Assumptions in Simple Linear Regression[1]

November 29, 2017

[1] HMS, 2017, v1.0

Chapter References

- Diez: Chapter 7
- Navidi, Chapter 7

Assumptions

- There is a linear relationship between x and y.
- The error terms (and thus the ys at each x) are normally distributed.
- The error terms (and thus the ys at each x) have constant variance (Homoscedasticity)
- The error terms are independent.

Assumption 1: Linearity

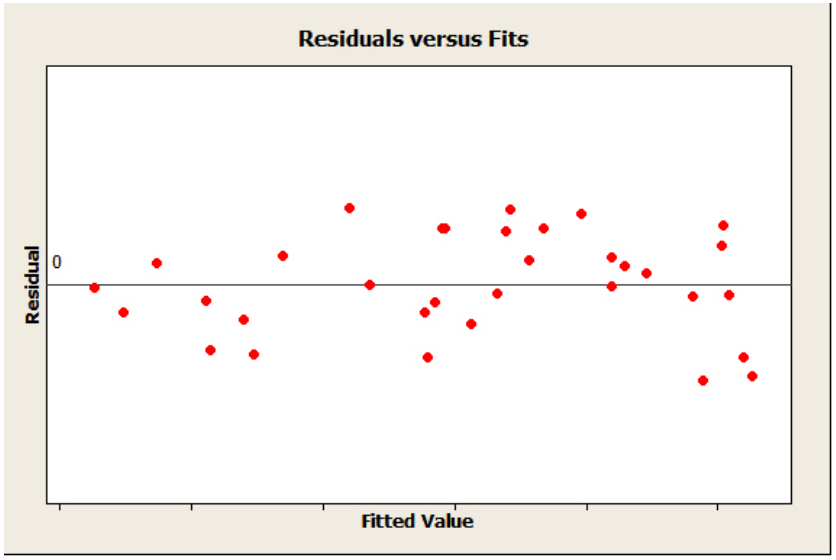
Assumption 1: Relationship is linear.

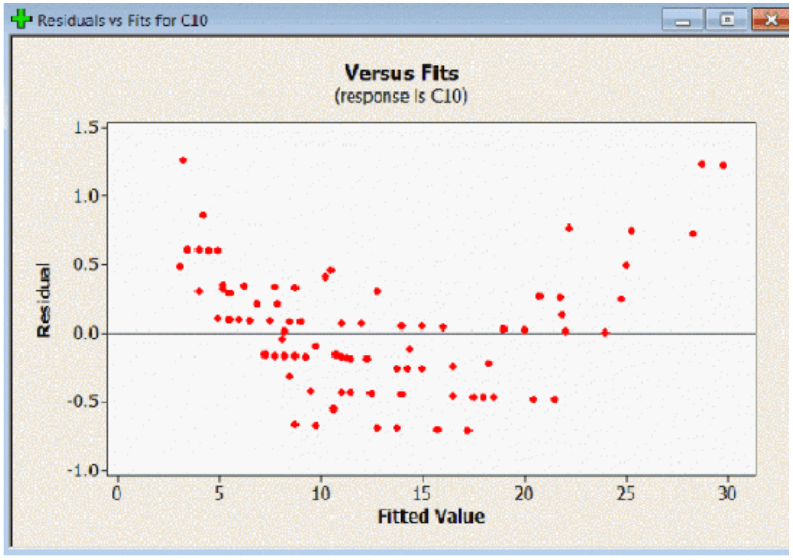
How to detect a problem:

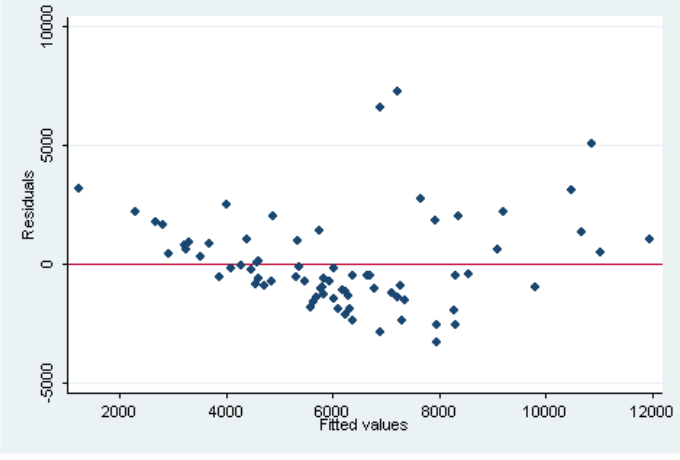
Plot residuals versus fitted values. If you see a pattern, there is a problem with the assumption.

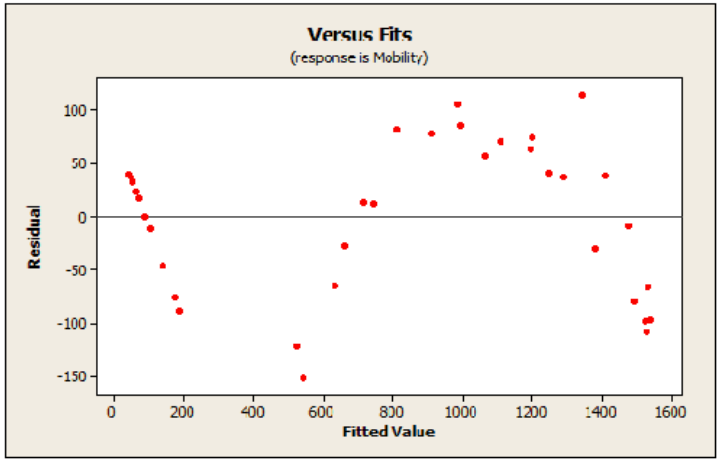
If linearity holds we expect the residuals to be a random cloud centered at 0.

