Old Faithful Geyser

1. Consider the simple linear regression model
$$y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2).$$
Suppose that all the x's are between 0 and 1 (i.e., x_i is in the interval [0,1]). Suppose you are going to run $n = 10$ experiments and you are able to design the experiment (i.e., you choose the choose the x's). What is the choice of (x_1, x_2, ..., x_{10}) that will minimize the variance of $\hat{\beta}_1$. (Hint: Write out the formula for the variance of the estimator of the slope, argue that minimizing the variance is equivalent to maximizing S_{xx}, and then maximize S_{xx}) (The answer is not unique). Recall,
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

2. For the multiple linear regression estimator,
$$y = X\beta + \epsilon,$$
derive the maximum likelihood estimator for the regression coefficients. Show all of your steps.

3. Consider the multiple linear regression estimator,
$$y = X\beta + \epsilon.$$
   a. For the fitted values, show that
   $$V(\hat{y}) = \sigma^2 H$$
   b. For a point x_0, show that.
   $$\hat{y}_0 = x_0' \hat{\beta}$$
   is an unbiased estimator of $E(y|x_0)$.

4. In this question, you will simulate data and perform various analyses. (The idea behind this question is to use simulation to evaluate the performance of your model and estimation approach).
    a. Generate two predictor variables of length n = 200 from a normal distribution with mean 0 and standard deviation of 2. Generate a response vector Y of length n = 200 using your predictors according to the model
    $$y = 1 + 2x_1 + 5x_2 + \epsilon$$
    where random error from a N(0,1) distribution. Estimate the least squares regression line to your data. What is the estimated line?
    b. Estimate the variance for each of the three regression coefficients. How do these compare to the theoretical values for the variance of these predictors (you can compute the theoretical values since you know the x's and also the variance of the errors)?
    c. For your data, test the hypothesis:
    $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0;$ assuming x_2 is in the model.
    Use a significance level of α=0.05.
    d. Repeat parts a. and c. 1000 times. What proportion of times do you reject H_0?