Econ 103: Homework 1

Manu Navjeevan

August 23, 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.

Answer: Not everyone at UCLA drinks the same number of cups of coffee. Depending on which UCLA student we select the answer to this question will vary.

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

Answer: We can see from the pmf that Y can only take on one of 6 values, $\{0,1,2,3,4,5\}$. So $\mathcal{O}_Y = \{1,2,3,4,5\}$.

(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?

Answer: We can interpret $\mathbb{P}_Y(A)$ as the proportion of UCLA students whose daily coffee consumption lies in the set A. For example, $\mathbb{P}_Y(\{0,1\})$ represents the share of UCLA students that drink either zero or one cup of coffee a day.

(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).

Answer: Because $\mathbb{P}_Y(\mathcal{O}_Y) = 1$ and $\mathbb{P}_Y(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}_Y(A_i)$ if all the $A_1, \ldots, A_n$ are pairwise disjoint we can decompose

$$\begin{aligned} 1 = \mathbb{P}_Y(\mathcal{O}_Y) &= \mathbb{P}_Y(\cup_{a\in\mathcal{O}_Y}\{a\}) \\ &= \sum_{a\in\mathcal{O}_Y} \mathbb{P}_Y(\{a\}) \\ &= \sum_{a\in\{0,1,2,3,4,5\}} p_Y(a) \\ &= \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{3}{32} + c \\ \implies c &= \frac{1}{32} \end{aligned}$$

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

Answer: Again using the property that $\mathbb{P}_Y(\bigcup_{i=1}^n A_i)$ for pairwise disjoint sets $A_1, A_2, \ldots, A_n$

$$\mathbb{P}_Y(Y \geq 3) = \mathbb{P}_Y(\{3,4,5\}) = \sum_{a \in \{3,4,5\}} \mathbb{P}_Y(\{a\}) = \underbrace{p_Y(3)}_{1/8} + \underbrace{p_Y(4)}_{3/32} + \underbrace{p_Y(5)}_{1/32} = \frac{1}{4}$$

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

Answer: Recall the formula for expected value, $\mathbb{E}[Y] = \sum_{a \in \mathcal{O}_Y} a \cdot p_Y(a)$.

$$\begin{aligned}\mathbb{E}[Y] = \sum_{a \in \{0,1,2,3,4,5\}} a \cdot p_Y(a) &= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{3}{32} + 5 \cdot \frac{1}{32} \\ &= \frac{8}{32} + \frac{16}{32} + \frac{12}{32} + \frac{12}{32} + \frac{5}{32} \\ &= \frac{53}{32} \approx 1.656\end{aligned}$$

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 unknown constant.

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

Answer: We can see that pdf is only non-zero on the interval $[0, 10]$ so $\mathcal{O}_X = [0, 10]$. Strictly speaking, we could define the outcome space to be any set that contains $[0, 10]$, it doesn't violate any assumption, so other answers may be accepted.

(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 $a \in \mathbb{R}$.

Answer: Suppose that the pdf $f_X(a)$ was strictly negative on some interval $[l, m]$, ($f_X(a) < 0$ for $l \leq a \leq m$). Then, using the relationship between probabilities and the pdf, we would find that $\mathbb{P}_X(l \leq X \leq m) < 0$, which violates the assumption that $0 \leq \mathbb{P}_X(A) \leq 1$ for any subset A.

(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.

Answer: The pdf must integrate to one so

$$1 = \int_0^{10} 0.11 - ca\, da = 0.11 \cdot 10 - c\frac{a^2}{2}\bigg|_0^{10} = 1.1 - 50c \implies c = 0.002$$

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

Answer: Recall that for a continuous random variable $\mathbb{E}[X] = \int_{\mathcal{O}_X} a \cdot f_X(a)\, da$. Using this

$$\begin{aligned}\mathbb{E}[X] = \int_0^{10} a(0.11 - 0.002a)\, da = \int_0^{10} 0.11a - 0.002a^2\, da \\ = 0.11\frac{a^2}{2}\Big|_0^{10} - 0.002\frac{a^3}{3}\Big|_0^{10} \\ = 0.11 \cdot 50 - 0.002\frac{1000}{3} \approx 4.8333\end{aligned}$$

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

Answer: Recall that $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. For a continuous random variable $\mathbb{E}[X^2] = \int_{\mathcal{O}_X} a^2 \cdot f_X(a)\, da$:

$$\begin{aligned}\mathbb{E}[X^2] = \int_0^{10} a^2(0.11 - 0.002a)\, da = \int_0^{10} 0.11 \cdot a^2 - 0.002a^3\, da \\ = 0.11\frac{a^3}{3}\Big|_0^{10} - 0.002\frac{a^4}{4}\Big|_0^{10} \\ = 0.11\frac{1000}{3} - 5 \approx 31.6666\end{aligned}$$

