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:

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 \text{ROE}_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility_i + u_i$$

where $\log(salary_i)$ = logarithm of CEO annual salary (in 1,000 USD)

X_1 $\log(sales_i)$ = logarithm of firms' sale (in 1 million USD)

X_2 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)

x_3 $finance_i$ = 1 if in financial industry, = 0 otherwise

x_4 $consprod_i$ = 1 if in consumer product industry, = 0 otherwise

x_5 $utility_i$ = 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
Model	23.8109943	5	4.76219887
Residual	42.9111689	203	.211385068
Total	66.7221632	208	.320779631

Number of obs	= 209
F(5, 203)	= 22.53
Prob > F	= 0.0000
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

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.

↳ P value

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     Omitted for the purpose of this exam.
       _cons |   118.5568   1.234278
------------------------------------------------------------------------------
```

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     Omitted for the purpose of this exam.
    fatheduc |   .4936695   .2832896
    motheduc |  -.4379234   .3197377
       _cons |   118.0741   3.500291
------------------------------------------------------------------------------
```

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 $\alpha = 0.05$)

k = 7   k-1 = 6

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
Model	35.339809	k-1 6	5.889968
Residual	187.987632	421 n-k	.446526442
Total	223.327441	427 n-1	6.336494

Number of obs	=	428
F(6 , 421)	=	13.19
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

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?

$k = 6$ $\qquad\qquad$ $\sum u_i^2 = RSS$

```
      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
------------------------------------------------------------------------------
```

X_1 X_2 X_3 X_4 X_5

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 \text{ROE}_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility_i + u_i$$

where $\log(salary_i)$ = logarithm of CEO annual salary (in 1,000 USD)

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 $\text{ROE}_i * utility_i$?

statiscally sig $\neq 0$

a) $\log(\text{sarlary}) = 4.5881 + 0.2572 \log X_1 + 0.0112 X_2 + 0.158 X_3 + 0.1809 X_4 - 0.2830 X_5$

b) Use t-test

β_0

① significant $\neq 0$

Hypothesis $H_0: \beta_0 = 0$

$H_A: \beta_0 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_0 - \beta_0}{S.E} = \dfrac{4.5881 - 0}{0.2950} = 15.5529$ #

③ $t_{209-6} = t_{203}$ $\alpha = 0.05$ $t_L = -1.960$ $t_U = 1.96$

④ 15.5529

$\therefore$ we can reject the null hypothesis at the significance level of 0.05

$\therefore$ we can say that the β_0 satatistically significant 95% of a time.

β_1

① significant $\neq 0$

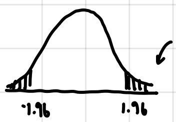
Hypothesis $H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_1 - \beta_1}{S.E} = \dfrac{0.2572}{0.0320} = 8.0375$

③ $t_{209-6} = t_{203}$ $\alpha = 0.05$ $t_L = -19.60$ $t_U = 1960$

④ 8.0375



$\therefore$ we can reject the null hypothesis at the significance level of 0.05

$\therefore$ we can say that the β_1 satatistically significant 95% of a time.

β_2

① significant $\neq 0$

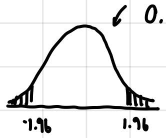
Hypothesis $H_0: \beta_2 = 0$

$H_a: \beta_2 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_2 - \beta_2}{S.E} = \dfrac{0.0112}{0.3343} = 0.0335$

③ $t_{209-6} = t_{203}$ $\alpha = 0.05$ $t_L = -19.60$ $t_U = 1960$

④ 0.0335



$\therefore$ we can not reject the null hypothesis at the significance level of 0.05

$\therefore$ we can say that β_2 may not be statiscally significant for 95% of the time

β_3

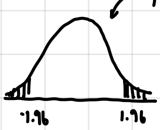
① Hypothesis $H_0: \beta_3 = 0$

$H_a: \beta_3 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_3 - \beta_3}{S.E} = \dfrac{0.158}{0.0890} = 1.7753$ #

③ $t_{209-6} = t_{203}$ $\alpha = 0.05$ $t_L = -19.60$ $t_U = 1960$

④ 1.7753



$\therefore$ we can not reject the null hypothesis at the significance level of 0.05

$\therefore$ we can say that β_3 may not be statiscally significant for 95% of the time

β_4

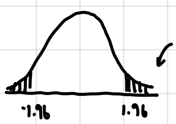
① Hypothesis $H_0: \beta_4 = 0$

$H_a: \beta_4 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_4 - \beta_4}{S.E} = \dfrac{0.1809}{0.0848} = 2.1333$

③ $t_{209-6} = t_{203}$ $\alpha = 0.05$ $t_L = -19.60$ $t_U = 1960$

④ 2.1333



$\therefore$ we can reject the null hypothesis at the significance level of 0.05

$\therefore$ we can say that the β_4 satatistically significant 95% of a time.

