Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 49(1) 81-86 (2017)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Controlled Data: Generation, Testing, Modeling and Impact Gauging

M. M. IQBAL, S. RASOOL, S. AFZAL

Department of Statistics, Bahauddin Zakariya University, Multan

Received 21st April 2016 and Revised 20th November 2016

Abstract: Several test procedures are available to check the admissibility of underlying assumptions of statistical procedures. Such procedures have potential to confirm or deny the adequacy of the assumption. In real data sets we do have indications and symptoms regarding assumptions but we are not sure about their presence and/or intensity. In the circumstances we have to rely on whatever is being told by the test procedures. To check the admissibility and potentials of these test procedures we need to apply them on the data which have been cloned to possess certain characteristics at known level. This article concentrates on the generation of controlled data with known level of multicollinearity, heteroscedasticity, and autocorrelation and establishing relation of common test procedures with the intensity of the violation of basic assumptions of linear models.

Keywords: Controlled Data, Cloned Data, Multicollinearity, Heteroscedasticity, Autocorrelation

1. INTRODUCTION

By controlled data we mean a data set whose generation has been controlled in the desired way. That is the generated data will have desired features built-in to it. In other words the generated dataset is cloned. The terms "Controlled Data" and "Cloned Data" are used interchangeably here but those are different from "Simulated Data" Morgan (1984) and "Random Data". The random data generation routines generate data that follow a particular distribution with specific construct of the underlying parameters whereas Simulated Data can be considered as the data used to imitate the happenings of a real-world system.

The controlled data is actually a blend of simulation and random data. The characteristics of both of the approaches are blended to exercise the control on the generated data. Thus controlled data provides basis for putting several of the procedures used for checking usual assumptions of various statistical techniques on test with higher level of clarity and objectivity.

In this piece of research we relied upon controlled data. To stay within the scope we restricted ourselves to three assumptions of linear regression models namely multicollinearity, heteroscedasticity, and autocorrelation. For the purpose we generated dataset by exercising our control on the level and extent of the departure of the data from each of the above-mentioned assumptions. The test procedures are then applied on the data sets to check if the test procedures are capable to detect the presence and to report the intensity of the departure.

A blend of computational approaches had to be employed to get the needful done. We mainly relied upon Minitab 16 and MS Excel 2007 for the purpose. Both of them facilitated us through their macros to clone the dataset ultimately generated for further processing.

It has always been told in the literature, lectures, notes, and discussions that departure from the basic assumptions of least square may put sever adverse effects on the quality of the regression estimates Gujarati (1978). Although this consequence is well known but the intensity of the affect has not been regulated. That is the relation between level of departure and the intensity of the consequence is unaddressed.

It was considered good to give it a go. For the purpose we started with three assumptions namely multicollinearity, heteroscedasticity, and autocorrelation. Essentially a similar approach was developed which was used for each of the assumptions.

2. MATERIAL AND METHODS

The generic approach is outlined as follows:

1. Draw sufficient number of sufficiently large samples with defined (and known) level of departure of the underlying assumption. For example a given level of multicollinearity.

++Corresponding author Email: mutahiriqbal@bzu.edu.pk, Shafqat553@gmail.com, saimaafzalbzu@bzu.edu.pk

2. Estimation in the presence of the built-in departure from the underlying assumption at the defined (and known) level of intensity. Has to be repeated for every single sample.
3. Application of the suitable test procedure to confirm its ability of detection of the problem and recording the respective p-value.
4. Rectification of the problem by using appropriate modification to the data or estimation procedure to define "control" group for comparison.
5. Establishment of relationship between these results.

