1. Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2(3x_i^2 - 2) + \epsilon_i,$$

for $i = 1, 2, 3$, where $x_1 = -1$, $x_2 = 0$, and $x_3 = 1$.
(a) Put this model into $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ form.
(b) Find the least squares estimates of β_0, β_1, and β_2. *Hint:* $\mathbf{X}'\mathbf{X}$ is diagonal, so inverting this matrix is easy.
(c) Show that the least squares estimates of β_0 and β_1 are unchanged if $\beta_2 = 0$. Why do you think this happens? *Hint:* What do you note about the 3 column vectors of $\mathbf{X}$ in part (a)?

2. Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is $n \times 1$ and $\mathbf{X}$ is $n \times p$. The $n \times 1$ error vector $\boldsymbol{\epsilon}$ satisfies $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. Let $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ denote the hat matrix. Recall $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$ and $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ are the vectors of least squares fitted values and residuals, respectively. Calculate the following and the note the dimension of each quantity.
(a) $E(\widehat{\mathbf{Y}})$
(b) $\text{Cov}(\widehat{\mathbf{Y}})$
(c) $E(\mathbf{e})$
(d) $\text{Cov}(\mathbf{e})$
(e) $\text{Cov}(\widehat{\mathbf{Y}}, \mathbf{e})$.
(f) Under the assumption that $\boldsymbol{\epsilon}$ is multivariate normal (which is not needed in the parts above), what are the sampling distributions of $\widehat{\mathbf{Y}}$ and $\mathbf{e}$?

3. 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

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

(c) Calculate rank($\mathbf{H}$) and tr($\mathbf{H}$) for simple linear regression.
(d) In regression analysis, an observation might be declared an "outlier" if its *studentized residual* is large in absolute value. Under our usual assumptions for the errors; i.e., $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, you showed on HW6 (Problem 2) that the ith residual

$$e_i \sim \mathcal{N}\left(0, \sigma^2(1 - h_{ii})\right),$$

where note that

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}$$

is the ith diagonal element of $\mathbf{H}$. Standardizing, we have

$$\frac{e_i}{\sqrt{\sigma^2(1 - h_{ii})}} \sim \mathcal{N}(0, 1) \quad \Longrightarrow \quad r_i = \frac{e_i}{\sqrt{\widehat{\sigma}^2(1 - h_{ii})}} \sim t(n - 2),$$

where $\widehat{\sigma}^2$ is the mean-squared error (MSE). The quantity r_i is the studentized residual of the ith observation. A reasonable level α decision rule is to declare the ith observation to be an outlier if

$$|r_i| > t_{n-2,\alpha/2},$$

where $t_{n-2,\alpha/2}$ is the upper $\alpha/2$ quantile of the $t(n-2)$ distribution. Calculate the studentized residuals for the maximum O_2 uptake exercise data in Example 12.1 (notes). Would any observations be declared "outliers" by using the decision rule above at $\alpha = 0.05$? Of course, we are ignoring the "multiple comparisons" problem here because there are $n = 24$ observations (so we are actually performing 24 hypothesis tests, one for each observation). However, let's worry about the multiple comparisons problem another day.

4. 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 = soluble dietary fiber percentage (`SDFP`) in soybean residue

to 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 overall 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, formulate an approach to test

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$$

versus

$$H_a : H_0 \text{ not true.}$$