Assignment #2

Instructions:

- For all questions, answer up to 4 decimal places.
- This assignment is due on **Thursday, May 20, 2021 before 23.59.**
- Write your answer in either digital or ordinary paper. For digital paper, export pages into a single PDF file. For ordinary paper, take photos of your writing and convert them into a single PDF file as well.
- There is no need to rewrite the question. Assign number item, i.e. 1 a., clearly before your answer is sufficient.
- Submit your assignment into Moodle.
- Name your file as StudentID_Nickname (in Thai) such as 123456789_โบ้. **Please follow this instruction strictly since it will help me a lot with file management.**

Question 1. The data set CEOSAL1.DTA contains information on 209 CEOs for the year 1990; these data were obtained from Business Week (5/6/1991). To study effect of firm performances and types of industry where CEOs work on CEO compensation, the CEO salary regression is proposed as follows:

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 \mathrm{ROE}_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility_i + u_i$$

where
$\log(salary_i)$ = logarithm of CEO annual salary (in 1,000 USD)
$\log(sales_i)$ = logarithm of firms' sale (in 1 million USD)
ROE_i = average return on equity for the CEO's firm for the previous three years (Return on equity is defined in terms of net income as a percentage of common equity)
$finance_i$ = 1 if in financial industry, = 0 otherwise
$consprod_i$ = 1 if in consumer product industry, = 0 otherwise
$utility_i$ = 1 if in utility industry, = 0 otherwise

Dummy –

($finance_i$, $consprod_i$, and $utility_i$ are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation.

Using STATA, the estimation result is shown below. Answer the following questions.

```
   Source |       SS       df       MS              Number of obs =     209
----------+------------------------------           F(  5,   203) =   22.53
    Model |  23.8109943     5  4.76219887           Prob > F      =  0.0000
 Residual |  42.9111689   203  .211385068           R-squared     =  0.3569
----------+------------------------------           Adj R-squared =  0.3410
    Total |  66.7221632   208  .320779631           Root MSE      =  .45977

------------------------------------------------------------------------------
  lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
   lsales |   .2571917   .0320348     8.03   0.000     .0194282    .3203553
      roe |   .0111517   .3342996     2.59   0.010     .0026742    .0196293
  finance |   .1579564   .0890017     1.77   0.077    -.0175299    .3334426
 consprod |   .1808917   .0847683     2.13   0.034     .0137524    .3480311
  utility |  -.2830015   .0992337    -2.85   0.005    -.4786624   -.0873405
    _cons |   4.588101   .2950221    15.55   0.000       4.0064    5.169801
------------------------------------------------------------------------------
```

a. Write out the estimated regression equation for $\log(salary_i)$. Interpret the estimated coefficient associated with $\log(sales_i)$.

b. What is the overall significance of the regression? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State **the critical value** for hypothesis testing to receive full points.

c. Compute the approximate percentage difference in estimated salary between the utility and transportation sector, holding $sales_i$ and ROE_i fixed.

d. Why can't we put all the sector dummies (i.e. $finance_i$, $consprod_i$, $utility_i$ and $transport_i$) in the equation? What would happen if we put all the sector dummies in the equation and use STATA run the regression anyway?

e. In the above model, is there any benefit if we add interaction terms between roe and sector dummies, i.e. $\text{ROE}_i * finance_i$ and/or $\text{ROE}_i * consprod_i$ and/or $\text{ROE}_i * utility_i$?

1a. $\widehat{\log(salary)} = 4.5881 + 0.2572 \log(sale) + 0.0115\ ROE + 0.1580\ finance + 0.1809\ consprod - 0.2830\ utility$

The relationship between salary and sales is double log, therefore the coefficient of sales is the elasticity.
If sale increases by 1 Percent, salary will increase by 0.257 percent.

b. Overall Significant indicates that all independent variables are jointly related to dependent variable on the regression or not.
We use F-statistic test to show overall significant.

At 5 percent significant level, t-critical value = 1.653 when df = 203

C-value of:

$B_{sales} = \frac{0.2572}{0.0320} = 8.0375$

$B_{roe} = \frac{0.0112}{0.3343} = 0.0335$

$B_{finance} = \frac{0.1580}{0.0890} = 1.7753$

$$B_{consprod} = \frac{0.1806}{0.0848} = 2.1333$$

$$B_{utility} = \frac{-0.2830}{0.9923} = -0.2852$$

C- value of Isales, finance, consprod is more than 1.t- table1 therefore they are significant.

c. 1 = Utility industry 0 = otherwise

B Utility = -0.2830

The estimated salary of people in the Utility sector is 28.30% less than transportation sector.

d. We can not put in all the sector dummies because a problem of perfect Multicolinearity will occur. This means that x will show a correlation with another x. If we run stata anyway, Newer version of stata will omit 1 dummy variable for us. The older version would show "error."

e. Interation terms = ROE x Sector

We will be able to see the impact of Roe of each sector on salary.

