Linear Regression: Statistical Inference

This file is meant for personal use by jemanuel.perez@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Outline

- Statistics Vs Machine Learning
- Assumptions of Linear regression
- Statistical Inference

This file is meant for personal use by jemanuel.perez@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Linear Regression Assumptions

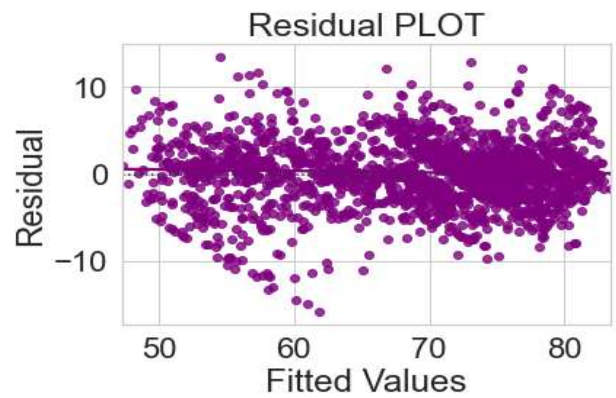
- Linearity: Independent and dependent variables are linearly related
- Independence: Residuals are independent
- Homoscedasticity: Equal Variance of residuals
- Normality: Residuals are normally distributed
- No (or little) Multicollinearity: Two or more independent variables have no (or little) correlation

If any of these assumptions are violated, then the forecasts, confidence intervals, and scientific insights yielded by the regression model may be seriously biased or misleading.

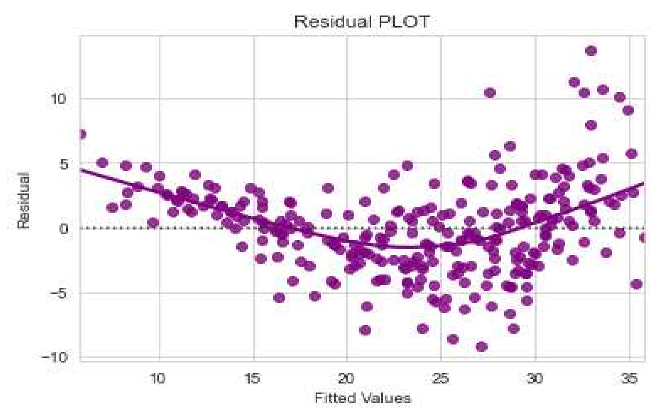
This file is meant for personal use by jemanuel.perez@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Linear Relationship

- After finding the best linear fit, a plot of the residuals will provide a good insight.
- If they don't follow any pattern, we say that the model is linear otherwise model is showing signs of non-linearity
- To deal with non-linearity, we can try transforming variables as per their relationship with target variable.



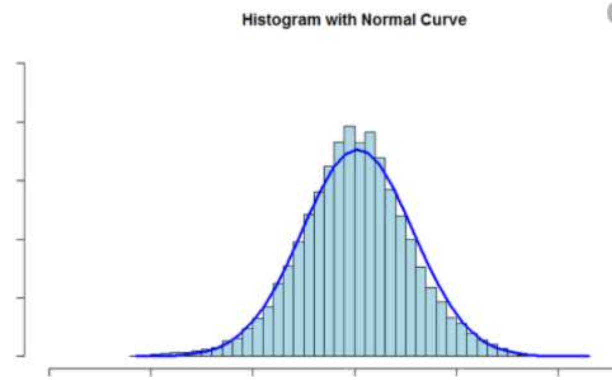
No pattern



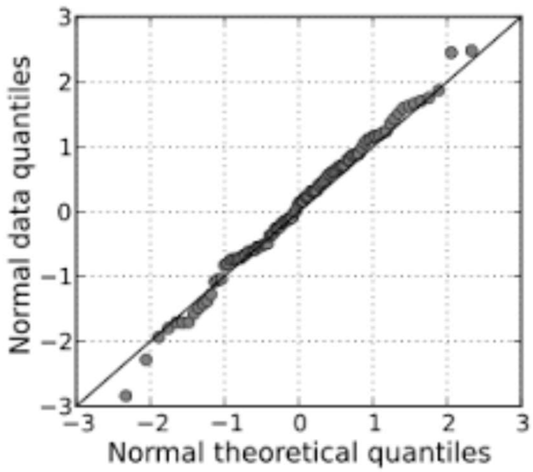
Some non-linearity

This file is meant for personal use by jemanuel.perez@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

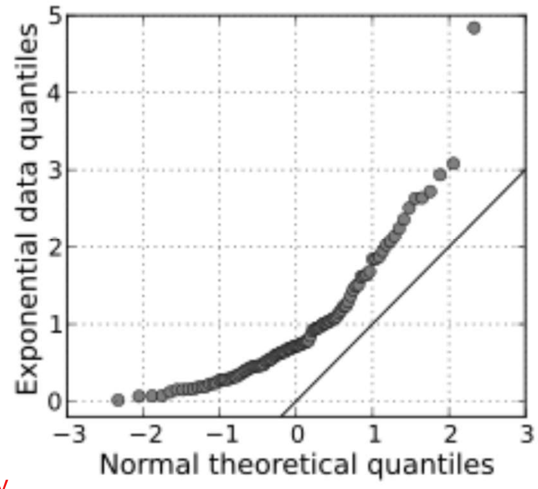
Normality



- Can be tested by plotting the distribution of residuals:
  - Histogram
  - Q-Q comparison plot
  - Tests for normality, like the Shapiro's test.
- When not normal
  - Transformations (log, exp etc.) of the dependent or independent variables can help



