1. Suppose that a researcher uses data of class size (CS) and average test scores from 100 third-grade classes to estimate the OLS regression

\[ \hat{TestScore} = 520.4 - 5.82 \times CS, \quad R^2 = 0.08, \quad SER = 11.5 \]

(a) A classroom has 22 students. What is the regression’s prediction for that classroom’s average test score?

(b) Last year a classroom had 19 students and this year it has 23 students. What is the regression prediction for that change in the classroom average test score?

(c) The sample average class size across the 100 classrooms is 21.4. What is the sample average of the test score across the 100 classrooms? (Hint: Review the formulas for the OLS estimators.)

(d) What is the sample standard deviation of test scores across the 100 classrooms? (Hint: Review the formula for the \( R^2 \) and \( SER \).)

2. A regression of average weekly earnings (AWE, measured in dollars) on age (measured in years) uses a random sample of college-educated full-time workers aged 25-65 yields the following:

\[ \hat{AWE} = 696.7 + 9.6 \times Age, \quad R^2 = 0.023, \quad SER = 624.1 \]

(a) Explain what the coefficient values 696.7 and 9.6 mean.

(b) What are the units of measurement for the \( SER \). dollars? years? or is \( SER \) unit-free?

(c) What are the units of measurement for the \( R^2 \) dollars? years? or is \( R^2 \) unit-free?

(d) What are the regression’s predicted earnings for a 25-year-old worker? And for a 45-year-old worker?

(e) Will the regression give reliable predictions for a 99-year-old worker? Why or why not?

(f) Given what you know about the distribution of earnings, do you think it is plausible that the distribution of errors in the regression is normal? (Hints: think about the symmetry of the distribution and the smallest value of earnings, then discuss if these attributes are consistent with a normal distribution).

(g) The average age in this sample is 41.6 years. What is the average value of \( AWE \) in the sample?

3. A professor decides to run an experiment to measure the effect of time pressure on final exam scores. He gives each of the 400 students in his course the same final exam, but some students have 90 minutes to complete the exam while others have 120 minutes. Each student is randomly assigned one of the examination times based on the flip of a coin. Let \( Y_i \) denote the number of points scored on the exam by the \( i^{th} \) student (0 ≤ \( Y_i \) ≤ 100). Let \( X_i \) denote the amount of time that the student has to complete the exam (\( X_i = 90 \) or \( X_i = 120 \)). Consider the regression model \( Y_i = \beta_0 + \beta_1 X_i + U_i \).

(a) Explain what the term \( U_i \) represents. Why will different students have different values of \( U_i \)?

(b) Explain why \( E(U_i | X_i) = 0 \) for this regression model.

(c) The estimated regression is \( \hat{Y}_i = 49 + 0.24X_i \).

1. Compute the estimated regression’s prediction for the average score of students given 90 minutes to complete the exam.

2. Compute the estimated gain in score for a student who is given an additional 10 minutes on the exam.


you will find a data file CollegeDistance that contains data from a random sample of high school seniors interviewed in 1980 and re-interviewed in 1986. In this exercise you will use these data to investigate the relationship between the number of completed years of education for young adults and the distance from each student’s high school to the nearest four-year college. (Proximity to college lowers the cost of education, so that students who live closer to a four-year college should, on average, complete more years of higher education.) A detailed description is given in CollegeDistance_Description.is available in the Web site.

(a) Run a regression of years of completed education (ED) on distance to the nearest college (Dist), where Dist is measured in tens of miles. (For example, Dist = 2 means that the distance is 20 miles.) What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How does the average value of years of completed schooling change when colleges are built close to where students go to high school?

(b) Bob’s high school was 20 miles from the nearest college. Predict Bob’s years of completed education using the estimated regression. How would the prediction change if Bob lived 10 miles from the nearest college?

(c) Does distance to college explain a large fraction of the variance in educational attainment across individuals? Explain.

(d) What is the value of the standard error of the regression? What are the units for the standard error (meters, grams, years, dollars,...)?


you will find a data file TeachingRatings that contains data on course evaluation, course characteristics and professor characteristics for 463 courses at the University of Texas Austin. A detailed description is given in TeachingRatings_Description, also available on the Web site. One of the characteristics is an index of the professor’s beauty (Beauty) as rated by a panel of six judges. In this exercise you will investigate how course evaluations are related to the professor’s beauty.

