Assignment #2

1. $\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ROE_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility_i + u_i$

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

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

ROE_i = average return on equity for the CEO's firm for the previous 3 years

$finance_i$ = 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

} dummies

($finance_i$, $consprod_i$, and $utility_i$ are binary variables indicating the financial, consumer products, and utilities industries.

The omitted industry is transportation

└ case that all dummies = 0

	$finance_i$	$consprod_i$	$utility_i$
financial	1	0	0
consumer products	0	1	0
utility	0	0	1
transportation	0	0	0

Source	SS	df	MS	
				Number of obs = 209
				F(5, 203) = 22.53 — แทนค่า F_{cal} ใน (b)
Model	ESS 23.8109943	5	4.76219887	Prob > F = 0.0000
Residual	RSS 42.9111689	203	0.211385068	R-squared = 0.3569
				Adj. R-squared = 0.3410
Total	TSS 66.7221632	208	0.320779631	Root MSE = 0.45977

lsalary	Coef.	Std. Err.	t (t_{cal})	P>|t|	[95% Conf. Interval]	
lsales	$\hat\beta_1$ 0.2571917	0.0320348	8.03	0.000	0.0194262	0.3203553
roe	$\hat\beta_2$ 0.0111517	0.3342996	2.59	0.010	0.0026742	0.0196293
finance	$\hat\beta_3$ 0.1579564	0.0890017	1.77	0.077	-0.0175299	0.3334426
consprod	$\hat\beta_4$ 0.1808917	0.0847683	2.13	0.034	0.0137524	0.3480311
utility	$\hat\beta_5$ -0.2830015	0.0992337	-2.85	0.005	-0.4786624	-0.0873405
_cons	4.588101 ($\hat\beta_0$)	0.2950221	15.55	0.000	4.0064	5.169801

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

$$\underset{Y_i}{\log(salary_i)} = \underset{\beta_0}{4.5881} + \underset{\beta_1}{0.2572}\, \underset{X_i}{\log(sales_i)} + \underset{\beta_2}{0.0112}\, ROE_i + \underset{\beta_3}{0.1580}\, finance_i + \underset{\beta_4}{0.1809}\, consprod_i + \underset{\beta_5}{0.2830}\, utility_i$$

$\boxed{\text{Intepret}}$ estimated coefficient of $\log(sales_i)$ → log-log form : 1% increase in X_i, on average, increases Y_i by $\hat\beta_1$ %.

∴ At 1% increase 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 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.

• use $\boxed{\text{F-test}}$ เพราะ test พร้อมกันหมด (ถ้า t-test จะ test ทีละตัว)

step1 state hypothesis

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

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

step2 calculate test statistics

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

step3 state decision rule

$\alpha = 0.01$ (เพราะ Prob > F = 0.000 ซึ่ง < 0.01)

$$\text{upper bound} = F_\alpha(k-1, n-k) = F_{0.01}(5, \underset{n}{209} - \underset{\#\text{ of }\beta}{6}) = F_{0.01}(5, 203) = 3.11$$

step4 conclude the test

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

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 (that $\hat\beta_i$)

So, $\hat\beta_0, \hat\beta_1, \hat\beta_2, \hat\beta_4, \hat\beta_5$ are statistically significant at 5% level.

$\hat\beta_0$ มี $t = 15.55 > 1.96$ ✓

$\hat\beta_1$ มี $t = 8.03 > 1.96$ ✓

$\hat\beta_2$ มี $t = 2.59 > 1.96$ ✓

$\hat\beta_3$ มี $t = 1.77$ ✗

$\hat\beta_4$ มี $t = 2.13 > 1.96$ ✓

$\hat\beta_5$ มี $t = -2.85 < -1.96$ ✓

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

$\log(salary_i) = 4.5881 + 0.2572\log(sales_i) + 0.0112\,ROE_i + 0.1580\,finance_i + 0.1809\,consprod_i - 0.2830\,utility_i$

แทนค่า dummies ของ utility กับ transport เข้าไป (0,0,1) กับ (0,0,0)

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 $\log(salary_i) = \underbrace{[4.3051 + 0.2572\log(sales_i) + 0.0112\,ROE_i]}_{utility} - \underbrace{[4.5881 + 0.2572\log(sales_i) + 0.0112\,ROE_i]}_{transport} = -0.2830$

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?

Why can't we put all dummies?

↳ Including all sector dummies as well as the intercept term would lead to a perfect multicollinearity problem.

What if we use STATA to run it anyway?

If we use STATA to run a regression with all sector dummies as well as the intercept term, the program would automatically drop one of the dummy variables

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 $ROE_i * utility_i$?

It is not likely that there would be benefit in adding interaction terms, because there is no apparent reason why the return on equity would affect the CEO annual salary in each sector differently.

