Tumnurat Pucksanonthachai 6204641077

Assignment 2

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 $\log(sales_i)$.

Log(salary) = 4.588101 + 0.2571917log($sales_i$)

0.2571917 is β_1 or coefficient meaning when log($sales_i$) increase by 1 unit, log($sales_i$) also increase by 0.2571917

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.

F-test

$H_0 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1 = Not all slope coefficients are 0

$F_{cal} = \frac{\frac{ESS}{K-1}}{\frac{RSS}{n-K}} = \frac{MS(E)}{MS(R} = \frac{4.76219887}{0.211385068} = 22.529$

At $\alpha = 0.05$, $n - k = 209 - 6 = 203$

$F_{cal\ upper,\beta(5,\ 203)} = 2.21$

$F_{cal} > F_{cri(0.05)}$, $22.529 > 2.21$ We can reject H_0 and we can make sure that β_2, β_3, β_4, β_5 are not 0

All independent variables can significantly explain the dependent variable

$\alpha = 0.05$, df = 203,

$t_{cri\ lower\ (203,0.05)\ two\ tail} = -1.9717$, $t_{cri\ upper} = 1.9717$

t_{cal}: when outside reject

$\log(sales_i) = 8.03$, reject H_0, significant

roe = 2.59, reject H_0, significant

finance = 1.77, can't reject H_0, not significant

consprod = 2.13, reject H_0, significant

utility = -2.85, reject H_0, significant

_cons = 15.55, reject H_0, significant

c. Compute the approximate percentage difference in estimated salary between the utility and transportation sector, holding $sales_i$ and ROE_i fixed.

Utility

$\log(salary_i) = \beta_0 + \beta_3 fincance_i + \beta_4 conspod_i + \beta_5 utility_i$

if 1 $= 4.588101 + 0.1579564(1) + 0.1808917(1) - 0.2830015 = 4.644$

if 0 $= 4.588101 + 0.1579564(0) + 0.1808917(0) - 0.2830015 = 4.3050995$

Transport sector, no industry

$\log(salaryi) = \beta_0 + \beta_3 fincancei + \beta_4 conspodi + \beta_5 utilityi$

$= 4.588101 + 0.1579564(0) + 0.1808917(0) - 0.2830015(0) = 4.5881$

Percentage change = $\frac{4.5881-4.644}{4.644}\ x\ 100$ = - 0.012, -1.2%

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?

Perfect collinerity will happens, all variables have same linear relation to each other. In STATA, it will ignore 1 dummy variable since it can't estimate.

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 $ROE_i * utility_i$?

$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ROE_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility + \beta_6(ROE_i * finance) + \beta_7(ROE*conspod) + \beta_8(ROE*utility) + u_i$

$E(\log(salary) \mid finance = 1) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ROE_i + \beta_3(1) + \beta_4 consprod_i + \beta_5 utility + \beta_6(ROE_i * 1) + \beta_7(ROE*conspod) + \beta_8(ROE*utility) + u_i$

$= (\beta_0 + \beta_3) + \beta_1 \log(sales) + (\beta_2 + \beta_6)(ROE_i) + \beta_4 conspond + \beta_5 utility + \beta_7(ROE*consprod) + \beta_8(ROE*utility) + u_i$

$E(\log(salary) \mid consprod = 1) = \beta_0 + \beta_1 \log(sales) + \beta_2 ROE + \beta_3 finance + \beta_4(1) + \beta_5 utility + \beta_6(ROE*finance) + \beta_7(ROE*(1)) + \beta_8 (ROE*utility) + u_i$

$= \beta_0 + \beta_4 + \beta_1 \log(sales) + (\beta_2 + \beta_7)(ROE) + \beta_3 finance + \beta_5 utility + \beta_6(ROE*finance) + \beta_8(ROE*utility) + u_i$

$E(\log(salary) \mid utility = 1) = \beta_0 + \beta_1 \log(sales) + \beta_2 ROE + \beta_3 finance + \beta_4 consprod + \beta_5(1) + \beta_6(ROE*finance) + \beta_7(ROE*consprod) + \beta_8 (ROE*1) + u_i$

