

ECN 102: Analysis of Economics Data

Homework 3

Remy Beauregard

Submission rules: homework must be submitted as *one pdf*; please do not submit multiple files. All submissions should include all Stata code and output produced, any figures generated, and answers to questions. For written questions, work must be shown for answers. Code need not be commented.

All datasets can be downloaded in Stata format from the [AED data page](#).

Question 1: Bivariate model assumptions

- a) What is the first population assumption of bivariate OLS regression? How might this assumption be violated? How could we account for this violation?
- b) What is the second population assumption of bivariate OLS regression? How might this assumption be violated? How could we account for this violation?
- c) What is the third population assumption of bivariate OLS regression? How might this assumption be violated? How could we account for this violation?
- d) What is the fourth population assumption of bivariate OLS regression? How might this assumption be violated? How could we account for this violation?

Question 2: Manual calculations

Consider a dataset with 3 observations for $(x, y) : (2, 4), (1, 3)$ and $(6, 2)$.

From first principles (using formulas **not Stata**), compute:

- a) the OLS intercept
- b) the OLS slope coefficient
- c) residuals for the three observations
- d) the R^2 of `reg y x` (or `reg x y`)

- e) the covariance of X and Y
- f) the correlation between X and Y

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \times \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x \times s_y}$$

$$\hat{y} = b_1 + b_2 x; \quad b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b_1 = \bar{y} - b_2 \bar{x}; \quad b_2 = r_{xy} \frac{s_y}{s_x}$$

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2; \quad ResSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2; \quad ExpSS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$R^2 = \frac{ExpSS}{TSS} = 1 - \frac{ResSS}{TSS}$$

Question 3: Bivariate regression

Use AED_SALARYSAT.DTA. This contains data on salary, test scores, and other demographic characteristics.

- a) What variables are present in our dataset? List them along with their labels.
- b) Suppose we want to run a regression of salary on SAT verbal score. Perform this regression and interpret the output.
- c) Generate the residuals from the above regression using `predict resid`, `resid` and plot these against SAT verbal scores in a scatter plot. Do our errors appear to be homoskedastic or heteroskedastic?
- d) Make changes to the regression in part (b) to account for your answer in part (c). How does this change your estimated coefficients? How does this change your estimated standard errors? Does this make sense?

Question 4: Bivariate regression

Using AED_CALELECTRICITY.DTA from hw2, we now want to investigate when electricity gets more and less expensive.

- a) Create a scatterplot of the spot price of electricity and the month. What do you notice?
- b) Create a scatterplot of the spot price and the day *only for the month of November*. What do you notice?
- c) Overlay a best fit line onto your scatterplot from part (b). Without calculations or running a regression, what would you expect b_1 and b_2 to be for a regression of day on spot price in the month of November?
- d) Between *day* and *month*, which series is more correlated with the spot price?
- e) Perform the regression from part (b) of the spot price on day during November. Interpret your results.
- f) Suppose we are concerned our data violates the third assumption of bivariate linear regression. What could we add to part (e) to address this? Make this change and interpret your new results.
- g) What would we expect the spot price of electricity to be on the first day of November? On the last?
- h) Find a day of the month in November with a positive residual. Find a day of the month in November with a negative residual.
- i) Interpret the t-stat, p-value, and confidence interval for b_2 in your regression from part (f). What test does this correspond to? What are its hypotheses? What is its conclusion and why?