Linear Regression

11/14/2021 12:46

# Abstract

Outcome variable: *year*  
Major predictor: *lifeExp*

A linear regression model is built to analyze the relationship between the outcome variable and the predictor(s).

### Linear Regression

Linear regression is a linear approach to modeling the relationship between a continuous scalar outcome (dependent) variable and one or more predict (independent) variables.

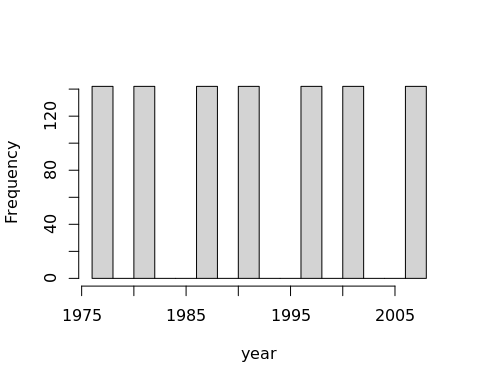
# Descriptive Statistics

Table 1 gives the basic information of the analyzing data set. Observations with missing values are removed when calculating. It shows that there is no missing value in the data.

Completeness of Data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total | Incomplete | Used |
| No. Observations | 994 | 0 | 994 |

For the outcome (dependent) variable *year*, the histogram is shown in Figure 1.



Histogram of year.

A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable.

For the numerical predictor *lifeExp*, some descriptive statistics including sample mean and sample standard deviation are shown in Table 2.

*Summary Statistics of Numerical Predictor Variable(s).*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | S.D. |
| **lifeExp** | 23.6 | 54.41 | 67.01 | 63.74 | 73.42 | 82.6 | 11.61 |

# Results

The linear regression model can be expressed as

where s are the parameters and is the error.

## Model Fitting

The estimated regression coefficients are shown in Table 3 along with their standard error, t-statistic, and *p*-value.

Regression Coefficients Estimation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | Stat. | p-value |  |
| (Intercept) | 1980.000 | 1.7400 | 1140.00 | 0 | \*\*\* |
| lifeExp | 0.172 | 0.0268 | 6.42 | 0 | \*\*\* |

The fitted model is

Test statistics are used to test whether the regression coefficients are significantly different from 0. That is whether the explanatory variables have significant effects on the response variable. If the *p*-value is smaller than a given threshold, e.g. 0.05, the corresponding regression coefficient is statistically significantly different from 0 under significance level 0.05, that is the predictor variable plays an important role in explaining the variation of the response. Otherwise, there is no sufficient evidence to conclude the regression coefficient departs from 0, i.e. the predictor has no significant effect on the response variable.

For example, the *p*-value of *lifeExp* is 2.12e-10. It means, *lifeExp* is statistically significant in this model and it has a significant effect on *year*.

## Model Interpretation

The interpretation for the regression coefficient of continuous predictor is controlling the effects from other predictors (if there exist), the change of the response variable given per unit increase of the predictor.

For example, the regression coefficient of *lifeExp* is 0.172. It means if *lifeExp* increases by one unit, *year* will increase by 0.172 unit(s). In addition, since its *p*-value is less than 0.05, we conclude the effect is significant under significance level 0.05.

## Goodness-of-Fit

Table 4 shows some important goodness-of-fit index of the model, including , adjusted , RMSE. The result of the F-test is also given.

Goodness-of-Fit Statistics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | R-Squared | Adj. R-Squared | RMSE | F(1,992) | p-value |
|  | 0.0399 | 0.0389 | 9.81 | 41.2 | 0 |

The *R*-squared of this model is 0.0399. It indicates that about 3.99 percent of turbulence of *year* is explained by the explanatory variables. If *R*-squared is small, some important predictor variables are probably missed and should be added to the model.

However, if irrelevant predictor variables are added to the model, *R*-squared always increases. To adjust model overfitting, adjusted *R*-squared is used when comparing models with different numbers of the explanatory variables. The adjusted *R*-squared of this model is 0.0389.

