LINEAR REGRESSION PARAMETERS EXTRA CREDIT

DUE DATE: SEPTEMBER 22, 2023

Recall that a linear regression model of Y on X for the data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is given by the equation

$$y_i = a + bx_i + e_i$$

where the e_i are the residuals for each of the points in the data set.

We will prove that for **the** linear regression of Y and X, that $b = \dfrac{\text{cov}(X, Y)}{\text{var}(X)}$ and $a = \mu_Y - b\mu_X$.

(a) Prove that for any linear regression model of Y on X that we can express each residual as

$$e_i = y_i - a - bx_i$$

(b) Using (a), show that

$$\underbrace{\sum_{i=1}^{n} e_i^2}_{\text{call } E(a,b)} = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

(c) (Calculus required, skip if you'd like) We need to find the constants a and b that minimizes the sum $E(a, b) = \sum_{i=1}^{n} e_i^2$.

(i) Prove that

$$\frac{\partial E(a, b)}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i)$$

(ii) Setting the derivative $= 0$, suppose α and β are the constants for a and b, respectively, that make the derivative 0...this means we have

$$-2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0.$$

Using this equation, prove that

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (\alpha + \beta x_i)$$

(d) Using the equation

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (\alpha + \beta x_i),$$

prove that $\mu_Y = \alpha + \beta\mu_X$ (Hint: start by taking the equation, then divide by n).

(e) Part (d) proves that (μ_X, μ_Y) must lie on **the** regression line. Using the slope-intercept form equation of a line, we get

$$y_i - \mu_Y = b(x - \mu_X) + e_i.$$

Using the sum-of-squares minimization property of **the** regression line, prove that $E(a,b) = \sum_{i=1}^{n} e_i^2$ can be rewritten as

$$E(a,b) = \sum_{i=1}^{n} ((y_i - \mu_Y) - b(x_i - \mu_X))^2.$$

(f) (Calculus required, skip if you'd like) Using the previous part, prove that

$$\frac{\partial E(a,b)}{\partial b} = -2\sum_{i=1}^{n} ((y_i - \mu_Y) - b(x_i - \mu_X))(x_i - \mu_X).$$

(g) Allowing β to be the constant for b that makes the derivative $= 0$, we get the expression

$$-2\sum_{i=1}^{n} ((y_i - \mu_Y) - \beta(x_i - \mu_X))(x_i - \mu_X).$$

Use this to prove that

$$\sum_{i=1}^{n} (y_i - \mu_Y)(x_i - \mu_X) = \beta \sum_{i=1}^{n} (x_i - \mu_X)^2.$$

(h) Using the equation above, prove that $\operatorname{cov}(X,Y) = \beta\sigma_X^2$ and conclude that $\beta = \dfrac{\operatorname{cov}(X,Y)}{\sigma_X^2}$.