Assignment 2

Tahir Muhammad — 1002537613

November 2019

1 Theoretical Problems

Question 1

a) True. We know that the formula for R^2 in a simple linear regression model is represented by:

$$R^2 = \frac{SSE}{SST}$$

When $R^2 = 1$, this means that we have a perfect predictive model. Note that this rarely happens in the empirical world. What it entails is that 100% of the variability in your dependent variable (Y) is explained by the variability in the independent variable (X), or in other words, explained by the model. This can further be understood as that our model graphs the data perfectly, with no residuals. Hence, the linear relationship between the variables is exact and the residuals are all indeed zero.

b) True. If we have $Var(X) = Var(Y)$, then we have the following:

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{Var(x)}$$

$$\Rightarrow \frac{Cov(X,Y)}{Var(X)}$$

$$\Rightarrow \frac{Cov(X,Y)}{Var(Y)}$$

which illustrates that the estimated slope, $\hat{B}_1$ will be the same in the linear regression model even if it is Y on X or X on Y, as the value for the numerator and denominator is equivalent under the condition $Var(X) = Var(Y)$.

c) False. This just means that no amount of variability in Y can be explained by the variability in X. In other words, our model does not explain any of the variability in the response variable Y around it's mean.

d) False. This is not true, as the assumptions of the linear regression model are: I. Linear in parameters, II. Random Sample from the population, III. The sample outcomes for a given X's are not the same value, IV. The error term, u, has an expected value of zero conditional on x. i.e. $E(u|x) = 0$. The last one, V. states that the error term is also homokedastic, i.e. same variance for all x. Thus, the sum of residuals is zero is not a critical assumption of the linear model, even though it ends up getting eliminated during the minimization of the OLS.

e) False. The reason why sum of residuals is zero is because we differentiate in terms of β estimates which minimize the sum of residuals. To mathematically see this, we do:

$$\sum(Y - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\Rightarrow \frac{d}{d\hat{\beta}_0}(\sum(Y - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2)$$

$$2\sum(Y - \hat{\beta}_0 - \hat{\beta}_1 x_i)(1) = 0$$

We have set the sum of residuals to zero in order to find the $\hat{\beta}$.

f) False. This is because in the linear model regression assumptions 1 - 4, we have proved OLS is unbiased which does not include the fact that error term needs to be normally distributed.

Question 2

a) We know that for the given model, $E(U) = 0$, $E(U|X) = 0$ and by using these facts and assumptions we get the following:

$$COV(X|U) = E(XU) - E(X)E(U)$$

$$\Rightarrow E(XU) - E(X)(0)$$

$$\Rightarrow E(XU) = E(E(XU|X))$$

$$\Rightarrow E(E(X(U|X))$$

$$\Rightarrow 0$$

Using $E(U) = 0$ and $E(U|X) = 0$, we have shown that $Cov(X, U) = 0$

b) Here, we assume that $Cov(X, U) = 0$ and need to find the value for β_1. Starting with the equation of our model, we have the following:

$$Y = \beta_0 + \beta_1(x) + U$$

If we subtract Y - E(Y), we get:

$$Y - E(Y) = \beta_0 + \beta_1(x) + U - E(\beta_0 + \beta_1(x) + U)$$

$$= \beta_0 + \beta_1(x) + U - E(\beta_0) - E(\beta_1(x)) - E(U)$$

$$= \beta_1(x - E(x)) + \beta_0 - E(B_0) + U - E(U)$$

$$= \beta_1(x - E(x)) + U - E(U)$$

Now if we multiply both sides by (X - E(X)) and take the expected value, we get:

$$E[(Y - E(Y))(X - E(X))] = E[((\beta_1(x - E(x)) + U - E(U)))(X - E(X))]$$

$$E[(Y{-}E(Y))(X{-}E(X))] = E[(\beta_1(x{-}E(x)))(X{-}E(X)){+}(U)(X{-}E(X)){-}(E(U))(X{-}E(X))]$$

$$E[(Y - E(Y))(X - E(X))] = E[\beta_1((x - E(x)))^2 + (U - E(U))X - E(X)]$$

By applying the assumption that $Cov(X, U) = 0$ we have:

$$E[(Y - E(Y))(X - E(X))] = E[\beta_1((x - E(x)))^2] + E((U - E(U))X - E(X))$$

$$E[(Y - E(Y))(X - E(X))] = E[\beta_1((x - E(x)))^2] + (0)$$

Simplifying for β_1 by expanding by re-writing the squared term we see that:

$$E[(Y - E(Y))(X - E(X))] = \beta_1[E((x - E(x)))^2]$$

$$\beta_1 = \frac{E[(Y - E(Y))(X - E(X))]}{E((x - E(x))(x - E(X))}$$

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)}$$

$\blacksquare$

c) The following proof will make use of two important facts as shown in class.

