DATA ANALYTICS: TOOLS AND

TECHNIQUES

Group 9

Source Code File

**List of** **Contents**

[Part 1- Issues of Multicollinearity 4](#_Toc121317325)

[Issues of Multicollinearity 4](#_Toc121317326)

[Recommended Multicollinearity Solutions 4](#_Toc121317327)

[Method 1: Sample correlations: 4](#_Toc121317328)

[Method 2: Variance Inflation Factor (VIF): 4](#_Toc121317329)

[Method 3: Tolerance: 4](#_Toc121317330)

[Part 2 – Analysis of Model 7](#_Toc121317331)

[Code: 7](#_Toc121317332)

[Findings 7](#_Toc121317333)

[Output of studentised Residuals against predicted values and 8 other explanatory variables 8](#_Toc121317334)

[Output of Normal Probability plot of the Studentized residuals: 13](#_Toc121317335)

[Output of Histogram of Studentized residuals: 14](#_Toc121317336)

[Part 3: Reduction of full regression model 15](#_Toc121317337)

[3.1. Backward Selection in Full Regression Model: 15](#_Toc121317338)

[Source Code 15](#_Toc121317339)

[Interpretation: 15](#_Toc121317340)

[Backward Elimination Results 15](#_Toc121317341)

[Forward Selection in Full Regression Model: 17](#_Toc121317342)

[Source Code 17](#_Toc121317343)

[Interpretation: 17](#_Toc121317344)

[Forward Elimination Results 17](#_Toc121317345)

[Part 4- Model Selection using the R2 / Error Mean Square Method 20](#_Toc121317346)

[Source Code 20](#_Toc121317347)

[Findings 20](#_Toc121317348)

[Part 5- Influential Point Analysis using Reduce Model 22](#_Toc121317349)

[Source Code 22](#_Toc121317350)

[Findings 22](#_Toc121317351)

[Model Analysis 23](#_Toc121317352)

[Parameter Estimates 23](#_Toc121317353)

[Temporary Dataset 23](#_Toc121317354)

[Index Creation 24](#_Toc121317355)

[Cut-off 24](#_Toc121317356)

[Plot 24](#_Toc121317357)

[Further Analysis 24](#_Toc121317358)

[Leverage H: 24](#_Toc121317359)

[Covariance Ratio 25](#_Toc121317360)

[Decision 25](#_Toc121317361)

[Part 6 – 27](#_Toc121317362)

[Source Code: 27](#_Toc121317363)

[Findings 27](#_Toc121317364)

# Part 1- Issues of Multicollinearity

## Issues of Multicollinearity

Multicollinearity is basically, it’s a problem, where several independent variables are correlated. Here we can easily see the exchange rate for United Kingdom depends upon the several depended variables which are similar like ER, Imp, Exp, GOR, MS, GNS, Bal, LTI, STI are quantitative, numeric data types.

Which is causing the following two basic types of problems:

1. The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.
2. Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of your regression model. You might not be able to trust the p-values to identify independent variables that are statistically significant.

## Recommended Multicollinearity Solutions

These problems are tackled down through following methods:

### Method 1: Sample correlations:

Sample correlations in which explanatory variables are first standardised. to measure effectively on same scale. So, that we could use the resulting regression coefficient to measure the strength of underlying relationships.

Sample correlation coefficient r known as Pearson correlation coefficient. Can be fairly calculated through formula in SAS. correlation coefficient r always lies between 1 and -1.

### Method 2: Variance Inflation Factor (VIF):

Another way to solve this problem is Variance Inflation Factor (VIF)and Tolerance. VIF is always calculated by this below formula.

VFj=1/1-Rj2

VIF is always positive and R 2 can’t exceed 1. Closer the R2 is to 1 greater the VIF will be and high relationship exist between the explanatory variables.

### Method 3: Tolerance:

Is measured through below formula:

Tolerance = 1/VIF

Higher the VIF is; lower the tolerance will be.

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

*Figure 1.1*

The correlation matrix indicates that the majority of the explanatory variable pairings have low correlation, but that some variable pairs, such as STI and LTI (0.7491), IMP and LTI (0.5914), and GNS and Bal (0.5841), have comparatively greater correlation.

