Assignment 1: ECON 2220 A (2020 Late Summer)

Due: 23 July 2020 (to be submitted at cuLearn)

Instructions:

1. You should use STATA software for any calculations. You must paste your log/output onto your assignment (in a word document) to show your use of STATA; however, this output does not replace any of the steps outlined below.

2. If you are performing a hypothesis test, make sure you state the hypotheses, the level of significance, the rejection region, the test statistic (and p-value, if requested), your decision (whether to reject or not to reject the null hypothesis), and a conclusion in managerial terms that answers the question posed. These steps must be completed in addition to any STATA output.

3. The required STATA data files can be found in the corresponding folder of cuLearn

4. Please do not forget to write your name and student ID on the cover page of the assignment.

Project 1:

The Transactional Records Access Clearinghouse at Syracuse University reported data showing the odds of an Internal Revenue Service (IRS) audit. The following table shows the average adjusted gross income reported and the percent of the returns that were audited for 20 selected IRS districts.

District	**Adjusted Gross Income ($)**	**Percent Audited**
Los Angeles	36,664	1.3
Sacramento	38,845	1.1
Atlanta	34,886	1.1
Boise	32,512	1.1
Dallas	34,531	1.0
Providence	35,995	1.0
San Jose	37,799	0.9
Cheyenne	33,876	0.9
Fargo	30,513	0.9
New Orleans	30,174	0.9
Oklahoma City	30,060	0.8
Houston	37,153	0.8
Portland	34,918	0.7
Phoenix	33,291	0.7
Augusta	31,504	0.7
Albuquerque	29,199	0.6
Greensboro	33,072	0.6
Columbia	30,859	0.5
Nashville	32,566	0.5
Buffalo	34,296	0.5

(a) Enter the data in STATA, use STATA to develop the estimated regression equation that could be used to predict the percent audited given the average adjusted gross income reported. Label your answers.

(b) At the 0.05 level of significance, determine whether the adjusted gross income and the percent audited are linearly related.

(c) Did the estimated regression equation provide a good fit? Explain.

(d) Use the estimated regression equation developed in part (a) to calculate a 95% confidence interval for the expected percent audited for districts with an average adjusted gross income of $35,000.

Project 2:

How much does education affect wage rates? The data file *cps4_small.dta* contains 1000 observations on hourly wage rates, education, and other variables from the 2008 Current Population Survey (CPS).

(a) Obtain the summary statistics and histograms for the variables *WAGE* and *EDUC*. Discuss the data characteristics, especially appropriate summary statistics (mean, median) and shape of the data.

(b) Estimate the linear regression $WAGE = \beta_1 + \beta_2 EDUC + \epsilon$ and discuss the results. Interpret the regression coefficients and comment on the quality of the fit.

(c) Calculate the least squares residuals and plot them against *EDUC*. Are any patterns evident? Comment on the assumptions (linearity, normality of errors, and constant variance of errors) of classical linear regression model.

(d) Estimate the quadratic regression $WAGE = \alpha_1 + \alpha_2 EDUC^2 + e$ and discuss the results.

(e) Plot the fitted linear model from part (b) and the fitted values from the quadratic model from part (d) either in the same or separate graph with the data on *WAGE and EDUC.* Which model appears to fit the data better? Compare the sum of squared residuals (*RSS*) from the models in (b) and (d). Which model has a lower *RSS?*

Project 3:

Suppose that you've been asked by the San Diego Padres baseball team to evaluate the economic impact of their new stadium by analyzing the team's attendance per game in the last year at their old stadium. After some research on the topic, you build the following model (standard errors in parentheses):

$$\widehat{ATT_i} = 25000 + \underset{(15000)}{15000\,WIN_i} + \underset{(2000)}{4000\,FREE_i} - \underset{(3000)}{3000\,DAY_i} - \underset{(3000)}{12000\,WEEK_i}$$

$$N = 35 \qquad \bar{R}^2 = .41$$

where:

- ATT_i = the attendance at the *i*th game
- WIN_i = the winning percentage of the opponent in the *i*th game
- $FREE_i$ = a dummy variable equal to 1 if the *i*th game was a "promotion" game at which something was given free to each fan, 0 otherwise
- DAY_i = a dummy variable equal to 1 if the *i*th game was a day game and equal to 0 if the game was a night or twilight game
- $WEEK_i$ = a dummy variable equal to 1 if the *i*th game was during the week and equal to 0 if it was on the weekend

a. You expect the variables WIN and FREE to have positive coefficients. Create and test the appropriate hypotheses to evaluate these expectations at the 5-percent level.

b. You expect WEEK to have a negative coefficient. Create and test the appropriate hypotheses to evaluate these expectations at the 1-percent level.

c. You've included the day game variable because your boss thinks it's important, but you're not sure about the impact of day games on attendance. Run a two-sided *t*-test around zero to test these expectations at the 5-percent level.

Project 4:

Consider the following model that relates the proportion of a household's budget spent on alcohol *WALC* to total expenditure *TOTEXP*, age of the household head *AGE*, and the number of children in the household *NK*.

$$WALC = \beta_1 + \beta_2 ln(TOTEXP) + \beta_3 AGE + \beta_4 NK + \epsilon$$

Note that only households with one or two children are being considered. Thus, *NK* takes only the values one or two. Output from estimating this equation appears in the following table.

Dependent Variable: *WALC* Included Observations: 1519				
Variable	**Coefficient**	**Std. Error (s_b)**	***t*-Statistic**	**Prob. (p-value)**
C	0.0091	0.0190	?	?
Ln(TOTEXP)	0.0276	?	6.6086	?
AGE	?	0.0002	-6.9624	?
NK	-0.0133	0.0033	-4.0750	?
R-squared (R^2)	?		**Mean** Dependent Var	0.0606
S.E. of regression ($s_\in$)	?		**S.D.** Dependent Var	0.0633
Sum of squared residual (*RSS*)	5.752896			

(a) Fill in the blank spaces (?) that appear in the table.
   (i) The *t*-statistic for b_1
   (ii) The standard error for b_2
   (iii) The estimate b_3
   (iv) R^2
   (v) $\hat{\sigma}$ (Standard Error of Regression Line or Estimate)
   (vi) All *p*-values

(b) Test the overall significance of the model.

(c) Interpret the coefficients b_3 and b_4.

(d) Compute a 95% confidence interval estimate for β_3. What does this interval tell you?

(e) Test the hypothesis that the budget proportion for alcohol does not depend on the number of children in the household.

Project 5:

A real estate agent wishes to determine the selling price of residences using the size (square feet) and whether the residence is a condominium or a single-family home. A sample of 20 residences was obtained with the following results:

Price ($)	Type	Square Feet	Price ($)	Type	Square Feet
199,700	Family	1,500	200,600	Condo	1,375
211,800	Condo	2,085	208,000	Condo	1,825
197,100	Family	1,450	210,500	Family	1,650
228,400	Family	1,836	233,300	Family	1,960
215,800	Family	1,730	187,200	Condo	1,360
190,900	Condo	1,726	185,200	Condo	1,200
312,200	Family	2,300	284,100	Family	2,000
313,600	Condo	1,650	207,200	Family	1,755
239,000	Family	1,950	258,200	Family	1,850
184,400	Condo	1,545	203,100	Family	1,630

(a) Produce a regression equation to predict the selling price for residences using a model of the following form:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where:

x_1=Square footage and $x_2 = \begin{cases} 1 & \text{if a condo} \\ 0 & \text{if a single-family home} \end{cases}$

(b) Interpret the estimated parameters $\widehat{\beta_1}$ and $\widehat{\beta_2}$ in the model give in part (a).

(c) Produce an equation that describes the relationship between the selling price and the square footage of (1) condominium and (2) single-family homes.

(d) Conduct a test of hypothesis to determine if the relationship between the selling price and the square footage is different between the condominiums and single-family homes.

(e) What change would you expect in the estimated regression model if you recode x_2 as the following? Rerun the regression model with the newly coded x_2 and verify your expectation.

$$x_2 = \begin{cases} 1 & \text{if a single-family home} \\ 0 & \text{if a condo} \end{cases}$$