I. $\frac{\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 0$

and II. $\frac{\sum_{i=1}^{n} X_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 1$

By the definition of $\hat{\beta}_1$, we have:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 X_i + u_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\Rightarrow \beta_0(\frac{\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}) + \beta_1(\frac{\sum_{i=1}^{n} X_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}) + \frac{\sum_{i=1}^{n} u_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

By substituting the fact I and II above, we get:

$$\Rightarrow \beta_0(0) + \beta_1(1) + \frac{\sum_{i=1}^{n} u_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\Rightarrow \hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} u_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

and as n $\rightarrow \infty$, we can see that $\hat{\beta}_1$ converges in probability to:

$$\beta_1 + \frac{Cov(X, U)}{Var(X)}$$

$\blacksquare$

d) In the following question, we make use of the following properties:

I. Taylor Series Approximation
II. $\Delta Log(W) = Log(W_1) - Log(W_0) = Log(\frac{W_1}{W_0})$
III. $\Delta X = X_1 - X_0$

Starting off with the regression model, we get:

$$Y = \beta_0 + \beta_1 X + U$$

If we now differentiate Y with respect to X and sub in Log(W) into Y we get:

$$\Delta Y = \beta_1 \Delta X$$

$$\Delta Log(W) = \beta_1 \Delta X$$

Using property number II and applying logarithmic rules, we achieve:

$$Log(W_1) - Log(W_0) = \beta_1 \Delta X$$

$$\Rightarrow Log(\frac{W_1}{W_0}) = \beta_1 \Delta X$$

$$\Rightarrow e^{Log(\frac{W_1}{W_0})} = e^{\beta_1 \Delta X}$$

$$\Rightarrow \frac{W_1}{W_0} = e^{\beta_1 \Delta X}$$

Subtracting one from each side, we get:

$$e^{\beta_1 \Delta X} - 1 = \frac{W_1}{W_0} - 1$$

$$\Rightarrow \frac{W_1 - W_0}{W_0}$$

Therefore, we can see that the approximate change is 100% of the β_1, and the exact change is 100% of $e^{\beta_1 \Delta X}$ where $\Delta X = X_1 - X_0$ from (III).

e) For an estimator, $\hat{\theta}$, to be biased, we know that $E(\hat{\theta}) \neq \theta$. In our case, we need to show that $E(e^{x\hat{\beta}_1} - 1) \neq e^{x\beta_1} - 1$. Then, for consistency we need to show that as $n \to \infty$ our estimator converges to it's true value. i.e $e^{x\hat{\beta}_1} \to e^{\beta_1} - 1$.

Lets start with showing the bias for this estimator.Since exponential functions are convex everywhere, we use the Jensen Inequality:

$$E(e^{x\hat{\beta}_1}|X) > e^{E(x\hat{\beta}_1|X)}$$

We know that x is a constant and the expectation of a constant is just a constant. Hence,

$$\Rightarrow E(e^{x\hat{\beta}_1}|X) > e^{xE(\hat{\beta}_1|X)}$$

$$\Rightarrow E(e^{x\hat{\beta}_1}|X) > e^{\beta_1 X}$$

Subtracting 1 from both sides, we have:

$$\Rightarrow E(e^{x\hat{\beta}_1}|X) - 1 > e^{\beta_1 X} - 1$$

and thus we can conclude that the estimator has a positive bias for $e^{\beta_1 X} - 1$.

To answer for consistency, we know that $\hat{\beta}_1$ converges in probability to $\beta_1 + \frac{Cov(X,U)}{Var(X)}$. Using the fact that $Cov(X,U) = 0$, we have $\hat{\beta} = \beta_1$ which implies that as $n \to \infty$, $e^{\hat{\beta}_1 X} - 1 = e^{\beta_1 X} - 1$, thus it is indeed consistent.

