

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_ชื่อ. **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:

$$\log(\text{salary}_i) = \beta_0 + \beta_1 \log(\text{sales}_i) + \beta_2 \text{ROE}_i + \beta_3 \text{finance}_i + \beta_4 \text{consprod}_i + \beta_5 \text{utility}_i + u_i$$

where

$\log(\text{salary}_i)$	= logarithm of CEO annual salary (in 1,000 USD)	
$\log(\text{sales}_i)$	= 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	} dummies
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.

↳ case that all dummies = 0

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

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

- Write out the estimated regression equation for $\log(\text{salary}_i)$. Interpret the estimated coefficient associated with $\log(\text{sales}_i)$.
- 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.
- Compute the approximate percentage difference in estimated salary between the utility and transportation sector, holding sales_i and ROE_i fixed.
- 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?
- In the above model, is there any benefit if we add interaction terms between roe and sector dummies, i.e. $\text{ROE}_i * \text{finance}_i$ and/or $\text{ROE}_i * \text{consprod}_i$ and/or $\text{ROE}_i * \text{utility}_i$?

- a. Write out the estimated regression equation for $\log(\text{salary}_i)$. Interpret the estimated coefficient associated with $\log(\text{sales}_i)$.

$$Y_i: \log(\text{Salary}_i) = 4.5881 + 0.2572 \log(\text{sales}_i) + 0.0112 \text{ROE}_i + 0.1580 \text{finance}_i + 0.1809 \text{consprod}_i - 0.2830 \text{utility}_i$$

Interpret estimated coefficient of $\log(\text{sales}_i) \rightarrow \log\text{-log form: 1\% increase in } X_i, \text{ on average, increase } Y_i \text{ by } \hat{\beta}_1\%$.

\therefore At 1% increase in firms' sale, on average, increases the CEO annual salary by 0.2572% given other factors remain constant.

- 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

step 1: state hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

H_a : Not all the slope coefficients are simultaneously equal to zero.

Step 2: calculate test statistics

$$F_{\text{cal}} = \frac{\text{ESS}/df}{\text{RSS}/df} = \frac{23.8109943/5}{92.911689/203} = 22.5285$$

Step 3: State decision rule

$$\alpha = 0.01$$

$$\begin{aligned} \text{upper bound} &= F_{\alpha}(k-1, n-k) \\ &= F_{0.01}(5, 209-6) \\ &= F_{0.01}(5, 203) = 3.11 \end{aligned}$$

Step 4: Conclude the test

$F_{\text{cal}} > \text{upper bound}$. The overall significance of the regression is at least 99%.

$$\text{critical value} = \pm t_{\frac{\alpha}{2}, n-k}$$

$$= \pm t_{0.025, 203}$$

$$= \pm 1.96$$

If $t > 1.96$ or $t < -1.96$, $\hat{\beta}_i$ will be a statistically significant.

So, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_4, \hat{\beta}_5$ are statistically significant at 5% level.

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

$$\log(\text{salary}_i) = 4.5881 + 0.2572 \log(\text{sales}_i) + 0.0112 ROE_i + 0.1580 \text{finance}_i + 0.1809 \text{consprod}_i - 0.2830 \text{utility}_i$$

$$\begin{aligned} \text{Utility: } E(\log(\text{Salary}_i) | \text{finance}_i = 0, \text{consprod}_i = 0, \text{utility}_i = 1) &= 4.5881 + 0.2572 \log(\text{Sales}_i) + 0.0112 ROE_i \\ &\quad + 0.1580(0) + 0.1809(0) - 0.2830(1) \\ &= 4.3051 + 0.2572 \log(\text{sales}_i) + 0.0112 ROE_i \end{aligned}$$

$$\begin{aligned} \text{Transport: } E(\log(\text{salary}_i) | \text{finance}_i = 0, \text{consprod}_i = 0, \text{utility}_i = 0) &= 4.5881 + 0.2572 \log(\text{sales}_i) + 0.0112 ROE_i + 0.1580(0) + 0.1809(0) - 0.2830(0) \\ &= 4.5881 + 0.2572 \log(\text{sales}_i) + 0.0112 ROE_i \end{aligned}$$

$$\text{Different in } \log(\text{Salary}_i) = [4.3051 + 0.2572 \log(\text{sales}_i) + 0.0112 ROE_i] - [4.5881 + 0.2572 \log(\text{sales}_i) + 0.0112 ROE_i] = -0.2830$$

- 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 in the equation because we will face a perfect multicollinearity problem.

If we put all the sector dummies in the equation and use STATA run the regression, STATA would drop one of the dummy variables.

