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

($finance_i$, $consprod_i$, 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 $\text{ROE}_i * utility_i$?

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

log(salary) = 4.588709 + 0.2571917 log(sales)

positive coefficient β_1 or log(sales). It means when log(sales) increase by 1%, log(salary) will increase by 0.2571917%.

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.

Use F-test

$H_0 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1 = Not slope coefficients are simultaneously 0

$$F_{cal} = \frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}} = \frac{MSE}{MSR} = \frac{4.762198857}{0.211385068} = 22.52854904$$

at $\alpha = 0.05$ 95%.

$F_{cal} > F_{cri,0.05}^{(5,203)}$

$F_{upper}^{(5,203)} = 2.21$

Reject H_0 and we sure that $\beta_2, \beta_3, \beta_4, \beta_5$ are not simultaneously zero

∴ According to the test it shows that the all independent variables can significantly explain the dependent variable

$\alpha = 0.05$ 95%, d.f. = 203, lower $t_{cri} = -1.96$, upper $t_{cri} = 1.96$

	t_{cal}	
1. log(sales)	8.03	∴ Reject H_0, significance
2. roe	2.59	∴ Reject H_0, significance
finance	1.77	∴ Cannot Reject H_0, not significance
consumer product	2.13	∴ Reject H_0, significance
utility	-2.85	∴ Reject H_0, significance
constant	15.55	∴ Reject H_0, significance

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

utility

$\log(salary) = \beta_0 + \beta_3 Finance_i + \beta_4 consprod_i + \beta_5 utility_i$

$= 4.588701 + 0.157964(1) + 0.1808917(1) - 0.283005(1)$

$\log(salary) = 4.643947$

Transport sector

$\log(salary) = \beta_0 + \beta_3 Finance_i + \beta_4 consprod_i + \beta_5 utility_i$

$= 4.588701 + 0.157964(0) + 0.1808917(0) - 0.283005(0)$

$= 4.588701$

$\Delta\% = \frac{4.588701 - 4.643947}{4.643947} = -0.012 \Rightarrow -1.2\%$ Ans

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?

Because if we put all dummy variables in equation, a case of perfect collinearity will occurs. Each variable will have exact linear relationship between them.

In stata, it will automatically reject or omit one dummy variable. Because it cannot be estimated.

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

$$\log(salary) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ROE_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility_i + \beta_6 (ROE, finance) + \beta_7 (ROE, consprod) + \beta_8 (ROE, utility) + u_i$$

$$E(\log salary \mid finance = 1) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ROE_i + \beta_3 (1) + \beta_4 consprod_i + \beta_5 utility_i + \beta_6 (ROE, 1) + \beta_7 (ROE, consprod) + \beta_8 (ROE, utility) + u_i$$

$$= (\beta_0 + \beta_3) + \beta_1 \log(sales_i) + (\beta_2 + \beta_6)(ROE_i) + \beta_4 consprod_i + \beta_5 utility_i + \beta_7 (ROE, consprod) + \beta_8 (ROE, utility) + u_i$$

$$E(\log salary \mid consprod = 1) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ROE_i + \beta_3 finance_i + \beta_4 (1) + \beta_5 utility_i + \beta_6 (ROE, finance) + \beta_7 (ROE, 1) + \beta_8 (ROE, utility) + u_i$$

$$= (\beta_0 + \beta_4) + \beta_1 \log(sales_i) + (\beta_2 + \beta_7) ROE_i + \beta_3 finance_i + \beta_5 utility_i + \beta_6 (ROE, finance) + \beta_8 (ROE, utility) + u_i$$

$$E(\log(salary) \mid Utility = 1)$$

$$= \beta_0 + \beta_1 \log(sales_i) + \beta_2 ROE_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 (1) + \beta_6 (ROE, finance)$$

$$+ \beta_2 (ROE, consprod) + \beta_8 (ROE, 1) + u_i$$

$$= (\beta_0 + \beta_5) + \beta_1 \log(sales_i) + (\beta_2 + \beta_8) ROE_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_6 (ROE, finance)$$

$$+ \beta_2 (ROE, consprod) + u_i$$

∴ When we add interaction term between roe and sector dummies. The intercept and slope of this regression will change. Y-intercept will increase and slope will be steeper or flatter is depends on roe and sector dummies that will significantly different from zero or not

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 $\alpha = 0.05$)

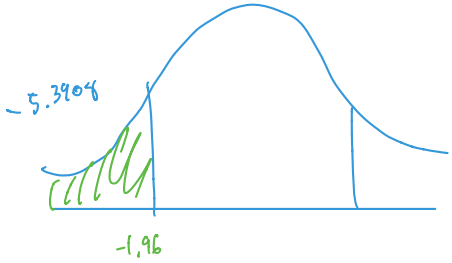
d. What is the overall significance of the regression from **Model 2.2**? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State the critical value for hypothesis testing to receive full points.

