Data Analytics in Business

Outliers vs. Influential Points

Sridhar Narasimhan
Professor
Scheller College of Business

Outliers and Influential Points

Outliers and Influential Points

- **Objective**: In fitting a model to a given body of data, we would like to ensure that the fit is not overly determined by one or a few observations, aka **outliers**.
- There are two types of outliers:
    1. **Y (response) outlier**
    2. **X (predictor) outlier, Leverage Point**
- An outlier has the potential to be identified as an **influential** data point if it unduly influences the regression analysis.
- The next few slides will define the two types of outliers and how to classify the outlier as an influential point.

Reference: Chatterjee S, Hadi AS (2012) *Regression Analysis by Example* 5 edition. (Wiley, Hoboken, New Jersey).

Outliers in the Response Y Variable

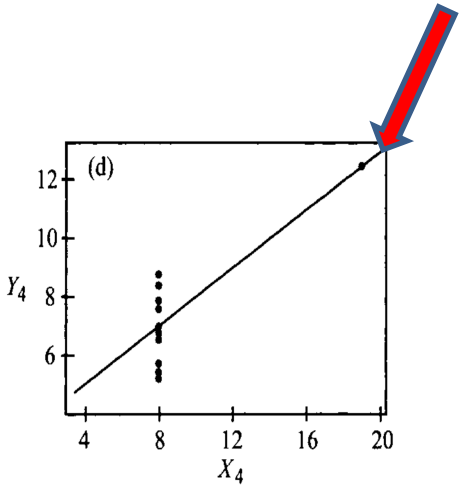
- An outlier is a point that has a y_i value that is far from its predicted value $\hat{y}_i$. It will have a large standardized residual.
    - Since the standardized residuals are approximately normally distributed with mean = 0 and a standard deviation = 1, points with standardized residuals larger than 2 or 3 standard deviation away from the mean (zero) are called outliers.
    - If removal of the outlier causes substantial change to the regression analysis, then it is an influential outlier (see influential point definition).
- **Detection**: One way to visualize/identify outliers is to plot residuals (or standardized residuals) against predicted values of y ($\hat{y}_i$)
- Why do outliers occur?
    - Outliers could occur because of incorrect data recording, because of real anomalous events recorded correctly or because the phenomenon could very well be non-linear.
    - Do not assume that an outlier observation should automatically be removed. It may signal a model deficiency (i.e. a missing predictor)

Outliers in the Predictor X Variable (Leverage Points)

- Extreme *x* values (*x* is the predictor variable) are high leverage points
  - The data point x_i will be unusually out of range of the other predictor X values
  - Does not have a large standardized residual
  - Can affect regression results
- **Detection**: Identify leverage points via index plot, dot plots, box plot, or Cook's Distance (next slide)
  - If the leverage point is flagged via Cook's distance, then it is also an influential point and thus has substantial influence to the fitted model (next slide)
- Why does the leverage point exist?
  - Requires a case-by-case data analysis
  - Often, it is best to analyze the leverage point by creating models with and without the data point to see the effect it has on the fitted line
  - This method of analysis applies to both *y* (response) outliers and influential points

Influential Points

- An outlier is an *influential point* if its deletion (by itself or with 2 or 3 other points) causes substantial changes in the fitted model (estimated coefficients, fitted values, slope, t-tests, hypothesis tests, etc.)
    - Deletion must cause large changes, thus the data point has undue influence
    - See plot below of (X,Y) least squares fitted line with an influential point.



Detection:

- With several variables, we cannot detect influential points graphically.
- Measure influential points via **Cook's Distance (C_i):** The difference between
    1. the regression coefficients obtained from the full data WITH observation point 'i'
    2. and the regression coefficients obtained by DELETING the 'ith' data point
        - Rule of thumb is to identify points with $C_i > 1$ as highly influential

GTx