Econometrics Exercise session for midterm preparation solutions

Problem 1

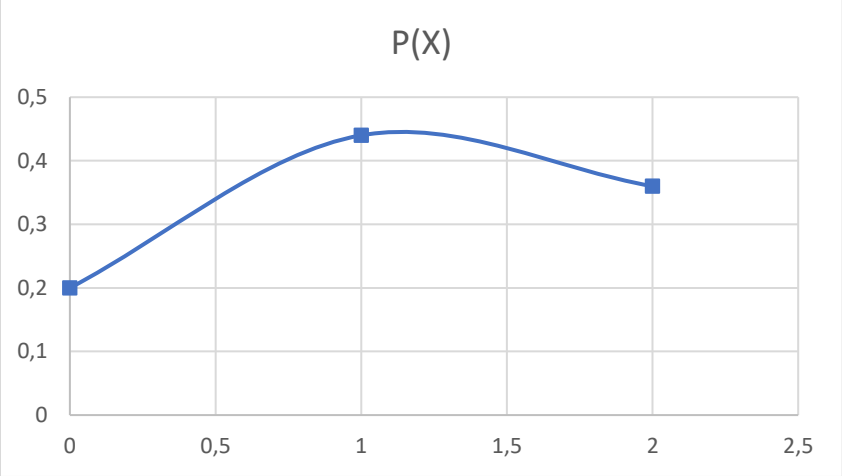
Suppose that X is the number of free throws made by a basketball player out of two attempts and assume that the individual probabilities for each outcome of X are the following:

pr(x=0)=0.2; pr(x=1)=0.44 and pr(x=2)=0.36

i) Define the random variable.

X	0	1	2
P(X)	0.2	0.44	0.36

ii) Draw the probability distribution associated to the above random variable.



iii) Calculate the expected value of the above random variable.
E(x)=0*0.2+1*0.44+2*0.36=1.16

iv) Calculate the probability that the player makes at least one free throw

P(X≥1)=1-P(X<1)=1-P(X=0)=1-0.2=0.8

Problem 2

We have a dataset containing data about births to women in the United States. Two variables of interest are the dependent variable, infant birth weight in ounces (bw), and an explanatory variable, average number of cigarettes the mother smoked per day during pregnancy (cigs). The following simple regression was estimated using data on 1,388 births:

$$\widehat{BW}_i = 119.77 - 0.514 cigs_i$$

i) Think about possible factors contained in the error term u_i .
Prenatal doctor visits, stressful environment of the mother, gender of a child, mother education, family income

ii) Interpret the above regression results.

If number of cigarettes smoked per day increases by one unit, then the birth weight of the child reduces by 0.5 ounces. Ceteris paribus, average birth weight of a child is 119.77 ounces

iii) What is the predicted birth weight when cigs =10? What about when cigs = 20 (one pack per day)? Comment on the difference.

119.77-0.514*10=114.63

119.77-0.514*20=109.49

Problem 3

We have information about mortality rates (MORT=total mortality rate per 100,000 population) in a specific year for 51 States of the United States combined with information about potential determinants: INCC (per capita income by State in Dollars), POV (proportion of families living below the poverty line), EDU (proportion of population completing 4 years of high school), TOBC (per capita consumption of cigarettes by State) and AGED (proportion of population over the age of 65). Estimation results are presented in the following table:

OLS Estimation Results

Variable	Model 1 coefficients	Model 2 coefficients	Model 3 coefficients
Constant	194.747	531.608	-9.231
	(53.915)	(94.409)	(176.795)
Aged	5,546.56	5,024.38	5,311.4
	(445.727)	(358.218)	(334.415)
Incc		0.014	0.015
		(0.0038)	(0.0037)
Edu		-682.591	-285.715
		(114.812)	(152.926)
Pov			854.178
			(302.345)
Tobc			0.989
			(0.342)
n	51	51	51
Adjusted R squared	0.759	0.856	0.884
SSR	228,770.3	128,260.1	99,303.73

i) Interpret the slope coefficient in Model 1 and validate it at 1% significance level.