Normal

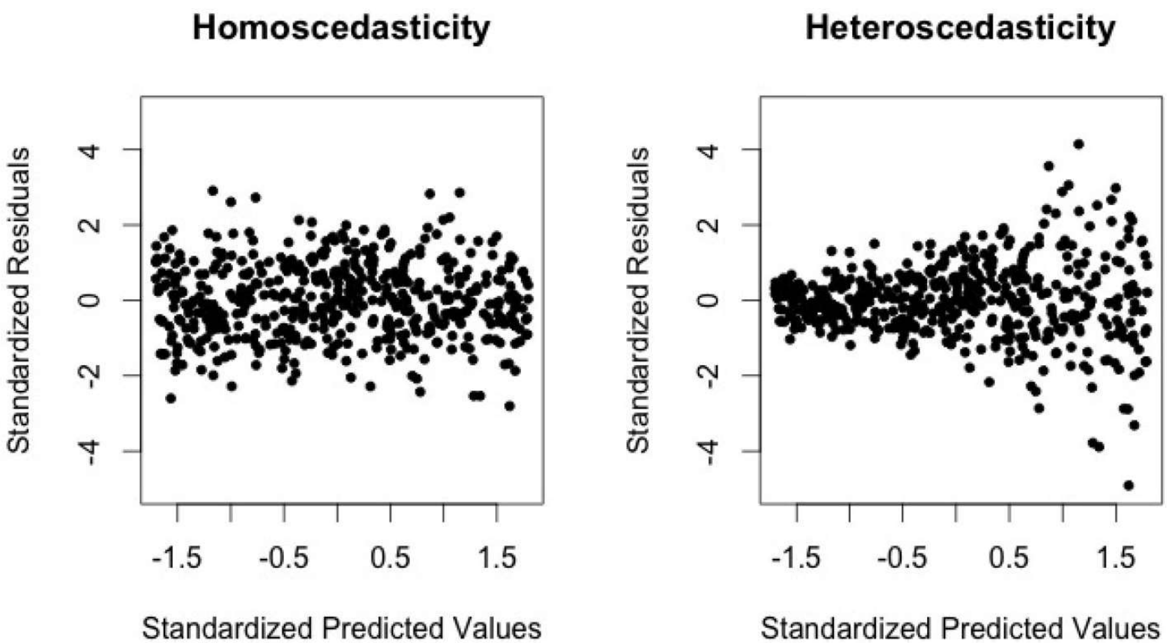


Not Normal

This file is meant for personal use by jemanuel.perez@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Homoscedasticity

- If the variance is not equal for the residuals across the regression line, then the data is said to be heteroscedastic.
- In this case the residuals can form a funnel shape or any other non symmetrical shape.
- Identifying the cause of hetroscedasticity is usually the best way to reason out ways to fix it



- Statistical test: The Goldfeld–Quandt test

This file is meant for personal use by jemanuel.perez@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Multicollinearity

- Multicollinearity occurs when independent variables in a regression model are correlated with each other.
- <u>When there is Multicollinearity, the relationship between any explanatory variable X and the response variable Y is not reflected by the coefficient of X</u>

- Could simply look at the entire correlation matrix
- Variance inflation factors also help identify multicollinearity
- Simply eliminating the linearly related variables or other dimensionality reduction techniques (like PCA) help reduce or eliminate multicollinearity.

This file is meant for personal use by jemanuel.perez@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Variance Inflation Factor (VIF)

$$VIF_j = \frac{1}{1 - R_j^2}$$

- R_j^2 is the R^2-value obtained by regressing the j-th independent variable on the other independent variables
- Variance inflation factors (VIFs) >= 1
- Tells you what percentage of the variance is inflated for each coefficient.
- For example, a VIF of 1.7 tells you that the variance of a particular coefficient is 70% bigger than what you would expect if there was no multicollinearity, i.e., if there was no correlation with other predictors.

This file is meant for personal use by jemanuel.perez@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Statistical Inference

- Given the best estimates from the data, what can we say about the unkown true model?
  - The unkown parameter? - confidence interval
  - Is there enough evidence in the data to say a coefficient is not zero? - hypothesis testing

This file is meant for personal use by jemanuel.perez@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Reviewing Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.814
Model:                            OLS   Adj. R-squared:                  0.809
Method:                 Least Squares   F-statistic:                     147.3
Date:                Wed, 09 Dec 2020   Prob (F-statistic):           1.20e-93
Time:                        12:48:42   Log-Likelihood:                -734.21
No. Observations:                 278   AIC:                             1486.
Df Residuals:                     269   BIC:                             1519.
Df Model:                           8
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const            -18.2835      5.549     -3.295      0.001     -29.209      -7.358
cylinders         -0.3948      0.423     -0.933      0.352      -1.228       0.439
displacement       0.0289      0.010      2.870      0.004       0.009       0.049
horsepower        -0.0218      0.016     -1.330      0.185      -0.054       0.010
weight            -0.0074      0.001     -8.726      0.000      -0.009      -0.006
acceleration       0.0619      0.118      0.524      0.601      -0.171       0.295
model year         0.8369      0.064     13.149      0.000       0.712       0.962
origin_america    -3.0013      0.704     -4.262      0.000      -4.388      -1.615
origin_asia       -0.6060      0.705     -0.860      0.391      -1.994       0.782
==============================================================================
Omnibus:                       13.244   Durbin-Watson:                   2.244
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               16.958
Skew:                           0.386   Prob(JB):                     0.000208
Kurtosis:                       3.932   Cond. No.                     8.26e+04
==============================================================================
```


This file is meant for personal use by jemanuel.perez@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.