

STAT 545/GS01 1233 193 5528:
Generalized Linear Models and Categorical Data
Analysis (Part II)
2: GLM Estimation, Inference and Model Checking

Ziyi Li

Department of Biostatistics
The University of Texas MD Anderson Cancer Center

ZLi16@mdanderson.org

Fall 2022

How to solve GLM: Estimation

- The solutions to the likelihood equations are the MLE of β , denoted by $\hat{\beta}$.
- Typically, the likelihood equations for the GLMs are nonlinear in β .
- Three ways to maximize these nonlinear equations and obtain $\hat{\beta}$:
 - 1 Newton-Raphson algorithm.
 - 2 Fisher scoring algorithms.
 - 3 Iteratively reweighted least squares (IRLS).

How to solve GLM: Estimation

1 Newton-Raphson algorithm

Let $\mathbf{H}(\boldsymbol{\beta}) = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}$ denote the Hessian matrix of $\ell(\boldsymbol{\beta})$. Given $\mathbf{U}(\hat{\boldsymbol{\beta}}) = 0$, a Taylor series expansion of $\mathbf{U}(\hat{\boldsymbol{\beta}})$ about the current value, $\boldsymbol{\beta}^{(t)}$, leads to

$$\begin{aligned} 0 &= \mathbf{U}(\boldsymbol{\beta}^{(t)}) + \mathbf{H}(\boldsymbol{\beta}^{(t)})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)}) \\ \rightarrow \hat{\boldsymbol{\beta}} &\approx \boldsymbol{\beta}^{(t)} - \mathbf{H}^{-1}(\boldsymbol{\beta}^{(t)})\mathbf{U}(\boldsymbol{\beta}^{(t)}) \end{aligned}$$

The last (approximate) equation suggests that $\hat{\boldsymbol{\beta}}$ can be obtained using an iterative procedure:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \mathbf{H}^{-1}(\boldsymbol{\beta}^{(t)})\mathbf{U}(\boldsymbol{\beta}^{(t)}) \quad \text{for } t = 0, 1, \dots$$

Once the algorithm converges, the last $\boldsymbol{\beta}^{(t)}$ is a good approximation to $\hat{\boldsymbol{\beta}}$.

2 Fisher scoring algorithm

The Newton-Raphson algorithm can be rewritten as

$$\beta^{(t+1)} = \beta^t + \mathbf{J}_n^{-1}(\beta^{(t)}) \mathbf{U}(\beta^{(t)}) \quad \text{for } t = 0, 1, \dots$$

where $\mathbf{J}_n = -\mathbf{H} = -\frac{\partial^2 \ell}{\partial \beta^T \partial \beta}$ is the observed information matrix.

If we replace \mathbf{J}_n with $\mathbf{I}_n = E(\mathbf{J}_n)$ the Fisher information, we have the Fisher scoring algorithm:

$$\beta^{(t+1)} = \beta^t + \mathbf{I}_n^{-1}(\beta^{(t)}) \mathbf{U}(\beta^{(t)}) \quad \text{for } t = 0, 1, \dots$$

3 IRLS algorithm

Given $I_n(\beta) = \mathbf{X}^T \mathbf{W} \mathbf{X}$ with $\mathbf{W} = \text{diag} \left\{ \frac{1}{\text{Var}(Y_i)(g'(\mu_i))^2} \right\}$, we can rewrite the Fisher scoring algorithm as follows:

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} + (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \Delta^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)}) \\ &= (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \{ (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X}) \beta^{(t)} + \mathbf{X}^T \Delta^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)}) \} \\ &= (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(t)} \{ \mathbf{X} \beta^{(t)} + (\mathbf{W}^{(t)})^{-1} \Delta^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)}) \}\end{aligned}$$

where $\mathbf{z}^{(t)} = \mathbf{X} \beta^{(t)} + (\mathbf{W}^{(t)})^{-1} \Delta^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)}) = \boldsymbol{\eta}^{(t)} + (\mathbf{W}^{(t)})^{-1} \Delta^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)})$ with the element $z_i^{(t)} = \eta_i^{(t)} + g'(\mu_i^{(t)})(y_i - \mu_i^{(t)})$ and the superscript (t) indicates that the quantity is evaluated at $\beta^{(t)}$.

