Section for Applied Statistics and Data Analysis

TA: Cong Mu

Office Hour: Wednesday 10:00AM - 12:00PM

October 25, 2019

Overview

1. Some Statistics
   - Finding Unusual Observations
     - Leverage
     - Outliers
     - Influential Observations
2. Some Programming
   - Examples and Exercises in Faraway

Diagnostics - Overview

- Recall
$$\epsilon \sim \mathcal{N}\left(0, \sigma^2 \mathrm{I}\right).$$
- Checking Error Assumptions
  - Constant Variance
  - Normality
  - Correlated Errors
- Finding Unusual Observations
  - Leverage
  - Outliers
  - Influential Observations
- Checking the Structure of the Model

Leverage

- **Leverage point**: potential to influence the fit
- Recall

$$\widehat{y} = X\widehat{\beta} = X(X^\top X)^{-1}X^\top y = Hy \qquad \text{where} \qquad H = X(X^\top X)^{-1}X^\top.$$

$$\widehat{\epsilon} = y - \widehat{y} = y - Hy = (I - H)y = (I - H)(X\beta + \epsilon) = (I - H)\epsilon.$$

$$Var\left[\widehat{\epsilon}\right] = Var\left[(I - H)\epsilon\right] = (I - H)Var\left[\epsilon\right](I - H)^\top = (I - H)\sigma^2.$$

- **Leverages**: $h_i = H_{ii}$ (`hatvalues` in R)

$$\sum_{i=1}^{n} h_i = \sum_{i=1}^{n} H_{ii} = p.$$

- **Rough rule**: check leverages of more than $\frac{2p}{n}$

Leverage

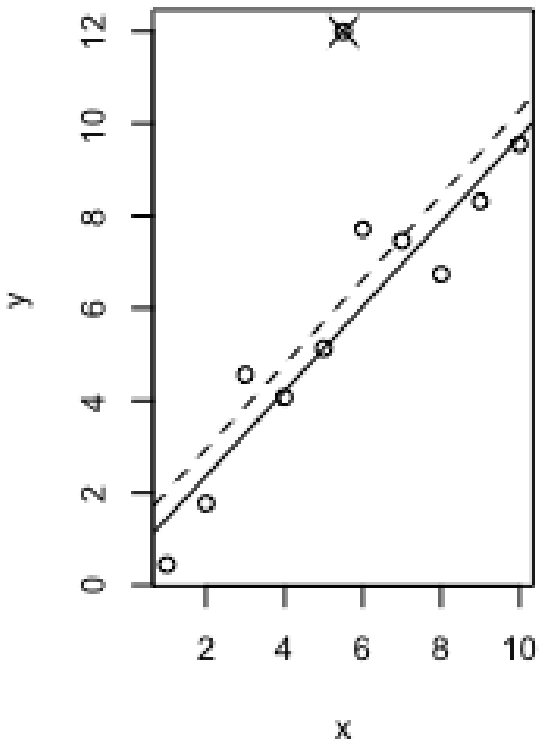
- **Half-normal plots** (`halfnorm` in R)
  - Sort the data: $x_{[1]} \leqslant \cdots \leqslant x_{[n]}$
  - Compute $u_i = \Phi^{-1}\left(\frac{n+i}{2n+1}\right)$
  - Plot $x_{[i]}$ against u_i
- **Standardized residuals** (`rstandard` in R)

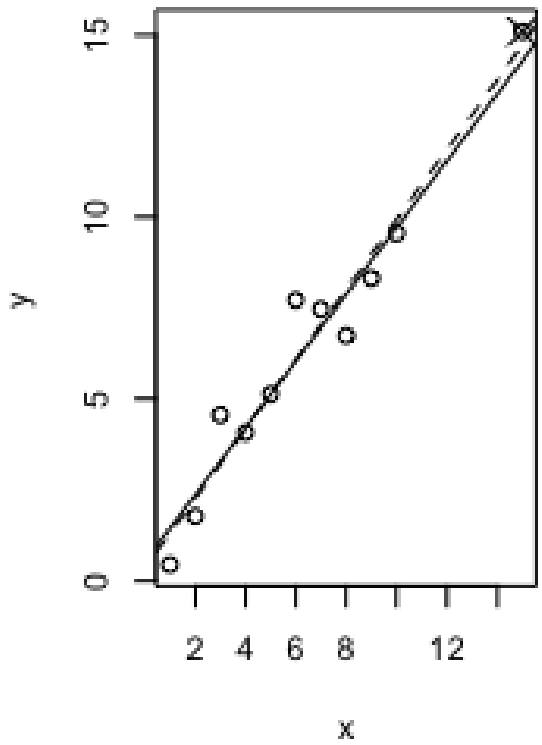
$$Var\left[\widehat{\epsilon_i}\right] = \sigma^2(1 - H_{ii}) = \sigma^2(1 - h_i).$$

