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 $\text{ROE}_i * utility_i$?

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.

มีทั้งแบบ sig & ไม่ sig

t-test

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 $\alpha = 0.05$)

To see if Model 2.2 is valid

→ ดูก่อน model ว่า compare

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              Number of obs =     428
----------+------------------------------           F(  6,  421   ) =   13.19
    Model |  ___________    6 v k-1  ___________     Prob > F      =  0.0000
 Residual |  ___________ n-k-1 421   .446526442     R-squared     =  0.1582
----------+------------------------------           Adj R-squared =  ______
    Total |  223.327441    427       ___________     Root MSE      =  .66823
                           1 n-1
------------------------------------------------------------------------------
       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?

1. (a) From stata

$\log(\text{salary}_i) = \beta_0 + \beta_1 \log(\text{sales}_2) + \beta_2\ \text{ROE}_i + \beta_3 \text{finance}_i + \beta_4 \text{consprod.} + \beta_5 \text{utility} + u_i$

so we get, $\log(\text{salary}_i) = 4.6881 + 0.2572 \log(\text{sales}_i) + 0.0112\ \text{ROE}_i + 0.1580 \text{finance}_i + 0.1809 \text{consprod}_i - 0.2830 \text{utility}_i$

(b) Use t-test to find coefficients that are individually statistically significant.

Step 1: State Hypothesis

β_0: $H_0 : \beta_0 = 0$, $H_1 : \beta_0 \neq 0$

β_1: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$

β_2: $H_0 : \beta_2 = 0$, $H_1 : \beta_2 \neq 0$

β_3: $H_0 : \beta_3 = 0$, $H_1 : \beta_3 \neq 0$

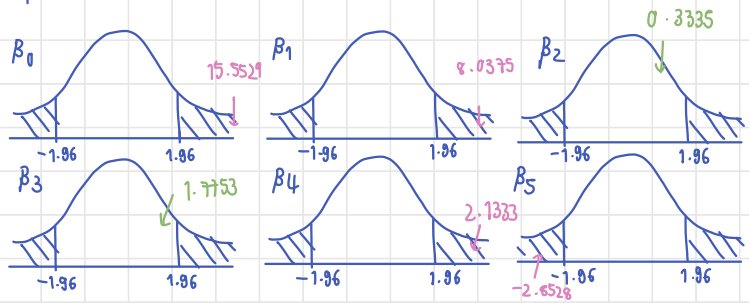
β_4: $H_0 : \beta_4 = 0$, $H_1 : \beta_4 \neq 0$

β_5: $H_0 : \beta_5 = 0$, $H_1 : \beta_5 \neq 0$

Step 3: Decision Rule: $\alpha = 0.05$, $df = n - k = 203$ (given)

$t_{upper} : 0 + 1.96 = 1.96$; $t_{lower} : 0 - 1.96 = -1.96$

Step 4: Conclusion



We cannot reject the null hypothesis for β_2 & β_3, so β_2 and β_3 is not statistically significant at $\alpha = 0.05$

Step 2:

$t_{cal}(\beta_0) = \frac{\hat{\beta}_0 - \beta_0}{se} = \frac{4.5881}{0.2950} = 15.5529$

$t_{cal}(\beta_1) = \frac{\hat{\beta}_1 - \beta_1}{se} = \frac{0.2572}{0.0320} = 8.0375$

$t_{cal}(\beta_2) = \frac{\hat{\beta}_2 - \beta_2}{se} = \frac{0.1115}{0.3343} = 0.3335$

$t_{cal}(\beta_3) = \frac{\hat{\beta}_3 - \beta_3}{se} = \frac{0.1580}{0.0890} = 1.7753$

$t_{cal}(\beta_4) = \frac{\hat{\beta}_4 - \beta_4}{se} = \frac{0.1809}{0.0848} = 2.1333$

$t_{cal}(\beta_5) = \frac{\hat{\beta}_5 - \beta_5}{se} = \frac{-0.2830}{0.0992} = -2.8528$

(c) Take a partial derivative of $\log(\text{salary})$ to utility

$\therefore \frac{\partial \log(\text{salary})}{\partial \text{utility}} = \beta_5 \text{utility}$

$= (-0.2830)\ \text{utility}$

→ Transportation is the base line (intercept) $= \ln(4.5881)$

→ $\beta_5 = -0.2830$ (intecept of utility) $= \ln(4.5881 - 0.2830) = \ln(4.3051)$

$e^{(4.5881)} = 98.3075$; $e^{(4.305)} = 74.0765$

$\therefore$ percentage difference $= \frac{98.3075 - 74.0765}{98.3075}$

$= 0.2465 = 24.65\%$

(d) We can't do because the transportation industry is already the base line of the regression function. It supposed to be omitted. If we put all dummies, including transport$_i$, it would cause an error by overestimation. Also, it is not often that there would be a case that all dummies are equal.

