Econometrics Exercise session for midterm preparation

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.
ii) Draw the probability distribution associated to the above random variable.
iii) Calculate the expected value of the above random variable.
iv) Calculate the probability that the player makes at least one free throw

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

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.

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

iii) Comment on the effect of INCC on MORT in the second model. Why do you think is a positive and significant effect?

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?

v) Are the effects of these two new variables the expected ones? Are they individually significant at 1% significance level?

vi) What about the individual significance of EDU in model 3 if compared with model 2? Why?

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.

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

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

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

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

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

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

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:

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

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.

(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.

(iii) 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?

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

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?

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

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

(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.

(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.

(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?

(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?