ROE = ค.สามารถในการทำกำไร

ถ้า บ.มี ค.สามารถในก.ทำกำไรดี CEO ก็ควรจะได้เงินเดือนเยอะ

[ถ้าเพิ่มในอัตราส่วนที่เท่าๆกัน (slope เท่ากัน) → จะไม่ benefit

แต่ถ้าเพิ่มในอัตราส่วนที่ไม่เท่ากัน (slope ไม่เท่ากัน) → จะ benefit]

สำหรับ interaction term

2. Birth weight has been used by officials as one of the main determinants of health. Data set B WGHT.DTA contains data on infant birth weight 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:

where $bwght_i$ = birth weight, ounces

$cigs_i$ = average number of cigarettes the mother smoked per day while pregnant

faminc = 1988 family income, $1000s

$fathereduc_i$ = father's years of education

$mothereduc_i$ = mother's years of education

Model 2.1 $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + u_i$ exclude (ตัวที่สนใจในข้อ(e) มีไม่ครบ)

regress bwght cigs faminc

Source	SS	df	MS
Model	14536.9538	2	7268.47691
Residual	468209.738	1188	394.115941
Total	482746.692	1190	405.669489

Number of obs = 1191
F(2,1188) = 18.44
Prob > F = 0.0000
R-squared = 0.301
Adj. R-squared = 0.0285
Root MSE = 19.852

bwght	Coef.	Std. Err. $se(\hat\beta_1)$	t	P>|t|	[95% conf. Interval]
cigs	$\hat\beta_1$ -0.5876985	0.1090181	Omitted for the purpose of this exam		
faminc	$\hat\beta_2$ 0.0624684	0.0324438			
_cons	$\hat\beta_0$ 118.5568	1.234278			

Model 2.2 : $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 fathereduc_i + \beta_4 mothereduc_i + u_i$ include (ตัวที่สนใจ

regress bwght cigs faminc fatheduc motheredu

Source	SS	df	MS
Model	15827.6593	4	3956.91482
Residual	466919.033	1186	393.69227
Total	482746.692	1190	405.669489

Number of obs = 1191
F (4, 1186) = 10.05
Prob > F = 0.0000
R-Squared = 0.0328
Adj. R-Squared = 0.0295
Root MSE = 19.842

bwght	Coef.	Std. Err. $se(\hat\beta_1)$	t	P>|t|	[95% conf. interval]
cigs	$\hat\beta_1$ -0.5894954	0.1106172			
faminc	$\hat\beta_2$ 0.0538254	0.0366502			
fatheduc	$\hat\beta_3$ 0.4936695	0.2832896	omitted for the purpose of this exam		
mothereduc	$\hat\beta_4$ -0.4379234	0.3197377			
_cons	$\hat\beta_0$ 118.0741	3.500291			

a) Based on Model 2.1, test whether smoking has an impact on birth weight. Show your work (use $\alpha = 0.05$)

t-test cigs; two tail test against zero

step1 : state hypothesis

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

step2 : calculate test statistics

$t_{cal} = \dfrac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \dfrac{-0.5876985 - 0}{0.1090181} = -5.3908$

(ตารางไม่มี 1188) เป็น ∞ เพราะ 180=

step3 : state decision rule

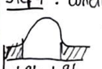
$\alpha = 0.05$ $df = n-k$

$df = 1191 - 3 = 1189$

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

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

step4 : conclude the test



-1.96 1.96

-5.3908

$t_{cal} <$ lower bound. We reject H_0 at 0.05 significance level. That 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 ← faminc;

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

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

$= 0.0624684 \pm 0.08357523$

$= (0.0211, 0.1460)$

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

t-test cigs; two tail test against zero

step1 : state hypothesis

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

step2 : calculate test statistics

$t_{cal} = \dfrac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \dfrac{-0.5894954 - 0}{0.1106172} = -5.3291$

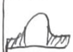
step3 : state Decision rule

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

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

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

step4 conclude test



-1.96 1.96

-5.3291

$t_{cal} <$ lower bound. We reject H_0 at 0.05 significance level. That is $\beta_1 \neq 0$. Smoking has a significant impact on infant birth weight.

The conclusion does not change from (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.

step1 : state hypothesis (F-test)

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

H_a: otherwise

step2 : calculate test statistics

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

step3 : State Decision Rule

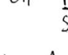
$\alpha = 0.01$

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

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

$= 3.32$

step4 conclude the test



3.32 10.05

$F_{cal} >$ upper bound. We reject H_0 at 0.01 significance level. That is the overall significance is at least 99%.