(e) The result would be "interaction dummies" and the function would cause an additional effect. The addition effect would be beneficial only if T& P-value of the additional estimator is statistically significance. If it's not, it would prevent the ability to predict of this regression model. In this case, t&P-value isn't given.

2. (a) Use t-test for model 2.1

Step 1: State Hypothesis

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Step 2: Statistic test

$$t_{cal}(\beta_1) = \frac{\hat{\beta}_1 - \beta_1}{se} = \frac{-0.5877}{0.1040} = -5.3977$$

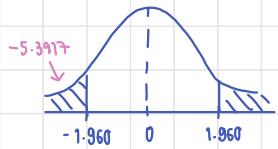
Step 3: Decision Rule : upper & lower

$\alpha = 0.05$

$t_{upper} = 0 + 1.960 = 1.960$

$t_{lower} = 0 - 1.960 = -1.960$

Step 4: Conclusion



t_{cal} is beyond the critical value, so we can reject the null hypothesis at the significance of 5%.

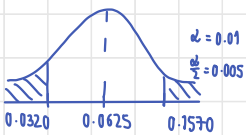
In other words, we can say the smoking has an effect on birth weight 95 out of 100.

(b) 99% confidence interval

2.2576

Lower bound : $\hat{\beta}_2 - t_{\frac{\alpha}{2}} \cdot \delta\hat{\beta}_2 = 0.0625 - t_{0.005}(0.0367) = 0.0320$

Upper bound : $\hat{\beta}_2 + t_{\frac{\alpha}{2}} \cdot \delta\hat{\beta}_2 = 0.0625 + t_{0.005}(0.0367) = 0.1570$



∴ 99% confidence interval of β_2 is 0.0320 until 0.1570.

(c) Use t-test for model 2.2

Step 1: State Hypothesis

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Step 2: Statistic test

$$t_{cal} = \frac{\hat{\beta}_1 - \beta}{se} = \frac{-0.5895 - 0}{0.1106} = -5.3300$$

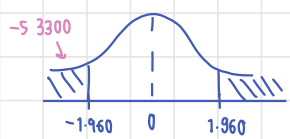
Step 3: Decision Rule: Upper & Lower

$\alpha = 0.05$

$t_{upper} = 0 + 1.960 = 1.960$

$t_{lower} = 0 - 1.960 = -1.960$

Step 4: Conclusion



Still, t_{cal} is beyond the critical value, so we can reject the null hypothesis at the significance level of 5%.

In other words, we can say the smoking has an effect on birth weight 95 out of 100 times. (conclusion is unchanged)

(d) Use t-test

Step 1: State Hypothesis

$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1 : Otherwise

Step 2: Statistic Test

$t_{cal}(\beta_0) = \frac{118.0741 - 0}{3.5003} = 33.7326$

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

$t_{cal}(\beta_2) = \frac{0.0538 - 0}{0.0367} = 1.4659$

$t_{cal}(\beta_3) = \frac{0.4937 - 0}{0.2833} = 1.7427$

$t_{cal}(\beta_4) = \frac{-0.4379 - 0}{0.3197} = -1.3697$

Step 3: Decision Rule: $\alpha = 0.05$

β_0 : $t_{upper} = 118.0741 + 1.960(3.5003) = 124.9347$; $t_{lower} = 111.2135$

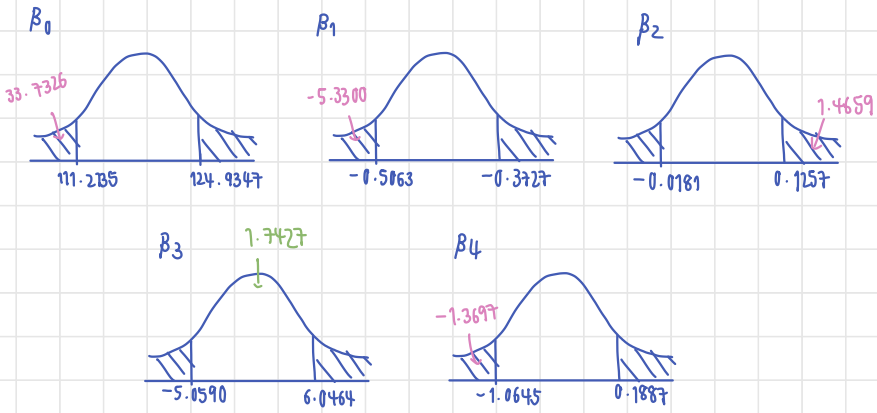
β_1 : $t_{upper} = -0.5894 + 1.960(0.1106) = -0.3727$; $t_{lower} = -0.5063$