3 IRLS algorithm (continue)

$$\beta^{(t+1)} = (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(t)} \{ \mathbf{X} \beta^{(t)} + (\mathbf{W}^{(t)})^{-1} \Delta^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)}) \}$$

where $\mathbf{z}^{(t)} = \boldsymbol{\eta}^{(t)} + (\mathbf{W}^{(t)})^{-1} \Delta^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)})$ with the element $z_i^{(t)} = \eta_i^{(t)} + g'(\mu_i^{(t)})(y_i - \mu_i^{(t)})$ and the superscript $^{(t)}$ indicates that the quantity is evaluated at $\beta^{(t)}$. This suggests the IRLS algorithm:

- 1 Start with an initial value $\beta^{(0)}$.
- 2 At the t -th iteration, compute the weighted least squares estimates $\beta^{(t+1)}$ with $\mathbf{z}^{(t)}$ as the “pseudo” outcome variable, \mathbf{X} the design matrix, and $\mathbf{W}^{(t)}$ the weight matrix.
- 3 Repeat 2 until convergence.

- To perform statistical inference, we need the information matrix for β ,

$$I_n(\beta) = -E \left(\frac{\partial^2 \ell}{\partial \beta^T \partial \beta} \right) = \text{Var} \left(\frac{\partial \ell}{\partial \beta} \right)$$

- Given $\frac{\partial \ell}{\partial \beta} = \mathbf{X}^T \Delta (\mathbf{y} - \boldsymbol{\mu})$ with $\Delta = \text{diag} \left\{ \frac{1}{\text{Var}(Y_i) g'(\mu_i)} \right\}$,

$$\begin{aligned} I_n(\beta) &= \text{Var}(\mathbf{U}(\beta)) \\ &= \text{Var}(\mathbf{X}^T \Delta (\mathbf{y} - \boldsymbol{\mu})) \\ &= \mathbf{X}^T \Delta \text{Var}(\mathbf{y}) (\mathbf{X}^T \Delta)^T \\ &= \mathbf{X}^T \Delta \text{Var}(\mathbf{y}) \Delta^T \mathbf{X} \\ &= \mathbf{X}^T \mathbf{W} \mathbf{X} \end{aligned}$$

where $\mathbf{W} = \text{diag} \left\{ \frac{1}{\text{Var}(Y_i) (g'(\mu_i))^2} \right\}$.

$$I_n(\beta) = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

- \mathbf{X} is the $n \times p$ design matrix
- \mathbf{W} is an $n \times n$ diagonal matrix: $\mathbf{W} = \text{diag}\left\{\frac{1}{\text{Var}(Y_i)(g'(\mu_i))^2}\right\}$ and each diagonal element is a function of parameter β .
- The asymptotic covariance matrix of $\hat{\beta}$ can be estimated by

$$\widehat{\text{Cov}}(\hat{\beta}) = \hat{\mathbf{I}}^{-1} = (\mathbf{X}' \widehat{\mathbf{W}} \mathbf{X})^{-1}$$

where $\widehat{\mathbf{W}}$ is evaluated at $\hat{\beta}$ and $\hat{\phi}$, a consistent estimator of ϕ , (e.g., $\hat{\phi} = \sum_i \frac{(y_i - \hat{\mu}_i)^2 / b''(\hat{\theta}_i)}{n-p}$).

- Given $\hat{\beta}$ and $\widehat{\text{Cov}}(\hat{\beta})$, we can construct confidence intervals for $\hat{\beta}$ and contrasts $\mathbf{c}'\beta$.

Example: Poisson loglinear model, $\log \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ with $\theta_i = \log \mu_i$ and $b(\theta_i) = \exp(\theta_i) = \mu_i$.

