$$e_i = Y_i - \widehat{Y}_i$$

Analysis of Residuals

Data = Fit + Residual

$$Y_i = \mathbf{x}_i' \widehat{\boldsymbol{\beta}} + e_i$$

Residual means left over

- Vertical distance of Y_i from the regression hyper-plane
- An error of “prediction”
- Big residuals merit further investigation
- Big compared to what?
- They are normally distributed (HW)
- Consider standardizing
- Maybe detect outliers

Standardized Residuals

- Could divide by square root of sample variance of $e_1, \ldots, e_n$
- "Semi-Studentized" (Kutner et al.)

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

- Studentized: Estimate $Var(e_i)$ and divide by square root of that

Studentized residuals

$$\begin{aligned} \mathbf{e} &= \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}, \text{ where} \\ \mathbf{H} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{aligned}$$

$$\begin{aligned} E(\mathbf{e}) &= \mathbf{0} \\ V(\mathbf{e}) &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned}$$

$$e_i^* = \frac{e_i}{\sqrt{MSE\,(1 - h_{i,i})}}$$

Studentized deleted residuals

- An outlier will make MSE big
- So Studentized residual will be too small – less noticeable
- So calculate Y-hat for each observation based on all the other observations, but not that one
- Basically, predict each observed Y based on all the others, and assess error of prediction (divide by standard error).

Deleted residual

$$d_i = Y_i - \widehat{Y}_{i(i)} = \frac{e_i}{1 - h_{i,i}}$$

$$s^2\{d_i\} = \frac{MSE_{(i)}}{1 - h_{i,i}}$$

Studentized deleted residual is
$t_i = \frac{d_i}{s\{d_i\}} \sim t(n - p - 1)$

Is it too big? Use a t-test.

Prediction interval

- Apply the same technology
- Think of Studentized deleted residual for case n+1
- So $$t_{n+1} = \frac{d_{n+1}}{s\{d_{n+1}\}} \sim t(n-p)$$

$$\begin{aligned} 1-\alpha &= Pr\left\{-t_{\alpha/2}(n-p) < \frac{Y_{n+1} - \mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}}}{s\{d_{n+1}\}} < t_{\alpha/2}(n-p)\right\} \\ &= Pr\left\{-t_{\alpha/2}\, s\{d_{n+1}\} < Y_{n+1} - \mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}} < t_{\alpha/2}\, s\{d_{n+1}\}\right\} \\ &= Pr\left\{\mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}} - t_{\alpha/2}\, s\{d_{n+1}\} < Y_{n+1} < \mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}} + t_{\alpha/2}\, s\{d_{n+1}\}\right\} \end{aligned}$$

Plotting residuals

- Against variables not in the equation
- Against variables in the equation
- Normal Q-Q plot to check approximate normality