AGEC5213: *ECONOMETRIC METHODS*
Spring 2019

PROBLEM SET NO. 1
Date Due: February 13, 2019

I. Estimate, Standard Error, t-statistic, and R-squared (5 points)

Researchers are often concerned with how consumers adjust their expenditures on some of addictive commodities (e.g., tobacco, alcohol) as their level of income increases. Consider the following econometric model that relates the proportion a household's income spent on alcohol (ALCOHOL), household's income (INCOME), age of the household head (AGE), and the number of children in the household (NUMKID).

$$ALCOHOL = \beta_1 + \beta_2 \ln(INCOME) + \beta_3 AGE + \beta_4 NUMKID + e$$

A household survey data were used to estimate this model, and the SAS output is kindly presented below. So you don't need to estimate this model by yourself (you can take a nap for a while). One minor problem is that my old laser printer has a problem with its ink cartridge, and as a result some parts of the print-out are left blank.

$t = \frac{b_K - \beta_K}{Se(b_K)}$

Dependent variable: ALCOHOL
Included Observation: 1519

Variable	Coefficient (b_K)	Std. Error ($Se(b_K)$)	t-Statistic (t)	Prob.
C	0.009052	0.024050	(i)	0.7347
Ln(INCOME)	0.527641	(ii)	6.608620	0.0000
AGE	(iii)	0.00208	-6.962389	0.0000
NUMKID	-0.013282	0.003259	-4.074993	0.0000
R-squared	(iv) = SSR/SST	Mean dependent var.		0.060596
Adjusted R-squared	0.053047	S.D. dependent var.		0.063325
S.E. of Regression (SSR)	(v) $\sqrt{\frac{SSE}{T-K}} = \sqrt{\frac{8.75}{1519-8}}$			
Sum squared error (SSE)	8.752896			

$CI = b_K \pm t_{\alpha/2} \cdot SE$ $\quad t_{\alpha/2} = 1.96$ at 95% CI

(a) Fill in the blank spaces that appear in the output.
(b) Interpret each of the estimates b_2, b_3, and b_4.
(c) Compute 95% interval estimates for β_2 and β_3. What do these interval estimates tell you?
(d) Test the hypothesis that the household income proportion for alcohol does not depend on the number of children in the household and interpret the test result.

II. Estimates, Standard Errors, and Hypothesis Test (7 points)

$SSR = \sqrt{\frac{SSE}{T-K}}$

$R^2 = \frac{SSR}{SST}$

Consider the regression model:

$R^2_{Adj} = 1 - \frac{(1-R^2)(T-1)}{(T-K)}$

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

$\widehat{Var}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$

$\hat{\sigma}^2 = \frac{e'e}{N-P}$ $e = Y - X\hat{\beta} = Y$

$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i; \quad e_i \sim N(0, \sigma^2)$

Data on this three-variable regression model yields the following results:

$$X'X = \begin{bmatrix} 25 & 0 & 0 \\ 0 & 10 & 20 \\ 0 & 20 & 50 \end{bmatrix}; \quad X'y \begin{bmatrix} 100 \\ 1 \\ 8 \end{bmatrix}; \quad \sum_i y_i^2 = 450$$

$b = \hat{\beta} - (X'X)^{-1} X'y$

$X' \Rightarrow X$ - transpose.

find inverse of $X'X$ matrix

(a) What is the sample size? 25

(b) Estimate parameters of the regression model.

(c) Estimate standard errors of b_2 and b_3.

(d) Test the hypothesis that $\beta_2 = 0$ against $\beta_2 \neq 0$; and the hypothesis that $\beta_3 = 1$ against $\beta_2 \neq 1$ (use 5% significance level for both tests).

(e) Compute R^2 and adjusted R^2 (both R^2 should be mean-centered).

