1. (a) Consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \iff E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Here is the R output from estimating this model:

```
> fit = lm(impurities ~ temp + conc)
> fit

Coefficients:
(Intercept)         temp         conc
   -13.8624       0.0996       0.5139
```

The (least-squares) prediction equation is

$$\widehat{Y} = -13.8624 + 0.0996x_1 + 0.5139x_2,$$

or, in other words,

$$\widehat{\texttt{Impurities}} = -13.8624 + 0.0996(\texttt{Temp}) + 0.5139(\texttt{Conc}).$$

(b) The R output below shows the ANOVA table:

```
> Model = cbind(temp,conc)
> fit = lm(impurities ~ Model)
> anova(fit)

Analysis of Variance Table
          Df Sum Sq Mean Sq F value Pr(>F)
Model      2  0.566  0.2828   0.723  0.507
Residuals 11  4.303  0.3912
```

(c) The F statistic tests

$$H_0 : \beta_1 = \beta_2 = 0$$

versus

$$H_1 : \text{at least one of } \beta_1 \text{ and } \beta_2 \text{ is nonzero.}$$

We know that large values of F (small p-values) are evidence against H_0, but that is not what we see here. At the $\alpha = 0.05$ level of significance, we would not reject H_0 because the p-value (0.507) is larger than 0.05. This means that neither `Temperature` nor `Concentration` is linearly related to the mean impurity percentage $E(Y)$ in the population of all production process measurements.

(d) We can use R's `confint` function to calculate 95% confidence intervals:

```
> fit = lm(impurities ~ temp + conc)
> confint(fit,level=0.95)

               2.5 %  97.5 %
(Intercept) -81.1128 53.3879
temp         -0.3330  0.5322
conc         -0.4303  1.4581
```

The output shows:

- We are 95% confident that β_1, the population parameter attached to `Temperature`, is between -0.3330 and 0.5322. Note that this interval includes "0," which is what we would expect from our analysis in part (c). We cannot conclude the population mean impurity percentage $E(Y)$ is linearly related to `Temperature` after accounting for the effect of `Concentration`.
- We are 95% confident that β_2, the population parameter attached to `Concentration`, is between -0.4303 and 1.4581. Note that this interval also includes "0," which is what we would expect from our analysis in part (c). We cannot conclude the population mean impurity percentage $E(Y)$ is linearly related to `Concentration` after accounting for the effect of `Temperature`.

(e) We now consider a multiple linear regression model that allows `Temperature` (x_1) and `Concentration` (x_2) to interact with each other:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \iff E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

Estimating this model simply uses an extra term in the `lm` statement:

```
> fit.2 = lm(impurities ~ temp + conc + temp*conc) # interaction model
> fit.2

Coefficients:
(Intercept)         temp         conc    temp:conc
   2205.458      -25.920      -50.553        0.599
```

The (least-squares) prediction equation is

$$\widehat{Y} = 2205.458 - 25.920x_1 - 50.553x_2 + 0.599x_1x_2,$$

or, in other words,

$$\widehat{\texttt{Impurities}} = 2205.458 - 25.920(\texttt{Temp}) - 50.553(\texttt{Conc}) + 0.599(\texttt{Temp*Conc})$$

Here are 95% confidence intervals for the regression parameters β_1, β_2, and β_3 in the interaction model:

```
> confint(fit.2,level=0.95)
                 2.5 %  97.5 %
(Intercept) -1099.1262 5510.04
temp          -64.6571   12.82
conc         -126.5822   25.48
temp:conc      -0.2926    1.49
```

The confidence interval for all parameters, including the interaction term parameter β_3, includes "0." This suggests that neither `Temperature`, `Concentration`, nor the interaction between them is linearly related to the mean impurity percentage $E(Y)$ in the population of all production process measurements.