The slope coefficient in Model 1 implies that increasing the proportion of population over the age 65 by 1 percentage points will increase the mortality rate per 100000 of population by 5 546. T statistics of the coefficient is t=$\frac{5556.56}{445.727} = 12.46$, which is > 2.7, therefore, it is significant at the 1% significance level

ii) Validate the joint significance of Model 2 in comparison to model 1 at 1% significance level?

Joint significance of model 2 in comparison to model 1 implies that we test joint significance of the coefficients Incc and Edu, according to F test we have $F = \frac{(228770.3-128260.1)/2}{128260.1/47} = \frac{50255.1}{2728.94} = 18.41$ comparing this with the critical value 5.1 we reject the hypothesis that the coefficients Incc and Edu are jointly insignificant

iii) Comment on the effect of INCC on MORT in the second model. Why do you think is a positive and significant effect? **When per capita income by state is higher, the mortality rate is also higher. This at one glance does not make any sense because rich people should be able to afford better health care and therefore, extend longevity of their lives. However, we can make an argument that generally, older people are more likely to have higher income, therefore, those states with high income probably also have proportion of old people higher and hence, the mortality rate is higher.**

iv) In Model 3 we add two new explanatory variables: POV and TOBC. Test whether this inclusion helps to improve the quality of the model at 1% significance level. Is model 3 the best in terms of goodness-of-fit? **This question indirectly asks to compare the model 3 to model 2, therefore, we need to test joint significance of the variables POV and TOBC**

Again we calculate an F testF= $\frac{(128260.1-99303.73)/2}{99303.73/45} = \frac{14478}{2206} = 6.56,$ **which is still larger than the critical value 5.1, therefore, these two variables are jointly significant at the 1% significance level. The model is the best in terms of the goodness of fit, because R^{adj} is the highest**

v) Are the effects of these two new variables the expected ones? Are they individually significant at 1% significance level? **These two new variables have "positive" impact on the mortality rate, which makes sense, more smokers- higher mortality, more poor people – higher mortality. T test for POV is 2.8 and for TOBC, 2.9, both are statistically significant at 1% significance level when comparing to critical value 2.7**

vi) What about the individual significance of EDU in model 3 if compared with model 2? Why? **Edu in model 2 is significant while it is not in model 3. The fact that POV and TOBC were omitted in the model 2 was causing a bias in the estimation of the coefficient Edu. The reason is that EDU is negatively correlated with both POV and TOBC – more educated people are less likely to smoke and less likely to be poor. Meanwhile, POV and TOBC are positively correlated with the explained variable MORT, meaning that the direction of bias is negative. Since coefficient on Edu is negative, this bias was making it more negative in absolute terms and this way it was making it significant.**

Problem 4

Suppose you are interested in studying the tradeoff between time spent sleeping and working and to look at other factors affecting sleep. You specify the following model:

$$sleep = \beta_0 + \beta_1 * totwrk + \beta_2 * educ + \beta_3 * age + u$$

where $sleep$ and $totwrk$ (total work) are measured in minutes per week and $educ$ and age are measured in years.

Suppose we estimated the following regression:

$$\widehat{sleep} = \underset{(112.28)}{3638.25} + \underset{(.017)}{0.148} * totwrk - \underset{(5.88)}{11.13} * educ + \underset{(1.45)}{2.2} * age$$

$n = 706,\ R^2 = .113$

where we report standard errors along with the estimates.

(i) Is either *educ* or *age* individually significant at the 5% level against a two-sided alternative? Show your work.

$$\boldsymbol{t_{educ} = \frac{11.13}{5.88} = 1.89}, \qquad \boldsymbol{t_{age} = \frac{2.2}{1.45} = 1.52}$$

Critical value at 5% with two tails and df=702 is $t_{cr} = 1.96$, therefore both age and educ are individually insignificant

(ii) Dropping *educ* and *age* from the equation gives

$$\widehat{sleep} = \underset{(38.91)}{3586.38} + \underset{(.017)}{0.151} * totwrk$$

$$n = 706,\ R^2 = .103$$

Are *educ* and *age* jointly significant in the original equation at the 5% level? Justify your answer.

