ECONOMETRICS ASSIGNMENT

Estimation of Average Annual Per Capita GDP growth rate in the last 25 years of Singapore

TEAM NO.:15

TEAM MEMBERS:
Dilipkumaar
EE/2023-25/008
Harish Chander P
EE/2023-25/011
Vasanth D
EE/2023-25/026
Ackshaiyaa V
EE/2023-25/001

REFERENCES

https://databank.worldbank.org/reports.aspx?source=world-development-indicators

https://www.econometrics-with-r.org/index.html

Topic: Estimation of Growth rate in per capita GDP using 15 years of data for a country.

Country: Singapore

INTRODUCTION:

This assignment is based on the data from World Development Indicators as provided by the World Bank of Singapore's Per Capita GDP(PCGDP) for the years from 1998 to 2023. Therefore, the data for 25 years of Singapore's PCGDP has been utilized for our econometric analysis using various statistical measures throughout the assignment.

REASON FOR SINGAPORE:

Every country has its own pros and cons in terms of its development in GDP. In this matter, Singapore stands unique in demographic and geographic characteristics and still being able to do well in a market as a whole:

For an econometric analysis we need a country who have seen it all, that is well diversified in all aspects, by which our inference can be made strong from the model. The certain diversified aspects include:

- An increase in GDP were primarily from strong export-oriented economy and at the same time it also prioritized strengths in finance, technology, and manufacturing can contribute to GDP growth, as it ensures a broad base of economic activities which many countries struggled.
- As our data is concerned with PCGDP, there are income inequality problems faced by the economy as a whole which is compensated by high cost of living bringing up equal proportion in the economy with its demographic characteristic.

(1) MODEL SPECIFICATION: Show how the residual term is added to the mathematical expression for growth rate estimation and then specified as a linear econometric model. The econometric model will enable us to estimate the (average) annual growth rate of per capita GDP in the last 25 years of that country. In the answer discuss the linearity assumption and the difference between population and sample regression functions.

Mathematical Model:

A time series dataset, can be mathematically modelled as

$$Y_t = Y_0(1+g)^t , \qquad t=(1,2,\ldots,25)$$

Y = Per Capita GDP

g = average annual growth rate

we can derive the deterministic part of the model from the mathematical model in the functional form of 'log – linear'.

$$\boldsymbol{lnY_t = lnY_0 + tln(1+g)}$$

where by taking $\alpha = \ln Y_0$; $\beta = \ln(1+g)$

$$\boldsymbol{lnY_t = \alpha + \beta t}$$

Average Annual Growth can be calculated by,

$$\beta = ln(1+g)$$

$$e^{\beta} = 1 + g$$

$$\boldsymbol{g = e^{\beta} - 1}$$

Econometric Model:

The econometric model can be constructed by adding error term to the above deterministic part constructed from the mathematical model.

$$\boldsymbol{Y_t^* = \alpha + \beta t + u_t} \qquad \textbf{for t = 1,\ldots,25}$$

Here u_t is the error term which denotes reasons other than the variables taken for the model that influence GDP. It can be political, geographical, natural, social problems.

Linearity Assumption:

Linearity assumption and the difference between population and sample regression functions:

(i) linearity of the relationship between dependent and independent variables:

(a) The expected value of the dependent variable is a straight-line function of each independent variable, holding the others fixed.

(b) The slope of that line does not depend on the values of the other variables.

The Population Regression Function (PRF) and the Sample Regression Function (SRF) are both derived from the linear regression model: $Y_i = \alpha + \beta X_i + u_i$

SRF: Satisfying the linearity assumption, the SRF constitutes for the whole model including the residuals with respect to the relationship with the years (X_i)

PRF: The PRF is derived from the assumption of $E(Y_i|X_i)$, which is the average of relationship between Y_i and X_i which can be derived using expression: $E(\alpha+\beta X_i+u_i|X_i)$ which yields us :