2 Computer Based Problems

Question 1

a) We model the given equation with regression:

```
. ** PART A **
.
. regress loginc female black age agesq educ1 educ2 educ3 educ4

      Source |       SS           df       MS      Number of obs   =     3,987
-------------+----------------------------------   F(8, 3978)      =    123.70
       Model |  1072.10954         8  134.013693   Prob > F        =    0.0000
    Residual |  4309.72891     3,978  1.08339088   R-squared       =    0.1992
-------------+----------------------------------   Adj R-squared   =    0.1976
       Total |  5381.83845     3,986  1.35018526   Root MSE        =    1.0409

------------------------------------------------------------------------------
      loginc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2672268   .0331098    -8.07   0.000    -.3321407    -.202313
       black |   -.552836   .0565014    -9.78   0.000    -.6636105   -.4420616
         age |   .0490182   .0053594     9.15   0.000     .0385108    .0595256
       agesq |  -.0004872   .0000526    -9.27   0.000    -.0005902   -.0003841
       educ1 |   .4489675   .0721566     6.22   0.000        .3075    .5904349
       educ2 |    .705606    .067872    10.40   0.000     .5725387    .8386732
       educ3 |   1.126566   .0713241    15.80   0.000     .9867309    1.266401
       educ4 |   1.503135   .0749474    20.06   0.000     1.356196    1.650074
       _cons |   9.044574   .1379751    65.55   0.000     8.774066    9.315083
------------------------------------------------------------------------------
```

Figure 1: Regression model of the given equation

The value for the educ1 coefficient is 0.4489675. This means that on average, high school graduates earn 45% more than students which drop out of high school. Furthermore, note that since the coefficients for gender being female and race is black variable turn out to be negative, we can infer that it is a Hispanic high school graduate male who will earn 45% more than a Hispanic high school drop out would have made, keeping age constant. Educ4 has a coefficient value of 1.503 and can be interpreted by adding up betas of educ1, educ2, educ3, educ4 and multiplying it by 100 to know how much percent your income will go up by if you have obtained a PhD. If an individual has a pHD then his income can be expected to increase by 378 percent.

Hence, In this case, we have: $(0.7 + 1.1 + 1.5)x100\% \approx 152\%$ increase in income if a student was to obtain a PhD degree.

b) Estimating the model with base case of education as completed high school. The estimated coefficient on educ4 still has a value of 1.5 but in this model, your income goes up by 198.82 percent. It may not be possible to obtain the same interpretation as in part because you are excluding a certain group(high school dropouts) in the regression model.

```
. ** PART B **
.
. * Regression model with education of a highschool dropout to be base case
. regress loginc female black age agesq educ0 educ2 educ3 educ4

      Source |       SS           df       MS      Number of obs   =     3,987
-------------+----------------------------------   F(8, 3978)      =    123.70
       Model |  1072.10954         8  134.013693   Prob > F        =    0.0000
    Residual |  4309.72891     3,978  1.08339088   R-squared       =    0.1992
-------------+----------------------------------   Adj R-squared   =    0.1976
       Total |  5381.83845     3,986  1.35018526   Root MSE        =    1.0409

------------------------------------------------------------------------------
      loginc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2672268   .0331098    -8.07   0.000    -.3321407    -.202313
       black |   -.552836   .0565014    -9.78   0.000    -.6636105   -.4420616
         age |   .0490182   .0053594     9.15   0.000     .0385108    .0595256
       agesq |  -.0004872   .0000526    -9.27   0.000    -.0005902   -.0003841
       educ0 |  -.4489675   .0721566    -6.22   0.000    -.5904349      -.3075
       educ2 |   .2566385   .0466763     5.50   0.000     .1651269    .3481502
       educ3 |   .6775986   .0514378    13.17   0.000     .5767517    .7784455
       educ4 |   1.054167   .0565341    18.65   0.000     .9433288    1.165006
       _cons |   9.493542   .1289343    73.63   0.000     9.240758    9.746325
------------------------------------------------------------------------------
```

Figure 2: Regression model of the modified equation from Part A

c)

Since the confidence interval does not include 0 for the age, this implies that age has a significant impact on income.

We have our regression model (from part A) as:

$$Log(inc) = \beta_0+\beta_1 female+\beta_2 black+\beta_3 age+\beta_4 age^2+\beta_5 educ1+\beta_6 educ2+\beta_7 educ3+\beta_8 educ4+\epsilon$$

$$\Rightarrow 9.493542+(-0.004872)+(-0.5528)+0.49+(-0.005)+(-0.44489)+0.25667+0.677598+1.05$$

$$\Rightarrow log(inc) = 10.96$$

Thus, the effect of income from age 34 to 35 would be approximately 109.6%.

