STAT 120 C

Introduction to Probability and Statistics III

Dustin Pluta
2019/05/14

Linear Regression Diagnostics

Assumptions of the linear regression model:

1. Errors are normally distributed.
2. Errors are independent.
3. Errors have constant variance.

- Diagnostic plots help us determine
  - if any of these assumptions are violated,
  - if the model adequately captures the relationship of X and Y,
  - if there are outliers in the data.

Linear Regression Diagnostics

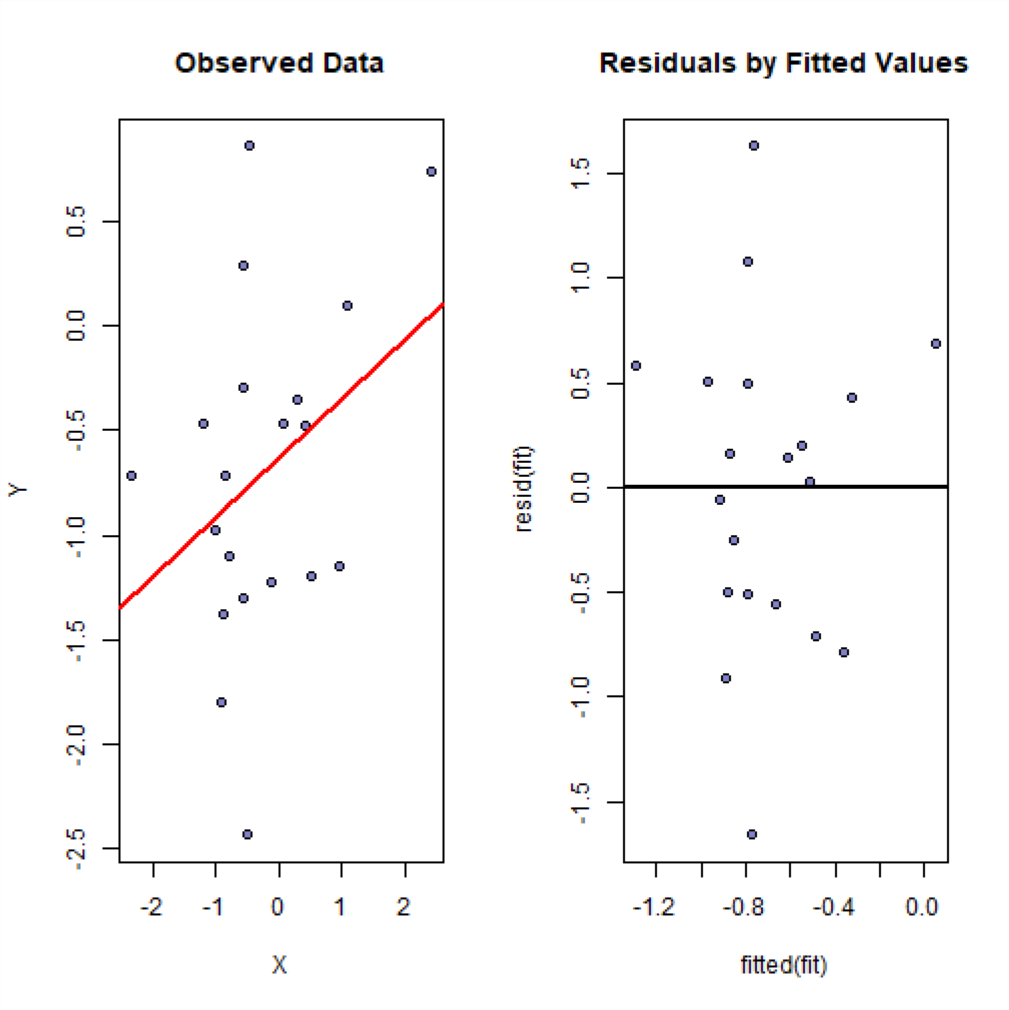
Consider $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$, where X_i is some observed (fixed) covariate for each $i = 1, \ldots, n$.

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the MLEs for the regression coefficients.