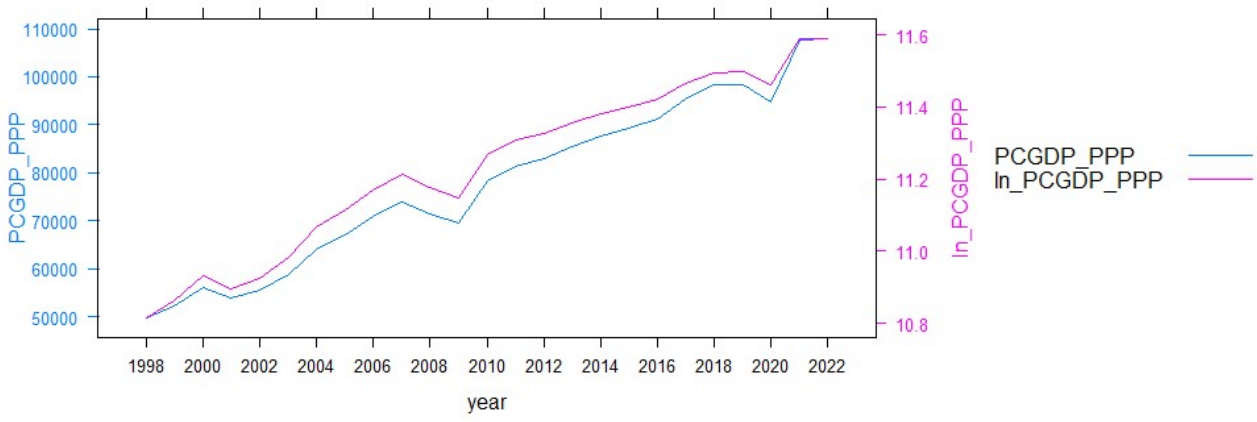
$\alpha + \beta X_i + E(u_i|X_i)$ from this we know that $E(u_i|X_i) = 0$.

The Difference: The PRF is the deterministic part of the model accounting for only the changes irrespective of stochastic part (U_i) of the model while SRF accounts also for the factors other than X_i responsible for the changes to Y_i.

(2) ASSESSING LINEARITY (visually):

(a) Plot a line graph of Y and lnY against Year (X-axis) on the same graph by choosing left vertical-axis for Y and right vertical-axis for lnY. Make sure that the axis are properly labelled and a title (as you see in the textbook); refer this as Figure 1.

Figure1: PCGDP & ln_PCGDP over years



(b) Discuss the pattern that is whether both graphs are linear or near linear? Do the two graphs have the same pattern? If so, why should that happen? and if not, what could be the reason for a different pattern?

By looking at the above result, we can say that both the per capita GDP in real terms and the log value of per capita GDP against years, deploy a linear graph and following a same pattern by the assumption of Normality as the log value of per capita GDP is the suppressed form of per capita GDP.

If the graph doesn't represent same pattern, this can be due to the violations of CLRM assumptions:

i. Homoscedasticity assumption would have been violated, as the variations of the error term from the independent variable (X_i) cannot be a constant that affects the dependant variable (Y_i)
ii. The model itself could have been a non-linear one in parameters α and β
iii. The model structure itself could have been a non-linear one (quadratic fn.)

(3) ESTIMATION:

(a) Read the J&D example and understand how to generate the time trend variable.
Note that the data set you are working with has year as the X variable and it has to be converted into a time trend variable so that the independent variable X is now transformed to takes on values 1 to 25.

TABLE 1: Per Capita GDP of Singapore(1998-2022)

PCGDP	Time Trend	ln_PCGDP	t	t^2	t*ln(y)
49789	1	10.81555	-12	144	-129.787
52217.16	2	10.86317	-11	121	-119.495
55959.03	3	10.93238	-10	100	-109.324
53886.5	4	10.89464	-9	81	-98.0517
55491.9	5	10.92399	-8	64	-87.3919
58877.61	6	10.98322	-7	49	-76.8825
63924.72	7	11.06546	-6	36	-66.3928
67039.2	8	11.11303	-5	25	-55.5652
70825.8	9	11.16798	-4	16	-44.6719
74064.65	10	11.21269	-3	9	-33.6381
71534.98	11	11.17794	-2	4	-22.3559
69498.54	12	11.14906	-1	1	-11.1491
78191.78	13	11.26692	0	0	0
81337.74	14	11.30637	1	1	11.30637
82886.78	15	11.32523	2	4	22.65046
85484.44	16	11.35609	3	9	34.06827
87702.52	17	11.38171	4	16	45.52682
89248.13	18	11.39918	5	25	56.99588
91270.64	19	11.42158	6	36	68.52951
95334.15	20	11.46514	7	49	80.256
98280.04	21	11.49558	8	64	91.96461
98455.33	22	11.49736	9	81	103.4762
94910.1	23	11.46069	10	100	114.6069
107741.1	24	11.58749	11	121	127.4624
108036.1	25	11.59022	12	144	139.0827
TOTAL		**280.8527**		**1300**	**41.22102**

(b) Now, irrespective of whether the graph is linear or non-linear, regress lnY on the time trend variable. Derive the estimators for the intercept and slope coefficients specifically for this model as in the example of J&D.