To get the the age where we see the maximum income level, we do:

$$\frac{dloginc}{dage} = 0.490182 + (-0.004872)age_i = 0$$

$$age_i = 0.490182/0.004872$$

$$\approx 100$$

Question 2

a) Estimation of Equation (1) from the given Question:

```
.
. * Create a new column as the ratio of GDP in 1995 over GDP in 1975
. generate _ratio_ = gdp1995/gdp1975

.
. * Create a new column for the log/ln of GDP in 1975
. generate ln_gdp_1975 = ln(gdp1975)

.
. ** Create a new column for the log of GDP in 1995 over GDP in 1975
. generate ln_ratio = ln(_ratio_)

.
. regress ln_ratio ln_gdp_1975

      Source |       SS           df       MS      Number of obs   =       104
-------------+----------------------------------   F(1, 102)       =      0.08
       Model |  .024514851         1  .024514851   Prob > F        =    0.7817
    Residual |  32.3795311       102  .317446383   R-squared       =    0.0008
-------------+----------------------------------   Adj R-squared   =   -0.0090
       Total |  32.4040459       103  .314602387   Root MSE        =    .56342

------------------------------------------------------------------------------
    ln_ratio |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 ln_gdp_1975 |  -.0083278   .0299676    -0.28   0.782    -.0677685    .0511128
       _cons |   .7134363   .3191008     2.24   0.028     .0805014    1.346371
------------------------------------------------------------------------------
```

Figure 3: Estimation of Equation 1

Beta-Convergence refers to the process of poor countries growing faster than the richer ones, and hence catch up. We can see here that the coefficient here is -.0083278, which implies that rich countries have been predicted to have a slower growth rate than the poorer countries. However, to be certain, we conduct a hypothesis test using a p-value approach A one tail hypothesis test is conducted using the significance level to be 0.05, i.e alpha = 0.05, as convention. Let $H_0 : B \leq 0$ and $H_a : B > 0$ with $\alpha = 0.05$. We can see here that the p-value is 0.782 which is extremely high and so we fail to reject the null hypothesis as the p-value is greater than our significance level. This means that our sample did not contain enough statistical evidence to conclude that Beta convergence is not present.

b) Estimation of Equation 2, given in the question

We define conditionally beta convergence to be true if the value of $\beta_1 < 0$ is true. In our estimation of equation 2, we see that our β_1 value is -.0288579,

```
. regress ln_ratio ln_gdp_1975 hci1975
```

Source	SS	df	MS
Model	.527511093	2	.263755547
Residual	31.8765348	101	.315609256
Total	32.4040459	103	.314602387

Number of obs	=	104
F(2, 101)	=	0.84
Prob > F	=	0.4365
R-squared	=	0.0163
Adj R-squared	=	-0.0032
Root MSE	=	.56179

ln_ratio	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_gdp_1975	-.0288579	.0340195	-0.85	0.398	-.0963435	.0386276
hci1975	.1272948	.1008331	1.26	0.210	-.072731	.3273207
_cons	.7038339	.318267	2.21	0.029	.0724777	1.33519

Figure 4: Estimation of Equation 2

a negative number. Thus, we might think that there is evidence in favour of conditional economic convergence. To be sure, we again conduct a Hypothesis test with $H_0 : B_1 \leq 0$ and $H_a : B_1 > 0$, and $\alpha = 0.05$. Although the p-value, 0.398, has dropped significantly, it is still greater than alpha and thus we conclude that there is not significant evidence against conditional economic convergence, i.e we fail to reject the null. Furthermore, in comparison to part a, we have the coefficient value is also more negative then before, along with the p-value. However, it is not low enough to statistically conclude against beta convergence. Moreover, it is important to note that Adjusted R^2 is negative and barely changed, which implies this variable might not be making the model better.

c) Estimation of Equation 3, given in the question.