2.1 Adapted Method for Multicollinearity

125 correlation matrices of order 4×4, mean vectors of order 4×1, and variance vectors of order 4×1 were generated and stored for further processing. The upper diagonal elements of the correlation matrices were generated using uniform random numbers over the interval 0 and 100 duly divided by 1000. The diagonal elements were obviously set at 1 and lower diagonal elements were copied appropriately to make it symmetric. The constant mean vector and constant variance vector were generated systematically using the pattern $0.01(0.04)4.97$ and $0.101(0.004)\ 0.597$ respectively

Using each of the 125 sets of correlation matrix, mean vector and variance vector a sample was drawn from multivariate normal distribution which has the correlation structure given in correlation matrix. That is multicollinear data with known correlation structure.

125 random samples each of size 3000 were drawn from standard normal distribution which were used as error term in fitting the regression model on multicollinear data.

125 Y-variables were generated using each of the 125 sets of collinear X-variables using same assumed values of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and one of the 125 error terms in turn.

A regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ was fit for each of the 125 samples and corresponding Standard Errors of each of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ were stored.

In each of the 125 cases, to define Variance Inflation Factor (VIF) for each of the regressor in turn was regressed on the remaining regressors and corresponding coefficient of determination was calculated and saved.

Each of the 125 collinear sets of regressors was orthogonalized by replacing them with their Principal Component Scores. Keeping rest of the data intact we repeated the whole process and fit the regression model $Y = \delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_3 Z_3 + \delta_4 Z_4 + \varepsilon$, where Z_i are the Principal Component Scores and Standard Errors of each of the $\hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3, \hat{\delta}_4$ were stored.

The above-mentioned results namely Standard Errors for $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ and $\hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3, \hat{\delta}_4$ were compared by taking their ratios. A perfect relation proportional to the VIF was expected.

Three samples of 1000 observations each were drawn from uniform distribution over the interval 0 and 5. These were regarded as x_1, x_2, x_3 and were used as explanatory variables in the subsequent iterative proceedings.

2.2 Adapted Method for Hetroscedasticity

100 samples each of size 1000 were drawn from normal distribution with zero mean but incremental value of the variance following the series $1(1)100$ for every subsequent sample. Each sample was multiplied by the product term $x_1 \times x_2$. Each of the resultant sample was used as error term for generating Y, the dependent variable. The Y_s were generating using each of the error term blended with the set of explanatory variable namely x_1, x_2, x_3 using arbitrary values of the four regression parameters using

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

This is how we end up in defining 100 Y_s which are duly heteroscedastic as per above-mentioned scheme.
The heteroscedasticity of the data was checked by means of Goldfeld-Quandt test (1972) and corresponding value of statistic and P Values were tabulated against level of heteroscedasticity. This provides us a table of three columns and 100 rows. Each row corresponds to the level of heteroscedasticity and each column represents value of test statistic and P Value.

2.3 Adapted Method for Autocorrelation

To check the autocorrelation in the data, three variables, having 100 number of observations each, namely, x_1, x_2, x_3 are generated randomly from uniform distribution over the interval between 1 and 100 as 100 samples are generated from standard normal distribution. Using a predefined value of the autocorrelation factor Rho(ρ), the autocorrelated error terms are defined using the relationship

$$\mu_i = \rho(\mu_{i-1}) + \varepsilon_i$$

Now we were in a position to generate Y, the dependent variable, by using x_1, x_2, x_3 as three independent variables and the above-mentioned autocorrelated error term using the model with arbitrary values of the regression parameters

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \mu_i$$

This is how we have 100 datasets of Y, the same set of dependent variable, x_1 , x_2 , x_3 explanatory variables.

By fitting the regression model on each of the 100 sets separately we obtained Mean Squared Error for each of the model and tabulated along with corresponding P Value.

3. THE ANALYTICAL PROCEDURE

After having generated controlled data to impersonate the problem of multicollinearity, heteroscedasticity, and autocorrelation with controlled level of problem we started analyzing the same separately for each type of the inherited problem.