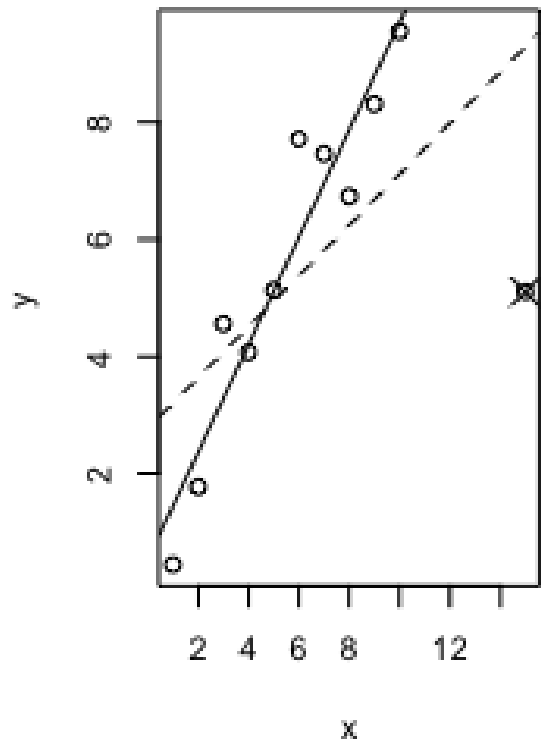
$$r_i = \frac{\widehat{\epsilon_i}}{\widehat{\sigma}\sqrt{1 - h_i}}.$$

Outliers

- **Outliers**: not fit the current model well
- Outliers may or may not affect the fit substantially







Outliers

- **Studentized residuals** (`rstudent` in R)

$$t_i = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2} \sim t_{n-p-1}.$$

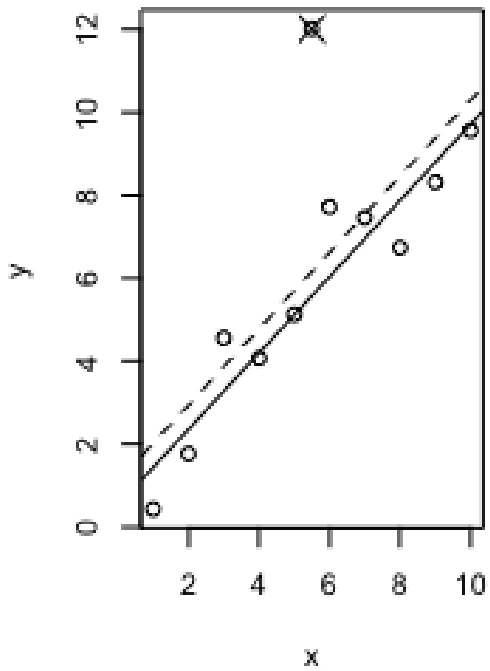
- **Bonferroni correction**: if an overall level α test is required, then a level α/n should be used in each of the tests

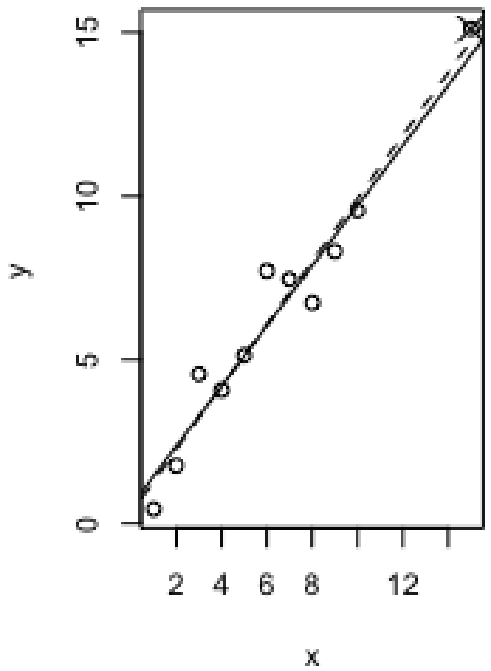
Outliers

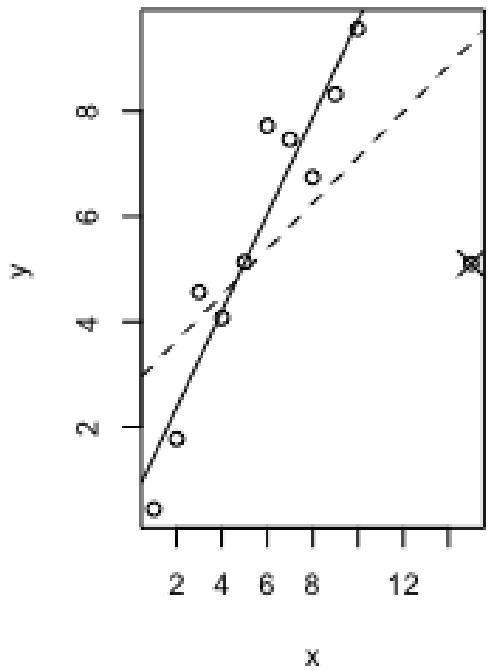
- Some notes about outliers
  - Two or more outliers next to each other can hide each other
  - An outlier in one model may not be an outlier in another
  - Error distribution may not be normal so may have larger residuals
  - Individual outliers are much less of a problem in larger datasets
- Some tips about outliers
  - Check for data-entry error first
  - Examine the physical context
  - Exclude the point from analysis but re-include it if model is changed

Influential Observations

- **Influential point**: removal from the dataset would cause a large change in the fit
- An influential point may or may not be an outlier and may or may not have large leverage but it will tend to have at least one of these two properties







Influential Observations

- **Cook statistics** (`cooks.distance` in R)

$$D_i = \frac{1}{p} r_i^2 \frac{h_i}{1 - h_i}.$$

- A half-normal plot of D_i can be used to identify influential observations

Examples and Exercises in Faraway Chapter 6

- **Example:** `savings` dataset
- **Example:** `star` dataset
- **Exercise:** `sat` dataset

Thanks for listening!