Lets add one more variable to our model, the share of gross capital in a country.

```
. regress ln_ratio ln_gdp_1975 gcf1975 hci1975

      Source |       SS           df       MS      Number of obs   =       104
-------------+----------------------------------   F(3, 100)       =      0.70
       Model |  .662809929         3  .220936643   Prob > F        =    0.5566
    Residual |   31.741236       100   .31741236   R-squared       =    0.0205
-------------+----------------------------------   Adj R-squared   =   -0.0089
       Total |  32.4040459       103  .314602387   Root MSE        =    .56339

------------------------------------------------------------------------------
    ln_ratio |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 ln_gdp_1975 |  -.0289217   .0341167    -0.85   0.399    -.0966082    .0387648
     gcf1975 |   .3298003   .5051448     0.65   0.515    -.6723927    1.331993
     hci1975 |   .1049866   .1067376     0.98   0.328    -.1067778     .316751
       _cons |   .6665799   .3242353     2.06   0.042     .0233062    1.309854
------------------------------------------------------------------------------
```

Figure 5: Estimation of Equation 3

Here we can see that the adjusted R^2 has decreased again, and it seems like the model is worse off. The results are similar to the answer in part B. They are not jointly important as adding the last one did not make the model much better off.

Question 3

a) A detailed summary of the Monte Carlo Model, testing for $H_0 : \beta_1 = 5$ and $H_1 : \beta_1 \neq 5$.

```
. * Provide a detailed summary for Model Number 1
. summarize p, detail

                            r(p)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0120835       .0007046
 5%     .0548334       .0032097
10%     .1163594        .004386       Obs               1,000
25%     .2626099       .0046347       Sum of Wgt.       1,000

50%     .5037142                      Mean           .5059368
                        Largest       Std. Dev.      .2835997
75%      .741937       .9938505
90%     .8989158       .9947109       Variance       .0804288
95%     .9502604       .9960333       Skewness       -.010742
99%     .9860045       .9974923       Kurtosis       1.828608

.
. summarize p if p < 0.05

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           p |         41     .024138    .0136896   .0007046   .0476492
```

Figure 6: Summary of Monte Carlo Model 1

It seems like only in the fraction of 1% can the null be rejected. Furthermore, approximately 95% of the observations should not be in the rejection region, which we can confirm is indeed the case as the observations are 41 out of 1000 that were rejected when p ¡ 0.05.

b) A detailed summary of the Monte Carlo Model Number 2, testing for $H_0 : \beta_1 = 4.5$ and $H_1 : \beta_1 \neq 4.5$.

```
. * Provide a detailed summary for Model Number 2
. summarize p, detail

                            r(p)
-------------------------------------------------------------
      Percentiles      Smallest
 1%      .005232       .0003912
 5%     .0259741       .0004445
10%     .0678381       .0009305       Obs                1,000
25%     .1678007        .001701       Sum of Wgt.        1,000

50%     .4001751                      Mean             .4371482
                        Largest       Std. Dev.        .2967907
75%      .689941       .9991254
90%     .8707902       .9993187       Variance         .0880847
95%     .9430179       .9998724       Skewness         .2925929
99%     .9911476       .9999812       Kurtosis         1.837409

.
. summarize p if p < 0.05

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           p |         80    .0218881     .014323   .0003912   .0487971
```

Figure 7: Summary of Monte Carlo Model 2

If we keep $\alpha = 0.05$, it seems like only in the fraction of 1% and 5% can the null be rejected. 10% would also be included, but it does not make the cut when alpha is 0.05. Furthermore, approximately 95% of the observations should not be in the rejection region, but with the beta value of 4.5, we get that 80 out of 1000 observations were rejected when p ¡ 0.05. This means that we are rejecting around 10% of the stimulations.

c) A detailed summary of the Monte Carlo Model Number 3, testing for $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$.

```
. * Provide a detailed summary for Model Number 3
. summarize p, detail

                            r(p)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     1.02e-14       2.91e-16
 5%     6.29e-13       6.03e-16
10%     5.65e-12       9.19e-16       Obs               1,000
25%     1.54e-10       1.54e-15       Sum of Wgt.       1,000

50%     5.68e-09                      Mean           .0000161
                        Largest       Std. Dev.      .0001735
75%     1.80e-07       .0015472
90%     3.34e-06       .0016745       Variance       3.01e-08
95%     .0000157       .0018517       Skewness       18.69145
99%     .0003339       .0043493       Kurtosis       418.6404

.
. summarize p if p < 0.05

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           p |      1,000    .0000161    .0001735   2.91e-16   .0043493
```

Figure 8: Summary of Monte Carlo Model 3

Looking at B1-hat = 0, we reject 100% of the simulations at the signifance level of alpa = 0.05. We can reject the null for all fractions at alpha = 0.05. These fractions are not close to 0.005, and it seems that alpha is greater. The results are expected as we know the true value to be 0 and these are the estimated values.