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.

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

```
      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 $\log(sales_i)$.

$\log(salary_i) = 4.588 + 0.2572 \log(sale_i) + 0.0112\, ROE_i + 0.1580\, finance_i + 0.1809\, consprod_i - 0.2830\, utility_i$

if logarithm of firms' sale increase by 1 million USD, logarithm of CEO annual salary will increase 0.2572 (in 1000 USD)

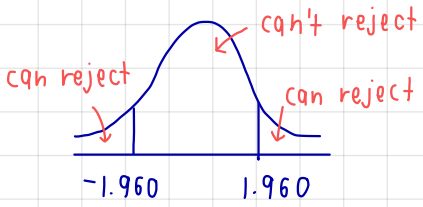
b. What is the overall significance of the regression? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State **the critical value** for hypothesis testing to receive full points.

Since $\alpha = 0.05$; $H_0 : \beta_1, \beta_2, \beta_3 = 0$
$H_a : \beta_1, \beta_2, \beta_3 \neq 0$

(1) Use p-value
- if $p > 0.05$, it means we need to accept H_0
- In this case, p-value of $finance_i$ is greater than 0.05 ($0.077 > 0.05$). So, we accept H_0 (not significant)

(2) use t-test ; $\alpha = 0.05$, $df = 203$ we directly open t-table for critical value which are

$t_{lower} = -1.960$ and $t_{upper} = 1.960$



Hence, β_0, β_1, β_2, β_4 and β_5 can reject H_0.

However, t_{cal} of β_3 is 1.77 so cannot reject H_0

lsalary	Coef.	Std. Err.	t	
lsales	.2571917	.0320348	8.03	Reject H_0
roe	.0111517	.3342996	2.59	Reject H_0
finance	.1579564	.0890017	1.77	can't reject H_0
consprod	.1808917	.0847683	2.13	Reject H_0
utility	-.2830015	.0992337	-2.85	Reject H_0
_cons	4.588101	.2950221	15.55	

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

$\log(salary_i) = 4.588 + 0.2572 \log(sale_i) + 0.0112 ROE_i + 0.1580 finance_i + 0.1809 consprod_i - 0.2830 utility$

(0 over $\log(sale_i)$, 0 over ROE_i)

Suppose fixed $ROE = 0$ and $sales_i = 0$

(1) $E(\hat{SS}_i \mid finance = 0;\ consprod = 0,\ utility = 1;\ transport = 0) = 4.588 - 0.2830(1) = 4.305$

(2) $E(\hat{SS}_i \mid finance = 0;\ consprod = 0,\ utility = 0;\ transport = 0) = 4.588$

$\therefore$ percentage difference in estimated salary between utility and transportation sector $= \frac{4.588 - 4.305}{4.305} \times 100 = 6.5738\%$

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?

We can't put all the sector dummies because it's overdetermined model or when k is larger than n. Moreover, overdetermined model causes multicollinearity. We need to calculate by using $n > k$.

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 $\text{ROE}_i * utility_i$?

For example, adding interaction terms between ROE and $finance_i$

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 \text{ROE}_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility_i + u_i + \beta_6 (\text{ROE}_i \, finance_i)$$

if β_2 is significantly differently from zero, ROE are different in terms of intercept

if β_6 is significantly differently from zero, financial industry has different slope compared to other industries. ROE affects salary for each industry differently.

By the ways, it shows that adding interaction term between ROE and dummies is benefit because the model is more realistic. The difference industry must have different of ROE to increase in 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 $\alpha = 0.05$)

$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + u_i$

$bwght_i = 118.5568 - 0.5877\ cigs_i + 0.0625\ faminc_i + u_i$

For β_1 $H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

$$t_{cal}(\beta_1) = \frac{-0.5877 - 0}{0.1090} = -5.3917$$

when $\alpha = 0.05$, $d.f = 1188$

$\Rightarrow t_{lower} = -1.960$, $t_{upper} = 1.960$

Hence, since t_{cal} is in reject region, we can reject H_0.
This shows that smoking has an impact on birth weight.

if average of mother smokes increase 1 day while pregnant, birth weight will decrease by 0.5877 ounces

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

step1 $\alpha = 0.01$; $d.f = 1188$

step2 $t_{\frac{0.01}{2}} = t_{0.005}$

step3 $\hat{\beta}_2 \pm t_{0.005} \cdot \sigma_{\hat{\beta}_2}$

$= 0.0625 - (2.576 \cdot 0.0324)$

$= 0.0625 - 0.0835$

$= -0.021$

$= 0.0625 + (2.576 \cdot 0.0324)$

$= 0.0625 + 0.0835$

$= 0.146$

So, 99% confidence interval are -0.021 and 0.146

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

$bwght_i = 118.0741 - 0.5895\ cigs_i + 0.0538\ faminc_i + 0.4937\ fathereduc_i - 0.4380\ mothereduc_i + u_i$

For β_1 $H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

$t_{cal}(\beta_1) = \frac{-0.5895 - 0}{0.1106} = -5.3300$

when $\alpha = 0.05$; d.f = 1186

$\Rightarrow t_{lower} = -1.960$, $t_{upper} = 1.960$

Hence, since t_{cal} is in reject region, we can reject H_0.
