1.

For the following data set for regression:

Index	X	Y
1	-1	-1
2	0	2
3	3	-2
4	1	3
5	-2	0

Assume that we want to construct a linear regression model in the form $\hat{y}_i = \hat{\beta} x_i$.

(a) Write down the RSS for this regression model.

(b) Find the β that minimizes the RSS, by taking a derivative of it with respect to β.

c. Calculate an estimate of

$$\mathrm{SE}\left(\widehat{\beta}_1\right).$$

2.

We did not show how exactly the coefficients in linear regression are calibrated, formulaically. This problem walks you through the procedure for the simplest case and connects linear regression to KNN. Assume that we have a linear regression problem with only one predictor, and the true model is linear *without an intercept*, i.e. $Y = \beta_1 X + \epsilon$. Assume that we have n samples, $(x_i, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and we want to find the least squares estimate $\hat{\beta}_1$ from the data.

(a) Formulate the residual sum of squares in terms of a candidate $\hat{\beta}_1$ and x_i's and y_i's, which are known. (5 pts)

(b) Show that the residual sum of squares is minimized by the following $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{j=1}^{n} x_j^2} \tag{1}$$

To do that, take a derivative of RSS with respect to the only unknown and set it equal to zero, i.e.:

$$\frac{d}{d\hat{\beta}_1} \text{RSS} = 0$$

Then solve for $\hat{\beta}_1$ to obtain equation (1). It might be easier to first focus on taking the derivative from each residual in RSS formula. (15 pts)

(c) Assume that you have a new test data point x_{new}. Predict its label $\hat{y}_{new}$ using the estimated coefficient in (1). (5 pts)

(d) Explain why your prediction in (2c) is a special case of KNN regression. Determine the k and the similarity measure $\text{sim}(x_{new}, x_i)$ in this case and interpret it. (10 pts)

3.

A regression analysis task relating test scores (Y) to leisure hours (X) produced the following fitted model: $\hat{y} = 25 - 0.5x$.

(a) What is the fitted value of the response variable corresponding to $x = 7$?

(b) What is the residual corresponding to the data point with $x = 3$ and $y = 30$?

(c) If x increases 3 units, how does $\hat{y}$ change?

(d) An additional test score is to be obtained for a new observation at $x = 6$. Would the test score for the new observation necessarily be 22? Explain.

(e) The residual sums of squares (RSS) for this model was found to be 7. If there were $n = 16$ observations, provide an estimate for $\sigma^2 = \text{Var}(\epsilon)$.

(f) If the standard error for $\hat{\beta}_1$ was calculated to be 0.1, is the predictor statistically significant?

(g) Rewrite the regression equation in terms of z where z is leisure hours measured in seconds plus one second. Is this new predictor statistically significant. You must argue using statistics, not subjectively.

If the TSS = 19 and RSS = 7, calculate the Pearson correlation r between x_i 's and y_i 's .

Also, build a 95% C. I. for $\widehat{\beta} = .5$

assuming the sample variance of x_i 's is 4.

4. Show that when the regression line is not dependent on x, then TSS=RSS

5.

A Machine Learning Engineer is working on modeling the amount of funding that companies obtain on a crwodsourcing website and has developed the following model. She used 26 companies to obtain the model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$
$$\hat{y} = 964.8 + 700.2 x_1 + 317.5 x_2 - 200.2 x_3 + 15.3 x_4 + 17.1 x_5$$

The standard errors are:

$$\text{SE}(\hat{\beta}_1) = 12$$
$$\text{SE}(\hat{\beta}_2) = 22.5$$
$$\text{SE}(\hat{\beta}_3) = 101.8$$
$$\text{SE}(\hat{\beta}_4) = 45.3$$
$$\text{SE}(\hat{\beta}_5) = 2.3$$

- $\hat{y}$: the amount of funding obtained by a company in 1000 dollars
- x_1: the average annual salary of the founders
- x_2: the number of employees the startup hired
- x_3: a dummy vaiable that is 1 when the company's field is information technology and 0 otherwise
- x_4: the age of the company
- x_5 is a dummy variable taking value 1 if the founders had previous failures and 0 otherwise

(a) Interpret each of the estimated coefficients $\hat{\beta}_i, i \in \{0, 1, \ldots, 5\}$. (15 pts)

(b) Test, at the $\alpha = 2\%$ significance level, the null hypothesis that the true coefficient on the dummy variable x_5 is 0 against the alternative that it is not 0. (10 pts)

(c) Find and interpret a 99.8% confidence interval for the parameter β_4. (10 pts)

(d) If for the model, RegSS=18147.5 (Regression Sum of Squares) and RSS = 17136.5 (Residual Sum of Squares), test the hypothesis that all the coefficients of the model are 0 (test overal significance of the model) using $\alpha = 5\%$. (10 pts).

The following least squares liner regression model was fitted to a sample of 25 students using data obtained at the end of their senior year in college. The aim was to explain students' weight gains:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$
$$\hat{y} = 7.35 + 0.653 x_1 - 1.345 x_2 + 0.613 x_3$$

The standard errors are:

$$\text{SE}(\hat{\beta}_1) = 0.189$$
$$\text{SE}(\hat{\beta}_2) = 0.565$$
$$\text{SE}(\hat{\beta}_3) = 0.243$$

- $\hat{y}$: weight gained, in pounds, during senior year
- x_1: average number of meals eaten per week
- x_2: average number of hours of exercise per week
- x_3: average number of beers consumed per week

(a) Interpret the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$

(b) Test, at the 2% level, the null hypothesis that the true coefficient on the variable x_3 is 0 against the alternative that it is not 0.

(c) Find and interpret a 99.8% confidence interval for the parameter β_1.

(d) If for the model, SSR=79.2 (Regression Sum of Squares) and SSE = 45.9 (Residual Sum of Squares), test the hypothesis that all the coefficients of the model are 0 (test overall significance of the model) using $\alpha = 1\%$. .

7.

Observe that both the F statistic and R^2 are functions of RegSS and RSS. Show that the F statistic can be described in terms of R^2 and the sample size n and the number of predictors p as: (15 pts)

$$F = \frac{n-p-1}{p}\frac{R^2}{1-R^2} \tag{2}$$

Assume that we estimate the coefficients in the linear regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_7 X_7 + \epsilon$ using 20 samples. Using equation (2), find a threshold t for R^2 such that if $R^2 > t$, we can reject the null hypothesis $\beta_1 = \beta_2 = \cdots = \beta_7 = 0$ with 95% confidence (15 pts)

8.

A multiplicative regression model is a model of the form:

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \dots X_p^{\beta_p} \times \epsilon \tag{3}$$

Explain how β_i's can be calibrated using linear regression. Hint: Think about a transformation that converts multiplication into addition (10 pts).

9.

Assume that in a regression problem with one independent variable $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where the sample size n is 20 and the Pearson correlation r between X and Y is 0.46. Test at $\alpha = 0.01$ the null hypothesis $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.