BIOSTAT 651
Notes: GLM Diagnostics

- Topics:
  - Goodness of fit
  - Residuals
  - Influence Measure

Goodness of Fit: General Considerations

- Measure goodness of fit by how well $\widehat{\mu}_i$ replace Y_i
  - $\mathbf{Y}$: $n-$dimensional
  - $\widehat{\boldsymbol{\beta}}$: $q-$dimensional
- *Saturated model*: n parameters (one per unique data point)
  - fits data perfectly
  - no data reduction
- *Null model*:
  - $\widehat{\mu}_i = \widehat{\mu}$ for all i
  - e.g., intercept-only model
  - maximum degree of data summarization
  - fit may be very poor
- The above models are use typically useful only for *judging the fit* of the current model

Deviance: Derivation

- *Deviance*: generalization of the sum of squares of residuals in linear regression.
- Derived by first comparing the likelihoods of the *fitted* and *saturated* models,

$$\left\{ \frac{L(\widetilde{\boldsymbol{\theta}})}{L(\widehat{\boldsymbol{\theta}})} \right\}^2$$

  where $\widetilde{\boldsymbol{\theta}}$ is based on the saturated model
- Then, work on the log scale,

$$2 \times \{\ell(\widetilde{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}})\}$$

- Now, consider a single data point:

$$\begin{aligned} \ell(\widehat{\theta}_i) &= \frac{Y_i\widehat{\theta}_i - b(\widehat{\theta}_i)}{a(\phi)} \\ \ell(\widetilde{\theta}_i) &= \frac{Y_i\widetilde{\theta}_i - b(\widetilde{\theta}_i)}{a(\phi)} \end{aligned}$$

Deviance: Derivation (continued)

- Take difference, sum over all subjects, remove scaling:

$$D \quad = \quad 2\sum_{i=1}^{n}\left[Y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - \{b(\widetilde{\theta}_i) - b(\widehat{\theta}_i)\}\right]$$

  which is known as the *Deviance*

- $D^* = D/a(\phi)$ is referred to as the *Scaled Deviance*
  - Note: In the book, $\frac{1}{a(\phi)} 2\sum_{i=1}^{n}\left[Y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - \{b(\widetilde{\theta}_i) - b(\widehat{\theta}_i)\}\right]$ is the Deviance, and $a(\phi)D$ is the Scaled Deviance.

- When the model fits well, $D^* \sim \chi^2_{n-q}$ asymptotically.

- Examples:

Normal
$$D = \sum_{i=1}^{n} (Y_i - \widehat{\mu}_i)^2$$

$$D^* = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \widehat{\mu}_i)^2$$

Poisson
$$D = D^* = 2 \left[\sum_{i=1}^{n} Y_i \ \log \frac{Y_i}{\widehat{\mu}_i} - \sum_{i=1}^{n} (Y_i - \widehat{\mu}_i) \right]$$

Binomial
$$D = D^* = 2 \sum_{j=1}^{m} \left[Y_j \ \log \left(\frac{Y_j}{\widehat{\mu}_j} \right) + (n_j - Y_j) \ \log \left(\frac{n_j - Y_j}{n_j - \widehat{\mu}_j} \right) \right]$$

Pearson Chi-Square Statistic

- Another measure of a model's fit, the *Pearson Chi-Square Statistic*,

$$X_P^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{V}(Y_i)}$$

- When the model fits well, $X_P^2 \sim \chi^2_{n-q}$ asymptotically.

Goodness of fit tests

- In principle, both (scaled) Deviance and Pearson statistics asymptotically follows χ^2_{n-q} distribution, so we can test GOF.
- However, it does not always work. Especially in logistic regression.
- There exists several modifications, including a test proposed by Hosmer and Lemeshow. (We will cover later)

Comparing GOF Statistics

- Deviance decreases when covariates are added to a model
  - note: applies to *nested* models
- Pearson X^2 has intuitive appeal
- Can carry out hypothesis testing using Deviance
  - applies to nested models
  - equivalent to Likelihood Ratio Test

