variation and for the leverage of the observations:
  \[
  R_{P_i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(Y_i|\eta_i)(1 - h_i)}}
  \]

- **Deviance residuals**, $D_i$, are the square-roots of the case-wise components of the residual deviance, attaching the sign of $Y_i - \hat{\mu}_i$.

- **Standardized deviance residuals** are
  \[
  R_{D_i} = \frac{D_i}{\sqrt{\phi(1 - h_i)}}
  \]

Several different approximations to studentized residuals have been suggested.

- To calculate exact studentized residuals would require literally refitting the model deleting each observation in turn, and noting the decline in the deviance.

- Here is an approximation due to Williams:
  \[
  E_i^* = \sqrt{(1 - h_i)R_{P_i}^2 + h_iR_{D_i}^2}
  \]
  where, once again, the sign is taken from $Y_i - \hat{\mu}_i$.

- A Bonferroni outlier test using the standard normal distribution may be based on the largest absolute studentized residual.
5.1.3 Influence Measures

- An approximation to Cook’s distance influence measure is
  \[ D_i = \frac{R^2_i}{\phi(k+1)} \times \frac{h_i}{1-h_i} \]

- Approximate values of dfbeta_{ij} and dfbeta_{sij} (influence and standardized influence on each coefficient) may be obtained directly from the final iteration of the IWLS procedure.

- There are two largely similar extensions of added-variable plots to generalized linear models, one due to Wang and another to Cook and Weisberg.

5.2 Nonlinearity Diagnostics

- Component-plus-residual plots extend straightforwardly to generalized linear models.
  - Nonparametric smoothing of the resulting scatterplots can be important to interpretation, especially in models for binary responses, where the discreteness of the response makes the plots difficult to examine.
  - Similar effects can occur for binomial and Poisson data.

- Component-plus-residual plots use the linearized model from the last step of the IWLS fit.
  - For example, the partial residual for \( X_j \) adds the working residual to \( B_jX_j \).
  - The component-plus-residual plot graphs the partial residual against \( X_j \).

An illustrative component+residual plot, for assets in the over-dispersed Poisson regression fit to Ornstein’s interlocking-directorate data appears in Figure 6.

- This plot is difficult to examine because of the large positive skew in assets, but it appears as if the assets slope is a good deal steeper at the left than at the right.
- I therefore investigated transforming assets down the ladder of powers and roots, eventually arriving at the log transformation, the component+residual plot for which appears quite straight (Figure 7).
Figure 7. Component+residual plot for log(assets) in the respecified over-dispersed Poisson regression model for Ornstein's data.