(c) Use the excel spreadsheet or the software based and create the Table similar to Table 2.1 page 54 in J&D. Make sure that title of the Table and the column headings are pertinent to this analysis.
(d) Report the results in equation form similar to equation 2.6 in Section 2.3 (pg.33) of Wooldridge. Interpret the slope and the growth rate coefficients as in J&D but appropriately modified for this data.

From the table available we can calculate the beta coefficient which can be used to find the estimated growth rate per year:

Slope(β):

$$\beta = (\sum t * \ln y)/(\sum t\text{^}2) = \frac{41.22102}{1300} = 0.0317$$

Intercept(α):

$$\alpha = \left(\frac{\sum ln\ y}{25}\right) - (\beta * t)$$

where t = mean of the years : 325/25 =13

$$\alpha = \frac{280.8527}{25} - (0.0317 * 13) = 10.82$$

ESTIMATED GROWTH RATE PER YEAR(g):

$$g = e^{\beta} - 1$$

$$g = e^{0.0317} - 1$$

therefore, **g = 0.03212**

In a growth rate model where **t** is the time trend and Y_t any continuous variable with non-zero values, then **β** (slope) is called as instantaneous growth rate (at a point in time) and ***g*** is called as the compound growth rate (over a period of time).

From the calculation, it is observed that with every year, the GDP increases by ***3.184%***, the compound growth rate in per capita GDP is observed to be ***3.23294%***.

The intercept (α) represents the estimated value of the per capita GDP when time trend is 0. In this case, it suggests that when time is at its initial point (year 1998) or when all other factors are absent or negligible, the per capita GDP is estimated to be around ***10.2955.***

THE ESTIMATED REGRESSION EQUATION:

ln_PCGDP = 10.82 + 0.0317 (time_trend)

(4) GOODNESS OF FIT:

(a) Refer to Table 1.6 (pg33 in J&D). Now create columns for observed lnY, fitted lnY, and the error term as the three columns based on the estimated coefficients for intercept and slope. Name this as Table 2 and give the table title as "Observed and fitted dependent variable and estimated error terms".

TABLE 2: observed, fitted dependant variable and estimated error term

ln_Y	Fit ln_Y	Error_term
10.86317	10.88531	-0.02214
10.93238	10.91702	0.01536
10.89464	10.94872	-0.05409
10.92399	10.98043	-0.05644
10.98322	11.01214	-0.02893
11.06546	11.04385	0.02161
11.11303	11.07556	0.037472
11.16798	11.10727	0.060709
11.21269	11.13898	0.073715
11.17794	11.17069	0.007254
11.14906	11.2024	-0.05334
11.26692	11.23411	0.032814
11.30637	11.26581	0.04055
11.32523	11.29752	0.027707
11.35609	11.32923	0.026857
11.38171	11.36094	0.020764
11.39918	11.39265	0.006525
11.42158	11.42436	-0.00278
11.46514	11.45607	0.009074
11.49558	11.48778	0.007798
11.49736	11.51949	-0.02213
11.46069	11.5512	-0.09051
11.58749	11.58291	0.004581
11.59022	11.61461	-0.02439
TOTAL = 270.0371	**TOTAL = 269.9991**	**TOTAL = 0.03804**

(b) Get the relevant column totals from this Table 2 and the previous Table 1 and use the decomposition of sum of squares to report the values for TSS, ESS, RSS and R^2; report the values in the word file the way it is in the J&D textbook. In a couple of sentences, explain what do you understand from R^2 value about the goodness of fit of this model?

TSS = $S_{yy} = \sum (Y_i - \bar{Y})^2$

S_{yy} = 1.344246

ESS = $\sum (\hat{Y}_i - \bar{Y})^2$ = 1.307104

RSS = TSS – ESS = 1.344246 – 1.307104

= 0.037142.

$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{0.037142}{1.34424} = 0.972384$

R^2 value suggests that the model is a 'good fit' by quantifying that the variances in PCGDP(per capita GDP) is strongly due to the variances in years by 97.24%.