- Since the log link is the canonical link, the score equations are

$$\sum_i (y_i - \mu_i) x_{ij} = 0, \quad j = 1, 2, \dots, p.$$

- The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ can be estimated by

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X} \right)^{-1}, \text{ where}$$

$$\mathbf{W} = \text{diag} \left\{ \frac{1}{\text{Var}(Y_i)(g'(\mu_i))^2} \right\} = \text{diag} \{ \mu_i \}, \text{ noting that } \text{Var}(Y_i) = \mu_i \text{ and } g(\mu_i) = \log(\mu_i).$$

Hypothesis Testing:

- **Simple $H_0 : \beta = \beta^0$** where β is of $p \times 1$,
 - 1 Wald Test: $T_1 = (\hat{\beta} - \beta^0)^T I_n(\hat{\beta}) (\hat{\beta} - \beta^0) \sim \chi_p^2$.
 - 2 Likelihood Ratio Test: $T_2 = 2 \log \frac{L(\hat{\beta})}{L(\beta^0)} \sim \chi_p^2$.
 - 3 Score Test: $T_3 = U(\beta^0)^T I_n^{-1}(\beta^0) U(\beta^0) \sim \chi_p^2$.

Hypothesis Testing:

- **Composite H_0** : $\beta_1 = \beta_1^0$ where $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ and β_1 and β_2 are of $p_1 \times 1$ and $p_2 \times 1$ respectively,
 - 1 Wald Test: $T_1 = (\hat{\beta}_1 - \beta_1^0) \{I_n^{11}(\hat{\beta})\}^{-1} (\hat{\beta}_1 - \beta_1^0) \sim \chi_{p_1}^2$.
 - 2 Likelihood Ratio Test: $T_2 = 2 \log \frac{L(\hat{\beta})}{L(\hat{\beta}^0)} \sim \chi_{p_1}^2$, where $\hat{\beta}^0 = \begin{pmatrix} \beta_1^0 \\ \hat{\beta}_2(\beta_1^0) \end{pmatrix}$ is the MLE under H_0 .
 - 3 Score Test: $T_3 = U_1(\hat{\beta}^0)^T \{I_n^{11}(\hat{\beta}^0)\}^{-1} U_1(\hat{\beta}^0) \sim \chi_{p_1}^2$, where $U_1(\hat{\beta}^0)$ is defined as $\mathbf{U} = \begin{pmatrix} U_1(\hat{\beta}^0) \\ U_2(\hat{\beta}^0) \end{pmatrix}$ and I_n^{11} is defined as $I_n = \begin{pmatrix} I_n^{11} & I_n^{12} \\ I_n^{21} & I_n^{22} \end{pmatrix}$.

GLM: Deviance and Goodness of Fit

- The **saturated model** is the model that has one separate parameter for each observation i (i.e., $p = n$).
- Let $\tilde{\mu}$ denote μ evaluated at $\tilde{\beta}$, the coefficient estimates under the saturated model. Let $\ell(\mu; \mathbf{y})$ denote the observed data log-likelihood under a given model.
- In the saturated model, it can be shown that $\tilde{\mu}_i = y_i$, noting that this achieves the maximum value of $\ell(\mu; \mathbf{y})$, among all possible models. In other words, the saturated model gives the perfect fit to the data and explains all variation through the systematic component of the model.
- Why not always use the saturated model?
- It can serve as baseline for comparison with other, more parsimonious models: balance between goodness of fit and the number of parameters.

GLM: Deviance and Goodness of Fit

- For an unsaturated model M_0 , denote the MLEs by $\hat{\theta}$ and $\hat{\mu}$.
- The likelihood ratio test statistic for testing the model of interest, M_0 against the alternative model M_1 (the saturated model) is:

$$\begin{aligned} & -2 \log \frac{\text{ML for model } M_0}{\text{ML for saturated model } M_1} \\ &= -2\{\ell(\hat{\mu}; y) - \ell(\tilde{\mu}; y)\} \\ &= 2\{\ell(\tilde{\mu}; y) - \ell(\hat{\mu}; y)\} \\ &= 2\{\ell(y; y) - \ell(\hat{\mu}; y)\} \end{aligned}$$

- **Deviance** is derived from this test statistic, and hence describes lack of fit compared to the saturated model.