High correlations exist between the explanatory variables GOR (0.9159) and GNS (-0.7074) and the response variable ER. As a result, we have an indication of the potential effectiveness of these two explanatory variables as a predictor of the response.

Table

Description automatically generated

*Figure 1.2*

No variance inflation factor (VIF) can be accurately predicted by another explanatory factor because none of them exceeds the cut off value of 10, which is 10. The hypothesised explanatory variables for the regression problem appear to have high collinearities, however there is currently little data to support this claim.

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated with medium confidence

*Figure 1.3*

The condition index in row 9 and 8 is the greatest and exceeds 30, but because it does not outperform the condition index in the row before it by a factor of three or more, we should proceed with extreme care.

Variable GNS (0.63372) has a heavier load in row 8. We disregard the second-highest load variable, Bal, because it considerably differs from the first variable.

When compared to other explanatory factors, Imp (0.83335) has the highest coefficient in row 9. We disregard this because the difference between the second highest load and the first variable is substantially higher. Hence, such a result can be effectively ignored.

Therefore, no major collinearity problem is likely to be present.

# Part 2 – Analysis of Model

## Code:

Graphical user interface, text

Description automatically generated

## Findings

F-test is valid at the significance level of 0.1%. Model is significant sufficient to interpret our explanatory variables Imp Exp GOR MS GNS Bal LTI STI against target variable ER.

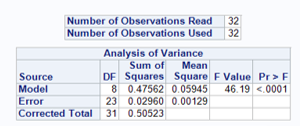


Figure 2.1 Total Observations

As further analysis R-Sq and Adj R-Sq is more than 70% Model is good in terms of explanatory variables.



Figure 2.2 R-SQ and Adj R-SQ

Furthermore, Imp Exp MS GNS LTI STI has value greater than 0.05 so these are least significant whereas GOR and Bal significant to our target variable.

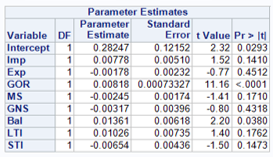


Figure 2.3 Parameter Estimates

In the first plot studentisd residuals appear to be randomly scatter about the mean value of zero a. have approximate constant range across the entire range of fitted values. There the plot is consistent with the adequacy the systematic component of multiple regression model.

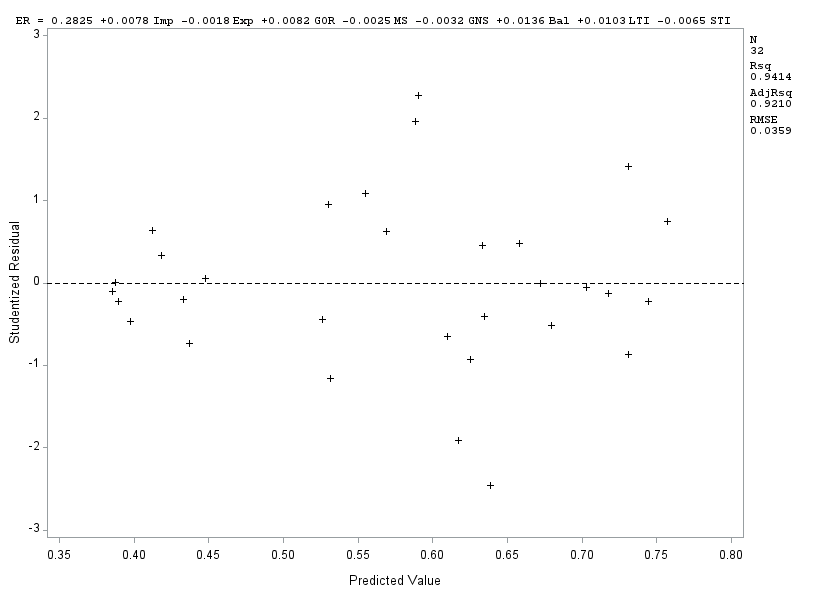


Figure 2. 4 Plot agaisnt Predicted Values and studentised Residuals

Same applies for the GOR and Bal plots; which has constant variance so plot has more likely a linear relationship for response variable ER.