(c) Based on the R2 values you can conclude that:

(i) Time trend explains large part of the variations in lnY when R2 is high OR the time trend does not explain large part of the variations in lnY when R2 is low.

(ii) There is a linear relationship between lnY and time trend when R2 is high OR linear model does not fit the data too well when lnY is regressed on time trend.

Do both these statements in (i) and (ii) make sense for this model or only one of them is meaningful. If it is only one of them then indicate, which one. Justify your choice(s) on the interpretation(s) with a reasoning in about 2-3 sentences.

(c) (i) Time trend explains large part of the variations in lnY when R^2 is high:

As years increase, the GDP also increases, but this increase is not only due to the years, but also in other factors excluding the years. So, irrespective of the R^2 value which can be high, it is safe to assume that years and GDP have a more or less a linear relationship normally.

(ii) There is a linear relationship between lnY and time trend when R2 is high:

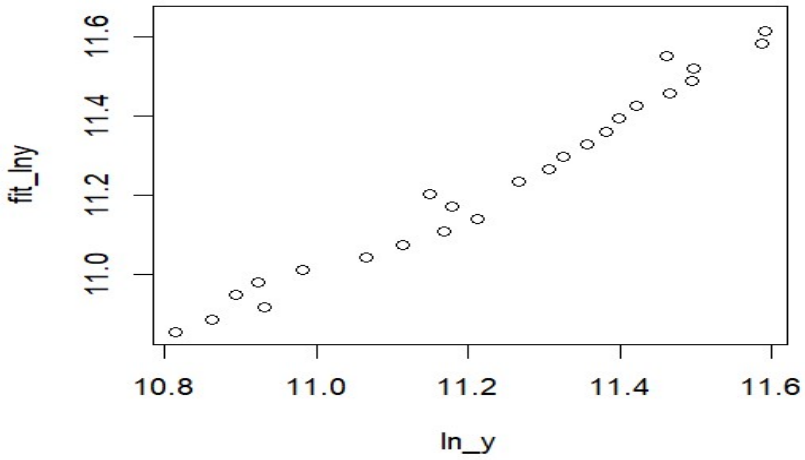
By the model specifications we can predict that the model is linear in relationship, and with R^2 value being so high by 97.23% further proves that there is a strong linear association.

(d) Plot the fitted lnY against observed lnY with appropriate labelling of the axis, a title and number this as Figure 2. Comment on the 'strength' of association by looking at the figure in terms of do you see that the 'linear association' as 'good'. Do the observed and fitted lnY have

an ‘association’ that also justifies the goodness fit value of R2? The comment should not exceed 3 sentences.

By looking at figure 2, we can infer that the model’s scatter plot is closely related in terms of the observed values and the predicted values saying that the model has predicted the values with the R^2 value being the representation of also the linearity between fit_lny and years as the obs_lny is close to fit_lny.

Figure 2: scatter plot of fit_lny over ln_y



(5) RESIDUAL ANALYSIS:

(a) Based on the data on error terms, verify the conditions arising from the two normal equations.

(a) FIRST NORMAL EQUATION:

$$\sum Y_t = n * \alpha + \beta \sum t$$

$$\sum Y_t - n * \alpha - \beta \sum t = 0$$

$$\sum (Y_t - \alpha - \beta * t) = 0$$

$$\sum (Y_t - \hat{Y}_t) = 0$$

$$\mathbf{\sum (U_t) = 0 = -2.664535e\text{-}14}$$

- This can be $\sum(Ut) \approx 0$, as the CLRM assumption suggests that the Expected value of the error term is 0.

(b) SECOND NORMAL EQUATION:

