

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_น้อย. **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
- 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
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 $\log(\text{sales}_i)$.

$$\log(\text{salary}_i) = 4.588401 + 0.2571917 \log(\text{sales}_i) + 0.0111517 \text{ROE}_i + 0.1579564 \text{finance}_i + 0.1808917 \text{Consprod}_i - 0.2830015 \text{utility}_i$$

$\log(\text{sales}_i)$ interpret that a 1% increase in firms' sale increase CEO annual salary by 0.2571917%, holding other thing 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.

using f-test

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_5$$

H_1 : Not all the slope coefficients are simultaneously zero (otherwise)

$$F_{\text{cal}} = \frac{\text{ESS} / k - 1}{\text{RSS} / n - k} = \frac{4.76219889}{0.211385068} = 22.5285$$

$$\alpha = 0.05$$

$\therefore F_{\text{cal}} > F_{\text{cri}}$ - we can reject H_0 .

$$F_{\text{upper}, \alpha(5, 203)} = 2.21$$

We can make sure that $\beta_1, \beta_2, \beta_3, \beta_4$ and β_5 are not simultaneously zero.

$\therefore$ All independent variables can significantly explain the dependent variable.

$$\alpha = 0.05$$

t_{cal}

$$d.f = 203$$

1. $\log(\text{sales})$

$$8.03$$

$\therefore$ Reject H_0 , significance

$$t_{\text{upper}} = 1.96$$

2. roe

$$2.59$$

$\therefore$ Reject H_0 , significance

$$t_{\text{lower}} = -1.96$$

3. finance

$$1.99$$

$\therefore$ can't reject H_0 , not significance

4. consumer product

$$2.13$$

$\therefore$ Reject H_0 , significance

5. utility

$$-2.85$$

$\therefore$ Reject H_0 , significance

6. constant

$$15.55$$

$\therefore$ 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**.

$$1. E(\log(\text{salary}_i) | \text{utility}_i = 1; \text{finance}_i = 0; \text{consprod}_i = 0) = \beta_0 + \beta_3 + \beta_4 + \beta_5$$

$$= 4.588101 + 0.1579564(0) + 0.1808919(0) - 0.2830015(1)$$

$$= 4.3050995 : \text{utility}$$

$$2. E(\log(\text{salary}_i) | \text{utility}_i = 0; \text{finance}_i = 0; \text{consprod}_i = 0) = \beta_0 + \beta_3 + \beta_4 + \beta_5$$

$$= 4.588101 + 0.1579564(0) + 0.1808919(0) - 0.2830015(0)$$

$$= 4.588101 : \text{transportation}$$

$$\Delta\% = \frac{4.588101 - 4.3050995}{4.3050995} \times 100 = 6.5736\%$$

- 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?

if put all sector dummies, assuming that β_6 is coefficient of transport,

$$\log(\text{salary}_i) = \beta_0 + \beta_1 \log(\text{sale}_i) + \beta_2 \text{ROE}_i + \beta_3 \text{finance}_i + \beta_4 \text{consprod}_i + \beta_5 \text{utility}_i + \beta_6 \text{transport}_i + u_i$$

if run the regression; $E(\log(\text{salary}_i) | \text{finance}_i = 0, \text{consprod}_i = 0, \text{utility}_i = 0, \text{transport}_i = 0)$

that no observation because it is impossible that all of the dummies are the same group of the sample. if the dummies are all zero, that mean it does not have $\beta_6 = 0$ (otherwise) for transport.

So, we can't put all the sector dummies, we only need $n-1$ dummy variables to incorporate n groups of the sample. In this case, $n = 4$ (finance_i , consprod_i , utility_i , transport_i) and then $n-1 \rightarrow 4-1 = 3$, so we need only 3 dummies.

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 $ROE_i * utility_i$?

$$\log(salary_i) = \beta_0 + \beta_1 \log(sale_i) + \beta_2 ROE_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility_i + \beta_6 (ROE_i * finance_i) + \beta_7 (ROE_i * consprod_i) + \beta_8 (ROE_i * utility_i) + u_i$$

$$\begin{aligned} E(\log(salary_i) | finance_i = 1) &= \beta_0 + \beta_1 \log(sale_i) + \beta_2 ROE_i + \beta_3 + \beta_4 consprod_i + \beta_5 utility_i + \beta_6 ROE_i + \beta_7 (ROE_i * consprod_i) + \beta_8 (ROE_i * utility_i) \\ &= \beta_0 + \beta_3 + \beta_1 \log(sale_i) + (\beta_2 + \beta_4) ROE_i + \beta_4 consprod_i + \beta_5 utility_i + \beta_7 (ROE_i * consprod_i) + \beta_8 (ROE_i * utility_i) \end{aligned}$$

