# Search for best Linear multiple regression model

Our assumptions test that will be used to remove complexity in the model will be based on the removing the

The Regression model with 127 model is susceptible to over fitting issue, because of the high complexity and introduced high variation due to the fitting the noise in model. Using SPSS this highly complex model has given us R-square 81%, showing as if the model can accurately forecast 81% variation with the sample data.

At the same time the adjusted R-square is 69%, which has given us 14% difference between the sample and population, this difference of sample and population values in coefficient of determination can be used to find the best fit model.

# The Approach

Here is the article that has been used as reference:

<http://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables>

The author argues the regression model with forecast good results and it will show some variables are highly significant is some cases. Here are the two learning that are useful:

* Use the adjusted R-square to compare models with different numbers of predictors
* Use the predicted R-square to determine how well the model predicts new observations and whether the model is too complicated

## Removing the variables that show violation of assumption

Our assumptions test that will be used to remove complexity in the model will be based on the removing the variables which violates the following tests and assumptions:

* **MULTICOLLINEARITY**
  + VIF < 10
* **Normality**
  + **Plots Histogram/NPP**
* **Independence**
  + **Durbin Watson**
* **Constant variance**
* **Linearity**
* **Randomness**

# Findings

Here are the findings after running 12 iterations by reducing the variables which are non-significant and having significant assumption violations such as VIF.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Iteration | Variables | Predicted R-Square | Adjusted R-Square | Difference (%) | Worst VIF | Improvement from previous model |
| 0 | 127 | 81.3 | 69.8 | 14.15 | 4398.09 | 0 |
| 1 | 101 | 78.6 | 67.9 | 13.61 | 207.7 | 1.90 |
| 2 | 50 | 69.9 | 64.4 | 7.87 | 145.13 | 3.50 |
| 7 | 7 | 61.2 | 61.1 | 0.16 | 62.8 | 3.30 |
| 11 | 4 | 62.8 | 62.3 | 0.80 | 2.4 | -1.20 |
| 12 | 3 | 62.6 | 62.3 | 0.48 | 2.27 | 0.00 |

The Final best score for R-square came out to be 62.6%. The improvement from the last iteration 11 did not improve thus we stopped at this iteration.