The F-test is employed to test the overall significance of the model. The F-ratio F(1,992) is 41.2. Since its *p*-value is less than 0.05, the entire set of variables taken collectively explain a significant part of the variation in the dependent variable. The model overall is reasonable.

## Assumption Checking

In this section, we will check the assumptions of the model. Four distinct assumptions are checked in this report (Peña and Slate 2006; Chatterjee and Hadi 2015):

1. ***Linearity***: The model that relates the response to the predictors is assumed to be linear in the regression parameters.
2. ***Normality***: The errors are assumed to have a normal distribution with a mean of 0.
3. ***Homoscedasticity***: The error terms are assumed to have the same variance.
4. ***Independent-errors assumption***: The errors are assumed to be independent of each other.

Assumption Checking Results.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test Stat. | p-value | Decision |
| Global Stat | 87.5000 | 0.000 | Assumptions NOT satisfied! |
| Skewness | 0.1440 | 0.705 | Assumptions acceptable. |
| Kurtosis | 54.8000 | 0.000 | Assumptions NOT satisfied! |
| Link Function | 32.6000 | 0.000 | Assumptions NOT satisfied! |
| Heteroscedasticity | 0.0102 | 0.920 | Assumptions acceptable. |

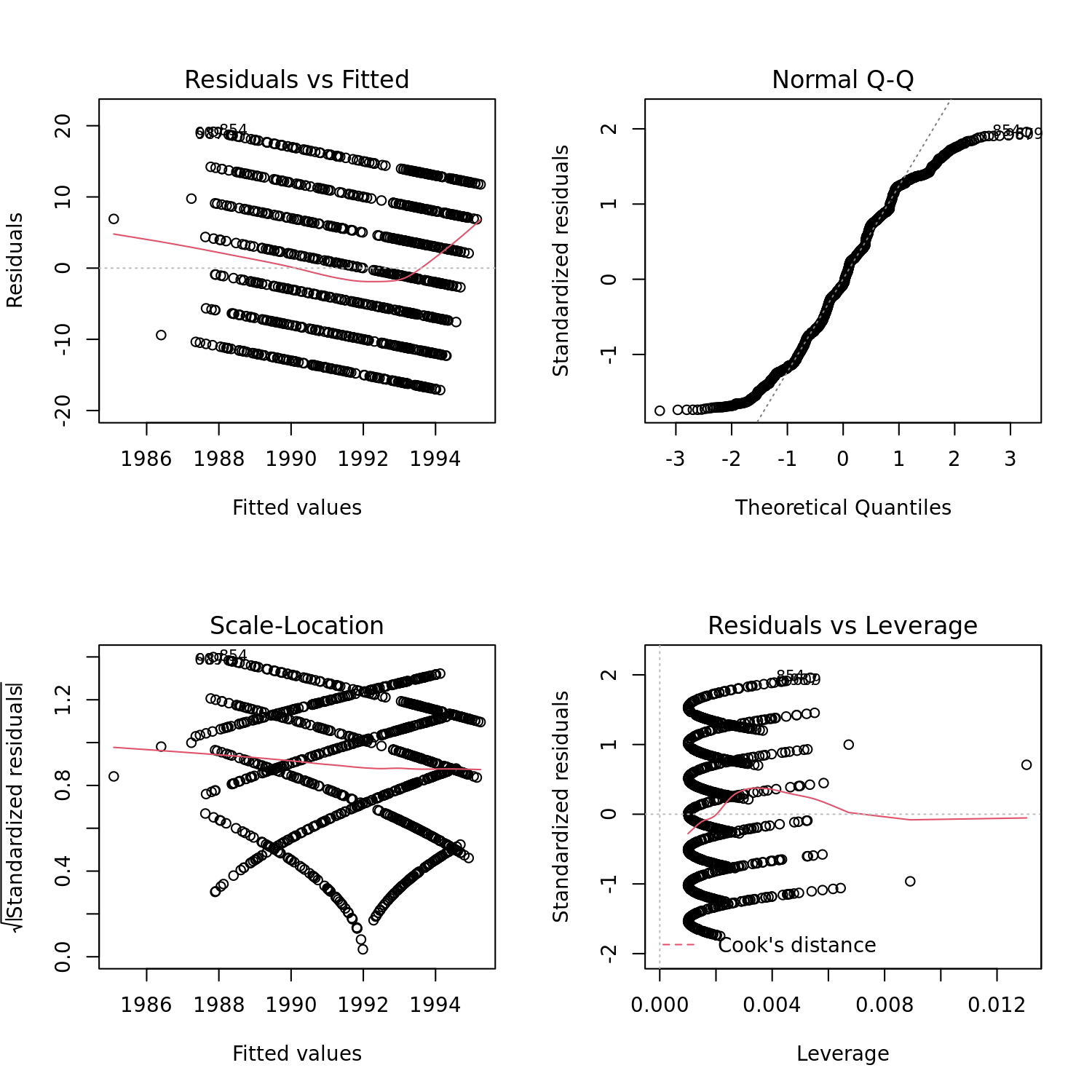
Table 5 summarizes the test results. For the Global test, the null hypothesis is all the four assumptions hold. Since its *p*-value is 0less than 0.05, at least one of the assumptions is violated.