Using this along with the solution to part (d) we can compute

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \approx 8.3056.$$

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

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

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

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

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

Answer: Using the formulas from above we can calculate

$$\mathbb{E}[Y] = \int_0^{10} 0.1a\, da = 5 \quad \text{and} \quad \text{Var}(Y) = \int_0^{10} (a-5)^2 \cdot 0.1\, da = 100/12$$

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

Answer: Using the formulas from above we calculate:

$$\begin{aligned} \mathbb{E}[X] &= 0.05(1+2+\cdots+12)+0.09(13+14+15+16)+0.004\cdot 17 = 9.8 \\ \mathbb{E}[X^2] &= 0.05(1^2+2^2+\cdots+12^2)+0.09(13^2+14^2+15^2+16^2)+0.004\cdot 17^2 = 120.2 \\ \implies \operatorname{Var}(X) &= \mathbb{E}[X^2]-(\mathbb{E}[X])^2 = 24.16 \end{aligned}$$

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

Answer: From lecture we showed that $\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. From part (a) we have that $\mathbb{E}[Y] = 5$ and from part (b) we have that $\mathbb{E}[X] = 9.8$. Using this we find that

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 60 - 9.8\cdot 5 = 11.$$

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

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

Answer: Using the above, we find that $\sigma_X = \sqrt{24.16}$ and $\sigma_Y = \sqrt{100/12}$. Since $\operatorname{Cov}(X,Y) = 11$ we can calculate

$$\rho_{XY} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{11}{\sqrt{24.16}\sqrt{100/12}} \approx 0.775.$$

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

Answer: Since the covariance is positive we can see that there is a positive association between education levels and income. In general, people with higher education levels may earn more than people with lower education levels.

(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?

Answer: We cannot necessarily interpret this as a causal relationship, it is purely associative. Some hidden factors (often referred to as confounding variables) that may affect both education and income may be technical skill and parental income level. Students that have high levels of education may have a high ability at some skill and so may have obtained high earnings otherwise. People who attend college also generally have parents that are better off than there non-college attending counterparts. This could allow them to get referrals to high paying jobs more easily even in the absence of a college degree.

(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 $h(\cdot)$,

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

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

Answer: Use $g(x) = x - \mu_X$ and $h(x) = y - \mu_Y$ so that we can write $\operatorname{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_X)] = \mathbb{E}[g(X)h(Y)]$. Then applying Cauchy-Schwarz

$$\big|\operatorname{Cov}(X,Y)\big| \le \sqrt{\mathbb{E}[g^2(X)]}\sqrt{\mathbb{E}[h^2(X)} = \underbrace{\sqrt{\mathbb{E}[(X-\mu_X)^2]}}_{\sigma_X}\underbrace{\sqrt{\mathbb{E}[(Y-\mu_Y)^2]}}_{\sigma_Y}.$$

Since $|\operatorname{Cov}(X,Y)| \le \sigma_X\sigma_Y$, $|\rho_{XY}| \le 1$.

Introduction to Single Linear Regression

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

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

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

Answer: Showing these in order:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X}) &= \frac{1}{n}\sum_{i=1}^{n}Y_iX_i - \bar{X}\frac{1}{n}\sum_{i=1}^{n}Y_i - \bar{Y}\frac{1}{n}\sum_{i=1}^{n}X_i + \bar{Y}\bar{X}\\ &= \frac{1}{n}\sum_{i=1}^{n}Y_iX_i - \bar{X}\bar{Y} - \bar{Y}\bar{X} + \bar{Y}\bar{X}\\ &= \frac{1}{n}\sum_{i=1}^{n}Y_iX_i - \bar{X}\bar{Y}\end{aligned}$$

The second one follows basically the same steps:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X} + (\bar{X})^2\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\bar{X}\frac{1}{n}\sum_{i=1}^{n}X_i + (\bar{X})^2\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2(\bar{X})^2 + (\bar{X})^2\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - (\bar{X})^2\end{aligned}$$

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]

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

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?

Answer: In this case the random sampling assumption would be violated. By only sampling people at UCLA campus we are not drawing from the population of interest, which is the population

[1] Recall that this regression specification corresponds to finding the line of best fit parameters $\beta_0, \beta_1 = \arg\min_{b_0,b_1}\mathbb{E}[(Y - b_0 - b_1X)^2]$ and defining $\epsilon = Y - \beta_0 - \beta_1 X$

of all (adult) Americans. People on UCLA campus are probably not representative of the entire country.

(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?