Whereas, plots for the Imp Exp MS GNS LTI and STI have less likely to be randomly scatted against the studentised residual and has less spread so, plot has approximate linear relationship between response variable ER.

## Output of studentised Residuals against predicted values and 8 other explanatory variables

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart

Description automatically generatedChart, scatter chart

Description automatically generatedCalendar

Description automatically generated with low confidence

### Output of Normal Probability plot of the Studentized residuals:

Chart, scatter chart

Description automatically generated

### Output of Histogram of Studentized residuals:

Chart, histogram

Description automatically generated

# Part 3: Reduction of full regression model

## 3.1. Backward Selection in Full Regression Model:

### Source Code

**proc** **reg** data=mydata. exchange;

model ER = Imp Exp GOR MS GNS Bal LTI STI / selection = backward slstay=**0.05**;

**run**;

**quit**;

### Interpretation:

Following is the out of above proc reg procedure.

Variable with highest p value will be removed first so, consequently Exp (P value =0.45), GNS (p value =0.42), STI (p-value 0.1867), LTI (0.23) has been deleted sequentially due to highest p value noted during backward selection.

Table

Description automatically generated

Figure 3.1. Summary of Backward Selection

### Backward Elimination Results

SAS has correctly identified the surviving explanatory variables IMP, GOR, MS, and Bal for our final model using backward selection.

Table

Description automatically generated

Figure 3.2. 1st table of Backward Selection

Table

Description automatically generated

Figure 3.3. 2nd table of Backward Selection

Table

Description automatically generated

Figure 3.4. 3rd table of Backward Selection

Table

Description automatically generated

Figure 3.5. 4th table of Backward Selection

Table, Excel

Description automatically generated

Figure 3.6. Final table of Backward Selection

Our corresponding fully fitted multiple regression equation can be drawn through our potential explanatory variables is below

**ER=0.15099+ 0.01232 IMP + 0.00839GOR -0.00288 MS+0.01726 Bal**

Our model perfectly predicts the response ER using above five explanatory variables.

## Forward Selection in Full Regression Model:

### Source Code

**proc** **reg** data=mydata. exchange;

model ER = Imp Exp GOR MS GNS Bal LTI STI / selection = forward slentry=**0.05**;

**run**;

**quit**;

### Interpretation:

First of all, our model is fitted with one of the most potential explanatory variable with the smallest p value. Here the most potential explanatory variable is GOR which is significant at 0.1% then we have LTI with p value (0.0004) and then we have MS with p value (0.0154) and STI with p value (0.0405). Hence the GOR, LTI, MS, STI has been considered as an explanatory variable in our regression model.

Table

Description automatically generated

Figure 3.7. Summary of Forward Selection

### Forward Elimination Results

Our corresponding fully fitted multiple regression equation can be drawn through our potential explanatory variables GOR, MS, LTI and STI as below

A picture containing graphical user interface

Description automatically generated

Figure 3.8. 1st table of Forward Selection

Table

Description automatically generated

Figure 3.9. 2nd table of Forward Selection

Table

Description automatically generated

Figure 3.10. 3rd table of Forward Selection

Table, Excel

Description automatically generated

Figure 3.11. Final table of Forward Selection

**ER=0.30270+ 0.00855 GOR - 0.00311 MS+ 0.02092 LTI -0.00759 STI**

Our model is able to predicts the response variable ER using above four potential explanatory variables.

* 1. No explanatory variable which was significant before the entry of new explanatory variable became unsignificant and no backward elimination sequence was initiated on variables which are already significant. If we compare fig.3.11. and fig.3.12, we can see that both the models of forward and stepwise procedure are the same. As there is no backward elimination we can safely say that forward and stepwise procedure are identical in this case.

Table

Description automatically generated

Figure 3.12. Final table of Stepwise Selection

* 1. If we compare forward selection with backward selection then we can clearly say that GOR, MS and LTI are the most potential explanatory variables for fullest multiple regression model. As there is no collinearity between

# Part 4- Model Selection using the R2 / Error Mean Square Method

## Source Code

**proc** **reg** data=mydata.exchange;

model ER = Imp Exp GOR MS GNS Bal LTI STI / selection = rsquare mse;

**run**;

**quit**;

## Findings

The below output will help in analysing, producing and comparing a list of models in terms of list

of explanatory variables.