The results of other tests are all displayed in the ‘*Decision*’ column. Each of them tests specific assumptions. One can determine exactly which assumption fails if there is any. The detail of these tests can be found in the *Terminologies* section.

## Regression Diagnostics

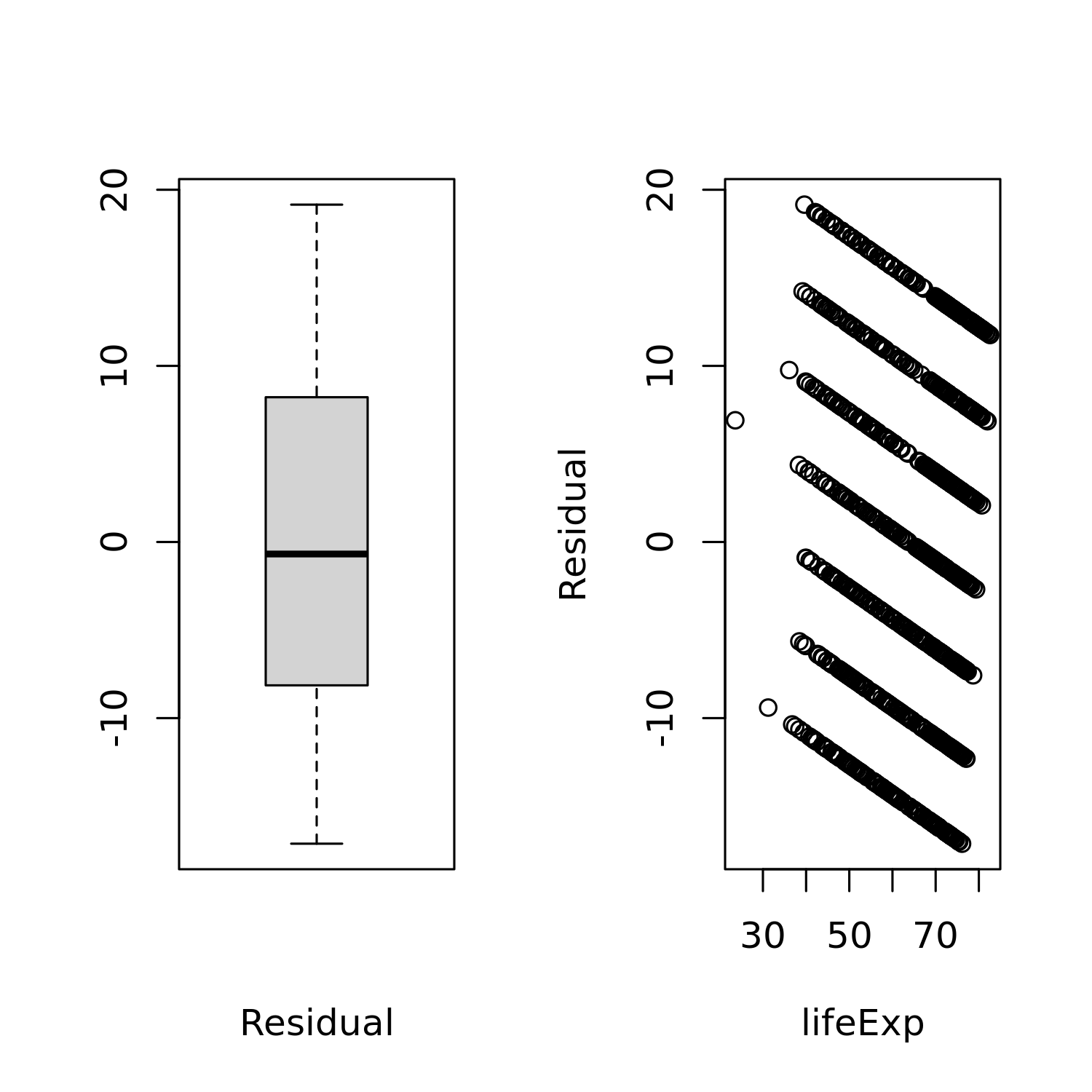
In statistics, a regression diagnostic is one of a set of procedures available for regression analysis that seek to assess the validity of a model in some different ways. This assessment may be an exploration of the model’s underlying statistical assumptions, an examination of the structure of the model by considering formulations that have fewer, more, or different explanatory variables, or a study of subgroups of observations, looking for those that are either poorly represented by the model (outliers) or that have a relatively large effect on the regression model’s predictions.

