Problem Set 1. Due: January 30

Include your code as an Appendix, not inline.

Problem 1 (Brushing up on statistics)

Suppose $X, Y$ and $Z$ are random variables with finite second moment and let $a, b \in \mathbb{R}$. Show:

(a) $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$.

(b) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.

(c) $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$.

(d) Write and simplify $\text{Var}(X - Y)$ in terms of covariances.

Problem 2 (Regression through the origin)

Suppose $(x_1, y_1), \ldots, (x_n, y_n)$ are observation pairs from the model

$$Y_i = \beta x_i + \varepsilon_i$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are iid $N(0, \sigma^2)$.

(a) Find the ordinary least squares estimator for $\beta$, call it $\hat{\beta}$.

(b) Find the expected value of your estimator from (a) – is it biased?

(c) Find the variance of your estimator from (a). Under what conditions is $\hat{\beta}$ a statistically consistent estimator?

(d) How would you interpret $\beta$ in this model? Note what happens when $x = 0$.

Problem 3 (Centering features)

Define $z_i = x_i - \bar{x}$ as the centered features, and consider the regression problem

$$Y_i = \beta_0 + \beta_1 z_i + \varepsilon_i, \quad i = 1, \ldots, n.$$ 

How do the OLS estimators of $\beta_0$ and $\beta_1$ change? Are they still unbiased? Additionally recall that the covariance between the original OLS estimators is

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$ 

How does this change for regressing on $z_i$ instead (hint: you don’t have to re-derive it)? Is there any reason we would want to do this in practice?
Problem 4 (Thinking about statistical assumptions)

Recall the differing sets of assumptions we made for the simple linear model

\[ Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \]  \hspace{1cm} (1)

Choose and fix values of \( \beta_0, \beta_1 \) and \( \sigma^2 \), and simulate 4 different datasets corresponding to each of the separate assumptions below. Include plots of all 4, and use the `abline` function to show the true regression line through each plot. Comment on the visual differences.

For each simulation, satisfy only the assumptions necessary for each A1, A2, A3, A4 but not the previous level – that is, for the A2 dataset, do not use normal errors, for the A3 dataset, do not use identically distributed errors, etc.

Report your choices for \( \beta_0, \beta_1, \sigma^2 \) and the distribution of \( \varepsilon \) for each scenario.

A1 1. The relationship (1) holds
2. \( \varepsilon_i \) are iid \( N(0, \sigma^2) \)

A2 1. The relationship (1) holds
2. \( \mathbb{E}\varepsilon_i = 0 \)
3. \( \text{Var}\varepsilon_i = \sigma^2 \) (homoskedasticity)
4. \( \varepsilon_i \) and \( \varepsilon_j \) are iid

A3 1. The relationship (1) holds
2. \( \mathbb{E}\varepsilon_i = 0 \)
3. \( \text{Var}\varepsilon_i = \sigma^2 \) (homoskedasticity)
4. \( \varepsilon_i \) and \( \varepsilon_j \) are uncorrelated for \( i \neq j \)

A4 1. The relationship (1) holds
2. \( \mathbb{E}\varepsilon_i = 0 \)
3. \( \text{Var}\varepsilon_i < \infty \)