3.1 The Problem of Multicollinearity

We had 125 sets of four explanatory variables with known level of collinearity along with 125 corresponding Y variables. That is we had 125 different sets of Y, X_1, X_2, X_3, X_4 for each scenario based upon mean vector, variance vector, correlation matrix to generate random samples from multivariate normal distribution. Each variable had 1000 observations. We applied Durbin-Watson test (1971) on each of the data set and recorded value of the test statistic. As the presence of multicollinearity is suppose to cause inflation of the variance of the estimates of regression coefficients thus standard errors of the regression coefficients of the model fitted to each of the 125 sets of Y, X_1, X_2, X_3, X_4 were recorded in tabular form.

The level of multicollinearity was brought to level nil by orthogonalizing each of the 125 sets. The explanatory variables X_1, X_2, X_3, X_4 were replaced with Principal Component Scores Z_1, Z_2, Z_3, Z_4which are orthogonal by principle. To check the impact of removing multicollinearity from the data is expected to be reflected in the standard error of the regression coefficients estimated using orthogonal data.

Ideally a dataset with no multicollinearity is used as base to gauge the impact of the collinearity on the standard errors of the regression estimates. The Variance Inflation Factor deals the situation in that direction i.e. from orthogonality to collinearity. We were dealing the situation in reverse direction so an inverse function of the Variance Inflation Factor was introduced to obtain the reverse effect of transforming data from multicollinearity to orthogonality. We termed it as Variance Deflating Factor.

$$VDF_j = \frac{1}{VIF_j} = \left(1 - R_j^2\right)$$

where $R_j^2 = \mathrm{R}^2$, in the regression of X_j on the remaining (*k-2*) regressions

The two tabulations were compared for each scenario to see if there is any role Variance Deflation Factor can play.

The Standard Errors of the estimates of regression coefficients for the multicollinear and orthogonal data corresponding to each of the scenarios are presented in **(Table-1)**.

Table 1: Standard Errors of the Estimates of the Regression Coefficients for Multicollinear and Orthogonal Data Sets

$SE(\hat{\beta}_0)$		$SE(\hat{\beta}_1)$		$SE(\hat{\beta}_2)$		$SE(\hat{\beta}_3)$		$SE(\hat{\beta}_4)$		Case
Mul	**Orth**	**Mul**	**Orth**	**Mul**	**Orth**	**Mul**	**Orth**	**Mul**	**Orth**	
0.03279	0.03264	0.33483	0.03060	0.32272	0.03258	0.03329	0.03329	0.03455	0.03455	1
0.04110	0.04092	0.41974	0.03837	0.40456	0.04084	0.04173	0.04173	0.04331	0.04331	2
0.04102	0.04084	0.41889	0.03829	0.40374	0.04076	0.04165	0.04165	0.04322	0.04322	3
0.04099	0.04081	0.41861	0.03826	0.40348	0.04073	0.04162	0.04162	0.04319	0.04319	4
0.04093	0.04075	0.41802	0.03821	0.40290	0.04067	0.04156	0.04156	0.04313	0.04313	5
...	...	...	...	...	...	...	...	...	...	...
0.03832	0.03815	0.39134	0.03577	0.37719	0.03808	0.03891	0.03891	0.04038	0.04038	121
0.03846	0.03830	0.39282	0.03590	0.37861	0.03822	0.03906	0.03906	0.04053	0.04053	122
0.03843	0.03826	0.39243	0.03587	0.37824	0.03818	0.03902	0.03902	0.04049	0.04049	123
0.03840	0.03824	0.39220	0.03585	0.37801	0.03816	0.03899	0.03899	0.04047	0.04047	124
0.03835	0.03818	0.39162	0.03580	0.37745	0.03810	0.03894	0.03894	0.04041	0.04041	125

To check the impact of the orthogonalization the ratio of the standard errors of the estimates of the regression coefficients obtained for the two types of the data sets in each of the 125 cases and tabulated in **(Table-2)**.

