1. A forest scientist is interested in studying the relationship between the terminal velocity of falling maple samaras and a measure of their size and weight known as disk loading. A samara is the winged fruit that falls to the ground with a helicopter-like motion. The disk loading is a quantity related to helicopter aerodynamics. In an experiment samaras were randomly selected from each of three distinct trees (of the same species). The disk loading (loading) for each samara was determined and the terminal velocity (velocity) was measured in the laboratory. The forest scientist conducting the study hypothesizes that there is a linear relationship between terminal velocity and disk loading. The scientist wishes to determine whether the relationship is the same for each tree. The data consist of 13 observations from Tree 1, 11 observations from Tree 2, and 12 observations from Tree 3. The file samara.txt contains the data. Investigate the data and determine an appropriate model. Support your selection of the best model. Comment on which regression coefficients are the same for each tree and which are different.

   Given the data set, we first graph the disk loading variable versus the terminal velocity using characters to differentiate between the three types of trees.
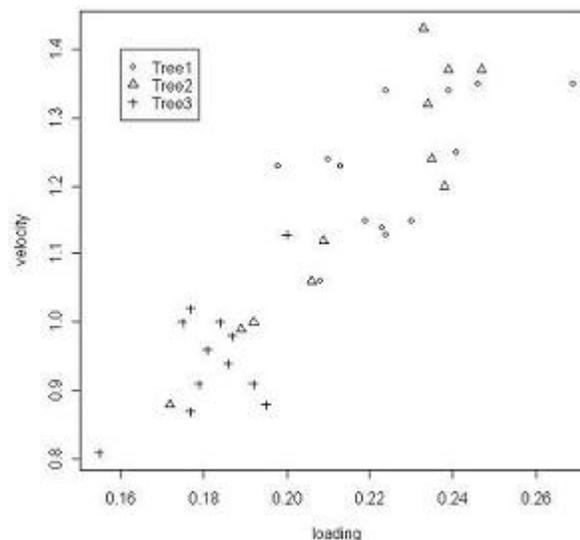
   According to this plot, it appears that the velocity and loading are positively correlated. Data for Tree1 appears in the upper right, Tree3 in the lower left, and Tree2 scattered in both areas. Overall from the graph, it appears that although they have varying levels for each variable, the trees as a whole follow the same linear pattern.

   The general model can be written as:

$$Velocity = \beta_0 + \beta_1 \times Loading + \beta_2 \times I(Tree2) + \beta_3 \times I(Tree3) + \beta_4 \times Loading \times I(Tree2) + \beta_5 \times Loading \times I(Tree3)$$

   According to the summary statistics below, it appears that Tree 3 does not differ significantly from Tree 1 as a main effect (p = 0.4960) nor as an interaction (p = 0.6984). This agrees with our general findings based on the plot. Although it does appear that Trees 1 and 3

follow the same linear trend, Tree 1 tends to have a higher disk loading and therefore a higher velocity than Tree 3. Tree 2 does differ significantly from Trees 1 and 3. Therefore the linear fit for Tree 2 has a significantly different slope and intercept.

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.5521     0.2652   2.082   0.0460 *
loading              2.9831     1.1673   2.555   0.0159 *
treeTree2           -0.8514     0.3382  -2.517   0.0174 *
treeTree3           -0.3094     0.4489  -0.689   0.4960
loading:treeTree2    3.8140     1.5106   2.525   0.0171 *
loading:treeTree3    0.9002     2.3010   0.391   0.6984
```

To verify that this model is plausible, we check the usual assumptions. There does not appear to be a blatant violation of the constant variance and normality assumptions.
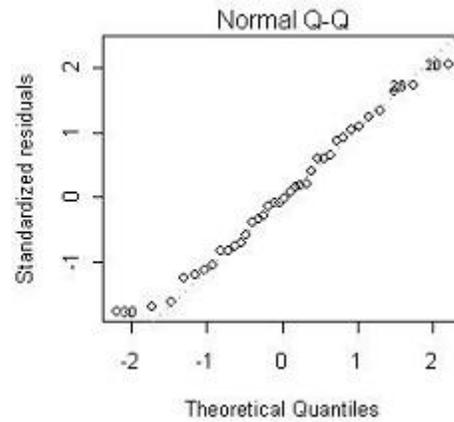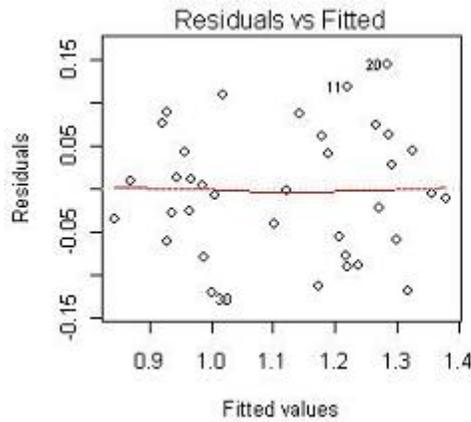
2. Consider the full data set from the file larch.txt. For this problem you do not need to consider tests for any possible transformations. The regression coefficients β_2, β_3, β_4, β_5 correspond to the explanatory variables stored in columns 25 (nitro, phos, potas, ash), respectively. For each test, report the F test statistic, the degrees of freedom for the reference distribution, and the p-value.

(a) Test the hypothesis: $H_o : (\beta_3 = \beta_5 = 0 | \beta_0, \beta_2, \beta_4)$.

