The coefficients of an OLS regression of Y_i on a constant and X_i when X_i is binary.

Assume that you have a sample with n units. Let Y_i denote the value of the variable y for unit i, and let X_i denote the value of the variable x for unit i. When you write "ls y c x", E-views computes $\widehat{\beta}_0$ and $\widehat{\beta}_1$, the coefficients of the constant and of X_i in the OLS regression of Y_i on a constant and X_i. It follows from a result you saw during the lectures that

$$\begin{aligned}\widehat{\beta}_1 = &\quad \frac{\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})(X_i-\overline{X})}{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2}\\ \widehat{\beta}_0 = &\quad \overline{Y}-\widehat{\beta}_1\overline{X},\end{aligned}$$

where $\overline{Y}$ denotes the average of the Y_is, while $\overline{X}$ denotes the average of the X_is.

Assume that X_i is a binary variable that is either equal to 0 or to 1. Let n_1 be the number of units with $X_i = 1$, and let $n_0 = n - n_1$ be the number of units with $X_i = 0$. The difference between the average of the Y_is among units with $X_i = 1$ and with $X_i = 0$ is $\frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i$. The goal of the exercise is to show that

$$\begin{aligned}\frac{\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})(X_i-\overline{X})}{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2} = &\quad \frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i\\ \overline{Y}-\widehat{\beta}_1\overline{X} = &\quad \frac{1}{n_0}\sum_{i:X_i=0} Y_i.\end{aligned}$$

Watch out, these results are only true when X_i is binary.

1) First, let's consider the denominator of $\frac{\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})(X_i-\overline{X})}{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2}$. Show that $\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2 = \overline{X}(1-\overline{X})$. Hint: remember that X_i is a binary variable.

2) Now, let's consider the numerator of $\frac{\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})(X_i-\overline{X})}{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2}$.

a) Show that $\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})(X_i-\overline{X}) = \frac{1}{n}\sum_{i=1}^n Y_iX_i - \overline{Y}\ \overline{X}$.

b) Show that $\frac{1}{n}\sum_{i=1}^n Y_iX_i = \overline{X}\frac{1}{n_1}\sum_{i:X_i=1} Y_i$.

c) Use questions a) and b) to show that $\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})(X_i-\overline{X}) = \overline{X}\left(\frac{1}{n_1}\sum_{i:X_i=1} Y_i - \overline{Y}\right)$.

d) Show that $\overline{Y} = \overline{X}\frac{1}{n_1}\sum_{i:X_i=1} Y_i + (1-\overline{X})\frac{1}{n_0}\sum_{i:X_i=0} Y_i$.

e) Use questions c) and d) to show that

$$\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})(X_i-\overline{X}) = \overline{X}(1-\overline{X})\left(\frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i\right).$$

3) Combine the results of questions 1) and 2) e) to show that

$$\widehat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})(X_i-\overline{X})}{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2} = \frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i.$$

4) Finally, use the result of question 3) to show that

$$\widehat{\beta}_0 = \overline{Y}-\widehat{\beta}_1\overline{X} = \frac{1}{n_0}\sum_{i:X_i=0} Y_i.$$

Conclusion of the exercise. This exercise shows that $\widehat{\beta}_1$, the coefficient of X_i in OLS regression of Y_i on a constant and X_i measures the difference between the average of the Y_is among units with $X_i = 1$ and with $X_i = 0$. Putting it in other words, $\widehat{\beta}_1$ measures the difference between the average Y_i of units whose X_i differs by one unit. $\widehat{\beta}_0$ measures the average of Y_i among units with $X_i = 0$.

Suppose that you estimate the following model:

$$\log income_i = \beta_0 + \beta_1 educ_i + \beta_2 male_i + \beta_3 \log IQ_i + e_i$$

where $educ_i$ is an individual's number of years of education, $male_i$ is a binary variable that is equal to 1 for males, and IQ_i is an individuals IQ. You find that $\hat{\beta}_1 = 1,535$, $\hat{\beta}_2 = 3,156$, $\hat{\beta}_3 = 3.8$.

1. Interpret each of these estimates.

2. Can we interpret β_1 as the causal effect of education on an individual's income? Why or why not?

The Center for Disease Control and Prevention (CDC) has hired you to examine the childhood obesity epidemic in the United States. One commonly proposed solution to reduce childhood obesity is encouraging exercise. Your assignment is to determine the causal effect of exercise on body mass index (bmi), which is a commonly used measure to determine obesity. A low bmi is considered healthy, while higher bmi's are considered unhealthy.

To determine the causal effect of exercise on bmi, you run the following regression for a random sample of U.S. children:

$$bmi_i = \beta_0 + \beta_1 exercise_i + u_i$$

where bmi_i represents child i's body mass index, and $exercise_i$ is the number of hours of exercise per week for child i. You obtain the following results:

Dependent Variable: bmi
Method: Least Squares
Date: 10/15/18 Time: 11:39
Sample: 1 10,135
Included observations: 10,135

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	33.5	0.55	60.91	0.0000
exercise	−0.15	0.025	−6	0.0000

You are worried about endogeneity. In particular, you think that children who exercise more are more likely to have a healthier diet compared with children that don't exercise as much, and that a healthy diet is an important factor for reducing bmi.

Given this worry, do you expect the estimated coefficient above to have a positive or negative bias? Explain your answer in words.

Suppose you are interested in the effect of hiring a private economics tutor on a student's Econ 10A final exam score. You collect a random sample (sample size $n = 321$) of Econ 10A students and gather information on whether they had a private tutor and their 10A final exam score. Let D_i be a binary variable equal to 1 if the student had a private tutor and 0 if they did not. Let Y_i denote the student's final exam score. Suppose that you estimate the following regression:

$$Y_i = \beta_0 + \beta_{1i} D_i + e_i$$

1. We are interested in the Average Treatement Effect on the Treated (ATT) of hiring a private tutor. With this binary treatment, **show that the coefficient on D_i is equal to the following**:

$$\hat{\beta}_1 = ATT + \frac{1}{n_1} \sum_{i:D_i=1} y_i(0) - \frac{1}{n - n_1} \sum_{i:D_i=0} y_i(0)$$

   Explain in words what the following terms represent:

   (a) $\frac{1}{n_1} \sum_{i:D_i=1} y_i(0)$

   (b) $\frac{1}{n-n_1} \sum_{i:D_i=0} y_i(0)$

   Using your descriptions of these terms and the above formula, explain why $\hat{\beta}_1$ may be a biased estimate of the Average Treatment Effect on the Treated. Do you think the estimate is biased upward or downward? Support your answer with a **specific** story or characteristic that describes the source of the bias.

2. (16 points) Under binary treatment, the Average Treatment Effect on the **Untreated** (ATU) is defined as follows:

$$ATU = \frac{1}{n - n_1} \sum_{i:D_i=0} y_i(1) - \frac{1}{n - n_1} \sum_{i:D_i=0} y_i(0)$$

If we use $\hat{\beta}_1$ from a regression of Y_i on a constant, and D_i, show that

$$\hat{\beta}_1 = ATU + \frac{1}{n_1} \sum_{i:D_i=1} y_i(1) - \frac{1}{n - n_1} \sum_{i:D_i=0} y_i(1)$$

and describe in words what is required for $\hat{\beta}_1$ to be an unbiased estimator of the ATU.