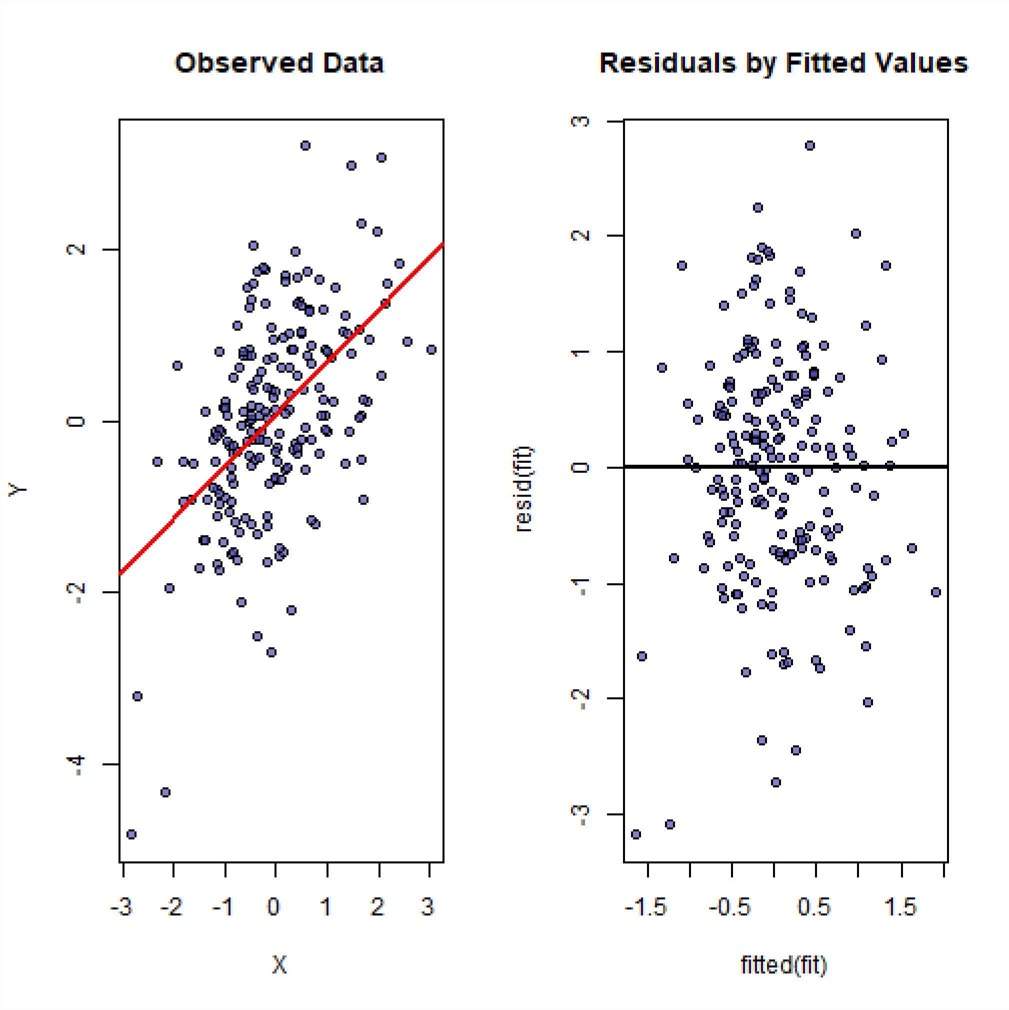
Denote the fitted values as $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, and the residuals as $\hat{e}_i = \hat{Y}_i - Y_i$.

- To check for violation of the constant variance assumption, we examine the *residuals vs fitted plot* (or the *residuals vs X plot*)

