Note: This homework assignment covers Chapter 11.

Disclaimer: If you use R, include all R code and output as attachments. Do not just "write in" the R code you used. Also, don't just write the answer and say this is what R gave you. If my grader can't see how you got an answer, it is wrong. I want to see your code and your answers accompanying your code (like in the notes).

1. Engineers studied the impact of temperature (x_1) and concentration (x_2) on the percentage of impurities Y for a chemical process. The data are below; there are $n = 14$ observations.

Observation	Y	x_1	x_2
1	14.9	85.8	42.3
2	16.9	83.8	43.4
3	17.4	84.5	42.7
4	16.9	86.3	43.6
5	16.9	85.2	43.2
6	16.7	83.8	43.7
7	17.1	86.1	43.3
8	16.9	85.9	43.4
9	16.7	85.7	43.3
10	16.9	86.3	42.6
11	16.7	83.5	44.0
12	17.1	85.8	42.8
13	17.6	85.9	43.1
14	16.9	84.2	43.5

The researchers would like to consider the statistical model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i,$$

for $i = 1, 2, ..., 14$, under the standard assumptions for the errors ϵ_i.

Note: In R (if you use R), use the following code when entering the data. Then, use my R code online as a guide to answer the questions above.

```
impurities = c(14.9, 16.9, ... 16.9)
temp = c(85.8, 83.8, ... 84.2)
concentration = c(42.3, 43.4, ... 43.5)
```

(a) Calculate the multiple linear regression equation that relates percentage of impurities (Y) to temperature (x_1) and concentration (x_2).
(b) Test to see whether the impurity percentage is linearly related to temperature in the population (after adjusting for the effect of concentration). You can do this by performing a relevant hypothesis test or by writing a relevant confidence interval.

(c) Test to see whether the impurity percentage is linearly related to concentration in the population (after adjusting for the effect of temperature). You can do this by performing a relevant hypothesis test or by writing a relevant confidence interval.
(d) Calculate R^2 and interpret its value clearly.
(e) Perform residual diagnostics for the model fit, specifically, construct a normal qq-plot for the residuals and plot the residuals versus the fitted values. Interpret each plot.

2. Researchers were interested in examining the relationship between the percentage of hardwood in a batch of pulp (x) and the tensile strength of Kraft paper made from the batch (Y, measured in psi). The data are below. There are $n = 19$ observations.

Y	x	Y	x
6.3	1.0	39.9	6.5
11.1	1.5	42.0	7.0
20.0	2.0	46.1	8.0
24.0	3.0	53.1	9.0
26.1	4.0	52.0	10.0
30.0	4.5	52.5	11.0
33.8	5.0	48.0	12.0
34.0	5.5	42.8	13.0
38.1	6.0	27.8	14.0
		21.9	15.0

(a) First consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., 19$, that is, a simple linear regression model that relates the percentage of hardwood to tensile strength. Fit this model, and superimpose the least squares line onto a scatterplot for the data.
(b) You should see in part (a) that the simple linear regression model is clearly not adequate for these data. From the scatterplot, the quadratic model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,$$

for $i = 1, 2, ..., n = 19$, appears to be much better. Therefore, fit a quadratic model to these data. Note that this is a multiple linear regression model with two independent variables: x and x^2. To create the x^2 variable in R, say, you can use the commands:

```
ten.strength = c(6.3, 11.1, ..., 21.9)
percentage = c(1.0, 1.5, ..., 15.0)
percentage.sq = (percentage)^2
```

Then you can regress Y on x and x^2 using the following command:

```
fit = lm(ten.strength~percentage+percentage.sq)
summary(fit)
```

Write out an equation for your fitted quadratic regression model in the form

$$\widehat{Y} = b_0 + b_1 x + b_2 x^2.$$

To display a graph of your fitted model, use the following commands:

```
x = percentage
plot(percentage,ten.strength,xlab = "Hardwood percentage",
      ylab = "Tensile strength", pch=16)
curve(expr = fit$coefficients[1] +
             fit$coefficients[2]*x +
             fit$coefficients[3]*x^2, col = "black",
             lty = "solid", lwd = 1, add = TRUE)
```

(c) Is β_2, the population-level regression parameter associated with the quadratic term, different from 0? Answer this question by performing a hypothesis test of

$$H_0 : \beta_2 = 0$$
versus
$$H_1 : \beta_2 \neq 0.$$

In the light of your finding, what does this say about the quadratic model (when compared to the simple linear regression model)?
(d) Based on your fitted quadratic model, what percentage of hardwood maximizes the tensile strength?
(e) Display residual plots for the quadratic model fit. What are your conclusions?

3. The brake horsepower (`HORSE`, Y) developed by an automobile engine is thought to be a function of the engine speed in revolutions per minute (`RPM`, x_1), the road octane number of the fuel (`OCT`, x_2), and the engine compression (`COM`, x_3). An experiment is run in a laboratory at twelve different times; on each run, the temperature (`TEMP`, x_4) is also recorded. The data from the experiment are below (see next page).
(a) Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

for $i = 1, 2, ..., 12$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Write this model in matrix notation, that is, of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Write out explicitly what each matrix/vector is.
(b) Calculate the vector of fitted values $\widehat{\mathbf{Y}}$ and the vector of residuals $\mathbf{e}$ from the least squares fit. Both $\widehat{\mathbf{Y}}$ and $\mathbf{e}$ are 12×1 vectors. Show numerically that $\sum_{i=1}^{12} e_i = 0$ and that $\widehat{\mathbf{Y}}'\mathbf{e} = 0$. Each of these is a byproduct of the least squares model fitting procedure.

Y	x_1	x_2	x_3	x_4
225	2000	90	100	71.2
212	1800	94	95	70.3
229	2400	88	110	72.3
222	1900	91	96	69.9
219	1600	86	100	73.2
278	2500	96	110	70.0
246	3000	94	98	70.7
237	3200	90	100	70.8
233	2800	88	105	72.1
224	3400	86	97	71.8
223	1800	90	100	71.1
230	2500	89	104	70.6

(c) Here is the ANOVA table for the fit of the model:

Analysis of Variance

Source	DF	SS	MS	F	Pr > F
Regression	4	2597.52	649.40	7.41	0.0117
Residual	7	613.48	87.64		
Total	11	3211.00			

Here are the least-squares estimates (`Parm.Est`), standard errors (`Std.Err.`), t statistics, and the associated p-values for the model fit:

Variable	DF	Parm.Est	Std.Err.	t value	Pr > |t|
Intercept	1	-402.8470	469.5873	-0.86	0.4194
RPM	1	0.0110	0.0049	2.26	0.0581
OCT	1	3.5253	1.5881	2.22	0.0619
COM	1	1.8005	0.6106	2.95	0.0214
TEMP	1	1.5127	5.0766	0.30	0.7744

Questions for you to answer (for part (c)):
(c1) Explain what the F statistic above is used to test. What is the conclusion reached from the value of this statistic?
(c2) Calculate R^2 and interpret its value clearly.
(c3) Use the information above to determine whether or not `TEMP` adds to the model (in the presence of the other three independent variables). State your conclusion in a well-written sentence.

(d) Engineers would like to use the model to predict the brake horsepower for a new engine at certain settings of x_1, x_2, x_3, and x_4. Would a confidence interval or prediction interval be appropriate? Explain.