Homework 2

Due Feb 10 at 10 pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages).

1. In this problem, we will work with the standard univariate normal density function:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$

   (a) Show that $f(x)$ is a valid probability density function by verifying that it integrates to 1 over $\mathbb{R}$.

   (b) Let X be a random variable with density $f(x)$. Prove that the expected value $E[X]$ is 0.

   (c) Prove that the variance of X, $\text{Var}(X)$, is equal to 1.

2. Someone proposes a two-stage approach to simple linear regression using a bivariate dataset $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. They first regress Y on X to obtain the estimated regression line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

   Using the newly formed triples $\{(X_1, Y_1, \hat{Y}_1), \ldots, (X_n, Y_n, \hat{Y}_n)\}$, they propose to regress $\hat{Y}$ on X, expecting this second regression to yield even better coefficient estimates. Using the geometric interpretation of regression, how would you explain to this person what they are not understanding about regression.

3. (Properties of covariance) Prove the following properties of covariance (you can use properties established in the chapter).

   For any random variables $\tilde{a}$ and $\tilde{b}$ with finite variance:

   (a) For any α, $\beta \in \mathbb{R}$

   $$\mathrm{Cov}[\beta\tilde{a} + \alpha, \tilde{b}] = \beta \mathrm{Cov}[\tilde{a}, \tilde{b}]. \tag{1}$$

   (b)

   $$\mathrm{Cov}[\tilde{a} + \tilde{b}, \tilde{a} - \tilde{b}] = \mathrm{Var}[\tilde{a}] - \mathrm{Var}[\tilde{b}]. \tag{2}$$

4. (Standardized variables and the sample correlation coefficient) We study the sample correlation coefficient $\rho_{X,Y}$ from Definition 8.10.

   We denote the OLS estimator of y_i, given x_i by $\ell_{\mathrm{OLS}}(x_i)$ and the corresponding residual by

   $$r_i := y_i - \ell_{\mathrm{OLS}}(x_i), \qquad 1 \leq i \leq n. \tag{3}$$

   (*Hint: For all the proofs, follow the same arguments as in the chapter, replacing the expectation operator by the averaging operator.*)

   (a) For the standardized data

   $$s(x_i) := \frac{x_i - m(X)}{\sqrt{v(X)}}, \tag{4}$$

   $$s(y_i) := \frac{y_i - m(Y)}{\sqrt{v(Y)}}, \quad 1 \leq i \leq n, \tag{5}$$

   where $m(X)$ and $m(Y)$ are the sample means of X and Y, and $v(X)$ and $v(Y)$ the sample variances, we define the standardized datasets as $S_X := \{s(x_1), s(x_2), \ldots, s(x_n)\}$ and $S_Y := \{s(y_1), s(y_2), \ldots, s(y_n)\}$. Show that the sample mean of the standardized data is zero,

   $$m(S_X) = m(S_Y) = 0, \tag{6}$$

   the sample variance is one,

   $$v(S_X) = \frac{1}{n-1} \sum_{i=1}^{n} s(x_i)^2 = 1, \tag{7}$$

   $$v(S_Y) = \frac{1}{n-1} \sum_{i=1}^{n} s(y_i)^2 = 1, \tag{8}$$

   and the sample covariance is equal to the sample correlation coefficient of the original data,

   $$c(S_X, S_Y) = \frac{1}{n-1} \sum_{i=1}^{n} s(x_i) s(y_i) = \rho_{X,Y}. \tag{9}$$

(b) Prove that

$$\frac{1}{n-1}\sum_{i=1}^{n} r_i^2 = \left(1-\rho_{X,Y}^2\right) v(Y). \tag{10}$$

(c) Prove that the sample correlation coefficient satisfies the same bounds as the correlation coefficient,

$$-1 \leq \rho_{X,Y} \leq 1, \tag{11}$$

as long as $v(Y)$ is not zero.

(d) Prove that if $\rho_{X,Y} = \pm 1$, then $y_i = \beta x_i + \alpha$, $1 \leq i \leq n$, for some constant α, $\beta \in \mathbb{R}$.

5. (Height and Weight) The table in `ANSUR II MALE Public.csv` reports physical measurements of members of the US army. In this problem, we will work with simple linear regression to estimate weight (*Weightlbs*) as a function of height (*Heightin*). Let h represent height (in inches) and w represent weight (in pounds).

(a) Compute the OLS estimator of weight given height. Add this line on a scatterplot (with h on the x-axis; w on the y-axis) and also provide the OLS estimator in equation form:

$$\hat{w} = \hat{\beta}_0 + \hat{\beta}_1 h$$

(b) Compute the sample covariance between the height and the residual of the OLS estimator. Are they correlated?

(c) Interpret the slope coefficient, $\hat{\beta}_1$.

(d) Compute the sample variance of the weight, the OLS estimator of the weight, and the residual. What relationship do you find among these three values?

(e) Compute and compare the sample coefficient of determination (using its definition as the fraction of the variance explained by the linear estimator), and compare it to the squared sample correlation coefficient.

(f) Compute the OLS estimator of h given w. Then rearrange this formula to express w as a function of h.

$$\tilde{w} = \tilde{\beta}_0 + \tilde{\beta}_1 h$$

Add this line also on the scatterplot (in a different color), and compare it with your initial OLS line.

6. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$$\hat{Y}_i = x_i\hat{\beta},$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i Y_i\right) / \left(\sum_{i'=1}^{n} x_{i'}^2\right).$$

show that we can write

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} Y_{i'}$$

What is $a_{i'}$? (Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.)