Four plots on the residuals are shown in Figure 2, including Residuals vs Fitted, Scale-Location, Normal Q-Q, and Residuals vs Leverage. The possible outliers are marked out if any.



“Diagnostics plots. (1) Residuals vs Fitted is a scatter plot of residuals and fitted values (estimated responses). The plot is used to detect non-linearity unequal error variances and outliers. (2) The Scale-Location plot shows whether the residuals are spread equally along with the predictor range i.e. homoscedastic. (3) The Normal Q-Q plot is used to learn whether it is reasonable to assume that the error terms are normally distributed. (4) The Residuals vs Leverage plots help to identify influential data points.”

Detailed information on these plots can be found in the *Terminologies* section.



Box plots of residuals and residual v.s. major predictors.

Observations with large residuals (absolute value) are potential outliers. Under the standard assumptions, the standardized residuals are uncorrelated with each of the predictor variables. If the assumptions hold, this plot should be a random scatter of points. Any discernible pattern in this plot may indicate a violation of some assumptions.

# Conclusions

Based on the above results, we can get the following conclusions:

* Based on the *p*-value, *lifeExp* is a significant major predictor in the model. It has significant effect on *year*.
* According to the adjusted *R*-squared value, the dependent variable *year* is not explained enough by the explanatory variables. Some informative predictors are not included in the model.
* Kurtosis test, Link Function test are not satisfied. The model is recommended to be refitted and the results of tests may lose credibility.

# Terminologies

***Regression coefficients***: Regression coefficients are estimates of the unknown population parameters and describe the relationship between a predictor variable and the response. In linear regression, coefficients are the values that multiply the predictor values, measuring how much one predictor variable could explain the outcome variable.

***Standard error***: The standard error of the coefficient estimator is a measure of how precisely the slope has been estimated. The smaller the standard error the more precise the estimator.

***t-statistic***: The t-statistic is defined by regression coefficient over its standard error. It is used to test whether the regression coefficient is equal to zero.

***p-value***: The *p*-value is the probability that the observations are observed if the true coefficient is equal to zero. If the *p*-value is too large, there is no significant relationship between the independent variable and the dependent variable. The variable is recommended to be deleted from the model. In the tables, “three stars” represents *p*<=0.001, “two stars” represents *p*<=0.01, and “one star” represents *p*<=0.05. If it was less than , it would be shown as ‘0’.