(f) I'm not really enthusiastic about either model because nothing appears to be significant. I decided to consider a third model where quadratic terms were included:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \epsilon \iff E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2.$$

Here is the R output from estimating this model:

```
> temp.sq = temp^2
> conc.sq = conc^2
> fit.3 = lm(impurities ~ temp + conc + temp.sq + conc.sq)
> summary(fit.3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1572.222   1660.682   -0.95   0.3685
temp          -66.768     49.716   -1.34   0.2122
conc          205.594     61.585    3.34   0.0087 **
temp.sq         0.392      0.292    1.34   0.2126
conc.sq        -2.384      0.716   -3.33   0.0088 **

Residual standard error: 0.453 on 9 degrees of freedom
Multiple R-squared:  0.62,      Adjusted R-squared:  0.451
F-statistic: 3.67 on 4 and 9 DF,  p-value: 0.0487
```

This model fit is more encouraging. We see now the linear and quadratic effects of `Concentration` are significant in the population of all production process measurements (small p-values). However, neither the linear nor quadratic effect of `Temperature` is significant. Therefore, I decided to estimate a quadratic model as a function of `Concentration` alone, that is,

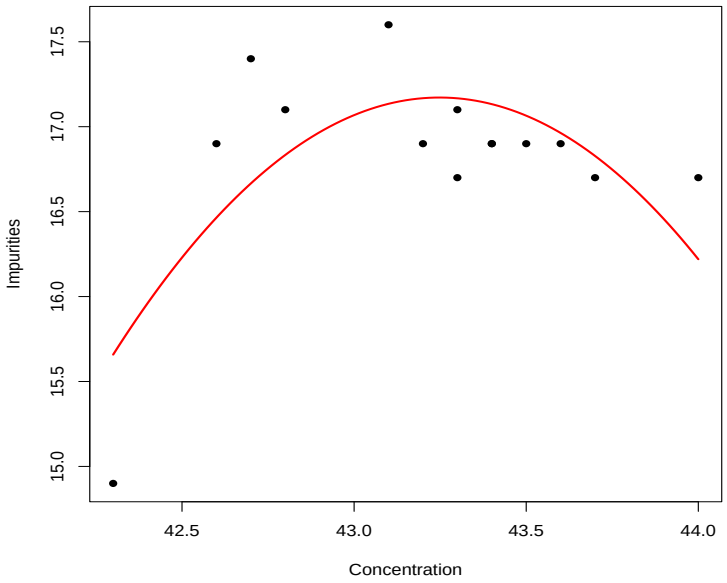
$$Y = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \epsilon \iff E(Y) = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2.$$

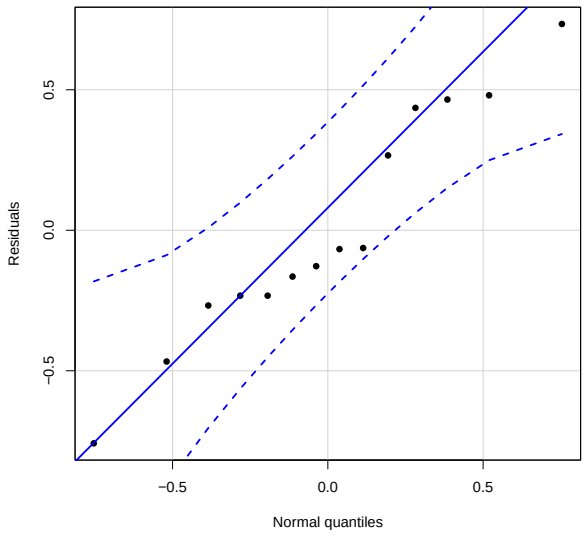
Here is the R output from estimating this model:

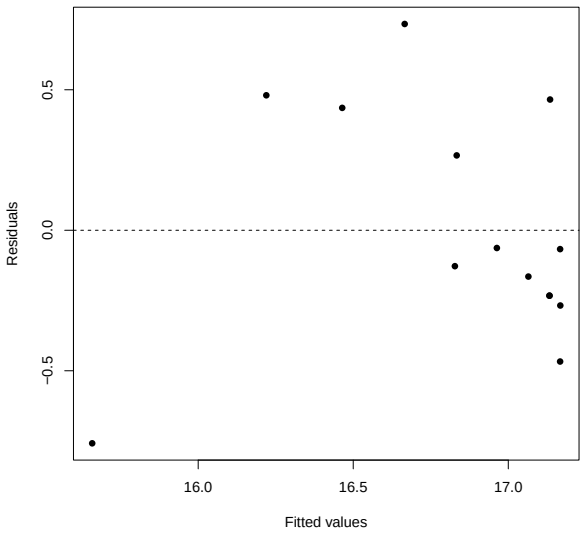
```
> fit.4 = lm(impurities ~ conc + conc.sq)
> summary(fit.4)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3132.144    974.829   -3.21   0.0083 **
conc          145.640     45.209    3.22   0.0081 **
conc.sq        -1.684      0.524   -3.21   0.0083 **

Residual standard error: 0.454 on 11 degrees of freedom
Multiple R-squared:  0.533,     Adjusted R-squared:  0.449
F-statistic: 6.29 on 2 and 11 DF,  p-value: 0.0151
```

The linear and quadratic effects of `Concentration` are significant. The R^2 statistic says that a little over 53% of the variation in the impurities percentage data can be explained by quadratic model fit. This may be as good as we can do.

Below is a scatterplot of the observations for `Concentration` and `Impurities` with the quadratic fit superimposed. The normal qq-plot and residual plot from the fit (`fit.4`) are also shown.







The qq-plot for the residuals looks fine (of course, there are only 14 observations), but the residual plot does not display the random appearance we would like to see. There appears to be one outlier (lower left, Observation 1) which may be making the modeling process more difficult and distorting the figures above. One could remove this outlier and refit the model (perhaps just a simple linear regression model with `Concentration`), but I have always had a problem with removing observations from a data analysis without good cause for doing so.

2. (a) Consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \iff E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Here is the R output from estimating this model:

```
> fit = lm(horse ~ rpm + oct + com + temp)
> fit

Coefficients:
(Intercept)          rpm          oct          com         temp
   -402.847        0.011        3.525        1.801        1.513
```

The (least-squares) prediction equation is

$$\widehat{Y} = -402.847 + 0.011x_1 + 3.525x_2 + 1.801x_3 + 1.513x_4$$

or, in other words,

$$\widehat{\texttt{BHP}} = -402.847 + 0.011(\texttt{RPM}) + 3.525(\texttt{OCT}) + 1.801(\texttt{COM}) + 1.513x_4(\texttt{TEMP})$$

(b) The R output below shows the ANOVA table:

```
> Model = cbind(rpm,oct,com,temp)
> fit = lm(horse ~ Model)
> anova(fit)
Analysis of Variance Table

Response: horse
          Df   Sum Sq Mean Sq F value   Pr(>F)
Model      4 2597.516 649.379 7.40958 0.011679 *
Residuals  7  613.484  87.641
```

The F statistic tests

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

versus

$$H_1 : \text{at least one of } \beta_j\text{'s is nonzero.}$$

We know that large values of F (small p-values) are evidence against H_0. At the $\alpha = 0.05$ significance level, we would reject H_0 in favor of H_1 (p-value < 0.05). We would conclude at least one of the independent variables is linearly related to brake horsepower in the population.
(c) From part (b), we see that the regression sum of squares is

$$\text{SS}_\text{R} = 2597.516$$

Here is the first set of sequential SS:

```
> # Sequential SS (set 1)
> fit = lm(horse ~ rpm + oct + com + temp)
```

```
> anova(fit)
Analysis of Variance Table

          Df   Sum Sq  Mean Sq  F value    Pr(>F)
rpm        1  509.354  509.354  5.81185 0.0467206 *
oct        1 1132.555 1132.555 12.92274 0.0088004 **
com        1  947.826  947.826 10.81493 0.0133281 *
temp       1    7.782    7.782  0.08879 0.7743718
Residuals  7  613.484   87.641
```

Note that, up to rounding error,

$$\text{SS}_\text{R} = 2597.516 = 509.354 + 1132.555 + 947.826 + 7.782.$$

Here is the second set of sequential SS:

```
> # Sequential SS (set 2)
> fit = lm(horse ~ temp + com + oct + rpm)
> anova(fit)
Analysis of Variance Table

Response: horse
          Df   Sum Sq  Mean Sq  F value    Pr(>F)
temp       1  375.248  375.248  4.28167 0.0773057 .
com        1 1497.911 1497.911 17.09153 0.0043809 **
oct        1  275.793  275.793  3.14687 0.1193470
rpm        1  448.565  448.565  5.11824 0.0581294 .
Residuals  7  613.484   87.641
```