GLM: Deviance and Goodness of Fit

- For GLMs, the **likelihood ratio test statistic** can be expressed as

$$\begin{aligned} & 2\{\ell(\mathbf{y}; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}; \mathbf{y})\} \\ &= 2\sum_i \{y_i \tilde{\theta}_i - b(\tilde{\theta}_i)\} / a(\phi) - 2\sum_i \{y_i \hat{\theta}_i - b(\hat{\theta}_i)\} / a(\phi) \\ &= \frac{2}{\phi} \sum_i \omega_i \{y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))\} \end{aligned}$$

where $a(\phi) = \phi/\omega_i$ often the case for GLMs. This defines the scaled deviance.

- The **deviance** for Model M_0 , $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$, is defined as:

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\mu}}) &\equiv 2\phi\{\ell(\mathbf{y}; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}; \mathbf{y})\} \\ &= 2\sum_i \omega_i \{y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))\} \end{aligned}$$

GLM: Deviance and Goodness of Fit

- The greater the deviance is, the poorer the fit is.
- When $a(\phi) = 1$, the scaled deviance is the same as the deviance.
- For most GLMs (but not always) the scaled deviance has an asymptotic χ^2 distribution:
 - Poisson count data
 - Grouped Binomial data

GLM: Deviance and Goodness of Fit

Example:

- Poisson Loglinear model $\log \mu = \mathbf{X}\beta$.
- We know that $\tilde{\theta}_i = \log \hat{\mu}_i$, $b(\hat{\theta}_i) = \exp(\hat{\theta}_i) = \hat{\mu}_i$, $\tilde{\theta}_i = \log \tilde{\mu}_i = \log y_i$ and $b(\tilde{\theta}_i) = \tilde{\mu}_i = y_i$.
- It follows that the deviance is

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\sum_i \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\},$$

which is equivalent to the G^2 statistic for two-way contingency tables.

GLM: Deviance and Goodness of Fit

The forms of the deviances for some common univariate distributions:

- Normal:

$$\sum_i (y_i - \hat{\mu}_i)^2$$

- Poisson:

$$2\sum_i \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$$

- Binomial:

$$2\sum_i \{y_i \log(y_i/\hat{\mu}_i) + (m_i - y_i) \log((m_i - y_i)/(m_i - \hat{\mu}_i))\}$$

- Gamma:

$$2\sum_i \{-\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i\}$$

- If $\hat{\mu}_i = y_i$, what happens to $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$?

GLM: Comparison of Models

How to compare any two nested models?

- **Nested models:** consider two models, M_0 with fitted values $\hat{\mu}_0$ and M_1 with fitted values $\hat{\mu}_1$; if M_0 is a special case of M_1 , M_0 is said to be nested within M_1 .
- Since M_0 is a reduced model (or sub-model) of M_1 , a smaller set of parameters, say $\beta_0 = (\beta_0, \beta_1, \dots, \beta_q)$, are used in M_0 than in M_1 , say $\beta_1 = (\beta_0, \beta_1, \dots, \beta_p)$, where $q < p$.
- Then,

$$\ell(\hat{\mu}_0; \mathbf{y}) \leq \ell(\hat{\mu}_1; \mathbf{y}),$$

and

$$D(\mathbf{y}; \hat{\mu}_0) \geq D(\mathbf{y}; \hat{\mu}_1).$$

Note: $D(\mathbf{y}; \hat{\mu}_0)$ ($D(\mathbf{y}; \hat{\mu}_1)$) compares model M_0 (M_1) with the saturated model and the saturated model is the same.

