1. The file glue.txt contains a data set with the results of an experiment on the dry sheer strength (in pounds per square inch) of birch plywood, bonded with 5 different resin glues A, B, C, D, and E. Eight pieces of plywood were tested with each glue type. Let $\mu_A, ..., \mu_E$ be the unknown true population mean strengths for the corresponding treatments. Analyze the data with a linear model. Summarize the linear model using both the summary function in R and the anova function.

   (a) The summary function provides a p-value for each of several regression parameters. In each case, state the hypothesis that is being tested and provide an interpretation of the regression parameter in terms of the unknown population means.

   The summary table can be seen below:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  482.250      6.210  77.659  < 2e-16 ***
GlueB         -7.500      8.782  -0.854 0.398904
GlueC         28.000      8.782   3.188 0.003011 **
GlueD         35.625      8.782   4.057 0.000265 ***
GlueE         23.750      8.782   2.704 0.010494 *
```

   According to this summary, each t-value is testing whether that particular glue type is significantly different than glue A. (NOTE: we are comparing all to glue A since alphabetically glue A appears first. If we want to test a different comparison, we must recode the dummy variables.) It is found that glues C through E differ significantly than glue A in the positive direction (since the coefficients are positive) while glue B does not appear to increase or decrease the bonding strength significantly as compared to glue A.

   (b) The ANOVA table has a single p-value. State the hypothesis that is being tested here. How does this hypothesis differ form the hypotheses in part (a)?

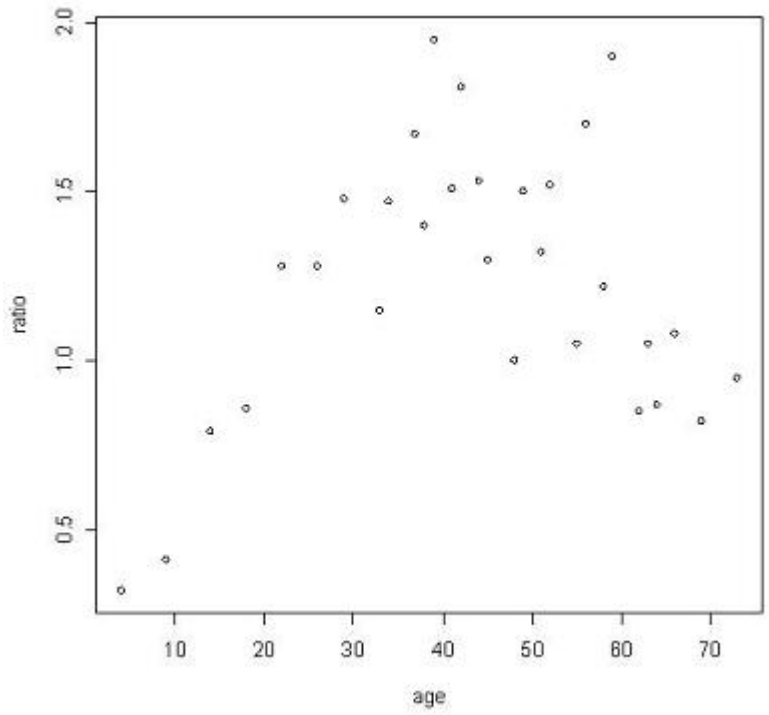
   The analysis of variance table is below:

```
Analysis of Variance Table

Response: Strength
          Df  Sum Sq Mean Sq F value   Pr(>F)
Glue       4 11179.6  2794.9  9.0597 3.93e-05 ***
Residuals 35 10797.4   308.5
```

   According to the F-statistic here, we are testing whether or not the mean bonding strength of the 5 types of glues are all equal. With a p-value of 3.93e-05, we are confident that there is a statistically significant difference in the bonding strength of the 5 glue types. This hypothesis differs from that in part (a) because here we are testing whether all are equal, whereas in part (a) we were testing if glues B through E are significantly different than glue A.

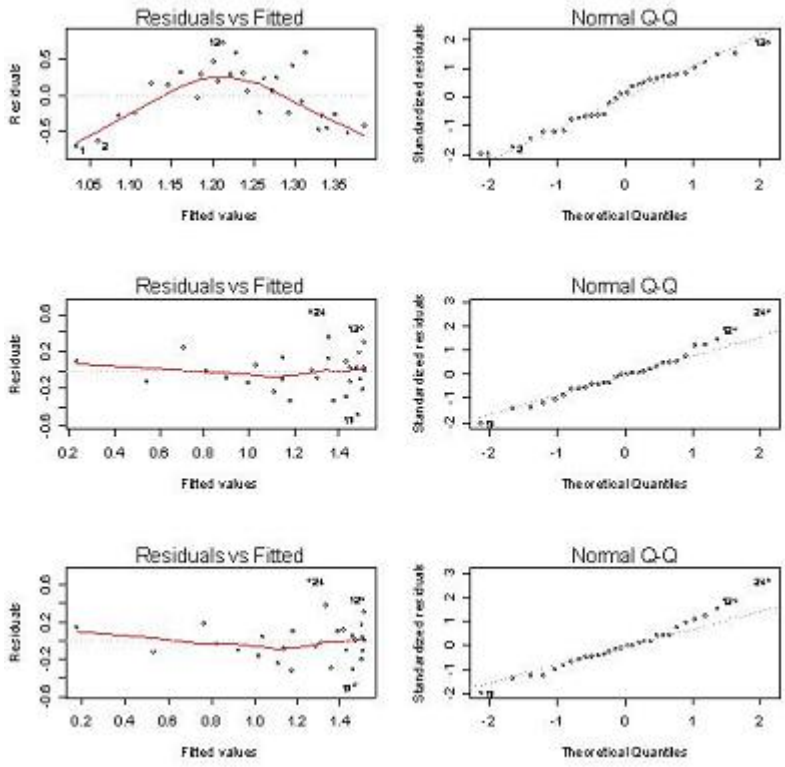
2. To study the relation between the age of oat plants and their nickel and iron content when grown in sand, the relative absorption of the two metals was measured in terms of the Ni/Fe ratio in individual plants ranging in age from 4 to 73 days. The data are in the file oat.txt with columns for age and ratio. Examine a scatterplot of these data before proceeding with the analysis.



There appears to be some sort of quadratic relationship between ratio and age.

(a) Fit a linear model, a quadratic model, and a cubic polynomial regression model for ratio as a function of age. Which of these three models is best? Explain briefly.

It appears that the quadratic model is the best. When compared with the linear model, the assumptions are better met (no longer parabolic trend in residuals versus fitted). When compared to the cubic polynomial, there was a decrease in the adjusted R-squared from 0.6204 to 0.6095. It also was seen that adding the cubic term was not significant ($p = 0.62659$). Therefore the best model appears to be the quadratic.

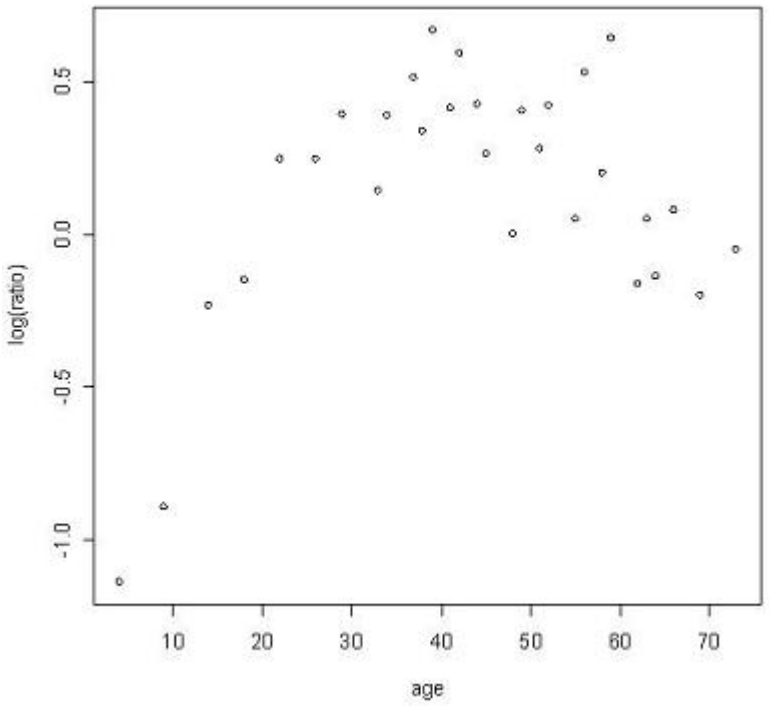


(b) Which model is preferred by AIC and which is preferred by BIC?

The AIC and BIC criteria was calculated for the four basic models (intercept only, linear, quadratic, and cubic polynomial). Based on the AIC and BIC values, the model that sufficiently minimized both was the quadratic (AIC = 6.374188, BIC = 11.97898). This matches our conclusion in part (a). Therefore the final model is found to be:

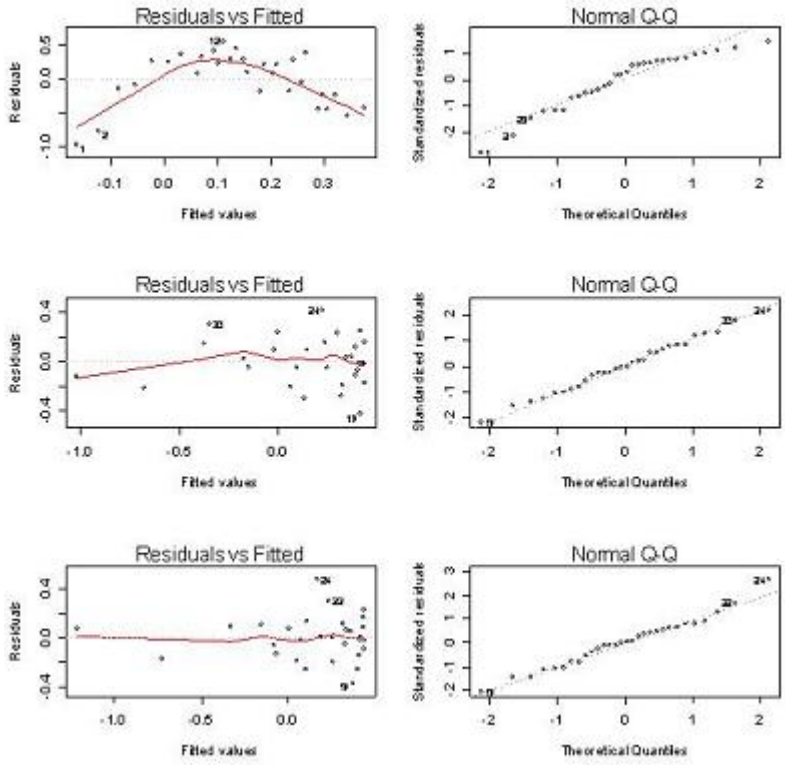
$$\text{Ni/Fe } ratio = -0.047 + 0.074 * age - 0.001 * age^2$$