Difference in Deviances: LRT

- Scaled deviance
  - for a given model, with MLE $\widehat{\boldsymbol{\beta}}$,

$$D^* \quad = \quad 2 \times \{\ell(\widetilde{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}})\}$$

    where $\widetilde{\boldsymbol{\beta}}$ corresponds to a *saturated* model

    i.e., one parameter for each unique covariate pattern

- If we let D_0^* and D_1^* denote the scaled deviances under H_0 and H_1, respectively, then the LRT can be computed as

$$X_L^2 \quad = \quad D_0^* - D_1^*$$

Residuals

- Deviance and Pearson X^2 are global measures of goodness-of-fit
  - summary of model's fit
- Also useful to evaluate the model's performance for individual subjects or groups of subjects
- Pearson residuals:

$$\widehat{r}_i^P = \frac{Y_i - \widehat{\mu}_i}{\widehat{V}(Y_i)^{1/2}}$$

  - Combining the Pearson residuals $\Rightarrow X_p^2$

$$X_P^2 = \sum_{i=1}^{n} \{\widehat{r}_i^P\}^2$$

- Deviance residuals:
  - $D = \sum_{i=1}^{n} D_i$

$$D_i = 2\left[Y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - \{b(\widetilde{\theta}_i) - b(\widehat{\theta}_i)\}\right]$$

  - then, define

$$\widehat{r}_i^D \quad = \quad \text{sign}(Y_i - \widehat{\mu}_i)\sqrt{|D_i|}$$

    i.e, such that $D = \sum_{i=1}^{n} \{\widehat{r}_i^D\}^2$

Examples

- Generate data from the following model

$$log(\lambda_i) = 1 + 0.5x + 0.5x^2, \quad 2 < x < 3$$

$$Y_i \sim Poisson(\lambda_i)$$

- Use the following model to fit the data
  - True model

$$log(\lambda_i) = \beta_0 + \beta_1 x + \beta_2 x^2$$

  - Missing x^2 term

$$log(\lambda_i) = \beta_0 + \beta_1 x$$

  - Identity link

$$\lambda_i = \beta_0 + \beta_1 x + \beta_2 x^2$$

True Model

The GENMOD Procedure

Model Information	
Data Set	WORK.A
Distribution	Poisson
Link Function	Log
Dependent Variable	Y

Number of Observations Read	1000
Number of Observations Used	1000

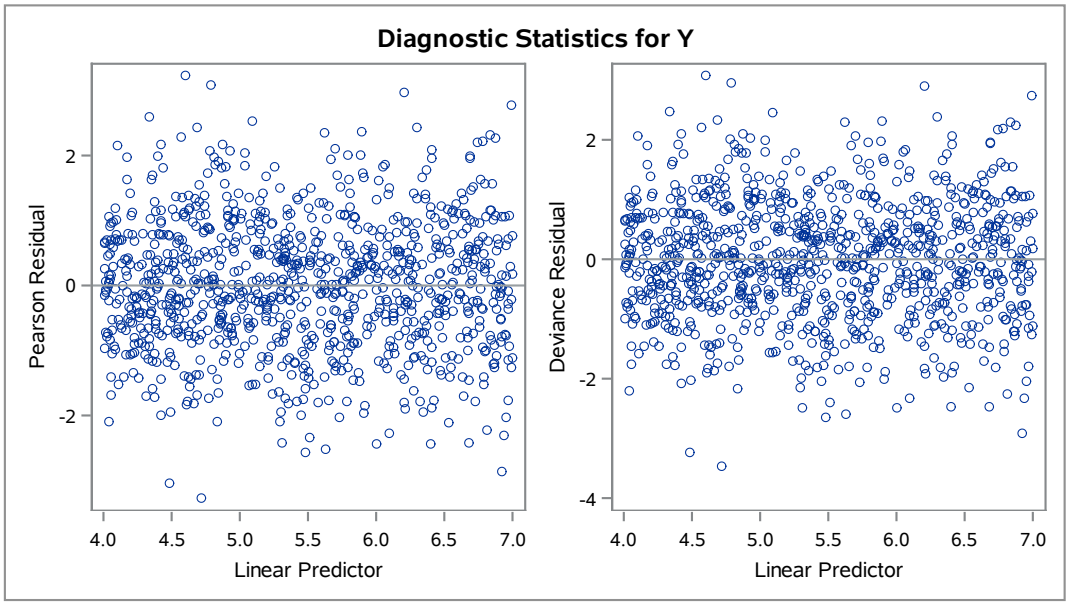
Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	997	974.5321	0.9775
Scaled Deviance	997	974.5321	0.9775
Pearson Chi-Square	997	972.9738	0.9759
Scaled Pearson X2	997	972.9738	0.9759
Log Likelihood		1708332.3290	
Full Log Likelihood		-4127.4402	
AIC (smaller is better)		8260.8803	
AICC (smaller is better)		8260.9044	
BIC (smaller is better)		8275.6036	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.9810	0.1904	0.6078	1.3543	26.54	<.0001
X	1	0.5049	0.1469	0.2171	0.7928	11.82	0.0006
X2	1	0.5005	0.0281	0.4454	0.5556	316.95	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