Table 2: Ratio of Standard Errors of Estimates of Regression Coefficients for Multicollinear and Orthogonal Data

Ratio of Standard Errors of Estimates of Regression Coefficient for Multicollinear to Orthogonal Data					Case
$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	
0.99565636521	0.09140323974	0.10094812307	0.10242239094	0.10370484984	1
0.99565636521	0.09140323974	0.10094812307	0.10242239094	0.10370484984	2
0.99565636521	0.09140323974	0.10094812307	0.10242239094	0.10370484984	3
0.99565636521	0.09140323974	0.10094812307	0.10242239094	0.10370484984	4
0.99565636521	0.09140323974	0.10094812307	0.10242239094	0.10370484984	5
...	...	...	...	...	...
0.99565636521	0.09140323974	0.10094812307	0.10242239094	0.10370484984	121
0.99565636521	0.09140323974	0.10094812307	0.10242239094	0.10370484984	122
0.99565636521	0.09140323974	0.10094812307	0.10242239094	0.10370484984	123
0.99565636521	0.09140323974	0.10094812307	0.10242239094	0.10370484984	124
0.99565636521	0.09140323974	0.10094812307	0.10242239094	0.10370484984	125

3.2 The Problem of Heteroscedasticity

The heteroscedastic errors are quite common in real life situations particularly when cross-sectional data is handled. Consequences of heteroscedasity is required to be taken care of otherwise regression estimates would lose their validity. For the purpose the intensity of the heteroscedasticity would play a vital role. We were interested to establish a way to gauge the intensity by establishing a relation between value of the test statistic and the intensity of the heteroscedasticity by means of an empirical study.

We drew 100 samples of size 1000 each from uniform distribution over the interval 1 and 100 and used as X_1. Similar samples were obtained for X_2 andX_3. That is we had 100 sets of X_1, X_2, X_3 which we used as sets of explanatory variables in the upcoming analysis. We used each set iteratively.

We then generated 100 samples of size 1000 from normal distribution with mean 10 and variance $1(1)100$. Each sample was used to provide basic error term. The error term so obtained was already heteroscedastic but to add the complexity and to intensify the level we multiply each error term with product of the respective X_1 and X_2. One hundred error terms are thus finalized.

We drew four samples of size 100 each from uniform distribution over the interval 1 and 10 and tabulated. Each row of the 4×100 table was used as the set of values assumed for β_0, β_1, β_2, β_3 in defining each of the 100 Y variables using the relation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + error\ term$$

This equipped us with 100 sets of Y, X_1, X_2, X_3 where presence as well as intensity of the heteroscedasticity is already known.

Using each of the above-mentioned set we attempted Goldfield Quandt test and compiled the P-Value, F-Ratio (the test statistic), Standard deviation of the residuals calculated earlier. The results produced from the analysis are presented as **(Table-3)**.

Table 3: Results of 100 Replications of Goldfield Quandt Test on Data with Known Level of Heteroscedasticity Available in SD Column

P-Value	F-Ratio	SD	Case
0.0001	1.3906	9129	1
0.0001	1.3960	8727	2
0.0254	1.1919	8987	3
0.0000	1.5053	8965	4
0.0103	1.2316	8855	5
...	...	....	…

P-Value	F-Ratio	SD	Case
...	...	...	…
0.1342	1.1046	9006	96
0.0001	1.4002	8793	97
0.0000	1.4931	9115	98
0.0015	1.3053	8844	99
0.0534	1.1560	8509	100

3.3 The Problem of Autocorrelation

The problem of autocorrelated error terms is more often encountered in time series data sets. Although tests procedures like Durbin Watson Test are available to check the presence of problem in the data. The test not only is able to detect its presence but it is capable to state the negative or positive direction of autocorrelation. Mere detection of presence and/or direction may not be enough as the strength of the autocorrelation does affect the quality of the least squares estimates of regression parameters.