III. Elasticity, R-Squared, Adjusted R-Squared, and Hypothesis Test (8 points)

Data on beginning salary (Y) and education in years (X) for 93 employees of Stillwater Bank can be found in the file, HW1-DATA.xls (see our D2L site). Estimate a log-log linear relationship between salary and education, i.e.,

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + e_i.$$

(a) Does the coefficient of education have the expected sign? What interpretation can you place on the slope coefficient?

(b) Predict the starting salary of an individual with 13 years of education.

(c) What is the interpretation of R-square from your output?

(d) Compute adjusted R-square and provide justification of using the adjusted R-square.

(e) Test if the elasticity of salary w.r.t. education is one. Test if all coefficients are zero. List null and alternative hypotheses, the choice of probability distribution, and degree of freedom for these tests.

(f) In addition to observations on Y and X, the data contains observations on number of months of previous work experience (E) and the number of months that the individual was hired for the Stillwater Bank (T). Run the regression using the same functional form, a double log form. Did b_1, b_2, R-squared, adjusted R-squared changed? What can you say about these different regression results?

(g) Test if elasticities of the number of months of previous work experience (E) and the number of months that the individual was hired for the Stillwater Bank (T) are the same. List null and alternative hypotheses, the choice of probability distribution, and degree of freedom for this test.

Ext: March 4

1. **Estimate, Standard Error, t-statistics, and R-Squared.**

Answers:

(A).

Dependent Variable: Alcohol Included Observation (N): 1519				
Variable	Coefficient (b_k)	Std. Error (SE_{bk})	T-Stat. (t)	Prob. (β_k)
C			-30.17247 (i)	0.3764
Ln(Income)		0.07984133 (ii)		
Age	-0.01448177 (iii)			
NUMKID				
R-squared	0.53139793 (iv)			
S.E. of Regression	0.07598472 (v)			

Equations used:

$t = (b_k - \beta_k)/ SE_{bk}$

S.E. Regression = √ {Sum Squared Error (SSE)/(N-k)} where, N = number of cases or observation and K is # of variables (4 in our case). (n-k) if degree of freedom.

1. (a).

(i) $t = \frac{0.009052 - (-0.7347)}{0.024050} = -30.17247$ $\beta_k = 0$ 0.3764

(ii) $6.608620 = \frac{0.52764 - 0}{SE}$ or $SE = 0.07984133$

(iii) $-6.962389 = \frac{b_k - 0}{0.10208}$ or $b_k = -0.01448177$

(v) $S.E.\ of\ Reg = \sqrt{\frac{SSE}{T-K}} = \sqrt{8.752896/(1519-4)} = 0.0759897228$

(iv) $R^2_{Adj} = 1 - \frac{SSE/(T-K)}{SST/(T-1)}$

$0.53047 = 1 - \frac{8.752896}{SST} \times \frac{1518}{1515}$; $SST = 18.67873931$

$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{8.752896}{18.67873931} = 0.5313979249$

$1 - \frac{R^2_{Adj}(T-1)}{(T-K)}$

(B)

Interpretation of b2: with the unit percentage increase in the household income keeping other variables constant, the proportion of household income spent on the alcohol increases (+ve sign) by 0.527641 unit.

Interpretation of b3: with the unit increase in the age of person in family keeping other variables constant, the proportion of household income spent on the alcohol decreases (-ve sign) by 0.0145 unit.

Interpretation of b4: with one additional child in the household and other situation remaining unchanged, the proportion of household income spent on the alcohol decreases (-ve sign) by 0.0133 unit.

C.

Solution c, b2

reg_coeff_b2 = 0.527641

st_err_b2 = 0.00208

Computing 95% CI for b2

ME_b2 = 1.96 * st_err_b2

CI_low_b2 = reg_coeff_b2 – ME_b2

CI_high_b2 = reg_coeff_b2 + ME_b2

 # CI_b2 = [0.3711519932, 0.6841300068]

$$CI = b_k \pm 1.96 \times SE(b)$$

If repeated sample of same size is drawn from same population, the true value lie between these two values interval for 95% of time.

