**Simulation Modeling with Excel Homework Assignment: *Impact of Non-Normality***

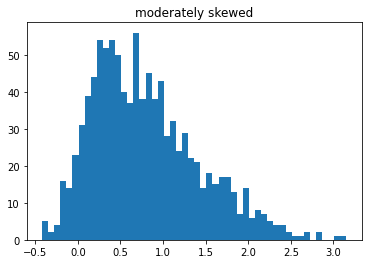
*by Lev Tsipes*

Firstly, with scipy package I generated four populations of different skewness levels, N = 1000 for each. Distributions are presented below

Chart, histogram

Description automatically generatedChart, histogram

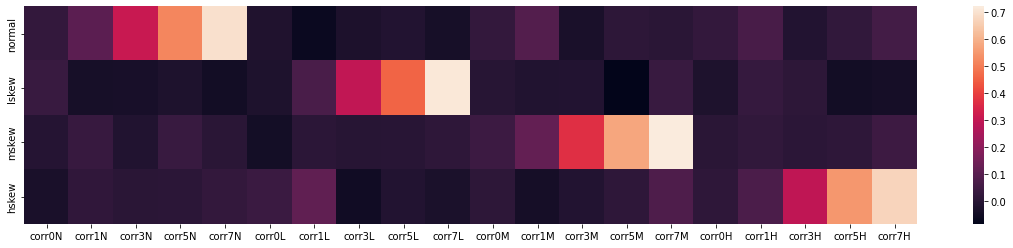
Description automatically generated

Chart, histogram

Description automatically generated

Secondly, for each of the populations I generated 5 correlated (corr) variables: correlations of

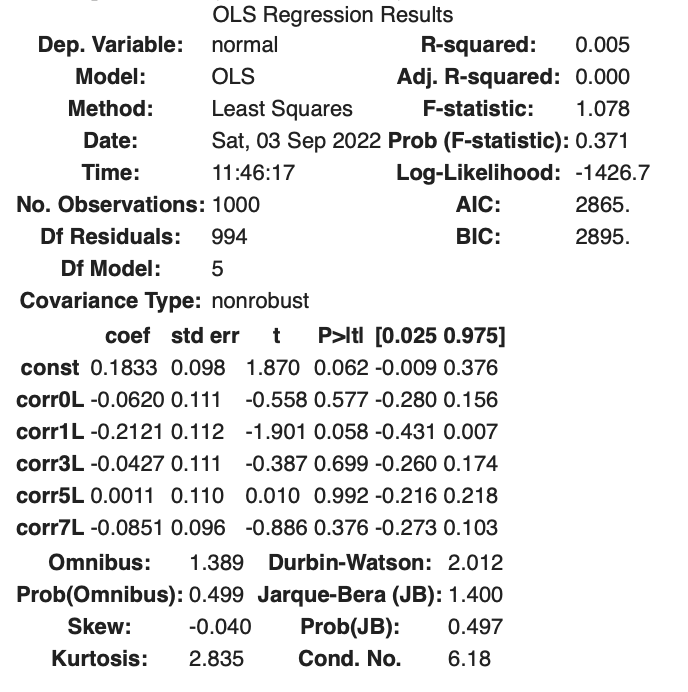
0; 0.1; 0.3; 0.5 and 0.7. So, in total there are 5\*4 = 20 variables.