$= \beta_0 + \beta_5 + \beta_1 \log(sales) + (\beta_2 + \beta_8)(ROE) + \beta_3 finance + \beta_4 consprod + \beta_6(ROE*finance) + \beta_7(ROE*consprod) + u_i$

Intercept and slope of regression change. Y-intercept increase and stronger slope (steep/flat) depends on roe and section dummy if it's significantly different from zero 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 ($fathereduc_i$), 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 $\alpha = 0.05$)

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

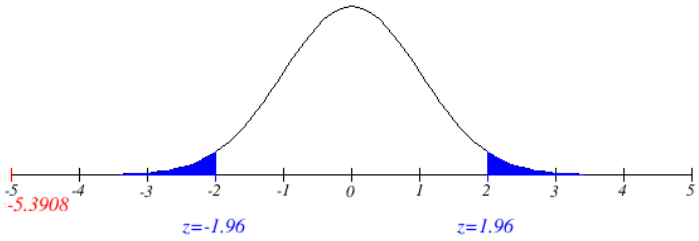
T-test

$t_{cal} = \frac{\widehat{\beta_1} - \beta_1}{se\widehat{\beta_1}} = \frac{cigs\ Coef.}{cigs\ Std.\ Err.} = \frac{-0.5876985}{0.1090181} = -5.3908$

α: f = n-k = 1191 – 3 = 1188

$t_{lower\ (3,\ 1188)} = -1.96$, $t_{upper} = 1.96$

Reject H_0, t_{cal} lower than t_{lower} we can make sure that β_1 is not 0, smoking has impact



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

$\widehat{\beta_2} \pm t_{\frac{\alpha}{2}}$, n – 3 x ó$\widehat{\beta}_2$; $t\frac{0.01}{2}$, 1188 = 2.576

= faminc coef ± 2.576 x faminc std. err.

= 0.62468a ± 2.576 x 0.324438

Lower = -0.2111

Upper = 1.4604

99% confidence interval for $\widehat{\beta_2}$ is (-0.2111 < $\widehat{\beta_2}$ < 1.4604) = 0.99, 99 out of 11 intervals will contain true $\widehat{\beta_2}$.

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

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

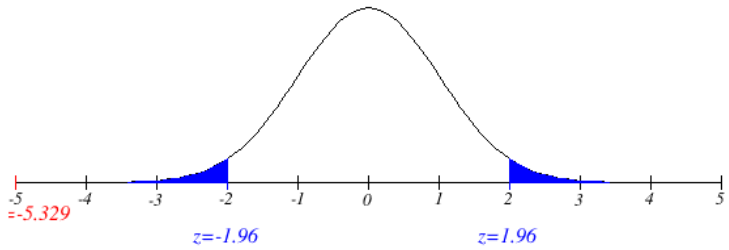
t-test

tcal = $\frac{\widehat{\beta_1} - \beta_1}{se\widehat{\beta_1}} = \frac{cigs\ coef.}{cigs\ std.\ err.} = \frac{-0.5894954}{0.1106172}$ = -5.329

df = n – k = 1191 -5 = 1186

$t_{lower\ (1186,\ 5)}$ = -1.96, t_{upper} = 1.96

Reject H_0, t_{cal} = -5.329 lower than t_{lower} = -1.96. We can make sure that β_1 is not 0, smoking has impact, conclusion don't change.



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.

F-test

H_0: $\beta_2 = \beta_3 = \beta_4 = 0$; $\alpha = 0.05$

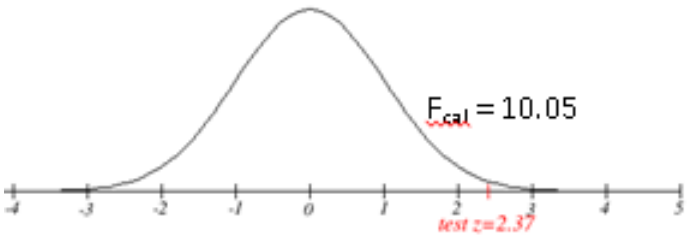
H_1: Not all slope are coefficient or simultaneously 0

$F_{cal\ (4,1186)} = \frac{Model\ Ms}{Residual\ MS} = \frac{3956.91482}{393.69227}$ = 10.051

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

$F_{cal} > F_{cri}$ = 10.05 > 2.37

We can reject H_0 and make sure that β_2, β_3, and β_4 are not 0.