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

t_{cal} ไม่ได้ให้มา (หาเอง)

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

critical value

$t_{cal}(\hat{\beta}_1) = \dfrac{-0.5894954 - 0}{0.1106172} = -5.3291 < -1.96$

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

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

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

critical value

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

At 5% level of significance, $\hat{\beta}_1$ 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 $\alpha = 0.05$)

ใช้ทั้ง 2 model 2.1 & 2.2 เมื่อเทียบกันแล้วดูผล (F_{cal})-test 2 ตัวขึ้นไป

step1 : state Hypothesis

H_0: parents' education does not have an impact on birth weight ($\beta_3 = \beta_4 = 0$)

H_a: otherwise

ตัวแปรที่เพิ่มเข้ามาใน model 2.2 (เทียบกับ 2.1) → มี father, mother เพิ่มมา

step2 : calculate test statistics

$F_{cal} = \dfrac{(ESS_{incl} - ESS_{excl}) / \text{no. of additional } X}{RSS_{incl} / (n - k_{incl})}$

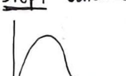
$= \dfrac{(15827.6593 - 14536.9538)/2}{466919.033/(1191-5)} = 1.639$

step3 state Decision Rule

$\alpha = 0.05$

upper bound $= F_{\alpha}(\text{no. of additional } X, n - k_{incl}) = F_{0.05}(2, 1186) = 3.00$

step4 conclude the test



1.639 3.00

$F_{cal} <$ upper bound. We fail to reject H_0 at 0.05 significance level. That is, parents' education does not have an impact on birth weight.

The impact of education on birth weight might already be reflected in variable family income

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
$kid6_i$ = number of children aged 0-6 in a household
$educ_i$ = years of education
age_i = age
$kid18_i$ = number of children aged 6-18 in a household

Source	SS	df	MS = SS/df
Model	ESS	K-1	ESS/K-1
Residual	RSS	n-K	ESS/n-K 0.446526442
Total	223.327441 TSS	n-1	TSS/n-1

Number of obs = 428
F(___ , ___) = 13.19
Prob > F = 0.0000
R-squared = 0.1582
Adj. R-squared = ___
Root MSE = 0.66823

lwage	Coef.	Std. Err.	t	P>|t|	[95% conf. Interval]	
exper	0.039819	0.013393	2.97	0.003	0.0134936	0.0661444
expersq	-0.0007812	0.0004022	-1.94	0.053	-0.0015718	9.37e-06
educ	-0.1078319	0.0144021	7.49	0.000	0.079523	0.1361409
age	-0.0014653	0.0052925	-0.28	0.782	-0.0118682	0.0089377
kidslt6	-0.0607106	0.0887626	-0.68	0.494	-0.2351836	0.1137625
kidsge6	-0.014591	0.278981	-0.52	0.601	-0.069428	0.0402459
_cons	-0.4209078	0.316905	-1.33	0.185	-1.043821	0.2020053

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 (MS) in this model

$MS = \frac{SS}{df}$

Residual MS = $\frac{RSS}{\text{df of RSS}}$

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

RSS = 0.446526442 (428-7)

RSS = 187.9876

ESS = TSS - RSS = 223.327441 - 187.9876
= 35.3398

Model MS = ESS / K-1 = $\frac{35.3398}{7-1}$ = 5.8900

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

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

Adjusted $R^2 = 1 - [(1-R^2)(\frac{n-1}{n-k})]$
$= 1 - [(1-0.1582)(\frac{427}{421})]$
$= 1 - 0.853797 = 0.1462$

3.2 ตัวแปรโดนตัดออกตัวนึง

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 $F_{1,421} = 3.84$)

ที่ให้มาคือ model 3.1, ส่วน model 3.2 ไม่ได้ให้มา (เป็น model ปริศนา)
จน.ตัวแปรใน 3.1 > จน.ตัวแปรใน 3.2
$\therefore R^2_{3.1} \geq R^2_{3.2}$
ถ้าตัวแปรที่โดนตัดไป สำคัญมาก, $R^2_{3.2}$ ก็จะน้อยกว่า $R^2_{3.1}$ มาก
โจทย์เลยถามว่า $R^2_{3.2}$ จะต้องมีค่าเท่าไร ถึงจะตัดสินว่า
ตัวแปรที่โดนตัดไป นั้นมีนัยสำคัญ (significant)

Model 3.1 (incl)
Model 3.2 (excl)

H_0: variable excluded in model 3.2 is not significant
H_a: otherwise

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

จน.ตัวแปรที่เพิ่มเข้ามา (Kincl - Kexcl)

$$\frac{(R^2_{incl} - R^2_{excl})/\text{no. of additional X}}{(1-R^2_{incl})/(n-K_{incl})} > 3.84$$

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

$$\frac{0.1582 - R^2_{3.2}}{0.00199952} > 3.84$$

$R^2_{3.2} < 0.1505$

$\therefore$ The maximum value of R^2 from model 3.2 that would lead to the conclusion that the excluded variable is significant at 0.05 significance level is 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?

Age has no impact on hourly wage. Does it make sense? What caused insignificance?

In my opinion, age could affect hourly wage. for example, as age increases, wage rises as workers gain expertise in their job and become more productive. However, the insignificance could be caused by the correlation with other variables such as experience or education. Moreover, the effect age has on wage could be non-linear and not captured by the model