Section for Applied Statistics and Data Analysis

TA: Cong Mu

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

November 15, 2019

Overview

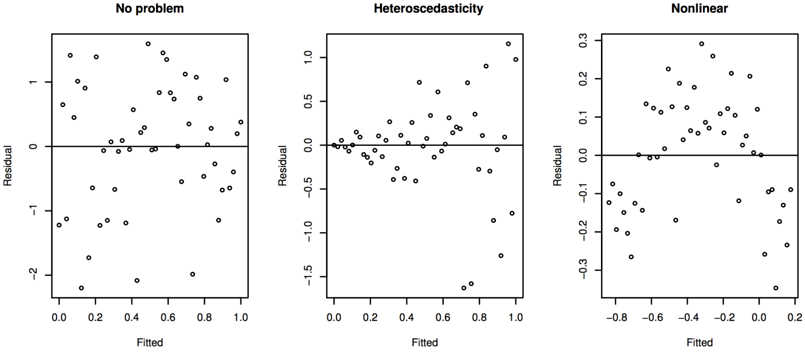
1. Some Statistics
   - Diagnostics
2. Some Programming
   - Problems in Homework

Diagnostics

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

Constant Variance

- **Residual Plot**



(Figure from Linear Models with R)

- **Brown-Forsythe Test**

Transformation

- **Box-Cox Transformation** (boxcox in R)

$$t_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

Normality

- **Q-Q Plot**
- **Tests**
  - Shapiro-Wilk test
  - Anderson-Darling test
  - etc
- **Some R Packages**
  - `nortest`
  - `normtest`
  - etc

Correlated Errors

- **Plot Successive Pairs of Residuals**
- **Tests**
  - Durbin-Watson test
  - etc
- **Some R Packages**
  - `lmtest`
  - etc

Leverage

- **Leverage point**: potential to influence the fit
- **Leverages**: $h_i = H_{ii}$ (`hatvalues` in R) where $H = X(X^\top X)^{-1}X^\top$

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

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

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_i}}.$$

Outliers

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

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
- **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

Checking the Structure of the Model

- **Partial regression or added variable plots**
  - Regress y on all x except x_i and get residuals $\widehat{\delta}$
  - Regress x_i on all x except x_i and get residuals $\widehat{\gamma}$
  - Plot $\widehat{\delta}$ against $\widehat{\gamma}$
- **Partial residual plots** (`termplot` in R)
  - Plot $x_i\widehat{\beta}_i + \widehat{\epsilon}$ against x_i
- Partial residual plots are believed to be better for non-linearity detection while added variable plots are better for outlier/influential detection.

Problems in Homework

- **Example:** `pipeline` dataset
- **Example:** `hills` dataset

Thanks for listening!