- 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 $ROE_i * utility_i$?

There is no benefit to add interaction terms between roe and sector dummies because ROE have no effect to CEO annual Salary in each sector.

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 ($fatheduc_i$), and mother's year of education ($motheduc_i$). 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		
Model	14536.9538	2	7268.47691	Number of obs =	1191
Residual	468209.738	1188	394.115941	F(2, 1188) =	18.44
Total	482746.692	1190	405.669489	Prob > F =	0.0000
				R-squared =	0.0301
				Adj R-squared =	0.0285
				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: $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 fatheduc_i + \beta_4 motheduc_i + u_i$

regress bwght cigs faminc fatheduc motheduc					
Source	SS	df	MS		
Model	15827.6593	4	3956.91482	Number of obs =	1191
Residual	466919.033	1186	393.69227	F(4, 1186) =	10.05
Total	482746.692	1190	405.669489	Prob > F =	0.0000
				R-squared =	0.0328
				Adj R-squared =	0.0295
				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
- $fatheduc_i$ = father's years of education
- $motheduc_i$ = mother's years of education

Answer the following questions.

- Based on **Model 2.1**, test whether smoking has an impact on birth weight. Show your work. (use $\alpha = 0.05$)
- Based on **Model 2.1**, construct a 99% confidence interval for β_2 .
- Would your conclusion in a) change if you use the result from **Model 2.2**? Show your work. (use $\alpha = 0.05$)
- 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.
- 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 $\alpha = 0.05$)

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

using t-test

step 1 : state hypothesis

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

Step 2 : Calculate test statistics

$$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{-0.5876985 - 0}{0.1090181} = -5.3908$$

Step 3 : State decision rule

$$\alpha = 0.05$$

$$df = n - k$$

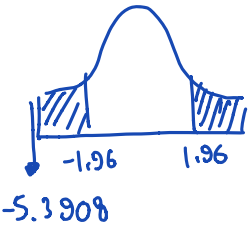
$$df = 1191 - 3 = 1189$$

$$\begin{array}{l} \text{upper bound} = + t_{\frac{\alpha}{2}, n-k} = + t_{0.025, \infty} = 1.96 \\ \text{lower bound} = - t_{\frac{\alpha}{2}, n-k} = - t_{0.025, \infty} = -1.96 \end{array}$$

Step 4: conclude the test

$t_{cal} < \text{lower bound}$. we rejected H_0 at 0.05 significance level.

That is $\beta_1 \neq 0$. Smoking has a significant impact on infant birth weight



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

$$df = n - k = 1191 - 3 = 1188$$

$$\begin{aligned} 99\% \text{ CI: } \hat{\beta}_2 \pm \left(t_{\frac{0.01}{2}, 1188} \cdot se(\hat{\beta}_2) \right) &= 0.0624684 \pm (2.576 \cdot 0.0324438) \\ &= 0.0624684 \pm 0.08357523 \\ &= (0.0211, 0.1460) \end{aligned}$$

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

t-test

step 1: state hypothesis

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

step 2: calculate test statistics

$$\begin{aligned} t_{cal} &= \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{-0.5894954 - 0}{0.1106172} \\ &= -5.3291 \end{aligned}$$

step 3: state Decision rule

$$\alpha = 0.05 \quad df = n - k = 1191 - 5 = 1186$$

$$\text{upper bound} = + t_{\frac{\alpha}{2}, n-k} = + t_{0.025, \infty} = 1.96$$

$$\text{lower bound} = - t_{\frac{\alpha}{2}, n-k} = - t_{0.025, \infty} = -1.96$$

step 4: conclude test

$t_{cal} < \text{lower bound}$ we reject H_0 at 0.05 significance level.

That is $\beta_1 \neq 0$.

Smoking has a significant impact on infant birth weight.

The conclusion does not change from (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.

Step 1: state hypothesis (F-test)

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

H_a : other wise

Step 2: calculate test statistics

$$F_{cal} = \frac{\beta^2 / k - 1}{(1 - R^2) / (n - k)} = \frac{0.0328 / (5 - 1)}{(1 - 0.0328) / (1191 - 5)} = 10.05$$

Step 3: State decision Rule

$$\alpha = 0.01$$

$$\text{upperbound} = F_{\alpha} (k - 1, n - k)$$

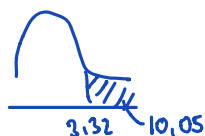
$$= F_{0.01} (4, 1186)$$

$$= 3.32$$

Step 4: conclude the test

$F_{cal} > \text{upper bound}$. We reject H_0 at 0.01 significance level.

