Regression Diagnostics- Data Analysis and Visualization

Dr. Farshid Alizadeh-Shabdiz

Spring 2021

Equation of the Individual F-Test

For an individual independent variable x_j, the F-statistic is given by:

$$F = \frac{(RSS_r - RSS_f)/(df_r - df_f)}{RSS_f/df_f}$$

where:

- RSS_r = Residual sum of squares for the **restricted model** (excluding x_j)
- RSS_f = Residual sum of squares for the **full model** (including x_j)
- df_r = Degrees of freedom of the restricted model
- df_f = Degrees of freedom of the full model

Since $df_r - df_f = 1$ (because we remove only one predictor), the equation simplifies to:

$$F = \frac{(RSS_r - RSS_f)}{RSS_f/(n - p - 1)}$$

where:

- n = Number of observations
- p = Number of predictors in the full model

How Measure Distance to Normal distribution – Visual inspection

1. Visual Inspection - Histogram & Q-Q Plot

- A **histogram** of residuals should resemble a bell curve.
- A **Q-Q plot** (quantile-quantile plot) compares quantiles of the error distribution with standard normal distribution.

How Measure Distance to Normal distribution - Shapiro-Wilk Test

- It checks normality, if
  - **p-value > 0.05**: Fail to reject normality (errors may be normal).
  - **p-value < 0.05**: Reject normality (errors are not normal).

- Note: Not suitable for large sample sizes.

How Measure Distance to Normal distribution – **Anderson-Darling Test**

- More powerful than Shapiro-Wilk for larger samples.
- If statistic > critical value at a given significance level, the data is not normal.

Measure Inflectional Point – Cook's distance

- Intuition:
  - We are finding a trend line through our data points. If removing **one data point drastically** changes the position of the line, that point is **influential**.
  - Cook's Distance tells us **how much the regression line would change** if we removed a particular observation.
- How It Works?
  - A point with **high leverage** (far from the mean of X) AND a **large residual** (not well predicted) has **high influence**.

There are different variations of Cook's Distance Equation

The formula for **Cook's Distance** for the ith observation is:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p \cdot \text{MSE}}$$

where:

- $\hat{y}_j$ = Predicted value for the jth observation in the **full model**
- $\hat{y}_{j(i)}$ = Predicted value for the jth observation when the ith data point is **removed**
- p = Number of predictors (including intercept)
- MSE = Mean Squared Error of the model

Cook's Distance D_i	Interpretation
$D_i < 0.5$	No significant influence
$0.5 \leq D_i < 1$	Moderate influence
$D_i > 1$	Highly influential (possible outlier)

- Outliers can be detected if their cook distances more than

  3*mean(cook's distances)

- Or Where the cook's distance is higher than 4/(n-k-1) – K is number of independent variables in the model.