GLM: Comparison of Models

- The likelihood ratio test statistic for $H_0 : M_0$ vs $H_a : M_1$ is:

$$\begin{aligned} & 2\{\ell(\hat{\boldsymbol{\mu}}_1; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}_0; \mathbf{y})\} \\ &= 2\{\ell(\hat{\boldsymbol{\mu}}_1; \mathbf{y}) - \ell(\tilde{\boldsymbol{\mu}} = \mathbf{y}; \mathbf{y})\} - 2\{\ell(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) - \ell(\tilde{\boldsymbol{\mu}} = \mathbf{y}; \mathbf{y})\} \\ &= D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1) \end{aligned}$$

- The likelihood ratio test statistic comparing two models is simply the difference between the deviances.
- Under some regularity conditions, the difference in deviances above is asymptotically χ^2 with degrees of freedom equal to the difference in the number of parameters in the two models, i.e., $p - q$.

GLM: Residuals

- When a GLM fits poorly according to an overall goodness-of-fit test, examination of residuals highlights where the fit is poor.
- Let $D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_i d_i$, where

$$d_i = 2\omega_i \{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i)) \}$$

- The **deviance residual** for observation i is defined as

$$\sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i)$$

- An alternative is the **Pearson residual**,

$$e_i = \frac{y_i - \hat{\mu}_i}{(\widehat{\text{Var}}(Y_i))^{1/2}}$$

- For instance, for a Poisson GLM, $\text{var}(Y_i) = \mu_i$, then the Pearson residual is

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

- Residual exists for GLM. Analogous to linear regression, it measures the difference between the outcome and predicted
- There are 3 types of residuals: Pearson, deviance, and standardized
- They are useful with categorical covariates and large sample size per category. 📖 Page 217, Table 6.5
- Could just do a visual check with grouping or use some measures or tests
- Not useful when there is a continuous covariate

GLM: Residuals

- When a GLM fits poorly according to an overall goodness-of-fit test, examination of residuals highlights where the fit is poor.
- Let $D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_i d_i$, where

$$d_i = 2\omega_i \{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i)) \}$$

- The **deviance residual** for observation i is defined as

$$\sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i)$$

- An alternative is the **Pearson residual**,

$$e_i = \frac{y_i - \hat{\mu}_i}{(\widehat{\text{Var}}(Y_i))^{1/2}}$$

- The standardized Pearson residual for the i th observation is

$$r_i = \frac{y_i - \hat{\mu}_i}{\{\widehat{\text{Var}}(Y_i)(1 - \hat{h}_i)\}^{1/2}} = \frac{e_i}{\sqrt{1 - \hat{h}_i}}.$$

GLM: Residuals

- Both the deviance residual and Pearson residual are less variable than the standard normal. **Why?**
- Standardized residuals divide the raw residuals by their asymptotic standard errors, which are not $\sqrt{\text{Var}(Y)}$ as shown in the Pearson residual.
- For GLMs, the asymptotic variance matrix of the raw residuals $\mathbf{y} - \hat{\boldsymbol{\mu}}$ is

$$\text{Var}(\mathbf{y} - \hat{\boldsymbol{\mu}}) = (\text{Var}(\mathbf{y}))^{1/2}(\mathbf{I} - \mathbf{H})(\text{Var}(\mathbf{y}))^{1/2}$$

where \mathbf{I} is the identity matrix and \mathbf{H} is the hat matrix, similar to the one in linear regression:

$$\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2}$$

- \mathbf{W} is the $n \times n$ diagonal matrix with diagonal elements $w_i = \frac{1}{\text{Var}(y_i)(g'(\mu_i))^2}$.

Residuals example: logistic regression

Let y_i denote the binomial outcome for n_i trials at level i of the categorical covariates (combination), $i = 1, 2, \dots, N$, $y_i = 0, 1, 2, \dots, n_i$. Let $\hat{\pi}_i$ denote the model estimate of $P(Y = 1)$ at level i .

