Assignment 2: Bivariate and Multiple Regression

ECO 321

DUE: Thursday March 8 in class

Instructions: You need to write a Stata do-file to answer question (2). Your do-file must produce a log file that contains your Stata output. You must submit your log file along with your answers to the assignment. Assignments that do not contain log files will be marked down. You may work in groups (up to 4 people), but each student must submit his/her own assignment. If you work in a group and do not submit your own assignment, you will receive a zero for the assignment. Remember that late assignments will not be accepted.

1. Consider the OLS estimates of the multiple regression model:

$$\hat{Y}_i = \underset{SE(\hat{\beta}_0)}{\hat{\beta}_0} + \underset{SE(\hat{\beta}_1)}{\hat{\beta}_1} X_{i1} + \underset{SE(\hat{\beta}_2)}{\hat{\beta}_2} X_{i2}$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are the estimates of the coefficients, and $SE(\hat{\beta}_0)$, $SE(\hat{\beta}_1)$, and $SE(\hat{\beta}_2)$ are the standard errors of the estimates.

a) What is the definition of consistency? Which assumptions do we need for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ to be consistent?

b) We want to test the statistical significance of β_2. Write down the null hypothesis, alternative hypothesis, select α, and write down the t-statistic.

c) Write down the steps you would use to find the p-value for the statistical significance test for β_2. Now suppose $\hat{\beta}_2 = 2.1$ and $SE(\hat{\beta}_2) = 1$. Find the p-value.

d) Write down the general expression for a 99% confidence interval for β_1.

e) Suppose $\hat{\beta}_1 = 3.5$ and $SE(\hat{\beta}_1) = 2.2$. What is the 99% confidence interval for a one-unit change in X_{i1} ($\Delta X_{i1} = 1$)? What is the 99% confidence interval for a 8-unit change in X_{i1} ($\Delta X_{i1} = 8$)? Using only the confidence intervals, what can we conclude about the statistical significance of β_1?

(f) Based on your answers to parts (c) and (e), how might you rewrite the model? (Hint: Would you exclude any variable, and if so, why?)

2. Smoking and birth outcomes. The data set bwght.dta contains two variables: the birth weight (in ounces) of a newborn (*bwght*) and the number of cigarettes the mother smoked per day during pregnancy (*cigs*).

a) Load the data set in Stata (via the `use` command). Construct a dummy variable (using the `gen` command) named *anycig* that equals one if the mother smoked at least one cigarette per day and zero otherwise. Use the `tab` command (or whatever command you like) to determine what share (percent) of moms smoked during pregnancy.

b) We want to test the null hypothesis that the average birth weight of babies born to smokers vs. non-smokers is the same. Use the `reg` command to estimate the model: $bwght_i = \beta_0 + \beta_1 anycig_i + u_i$. How do you interpret $\hat{\beta}_1$ in this regression? How do you interpret $\hat{\beta}_0$? Can you reject the null hypothesis that the average birth weights of babies born to smokers vs. non-smokers is the same (use $\alpha = 0.05$)?

c) Rather than looking at the effect of smoking versus not smoking, we want to determine how smoking intensity affects birth weights. Use the `reg` command to estimate: $bwght_i = \beta_0 + \beta_1 cigs_i + u_i$. Report your regression results in standard form. What is the marginal effect of smoking an additional cigarette on birthweight? Is the effect statistically significant?

d) Using your regression results in part (c), compute the predicted birth weight of a child whose mother does not smoke, and of a child whose mother smokes a pack a day (assume 20 cigarettes per pack). Is the difference between these values statistically significant?

e) Does the simple regression in part (c) necessarily capture a causal relationship between the child's birth weight and the mother's smoking habits? Explain. (Hint: What is contained in u_i?)

f) What is homoskedasticity, and is it likely to be a valid assumption in this example?

g) Estimate the regression from part (c) without assuming homoskedasticity (i.e. use the `reg` command, but specify the option `robust` to account for heteroskedasticity in the error terms). Compare your standard errors to the standard errors from the regression that assumes homoskedasticity. What is the difference? Do you suspect heteroskedasticity is a problem in this data? Is there a benefit or cost to using robust standard errors?

3. In this question we want to explore the importance (or lack thereof) of unionization on household earnings. In particular, some economic theories suggest that collective bargaining (such as unionization) matters more for traditionally marginalized groups—female and minority workers. Consider the models:

$$FamIncome_i = \beta_0 + \beta_1 HAge_i + \beta_2 WAge_i + u_i \qquad (1)$$

$$FamIncome_i = \beta_0 + \beta_1 HAge_i + \beta_2 WAge_i + \beta_3 HUnion_I + \beta_4 WUnion_i + u_i \qquad (2)$$

where $FamIncome_i$ is a family's annual income, $HAge_i$ is the husband's age, $WAge_i$ is the wife's age, $HUnion_i$ is a dummy variable indicating that the husband is in a union, and $WUnion_i$ is a dummy variable for the wife's union status. Suppose average family income in the sample is \$50,000. Estimating these models using OLS gives the following results:

	(1)	(2)
$HAge$	-90.8	205
	(101)	(125)
$WAge$	495	322
	(112)	(134)
$HUnion$		-1805
		(1351)
$WUnion$		8972
		(1594)
$Constant$		24164
		(2380)
R^2	0.0173	0.0441
n	2574	2574

a) Consider regression (1). How do you interpret the coefficient on $WAge$? Is the coefficient estimate on $WAge$ in regression (1) economically significant?

b) According to regression (2), how much more (in dollars) can the average (married) family expect to earn in ten years (holding union status constant)?

c) Construct a t-statistic for the significance test of $HAge$ in regression (2). Is the relationship between husband's age and family earnings statistically significant?

d) How do you interpret the coefficient on $HUnion$?

e) What is the average effect on family income of a husband and wife both joining a labor union, if neither were previously in a union?

f) Extra Credit: Assuming homoskedasticity, are the coefficients on unionization jointly statistically significant?