Econ 103: Homework 1

Manu Navjeevan

September 30, 2021

Econ 41 Review

1. Discrete Random Variables. Suppose that we are interested in the number of cups of coffee drank by a (randomly selected) student at UCLA. This quantity can be represented as a random variable Y with probability mass function:

$$p_Y(a) = \begin{cases} \frac{1}{4} & \text{if } a \in \{0, 1, 2\} \\ \frac{1}{8} & \text{if } a = 3 \\ \frac{3}{32} & \text{if } a = 4 \\ c & \text{if } a = 5 \\ 0 & \text{otherwise} \end{cases},$$

where c is an unknown constant.

   (a) Explain why the number of cups of coffee drank in a day by a randomly selected student at UCLA is a random variable.

   (b) What is the relevant outcome space of the random variable Y?

   (c) Explain what the distribution of this random variable represents. In other words distribution of Y assigns a probability to any subset of the outcome space. How do we interpret this probability?

   (d) Solve for c. (*Hint:* Recall that $\mathbb{P}_Y(\mathcal{O}_Y) = 1$ so that $\sum_{a \in \mathcal{O}_Y} p_Y(a)$ must equal one).

   (e) What is the probability that a randomly selected student at UCLA drinks at least 3 cups of coffee a day, $\mathbb{P}_Y(Y \geq 3)$?

   (f) What is the expected number of cups of coffee drank per day for a randomly selected student at UCLA?

2. Continuous Random Variables. Suppose that we are interested in the income of a randomly selected Angeleno. The distribution of incomes (in tens of thousands of dollars) for residents of Los Angeles can be described as a random variable, X, with the following pdf.

$$f_X(a) = \begin{cases} 0.11 - ca & \text{if } 0 \leq a \leq 10 \\ 0 & \text{otherwise} \end{cases},$$

where c is an unkown constant.

   (a) What is the outcome space of X, $\mathcal{O}_X$?

   (b) Using the relationship

$$\mathbb{P}_X(l \leq X \leq m) = \int_l^m f_X(a)\, da,$$

   explain why the pdf must always be weakly positive, $f_X(a) \geq 0$, for any $a \in \mathbb{R}$.

   (c) Because $\mathbb{P}_X(\mathcal{O}_X) = 1$ we must have that $\int_0^{10} f_X(a)\, da = 1$. Using this fact, solve for c.

(d) What is the expected value of X, $\mathbb{E}[X]$?

(e) What is the variance of X, $\text{Var}(X)$?

3. Variance and Covariance. Let Y be a random variable representing income (in tens of thousands of dollars) and X be a random variable representing years of education. Suppose that the marginal distribution of X is described by its probability mass function

$$p_X(x) = \begin{cases} 0.05 & \text{if } x \in \{1, 2, \dots, 12\} \\ 0.09 & \text{if } x \in \{13, 14, 15, 16\} \\ 0.04 & \text{if } x \in \{17\} \\ 0 & \text{otherwise} \end{cases}.$$

The marginal distribution of Y is described by its probability density function

$$f_Y(y) = \begin{cases} 0.1 & \text{if } 0 \leq y \leq 10 \\ 0 & \text{otherwise} \end{cases}.$$

(a) What is the expectation of Y, $\mathbb{E}[Y]$? What is its variance, $\text{Var}(Y)$?

(b) What is the expectation of X, $\mathbb{E}[X]$? What is its variance, $\text{Var}(X)$?

(c) Using $\mathbb{E}[YX] = 60$ compute the covariance between Y and X, $\text{Cov}(X, Y)$.

(d) Calculate the correlation coefficient between X and Y.

$$\rho_{YX} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

(e) What does this covariance tell us about the relationship between education levels and income? Is there a positive or negative association?

(f) Should we interpret this result as a *causal* relationship between education and income? What are some reasons we may want to refrain from this interpretation?