Note that, up to rounding error,

$$\text{SS}_\text{R} = 2597.516 = 375.248 + 1497.911 + 275.793 + 448.565.$$

There are $4! = 24$ different sets of sequential sums of squares. This is number of ways you can permute the four predictors `rpm`, `oct`, `com`, and `temp`.
(d) We will use R's predict function to calculate both intervals:

```
fit = lm(horse ~ rpm + oct + com + temp)
> x.0 = data.frame(rpm=20,oct=90,com=100,temp=72)
> predict(fit,x.0,conf.level=0.95,interval="confidence")
      fit     lwr     upr
1 203.618 175.738 231.498
> predict(fit,x.0,conf.level=0.95,interval="prediction")
      fit     lwr     upr
1 203.618 168.019 239.218
```

Confidence interval: For the population of all experimental runs where

$$\mathbf{x}_0 = \begin{pmatrix} x_{10} \\ x_{20} \\ x_{30} \\ x_{40} \end{pmatrix} = \begin{pmatrix} 2500 \\ 90 \\ 100 \\ 72 \end{pmatrix},$$

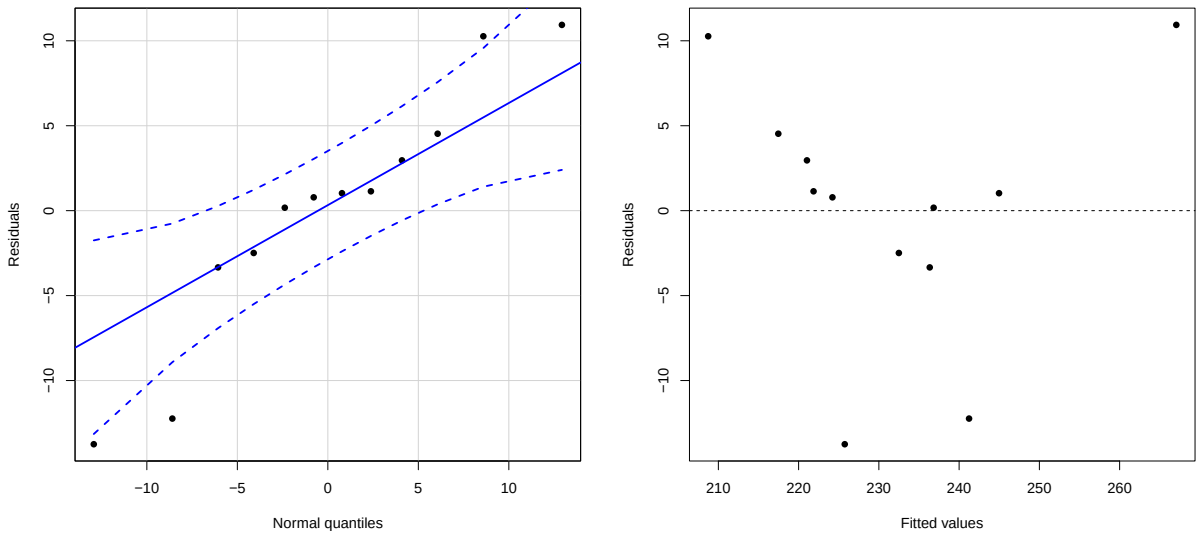
we are 95% confident the population mean brake horsepower $E(Y)$ is between 175.738 and 231.498.

Prediction interval: For a single experimental run where

$$\mathbf{x}_0 = \begin{pmatrix} x_{10} \\ x_{20} \\ x_{30} \\ x_{40} \end{pmatrix} = \begin{pmatrix} 2500 \\ 90 \\ 100 \\ 72 \end{pmatrix},$$

we would predict the brake horsepower Y^* to be between 168.019 and 239.218 with probability 0.95.