The GENMOD Procedure



The GENMOD Procedure

Model Information	
Data Set	WORK.A
Distribution	Poisson
Link Function	Log
Dependent Variable	Y

Number of Observations Read	1000
Number of Observations Used	1000

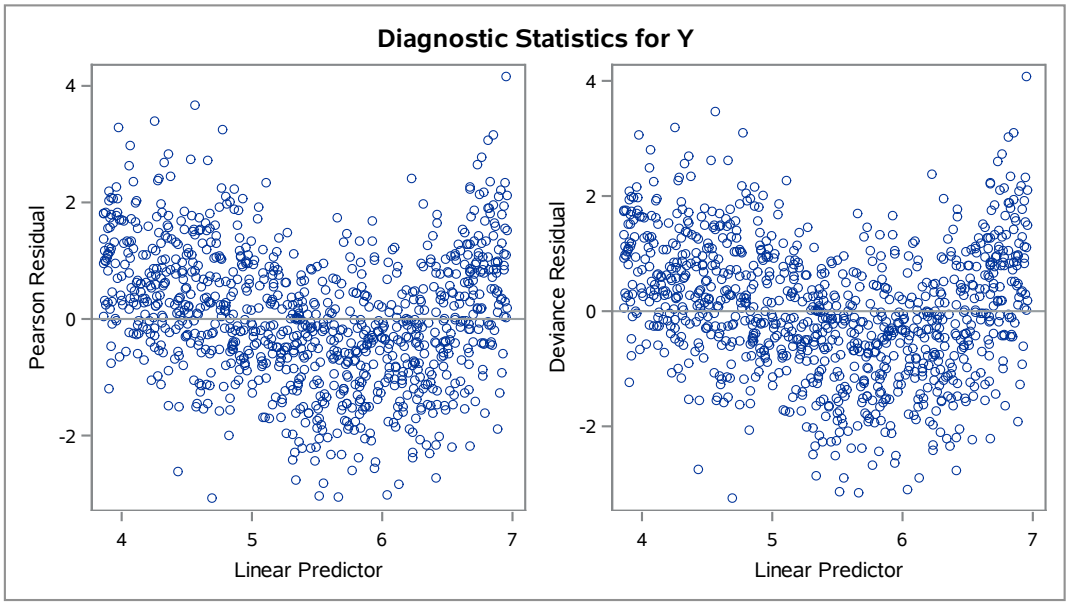
Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	998	1287.4319	1.2900
Scaled Deviance	998	1287.4319	1.2900
Pearson Chi-Square	998	1299.1010	1.3017
Scaled Pearson X2	998	1299.1010	1.3017
Log Likelihood		1708175.8791	
Full Log Likelihood		-4283.8901	
AIC (smaller is better)		8571.7802	
AICC (smaller is better)		8571.7922	
BIC (smaller is better)		8581.5957	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.3963	0.0205	-2.4365	-2.3561	13649.5	<.0001
X	1	3.1187	0.0075	3.1040	3.1333	174126	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

The GENMOD Procedure



Identity link

Sunday, Ja

The GENMOD Procedure

Model Information	
Data Set	WORK.A
Distribution	Poisson
Link Function	Identity
Dependent Variable	Y

