

**ANL302**

**Selected Topics in Regression**

**Group Based Assignment 1**

**July 2017 Presentation**

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| Gareth Kwan Jian Hong | K1671800 |
| Mohamad Faisal Bin Maun | Z1511649 |
| Lim Hui Qin | Z1610579 |

Contents

[Question 1 3](#_Toc491202119)

[Part (a) 3](#_Toc491202120)

[Part (b) 4](#_Toc491202121)

[Question 2 6](#_Toc491202122)

[Part (a) 6](#_Toc491202123)

[Part (b) 7](#_Toc491202124)

[Question 3 9](#_Toc491202125)

[Part (a) 9](#_Toc491202126)

[Part (b) 10](#_Toc491202127)

[Part (c) 10](#_Toc491202128)

[Part (d) 10](#_Toc491202129)

[Question 4 11](#_Toc491202130)

[Part (a) 11](#_Toc491202131)

[Part (b) 12](#_Toc491202132)

[Part (c) 13](#_Toc491202133)

# Question 1

## Part (a)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | | | |  | | |  | | |  | | |  | | |  | |  | | |  |
|  |  | | |  | | |  | | |  | | |  | | |  | |  | | |  |
| *Regression Statistics* | | |  | | |  | | |  | | |  | | |  | | | |  | | |
| Multiple R | 0.9979026 |  | | |  | | |  | | |  | | |  | | |  | | |  | |
| R Square | 0.9958096 |  | | |  | | |  | | |  | | |  | | |  | | |  | |
| Adjusted R2 | 0.9952381 |  | | |  | | |  | | |  | | |  | | |  | | |  | |
| Standard Error | 692.46733 |  | | |  | | |  | | |  | | |  | | |  | | |  | |
| Observations | 26 |  | | |  | | |  | | |  | | |  | | |  | | |  | |
|  |  |  | | |  | | |  | | |  | | |  | | |  | | |  | |
| ANOVA |  |  | | |  | | |  | | |  | | |  | | |  | | |  | |
|  | *df* | *SS* | | | *MS* | | | *F* | | | *Significance F* | | |  | | |  | | |  | |
| Regression | 3 | 2.507E+09 | | | 835635993 | | | 1742.6837 | | | 2.70E-26 | | |  | | |  | | |  | |
| Residual | 22 | 10549242 | | | 479511 | | |  | | |  | | |  | | |  | | |  | |
| Total | 25 | 2.517E+09 | | |  | | |  | | |  | | |  | | |  | | |  | |
|  |  |  | | |  | | |  | | |  | | |  | | |  | | |  | |
|  | *Coefficients* | *Standard Error* | | | *t Stat* | | | *P-value* | | | *Lower 95%* | | | *Upper 95%* | | | *Lower 95.0%* | | | *Upper 95.0%* | |
| Intercept | 4999.0991 | 1193.7841 | | | 4.1876 | | | 0.0004 | | | 2523.3423 | | | 7474.8558 | | | 2523.3423 | | | 7474.8558 | |
| X Variable 1 | 0.5386 | 0.0179 | | | 30.1472 | | | 0.0000 | | | 0.5016 | | | 0.5757 | | | 0.5016 | | | 0.5757 | |
| X Variable 2 | -80.8757 | 106.2054 | | | -0.7615 | | | 0.4544 | | | -301.1321 | | | 139.3807 | | | -301.1321 | | | 139.3807 | |
| X Variable 3 | 15.8543 | 4.5147 | | | 3.5117 | | | 0.0020 | | | 6.4914 | | | 25.2172 | | | 6.4914 | | | 25.2172 | |

Ct = 4999.0991 + 0.5386Yt – 80.8757Rt + 15.8543Pt

The intercept is 4999.0991.

The slope coefficient for Income Yt is 0.5386. It means that in the event where the income goes up by $1, while the Interest Rate and Price Index remains constant, consumption of the country is expected to increase by 0.5386. This is consistent with expectations as when Income increases, the spending power goes up too, thus pushing the consumption figure upwards.

The slope coefficient for Interest Rate Rt is -80.8757. It implies that in the event where the Interest Rate goes up by 1%, while the Income and Price Index remains constant, consumption of the country is expected to decrease by 80.8757. This is consistent with expectations as when Interest Rate goes up, it meant that the repayment amount on loans will increased, thus pushing the consumption figure to go down.

The slope coefficient for Price Index Pt is 15.8543. It shows that in the event where Price Index goes up by 1, while Income and Interest Rate remains constant, consumption of the country is expected to increase by 15.8543. This is consistent with expectations as when Price Index goes up, it generally means that prices of items will be increased, thus causing the consumption goes up as well.