We know that $F = \frac{(SSR_r - SSR_{ur})/q}{\frac{SSR_{ur}}{df}}$, here q is the number of restrictions. We also know that

$$\boldsymbol{R^2 = \frac{ESS}{TSS} = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS} \Rightarrow SSR = (1 - R^2) * TSS}$$

TSS will be the same for both restricted and the unrestricted models, therefore it will cancel out. We will have:

$$\boldsymbol{F = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} * \frac{df}{q} = \frac{0.113 - 0.103}{1 - .113} * \frac{702}{2} = 3.96}$$

The 5% critical value in the F table at $F_{2,702} = 3$, Therefore, we reject the hypothesis that age and education are jointly insignificant at the 5% level (3.96 > 3.00). In fact, the p-value is about .019, and so educ and age are jointly significant at the 2% level.

(iii) Does including *educ* and *age* in the model greatly affect the estimated tradeoff between sleeping and working?

Not really. These variables are jointly significant, but including them only changes the coefficient on totwrk from –.151 to –.148.

(iv) Suppose that the sleep equation contains heteroskedasticity. What does this mean about the tests computed in parts (i) and (ii)?

The standard t and F statistics that we used assume homoskedasticity. If there is heteroskedasticity in the equation, the tests are no longer valid. In fact, standard errors without controlling heteroskedasticity are smaller than what it should be - increasing the significance of the estimated parameters, which is wrong.

Problem 5

consider a simple model to compare the returns to education at junior colleges and four-year colleges; for simplicity, we refer to the latter as "universities." The population includes working people with a high school degree, and the model is:

$\log(wage) = \alpha_0 + \alpha_1 jc + \alpha_2 univ + \alpha_3 exper + u$ (1)

where

jc is number of years attending a two-year college, *univ* is number of years at a four-year college. *exper* is months in the workforce.

Note that any combination of junior college and four-year college is allowed, including *jc* =0 and *univ* = 0. Use the data ***twoyear.dta***

a) Test the hypothesis that $\alpha_1 = \alpha_2$. The hypothesis of interest is whether one year at a junior college is worth one year at a university.
To test this hypothesis we instead want to test $\theta = \alpha_1 - \alpha_2 = 0$ and plug it in the original regression:

$$\log(wage) = \alpha_0 + (\theta + \alpha_2) jc + \alpha_2 univ + \alpha_3 exper + u$$
$$\log(wage) = \alpha_0 + \theta jc + \alpha_2 (univ + jc) + \alpha_3 exper + u \quad \textbf{(2)}$$

Now run:
genr unjc=univ+jc
ols lwage const jc unjc exper

	coefficient	std. error	t-ratio	p-value	
const	1.47233	0.0210602	69.91	0.0000	***
jc	-0.0101795	0.00693591	-1.468	0.1422	
unjc	0.0768762	0.00230873	33.30	2.96e-225	***
exper	0.00494422	0.000157474	31.40	4.12e-202	***

Mean dependent var	2.248096	S.D. dependent var	0.487692
Sum squared resid	1250.544	S.E. of regression	0.430138
R-squared	0.222442	Adjusted R-squared	0.222097
F(3, 6759)	644.5330	P-value(F)	0.000000
Log-likelihood	-3888.687	Akaike criterion	7785.374
Schwarz criterion	7812.651	Hannan-Quinn	7794.789

$\widehat{\alpha_1} - \widehat{\alpha_2} = -0.0102$ **so the return to a year at a junior college is about one percentage point less than a year at a university.**

Test statistic on jc t=0.0102/.0069 =-1.48. We need to compare this with one sided alternative critical value. At 10% one-sided significance level, critical value is -1.282. Therefore, there is some but not strong evidence against the null hypothesis.

Check also command: ols lwage const jc univ exper. Make your own observation!