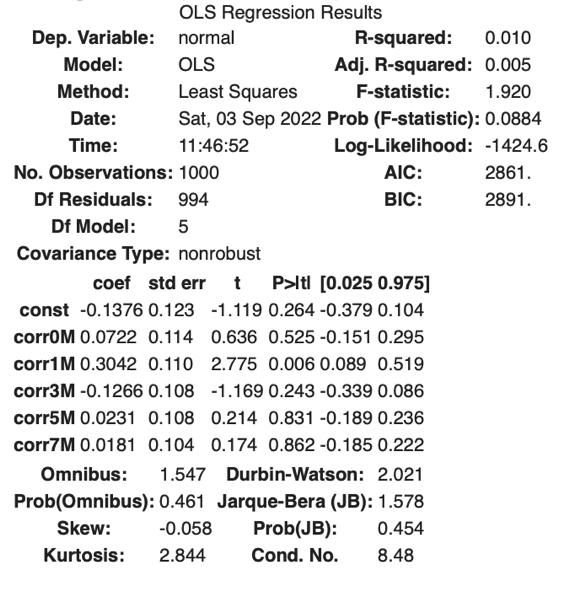
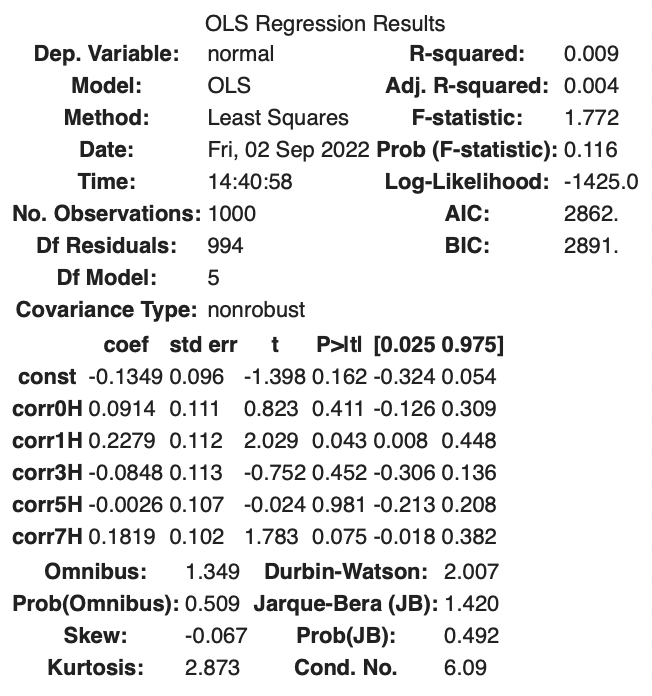
Then, I generate 4 classic OLS model, where one of the four initial variables is a DV and according corr variables are IVs.

Normal distribution Lightly skewed distribution

Table

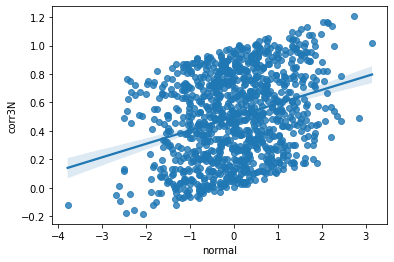
Description automatically generated 

Moderately skewed distribution Highly skewed distribution

Here, we see that only the model with normally distributed variables gets adequate results. R-scores, p-values and F-statistics for models with skewed data are inappropriate. Also, it is interesting that for normally distributed expectedly the more correlated the IV is with the DV the smaller is the p-value i.e., the more important the IV is, whereas the opposite is not true for the skewed data. Less correlated IV are identified as important by OLS on skewed data. This is explained by the following. Skewed IVs with 0.1 correlation level get the best p-vals, as soon as they are both distributed normally (since the IVs were generated by mixing DVs with random variables distributed normally) and contain some information about the corresponding DVs

These tables highlight the importance of the assumption of normal distribution of residuals, that cannot be distributed in Laplace–Gauss way when IVs are distributed skewedly. The consequence of breaking assumption is metrics showing scores that do not reflect reality. The regression lines though, may still be representative as illustrated below

 Chart, scatter chart

Description automatically generated

However, when we get a smaller sample, the picture changes. I generated 1000 samples of size of 10, 15, 30 or 1000 for each set of the variables and got the following results.

Normal Lightly Skewed Moderately Skewed Highly Skewed

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

A picture containing shape

Description automatically generated A picture containing shape

Description automatically generated A picture containing shape

Description automatically generated Shape

Description automatically generated with medium confidence

We can see that the results can be hardly told apart. The reason to such effect is that the sample size indirectly affects all the scores. It affects mean y in the formulas of R-squared & F-s (hence, p-val, type I error). This illustrates how important the sample size is and how undersampling can produce results with misleading metric scores.