At 5% level of significance, only Income and Price Index is statistically significant. P value for Income is 2.1852E-19 while P value for Price Index is 0.00196805 which is lower than 0.05.

The R square of the model is 0.9958 meaning that 99% of the variation in consumption can be explained by the variation in Income, Interest Rate and Price Index in the period.

## Part (b)

From the residual plot against time period, there is an indication that the residual plot follows a cyclical pattern. Having a cyclical pattern in the residual plot suggests that there could be Autocorrelation existing in the regression model as the residual in each time period seems to be linked to each other.

Having Autocorrelation existing in the model will cause misleading conclusions towards the identified variables as the R2 in regression model is inflated towards a high percentage of almost 99%. Hence giving the impression that the Income, Interest Rate, and Price Index is enough to explain the change in Consumption, as well as the significance of each identified variable could also be possibly inflated.

As the data is of time series nature, we can further test and confirm the existence of Autocorrelation by doing a Durbin-Watson d Test on the residuals.

Using the formula



We are able to compute the following value of 0.770999 for the Durbin Watson d Test. At k’ = 3 with n = 26, at 5% level, we will have the dL = 1.143 and dU = 1.652. Since the Durbin Watson d statistics computed above is 0.770999 which is lower than the dL of 1.143 and greater than 0, therefore we reject the null hypothesis that there is no positive Autocorrelation, in another words, we are able to accept that autocorrelation is present in the regression model.

# Question 2

## Part (a)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SUMMARY OUTPUT – Unrestricted Model** | | | | | |  | | |  | | |  | | |  |
|  |  | |  | | |  | | |  | | |  | | |  |
| *Regression Statistics* | | | |  | | |  | | |  | | |  | | |
| Multiple R | 0.997902583 |  | | |  | | |  | | |  | | |  | |
| R Square | 0.995809565 |  | | |  | | |  | | |  | | |  | |
| Adjusted R2 | 0.995238142 |  | | |  | | |  | | |  | | |  | |
| Standard Error | 692.4673269 |  | | |  | | |  | | |  | | |  | |
| Observations | 26 |  | | |  | | |  | | |  | | |  | |
|  |  |  | | |  | | |  | | |  | | |  | |
| ANOVA |  |  | | |  | | |  | | |  | | |  | |
|  | *df* | *SS* | | | *MS* | | | *F* | | | *Significance F* | | |  | |
| Regression | 3 | 2506907980 | | | 835635993.4 | | | 1742.683683 | | | 2.70108E-26 | | |  | |
| Residual | 22 | 10549241.98 | | | 479510.9989 | | |  | | |  | | |  | |
| Total | 25 | 2517457222 | | |  | | |  | | |  | | |  | |
|  |  |  | | |  | | |  | | |  | | |  | |
|  | *Coefficients* | *Standard Error* | | | *t Stat* | | | *P-value* | | | *Lower 95%* | | | *Upper 95%* | |
| Intercept | 4999.0991 | 1193.7841 | | | 4.1876 | | | 0.0004 | | | 2523.3423 | | | 7474.8558 | |
| Income | 0.5386 | 0.0179 | | | 30.1472 | | | 0.0000 | | | 0.5016 | | | 0.5757 | |
| Interest Rate | -80.8757 | 106.2054 | | | -0.7615 | | | 0.4544 | | | -301.1321 | | | 139.3807 | |
| Price Index | 15.8543 | 4.5147 | | | 3.5117 | | | 0.0020 | | | 6.4914 | | | 25.2172 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SUMMARY OUTPUT – Restricted Model** | | | | | |  | | |  | |  | |  | | |
|  |  | |  | | |  | | |  | |  | |  | | |
| *Regression Statistics* | | | |  | | |  | | |  | | | |  | |
| Multiple R | 0.9964533 |  | | |  | | |  | | | |  | | |  |
| R Square | 0.9929192 |  | | |  | | |  | | | |  | | |  |
| Adjusted R2 | 0.9926242 |  | | |  | | |  | | | |  | | |  |
| Standard Error | 861.8182905 |  | | |  | | |  | | | |  | | |  |
| Observations | 26 |  | | |  | | |  | | | |  | | |  |
|  |  |  | | |  | | |  | | | |  | | |  |
| ANOVA |  |  | | |  | | |  | | | |  | | |  |
|  | *df* | *SS* | | | *MS* | | | *F* | | | | *Significance F* | | |  |
| Regression | 1 | 2499631683.66 | | | 2499631683.66 | | | 3365.46 | | | | 2.56864E-27 | | |  |
| Residual | 24 | 17825538.38 | | | 742730.77 | | |  | | | |  | | |  |
| Total | 25 | 2517457222.04 | | |  | | |  | | | |  | | |  |
|  |  |  | | |  | | |  | | | |  | | |  |
|  | *Coefficients* | *Standard Error* | | | *t Stat* | | | *P-value* | | | | *Lower 95%* | | | *Upper 95%* |
| Intercept | 1517.1376 | 963.1310 | | | 1.5752 | | | 0.1283 | | | | -470.6670 | | | 3504.9422 |
| Income | 0.5924 | 0.0102 | | | 58.0126 | | | 0.0000 | | | | 0.5714 | | | 0.6135 |

