11/4/2016

**Residual Analysis**

Residual is the difference between the observed value of dependent variable and the predicted value. A residual plot is a graph that shows residual on the vertical axis and the independent variable on the horizontal axis.

Residual analysis typically is used in a linear regression model; we check residual plot to verify if linear relationship is appropriate for the model. If residuals are randomly dispersed around horizontal axis with mean of 0, linear is appropriate, otherwise non-linear is appropriate.

We can also understand residuals from randomness point of view. There are two basic components of valid regression model:

Response = (Constant + Predictors) + Error = Deterministic + Stochastic

Stochastic means random and unpredictable, residual should be stochastic, in other words, if you observe explanatory or predictive power in residual, you know that your predictors are missing some of the predictive information. The process is easy to understand with a die-rolling analogy, when you roll a die you shouldn’t be able to predict which number on any given toss, if the number six shows up more frequently than randomness dictates, you know sth wrong with your understanding of how the die actually rolls, if a gambler analyzes at the die rolls, he could adjust his mental model and play style to factor in the higher frequency of sixes, and his new mental model better reflects the outcome. The same principles apply to regression models, you shouldn’t be able to predict error for any given observation, just like the die, if the residuals suggest that your model is systematically incorrect, you have an opportunity to improve your model.

Non-random pattern of residuals indicates the predictor variables of the model did not capture some explanatory information that is leaking into the residuals, there are several possibilities to the non-randomness:

Missing variable

Missing higher-order term of variable in the model to explain curvature

Missing interaction between terms

In addition to above, there are two more ways that predictive information can sneak into the residuals:

The residuals should not be correlated with another variable

Adjacent residuals should not be correlated with each other(autocorrelation): if we can use one residual to predict the next residual, there is some specific information present that is not captured by the predictors, typically this situation involves time ordered observations. For example, if a residual is more likely to be followed by another residual that has the same sign, adjacent residuals are positively correlated, we can use a time variable to capture relevant time correlated information or use time series model.

**Heteroscedasticity** refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it. The opposite of heteroscedasticity is **homoscedasticity**.

11/6/16

**Residual Analysis Tests**

<http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm>

Test for Residual Normality, the two most common types are histogram, normal probability plot (QQ plot, an approximately straight line should be produced if the points come from a normal distribution), QQ plots work better than histogram on small sample size of residual.

Small departures from the straight line in the normal probability plot are common, but a clearly S curve suggests a bimodal distribution of residuals.

We can also test independence of residuals over time, if the order of observation represents the order of execution of each treatment combination.

We also plot residuals versus corresponding predicted values (residuals vs fit), this should produce a distribution of points scattered randomly about 0, commonly though residuals may increase as the value of fit increases. When this happens, it suggests one should transform the response, perhaps by modelling its logarithm or square root.

Another important test of residuals is to test it against predictor variables to assure no explanatory information sneaked into predictors from residuals.

**Autocorrelation in the residuals**(Time Series Residual Analysis)

<https://www.otexts.org/fpp/5/4>

When the data are a time series, you should look at ACF plot of the residuals. They will reveal if there is autocorrelation in the residuals, another test of autocorrelation is Durbin-waston test, it will report P value, if p is small, there is significant correlation remaining in the residuals.

**Sensitivity Analysis**

Sensitivity analysis is an analysis method that is used to identify how much variations in the input values for a given variable will impact the result. It is concerned with the uncertainty inherent in mathematical models where the values for inputs can vary, conclusion drawn from studies can be significantly altered depending on such things as how a certain variable is defined or observed or measured for a study, when results of study do not significantly change due to variations in underlying assumptions, they are considered robust.

How does auto.arima work?

Selecting predictors