$$\sum t * Y_t = \alpha \sum t + \beta \sum t^2$$

$$\sum t * Y_t - \alpha \sum t - \beta \sum t^2 = 0$$

$$\sum (t * Y_t - \alpha * t - \beta * t^2) = 0$$

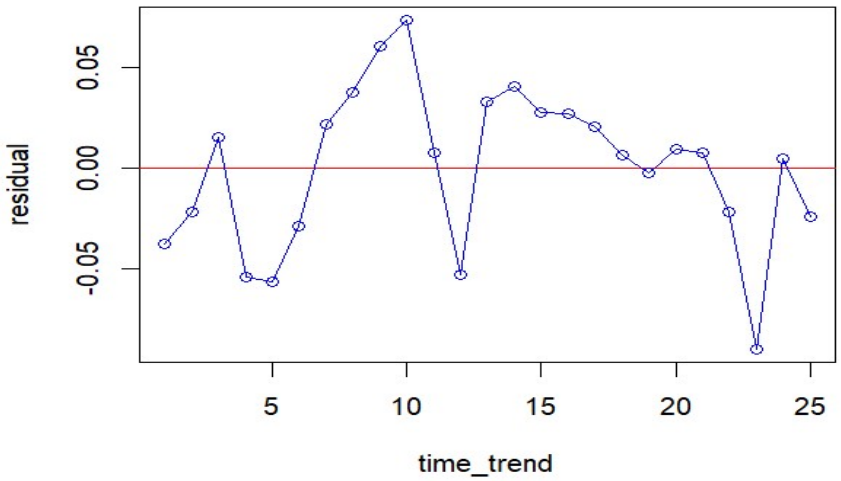
$$\sum t * (Y_t - \alpha - \beta * t) = 0$$

$$\mathbf{\sum (t * U_t) = 0 = 7.105427e\text{-}14}$$

- This can also be calculated as ∑(t*Ut) ≈ 0 as there can be no autocorrelation between the error term and the dependent variable i.e., Cov(Ut,t) = 0.

(b) Draw a scatter plot of the residual term against the time trend (or year) variable. Names the axes, and the titles for this graph and name it Figure 3. Do you observe random pattern of error term with time trend? If yes, is it justified by theory? If you observe a pattern like an upward sloping or downward sloping scatter instead of a random pattern, what do you think could be the reason [hint: connect it with the R2 value]

Figure 3: Residual values over time trend



By looking at figure 3, we can infer randomness on the pattern of U_i, despite of a strong upward pattern from 5th year but later on it depleted down, showing randomness in forth coming years. As from the CLRM assumption suggesting: $Cov(U_t,U_s) \approx 0$, showing that the variations in GDP is only due to the variations in years which is obvious that the model proved to have a R^2 value of 97.24%.

(6) INFERENCE:

(a) Based on the output, what is your conclusion with regard to the statistical significance of the intercept and slope coefficient?

Standard Error of intercept coefficient 'α' = 0.00021

Standard Error of slope coefficient 'β' = 0.015

To test the level of significance at 5%, two tailed hypothesis testing has to be carried out. Primarily, null hypothesis and alternate hypothesis have to be constructed and test statistic have to be formulated.

For Intercept coefficient 'α':

Null hypothesis: H_0: $\alpha = 0$

Alternate hypothesis: H_1: $\alpha \neq 0$

Test Statistic: $t_{cal} = \frac{\alpha - \alpha*}{se(\alpha)}$

$\mathbf{t_{cal} = 51{,}523.80}$

The tabulated value at 5% level of significance, $\mathbf{t_{tab,0.025,23} = 2.068658}$

As the calculated t value is greater than the tabulated t value, we reject H_0 and accept H_1, as the intercept coefficient can't be zero. Thus, the intercept α is statistically significant at 95% confidence interval.

For Slope Coefficient 'β'

Null hypothesis: H_0: $\beta = 0$

Alternate hypothesis: H_1: $\beta \neq 0$

Test Statistic: $t_{cal} = \frac{\beta - \beta*}{se(\beta)}$

$\mathbf{t_{cal} = 2.11}$

The tabulated value at 5% level of significance, $\mathbf{t_{tab,0.025,23} = 2.068658}$

As the calculated t value is greater than the tabulated t value, we reject H_0 and accept H_1, as the slope coefficient can't be zero. Thus, the slope β is statistically significant at 95% confidence interval.

(b) What all will change if the alternative hypothesis is of the type ">0" and give your conclusion regarding accepting or rejecting the null hypothesis that the slope coefficient is zero.

For Intercept coefficient 'α':

Null hypothesis: $H_0: \alpha = 0$

Alternate hypothesis: $H_1: \alpha > 0$

Test Statistic:

$t_{cal} = 51{,}523.80$

The tabulated value at 5% level of significance, $t_{tab,0.05,23} = 1.713872$

As the calculated t value is greater than tabulated t value, we reject H0 and accept H1. Therefore, the intercept coefficient is greater than 0.

For Slope coefficient 'β':

Null hypothesis: $H_0: \beta = 0$

Alternate hypothesis: $H_1: \beta > 0$

Test Statistic:

$t_{cal} = 2.11$

Tabulated value at 5% level of significance, $t_{tab,0.05,23} = 1.713872$

Since, calculated t value is greater than tabulated t value, we reject H0 and accept H1. Therefore, the slope coefficient is greater than zero. There is a statistically significant rate of increase in ln GDP over the years.

