

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

Analysis of Residuals

Data = Fit + Residual

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

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, \dots, e_n
- “Semi-Studentized” (Kutner et al.)

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

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

Studentized residuals

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}, \text{ where}$$
$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$E(\mathbf{e}) = \mathbf{0}$$

$$V(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

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

Studentized deleted residuals

- An outlier will make MSE big
- So Studentized residual will be too small – less noticeable
- So calculate \hat{Y} 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

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

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

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.

Plotting residuals

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