Number of Observations Read	1000
Number of Observations Used	1000

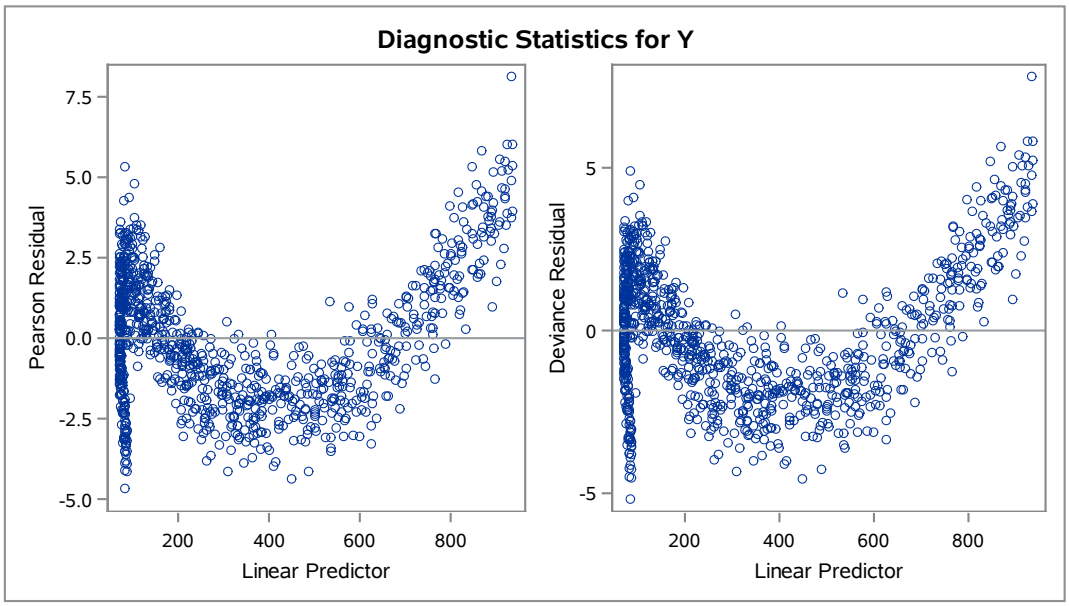
Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	997	4143.7880	4.1563
Scaled Deviance	997	4143.7880	4.1563
Pearson Chi-Square	997	4142.1266	4.1546
Scaled Pearson X2	997	4142.1266	4.1546
Log Likelihood		1706747.7011	
Full Log Likelihood		-5712.0681	
AIC (smaller is better)		11430.1363	
AICC (smaller is better)		11430.1603	
BIC (smaller is better)		11444.8595	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	5197.430	39.1953	5120.608	5274.251	17583.6	<.0001
X	1	-4823.30	32.4826	-4886.96	-4759.63	22049.0	<.0001
X2	1	1134.439	6.6743	1121.357	1147.520	28890.5	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

The GENMOD Procedure



Leverage

- In linear regression, the projection matrix (Hat matrix) is

$$H = X(X^T X)^{-1} X^T$$

  - h_{ii}, ith diagonal element of H, is called the leverage of the ith observation.

- In GLM, the projection matrix (from IRWLS)

$$H = V^{1/2} X(X^T V X)^{-1} X^T V^{1/2}$$

  - As the same as the linear regression, h_{ii} is leverage.

- The first order approximation of the variance of raw Pearson residual

$$Var(Y_i - \widehat{\mu}_i) \approx (1 - h_{ii}) Var(Y_i)$$

- Standardized Pearson residual

$$\widehat{r}_i^{PS} = \frac{\widehat{r}_i^{P}}{\sqrt{1 - h_{ii}}}$$

- Similarly, standardized Deviance residual

$$\widehat{r}_i^{DS} = \frac{\widehat{r}_i^{D}}{\sqrt{1 - h_{ii}}}$$

Influence measure

- In linear regression, there are a number of diagnostic measures for the influence of one observation based on leave it out, refitting the model, and checking the changes.
  - DFBETA

$$DFBETA_i \approx \widehat{\beta} - \widehat{\beta}_{-i}$$

  - Cook's distance

$$\begin{aligned} D_i &= \frac{1}{q\widehat{\sigma}^2}(\widehat{\beta} - \widehat{\beta}_{-i})^T X^T X(\widehat{\beta} - \widehat{\beta}_{-i}) \\ &= \frac{1}{q}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_i^2 \end{aligned} \tag{1}$$

- In the linear regression, these statistics can be calculated without refitting the model n times. Explicit shortcut is available based on H.
- In GLM, the exact solution for the explicit shortcut is not available. But the one-step approximation method has been developed to avoid to fitting n times.

- One-step approximation:
  - Cook's distance:

$$D_i = \frac{1}{q}\left(\frac{h_{ii}}{1-h_{ii}}\right)(r_i^{PS})^2$$

  - One-step approximation for DFBETA is also available.