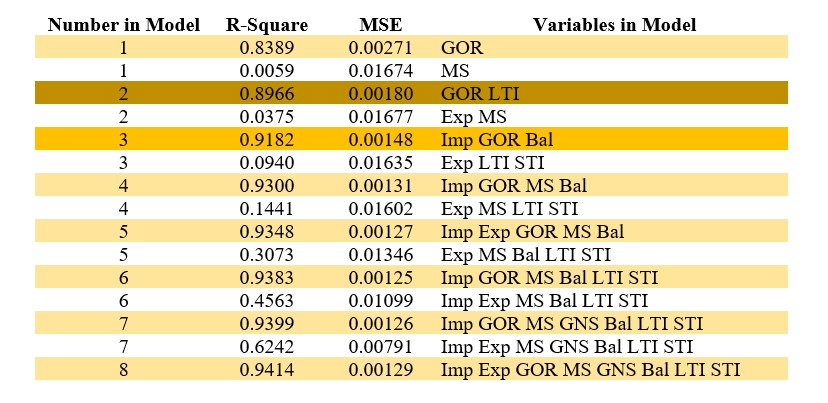


Figure 4.1 List of Model

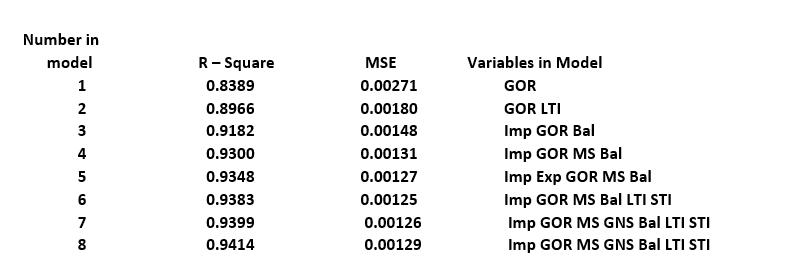


Figure 4.2 Best Model

The best model from the all given k model is the one which is at the top of group which highlighted by yellow colour. We choose the bivariate models when the model contains k=2 and explanatory variables are GOR and LTI with the R2 (0.8966) and MSE (0.00180). This model is chosen by the both backward and stepwise elimination.

This model can be further re fitted using proc reg without selection = option in order to obtain further relevant information, in particular the parameter estimates. The fitted regression equation for the model is,

**ER= 0.15099+0.01232 imp +0.00839 GOR – 0.00288 MS +0.01726 Bal**

With R Sq=0.9300 the model perfectly predicts the target variable ER with the above four explanatory variables. The variation coefficient R Sq reflects the percentage of the response variable's overall variation that the specific multiple regression model accounts for. However, it is well known that R Sq will only grow by chance when the model's explanatory variable count, k, rises.

An improved model fit is shown by a higher R squared, whereas a poor statistical structure in the data is indicated by a lower R squared.

However, R Sq will give a fair comparison between models of the same size, those with larger R2 measures being considered “better “.

The idea here is first to use R2 to find the “best” model of each possible size, and then to use the error mean square associated with each of these “best” models to choose the most parsimonious of them.

# Part 5- Influential Point Analysis using Reduce Model

## Source Code

**proc** **reg** data = mydata.exchange noprint;

model ER=Imp GOR MS Bal;

output out = INFL3 p = Pred dffits = Dff;

**run**;

**quit**;

**data** INFL3; set INFL3;

Adff = abs(Dff);

**run**;

**proc** **sort** data = INFL3;

by descending Adff;

**run**;

**data** INFL3; set INFL3;

Index = \_N\_;

**run**;

**proc** **gplot** data = INFL3;

plot ADff\*Index / vref = **0.7905** vref = **2**;

**run**;

**quit**;

/\* Further Diagnostics for Potential Influential Points \*/

**proc** **reg** data = INFL3;

model ER=Imp GOR MS Bal;

output out = INFL4 h = H rstudent = Dresid covratio = C;

