Phitak Phiencharoen

6204640962

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 $\text{ROE}_i * utility_i$?

a. Write out the estimated regression equation for $\log(salary_i)$. Interpret the estimated coefficient associated with $\log(sales_i)$.

$Y_i \log(salary_i) = 4.588101 + 0.2572 \log(sales_i) + 0.0112\,ROE_i + 0.1580\,finance_i + 0.1809\,consprod_i - 0.2830\,utility_i$

Intepret estimated coefficient of $\log(sales_i)$ ⟶ log - log form : 1 % increase in X_i, on average, increases Y_i by $\hat{B}_1$ %

∴ At 1 % increases 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.

Using F - test

Step 1 : State the hypothesis

$H_0 : B_1 = B_2 = B_3 = B_4 = B_5 = 0$

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

Step 2 : Calculate test statistics

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

Step 3 : State decision rule

$\alpha = 0.01$

upper bound $= F_\alpha(k-1, n-k)$

$= F_{0.01}(5, 209-6)$

$= F_{0.01}(5, 203) = 3.11$

Step 4 : Conclude the test

F_{cal} > upper bound. The overall significance of the regression is at least 99 % since, $\alpha = 0.01$.

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

$= \pm t_{0.025, 203}$

$= \pm 1.96$

If $t > 1.96$ or $t < -1.96$, it will be a statistically significant.

Therefore, $\hat{B}_0, \hat{B}_1, \hat{B}_2, \hat{B}_4, \hat{B}_5$ 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.

$\log(salary_i) = 4.588101 + 0.2572 \log(sales_i) + 0.0112 ROE_i + 0.1580 finance_i + 0.1809 consprod_i - 0.2830 utility_i$

Plug in dummies;

utility : $E(\log(salary_i) \mid finance_i = 0, consprod_i = 0, utility_i = 1)$

$= 4.5881 + 0.2572 \log(sales_i) + 0.0112 ROE_i + 0.1580(0) + 0.1809(0) - 0.2830(1)$

$= 4.3051 + 0.2572 \log(sales_i) + 0.0112 ROE_i$

transport : $E(\log(salary_i) \mid finance_i = 0, consprod_i = 0, utility_i = 0)$

$= 4.5881 + 0.2572 \log(sales_i) + 0.0112 ROE_i + 0.1580(0) + 0.1809(0) - 0.2830(0)$

$= 4.5881 + 0.2572 \log(sales_i) + 0.0112 ROE_i$

Difference in $\log(salary_i) = [4.3051 + 0.2572 \log(sales_i) + 0.0112 ROE_i]$

$- [4.5881 + 0.2572 \log(sales_i) + 0.0112 ROE_i] = -0.2830$

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?

Because when we put all the sector dummies as well as the intercept will lead to a perfect multicollinearity problem.

If we use STATA to run all the sector dummies as well as the intercept, the STATA program would automatically drop out one of the variables.

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 $\text{ROE}_i * utility_i$?

No, it is not going to have any benefit in adding interaction terms because there is no obvious reason of why the return on equity would affect the CEO annual salary in each sector differently.

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 $\alpha = 0.05$)

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 $\alpha = 0.05$)

Step 1 : State the 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

$\alpha = 0.05$ $df = n-k$

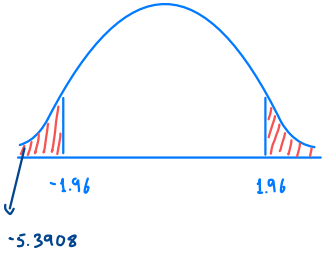
$df = 1191 - 3 = 1189$

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

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

Step 4 : Conclude the test

$t_{cal} <$ lower bound, we reject H_0 at 0.05 significant level, which 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.