Use restricted least-square. H0: β2 = β3 = 0 and H1: Otherwise.



The critical value is F(0.05, 2, 22) = 3.44. Reject H1. Rt and Pt should be excluded in the model.

Testing H0: 2 = 0 against H1: 2 ≠ 0 at 5% level of significance, reject H0 if │t │> t 0.025, n-k

| t | = | 0.7615 |

t0.025, 22 = 2.074

Do not reject H0. Minimum lending rate (Rt) is an insignificant indicator of consumption, hence it should be excluded in the regression model.

Testing H0: 3 = 0 against H1: 3 ≠ 0 at 5% level of significance, reject H0 if │t │> t 0.025, n-k

| t | = | 3.5117 |

t0.025, 22 = 2.074

Reject H0. Expenditure deflator (Pt) is a significant indicator of consumption, hence it should be included in the regression model.

## Part (b)

Model 1: Ct = β0 + β1Yt + β2Rt + β3Pt + ut

Model 2: Ct = β0 + β1Yt + ut

Based on Model 1, it is concluded that only National Income (Yt) and Expenditure deflator (Pt) are significant variables at 5% level of significance. The restricted F test performed on Model 2, it shows that having the variable National Income (Yt) alone in the model is adequate. However upon performing t test on individual coefficients, it is noted that Consumption expenditure deflator (Pt) is a significant indicator while Minimum lending rate (Rt) is insignificant.

As such, to improve the model, it is recommended to include National Income (Yt) and Expenditure deflator (Pt) in the final model and exclude Minimum lending rate (Rt). The t test performed on individual coefficients substantiate the significance of having Expenditure deflator (Pt) in the model.

Final Model: Ct = β0 + β1Yt + β3Pt +ut

From the perspective of consumption theory, it is logical to explain consumption based on income and expenditure. Income is an indication of spending power while expenditure is an action of consumers. The sign of the two variables are also in synced with expectation. It is a logical understanding that higher income and higher expenditure leads to greater consumption. Income and expenditure are positively correlated to expenditure, hence they make good estimator to determine consumption.

# Question 3

OLS method was performed as required for

lnWage = 1 + 2Age + 3DE2 + 4DE3 + 5DE4 + 6DPT+ 7DGender +ui

which the following equation and output were derived:

lnWage = 3.621 + 0.029Age + 0.157DE2 + 0.505DE3 + 0.622DE4 + 0.319DPT– 0.667DGender

The Significance F figure is less than 0.05 which suggests that the model is statistically significant or reliable. Also, the R Square and Adjusted R Square figures return a figure close to 0.5 which indicates that the model is 50% representative of the population it represents.

## Part (a)

Education is represented by variables: DE2, DE3 & DE4.

DE2 is statistically insignificant (P value = 0.342, which is >0.05) and have a possibility of zero value (Lower 95% = -0.169, Upper 95% = 0.485). This means wage is most likely not to be affected by those possessing a primary education.

DE3, on the other hand, is statistically significant (P value = 0.002, which is <0.05) and has positive lower and upper 95% values which are logical (being significant, DE3 cannot be zero in value). This means wage is most likely to be affected by those possessing a secondary education.

DE4, on the other hand, is also statistically significant (P value = 0.0214, which is <0.05) and has positive lower and upper 95% values which are logical (being significant, DE3 cannot be zero in value). This means wage is also most likely to be affected by those possessing a higher than secondary education.

In conclusion, the regression suggests that education (secondary and higher) is important in influencing wage.

## Part (b)

Education: DE3 and DE4 are important in influencing wage (being statistically significant).

From the regression results, the coefficients of DE3 & DE4 are both positive (+ 0.505DE3 + 0.622DE4) & exponentially linear, and therefore, based on education, the model does not suggest any diminishing return.