Solution c, b3

reg_coeff_b3 = -0.01448177

st_err_b3 = 0.00208

Computing 95% CI for b3

ME_b3 = 1.96 * st_err_b3

CI_low_b3 = reg_coeff_b3 - ME_b3

CI_high_b3 = reg_coeff_b3 + ME_b3

 # CI_b3 = [-0.01855857, -0.01040497]

=> Intervals tell that if repeated samples of the same size were taken from the population, the true value of the estimates would lie between the referred intervals 95% of the time.

All the t statistics values that falls inside the range of confidence interval tells that these values are not significantly different than the null value. T-statistics that falls outside this interval are significantly different than null value and thus we can reject null hypothesis at 95% confidence interval. Here both, Ln(Income) and Age are significantly different than their null values as they fall outside the these

confidence interval. So, we reject null hypothesis (here absolute value of t-stat is greater than t-cric) in favor of alternative hypothesis for both b2 and b3 (ln(income) and Age). The ln(age) and the income has statistically significant effect in the proportion of household income spent on alcohol.

D. -0.2

Solution d, b4

reg_coeff_b4 = -0.013282

st_err_b4 = 0.003259

Computing 95% CI for b4

ME_b4 = 1.96 * st_err_b4

CI_low_b4 = reg_coeff_b4 - ME_b4

CI_high_b4 = reg_coeff_b4 + ME_b4

 # CI_b4 = [-0.01966964, -0.00689436]

$H_0: \beta = 0$

$H_A: \beta \neq 0$

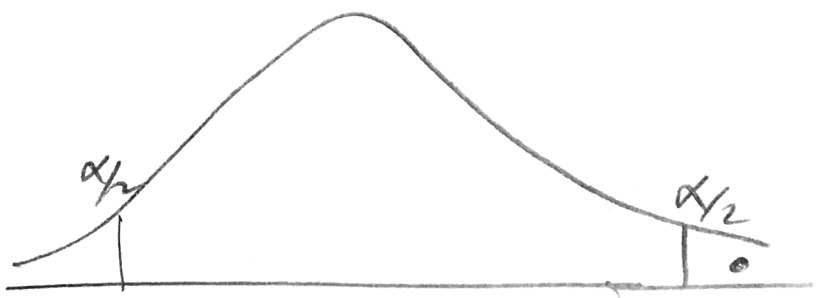
$t = \frac{b_k - \beta}{SE} = \frac{0.527641}{0.07984133}$

$= 6.66.$

$t_{crit_{\alpha/2}} = 1.96 < t_{calc}$

The higher and lower critical values at 95% confidence interval is as given above. The calculated t-statistics (-4.074993) does not lies within this range of values in the probability density function and thus the null hypothesis that the household income proportion for alcohol does not depend upon the number of children in the household is rejected. The result suggests that the household income proportion for the alcohol indeed depends upon the number of children in the household.

Ok; In this problem, You should use t-test not interval.



Reject H_0.

II. Estimates, Standard Errors, and Hypothesis Test

a. What is the sample size? Answer: 25.

b.

	B	C	D	E	F	G	H
17							
18		X'X	25	0	0		
19			0	10	20		
20			0	20	50		
21		Highlight area to put inverse matrix. X'X Inverse = =MINVERSE(D18:F20). Then hit Ctrl+Shift+Enter at a time.					X'Y
22		(X'X) Inverse	0.04	0	0		100 ← ΣY_t
23			0	0.5	-0.2		1
24			0	-0.2	0.1		8
25		Determinant of Matrix = mdeterm(array) and press enter. Det =MDETERM(D18:F20)					
26		Det. Mat X'X	2500				
27							
28		Multiply Matrics: =MMULT(D22:F24,H22:H24)					
29		Multiply two matrices	4	b1			
30			-1.1	b2			
31			0.6	b3			

$\bar{Y} = \frac{\Sigma Y_t}{N}$

$= XX'X'y$

B1, b2 and b3 are parameters.

