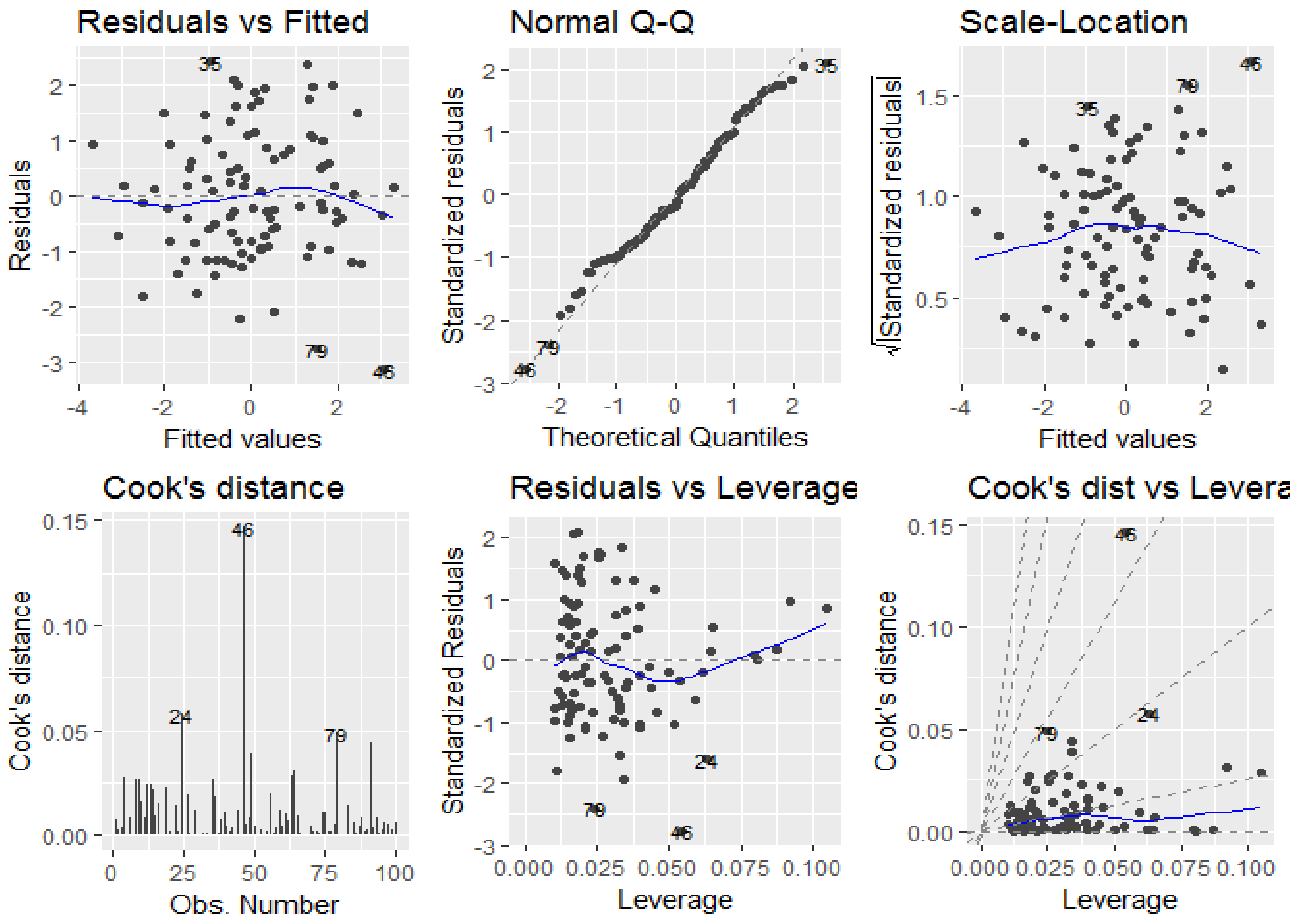
Lecture 6: Residual Analysis

Instructor: Prof. Shuai Huang

Industrial and Systems Engineering

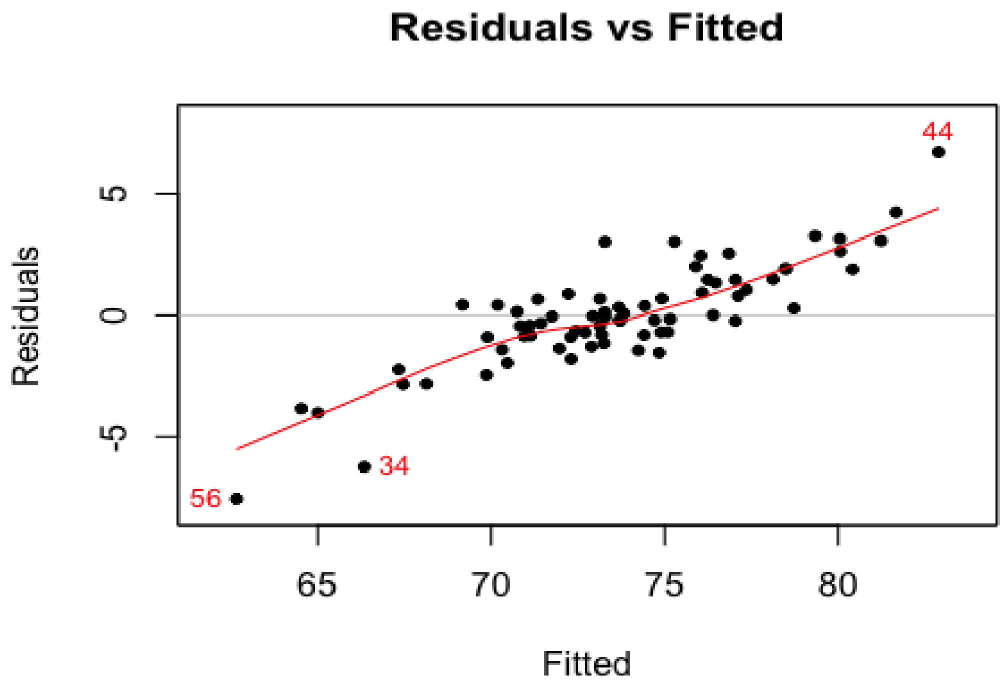
University of Washington

Residual Analysis (a.k.a. Model Diagnostics)



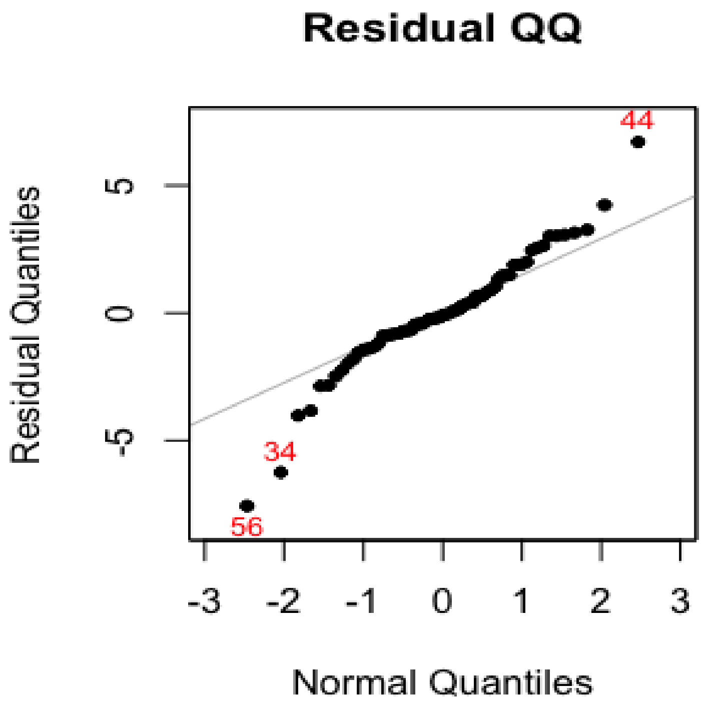
Residual versus fitted values

- The residuals, by definition, form the “unsystematic” part of the data, that suppose to be noise and random (any nonrandom behavior raises a red flag)



Q-Q Plot

- Q-Q plot is to validate that the residuals follow a certain distribution (e.g., a normal distribution)



Cook’s distance

- The Cook’s distance shows the influential data points that have larger than average influence on the parameter estimation.
- The Cook’s distance of a data point is built on the idea of how much change will be induced on the estimated parameters if the data point is deleted.