**run**;

**quit**;

**proc** **print** data = INFL4;

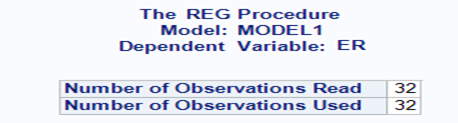
var ID ER Dff H Dresid C;

where Index <= **3**;

**run**;

## Findings

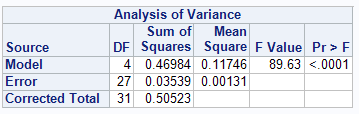
The first proc reg will produce the following results for the dependent variable ER. 32 observations are used for our analysis.



*Figure 5.1 Total Observations*

### Model Analysis

Here our model has p –value for the F Test Variance analysis (0.0001) which means our model is significant at 0.1% level with DF equal to 4 against our explanatory variables.



*Figure 5.2 Variance Analysis*

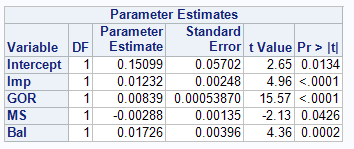
Moreover, R2 and Adj R2 > 90 %; so our model better fits the regressions of the response variable ER against our explanatory variables.



*Figure 5.3 RSE and RSq*

### Parameter Estimates

By looking at parameter estimates it is clearly seen Imp, GOR, MS, and Bal have p-value < 0.05 which is significant.



*Figure 5.4 Parameter Estimates*

**But in this question, we are only looking for influential points.**

### Temporary Dataset

Temporary dataset with the name “INFL3” has been created containing DFFITS in the variable DFF, and Pred (predicted values), with the explanatory variables.

### Index Creation

A variable Adff is created for storing absolute DFFITS values. Then index is created for the N = 32 total number of observations. Finally, we perform sorting in descending order by Adff, so we can sort higher values in our beginning indexes.

### Cut-off

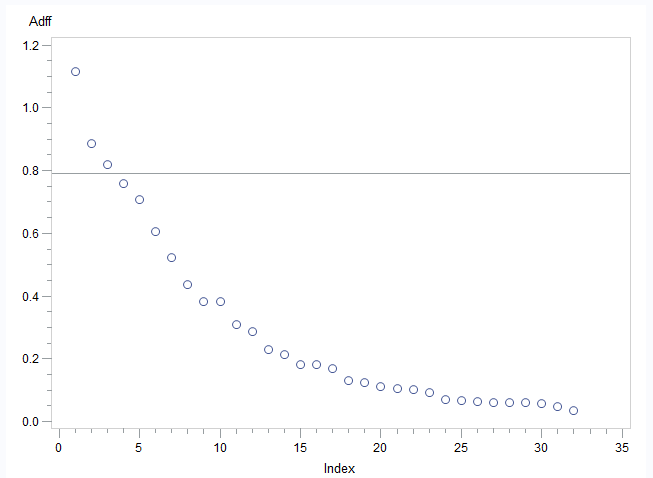
As we have N = 32 observations and K = 4, explanatory variables and p is fitted parameters therefore P = K + 1 = 5. So cut-off for

DFFITS= = 2 \* √(5/32)= 0.79052

and our absolute DFFITS value is 2.

### Plot

Our plot show that the three left most points which are above than our cut-off (0.79052) for DFFITS. Both these points lie above 0.79052 but does not exceeds 2, which is potential influential points but we do not have marginal evidence for this. So we do further analysis.



*Figure 5.5 Plot Against Index and Adff*

### Further Analysis

Analyzing further for the evidence of potential influential points the proc reg procedure is used with leverage H, deleted residuals, and covariance ratio.

There is no option in proc reg to look at DFBETAS but it can be looked with the \ influence . But here it is not required to analyze our potential influential points through DFBETAS. But the most three most influential observations are here.

### Leverage H:

Average leverage is H = p/n

P= K+1

P= 4+1=5

N =32

Average H =p/n= 0.15625

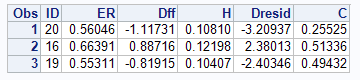
Cutt off=3\*(p/n) = 0.46875

### Covariance Ratio

Upper limit C= 1+ = 1+0.46875=1.46875

Lower limit C= 1- = 1-0.46875=0.53125

So , proc print step will print the additional influence statistics.