(c) Choose a non-zero hypothesized value for the growth rate. Test the null hypothesis with this hypothesized value such that the alternate hypothesis is "less than this hypothesized vale".

For Intercept coefficient 'α':

Null hypothesis: $H_0: \alpha = 10.9$

Alternate hypothesis: $H_1: \alpha < 10.9$

Test Statistic:

$t_{cal} = -380.95$

Tabulated value at 5% level of significance, $t_{tab,0.05,23} = 1.713872$

As the calculated t value is lesser than the tabulated t value, we reject H_0 and accept H_1. Therefore, the intercept coefficient is lesser than 10.70.

For slope coefficient 'β'

Null hypothesis: $H_0: \beta = 0.04$

Alternate hypothesis: $H_1: \beta < 0.04$

Test Statistic:

$t_{cal} = -2.63$

The tabulated value at 5% level of significance, $t_{tab,0.05,23} = 1.713872$

As the calculated t value is lesser than the tabulated t value, we reject H_0 and accept H_1. Therefore, the slope coefficient is lesser than 0.03.

(d) From the ANOVA table in the STATA output give the mathematical expressions for how the different components of the table are calculated. What is the test of hypothesis that the ANOVA table has information about? State the null and alternative hypotheses, test statistic along with its value, the degrees of freedom and decision about the test of hypothesis.

ANOVA TABLE

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F- value	Pr(>F)
X	ESS = 1.307	1	ESS/1 = 1.307	(ESS/RSS) (n-2) = 809.4	<2e-16
Residual	RSS = 0.037	(n-2) = 23	RSS/(n-2) = 0.0016		
Total	TSS = 1.344	(n-1) = 24			

(e) Give the mathematical expression to calculate the p-value in all the above tests of hypothesis. Write the code to calculate the p-value within R and show that it matches with what is given in the output of the software. Is the decision in favour of or against rejecting the null based on p-values?

P-value can be calculated by the following mathematical expressions.

$\boldsymbol{P\ value = Pr\ (t > |t_{cal}|\ |\ H_0)}$ for one tailed

$\boldsymbol{P\ value = 2[Pr\ (t > |t_{cal}|\ |\ H_0)]}$ for two tailed

R Code:

summary(model)$coefficients[,4]

Intercept p-value: 1.351366e-50

Slope p-value : 1.985747e-19

At 95% confidence interval, the p-values are less than 0.05 making the decision in favour of **rejecting the null hypothesis.**

(f) Obtain the 90% confidence interval (CI) for the intercept and the slope coefficient. Based on the CI approach, is the decision in favour of or against rejecting the null in (a) for the intercept and the slope coefficient respectively? Write your findings in two separate sentences for the intercept and slope coefficients.

Confidence Interval for Intercept Coefficient:

$\boldsymbol{\alpha = \alpha \pm t_{tab,0.025,23} * se(\alpha)}$

CI = 0.001913 < α < 0.002631

We reject the Null Hypothesis(H_0) as the Confidence Interval for the Intercept Coefficient does not include zero.

Confidence Interval for Slope Coefficient:

$\boldsymbol{\beta = \beta \pm t_{tab,0.025,23} * se(\beta)}$

CI = -0.02628 < β < 0.03579

We accept the Null Hypothesis(H_0) as the Confidence Interval for the Slope Coefficient includes zero.

(7) PREDICTION: Predict the per capita GDP for one year after the last year in your sample, based on the estimated growth rate using the mathematical expression for growth rate in equation 2.6 (pg53) and then the econometric model. Are there any differences or are they the same?

MATHEMATICAL EXPRESSION :

$$\mathbf{Y_T = Y_0 * (1.0323)^t}$$

For t = 26 and Y_0 to be 49789:

Y_{26} = $113,785

ln_Y_{26} = 11.6422

ECONOMETRIC MODEL:

ln_PCGDP = 10.82 + 0.0317 (time_trend)

PCGDP = $108,772.41

ln_PCGDP = 11.5979

From the values above we see a difference between the mathematical output and econometric result.

However the log values of both the models tend to be the same, we can't infer that both the models are approximately equal as the real PCGDP shows to have a considerable difference.

End of Assignment