We were interested to gauge the impact of autocorrelation on the least squares estimates in terms of

its intensity. We were, therefore, intended to establish a relation between level of autocorrelation present in the error term and its impact on regression estimates. As it has already been established that

$$Var(\hat{\beta})_{AR(1)} = Var(\hat{\beta})_{OLS}\left(\frac{1+r\rho}{1-r\rho}\right)$$

So the knowledge of the level of first order autocorrelation ρ would be enough to gauge its impact on the variance of the estimate.

For the purpose we generated data with known level of first order autocorrelation as follows:
Suppose in the two variable model the true values of intercept and slope coefficient are known thus

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

Thus

$$E(Y_t|X_t) = \alpha + \beta X_t$$

defines population regression function. If it can be assumed that ε_t is generated by the first order autoregressive scheme following the pattern

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$$

Assuming $\nu_t$follows all assumptions of ordinary least squares estimation procedure.
Now a sample of appropriate size is drawn from standard normal distribution and ε_t are generated using

$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t.$

This requires an initial value to begin with.
The Y variable is then generated using

□□ = □ + □□□ + □□

Twenty different scenarios were used which are available in **(Table-4)**.

Table 4: Twenty Scenarios used to Generate First Order Autoregressive Data and Analysis

Scenario	α	β	ρ	Scenario	α	β	ρ
1	3	0.66	+0.1	11	3	0.66	-0.1
2	6	1.32	+0.2	12	6	1.32	-0.2
3	9	1.98	+0.3	13	9	1.98	-0.3
4	12	2.64	+0.4	14	12	2.64	-0.4
5	15	3.30	+0.5	15	15	3.30	-0.5
6	18	3.96	+0.6	16	18	3.96	-0.6
7	21	4.62	+0.7	17	21	4.62	-0.7
8	24	5.28	+0.8	18	24	5.28	-0.8
9	27	5.94	+0.9	19	27	5.94	-0.9
10	30	6.60	+1.0	20	30	6.60	-1.0

4. DISCUSSION

Three subject areas are discussed in turn.
The standard errors of the least squares estimates of the regression coefficients were obtained from collinear data. The data was then orthogonalized and standard errors of the least squares estimates of the same regression coefficients were obtained once again.

As per established theory the variance of the standard error of the OLS estimates of the regressors gets increased which can be defined as a factor known as Variance Inflation Factor. As the multicollinear data was first generated and the standard errors of the OLS estimates of the four regression coefficients (excluding intercept) were obtained. These standard errors were used as base for the comparison which is why a reversing factor termed as Variance Deflating Factor was introduced to see how removal of the collinearity deflates the standard error of the OLS estimates of the regression coefficients. **(Table-3)** summarizes the comparison by presenting the ratio of the two types of standard errors.

The data used in this research was special in the sense that control was exercised by the researcher while the data was generated. It was found that the empirical study performed in this research not only confirms the theoretical results but also lets us establish a direction relation between intensity of the collinearity and the size of the standard error. A uniform deflation factor (10%) was observed throughout the study. The standard error for the intercept term is found unchanged. One hundred heteroscedastic data sets were generated with known level of heteroscedasticity and were analysed using Goldfield Quandt test. The results were summarized in **(Table-4)**. It is apparent from these results that a close relation is available in level of heteroscedaticity and value of test statistics. That is the level of heteroscedasticity and value of test statistics are statistically related.

Twenty replications of 100 samples of 1000 observations were performed for autocorrelated scenarios. That is 200 situations were analyzed by means of Durbin Watson test. The results are summarized in **(Table-5)**.

A clear relation between intensity of the first order auto correlation and the value of the test statistic has been established.