(a) Construct a scatter plot of average course evaluations (Course_Eval) on the professor’s beauty (Beauty). Does there appear to be a relationship between the variables?

(b) Run a regression of average course evaluations (Course_Eval) on the professor’s beauty (Beauty). What is the estimated intercept? What is the estimated slope? Explain why the estimated intercept is equal to the sample mean of Course_Eval (Hint: What is the sample mean of Beauty?)

(c) Professor A has an average value of Beauty, while Professor B has a value of Beauty of one standard deviation above the average. Predict Professor A and Professor B course evaluations.

(d) Comment on the size of the regression’s slope. Is the estimated effect of Beauty on Course_Eval large or small? Explain what you mean by "large" and "small."

(e) Does Beauty explain a large fraction of the variance in evaluations across courses? Explain.

6. Show that the least squares estimator of the slope and constant of a simple regression model are unbiased under the classical assumptions.

(a) A linear regression gets $\hat{\beta}_1 = 0$. Show that $R^2 = 0$.

(b) A linear regression gets that $R^2 = 0$. Does it imply that $\hat{\beta}_1 = 0$?

7. Suppose that $Y_i = \beta_0 + \beta_1 X_i + U_i$, where $\{X_i, U_i\}_{i=1}^n$ are iid and $X_i$ is distributed as a random variable Bernoulli with $Pr (X_i = 1) = 0.2$. We know that $U_i = Z_i (2 - X_i)$ where $Z_i$ is iid standard normal independently distributed of $\{X_i\}_{i=1}^n$.

(a) Check if the following statements hold: $E (U_i | X_i) = 0$ and $Var (U_i | X_i) = \sigma^2$ (constant).

(b) Obtain an expression for large samples of the variance of $\hat{\beta}_1$.

8. In the linear regression model $Y_i = \beta_0 + \beta_1 X_i + U_i$, $i = 1, \ldots, n$,
(a) Show that the $R^2$ of the regression of $Y$ on $X$ is the squared value of the sample correlation between $X$ and $Y$.

(b) Show that the $R^2$ of the regression of $Y$ on $X$ is the same as the $R^2$ of $X$ on $Y$.

(c) Show that $\hat{\beta}_1 = r_{xy} (s_y/s_x)$, where $r_{xy}$ is the sample correlation between $X$ and $Y$, and $s_y$ and $s_x$ are the sample standard deviations of $Y$ and $X$, respectively.

(d) Prove that the regression sample line passes through the point $(\bar{Y}, \bar{X})$.

**ANSWERS:**

1. a) 392.36; b) 23.28; c) 395.85; d) 11.93.

2. b) Dollars per week; c) It is unit-free; d) i) $936.7$, ii) $1,128.7$; e) No. The oldest worker in the sample is 65 years old.; f) No. The distribution of earning is positively skewed and has kurtosis larger than the normal; g) $1,096.06$.

3. b) Because of random assignment $U_i$ and $X_i$ are independent. Thus, $E(U_i|X_i) = E(U_i) = 0$; c) i) 70.06, ii) 2.4.

4. a) $\hat{E}d = 13.96 - 0.073 \times Dist$. The regression predicts that if colleges are built 10 miles closer to high schools, years of completed education increases by 0.073 years; b) 13.81; c) $R^2 = 0.007$, thus, the distance explains only a very small percentage of the years; d) 1.8 years.

5. a) It seems there’s no relation; b) $\text{Course\_Eval} = 4.00 + 0.133 \times \text{Beauty}$. The sample mean of the variable $\text{Beauty}$ is 0; the intercept is the mean of the dependent variable ($\text{Course\_Eval}$) minus the estimated slope (0.133) times the mean of the regressor ($\text{Beauty}$). Therefore, the estimator of the constant is equal to the mean of $\text{Course\_Eval}$; c) A 4.00 and B 4.105; d) The standard deviation of $\text{Course\_Eval}$ is 0.55 and the standard deviation of $\text{Beauty}$ is 0.789. It is expected that an increase in the standard deviation of $\text{Beauty}$ increases $\text{Course\_Eval}$ in $0.133 \times 0.789 = 0.105$, or $1/5$ of the standard deviation $\text{Course\_Eval}$. The effect is small. e) $R^2 = 0.036$, then, $\text{Beauty}$ explains only 3.6% of the variance of $\text{Course\_Eval}$.

7. a) Yes and no.