$$(XX')^{-1} = \begin{bmatrix} 0.04 & 0 & 0 \\ 0 & 0.5 & -0.2 \\ 0 & -0.2 & 0.1 \end{bmatrix} \quad X'y = \begin{bmatrix} 100 \\ 1 \\ 8 \end{bmatrix}$$

$3 \times 3 \qquad 3 \times 1$

$$b = \begin{bmatrix} 4+0+0 \\ 0+0.5-1.6 \\ 0-0.2+0.8 \end{bmatrix} = \begin{bmatrix} 4 \\ -1.1 \\ 0.6 \end{bmatrix} = \begin{matrix} b_1 \\ b_2 \\ b_3 \end{matrix}$$

3×1

$$b = (XX')^{-1}X'y$$

c.

$SEE = yy - bx'y$

II. © standard errors of b_2 & b_2

$T = 25 \quad K = 3$

b_2

$y'y = \sum_t y_t^2 = 450$

$SSE = \hat{e}\hat{e}' = y'y - \hat{b}'x'y$

$$= 450 - \begin{bmatrix} 4 & -1.1 & 0.6 \end{bmatrix} \begin{bmatrix} 100 \\ 1 \\ 8 \end{bmatrix} = 450 - 400 + 1.1 - 4.8$$

$$= 96.3$$

$$SEE = \hat{\sigma}^2 = \frac{\sum \hat{e}^2}{T-K} = \frac{SSE}{T-K} = \frac{46.3}{22} = 2.104545$$

$$\text{Var}(b) = \hat{\sigma}^2 (x'x)^{-1} = 2.104545 \begin{bmatrix} 0.04 & 0 & 0 \\ 0 & 0.5 & -0.2 \\ 0 & -0.2 & 0.1 \end{bmatrix}$$

Cov-cov matrix

$$= \begin{bmatrix} 0.0841818 & 0 & 0 \\ 0 & 1.0522725 & -0.420909 \\ 0 & -0.420909 & 0.2104545 \end{bmatrix}$$

$$Se(b_2) = \sqrt{1.0522725} = 1.025803344$$

$$se(b_3) = \sqrt{0.2104545} = 0.4587532016$$

$$SSE = \hat{e}\hat{e}' - \hat{b}'x'y$$

$$= 450 - \begin{bmatrix} 4 & -1.1 & 0.6 \end{bmatrix} \begin{bmatrix} 100 \\ 1 \\ 8 \end{bmatrix} = 96.3$$

$$\hat{\sigma}^2 = \frac{SSE}{T-K} \quad \& \quad \sigma = \sqrt{\frac{SSE}{T-K}}$$

$$\text{Var}(b) = (XX')\hat{\sigma}^2$$

d.

II d.

$$t_{stat} = \frac{b_k - \beta_k}{se_{(bk)}}$$

$H_0 : \beta_2 = 0$

$H_A : \beta_2 \neq 0$

$$t_{stat} = \frac{-1.1 - 0}{1.0258} = -1.07$$

You should test using t-value and t critical value not Interval

$$CI = b_k \pm t_{\alpha/2} \cdot SE$$

$$= -1.1 \pm 1.967 \times 1.0258 = [-3.110568, 0.910568]$$

Since t_{calc} lies within the CI, we failed to reject null hypothesis

$H_0 : \beta_3 = 1$

$H_0 : \beta_3 \neq 1$

$$t_{stat} = \frac{0.6 - 1}{0.45875} = 0.87193$$

$$CI = b_k \pm t_{\alpha/2} \cdot SE$$

$$= 0.6 \pm 1.967 \times 0.45875 = [-0.30236, 1.50236]$$

Since t_{calc} falls within the confidence interval, we failed to reject null hypothesis

e.