Leverage

- Mathematically, the leverage of a data point is $\frac{\partial \hat{y}_i}{\partial y_i}$, reflecting how sensitive the prediction on the data point by the model is decided by the observed outcome value y_i.
- For data points that are surrounded by many close-by data points, their leverages won't be large.
- Thus, we could infer that the data points that sparsely occupy their neighbor areas will have large leverages.
- These data points could either be outliers that severely derivate from the linear trend represented by the majority of the data points, or could be valuable data points that align with the linear trend but lack neighbor data points.

Multicollinearity analysis

- Suppose the data is generated by this model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon,\ \ \varepsilon \sim N(0, \sigma_\varepsilon^2),$$

$$x_1 = 2x_2 + \epsilon,\ \ \epsilon \sim N(0,0.1\sigma_\varepsilon^2)$$

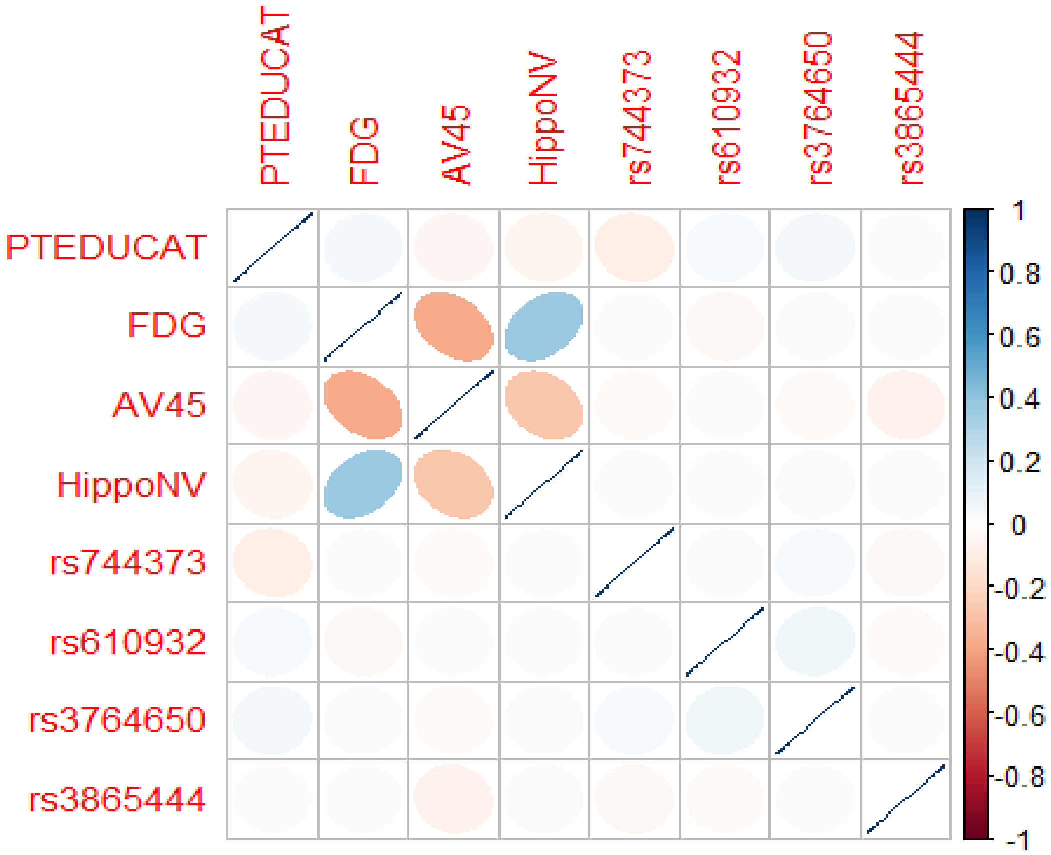
- Theoretically, we could value the regression model that is shown in above as the ground truth model equally as we value the following models:

$$y = \beta_0 + (2\beta_1 + \beta_2)x_2 + \beta_3 x_3 \ldots + \beta_p x_p,$$

$$y = \beta_0 + (\beta_1 + 0.5\beta_2)x_1 + \beta_3 x_3 + \cdots + \beta_p x_p,$$

$$y = \beta_0 + 1000x_1 + (\beta_2 + \beta_1 - 2000)x_2 + \beta_3 x_3 + \cdots + \beta_p x_p.$$

Correplot Package



Remarks

- Important to understand that, residual analysis is “opportunistic” checking of the model
- Like patient checks in hospital for screening or examination. Negative results don’t mean that the patient is healthy
- It is a significant focus on regression models, but less developed in machine learning community

R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets