Yeoun Chan Kim

April 2020

Detecting unusual obeservations

**Abstract**

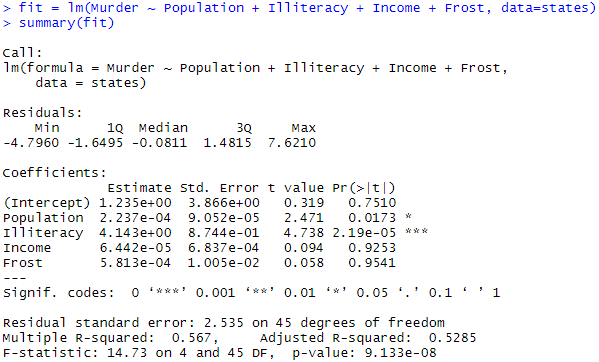
Leverage, outlier, and influential observations are considered unusual observations. In this project, methods to detect unusual data values are explained. Also, impact of those values to a model is described by looking at a dataset.

**Introduction**

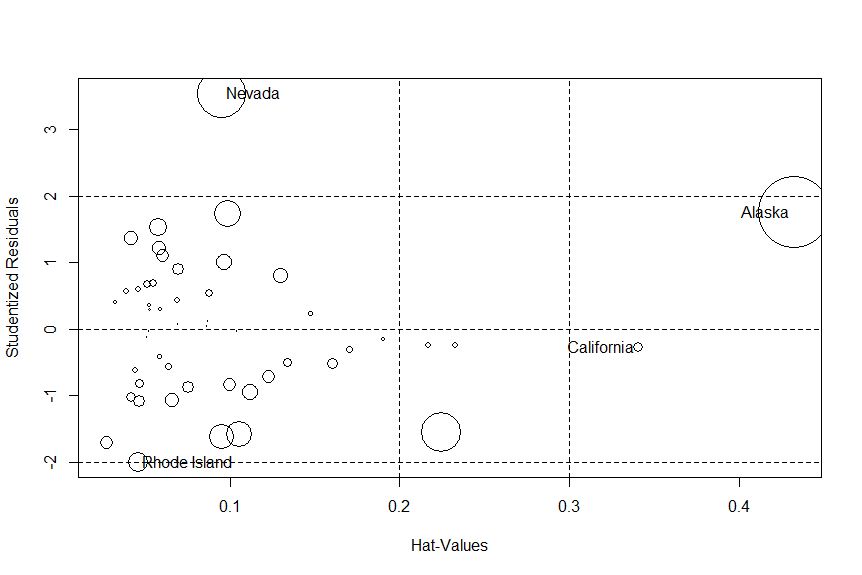
In class, we have learned various ways to analyze data using R. Also, we have done assignments on given data. Given examples are tested to provide outcome that is expected for a specific purpose of analysis. However, real data often does not provide outcome that one expects. Few outliers can totally change a result of analysis, especially in regression. Simple linear regression is considered to be appropriate when following conditions are satisfied: linear relationship, homoscedasticity, and independent errors. Addition to these conditions, outliers and normally distributed residuals are important issues to consider. First step to make appropriate regression is to determine usual and unusual data values. Frequently used ways to determine unusual data values are looking for outlier, high leverage points, and influential observation. Outlier is an observation that lies abnormal distance from other values and we use standard deviation to distinguish it. To distinguish high leverage points, mean of observable estimates is used. Influential observation is determined using Cook distance or D statistics. This project focuses on how these three ways are used in determining unusual data values to make appropriate regression using state dataset.

**Materials and Methods**

The state dataset has 50 rows of states and 8 attributes. 8 attributes are population estimate, per capita income, illiteracy, life expectancy in years, murder rate per 100,000 population, percent high-school graduates, mean number of days with temperature below freezing, and land area in square miles. For convenience, five attributes (population, illiteracy, murder rate, income, frost) are extracted and saved in a new data frame states. Without determining unusual data values, regression model is made on murder by other four attributes.



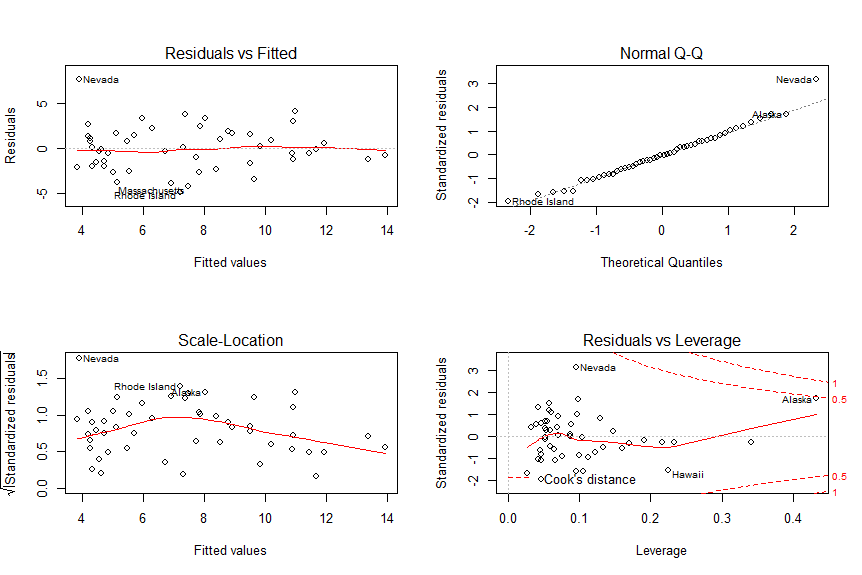
In order to find unusual data values in this regression, influencePlot from car library is used and the plot is



Y axes represents standard deviation and data value lies above +2 standard deviation or below -2 standard deviation is considered as outlier. In this case Nevada and Rhode Island are outliers. Even though Rhode Island is on the border of -2 standard deviation, it is important to check the data. X axes represents mean of observable estimates(hat) which is calculated by where p represents sum of the leverage values and n is number of observations. In our case, mean of hat-values is and data value exceeds two or three multiplied by the mean is considered to be an outlier. This plot used and California and Alaska exceed this value. The diameter of each circle represents Cook’s distance and if Cook’s distance is greater than , where n is number of observations and k is number of predictors, is influential. Hair, Anderson, Tatham and Black (1998) suggested Cook’s distance greater than 1 and Bollen and Jackman (1990) suggested Cook’s distance greater than to be influential. These three thresholds are generally used. It is impossible to distinguish by looking at diameter of each circle, and it is recommended to look at data values with large circles. This plot suggests that Alaska, California, Nevada, and Rhode Island can be possible unusual data values.

**Results**

After finding possible unusual data values, plot function is used to determine if there were issue from data collection.



The plot of the regression model represents that this model satisfies linear relationship, homoscedasticity, and independent errors. In other words, those four possible unusual data values had no issues on data collection and the data values of the outliers will be kept.

**Discussion**

Detecting unusual data values and taking care of them are important in various statistical analysis. As real data contains tremendous amount of information, it is nearly impossible to detect unusual data values by looking row by row. The methods above provide efficient and accurate way to detect unusual data values. If plot function suggests that outliers had issues on data collection, trimming, winsorizing, or other methods can be considered for appropriate statistical analysis of a model.

**Appendix**

Watson, Peter. “Checking for Outliers in Regression.” imaging.mrc\_cbu.cam.ac.uk,

MRC Cognition and Brain Sciences Unit, 6 May, 2015.

“Assumptions of Simple Linear Regression.” restore.ac.uk,

University of Southampton, 21 Jul, 2011.