***R-squared ()***: *R*-squared, also called the coefficient of determination, is a goodness-of-fit index. It can be interpreted as the proportion of the total variability in the response variable that is accounted for by the predictor variable. If *R*-squared is small, some important predictor variables are probably missed and should be added to the model.

***Adjusted R-squared***: Adjusted *R*-squared is used to compare models having different numbers of predictor variables. If irrelevant predictor variables are added to the model, *R*-squared always increases. Adjusted *R*-squared tries to ‘adjust’ for the unequal number of variables in the different models.

***Square root of mean squared error (RMSE)***: The RMSE represents the square root of the second sample moment of the differences between predicted values and observed values. RMSE is a measure of accuracy. The smaller the RMSE is, the better the model fits.

***Degree of freedom (df)***: Degree of freedom is the difference between the number of observations and the number of unknown parameters.

***F-statistic (F-ratio)***: An F-statistic (F-ratio) is a ratio of the between-group variability and within-group variability. It is used to test if the means between two populations are significantly different. An F-test in regression tells if a group of variables are jointly significant.

***Skewness test***: The skewness test is to test whether the error terms have skewed distribution, or violate normality.

***Kurtosis test***: The kurtosis test is to measure deviations from the normal distribution kurtosis of the true error distribution. It is also used to test normality.

***Link Function test***: The link Function test is to test whether the model is mis-specified or other predictor variables are absent in the model. In other words, it is used to test the linearity assumption.

***Heteroscedasticity test***: The heteroscedasticity test is to test the presence of heteroscedastic errors and/or dependent errors. Both homoscedasticity and independent-errors assumptions are tested.

***Residuals vs Fitted values plot***: The Residuals vs Fitted values plot is a scatter plot of residuals and fitted values (estimated values of the response variable). The plot is used to detect non-linearity, heteroscedasticity, and outliers. Ideally, this plot should be a random scatter of points. If points form a pattern, for example, a quadratic pattern, it suggests that the linearity assumption is not reasonable. If the residuals increase (decrease) as fitted values increase, this suggests that heteroscedasticity exists. If some residuals stand out from the average level, it suggests that they could be outliers.

***Scale-Location plot***: The Scale-Location plot shows whether the residuals are spread equally along with the predictor range, i.e. homoscedastic. We want the line on this plot to be horizontal with randomly spread points on the plot.

***Normal Q-Q plot***: The Normal Q-Q plot is used to check whether the normality assumption satisfies. If the points are approximately linear, not departing from the dashed line, we proceed assuming that the error terms are normally distributed. Some points that are far away from the dashed line could be outliers.

***Residuals vs Leverage plot***: The Residuals vs Leverage plot helps to identify influential observations. Possible influential observations are points on the upper right or lower right corner, which are outside the red dashed Cook’s distance line. These points are so influential that should be paid more attention to. The regression results would noticeably change if removing them from the model.

***Outlier***: An outlier is a data point whose residuals differ significantly from other observations. An outlier can cause serious problems in statistical analyses. However, a careful decision should be made to deal with outliers instead of simply deleting them from the model.

***Influential point***: A point is an influential point if its deletion, singly or in combination with others, causes substantial changes in the fitted model (regression coefficients, fitted values, t-statistics, etc.).

***Variance inflation factor (VIF)***: Variance inflation factor (VIF) quantifies the severity of multicollinearity among the predictor variables in an ordinary least squares regression analysis. If this value is large, strong collinearity exists in the model and the estimation of the coefficient has low credibility.

***Generalized variance inflation factor (GVIF)***: In statistics, the variance inflation factor (VIF) quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate’s standard deviation) of an estimated regression coefficient is increased because of collinearity. Generalized variance inflation factor (GVIF) expands VIF to polytomous categorical variables.

***Akaike information criterion (AIC)***: The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

# References

Chatterjee, Samprit, and Ali S Hadi. 2015. *Regression Analysis by Example*. John Wiley & Sons.

Peña, Edsel A, and Elizabeth H Slate. 2006. “Global Validation of Linear Model Assumptions.” *Journal of the American Statistical Association* 101 (473): 341–54.