This shows that smoking has an impact on birth weight.
However, using model 2.2 doesn't change the conclusion a)

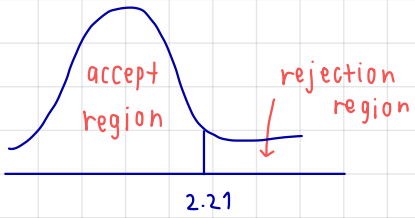
d. What is the overall significance of the regression from **Model 2.2**? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State the critical value for hypothesis testing to receive full points.

» Using general F-testing to test overall significance of regression from Model 2.2

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ $H_a : \beta_1 = \beta_2 = \beta_3 = \beta_4 \neq 0$

$F_{cal} = \frac{MS(E)}{MS(R)} = \frac{3956.9148}{393.6923} = 10.0508$

when $\alpha = 0.05$, $F_{upper,\alpha}(4, 1186) = 2.37$



$\therefore F_{cal} > F_{cri(0.05)}(4, 1186)$

we can reject H_0. We can make sure that $\beta_1, \beta_2, \beta_3$ and β_4 are not simultaneously zero

» using t-test

For β_0 $H_0 : \beta_0 = 0$
$H_a : \beta_0 \neq 0$

For β_4 $H_0 : \beta_4 = 0$
$H_a : \beta_4 \neq 0$

For β_1 $H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

For β_2 $H_0 : \beta_2 = 0$
$H_a : \beta_2 \neq 0$

For β_3 $H_0 : \beta_3 = 0$
$H_a : \beta_3 \neq 0$

For β_0 : $t_{cal}(\beta_0) = \frac{118.0741 - 0}{3.500} = 33.7355$: Reject H_0

For β_1 : $t_{cal}(\beta_1) = \frac{-0.5895 - 0}{0.1106} = -5.3300$: Reject H_0

For β_2 : $t_{cal}(\beta_2) = \frac{0.0538 - 0}{0.0367} = 1.4659$: Not reject H_0

For β_3 : $t_{cal}(\beta_3) = \frac{0.4937 - 0}{0.2833} = 1.7427$: Not reject H_0

For β_4 : $t_{cal}(\beta_4) = \frac{-0.4379 - 0}{0.3197} = -1.3697$: Not reject H_0

>> from $\alpha = 0.05 \longrightarrow t_{lower} = -1.960$ $t_{upper} = 1.960$

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 $\alpha = 0.05$)

We use "marginal contribution" to make sure that we should add parent education into model or not

> Excluding : $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + u_i$

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

1) H_0 : parent education has no marginal contribution to the model
H_a : otherwise

2) from $F_{cal} = \frac{R^2_{new} - R^2_{old} / (\text{number of new regressors})}{1 - R^2_{new} / (n - k_{new})} = \frac{(0.0328 - 0.0301)/2}{(1-0.0328)/(1191-5)} = 1.6554$

3) $\alpha = 0.05$, $F_{upper,\alpha}(2, 1186) = 3.00$

$F_{cal} < F_{upper}$

∴ The addition of parent education has no marginal contribution to the model (cannot reject H_0)

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 = SS/df		
Model	_____	k-1	_____	Number of obs	428
Residual	_____	n-k	.446526442	F(___,_____)	13.19
Total	223.327441	n-1	_____	Prob > F	0.0000
				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?

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		
ESS Model	35.3304012	6	5.8884	Number of obs	= 428
RSS Residual	187.99704	421	.446526442	F(6 , 421)	= 13.19
				Prob > F	= 0.0000
				R-squared	= 0.1582
				Adj R-squared	= 0.1462
TSS Total	223.327441	427	0.5230	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.

df for ESS = k-1 = 7-1 = 6

df for RSS = n-k = 428-7 = 421

df for TSS = n-1 = 428-1 = 427

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

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

$0.1582 = \frac{ESS}{223.3274441}$

ESS = 35.3304012

from TSS = RSS + ESS

233.3274441 = RSS - 35.3304012

RSS = 187.99704

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

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

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

$$\bar{R}^2 = 0.1462$$

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 $F_{1,421} = 3.84$)

According to marginal contribution, H_0 : new variable has no marginal contribution
H_a : otherwise

$$F_{cal} = \frac{R^2_{3.1} - R^2_{3.2} / (\text{number of new regressors})}{1-R^2_{3.1} / (n-k_{3.1})}$$

$$3.84 < \frac{0.1582 - R^2_{3.2} / 5}{(1-0.1582)/(428-7)} \quad ;\ \text{we need to find } F_{cal} > F_{cri} \text{ because we want to reject } H_0$$

$$R^2_{3.2} < 0.1198$$

Hence, the maximum value of R^2 of model 3.2 is 0.1198

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?

From the model, since age is not significantly different from zero, age doesn't impact natural log of hourly wage of married women.

In my opinion, it doesn't make economic sense since age and hourly wage are substituted each other. When we're older, we need to have higher wage because we have more experiences leading to more specialize in that job

However, in this model, the cause of insignificance is multicollinearity problem since the model already consists of year of experience (exp_i), year of experience squared and year of education So, when put age in model it doesn't add more information.