

## 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)
- $\text{ROE}_i$  = 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)
- $\text{finance}_i$  = 1 if in financial industry, = 0 otherwise
- $\text{consprod}_i$  = 1 if in consumer product industry, = 0 otherwise
- $\text{utility}_i$  = 1 if in utility industry, = 0 otherwise

( $\text{finance}_i$ ,  $\text{consprod}_i$ , and  $\text{utility}_i$  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  $\text{sales}_i$  and  $\text{ROE}_i$  fixed.
- Why can't we put all the sector dummies (i.e.  $\text{finance}_i$ ,  $\text{consprod}_i$ ,  $\text{utility}_i$  and  $\text{transport}_i$ ) 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  $\text{ROE}_i * \text{utility}_i$ ?

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

since  $\text{finance}_i$ ,  $\text{consprod}_i$ , and  $\text{utility}_i$  are binary variables

$D_{3i} = 1$  if in financial industry,  $= 0$  otherwise

$D_{4i} = 1$  if in consumer product industry,  $= 0$  otherwise

$D_{5i} = 1$  if in utility industry,  $= 0$  otherwise

Hence, the estimated regression equation is

$$\log(\text{salary}_i) = 4.588101 + 0.2571917 \log(\text{sales}_i) + 0.0111517 \text{ROE}_i + 0.1579564 D_{3i} + 0.1808917 D_{4i} - 0.2830015 D_{5i} + u_i$$

In the sense of  $\log(\text{sales}_i)$ , if there is an increase in sales by one percent, the salary will be increased by 0.2571917 as the value of  $\beta_1$ .

1.6 since  $\text{Prob} > F = 0.0000$ , the overall significance of regression is about 99%.

By using the p-value of t-test with the level of significance at 5% and degree of freedom  $n-k = 209-6 = 203$ , the critical value  $\approx \pm 1.960$ . " $k = 6: \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ "

1.7 since the transportation sector is baseline for this model, the approximate percentage difference in estimated salary between the utility and transportation sector is about 28.30015%.

1.8 If we put all the variables to the equation without any dummy variables ( $\text{dummies} = 0$ ), there will be no baseline and sector for the regression. Moreover, the STATA will be systematically omitted.

1.9 If we add, for instance,  $\text{ROE} \cdot \text{finance}$  to the model, the average salary in finance sector will be much higher than those of transportation sector in percentage.

**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	Number of obs = 1191	
Model	14536.9538	2	7268.47691	F( 2, 1188) = 18.44	
Residual	468209.738	1188	394.115941	Prob > F = 0.0000	
Total	482746.692	1190	405.669489	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	Number of obs = 1191	
Model	15827.6593	4	3956.91482	F( 4, 1186) = 10.05	
Residual	466919.033	1186	393.69227	Prob > F = 0.0000	
Total	482746.692	1190	405.669489	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.



- 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  $\beta_2$ .
- c. Would your conclusion in a) change if you use the result from **Model 2.2**? Show your work. (use  $\alpha = 0.05$ )
- 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  $\alpha = 0.05$ )

$$2, a \quad H_0 : \hat{\beta}_1 = 0$$

$$H_a : \hat{\beta}_1 \neq 0$$

$$\alpha = 0,05$$

$$t_{cal} = \frac{-0,5876985}{0,1090181} = -5,3908$$

$$d.f. = n - k ; k = 3$$

$$= 1191 - 3 = 1188$$

the critical value of  $t_{\alpha/2} = \pm 1,960$

as a result,  $t_{cal} > \text{critical value}$

$t_{cal}$  is in the area of rejection, meaning that  $H_0$  is rejected by 95% of the time we are sure the smoking affects birth weight.