(e) Here are the normal qq-plot and residual plot from the fit:



The qq-plot looks fine (there are only 12 observations). The residual plot doesn't quite have the random appearance I would like to see, but, again with only 12 experimental runs, it is hard to discern whether this is truly random (no model violations) or not.

R CODE:

```
# Problem 1
# Enter the data
impurities = c(14.9,16.9,17.4,16.9,16.9,16.7,17.1,16.9,16.7,16.9,16.7,17.1,17.6,16.9)
temp = c(85.8,83.8,84.5,86.3,85.2,83.8,86.1,85.9,85.7,86.3,83.5,85.8,85.9,84.2)
conc = c(42.3,43.4,42.7,43.6,43.2,43.7,43.3,43.4,43.3,42.6,44.0,42.8,43.1,43.5)

# Estimate the model
fit = lm(impurities ~ temp + conc)
fit

# ANOVA table
Model = cbind(temp,conc)
fit = lm(impurities ~ Model)
anova(fit)

# Confidence intervals for regression parameters
fit = lm(impurities ~ temp + conc)
confint(fit,level=0.95)

# Estimate the interaction model
fit.2 = lm(impurities ~ temp + conc + temp:conc)
fit.2
confint(fit.2,level=0.95)

temp.sq = temp^2
conc.sq = conc^2
# Estimate the quadratic model (with both predictors)
fit.3 = lm(impurities ~ temp + conc + temp.sq + conc.sq)
summary(fit.3)

# Estimate the quadatic model
fit.4 = lm(impurities ~ conc + conc.sq)
summary(fit.4)

# Scatterplot with quadratic model fit superimposed
plot(conc,impurities,xlab = "Concentration",ylab = "Impurities",pch=16)
curve(expr = fit.4$coefficients[1] +
    fit.4$coefficients[2]*x +
    fit.4$coefficients[3]*x^2, col="red",lwd=2,add=TRUE)

# QQ plot of the residuals
resid = resid(fit.4)
library(car)
qqPlot(resid,distribution="norm",mean=mean(resid),sd=sd(resid),
    xlab="Normal quantiles",ylab="Residuals",pch=16,
    envelope=list(border=TRUE,style="lines"),id=FALSE)

# Residual plot
plot(fitted(fit.4),resid(fit.4),pch=16,xlab="Fitted values",ylab="Residuals")
abline(h=0,lty=2)
```

```
# Problem 2
# Enter the data
horse = c(225,212,229,222,219,278,246,237,233,224,223,230)
rpm = c(2000,1800,2400,1900,1600,2500,3000,3200,2800,3400,1800,2500)
oct = c(90,94,88,91,86,96,94,90,88,86,90,89)
com = c(100,95,110,96,100,110,98,100,105,97,100,104)
temp = c(71.2,70.3,72.3,69.9,73.2,70.0,70.7,70.8,72.1,71.8,71.1,70.6)

# Estimate the model
fit = lm(horse ~ rpm + oct + com + temp)
fit

# ANOVA table
Model = cbind(rpm,oct,com,temp)
fit = lm(horse ~ Model)
anova(fit)

# Sequential SS (set 1)
fit = lm(horse ~ rpm + oct + com + temp)
anova(fit)

# Sequential SS (set 2)
fit = lm(horse ~ temp + com + oct + rpm)
anova(fit)

# Confidence interval and prediction interval
fit = lm(horse ~ temp + com + oct + rpm)
x.0 = data.frame(rpm=20,oct=90,com=100,temp=72)
predict(fit,x.0,conf.level=0.95,interval="confidence")
predict(fit,x.0,conf.level=0.95,interval="prediction")

# QQ plot of the residuals
resid = resid(fit)
qqPlot(resid,distribution="norm",mean=mean(resid),sd=sd(resid),
    xlab="Normal quantiles",ylab="Residuals",pch=16,
    envelope=list(border=TRUE,style="lines"),id=FALSE)

# Residual plot
plot(fitted(fit),resid(fit),pch=16,xlab="Fitted values",ylab="Residuals")
abline(h=0,lty=2)
```