(g) (Challenge) A common inequality used in econometrics is the *Cauchy-Schwarz* inequality. It states that, for any random variables X and Y, and any functions $g(\cdot)$ and $h(\cdot)$,

$$\big|\mathbb{E}[g(X)h(Y)]\big| \leq \sqrt{\mathbb{E}[g^2(X)]}\sqrt{\mathbb{E}[h^2(Y)]}.$$

Use this inequality to show why the correlation coefficient is bounded between negative one and one, $-1 \leq \rho_{XY} \leq 1$. (*Hint*: Try $g(x) = x - \mu_X$ and $h(y) = y - \mu_Y$).

Introduction to Single Linear Regression

1. Useful Equalities. Recall that in deriving the form of $\hat{\beta}_1$ we used the following equalities

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i - \bar{Y}\bar{X} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2.$$

Show either one of these equalities (only have to show one or the other).

2. Assumptions for Inference. Suppose we are interested in the relationship between the size of the average American's social circle, X, and whether or not they are unemployed, Y. To investigate this relationship we want to estimate the following regression equation[1]

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \mathbb{E}[\epsilon] = \mathbb{E}[\epsilon X] = 0.$$

[1] Recall that this regression specification corresponds to finding the line of best fit parameters $\beta_0, \beta_1 = \arg\min_{b_0, b_1} \mathbb{E}[(Y - b_0 - b_1 X)^2]$ and defining $\epsilon = Y - \beta_0 - \beta_1 X$

To estimate the regression coefficient parameters we collect a sample of size n, $\{Y_i, X_i\}_{i=1}^{n}$. Recall that for valid asymptotic inference on our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ we require the following assumptions: Random Sampling, Homoskedasticity, and Rank condition.

- Random Sampling: Assume that $\{Y_, X_i\}$ are independently and identically distributed from the population of interest, $(Y_i, X_i) \overset{\text{i.i.d}}{\sim} (Y, X)$.
- Homoskedasticity: Assume that $\text{Var}(\epsilon | X = x) = \sigma_\epsilon^2$ for all possible values of x.
- Rank Condition: There must be at least two distinct values of X that appear in the population.

(a) Suppose we collect our sample by only randomly surveying people on UCLA campus. Which assumption would be violated?

(b) Suppose we collect our sample and find that everyone appears to have exactly one friend. Which assumption would be violated? Why is this a problem when computing the line of best fit through our sample?

(c) Suppose random sampling, homoskedasticity, and the rank condition are all satisfied, but $n = 10$. Why might inferences based on the approximation

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}/\sqrt{n}} \sim N(0, 1)$$

not be valid?

3. Hypothesis Testing. Suppose now that we are interested in investigating the relationship between the size of someone's social circle, X, and their income (in tens of thousands of dollars), Y. We want to estimate the following linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \mathbb{E}[\epsilon] = \mathbb{E}[\epsilon X] = 0.$$

To do so we collect a random sample of size $n = 64$, $\{Y_i, X_i\}_{i=1}^{64}$ and find that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 100$, $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) = 225$, $\bar{Y} = 5.5$, and $\bar{X} = 1.5$.

(a) Using this information find and interpret $\hat{\beta}_1$ and $\hat{\beta}_0$.

(b) After finding $\hat{\beta}_1$ and $\hat{\beta}_1$ describe how you would construct the estimated residuals $\hat{\epsilon}_i$.

(c) We find that $\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = 36$. Use this and the result that, for n large,

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}/\sqrt{n}} \sim N(0, 1),$$

to compute the (approximate) probability that, if the true value was given $\beta_1 = 0$, we would see a value of $|\hat{\beta}_1|$ equal to or larger than the one that we observed.

(d) Use this result to test, at level $\alpha = 0.1$, the hypotheses

$$H_0 : \beta_1 = 0 \ \text{ vs. } \ H_1 : \beta_1 \neq 0$$

(e) Conduct this test in another fashion by constructing the test statistic t^* and comparing to either $z_{0.95} = 1.64$ or $z_{0.9} = 1.24$ (indicate which value you are comparing the test statistic to).

(f) Construct a 90% confidence interval for β_1. How could we use this to conduct the hypothesis test in part (d)?

(g) Suppose that we find we made an error in our calculation and actually $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 1$. If all other values stayed the same, how would this change the result of the hypothesis test in part (d)?