1. Simple linear regression is a special case of multiple linear regression, so everything we have talked about in multiple linear regression applies to this special case. Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$, or, in matrix notation, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta}_{2\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\epsilon}_{n\times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

(a) Show

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \quad \text{and} \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\overline{x})^2} & -\frac{\overline{x}}{\sum_{i=1}^n (x_i-\overline{x})^2} \\ -\frac{\overline{x}}{\sum_{i=1}^n (x_i-\overline{x})^2} & \frac{1}{\sum_{i=1}^n (x_i-\overline{x})^2} \end{pmatrix}.$$

(b) Show the hat matrix for simple linear regression is

$$\mathbf{H} = \begin{pmatrix} \frac{1}{n} + \frac{(x_1-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2} & \frac{1}{n} + \frac{(x_1-\overline{x})(x_2-\overline{x})}{\sum_{i=1}^n (x_i-\overline{x})^2} & \cdots & \frac{1}{n} + \frac{(x_1-\overline{x})(x_n-\overline{x})}{\sum_{i=1}^n (x_i-\overline{x})^2} \\ \frac{1}{n} + \frac{(x_1-\overline{x})(x_2-\overline{x})}{\sum_{i=1}^n (x_i-\overline{x})^2} & \frac{1}{n} + \frac{(x_2-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2} & \cdots & \frac{1}{n} + \frac{(x_2-\overline{x})(x_n-\overline{x})}{\sum_{i=1}^n (x_i-\overline{x})^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} + \frac{(x_1-\overline{x})(x_n-\overline{x})}{\sum_{i=1}^n (x_i-\overline{x})^2} & \frac{1}{n} + \frac{(x_2-\overline{x})(x_n-\overline{x})}{\sum_{i=1}^n (x_i-\overline{x})^2} & \cdots & \frac{1}{n} + \frac{(x_n-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2} \end{pmatrix}.$$

(c) Calculate rank($\mathbf{H}$) and tr($\mathbf{H}$) for simple linear regression.
(d) Let $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ denote the vector of residuals from estimating the simple linear regression model using least squares. Under our usual assumptions for the error vector $\boldsymbol{\epsilon}$, calculate the diagonal entries of the variance-covariance matrix of $\mathbf{e}$.

2. Consider the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. The design matrix $\mathbf{X}$ is $n \times p$ and has rank($\mathbf{X}$) $= p$. Let $\mathbf{H}$ denote the hat matrix. Define

$$\widetilde{\sigma}^2 = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}/c,$$

and consider using $\widetilde{\sigma}^2$ as an estimator for σ^2.
(a) Show the mean-squared error of $\widetilde{\sigma}^2$ is

$$\text{MSE}(\widetilde{\sigma}^2) = \sigma^4 \left[\frac{2(n-p)^2 + (n-p-c)^2}{c^2} \right].$$

Hint: Recall MSE($\widetilde{\sigma}^2$) $= V(\widetilde{\sigma}^2) + [B(\widetilde{\sigma}^2)]^2$, where $B(\widetilde{\sigma}^2) = E(\widetilde{\sigma}^2) - \sigma^2$ is the bias. Use the formulas for the mean and variance of a quadratic form; see HW8, Problem 5.
(b) Find the value of c that minimizes MSE($\widetilde{\sigma}^2$).

3. This problem deals with an extrusion process used in soybeans; basically "extrusion" refers to the process by which certain materials are extracted from the soybeans (e.g., fiber, oil, etc.) to be used in other products (e.g., cattle feed, flour, etc.). An experiment was performed to investigate the relationship between

$$Y = \text{soluble dietary fiber percentage (SDFP) in soybean residue}$$

and three independent variables

- x_1 = extrusion temperature, (`temp`, in deg C)
- x_2 = feed moisture (`moisture`, in %)
- x_3 = extrusion screw speed (`speed`, in rpm).

Here are the data recorded in the experiment:

Observation	x_1	x_2	x_3	Y
1	35	110	160	11.13
2	25	130	180	10.98
3	30	110	180	12.56
4	30	130	200	11.46
5	30	110	180	12.38
6	30	110	180	12.43
7	30	110	180	12.55
8	25	110	160	10.59
9	30	130	160	11.15
10	30	90	200	10.55
11	30	90	160	9.25
12	25	90	180	9.58
13	35	110	200	11.59
14	35	90	180	10.68
15	35	130	180	11.73
16	25	110	200	10.81
17	30	110	180	12.68

(a) Experimenters initially considered the multiple linear regression model to relate `SDFP` to the three independent variables:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 17$. Calculate the ANOVA table for this analysis using R (entering the independent variables in the order they appear in the model above). Interpret each term's sum of squares contribution (in words) and then assess the overall fit of the model using an F test. What are your conclusions from this analysis?

(b) Experimenters also considered a multiple linear regression model with quadratic terms:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2 + \beta_6 x_{i3}^2 + \epsilon_i,$$

for $i = 1, 2, ..., 17$. The extra independent variables are the squared versions of x_1, x_2, and x_3, respectively. Analyze these data under this population-level model.

(c) We see the regression model in part (a) is a special case of the model in part (b) when $\beta_4 = \beta_5 = \beta_6 = 0$. Using sequential sum of squares, describe how you could test

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$$
versus
$$H_a : H_0 \text{ not true.}$$

Note that if H_0 is true, then the simpler model in part (a) is an adequate description.