$$\begin{aligned} E(\log(salary_i) | consprod_i = 1) &= \beta_0 + \beta_4 + \beta_1 \log(sale_i) + (\beta_2 + \beta_7) ROE_i + \beta_3 finance_i + \beta_5 utility_i + \beta_6 (ROE_i * finance_i) + \beta_8 (ROE_i * utility_i) \end{aligned}$$

$$\begin{aligned} E(\log(salary_i) | utility_i = 1) &= \beta_0 + \beta_5 + \beta_1 \log(sale_i) + \beta_3 finance_i + \beta_4 consprod_i + (\beta_2 + \beta_8) utility_i + \beta_6 (ROE_i * finance_i) + \beta_7 (ROE_i * consprod_i) \end{aligned}$$

$\therefore$ if we add interaction terms between roe and sector dummies, the y-intercept of CEO salary regression will be increased with slope that also change to be steeper or flatter. So, we can test that both the intercepts and slopes are significantly different simultaneously 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 ($fatheduc_i$), 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
				Prob > F =	0.0000
				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: $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 fatheduc_i + \beta_4 motheduc_i + u_i$

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
				Prob > F =	0.0000
				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
 $fatheduc_i$ = father's years of education
 $motheduc_i$ = 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 $\alpha = 0.05$)

$$\text{bwght}_i = 118.5568 - 0.5876985 \text{ cig}_i + 0.0624684 \text{ faminc}_i$$

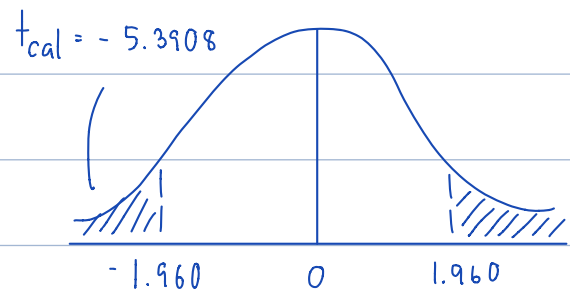
$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

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

$$\alpha = 0.05 ; t_{\text{upper}} = 1.960$$

$$d.f = 1191 - 3 \quad t_{\text{lower}} = -1.960$$



$\therefore t_{\text{cal}}$ lie beyond any rejection region, we can reject H_0 that β_1 is not zero.
It mean that smoking has impact on birth weight.

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

$$\hat{\beta}_2 + t_{\frac{\alpha}{2}, n-3} \cdot \text{se}(\hat{\beta}_2) ; 0.0624684 \pm t_{0.005, 1188} \cdot (0.0324438)$$

$$0.0624684 \pm 2.576 \cdot (0.0324438)$$

$$P(-0.0211 < \beta_2 < 0.1460) = 0.99$$

$\therefore$ Given the confidence β_2 of 99%, the true parameter is between -0.0211 and 0.1460.

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

$$\text{bwght}_i = 118.0741 - 0.5894954 \text{cig}_i + 0.0538254 \text{faminc}_i + 0.4936695 \text{fatheduc}_i - 0.4379234 \text{motheduc}_i$$

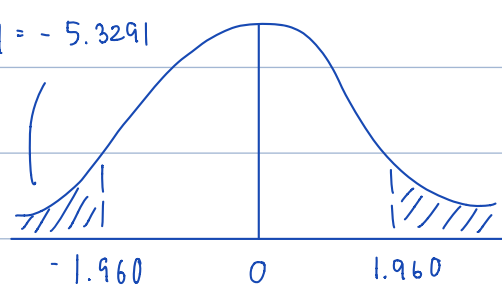
$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

$$t_{\text{cal}}(\beta_1) = \frac{\hat{\beta}_1 - \beta_1}{\text{se}_{\hat{\beta}_1}} = \frac{-0.5894954 - 0}{0.1106172} = -5.3291$$

$$\alpha = 0.05; t_{\text{upper}} = 1.960$$

$$d-f = 1191-5 \quad t_{\text{lower}} = -1.960$$



t_{cal} lie beyond any rejection region, we can reject H_0 that β_1 is not zero

It mean that smoking has impact on birth weight.

$\therefore$ the conclusion is the same as Model 2.1.