$$2, b \quad \alpha = 0,01 \quad t_{\alpha/2} = 2,576 \quad \text{where } d.f. = 1188$$

$$[\hat{\beta}_2 - t_{\alpha/2} \text{se}(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} \text{se}(\hat{\beta}_2)] = 1 - \alpha$$

$$[0,0624684 - (2,576)(0,0324438) \leq \beta_2 \leq 0,0624684 + (2,576)(0,0324438)] = 1 - 0,01$$

$$[-0,02110683 \leq \beta_2 \leq 0,14604363] = 0,99$$

$$2, c \quad \bar{R}^2 = 1 - (1 - R^2)(n-1)/(n-k)$$

$$= 1 - (1 - 0,0301)(1190)/(1188)$$

$$= 0,0285$$

2, d since  $\text{Prob} > F = 0,0000$ , the overall significance of the regression is 99% significant

$$\alpha = 0,05 \quad d.f. = n - k = 1191 - 5 = 1186$$

$$\text{the critical value from } F\text{-test} = F_{0,05,4,1186} = 2,37$$

$$\text{the critical value from } t\text{-test} = t_{\alpha/2, 1186} = \pm 1,960$$

$$t_{cal} : \hat{\beta}_0 = 118,0741/3,500291 = 33,7327 > 1,960 ; \text{significant}$$

$$\hat{\beta}_1 = -0,5894954/0,1106172 = -5,3291 > 1,960 ; \text{significant}$$

$$\hat{\beta}_2 = 0,0538254/0,0366502 = 1,4686 \quad \text{where } -1,960 < \beta_2 = 1,4686 < 1,960 ; \text{not significant}$$

$$\hat{\beta}_3 = 0,4936695/0,2832896 = 1,7426 \quad \text{where } -1,960 < \beta_2 = 1,7426 < 1,960 ; \text{not significant}$$

$$\hat{\beta}_4 = -0,4379234/0,3197377 = -1,3696 \quad \text{where } -1,960 < \beta_2 = 1,3696 < 1,960 ; \text{not significant}$$

2.e.  $H_0$ : father education & math education has no marginal contribution to the model.

$H_a$ : otherwise

$$F_{\text{cal}} = \frac{\text{ESS}_{\text{new}} - \text{ESS}_{\text{old}}}{\text{RSS}_{\text{new}} / (n - K_{\text{new}})} \quad (\text{number of new regressors})$$

$$= \frac{15827.6503 - 14536.0538}{466919.033 / (1191 - 5)} \quad \frac{2}{393.692271} = 1.63923145$$

$$F_{\text{upper } \alpha}(2, 1186) = 3.00 > F_{\text{cal}} \quad \alpha = 0.05$$

$\therefore$  95% of the time that the education of father and mother has no marginal contribution to the model has no impact on birth weight.

**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(____,____) = 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?

Source	SS	df	MS
Model	34.893283	6	5.815547167
Residual	188.434158	421	.446526442
Total	223.327441	427	0.523015084

Number of obs = 428  
 F(6, 421) = 13.19  
 Prob > F = 0.0000  
 R-squared = 0.1582  
 Adj R-squared = 0.1462  
 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

3.c  $\bar{R}^2 = 1 - (1 - R)(n-1)/(n-k)$

$$= 1 - (1 - 0.1582)(428-1)/(428-7)$$

$$= 0.1462$$

3.d let new = model 3.1  
old = model 3.2

$H_0$ : the new regressor has no marginal contribution to the model.

$H_a$ : otherwise

$$F_{cgl} = \frac{R_{new}^2 - R_{old}^2}{\text{number of regressor}} \div \frac{1 - R_{new}^2}{(n - k_{new})}; \text{ given } F_{cgl} = 3.84$$

$$\frac{0.1582 - R_{old}^2}{1} > 3.84$$

$$(1 - 0.1582)/(428 - 7)$$

$$\frac{0.1582 - R_{old}^2}{0.0020} > 3.84$$

$$0.1582 - R_{old}^2 > 0.00768$$

$$R_{old}^2 < 0.15052$$

3.e No, age and experience are correlated as multicollinearity and cause of insignificance.