Assumption 1: Linearity







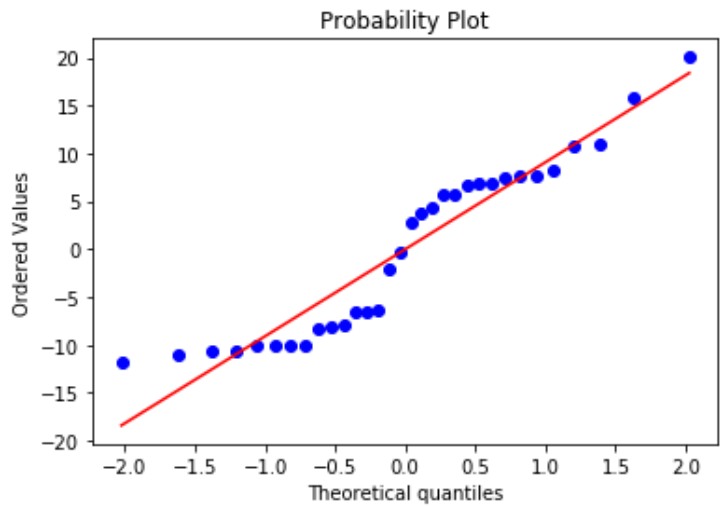


Assumption 2: Normality

Assumption 2: The residuals are distributed normally

How to detect a problem:

Do a Q-Q plot to determine whether the residuals are normally distributed. The following plot shows the QQ plot for the Burger King data:



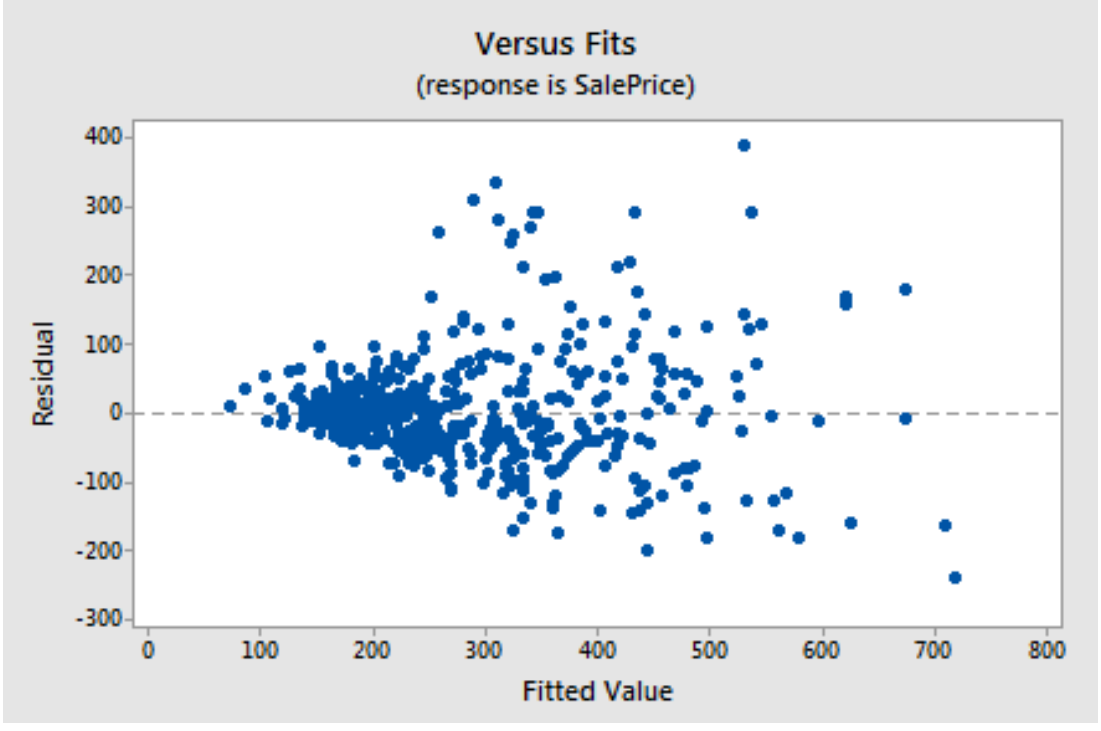
If you have a lot of residuals (eg ¿100) you can also do a frequency plot and see by eye whether the plot shows non-normal tendencies.

Assumption 3: Constant variance

Assumption 3: Constant variance of the errors across X values.

How to detect a problem:

Plot residuals versus fitted values. If you see increasing or decreasing spread, there is a problem with the assumption.



Assumption 4: Independent errors

Assumption 4 : Independent errors

How to detect a problem: Plot residuals versus fitted values. If you see

any pattern in the residuals this could indicate that the errors are not independent

a) Plot ε_i versus ε_{i+1} to see if there is a pattern.

b) More sophisticated way is to plot the autocorrelation as a function of different lags. If k is the lag, then:

$$r_k = \frac{\sum_{i=k+1}^{n}(y_i - \bar{y})(y_{i-k} - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

If there is no autocorrelation then the plot should be flat and close to zero.