Lesson 14

Machine Learning – Linear Regression Assumptions

Kush Kulshrestha

Assumptions of Regression Analysis

There are 5 assumptions in regression analysis:

1. There should exist a linear relationship between the dependent variable (Y) and independent variables (X). This means that change in Y due to unit change in X should be independent of the value of X.
2. There should be no correlation between the residual terms. It is also called autocorrelation.
3. Independent variables (Xs) should not have high correlation with each other. If they are, we end up counting the effect multiple times. This is called multicollinearity.
4. Error term must have constant variance. This phenomenon is called homoscedasticity. Non constant variance is called hetroskedasticity.
5. The errors must be normally distributed.

1. Existence of Linear Relationship

There should exist a linear relationship between the dependent variable (Y) and independent variables (X). This means that change in Y due to unit change in X should be independent of the value of X.

If this is not true, then we end up modelling a linear model on non linear data. This will result in a poor fit model.

How to fix:

Consider applying a nonlinear transformation to the dependent and/or independent variables if you can think of a transformation that seems appropriate.

Another possibility to consider is adding *another regressor* that is a nonlinear function of one of the other variables. For example, if you have regressed Y on X, and the graph of residuals versus predicted values suggests a parabolic curve, then it may make sense to regress Y on both X and X^2 (i.e., X-squared). The latter transformation is possible even when X and/or Y have negative values, whereas log transform is not. Higher-order terms of this kind (cubic, etc.) might also be considered in some cases, but are not preferred.

2. Correlation in the Residuals

This is mostly applicable for time series regression models.

Serial correlation in the errors means that there is room for improvement in the model. This type of error is basically some correlation between consecutive error terms. This can be checked with the help of ACF and PACF plots.

In general, after fitting a time series model the residuals should be white noise. So they should have no autocorrelation.

This can be tested using Durbin-Watson test.

The *Durbin-Watson statistic* provides a test for significant residual autocorrelation at lag 1: the DW stat is approximately equal to 2(1-a) where a is the lag-1 residual autocorrelation, so ideally it should be close to 2.0--say, between 1.4 and 2.6 for a sample size of 50.

3. Multicolinearity

Multicolinearity refers to the condition when the input independent variables are highly correlated with each other. Check the correlation coefficients for all of the input pairs to check if there is multicolinearity present in the dataset.

According to Tabachnick & Fidell (1996) the independent variables with a bivariate correlation more than .70 should not be included in multiple regression analysis.

If you do find high collinearity, it means that your parameter estimates are unstable. That is, small changes (sometimes in the 4th significant figure) in your data can cause big changes in your parameter estimates (sometimes even reversing their sign). This is a bad thing.

If the correlations are high, this does not guarantee that there is multicolinearity. In order to confirm this, you have to check the partial correlations between the input variables. Having a value of >0.4 of partial correlation along with high correlation indicates the presence of multicolinearity.

If there are two variables showing multicolinearity, one of them should be excluded from the analysis as it is not adding any extra value or information in the analysis but is also making the regression coefficients unstable.

4. Homoscedasticity

Violations of homoscedasticity (which are called "heteroscedasticity") make it difficult to gauge the true standard deviation of the forecast errors, usually resulting in confidence intervals that are too wide or too narrow. In particular, if the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow.

Heteroscedasticity may also have the effect of giving too much weight to a small subset of the data (namely the subset where the error variance was largest) when estimating coefficients.

To diagnose: look at a plot of residuals versus predicted values. Be alert for evidence of residuals that grow larger either as a function of time or as a function of the predicted value.

To fix: If the dependent variable is strictly positive and if the residual-versus-predicted plot shows that the size of the errors is proportional to the size of the predictions, a log transformation applied to the dependent variable may be appropriate.