Question 2. Birth weight has been used by officials as one of the main determinants of health. Data set BWGHT.DTA contains data on infant birth weights in ounces ($bwght_i$), average number of cigarettes mother smoked per day during pregnancy (cigs), family income ($faminc_i$), father's year of education ($fathereduc_i$), and mother's year of education ($mothereduc_i$). The following two regressions were estimated using data on n = 1191 births:

Model 2.1: $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + u_i$

```
regress  bwght  cigs   faminc

    Source |       SS       df       MS              Number of obs =    1191
-----------+------------------------------           F(  2,  1188) =   18.44
     Model |  14536.9538     2  7268.47691            Prob > F      =  0.0000
  Residual |  468209.738  1188  394.115941            R-squared     =  0.0301
-----------+------------------------------           Adj R-squared =  0.0285
     Total |  482746.692  1190  405.669489            Root MSE      =  19.852

------------------------------------------------------------------------------
     bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
      cigs |  -.5876985   .1090181
    faminc |   .0624684   .0324438
     _cons |   118.5568   1.234278
------------------------------------------------------------------------------
```

Omitted for the purpose of this exam.

Model 2.2: $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 fathereduc_i + \beta_4 mothereduc_i + u_i$

```
regress  bwght  cigs   faminc fatheduc motheduc

    Source |       SS       df       MS              Number of obs =    1191
-----------+------------------------------           F(  4,  1186) =   10.05
     Model |  15827.6593     4  3956.91482            Prob > F      =  0.0000
  Residual |  466919.033  1186   393.69227            R-squared     =  0.0328
-----------+------------------------------           Adj R-squared =  0.0295
     Total |  482746.692  1190  405.669489            Root MSE      =  19.842

------------------------------------------------------------------------------
     bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
      cigs |  -.5894954   .1106172
    faminc |   .0538254   .0366502
  fatheduc |   .4936695   .2832896
  motheduc |  -.4379234   .3197377
     _cons |   118.0741   3.500291
------------------------------------------------------------------------------
```

Omitted for the purpose of this exam.

where $bwght_i$ = birth weight, ounces
$cigs_i$ = average number of cigarettes the mother smoked per day while pregnant
$faminc_i$ = 1988 family income, $1000s
$fathereduc_i$ = father's years of education
$mothereduc_i$ = mother's years of education

Answer the following questions.

a. Based on **Model 2.1**, test whether smoking has an impact on birth weight. Show your work. (use $\alpha = 0.05$)

b. Based on **Model 2.1**, construct a 99% confidence interval for β_2.

c. Would your conclusion in a) change if you use the result from **Model 2.2**? Show your work. (use $\alpha = 0.05$)

d. What is the overall significance of the regression from **Model 2.2**? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State the critical value for hypothesis testing to receive full points.

e. If we are interested in testing whether "**parents' education**" has an impact on birth weight at all, what kind of null/alternative hypothesis would we be testing? Perform the test and discuss your finding. (use $\alpha = 0.05$)

a. Hypothesis testing:

H_0 : $\beta_{cig} = 0$ [No impact]
H_1 : $\beta_{cig} \neq 0$ [Impact]

Significant level $\alpha = 0.05$ [two tail]
$\alpha = 0.025$ [one tail]

$$t\text{-stat} = \frac{\hat{\beta} - \beta}{SE(\beta)} = \frac{-0.5877 - 0}{0.1090} = -5.3917$$

t-table : df-k = 1191 - 3 = 1188
t = 1.962, -1.962 at $\alpha = 0.025$

$|t\text{-stat}| > |t\text{-table}|$ Therefore we reject H_0 and accept H_1.
Smoking have impact on Bweight at 0.05 sig level.

b. Confident Interval = $\hat{\beta} \pm t_{\alpha/2} \cdot SE(\beta)$

β_2 = fanInc
$\hat{\beta_2} = 0.0624$
$SE(\hat{\beta_2}) = 0.0324$
$t_{0.01/2} = t_{0.005}$

df = n - k = 1191 - 3 = 1188 t = 2.581 at 0.005

CI = $0.0624 \pm 2.581 \cdot 0.0324 = 0.0212, 0.1460$

c. H_0: $B_{cig} = 0$ [No impact]
H_1: $B_{cig} \neq 0$ [impact]

At significant level = 0.05 (two tails)

$$T\text{-stat} = \frac{\hat{B}_2 - B}{SE(B_2)} \qquad \frac{-0.5895 - 0}{0.1106} = -5.3300$$

T-table: $df = n - k = 1191 - 5 = 1186$
T-table: 1.962, -1.962 at $\alpha = 0.025$

$|T\text{-stat}| > |T\text{-call}|$ = Therefore no change in answer.

d. For overall significant, we are using F-test.