- Pearson residual is like the residual for linear regression but with standardization:
$$e_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$
- Deviance residual is motivated from the likelihood and deviance (which resembles the sum of squares in linear regression): $\sqrt{d_i} \times \text{sign}(y_i - n_i \hat{\pi}_i)$ with
$$d_i = 2 \left(y_i \log \frac{y_i}{n_i \hat{\pi}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - n_i \hat{\pi}_i} \right)$$
- Standardized Pearson residual is the Pearson residual with further adjustment of its variance to make it closer to a $N(0, 1)$ distribution:
$$r_i = \frac{e_i}{\sqrt{1 - \hat{h}_i}}. \text{ (most useful of the three; } r_i > 2 \text{ or } 3 \text{ indicates lack of fit)}$$

These residuals are more useful when the outcome variable is more granular (e.g., continuous or Poisson rather than binary) or y_i is large

GLM: Influence Diagnostics

Deletion Diagnostics

- Several deletion diagnostic measures are available, which are computed by removing one observation (say, the i -th) at a time, including:
 - ① $\Delta\beta_j^{(-i)} \equiv \hat{\beta}_j - \hat{\beta}_j^{(-i)}$, where $\hat{\beta}_j^{(-i)}$ denotes the parameter estimate with the i -th observation removed.
 - ② $\Delta\mu_j^{(-i)} \equiv \hat{\mu}_j - \hat{\mu}_j^{(-i)}$, where μ_j is the mean of Y_i .
 - ③ The change in G^2 or D GOF statistics when the observation i is deleted.
 - ④ A measure of the change in a joint confidence interval for the parameters produced by deleting the observation.
 - ⑤ Cook's distance, a summary measure of the change in predicted values for all observations after deleting the observation.
- The larger the value of each of the above measures is for a particular observation, the greater the influence that observation has on the model fit.

These residuals are more useful when the outcome variable is more granular (e.g., continuous or Poisson rather than binary) or y_i is large


GLM: Goals in Model Selection

- With several explanatory variables, there are many potential models.
- Two competing goals:
 - ① The final model should be complex enough to fit the data well.
 - ② The final model should be simple enough to interpret.
- Every model is an approximation to reality.
- Some are better than others, but we never know which is better.
- Parsimonious models, that fit the data adequately, have the advantage of ease of interpretation and more accuracy of predicting future observations.

Forward, backward, and stepwise model selection

- Forward procedure: (1) start with just the intercept (2) at each step, add the covariate with the smallest p-value in likelihood ratio or Wald test (3) stop when no more significant covariate is available (However, it can stop prematurely due to lack of power)
- Stepwise procedure: at each step, retest the significance of the terms added at previous stages
- Backward procedure: (1) start with full model (2) at each step, remove the covariate with the largest p-value (3) stop when all remaining covariates are significant. (However, full model may not be stable)
- The dummy variables for a single categorical covariate should be added or removed together (likelihood ratio test); do not place an interaction in the model without the main effect terms
- E.g., SAS PROC LOGISTIC offers additional entry and exit p-value criteria

Further comment on forward, backward, and stepwise model selection


-  Page 211, Table 6.2 illustrates
 - three-way interaction is usually not significant (e.g., lack of power) and not desirable (hard to interpret)
 - dropping multiple covariates at once using likelihood ratio test (LRT) or dropping them one at a time (Wald or LRT)
- All these procedures are not rigorously justified (*ad hoc*); use with caution!
- Modern approaches are available (LASSO, bagging, etc.)
- Philosophically, there is no such thing as “the correct model” or “the true model”: ALL MODELS ARE WRONG, SOME ARE USEFUL — George Box

AIC and BIC

Select the model with smaller AIC or BIC (L : maximized log likelihood; p : number of parameters in the model; n : sample size)

$$AIC = -2L + 2p$$

$$BIC = -2L + \log(n)p$$

- **Rationale:** Including more covariates will always include the log likelihood, but may cause overfitting; so we put a “penalty” by adjusting for the size of the model. There are mathematical reasons why the penalty must take this form.
- Other penalties are available: HQ, DIC, etc.
- BIC puts more penalty on larger model, and therefore tends to select the simpler model  Page 213
- Like scatter plot smoothing, the “desired” amount of penalty is a somewhat subjective choice
- Need a comprehensive assessment of AIC/BIC, significance, residuals, scientific rationale, parsimony and interpretability, etc.