99% CI : $\hat{\beta}_2 \pm (t_{\frac{0.01}{2}, 1188} \cdot se(\hat{\beta}_2)) = 0.0624684 \pm (2.576 \cdot 0.0324438)$

$= 0.0624684 \pm 0.08357523$

$= (0.0211, 0.1460)$

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

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

Step 1 : State the hypothesis

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

Step 2 : Calculate test statistics

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

Step 3 : State decision rule

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

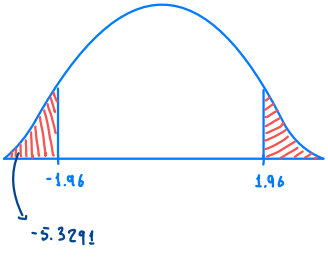
upper bound $= +t_{\frac{\alpha}{2}, n-k} = +t_{0.025, \infty} = 1.96$

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

Step 4 : Conclude the test

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

$\beta_1 \neq 0$ hence, somking has a significant impact on infant birth weight. [The conclusion is the same as (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 the hypothesis

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

H_a : otherwise

Step 2 : Calculate test statistics

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

Step 3 : State decision rule

$\alpha = 0.01$

upper bound $= F_\alpha(k-1, n-k)$

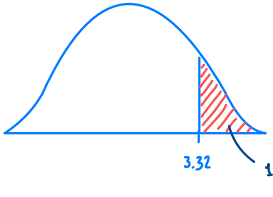
$= F_{0.01}(4, 1186)$

$= 3.32$

Step 4: Conclude the test

F_{cal} > upper bound, we reject H_0 at 0.01 significance level.

Meaning that the overall significance is at least 99 %.



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

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

$t_{cal}(\hat{\beta}_1) = \frac{-0.6894954 - 0}{0.1106172} = -5.3291 < -1.96$

$t_{cal}(\hat{\beta}_2) = \frac{0.0538254 - 0}{0.0366502} = 1.4686$

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

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

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

At 5 % significance level, $\hat{\beta}_1$ 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 the hypothesis

H_0: Parent's education doesn't have impact on birth weight

H_a: otherwise

Step 2: Calculate test statistics

$$F_{cal} = \frac{(ESS_{incl} - ESS_{excl}) / \#\text{ of additional } X}{RSS_{incl} / (n - K_{incl})} = \frac{(15827.6593 - 14536.9538)/2}{466919.033/(1191-5)} = 1.639$$

Step 3: State decision rule

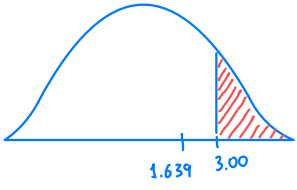
$\alpha = 0.05$

upper bound $= F_{\alpha}(\#\text{ of additional } X, n-k_{incl}) = F_{0.05}(2, 1186) = 3.00$

Step 4: Conclude the test

$F_{cal} <$ upper bound, we fail to reject H_0 at 0.05 level of significance.

Parent's education doesn't have impact on birth weight, but it might be reflected on family income.



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(___,______) =   13.19
    Model |  __________    ___  __________           Prob > F      =  0.0000
 Residual |  __________    ___   .446526442         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
     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?

a) Figure out all the degrees of freedom in this model.

df of ESS = $K-1 = 7-1 = 6$

df of RSS = $n-K = 428-7 = 421$

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.

Mean square = $\frac{SS}{df}$, Residual mean square = $\frac{RSS}{df \text{ of } RSS}$

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

$$RSS = 0.446526442(428-7) = 187.9876$$

ESS = TSS − RSS = $223.327441 - 187.9876 = 35.3398$

Model mean square = ESS / $k-1$ = $\frac{35.3398}{7-1} = 5.8900$

Total mean square = TSS / $n-1$ = $\frac{223.327441}{428-1} = 0.5230$

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

The adjusted 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 $F_{1,421} = 3.84$)

From the given information # of variation in model 3.1 is more than 3.2

Meaning that $R^2_{3.1} \geq R^2_{3.2}$

H_0 : variable excluded in model 3.2 is not significant

H_a : otherwise

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

$$\frac{(R^2_{incl} - R^2_{excl}) / \#\text{ of additional } X}{(1 - R^2_{incl})/(n - K_{incl})} > 3.84$$

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

$$\frac{0.1582 - R^2_{3.2}}{0.0019995} > 3.84$$

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

∴ 0.1505 is the maximum value of R^2 in model 3.2 that would lead to the conclusion that the excluded variable is significant at 0.05 significance level.

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?

I think age can make an impact on hourly wage. As an older age, worker could have more specialized in the job. However, an insignificance could be occured by the correlation with other variables like, education or experience. The effect that age has on wage could be a non-linear and not apprehended by the model.