F-test:
H_0: all coefficient = 0
H_1: atleast one coefficient $\neq 0$
From the table F stat = 10.05 with p-value = 0
We found that p-value of F < Significant level so the regression is overall significant

T-table: $df = n - k = 1191 - 5 = 1186$
T-table: 1.962, -1.962 at $\alpha = 0.025$

$B_{cigs} = \frac{-0.5895}{0.1106} = -5.3300$

β faminc $= \frac{0.0538}{0.0367} = 1.4659$

β fatheduc $= \frac{0.4937}{0.2833} = 1.7427$

β motheduc $= \frac{-0.4379}{0.3197} = 1.3697$

Only C value of β Cigs is more than |T-table| $\therefore$ Only β Cigs is individually, statistically significant.

e. H_0 : β fatheduc $= 0$ [has no impact]
H_1 : β fatheduc $\neq 0$ [has impact]
At $\alpha = 0.05$

T-stat: $\frac{\hat{\beta} - \beta}{SE(\beta)} = \frac{0.4937 - 0}{0.2833} = 1.7427$

T-table : $df = n - k = 1191 - 5 = 1186$
T-table : $1.962, -1.962$ at $\alpha = 0.025$

$|T\text{-stat}| < |T\text{-table}|$ $\therefore$ Father's education has no impact on birth's weight at 0.05 α.

Question 3. A model of wage equation is given by

$$lawge_i = \beta_1 + \beta_2 exp_i + +\beta_3 expsq_i + \beta_4 educ_i + \beta_5 age_i + \beta_6 kid6_i + \beta_7 kid18_i + u_i$$

where $lwage_i$ = natural log of hourly wage of married women
exp_i = years of experience
$expsq_i$ = years of experience squared
$educ_i$ = years of education
age_i = age
$kid6_i$ = number of children aged 0-6 in a household
$kid18_i$ = number of children aged 6-18 in a household

The regression result from OLS is shown in the table below and answer the following questions.

Source	SS	df	MS		
Model	ESS	___	_____	Number of obs =	428
Residual	RSS	___	.446526442	F(___,_____) =	13.19
Total	223.327441	___	_____	Prob > F =	0.0000
	TSS			R-squared =	0.1582
				Adj R-squared =	_____
				Root MSE =	.66823

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.039819	.013393	2.97	0.003	.0134936	.0661444
expersq	-.0007812	.0004022	-1.94	0.053	-.0015718	9.37e-06
educ	.1078319	.0144021	7.49	0.000	.079523	.1361409
age	-.0014653	.0052925	-0.28	0.782	-.0118682	.0089377
kidslt6	-.0607106	.0887626	-0.68	0.494	-.2351836	.1137625
kidsge6	-.014591	.0278981	-0.52	0.601	-.069428	.0402459
_cons	-.4209078	.316905	-1.33	0.185	-1.043821	.2020053

a) Figure out all the degrees of freedom in this model.

b) Figure out all the sum of squares (ESS and RSS) and mean squares in this model.

c) Figure out the adjusted R-squared ($\bar{R}^2$)

d) Given that the model above is called '**Model 3.1**', there is another competing model called '**Model 3.2**' which **an explanatory variable is excluded**, compared to '**Model 3.1**'. Though the result of estimating '**Model 3.2**' is not shown here, **what is the maximum value of R^2 from 'Model 3.2'** which will make you conclude that the excluded variable has a significant contribution in '**Model 3.1**', at the significance level of 0.05. (**Hint:** the critical value of the F-test at the significance level of 0.05 is $F_{1,421} = 3.84$)

e) As you can see from the result, age is not significantly different from zero. In other words, age does not determine how much hourly wage would be. Does this make economic sense in your opinion? What do you think cause this insignificance?

3a. df Model = k-1 = 7-1 = 6
df residual = n-k = 428-7 = 421
df total = n-1 = 428-1 = 427

b. MS = SS/df 0.4463 = RSS/421 RSS = 187.8923

ESS + 187.8923 = 223.3274

ESS = 35.4351

c. $\bar{R}^2 = 1 - \frac{(1-R^2)(N-1)}{N-k}$

$\bar{R}^2 = 1 - \frac{(1-0.1582)(428-1)}{428-7}$

$\bar{R}^2 = 1 - 0.8538 = 0.1462$

d. If the value of the R^2 in Model 3.2 decreased, this means that the excluded variable is significant.

e. It does not make economic sense because as you age from child to adult, you gained knowledge and experience to work. And as you grow older, you gained more skills. The cause of this insignificance is because there is a problem of co-linearity in the regression. Both the "exper" and "age" resembles the same type of informations.