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

$\alpha = 0.05$ $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$



use T-test

$$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se\,\hat{\beta}_1} = \frac{-0.5876985}{0.1090187} = -5.39048$$

$d.f = n-3 = 1184$ $t_{lower} = -1.96$, $t_{upper} = 1.96$

∴ reject H_0. We can make sure that β_1 is not simultaneously zero.

so smoking has an impact on birth weight

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

B. $\hat{\beta}_2 \pm t_{\frac{\alpha}{2}, n-3} \cdot se(\hat{\beta}_2)$; $t_{\frac{0.01}{2}, 1188} = 2.576$

$= 0.624684 \pm 2.576 \cdot (0.324438)$

lower $= -0.2111$

upper $= 1.4604$

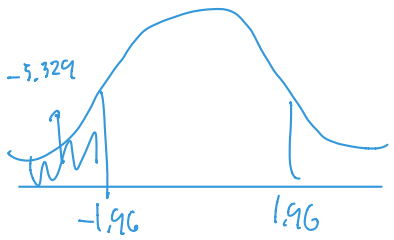
∴ Confidence interval for $\hat{\beta}_1$ is $Pr(-0.2110 < \hat{\beta}_2 < 1.4604) = 0.99$

given this confidence of 99%. 99 out of 100 case this interval will certain the true $\hat{\beta}_2$.

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

$\alpha = 0.05$ $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$



t-test

$$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se\hat{\beta}_1} = \frac{-0.5894954}{0.1106772} = -5.329$$

d.f. $= n-5 = 1191-5 = 1186$

$t_{lower} = -1.96$, $t_{upper} = 1.96$

∴ We reject H_0, we can make sure that β_1 is not simultaneously zero.

so smoking has an impact on birth and conclusion in a. not change

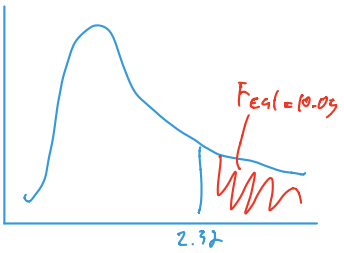
d. What is the overall significance of the regression from **Model 2.2**? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State the critical value for hypothesis testing to receive full points.

F-test

$H_0: \beta_2 = \beta_3 = \beta_4 = 0$; $\alpha = 0.05$

H_1: Not all slope coefficients are simultaneously zero

From stata $F^{cal}_{(4,1986)} = 10.05$



$F_{crit(4,1986)} = 2.32$ $\quad F_{cal} > F_{crit(4,1986)}$

∴ we can reject H_0 and make sure that β_2, β_3 and β_4 are not simultaneously zero.

T-test : $\alpha = 0.05$; $t_{lower} = -1.96$, $t_{upper} = 1.96$

$H_0: \beta_0 = 0$
$H_1: \beta_0 \neq 0$

$t_{cal} = \frac{\hat{\beta}_0 - \beta_0}{se\hat{\beta}_0} = \frac{118.0741}{3.500291} = 33.7326$ ∴ Reject H_0, significance

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se\hat{\beta}_1} = \frac{-0.5894954}{0.1106172} = -5.329$ ∴ Reject H_0, significance

$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

$t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{se\hat{\beta}_2} = \frac{0.5382541}{0.3665021} = 1.4686$ ∴ cannot reject H_0, not significance

$H_0: \beta_3 = 0$
$H_1: \beta_3 \neq 0$

$t_{cal} = \frac{\hat{\beta}_3 - \beta_3}{se\hat{\beta}_3} = \frac{0.4936695}{0.2832896} = 1.7426$ ∴ cannot reject H_0, not significance

$H_0: \beta_4 = 0$
$H_1: \beta_4 \neq 0$

$t_{cal} = \frac{\hat{\beta}_4 - \beta_4}{se\beta_4} = \frac{-0.4329234}{0.3297377} = -1.3696$ ∴ cannot reject H_0, not significance

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 $\alpha = 0.05$)

E. $H_0: \beta_3 = \beta_4 = 0$

$H_1:$ Not all slope coefficients are simultaneously zero

$\alpha = 0.05 \quad d.f = 4, n = 1180$

$$F_{cal} = \frac{MSCE1}{MSCR1} = \frac{3956.91482}{393.69227} = 10.05$$

$F_{cri(4,1186)} = 2.37$

$\therefore F_{cal} > F_{cri(4,1186)}$ we can reject H_0 and we know that both β_3 and β_4 has an impact on birth weight.

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	35.3304012	6	5.88840019	Number of obs	= 428
Residual	187.997304	427	.446526442	F(6, 421)	= 13.19
				Prob > F	= 0.0000
				R-squared	= 0.1582
				Adj R-squared	= 0.1462
Total	223.327441	427	0.523015	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?

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

	ESS	RSS	TSS
df =	6	421	422

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

$R^2 = \frac{ESS}{TSS}$

$0.1582 = \frac{ESS}{223.327441}$

$ESS = 35.3304012$

$RSS = TSS - ESS$

$RSS = 223.327441 - 35.3304012$

$RSS = 187.99704$

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

$\bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-k}$

$\bar{R}^2 = 1-(1-0.1582)\frac{427}{421} \Rightarrow \bar{R}^2 = 0.1462$

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 $F_{1,421} = 3.84$)

$$F_{(1,421)} = \frac{\frac{R^2_{3.1} - R^2_{3.2}}{\text{(number of new regressions)}}}{\frac{1-R^2_{3.1}}{(n-k_{3.1})}} = \frac{\frac{0.1582 - R^2_{3.2}}{1}}{\frac{1-0.1582}{(428-7)}}$$

$$3.84 = \frac{(0.1582 - R^2_{3.2})(421)}{0.8418} \qquad R^2_{3.2} = 0.1505$$

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?

In terms of economic sense, we would say No. Because in reality age is one of factor that will determines wage. For example, people with higher age may have work experience than people with lower age. So wage most be higher than those people with lower age. on the other hand, age variable is insignificance in this model because this term may have multicollinearity with other terms. Therefore, age variable will be omitted because it unnecessary to have this variable in this variable in this.