(ii) The variable phsrank is the person's high school percentile. (A higher number is better. For example, 90 means you are ranked better than 90 percent of your graduating class.) Find the smallest, largest, and average phsrank in the sample.

summary phsrank

(ii) Add phsrank to regression (2) and report the OLS estimates in the usual form. Is phsrank statistically significant? How much is 10 percentage points of high school rank worth in terms of wage?

ols lwage const jc unjc exper phsrank

***phsrank* has a *t* statistic equal to only 1.25; it is not statistically significant. If we increase *phsrank* by 10, log(*wage*) is predicted to increase by (.0003)10 = .003. This implies a .3% increase in *wage*, which seems a modest increase given a 10 percentage point increase in *phsrank*.**

(iii) Does adding phsrank to regression (2) substantively change the conclusions on the returns to two- and four-year colleges? Explain.

Adding *phsrank* makes the *t* statistic on *jc* even smaller in absolute value, about 1.33, but the coefficient magnitude is similar to (2). Therefore, the base point remains unchanged: the return to a junior college is estimated to be somewhat smaller, but the difference is barely significant with one-sided test.

Problem 6

A soda vendor at Louisiana State University football games observes that more sodas are sold the warmer the temperature at game time is. Based on 32 home games covering five years, the vendor estimates the relationship between soda sales and temperature to be $\hat{y} = -240 + 8x$ where y is the number of sodas she sells and x is temperature in degrees Fahrenheit,

(a) Interpret the estimated slope and intercept. Do the estimates make sense? Why, or why not?

The intercept estimate b_1 =-240 is an estimate of the number of sodas sold when the temperature is 0 degrees Fahrenheit. A common problem when interpreting the estimated intercept is that we often do not have any data points near x =0 . If we have no observations in the region where temperature is 0, then the estimated relationship may not be a good approximation to reality in that region. Clearly, it is impossible to sell -240 sodas and so this estimate should not be accepted as a sensible one.

The slope estimate b_2=8 is an estimate of the increase in sodas sold when temperature increases by 1 Fahrenheit degree. This estimate does make sense. One would expect the number of sodas sold to increase as temperature increases.

(b) On a day when the temperature at game time is forecast to be 80^0F, predict how many sodas the vendor will sell.

If temperature is 80^{0}F, the predicted number of sodas sold is

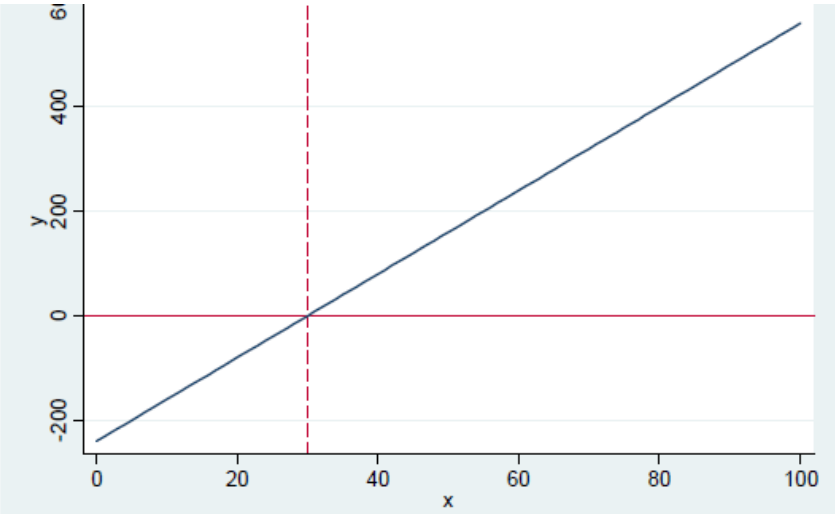
$\hat{y} = -240 + 8 * 80 = 400$

(c) Below what temperature are the predicted sales zero?

If no sodas are sold, y =0, and

$0 = -240 + 8 * x \Rightarrow x = 30$

(d) Sketch a graph of the estimated regression line.



Problem 7

Data on the weekly sales of a major brand of canned tuna by a supermarket chain in a large midwestern U.S. city during a mid-1990s calendar year are contained in the file *tuna.dat.* There are 52 observations on the variables. The variable SAL1= unit sales of brand no. 1 canned tuna, APR1= price per can of brand no. 1 canned tuna, APR2, APR3= price per can of brands nos. 2 and 3 of canned tuna.