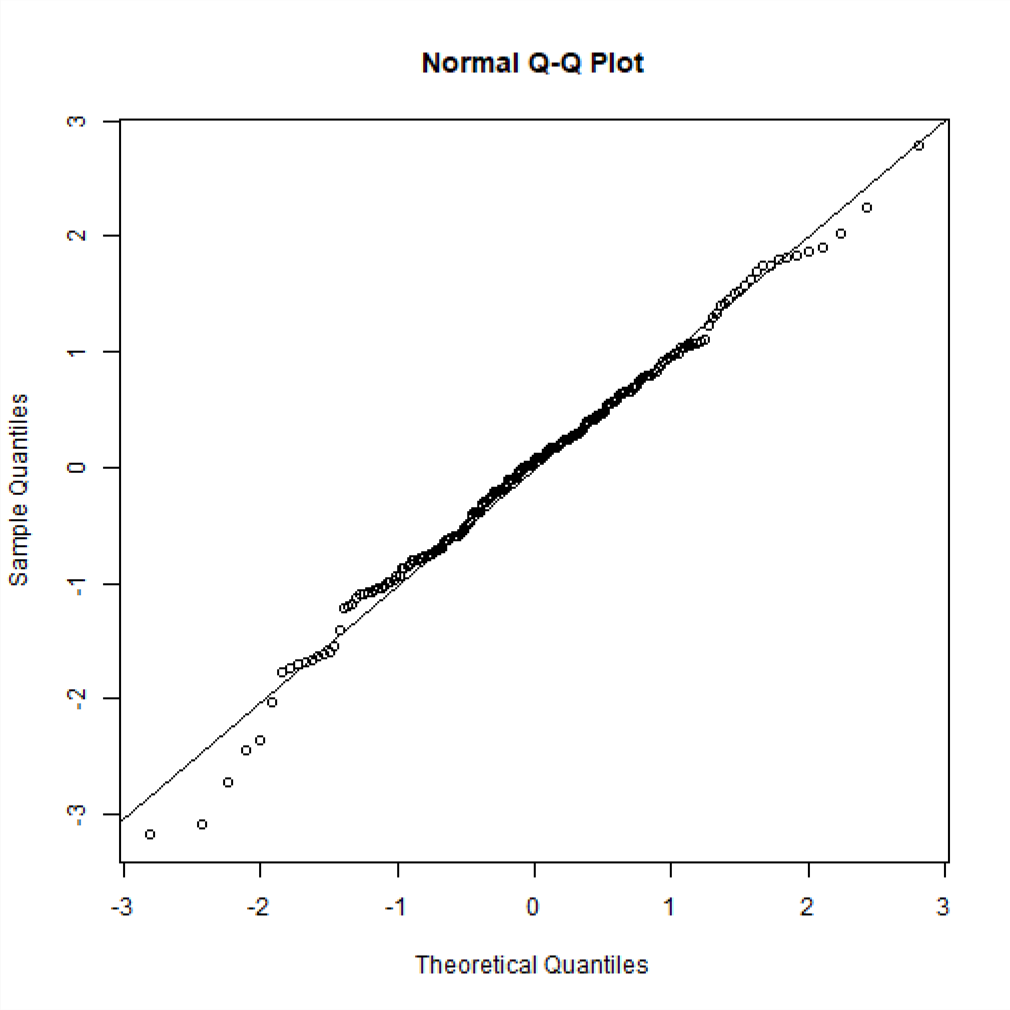
Linear Regression Diagnostics



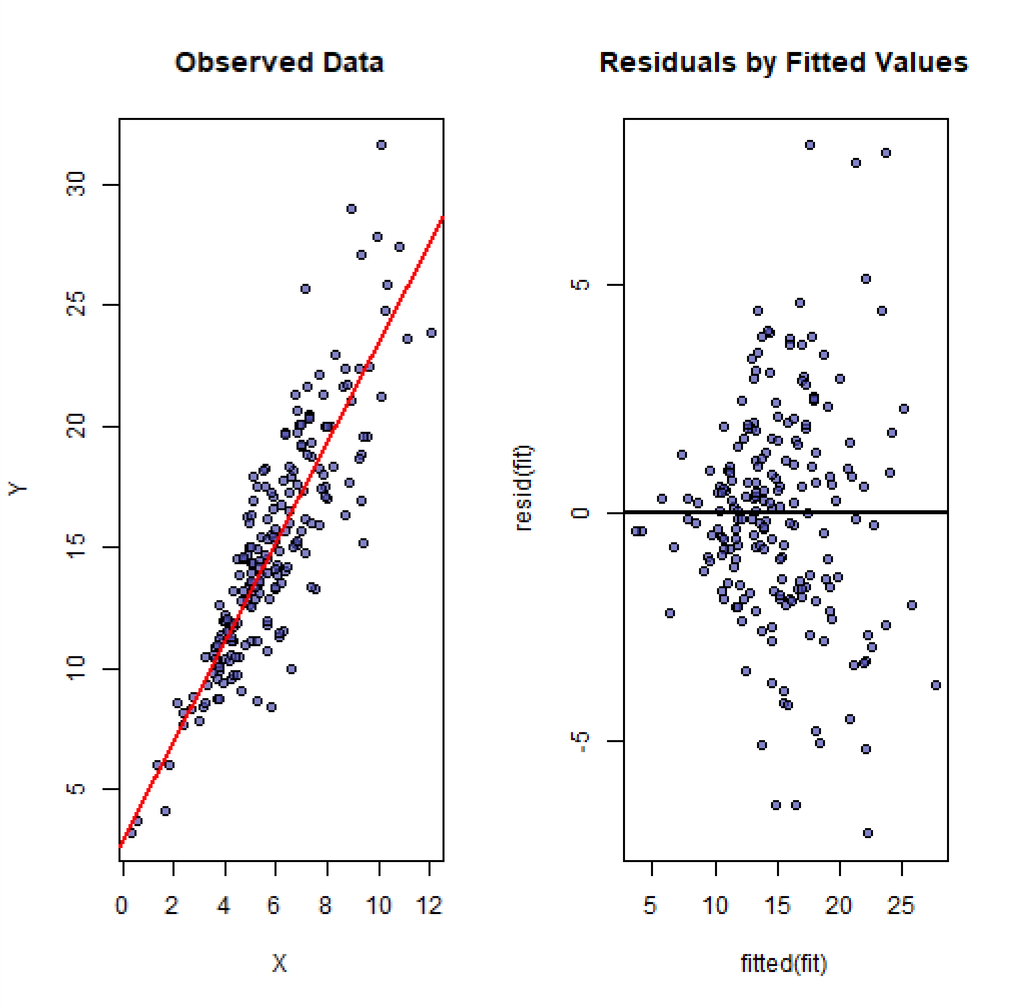
Linear Regression Diagnostics



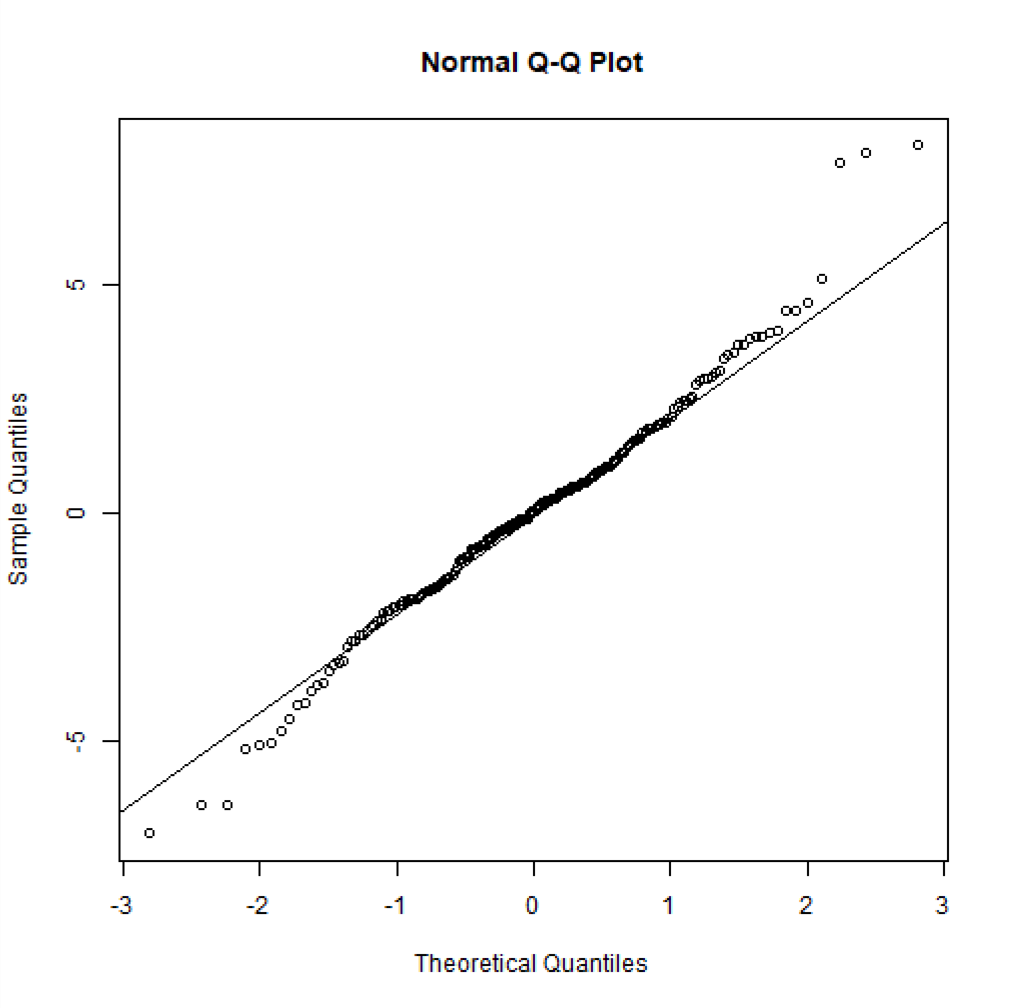
Linear Regression Diagnostics



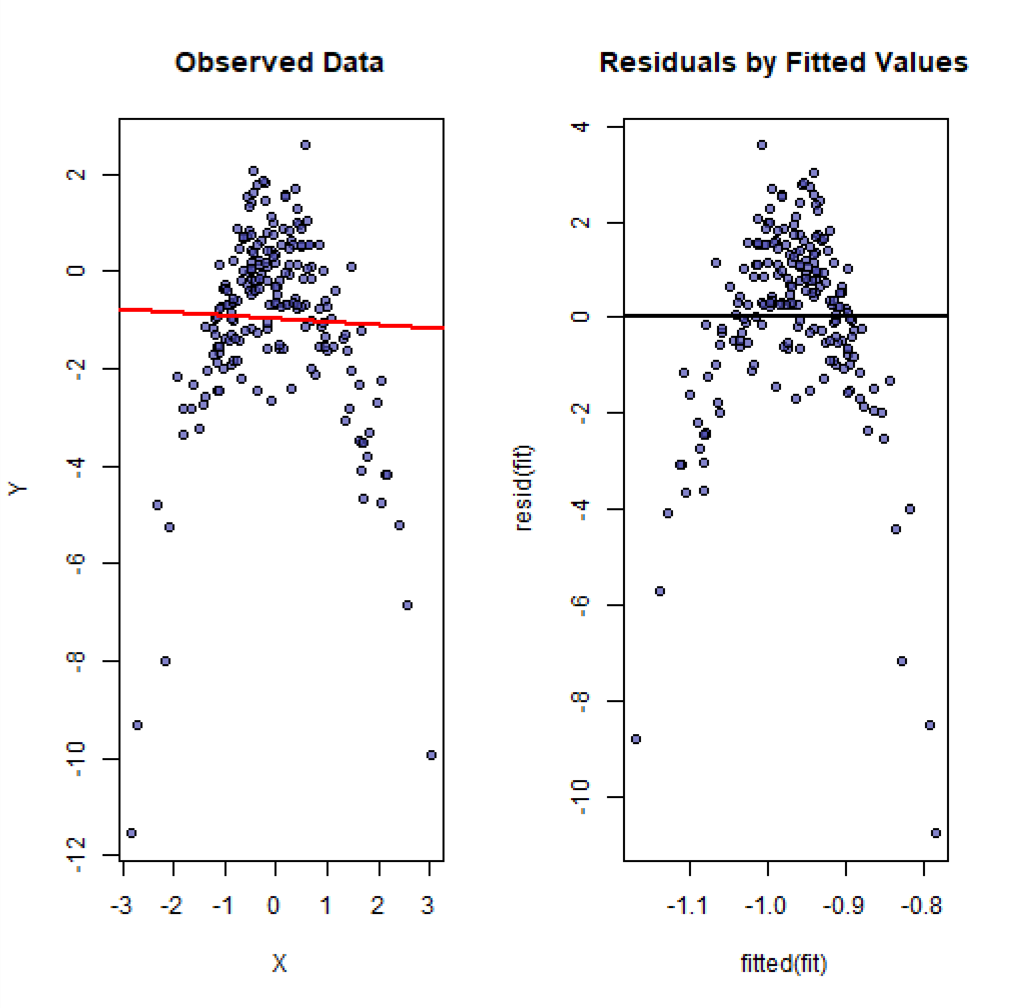