- 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 F-test

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

H_1 : Not all slope coefficients are simultaneously zero

$$F_{\text{cal}} = 10.05 \quad \therefore F_{\text{cal}} > F_{\text{cri}}, \text{ we can reject } H_0. \text{ We can make sure}$$

$$F_{\text{cri}} = 2.37 \quad \text{that } \beta_1, \beta_2, \beta_3, \beta_4 \text{ are not simultaneously zero.}$$

$$t\text{-test}: \alpha = 0.05, t_{\text{upper}} = 1.96, t_{\text{lower}} = -1.96$$

t - test

$$H_0: \beta_0 = 0 \quad t_{cal} = \frac{118.0791}{3.500291} = 33.7326 \quad \therefore \text{reject } H_0, \text{ significance}$$

$$H_1: \beta_0 \neq 0$$

$$H_0: \beta_1 = 0 \quad t_{cal} = \frac{-0.5894954}{0.1106172} = -5.329 \quad \therefore \text{reject } H_0, \text{ significance}$$

$$H_1: \beta_1 \neq 0$$

$$H_0: \beta_2 = 0 \quad t_{cal} = \frac{0.538259}{0.366502} = 1.4686 \quad \therefore \text{can't reject } H_0, \text{ not significance}$$

$$H_1: \beta_2 \neq 0$$

$$H_0: \beta_3 = 0 \quad t_{cal} = \frac{0.4936695}{0.2832896} = 1.7426 \quad \therefore \text{can't reject } H_0, \text{ not significance}$$

$$H_1: \beta_3 \neq 0$$

$$H_0: \beta_4 = 0 \quad t_{cal} = \frac{-0.4399234}{0.3197377} = -1.3696 \quad \therefore \text{can't reject } H_0, \text{ not significance}$$

$$H_1: \beta_4 \neq 0$$

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 $\alpha = 0.05$)

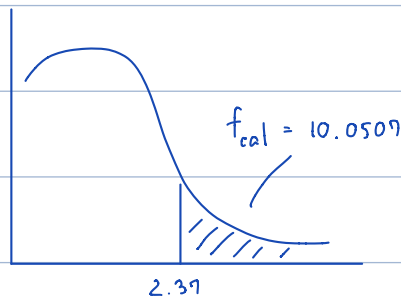
$$H_0: \beta_3 = \beta_4 = 0$$

$$H_1: \text{Not all the slope coefficients are simultaneously zero (otherwise)}$$

$$F_{cal} = \frac{ESS/u-1}{RSS/n-k} = \frac{3956.91482}{393.69227} = 10.0507$$

$$\alpha = 0.05$$

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



$\therefore f_{cal}$ lie beyond any rejection region, we can reject H_0 that β_3 and β_4 are not simultaneously zero.

Parent education has an impact on birth weight at all.

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	Number of obs = 428		
Model				F(6 , 421) = 13.19		
Residual			.446526442	Prob > F = 0.0000		
				R-squared = 0.1582		
				Adj R-squared =		
Total	223.327441			Root MSE = .66823		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.039819	.013393	2.97	0.003	.0134936	.0661444
expsq	-.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.

$$n = 428, \quad k = 7$$

$$k - 1 = 6 \quad \#$$

$$n - k = 421 \quad \#$$

$$n - 1 = 427 \quad \#$$

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

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

$$ESS = TSS - RSS$$

$$35.3304 = 223.321441 - RSS$$

$$0.1582 = \frac{ESS}{223.321441}$$

$$RSS = 187.9970 \quad \#$$

$$ESS = 35.3304 \quad \#$$

$$MS_{(E)} = \frac{ESS}{k - 1}$$

$$= \frac{35.3304}{6} = 5.8884 \quad \#$$

$$MS_{(R)} = \frac{RSS}{n - k}$$

$$= \frac{187.9970}{421} = 0.4465 \quad \#$$

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

$$\bar{R}^2 = 1 - \frac{(1 - R^2) \cdot n - 1}{n - k} = 1 - \frac{(1 - 0.1582) \cdot 427}{421} = 0.1462 \quad \#$$

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

H_0 : the excluded has no marginal contribution to the model

H_1 : otherwise

$$F_{(1,421)} < \frac{R^2_{3.1} - R^2_{3.2} / \text{number of new regression}}{1 - R^2_{3.1} / n - k_{3.1}}$$

$$3.84 < \frac{0.1582 - R^2_{3.2} / 1}{1 - 0.1582 / 428 - 1}$$

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

$\therefore$ To reject H_0 , F_{cal} need to be more than F_{cri} . So $R^2_{3.2}$ need to not exceed 0.1505, the maximum value of $R^2_{3.2}$ is $R^2_{3.2} \in (-\infty, 0.1505)$.

- 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?**

For about the economic sense, in the reality, age is the major factor of wage that you will receive. People with high age cannot work as people who have lower age that the wage of high age people will lower than young people. On the other hand, higher age people can get more wage than young people due to the work experience that increase with age. So, age determine the hourly wage.

In this model, age is insignificant because there has multicollinearity with other terms. Age variable is not necessary.