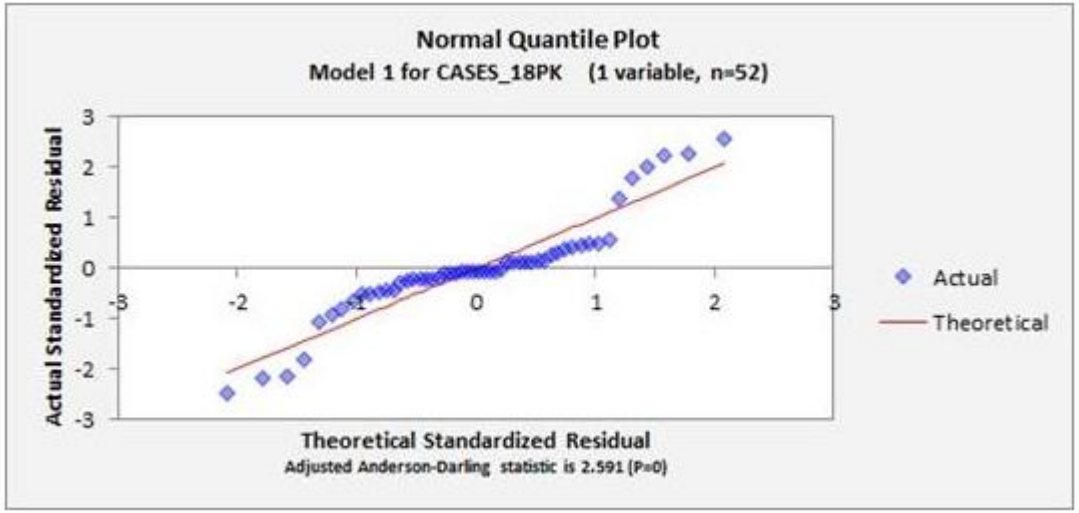
5. Normality of Residuals

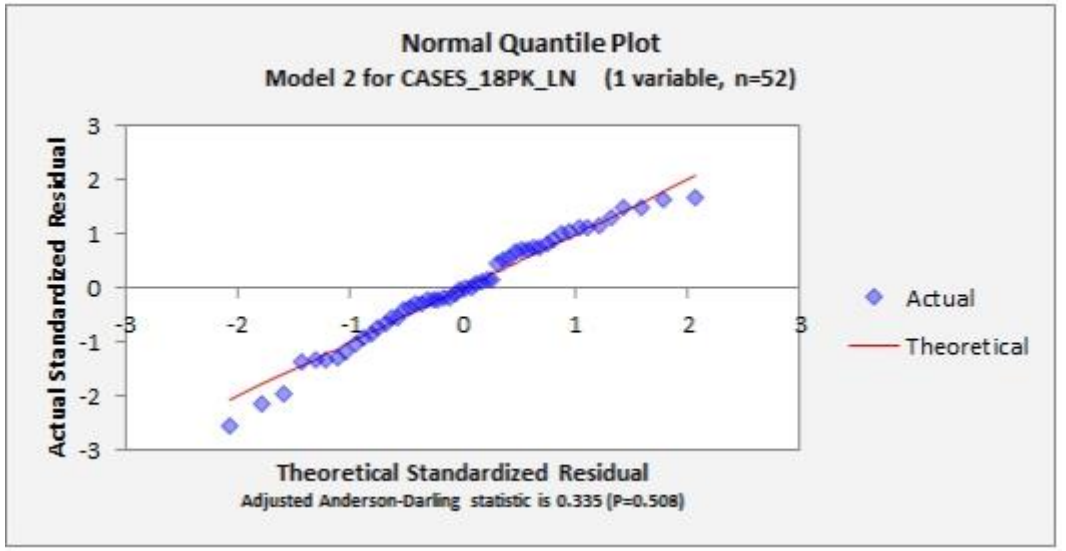
Violations of normality create problems for determining whether model coefficients are significantly different from zero and for calculating confidence intervals for forecasts. Calculation of confidence intervals and various significance tests for coefficients are all based on the assumptions of normally distributed errors. If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow.

The best test for normally distributed errors is a **normal probability plot** or **normal quantile plot** of the residuals. If the distribution is normal, the points on such a plot should fall close to the diagonal reference line.

There are also a variety of statistical tests for normality, Jarque-Bera test is the one we have used.

Here is an example of a bad-looking and a good looking normal quantile plot:





5. Normality of Residuals

In such cases, a nonlinear transformation of variables might cure both problems. In the case of the two normal quantile plots above, the second model was obtained applying a natural log transformation to the variables in the first one.

The dependent and independent variables in a regression model do not need to be normally distributed by themselves--only the prediction errors need to be normally distributed.