Linear Regression Diagnostics



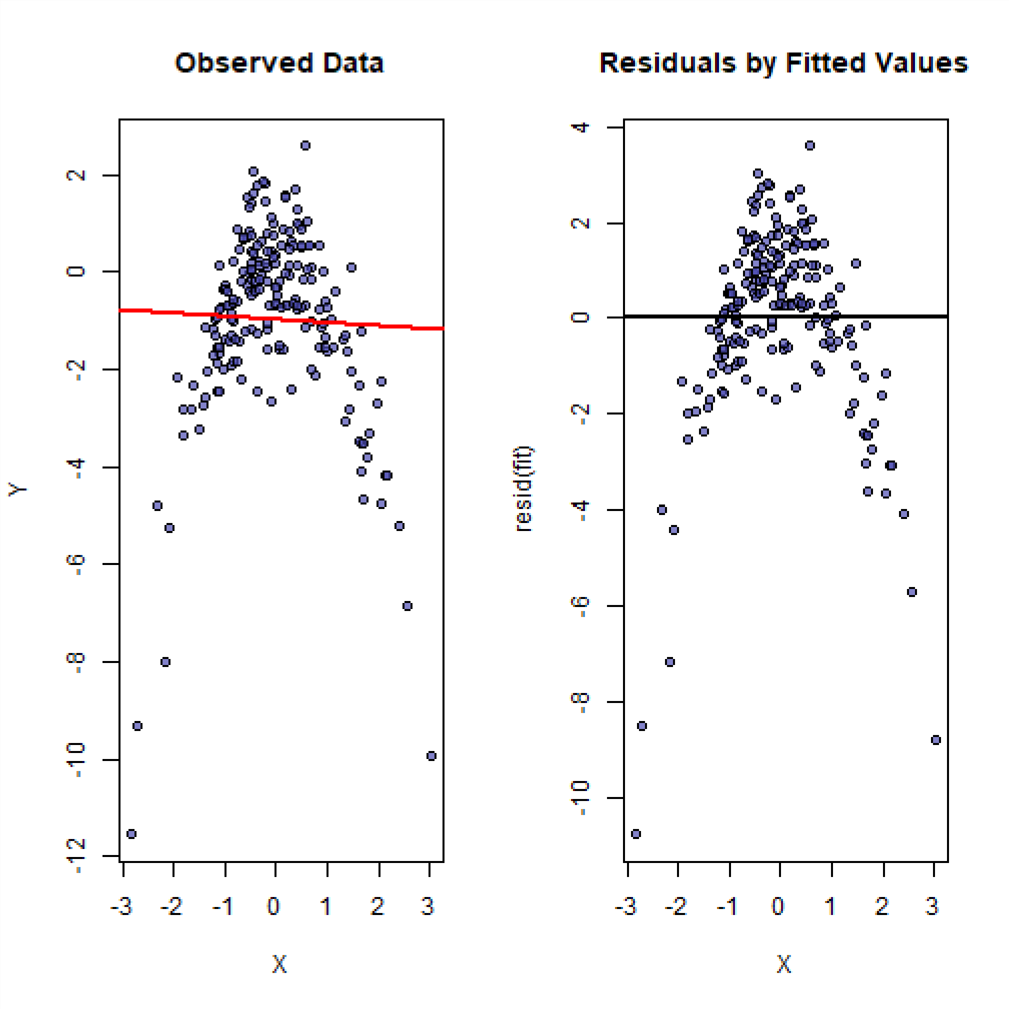
Linear Regression Diagnostics



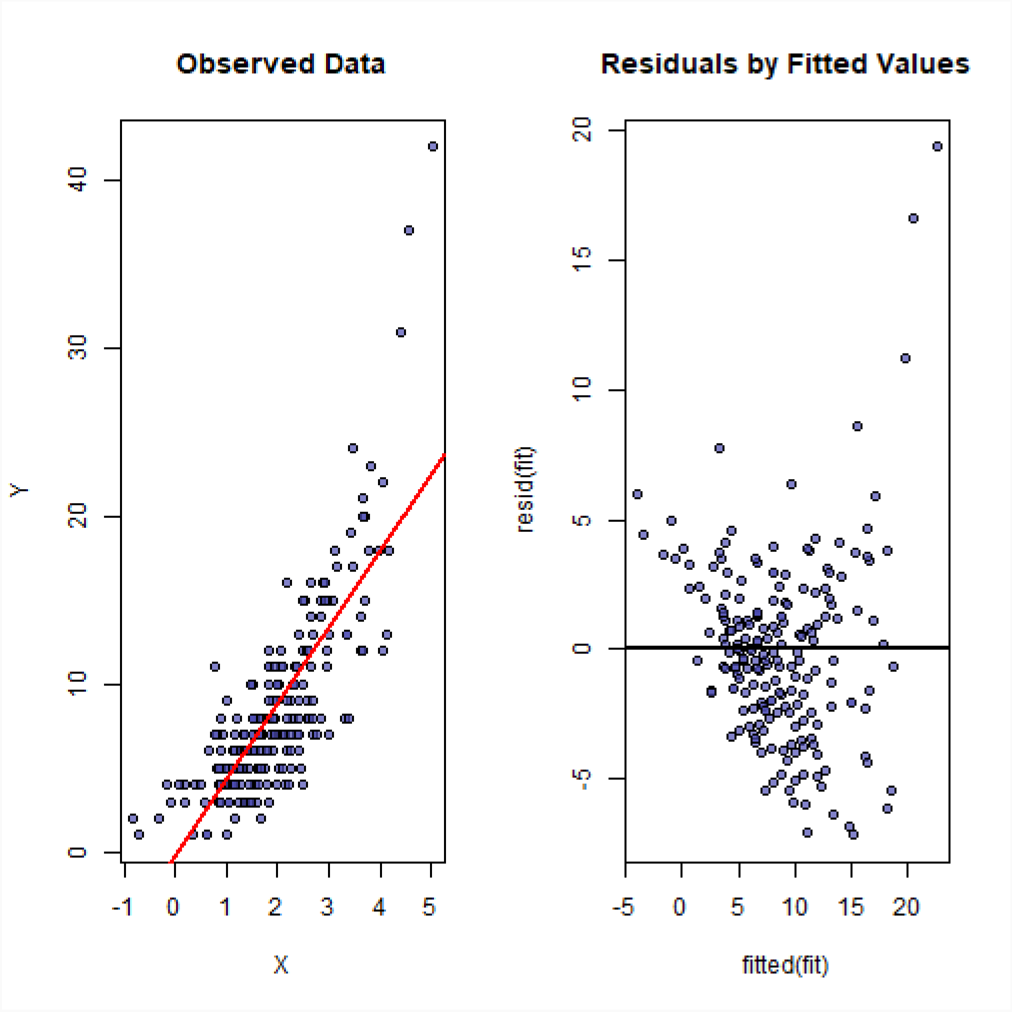
Linear Regression Diagnostics



Linear Regression Diagnostics



Linear Regression Diagnostics



Variance Stabilizing Transformation

In some cases, if the response follows a non-normal distribution with a mean-variance relationship, a variance stabilizing transformation can be applied.

Suppose $\mathbf{Var}(Y_i|X_i) = \sigma(\mathbb{E}[Y_i|X_i])$ for some function $\sigma(\cdot)$.

By the Central Limit Theorem

$$\sqrt{n}(\bar{Y} - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2(\mu)).$$

Variance Stabilizing Transformation

We want to transform the data by some function g so that the resulting variance doesn't depend on μ. By the Delta Method

$$\sqrt{n}[g(\bar{Y}) - g(\mu)] \xrightarrow{D} \mathcal{N}(0, [g'(\mu)]^2\sigma^2(\mu)).$$

Consequently, we can find a variance stabilizing g by the following:

$$\begin{aligned}
(g'(\mu))^2\sigma^2(\mu) &= 1 \\
g'(\mu) &= \frac{1}{\sigma(\mu)} \\
g(\mu) &= \int \frac{1}{\sigma(\mu)}\, d\mu
\end{aligned}$$