β_2 : $t_{upper} = 0.0538 + 1.960(0.0367) = 0.1257$; $t_{lower} = -0.0181$

β_3 : $t_{upper} = 0.4937 + 1.960\,(2.833) = 6.0464$; $t_{lower} = -5.0590$

β_4 : $t_{upper} = -0.4379 + 1.960\,(0.3197) = 0.1887$; $t_{lower} = -1.0645$

Step 4 : Conclusion



For β_3, we cannot reject the null hypothesis at the significance level of 0.05, as t_{cal} is still within the critical value. So, we cannot say for sure that father's education has an impact on birth weight for 95 out of 100 times.

To test the overall significance of regression, use F-test

Step 1 : state Hypothesis

$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1 : Otherwise

Step 2 : $F_{cal} = F_{(2,1188)} = 18.44$

Step 3 : $\alpha = 0.05$, $df_{num.} = 2$, $df_{denum} = 1188$

$F_{upper, \alpha} = 3.00$

Step 4 : Conclusion

We can reject the null hypothesis at the significance level of 0.05 We can say for sure that the model is statistically significant 95 out of 100 times.

(e) Parents' education consists of father's and mother's one. Therefore, we have to test if two estimators work in the same or not, where, the two estimators have to be simultaneously significance to make sure that these are impactful (ANOVA)

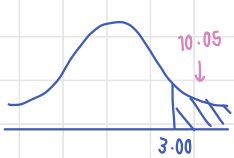
Step 1 : $H_0 : \beta_3 = \beta_4 = 0$

H_1 : Otherwise

Step 2 : $F_{cal} = F_{(4,1186)} = 10.05$

Step 3 : $\alpha = 0.05$, $df_{num} = 4$, $df_{denum} = 1186$

$F_{upper} = 2.37$



We can reject the null hypothesis at the signifance level of 0.05. So, we can say for sure that

3. (a) $df_{model} = k-1 = 7-1 = 6$

$df_{residual} = n-k = 428-7 = 421$

$df_{total} = n-1 = 428-1 = 427$

(b) $F_{cal} = \dfrac{ESS/TSS}{RSS/TSS}\Big/\dfrac{(k-1)}{(n-k)}$ $\quad\therefore$ $ESS = 35.3384$

$RSS = 223.3274 - 35.3384$

$= 187.9890$

$13.19 = \dfrac{\frac{ESS}{223.3274}/(6)}{1-\frac{ESS}{223.3274}/421}$

(c) $\bar{R}^2 = 1-(1-R^2) = \dfrac{n-1}{n-k} = 1-(1-0.1582)\dfrac{427}{421} = 0.1462$

(d) Marginal Contribution → there's one less estimator

Step 1: State Hypothesis

H_0: the estimator has no marginal contribution to the model

H_1: Otherwise

Step 2: $F_{cal} = \dfrac{R^2_{new} - R^2_{old}/(\#\text{ new regressors})}{1-R^2_{new}/(n-k_{new})}$

$= \dfrac{0.1600-0.1582/(1)}{1-0.1600/(422)}$

$= 0.9043$

$R^2_{new} = 1-(1-R^2_{old})\dfrac{n-1}{n-k}$

$= 1-(1-0.1582)\dfrac{427}{422}$

$= 0.1600$

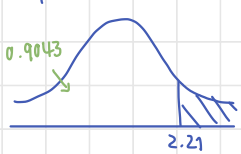
Step 3: $F_{(1,422)} = 3.84$

$F_{(6,422)} = 13.19$

$\therefore F_{(5,421)} = 2.21$

we can't reject the null hypothesis → at the significance level of 0.05.

step 4: Conclusion



Model 3.1, that has more estimators, is not better than model 3.2, that exclude the variables, due to the reason that the variable has no contribution to the model.

(e) Yes, it makes economic sense, because each person has different background and opportunities. With the higher wage, it means that in order to qualify the higher reward, there must be higher experience or higher educational level, as the reward is related to the responsibilities and difficulty of that particular job. Especially with woman, on these days, it can't be denied that the inequality between men and women still exist. Thus, the rise of women's wage might be less if compared to men, especially married woman