That is the overall significance is at least 99%.



$$\text{critical value} = \pm t_{\frac{\alpha}{2}, n - k} = \pm t_{0.025, 1186} = \pm 1.96$$

$$t_{cal} = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)}$$

$$t_{cal}(\hat{\beta}_1) = \frac{-0.5894954 - 0}{0.1106172} = -5.3291 < -1.96$$

$$t_{cal}(\hat{\beta}_2) = \frac{0.0538234 - 0}{0.0366502} = 1.4686$$

$$t_{cal}(\hat{\beta}_3) = \frac{0.4936695 - 0}{0.2832896} = 1.7426$$

$$t_{cal}(\hat{\beta}_4) = \frac{-0.4379234 - 0}{0.3197377} = -1.3696$$

$$t_{cal}(\hat{\beta}_0) = \frac{118.0741 - 0}{3.500291} = 33.73 > 1.96$$

At 5% level of significance, $\hat{\beta}_1$ and $\hat{\beta}_0$ are statistically significant.

- 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 $\alpha = 0.05$)

Step 1: state Hypothesis

H_0 : parents education does not have an impact on birth weight ($\beta_3 = \beta_4$)

H_a : other wise

Step 2: calculate test statistics

$$F_{cal} = \frac{(ESS_{incl} - ESS_{excl}) / \text{no. of additional } X}{RSS_{incl} / (n - k_{incl})}$$

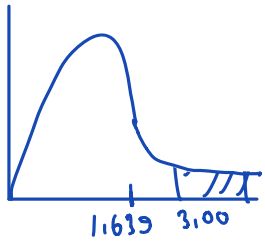
$$= \frac{(15827.6593 - 14536.9538) / 2}{466919.033 / (1191 - 5)} = 1.639$$

Step 3: State Decision Rule

$$\alpha = 0.05$$

upper bound = $F_{\alpha}(\text{no. of additional } X, n - k_{incl}) = F_{0.05}(2, 1186) = 3.00$

Step 4 conclude the test



$F_{cal} < \text{upper bound}$. We fail to reject H_0 at 0.05 significance level.
That is, parents' education does not have an impact on birth weight.
The impact of education on birth weight might already be reflected in variable family income.

Question 3. A model of wage equation is given by

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

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	_____	_____	_____	Number of obs =	428	
Residual	_____	_____	.446526442	F(____, _____) =	13.19	
Total	223.327441	_____	_____	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

- Figure out all the degrees of freedom in this model.
- Figure out all the sum of squares (ESS and RSS) and mean squares in this model.
- Figure out the adjusted R-squared (\bar{R}^2)
- 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 $F_{1,421} = 3.84$)
- 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.

$$\text{df of ESS} = k - 1 = 7 - 1 = 6$$

$$\text{df of RSS} = n - k = 428 - 7 = 421$$

$$\text{df of TSS} = n - 1 = 428 - 1 = 427.$$

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

$$MS = \frac{SS}{df}$$

$$\text{Residual } MS = \frac{RSS}{\text{df of RSS}}$$

$$0.446526442 = \frac{RSS}{n - k}$$

$$RSS = 0.446526442(428 - 7)$$

$$RSS = 187,9896$$

$$ESS = TSS - RSS = 223.329441 - 187.9896$$

$$= 35.3398$$

$$\text{Model } MS = TSS/n - 1 = \frac{223.329441}{428 - 1}$$

$$= 0.5230$$

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

$$R^2 = 1 - \left[(1 - R^2) \left(\frac{n - 1}{n - k} \right) \right]$$

$$= 1 - \left[(1 - 0.1582) \left(\frac{427}{421} \right) \right]$$

$$= 1 - 0.853797 = 0.1462$$

- 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 $F_{1,421} = 3.84$)

Model 3.1 (incl)

Model 3.2 (excl)

H_0 : variable excluded in Model 3.2 is not significant

H_a : otherwise

To reject H_0 , $F_{cal} > 3.84$

$$\frac{(R^2_{incl} - R^2_{excl}) / \text{no. of additional X}}{(1 - R^2_{incl}) / (n - k_{incl})} > 3.84$$

$$\frac{(0.1582 - R^2_{3.2}) / 1}{(1 - 0.1582) / (428 - 7)} > 3.84$$

$$0.1582 - R^2_{3.2} > 3.84$$

$$0.00199952$$

$$R^2_{3.2} < 0.1505$$

\therefore The maximum value of R^2 from model 3.2 that would lead the conclusion that the excluded variable is significant at 0.05 significance level is 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 my opinion age should affect hourly wage but it does not affect directly. Age does have correlation with other variables such as experience or education so the effect age has on wage is non-linear and isn't shown in the Model.