II e

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad \text{Adj } R^2 = 1 - \frac{SSE/(T-K)}{SST/(T-1)}$$

$$X'y = \begin{bmatrix} \Sigma Y_t \\ \Sigma X_{2t} \cdot Y_t \\ \Sigma X_{3t} \cdot Y_t \end{bmatrix} = \begin{bmatrix} 100 \\ 1 \\ 8 \end{bmatrix}$$

ee' or $\hat{e}^2$; Average of Y.

$$\Sigma_n Y_t = 100, \quad \Sigma Y_t^2 = 450 \quad \bar{y} = \frac{\Sigma Y_t}{n} = \frac{100}{25} = 4$$

$$SST = \sum_n (Y_t - \bar{Y})^2$$

$$= \sum_n (Y_t^2 - 2Y_t\bar{Y} + \bar{Y}^2) = \sum_n (Y_t^2 - 2Y_t \cdot 4 + 4^2)$$

$$= \sum_n Y_t^2 - 8\sum_n Y_t + 16\sum_n$$

$$= 450 - 8 \cdot 100 + 16 \times 25$$

$$= 50$$

$SSE = 46.3$

$SSR = SST - SSE = 50 - 46.3 = 3.7$

$$R^2 = \frac{SSR}{SST} = \frac{3.7}{50} = 0.074$$

$$R^2_{Adj} = 1 - \frac{SSE/(T-K)}{SST/(T-1)} = 1 - \frac{(46.3)/22}{50/24} = 1 - 1.01018$$

$$= -0.010182$$

III. Elasticity, R-Squared, Adjusted R-Squared, and Hypothesis Test

a.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: lny

Number of Observations Read	94
Number of Observations Used	93
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.22901	0.22901	15.94	0.0001
Error	91	1.30703	0.01436		
Corrected Total	92	1.53604			

Root MSE	0.11985	R-Square	0.1491
Dependent Mean	8.58961	Adj R-Sq	0.1397
Coeff Var	1.39524		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.95996	0.15817	50.32	<.0001
lnX	1	0.25108	0.06288	3.99	0.0001

a. I expect the rise in salary with the higher level of education i.e. positive relationship between these two variables. So, yes, the education has the expected sign with the salary. With the unit percentage change (say, increase) in the education level, the change in the salary (increase in this case) is by 0.25108% unit.

b. ln(y) = 7.95996 + 0.25108 Ln(x)

When x = 13, ln(y) = 7.95996 + 0.25108 ln(13) = 8.603967.

Y-hatt = EXP(8.603967) = 5453.252

c. The value of R^2 = 0.1491 (14.91%) means the ln(x) (natural log of education in years) explains 14.91% of variation in the ln(y) (natural log of salary).

d. 0.4

$$R^2 = 1 - \frac{SSE/(T-K)}{SST/(T-1)}$$

$$SSE = 1 - \frac{SSE}{SST} \times \frac{(T-1)}{(T-K)}$$

$$= 1 - \frac{1.30703}{1.53604} \times \frac{92}{91}$$

$$= 1 - 0.860259477$$

$$= 0.13974$$

R squared explains the variation to all the independent variables in the model. So, as we add more variables into the model, we will get higher R-squared value. However, adjusted R-squared only explains the variation in the model due to significant variables in the model. So, it is not possible to increase the value of adjusted R-squared just by adding new variable into the model.

$\bar{R}^2$ value is corrected for the numbers of regressors (independent variable) included. (You should mention the relationships between the number of independent variable and value of R^2 and $\bar{R}^2$)

The value of ##

e. -0.5

The SAS System

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable lny				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	2.03761	141.87	<.0001
Denominator	91	0.01436		

Null Hypothesis: H_0: $\beta_2 = 1$

Alternative Hypothesis H_a: $\beta_2 \neq 1$

Choice of Probability Distribution: F Distribution

Degree of Freedom = f (1, 91)

The result shows that the elasticity of salary w.r.t. education is not one (Pr < 0.001), reject null hypothesis.