β_5

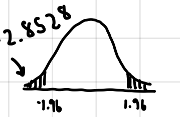
① Hypothesis $H_0: \beta_5 = 0$

$H_a: \beta_5 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_5 - \beta_5}{S.E} = \dfrac{-0.283}{0.0992} = -2.8528$

③ $t_{209-6} = t_{203}$ $\alpha = 0.05$ $t_L = -19.60$ $t_U = 1960$

④ -2.8528



$\therefore$ we can reject the null hypothesis at the significance level of 0.05

$\therefore$ we can say that the β_1 satatistically significant 95% of a time.

Use F-test

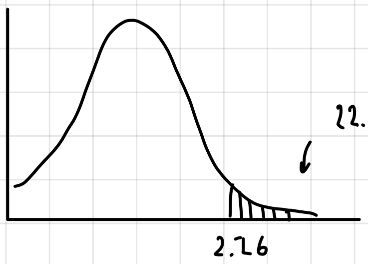
① $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_a : other wise

② $F_{cal} = \dfrac{ESS/df}{RSS/df} = \dfrac{4.7622}{0.2114} = 22.527$ #

③ $\alpha = 0.05 \quad F_{(5,203)} = 2.26$

④ Conclude the test result



∴ We can reject H_0 at the significance level of 0.05

∴ we can make sure that $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are not simultaneously zero.

∴ we can say for sure that the model is statistically significant for 95%.

c) $\hat{\beta}_5 = -0.283$; $100(e^{(-0.2833)} - 1) = -24.648\%$

on average, the ceo annual salary are lower than the transportation sector by about 24.648%. holding sales and ROE_i fixed

d) we cannot put all the dummies variable because transportation is a based group and exclude the error. If we still put all the sector dummies, it will cause the perfect multicollinearity

e) If we add another dummy then making the adjusted R squared, higher F or absolute value of t-value is more than one, then we should add the interaction terms.

n-k = 1188

Model 2.1: $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + u_i$

k=3

```
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     Omitted for the purpose of this exam.
     _cons |   118.5568   1.234278
------------------------------------------------------------------------------
```