Table 5: Values of Test Statistic for 100 replications of each of the 20 Scenarios (Case of First Order Autoregressive error term)

α	3	6	9	12	15	18	21	24	27	30	3	6	9	12	15	18	21	24	27	30
β	**0.66**	**1.32**	**1.98**	**2.64**	**3.30**	**3.96**	**4.62**	**5.28**	**5.94**	**6.60**	**0.66**	**1.32**	**1.98**	**2.64**	**3.30**	**3.96**	**4.62**	**5.28**	**5.94**	**6.60**
ρ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	**-0.1**	**-0.2**	**-0.3**	**-0.4**	**-0.5**	**-0.6**	**-0.7**	**-0.8**	**-0.9**	**-1.0**
1	1.98	1.71	1.35	1.24	0.95	0.79	0.65	0.39	0.18	0.01	2.24	2.42	2.51	2.79	2.95	3.10	3.34	3.52	3.77	3.99
2	1.79	1.61	1.36	1.25	0.96	0.80	0.63	0.32	0.17	0.01	2.20	2.40	2.55	2.87	2.97	3.13	3.35	3.55	3.77	3.99
3	1.81	1.74	1.32	1.16	0.98	0.80	0.53	0.43	0.16	0.01	2.16	2.47	2.72	2.76	2.99	3.18	3.40	3.54	3.72	4.00
4	1.77	1.57	1.39	1.29	0.93	0.73	0.65	0.36	0.19	0.01	2.28	2.42	2.56	2.74	3.05	3.12	3.34	3.58	3.77	3.99
5	1.82	1.60	1.30	1.21	1.03	0.84	0.57	0.37	0.18	0.00	2.21	2.36	2.64	2.75	3.03	3.18	3.28	3.57	3.78	3.99
…	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
96	1.84	1.78	1.53	1.27	0.90	0.80	0.54	0.35	0.20	0.01	2.12	2.38	2.66	2.86	3.06	3.18	3.31	3.57	3.79	3.99
97	1.86	1.65	1.45	1.18	0.90	0.73	0.56	0.40	0.18	0.01	2.36	2.33	2.57	2.79	2.98	3.20	3.35	3.58	3.71	3.99
98	1.76	1.58	1.34	1.18	0.99	0.68	0.62	0.31	0.16	0.00	2.13	2.44	2.68	2.79	2.96	3.09	3.35	3.56	3.78	3.99
99	1.87	1.52	1.40	1.18	1.03	0.77	0.55	0.41	0.16	0.00	2.13	2.40	2.53	2.76	2.88	3.21	3.47	3.58	3.77	3.99
100	1.83	1.70	1.43	1.15	1.07	0.78	0.55	0.39	0.18	0.03	2.04	2.29	2.61	2.81	2.98	3.20	3.34	3.53	3.78	4.00

5. CONCLUSION

To conclude all of our work related to the violation of the assumptions of normality namely, multicollinearity, heteroscedastity, and autocorrelation, it is very clear that the violation of these three assumptions affects the interpretations of the results seriously. To check the violation of these three assumptions cloning of datasets is made and results are compared accordingly. It is very interesting to conclude that the generation of datasets is appeared to be very useful as we want to check accordingly. The automatic process (Minitab and MS Excel 2007 Macros) of data generation, for the testing of the violation of these assumptions and for the comparison, is adapted. After putting the desired level of multicollinearity, heteroscedastity, and autocorrelation in the datasets the comparison of the results is made which is according to the expected results. These results are compared head to dead and it is concluded that the violations of these assumptions at the desire level of multicollinearity, heteroscedastity, and autocorrelation how much affect the results and the interpretation of these results.

As a result, a relationship between the given level of multicollinearity, heteroscedastity, and autocorrelation is gauged and a serious and proper impact is gauged as well. For all the three assumptions an amount of violations at specific level is gauged. A very interesting process of data cloning is adapted to gauge the expected impact on the results which gave us the results as we want to compare and interpret.

REFERENCES:

Durbin, J., G. S. Watson, (1971). Testing for serial correlation in least squares regression. Biometrika, 58(1), 1-19.

Goldfeld, S. M., R. E. Quandt, (1972). Nonlinear Methods in Econometrics. Amsterdam: North-Holland.

Gujarati, D. N., (1978). Basic Econometrics: Mc Graw Hill.

Morgan, B. J. T., (1984). Elements of Simulation: Chapman & Hall.