(a) Create the relative price variables RPRICE2= APR1/APR2 and RPRICE3=APR1/APR3. What do you anticipate the relationship between sales (SAL1) and the relative price variables to be? Explain your reasoning.

The relationship between sales (SAL1) and the relative price variables is expected to be a negative one. Since brands 2 and 3 are substitutes for brand 1, an increase in the price of brand 1 relative to the price of brand 2, or relative to the price of brand 3, will lead to a decline in the sales of brand 1.

(b) Estimate the log-linear model $\ln(SAL1) = \beta_0 + \beta_1 PRICE2 + \varepsilon$. Interpret the estimate of β_1. Construct and interpret a 95% interval estimate of the parameter.

genr PRICE2=apr1/apr2
genr PRICE3=apr1/apr3
genr lnsal1=ln(sal1)
ols lnsal1 PRICE2 const

```
Model 1: OLS, using observations 1-52
Dependent variable: lnsall

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       10.2758       0.518528    19.82     2.49e-025 ***
  PRICE2      -1.85807      0.513899    -3.616    0.0007    ***

Mean dependent var   8.437187   S.D. dependent var   0.813654
Sum squared resid    26.76564   S.E. of regression   0.731651
R-squared            0.207265   Adjusted R-squared   0.191410
F(1, 50)             13.07278   P-value(F)           0.000696
Log-likelihood      -56.51756   Akaike criterion     117.0351
Schwarz criterion    120.9376   Hannan-Quinn         118.5312
```

The estimated log-linear regression is:

$$\ln(SAL1) = 10.2758 - 1.85807 PRICE2$$

The typical interpretation of β_1 in a log-linear model is that 1-unit increase in x will lead to a 100 *β_1% increase in y. In this particular case, where RPRICE2 is a unit-free relative price variable, it is not so meaningful to talk about a 1-unit increase in RPRICE2. Instead, we consider the elasticity

$$\frac{\partial(SAL1)}{\partial PRICE2} * \frac{PRICE2}{SAL1} = \beta_1 * PRICE2$$

We can interpret β_1 as the percentage change in sales from a 1% increase in the relative price when the prices of the two brands are identical (RPRICE2 =1). In terms of our estimate, and considering a price change of a realistic magnitude: If the prices of brands 1 and 2 are the same, and the relative price of brand 1 to brand 2 increases by 10%, the sales of brand 1 will decline by 18.58%. Demand is elastic.

A 95% interval estimate for β_1 from the regression is:

$$b_1 \pm t_{(0.975,50)} * se(b_1) = -1.85807 \pm 2.009 * 0.5139 = (-2.890, -0.826)$$

This interval estimate suggests that, with 95% confidence, when the two prices are the same, a 10% increase in the relative price of brand 1 tuna to brand 2 tuna will decrease sales of brand 1 by between 8.26% and 28.90%.

(c) Test the null hypothesis that the slope of the relationship in (b) is zero. Create the alternative hypothesis based on your answer to part (a). Use the 1% level of significance and draw a sketch of the rejection region. Is your result consistent with economic theory?

We set up the following hypothesis test:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 < 0$$

The test statistic, given H0 is true, is

$$\boldsymbol{t = \frac{b_1}{se(b_1)} = -3.616}$$

Rejection region is $\boldsymbol{t < -2.403 = t_{(0.01,50)}}$
Because -3.616<-2.403, our test statistic falls in the rejection region
We conclude that there is a statistically significant inverse relationship between the unit sales of brand 1 tuna and the relative price of brand 1 tuna to brand 2 tuna. This result is consistent with economic theory, as it is expected that demand for a good should be inversely related to the relative price of that good to a substitute good.

(d) Estimate the log-linear model ln $\ln(SAL1) = \gamma_0 + \gamma_1 PRICE3 + \epsilon$. Interpret the estimate of γ_1. Construct and interpret a 95%interval estimate of the parameter.
(e) Test the null hypothesis that the slope of this relationship is zero. Create the alternative hypothesis based on your answer to part (a). Use the 1% level of significance and draw a sketch of the rejection region. Is your result consistent with economic theory?