T-test: $t_{lower} = -1.96$, $t_{upper} = 1.96$

H_0: $\beta_0 = 0$

H_1: $\beta_0 \neq 0$

$t_{cal} = \frac{\widehat{\beta}_0 - \beta_0}{se\widehat{\beta}_0} = \frac{cons\ coef}{cons\ std.\ err.} = \frac{118.0741}{3.500291} = 33.7326$ Reject H_0, significance

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

$t_{cal} = \frac{\widehat{\beta}_1 - \beta_1}{se\widehat{\beta}_1} = \frac{cigs\ coef}{cigs\ std.\ err.} = \frac{-0.5894954}{0.1106172} = -5.329$ Reject H_0, significance

H_0: $\beta_2 = 0$

H_1: $\beta_2 \neq 0$

$t_{cal} = \frac{\widehat{\beta}_2 - \beta_2}{se\widehat{\beta}_2} = \frac{faminc\ coef}{faminc\ std.\ err.} = \frac{0.538254}{0.366502} = 1.4686$ Cannot reject H_0, not significance

H_0: $\beta_3 = 0$

H_1: $\beta_3 \neq 0$

$t_{cal} = \frac{\widehat{\beta}_3 - \beta_3}{se\widehat{\beta}_3} = \frac{fatheduc\ coef}{fatheduc\ std.\ err.} = \frac{0.4936695}{0.2832896} = 1.7726$ Cannot reject H_0, not significance

H_0: $\beta_4 = 0$

H_1: $\beta_4 \neq 0$

$t_{cal} = \frac{\widehat{\beta}_4 - \beta_4}{se\widehat{\beta}_4} = \frac{motheduc\ coef}{motheduc\ std.\ err.} = \frac{-0.4379234}{0.3197377} = -1.3696$ Cannot reject H_0, not significance

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 $\alpha = 0.05$)

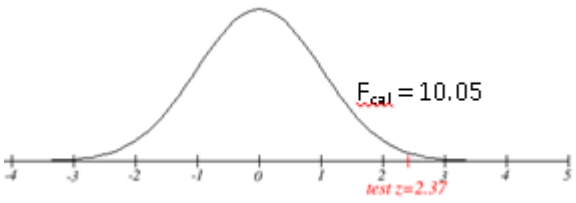
H_0: $\beta_3 = \beta_4 = 0$

H_1: Not all slope coefficients are simultaneously 0

$F_{cal} = \frac{Model\ MS}{Residual\ MS} = \frac{3,956.91482}{393.69227} = 10.05$

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

$F_{cal} > F_{cri(4,\ 1186)}$, we can reject H_0 both β_3 and β_4 have impact on birth wage



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
Model	_____	___	_____	F(___,_____) =	13.19
Residual	_____	___	.446526442	Prob > F =	0.0000
Total	223.327441	___	_____	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.

df: k - 1

ESS = no.of β – 1 = 7 – 1 = 6

RSS = n – k = 428 – 7 = 421

TSS = n – 1 = 428 – 1 = 427

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

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

0.1582 = ESS/223.327441

ESS = 35.3304

RSS = TSS – ESS

RSS = 223.327441 – 35.3304012 = 187.9970

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

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

$= 1 - (1 - 0.1582)\frac{428-1}{428-7} = 0.1462$

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 $F_{1,421} = 3.84$)

$F_{(1,\,421)} = \frac{\frac{R^2_{3.1} - R^2_{3.2}}{no.of\ regressor}}{\frac{1-R^2_{3.1}}{n-k_{3.2}}} = \frac{\frac{0.1582-R^2_{3.2}}{1}}{\frac{1-0.1582}{428-7}}$

$3.84 = \frac{(0.1582-R^2_{3.2})(421)}{0.8418}$

$R^2_{3.2} = 0.1505$

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?

Wage has no economic sense as when higher age means more experience. In term of salary difference between age 20 and 21 are not much, but salary change more between age 25 and 30. Age variable is significance because of multicollinearity with other terms, age variable will be ignore due to its unnecessarily in the model