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     Omitted for the purpose of this exam.
  fatheduc |   .4936695   .2832896
  motheduc |  -.4379234   .3197377
     _cons |   118.0741   3.500291
------------------------------------------------------------------------------
```

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 $\alpha = 0.05$)

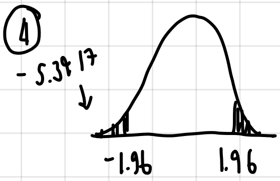
a) Test β_1 t-test

① $H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

② $t_{cal} \quad \frac{\hat{\beta}_1 - \beta_1}{S.E} = \frac{-0.5877}{0.1090} = -5.3917$

③ $t_{cal} = t_{1188} \quad \alpha = 0.05 \quad = 1.960$

④



∴ we can reject the hypothesis at the significance level of 0.05

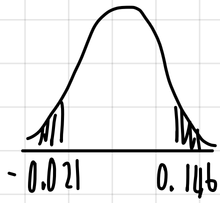
∴ we can say for sure that cigarette affects birth weight for 95% of the times.

b) Confidence interval

$\beta_2 \pm t_{\frac{\alpha}{2}} \cdot SE$

$0.0625 \pm \underset{2.576}{t_{\frac{0.01}{2}}} \cdot 0.0324$

Upper 0.146
lower -0.021



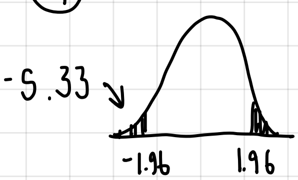
c) Test β_1 t-test

① $H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

② $t_{cal} \quad \frac{\hat{\beta}_1 - \beta_1}{S.E} = \frac{-0.5895}{0.1106} = -5.3300$ #

③ $t_{cal} = t_{1186} \quad \alpha = 0.05 \quad = 1.960$

④



∴ we can reject the hypothesis at the significance level of 0.05

∴ we can still say for sure that cigarette affects birth weight for 95% of the times.

d) use t-test

β_0

b) ① significant to

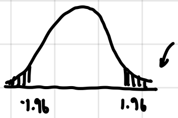
Hypothesis $H_0 : \beta_0 = 0$

$H_A : \beta_0 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_0 - \beta_0}{S.E} = \dfrac{118.0741}{3.5003} = 33.7326$ #

③ $t_{cal} : t_{1186}$, $\alpha = 0.05$ $t_L = -1.960$ $t_U = 1.96$

④ 33.7326



∴ we can reject the null hypothesis at the significance level of 0.05

∴ we can say that the β_0 satatistically significant 95 % of a time.

β_1

① significant to

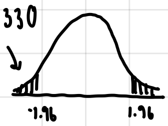
Hypothesis $H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_1 - \beta_1}{S.E} = \dfrac{-0.5895}{0.1106} = -5.330$ #

③ $t_{cal} = t_{1186}$ $\alpha = 0.05$ $t_L = -19.60$ $t_U = 1960$

④ -5.330



∴ we can reject the null hypothesis at the significance level of 0.05

∴ we can say that the β_1 satatistically significant 95 % of a time.

β_2

① significant to

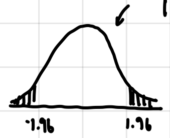
Hypothesis $H_0 : \beta_2 = 0$

$H_a : \beta_2 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_2 - \beta_2}{S.E} = \dfrac{0.0538}{0.0367} = 1.4659$ #

③ $t_{cal} = t_{1186}$ $\alpha = 0.05$ $t_L = -19.60$ $t_U = 196$

④ 1.4659



∴ we can not reject the null hypothesis at the significance level of 0.05

∴ we can say that β_2 may not be statiscally significant for 95% of the time

β_3

① Hypothesis $H_0 : \beta_3 = 0$

$H_a : \beta_3 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_3 - \beta_3}{S.E} = \dfrac{0.4937}{0.2833} = 1.7427$ #

③ $t_{cal} = t_{1186}$ $\alpha = 0.05$ $t_L = -19.60$ $t_U = 1960$

④ 1.7427

∴ we can not reject the null hypothesis at the significance level of 0.05

∴ we can say that β_3 may not be statiscally significant for 95% of the time

β_4

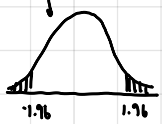
① Hypothesis $H_0 : \beta_4 = 0$

$H_a : \beta_4 \neq 0$

② $t_{cal} = \dfrac{\hat{\beta}_4 - \beta_4}{S.E} = \dfrac{-0.4379}{0.3197} = -1.3697$ #

③ $t_{cal} = t_{1186}$ $\alpha = 0.05$ $t_L = -19.60$ $t_U = 1960$

④ -1.3697



∴ we can not reject the null hypothesis at the significance level of 0.05

∴ we can say that β_4 may not be statiscally significant for 95% of the time

use F-test

① $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

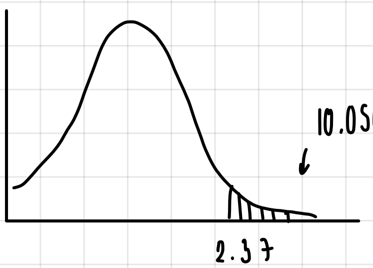
H_a : other wise

② $F_{cal} = \dfrac{ESS / df}{RSS / df} = \dfrac{3956.9148}{393.6923} = 10.0508$

③ $\alpha = 0.05$ $F_{(4, 1186)} = 2.37$

④ Conclude the test result

10.0508



∴ We can reject H_0 at the significance level of 0.05

∴ we can make sure that $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are not simultaneously zero.

∴ we can say for sure that the model is satistically significant for 95%.

e) The marginal Distribution

worth to add to the model or not

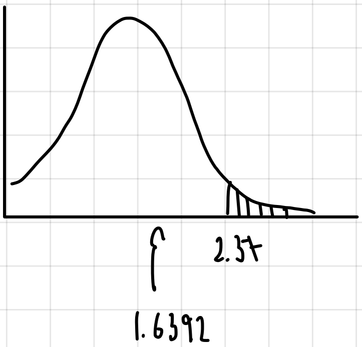
① H_0 : parents education has no marginal contribution to the model

H_1 : otherwise

② $F_{cal} = \dfrac{ESS_{new} - ESS_{old} / (\text{number of new regression})}{RSS_{new}\ (n-k_{new})} = \dfrac{15827.6593 - 14536.9538/(2)}{393.6923}$

$= 1.6392$ #

③ $\alpha = 0.05$ $F_{(4,\ 1186)} = 2.37$



∴ we cannot reject the hypothesis at the significance level of 0.05

∴ We can not say for sure that parents education has no marginal contribution to the model

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
Model	35.339809	6	5.889968
Residual	187.987632	421	.446526442
Total	223.327441	427	6.336494

df: Model $k-1$, Residual $n-k$, Total $n-1$

Number of obs	=	428
F(6 , 421)	=	13.19
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

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) df for model $k-1$: ; $k = 7$ so df for model 6
df for residual $n-k$: ; $n = 428$, $k = 7$ so $n-k = 421$
dk for Total $n-1$: $n = 428$ so $n-1 = 427$

b)

M_s for total $= \dfrac{223.327441}{427} = 6.336494$

M_s for model $\dfrac{35.339809}{6} = 5.889968$

ESS : 35.339809

M_s for residual $\dfrac{187.987632}{421} : 0.446526442$

RSS = 187.987632

c) Adjust R square

$$\bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-k}$$

$$= 1-(1-0.1582)\left(\frac{427}{421}\right) = 0.1462$$

d) Model 3.1 new $R^2 = ?$ $k_{new} = 7$ $F(1,421) = 3.84$
Model 3.2 old $R^2 = 0.1582$ $k_{old} = 6$

1. H_0 : exclude variable has no marginal contribution to the model

H_a . otherwise

2. $F_{cal} = \dfrac{R^2_{new} - R^2_{old} \;/ (\text{number of new regression})}{1 - R^2_{old}\,(n-k_{new})} = 3.84 = \dfrac{R^2_{new} - 0.1582 \;/\; (1)}{(1-0.15482)(421)}$

$R^2 = 0.1505$ # max R^2 of model 3.2

e) Yes, in a sense of economic, Work didn't required the age but prefer practice, educational level instead.