## Part (c)

From the regression results, DGender has a negative coefficient which may suggest that a Male worker will earn lesser than a Female worker, leading to an interpretation that there is discrimination in wage favouring Female workers over Male workers.

However, since this variable is statistically insignificant (P value = 1.031, > 0.05), any interpretation of wage discrimination concerning gender is non-consequential, and may also be misleading.

## Part (d)

From the regression results, DPT is statistically significant (P value = 0.012, which is <0.05) and has positive lower and upper 95% values which is logical, as a progressive career (through permanent jobs) would provide for salary appreciation as compared to temporary jobs which are usually remunerated based on Minimum Wage (in some countries) or low entry-level salary without substantial opportunity for appreciation.

This means wage is most likely to be influenced by having a permanent job, and it is, therefore, better for a person to have a permanent job so that he can expect better wage.

# Question 4

## Part (a)

If the following is the correct model:

Model 1 -- *Ct=0 + 1Yt + 2Pt + εt*

Where:

*Ct* = consumption (consumption) at period t

*Yt* = national income (income) at period t

*Ct* = consumption expenditure deflator (price index) at period t

However, if a researcher is to run the following model:

Model 2 -- *Ct=0 + 1Yt + 2Rt + 3Pt + εt*

Where there is an additional variable:

*Rt* = minimum lending rate (interest rate) at period t

In other words, the researcher has decided to add a new variable - Rt , into Model 1 to derive Model 2. Since it has been determined that Model 1 is the correct model, any additional variable is deemed unnecessary.

By adding an unnecessary/irrelevant variable(s), it would render the model to be problematic (model specification error) as the model is now overfitted. This departs from the primary goal of attaining a model which is a best/good/right fit. If the researcher continues to rely on Model 2, then the following consequences shall ensue:

1. The model will produce results which are still unbiased and consistent.

This suggests that Model 2 will produce estimates similar to those produced by the coefficients of Model 1.

1. The error variance is correctly estimated.

This suggests that Model 2 will result in the correct error variance as Model 1, which indicates how far each variable varies around the mean.

1. Coefficients under Model 2 are deemed to be inefficient.

However, by adding an unnecessary/irrelevant variable(s), this will result in coefficients in Model 2 to have larger variances which will render any probability inference about the parameters to become less precise.

1. Confidence intervals and hypothesis-testing procedures remain valid.

Despite this, Model 2 can still be relied on in terms of confidence intervals and how it can be used in hypothesis-testing situations, similar to Model 1.

In conclusion, it is not encouraged to use Model 2 unless the researcher has a good reason to include those additional variables for the purposes intended. The researcher should bear in mind the above consequences when interpreting any outcome based on Model 2 vis-à-vis Model 1.

## Part (b)

If the researcher runs the regression model as Ct = β0 + β1Yt + εt, the researcher would have run into the problem of Underfitting. Underfitting occurs when a relevant variable to the model is being excluded from the model. If the missing variable is significantly related to the variations of Consumption, then the error term of the regression will include the value linked to the missing variable. When a residual plot is plotted using the residuals value, there will be a systematic pattern shown in the plot, giving an indication that an important variable that can influence the value of the dependent variable is present inside the error term.

As a result, the error variance and variance of the model to predict the value of Consumption will be huge and there will be biased and inconsistent coefficients linked to the remaining intercept and explanatory variable of Yt which is the national income at period t. With the presence of biased and inconsistent coefficients in the event where the omitted variable is correlated to the included variable, the model will be highly inaccurate even as the sample size increases. This is due to the inability to correctly measure the influence of national income towards the Consumption in a period t.

In addition, computation of the standard error of the remaining regression coefficient will have an incorrect figure. When performing hypothesis testing, there will be misleading conclusions as the t statistic and f statistics of the regression model will be wrong. Having misleading conclusions will be detrimental to any business decisions made based on the model, as the impact of any change in the remaining explanatory variables is likely to be wrongly predicted.

## Part (c)

Applying Ramsey Reset Test,

R2OLD = 0.9929

R2NEW = 0.9998

Number of new regressors = 2

Sample size n = 26

Parameters in the new model = 4

Thus we have 

The critical F value at 5% level is F(2,22), 0.05 = 3.44. Since the computed F value exceeds the critical F value, we can conclude that the model is mis-specified.

Using the residuals, we can compute the Durbin Watson statistics as follows:



At k' = 3 with n = 26, at 5% level, we have dL = 1.143 and dU = 1.652

Since the DW d statistics is less than dL, we conclude that there could be first order positive autocorrelation or specification error. The model is inadequate.