*Figure 5.6 Influential Observations*

The observation 1, ID is equal to 20 has average leverage H = 0.10810 which is moderately below than average leverage H = 0.15625 and covratio (0.25525) is below the expected limit (0.53125 to 1.46875) and less than 1. Tut the deleted residual is greater than 2. So, the inclusion of this observation reduces the fitted regression model equation. On further investigation, this point certainly causes concern.

Similarly, the observation 2, ID is equal to 16 has average leverage H = 0.12198 which is moderately lower than average leverage H = 0.15625 and covratio (0.51336) is just slightly below the limit (0.53125 to 1.46875) and less than 1. The deleted residual is greater than 2. So, the inclusion of this observation reduces the fitted regression model equation. On further investigation, this point certainly causes concern.

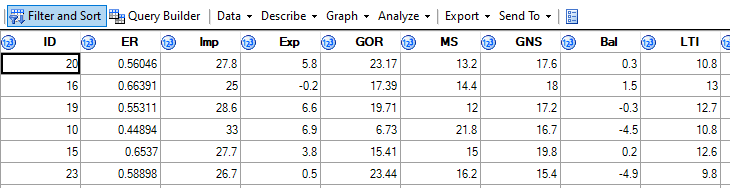
Similarly, the observation 3, ID is equal to 19 has average leverage H = 0.10407 which is moderately lower than average leverage H =0.15625 and covratio (0.49432) is just below the limit (0.53125 to 1.46875) and less than 1. The deleted residual is greater than 2. So, the inclusion of this observation reduces the fitted regression model equation. On further investigation, this point certainly causes concern.

### Decision

In conclusion, the three observations have deleted residual exceeding 2 which shows some abnormality, covariance ratio is below the limit but less than one and Leverage H is below than average leverage. Three potential influence point were identified none is sufficiently abnormal to given much cause for concern. The multiple regression analysis should therefore proceed retaining these three points.

So, the inclusion of all three observation reduces the fitted regression model equation. On further investigation, this point certainly causes concern.

Its better to exclude them from the regression and apply the model only to observations having less precisions.



*Figure 5.6 Influential Observations from Data Set*

# Part 6 –

## Source Code:

**data** EXCH;

Imp = **16**; Exp = **4.6**; GOR = **50**; MS = **22**; GNS = **18**; Bal = -**10**; LTI = **10**; STI = **10**;

**run**;

**data** EXCH;

set mydata. exchange EXCH;

**run**;

/\*Reduced model for our Explanatory variable GOR, Bal \*/

**proc** **reg** data = EXCH noprint;

model ER = Imp GOR MS Bal;

output out = DataSet2 p = Pred h = H lcl = Lower ucl = Upper;

**run**;

**quit**;

**proc** **print** data =work.dataset2;

var Imp GOR MS Bal H Pred Lower Upper;

where ER =.;

**run**;

## Findings

Output of the above code is as below

A picture containing table

Description automatically generated

Figure 6.1 New Observation

First two data set lines are adding values for the above eight variables and whom predictions are added into EXCH data set.

The proc reg procedure is now follows and model is fitted for the target variable ER against the explanatory variables which were reduced in backward elimination step.

Temporary dataset DataSet2 contains all the information related to predicted residuals, leverage h, and at upper and lower confidence limit of 95%.

Current model has K = 4 explanatory variables   
so,  
Parameters P =K +1 =5  
Now current data after the addition of a new observation is N=33 observations.  
The average leverage H for this data is

*Averag*e = 5/ 33 = 0.1515

= 0.4545

Analyzing the 33th observation it has leverage (2.03413) which is greater than average leverage point so model will not produce reliable predictions for this observation at the 95% confidence interval.

The predicted exchange rate for the pred 0.53162 with associated 95% prediction limit of 0.40223 to 0.66100.

As the exchange rate for our 33th observation is not reliable so we can re-run the code by adding more values and perform further analysis.