Answer: In this condition we are violating the rank condition. We need at least two x values to make a well defined line, otherwise the slope would be infinite.

(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?

Answer: This approximation is based on the Central Limit Theorem, which only applies when n is large. For small n the normal distribution may not be a good approximation for the distribution of $\hat{\beta}_1$.

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 $\hat{\beta}_0$.

Answer: Recall from lecture

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} \implies \hat{\beta}_1 = \frac{225}{100} = 2.25$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \implies \hat{\beta}_0 = 5.5 - 2.25 \cdot 1.5 = 2.125$$

Interpreting $\hat{\beta}_0$ in context we can say that we estimate that the average value of income for someone with no friends to be \$21,250. We interpret $\hat{\beta}_1$ in context by saying that we estimate that having an additional friend is associated with a \$22,500 increase in income.

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

Answer: We would construct the estimated residuals $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ for each (Y_i, X_i) in our sample.

(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.

Answer: Under the assumption that $\beta_1 = 0$, we want to compute $\Pr(|\hat{\beta}_1| \geq 2.25)$. To compute this see that

$$\Pr(|\hat{\beta}_1| \geq 2.25) = \Pr\left(\left|\frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}/\sqrt{n}}\right| \geq \frac{2.25}{\hat{\sigma}_{\beta_1}/\sqrt{n}}\right) \approx \Pr\left(|Z| \geq \frac{2.25}{\hat{\sigma}_{\beta_1}/\sqrt{n}}\right).$$

What remains is to compute $\hat{\sigma}_{\beta_1}$. Recall from lecture that

$$\hat{\sigma}^2_{\beta_1} = \frac{\hat{\sigma}^2_{\epsilon}}{\hat{\sigma}^2_X} = \frac{36}{100} \implies \hat{\sigma}_{\beta_1}/\sqrt{n} = \frac{6}{10}\frac{1}{8} = 0.075.$$

So $\Pr\left(|Z| \geq \frac{2.25}{\hat{\sigma}_{\beta_1}/\sqrt{n}}\right) = \Pr\left(|Z| \geq 30\right) \approx 0$.

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

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

Answer: In part (c) we found that the p-value for this test was very close to zero. Since $0 \leq \alpha = 0.1$ we can reject this null hypothesis in favor of the alternative hypothesis and conclude that there is a relationship between the size of someone's circle and their income.

(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 too).

Answer: We can also conduct this test by constructing the test statistic t^* and comparing its absolute value to $z_{1-\alpha/2} = z_{0.95}$ (since we are running a two-sided test at $\alpha = 0.1$). The test statistic is constructed:

$$t^* = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}/\sqrt{n}} = \frac{2.25}{0.6/8} = 30.$$

Since $|t^*| = 30 \gg 1.64 = z_{0.95}$ we reject the null hypothesis that there is no relationship between the size of one's social circle and income in favor of the alternative hypothesis that there is a relationship between the two.

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

Answer: Recall that a 90% confidence interval for β_1 consists of all the values b for which we would not reject, at level $\alpha = 1 - 0.9 = 0.1$, the null hypothesis $H_0 : \beta_1 = b$ in favor of an alternative $H_1 : \beta_1 \neq b$. We fail to reject this null hypothesis in favor of the two sided alternative if $|(\hat{\beta} - b)/\hat{\sigma}_{\beta_1}/\sqrt{n}| = |t^*| \leq z_{1-\alpha/2} = z_{0.95} \implies b \in [\hat{\beta} - z_{1-\alpha/2}\hat{\sigma}_{\beta_1}\sqrt{n}, \hat{\beta} + z_{1-\alpha/2}\hat{\sigma}_{\beta_1}/\sqrt{n}]$. In part (c) we found that $\hat{\sigma}_{\beta_1}/\sqrt{n} = 0.075$ and we know $z_{0.95} = 1.64$ so that our 90% confidence interval for β_1 is given:

$$2.25 \pm 1.64 \cdot 0.075 = [2.127, 2.373].$$

Since zero is not contained in this interval we can conclude that we would reject the null hypothesis 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)?

Answer: To rerun this test we will need to recompute $\hat{\beta}_1$ and $\hat{\sigma}_{\beta_1}$. Using the formula

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \implies \hat{\beta}_1 = \frac{225}{1} = 225.$$

Now we recompute $\hat{\sigma}_{\beta_1}$ via

$$\hat{\sigma}^2_{\beta_1} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \implies \hat{\sigma}^2_{\beta_1} = 36.$$

The standard error is now given $\hat{\sigma}_{\beta_1}/\sqrt{n} = 6/8 = 0.75$. Our test statistic can now be computed

$$t^* = \frac{225}{0.75} = 300.$$

Clearly, we still reject our null hypothesis.