Examples

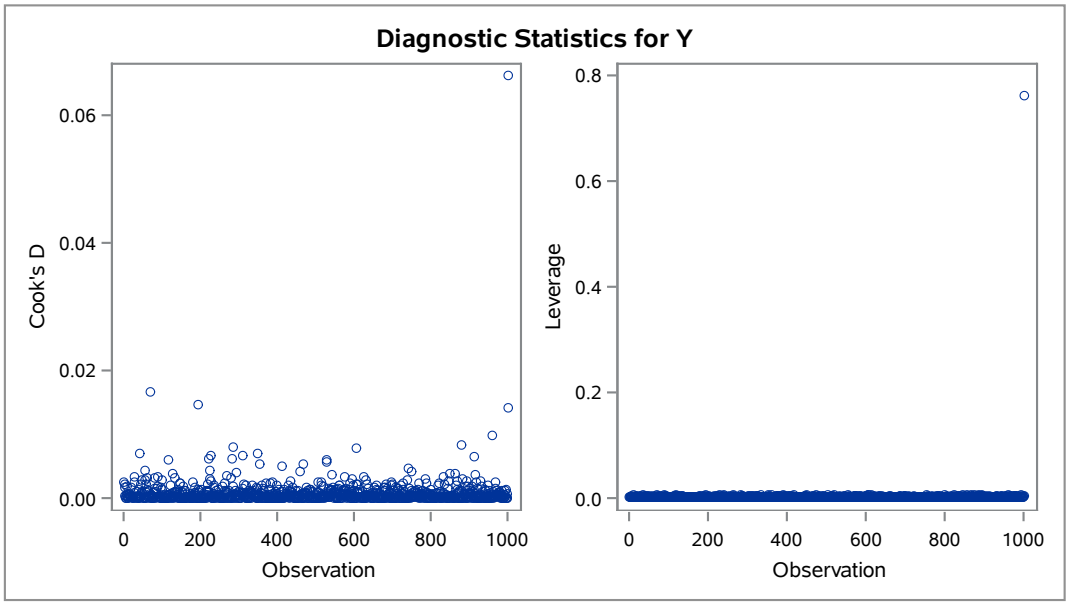
- Previous example:

