Analysis of Residuals[1]

STA442/2101 Fall 2016

[1] See last slide for copyright information.

Residual means left over: $e_i = Y_i - \widehat{Y}_i$

- 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
- Consider standardizing
- Maybe detect outliers
- Plots can also be informative.

Residuals are like estimated error terms

$$e_i = Y_i - \widehat{Y}_i \quad \Leftrightarrow \quad Y_i = \widehat{Y}_i + e_i$$

$$\begin{aligned} Y_i &= \widehat{Y}_i + e_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \cdots + \widehat{\beta}_{p-1} x_{i,p-1} + e_i \\ &= \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \end{aligned}$$

Normal distribution of ϵ_i implies normal distribution of e_i, but the e_i are not independent, and they do not have equal variance.

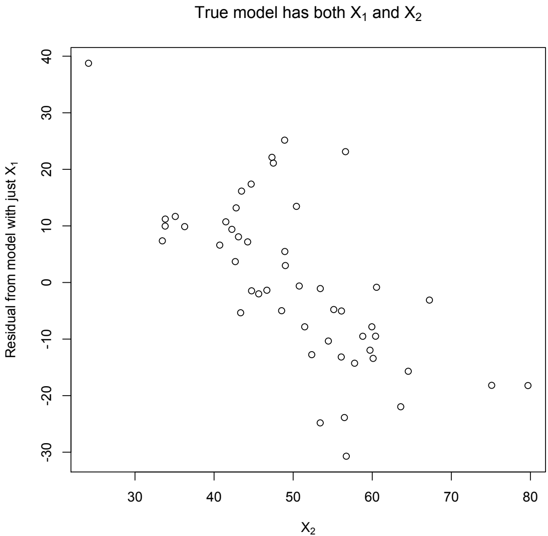
Data = Fit + Residual

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

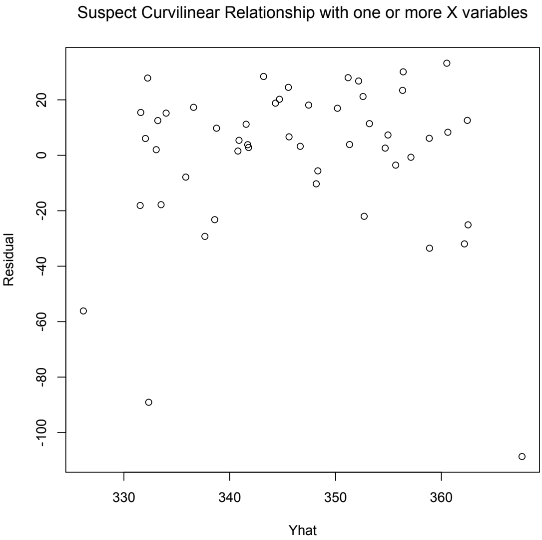
Plotting residuals can be helpful

- Against predicted Y.
- Against explanatory variables not in the equation.
- Against explanatory variables in the equation.
- Look for serious departures from normality.

Plot Residuals Against Explanatory Variables Not in the Equation

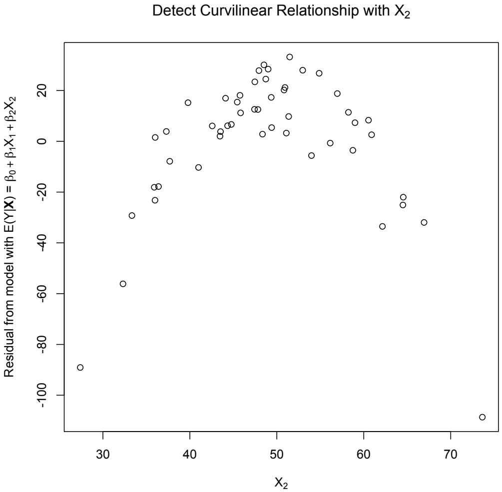


Plot Residuals Against $\widehat{Y}$



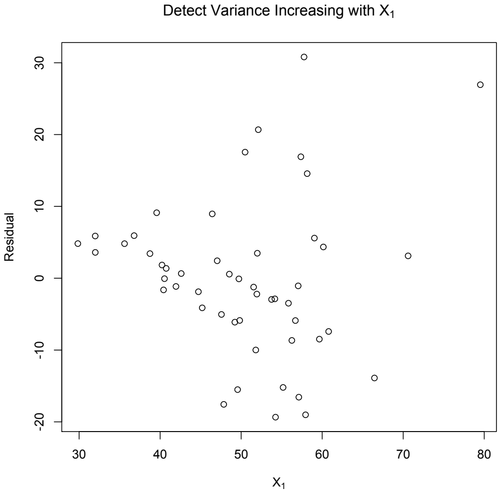
Plot Residuals Against Explanatory Variables in the Equation

Plot versus X_1 showed nothing



Plot Residuals Against Explanatory Variables in the Equation

Can show non-constant variance



Outlier detection

- Big residuals may be outliers. What's "big?"
- Consider standardizing.
- Could divide by square root of sample variance of $e_1, \ldots, e_n$.
- Semi-Studentized: Estimate $Var(e_i)$ and divide by square root of that. $e_i^* = \frac{e_i}{\sqrt{MSE\,(1-h_{i,i})}}$
- In R, this is produced with `rstandard`.

Studentized deleted residuals

The idea

- An outlier will make MSE big.
- In that case, the standardized (Semi-Studentized) residual will be too small – less noticeable.
- So calculate $\widehat{Y}$ for each observation based on all the other observations, but not that one.
- Predict each observed Y based on all the others, and assess error of prediction (divided by standard error).

Apply prediction interval technology

$$T = \frac{Y_{n+1} - \mathbf{x}_{n+1}^\top \widehat{\boldsymbol{\beta}}}{\sqrt{MSE(1 + \mathbf{x}_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{n+1})}} \sim t(n-p)$$

- Note that Y_i is now being called Y_{n+1}.
- If the "prediction" is too far off there is trouble.
- Use T as a test statistic.
- Need to change the notation.

Studentized deleted residual

$$T_i = \frac{Y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{(i)}}{\sqrt{MSE_{(i)}(1 + \mathbf{x}_i^\top (\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1} \mathbf{x}_i)}} \sim t(n - 1 - p)$$

- In R, this is produced with `rstudent`.
- There is a more efficient formula.
- Use T_i as a test statistic of $H_0 : E(Y_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$.
- If H_0 is rejected, investigate.
- We are doing n tests.
- Type I errors are very time consuming and disturbing.
- If independent, probability of no false positives would be $(1 - \alpha)^n \to 0$.
- But they are not independent.
- How about a Bonferroni correction?

Copyright Information