```
Analysis of Variance Table

Model 1: height ~ nitro + potas
Model 2: height ~ nitro + phos + potas + ash

  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     23 36771
2     21 30122  2      6649 2.3177 0.1232
```

With a p-value of 0.1232, we fail to reject the null hypothesis. Therefore, the reduced model (excluding phosphorous and ash) is statistically significant.

(b) Test the hypothesis $H_o : (\beta_4 = 3 \times \beta_2 | \beta_0)$.

```
Analysis of Variance Table

Model 1: height ~ I(potas + 3 * nitro)
Model 2: height ~ nitro + potas

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     24 55657
2     23 36771  1     18886 11.813 0.002247 **
```

With a p-value of 0.002247, we reject the null hypothesis. Therefore, there is evidence that the amount of potassium in the soil is not 3 times that of the amount of nitrogen.

3. Consider the following experimental design. A dairy scientist is interested in testing the effect of diet on nitrogen content in manure. He would like to generalize the results he finds to all dairy cows of a given species in the US. The scientist designs four diets which differ in content. The experimental setup allows for individual measurement of nitrogen content of the manure, but cows in each diet group eat from a common source. The scientist has access to a large research herd. For the experiment, he selects 8 primiparous cows (cows that have given birth once) and 8 multiparous cows (cows that have given birth multiple times). The associated variable is called parity. The researcher assigns two cows from each parity level into each of the four treatment groups. Each cow is measured at the beginning of the experiment and after one and two months on the assigned diet. (Realistic designs would be different.) For each of the following variables, indicate if it should be treated as a fixed effect or a random effect. Briefly explain.

    (a) **Diet** should be treated as a *fixed* effect. This is because we have four diets that must stay consistent.

    (b) **Parity** must be treated as a *fixed* effect because it would be impossible to randomly assign a cow that gave birth more than one time to the group that has only given birth once.

    (c) **Period** (time of measurement) should be treated as a *fixed* effect. It is known that the amount of milk produced by a female cow varies after giving birth. Therefore we want to be able to predict the nitrogen content based on diet at a specific period in a female's milking cycle.

    (d) **ID** (unique identifier for each cow) should be treated as a *random* effect. We want to generalize the results to all dairy cows of a given species in the US.

R Code

```
###############################
## Assignment 6 - Problem 1 ##
###############################

##Read in table
> samara = read.table("samara.txt", header = T)
> str(samara)

##Plot of data
> plot(velocity~loading, pch = unclass(tree), data = samara)
> legend(0.16, 1.4, levels(samara$tree), pch = 1:3)

##Full model analysis
> lm1 = lm(velocity ~ loading * tree, samara)
> summary(lm1)

Call:
lm(formula = velocity ~ loading * tree, data = samara)

Residuals:
      Min       1Q    Median        3Q       Max
-0.120023 -0.056081 -0.002943  0.048945  0.145571

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.5521     0.2652   2.082   0.0460 *
loading              2.9831     1.1673   2.555   0.0159 *
treeTree2           -0.8514     0.3382  -2.517   0.0174 *
treeTree3           -0.3094     0.4489  -0.689   0.4960
loading:treeTree2    3.8140     1.5106   2.525   0.0171 *
loading:treeTree3    0.9002     2.3010   0.391   0.6984
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07615 on 30 degrees of freedom
Multiple R-Squared: 0.8358,     Adjusted R-squared: 0.8084
F-statistic: 30.54 on 5 and 30 DF,  p-value: 6.532e-11

##Plot of assumptions
> par(mfrow = c(2,2))
> plot(lm1)
```

```
###############################
## Assignment 6 - Problem 2 ##
###############################

##Read in table
> larch = read.table("larch.txt", header = T)

##PART A: beta3 = beta5 = 0
> lm.full = lm(height ~ nitro + phos + potas + ash, data = larch)
> lm.reduced = lm(height ~ nitro + potas, data = larch)
> anova(lm.reduced, lm.full)

##PART B: beta4 = 3*beta2 given beta0 in model
> lm.full = lm(height ~ nitro + potas, data = larch)
> lm.reduced = lm(height ~ I(potas + 3*nitro), data = larch)
> anova(lm.reduced, lm.full)
```