$$log(\lambda_i) = 1 + 0.5x + 0.5x^2, \quad 2 < x < 3$$

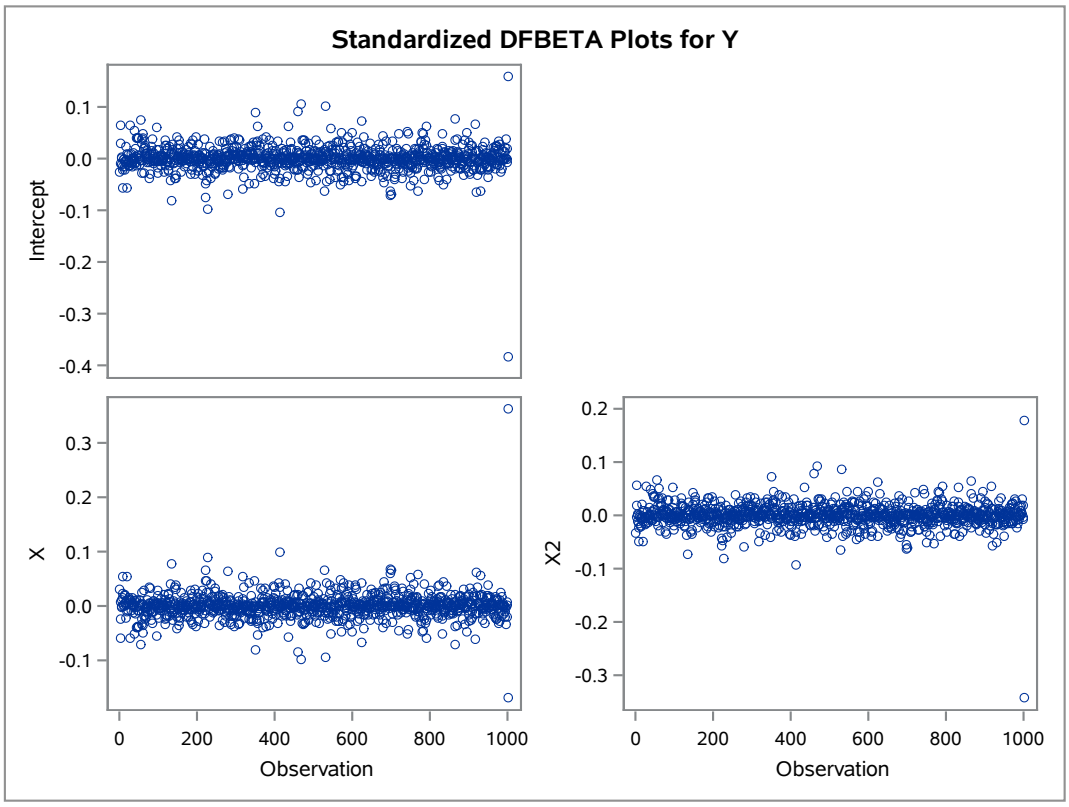
$$Y_i \sim Poisson(\lambda_i)$$

- Add two outliers (Obervation 1001 and 1002)
  - Obs 1001: X=2, Y=0
  - Obs 1002: X=3.5, Y from the true model

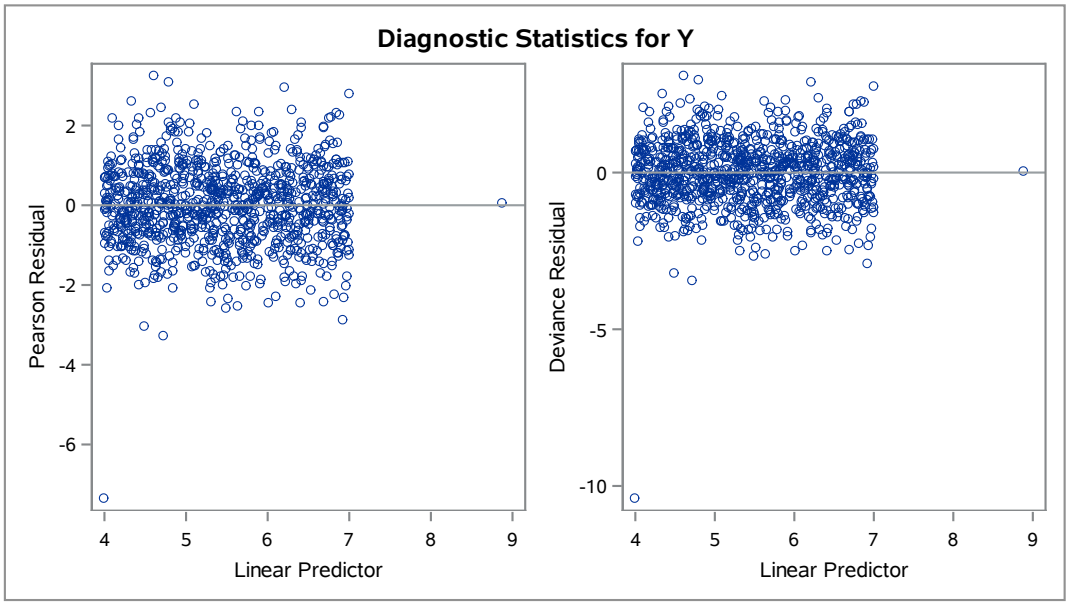
The GENMOD Procedure



The GENMOD Procedure



The GENMOD Procedure



- Obs 1001: X=2, Y=0
  - Leverage: 0.0037
  - Cook's distance: 0.066
- Obs 1002: X=3.5, Y from the true model
  - Leverage: 0.76
  - Cook's distance: 0.014

Multicollinearity

- Explanatory variable (X) are highly correlated with one another.
- Can cause several undesirable consequences.
  - $\widehat{\beta}$ will be very unstable.
  - Variances of some $\widehat{\beta}$ can be very large.
- Variance inflation factor

$$VIF_j = \frac{1}{1 - R^2_{(j)}}$$

  - $R^2_{(j)}$: R^2 obtained from regressing the jth variable against all other variables.
  - $VIF = 1$: Not correlated
  - $1 < VIF < 5$: moderately correlated
  - $VIF >$5 to 10: highly correlated

- In linear regression, we are concerning about the collinearity in the predictors (X)
- In GLM, we are concerning about the collinearity in the weighted predictor ($V^{1/2}X$)
- SAS proc genmod does not provide VIF, so you have to calculate it using proc reg with the weight statement.