The SAS System

The REG Procedure
Model: MODEL1

Test 2 Results for Dependent Variable lny				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	36.37462	2532.53	<.0001
Denominator	91	0.01436		

Null Hypothesis: H_0: $\beta_1 = 0$

Alternative Hypothesis H_a: $\beta_1 \neq 0$

Choice of Probability Distribution: F Distribution

Degree of Freedom = f (1, 91)

We reject null hypothesis that intercept is zero (P < 0.0001).

The SAS System

The REG Procedure
Model: MODEL1

Test 3 Results for Dependent Variable lny				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.22901	15.94	0.0001
Denominator	91	0.01436		

Null Hypothesis: H_0: $\beta_2 = 0$;

Alternative Hypothesis H_a: $\beta_2 \neq 0$

$H_0: \beta_1 = 0, \beta_2 = 0$

$H_1: \beta_1 \neq 0$ and or $\beta_2 \neq 0$

Choice of Probability Distribution: F Distribution

Degree of Freedom = f (1, 91)

The result shows that the lnx is not zero (Pr < 0.001). We can reject null hypothesis.

f.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: lny

Number of Observations Read	94
Number of Observations Used	90
Number of Observations with Missing Values	4

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.44972	0.14991	14.43	<.0001
Error	86	0.89354	0.01039		
Corrected Total	89	1.34326			

Root MSE	0.10193	R-Square	0.3348
Dependent Mean	8.59773	Adj R-Sq	0.3116
Coeff Var	1.18556		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.72596	0.14504	53.27	<.0001
lnX	1	0.25183	0.05350	4.71	<.0001
lne	1	0.03694	0.00990	3.73	0.0003
lnt	1	0.03354	0.01177	2.85	0.0055

Yes, b_1, b_2, R-squared and Adjusted R-squared value changed. The addition and removal of independent variables in the analysis changes the model that we are developing or using to predict the dependent variable which also changes these values.

The addition of two more variables in the model increased the value of adjusted R-squared which signifies that addition of these variables help to develop better model than before.

g.

The SAS System

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable lny				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.00050707	0.05	0.8257
Denominator	86	0.01039		

Null Hypothesis: H_0: $\beta_2 - \beta_3 = 0$; $H_0: \beta_3 = \beta_4$

Alternative Hypothesis H_a: $\beta_2 - \beta_3 \neq 0$ $H_a: \beta_3 - \beta_4 \neq 0$

Choice of Probability Distribution: F Distribution

Degree of Freedom = f (1, 86)

We failed to reject null hypothesis based on the P value (0.8257) which is greater than 0.05 at 95% confidence interval.

$F_{1,86} = 3.94$

need conclusion

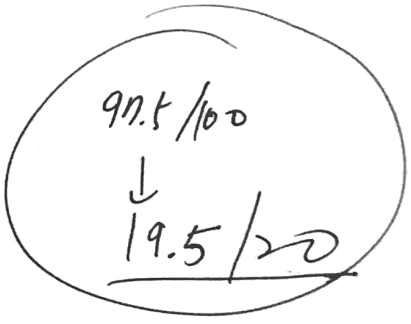
SAS Code:

Import Data and Open Project:

```
PROC IMPORT OUT= WORK.bm
        DATAFILE= "C:\Users\casnrlab_agh128\Desktop\EconHW\HW1-DATA.xls"
        DBMS= EXCEL REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;
dbms = excel replace;
range = sheet1$
getnames = yes;
mixed = no;
scantext = yes;
scantime = yes;
run;

data bm; set bm;
lny = log (y);
lnX = log(x);
run;

# a. # e.
proc reg data = bm;
model lny = ln(x);
test lnx = 1;
test intercept = 0;
test lnx = 0;
run;
proc print;
```



```
run;

#f #g
data bm; set bm;
lny = log(y);
lnx = log(x);
lne = log(e);
lnt = log(t);

proc reg data = bm;
model lny = lnx lne lnt;
test lne-lnt = 0;

run;
proc print;
run;
```