Variance Stabilizing Transformation

Poisson Example

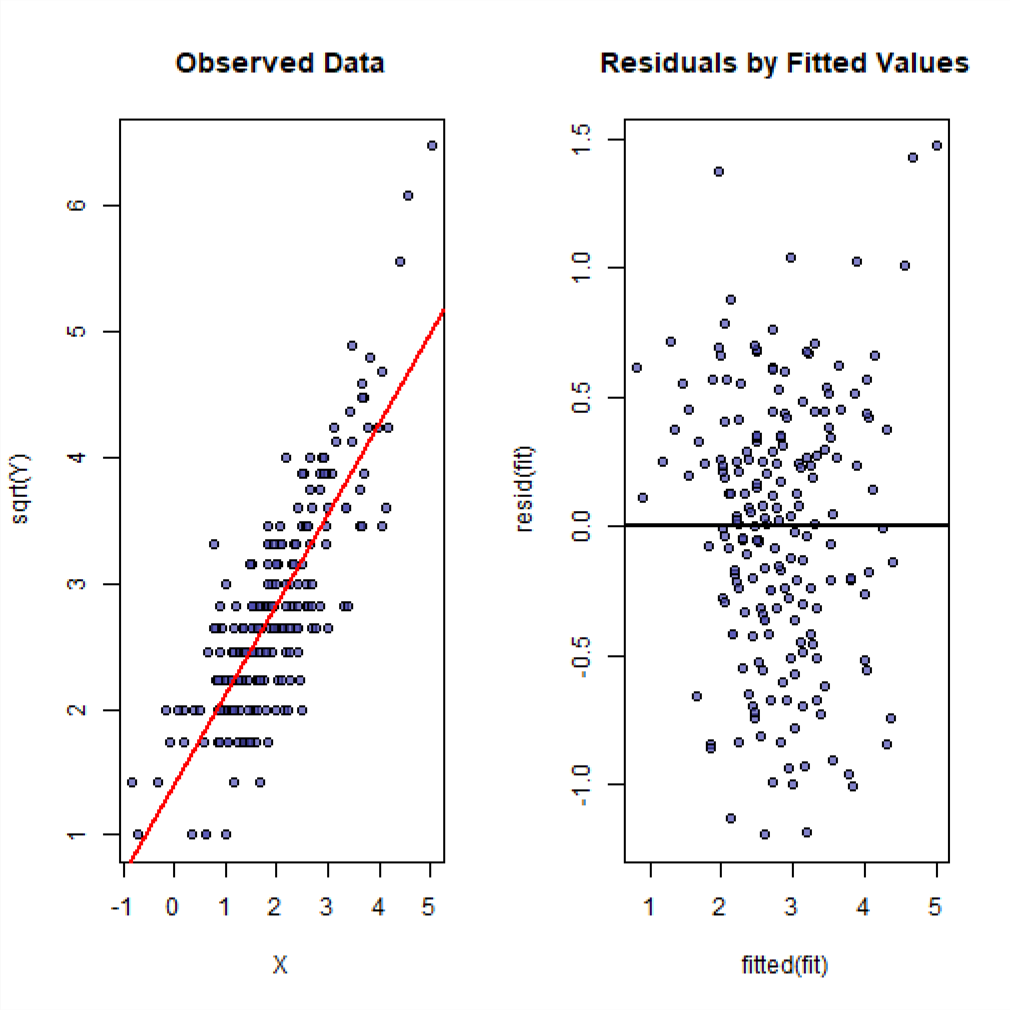
If $Y \sim Pois$, then $\sigma^2(\mu) = \mu$. Plugging this into the formula for g, we get

$$g(\mu) = \int \frac{1}{\sqrt{\mu}}\, d\mu = 2\sqrt{\mu}$$

If we transform the data by the square root, the resulting asymptotic distribution is

$$\sqrt{n}(\sqrt{\bar{Y}} - \sqrt{\mu}) \xrightarrow{D} \mathcal{N}(0, 1/4)$$

Variance Stabilizing Transformation



Confidence and Prediction Intervals

- The distribution of the fitted values $\hat{Y}_i$ is

$$\hat{\mu}(X_i) \equiv \hat{Y}_i \sim \mathcal{N}\left(\beta_0 + \beta_1 X_i, \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]\right)$$

- From this, we can calculate a $(1-\alpha)$ 100% **Confidence Interval** for the mean regression line at covariate value X is

$$\hat{\mu}(X) \pm t_{1-\alpha/2}(n-2)\sqrt{s^2\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]}$$

- A $(1-\alpha)$ 100% **Prediction Interval** for a new observation at covariate value X_{new} is

$$\hat{Y}_{new} \pm t_{1-\alpha/2}(n-2)\sqrt{s^2\left[1 + \frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]}$$

- See the code *LinearRegressionExample_Part2.R* for a comparison of the confidence and prediction intervals.