ESI 4606 Analytics I - Foundations of Data Science
Homework 3
Due: September 28$^{\text{st}}$ (11:00AM), 2022

Problem 1 (2.5 points)

Consider the following data on the propagation velocity of an ultrasonic stress wave through a substance, y (km/s), and the tensile strength of substance, x (MPa).

Table 1: Hypothetical data on the propagation velocity

x, MPa	12	30	36	40	45	57	62	67	71	78	93	94	100	105
y, km/s	3.3	3.2	3.4	3.0	2.8	2.9	2.7	2.6	2.5	2.6	2.2	2.0	2.3	2.1

Suppose a simple linear regression, i.e., $y = \beta_0 + \beta_1 x + \varepsilon$, is used to fit the data, where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. Least squares estimation is employed to estimate the model parameters. Compute the following through **hand calculation**.

(a) What are estimated values for $\hat{\beta}_0$ and $\hat{\beta}_1$? What are their corresponding interpretations.

(b) For a two-sided hypothesis test: $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$, use t-test approach and a significance level of $\alpha = 0.05$ to perform the hypothesis testing and draw the conclusion.

(c) What is the 95% confidence interval for β_1? What is the corresponding interpretation?

(d) What are values for R^2 and $\hat{\sigma}$? If tensile strength of substance is 50 MPa, what is the predicted propagation velocity of an ultrasonic stress wave through the substance based on this model?

Note: (i) **Use R only when** computing the t critical value. **Do not use R** for the rest of calculation.

(ii) **To get full points, include intermediate steps.**

Problem 2 (1 point)

Prove that the fitted least squares line, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, will always pass through the point $(\bar{x}, \bar{y})$, where $\bar{x}$ and $\bar{y}$ are sample averages.

Problem 3 (1.5 points)

This question involves **using R** to perform the multiple linear regression using the "**Carseats**" datain from R library of "**ISLR**".

(a) Fit a multiple regression model to predict "Sales" using "Price", "Urban", and "US". Use the summary() function to print the results.

(b) Provide an interpretation of each coefficient in the model. It is noted that some of the input variables in the model are qualitative.

(c) Based on significant level of $\alpha = 0.05$, for which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(d) On the basis of your response to question (c), fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

(e) How well do the models in (a) and (d) fit the data?

(f) Using the model from (d), obtain 95% confidence intervals for the coefficient(s).

Note: **To get full points, include R codes in the appendix sections**