(c) Repeat (a) using a log transformation of ratio as the response variable. Which of these three models is best? Explain briefly.



Again, the data appears to be quadratic. There could be potential for a cubic polynomial as the right-hand of the graph appears to level off and may increase thereafter. More data may need to be collected on older oat plants in order to verify this possibility.

The four new models were fit using log(ratio) as the response. The cubic polynomial model is chosen as best since all terms are significant, and there are no distinct violations seen in the assumption plots.



(d) Which model from (c) is preferred by AIC and which is preferred by BIC? Is the best model in (a) or the best model in (c) better? Provide reasons.

The four new models were fit using log(ratio) as the response. Based on the AIC and BIC criterion, the model that minimized both was the cubic polynomial (AIC = -9.829357, BIC = -2.823370).

$$\log(\text{Ni/Fe } ratio) = -1.679 + 0.126 * age - 0.002 * age^2 + 0.00001 * age^3$$

I prefer the quadratic untransformed model because the log transformation did not do much to help the heteroscedasticity assumption. The normal q-q plot of the residuals for the untransformed model is not bad in the middle. The untransformed model is also simpler leading to simpler inferences and interpretations.

(e) Explain why it is inappropriate to compare models in (a) and in (c) on the basis of AIC or BIC.

We cannot compare the models based on AIC and BIC because the response variable is different in each case.

R Code

```
##############################
## Assignment 5 - Problem 1 ##
##############################

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

##Review structure of data
> str(glue)
'data.frame':   40 obs. of  2 variables:
 $Glue : Factor w/ 5 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 2 2 ...
 $Strength : int  504 498 471 458 479 468 463 517 455 482 ...

##Attach data set
> attach(glue)

##Define linear model
> lm.glue = lm(Strength~Glue)


##PART A - Summary of linear model
> summary(lm.glue)

Call: lm(formula = Strength ~ Glue)

Residuals:
    Min      1Q  Median      3Q     Max
-30.250 -10.500  -2.125  10.719  34.750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  482.250      6.210  77.659  < 2e-16 ***
GlueB         -7.500      8.782  -0.854 0.398904
GlueC         28.000      8.782   3.188 0.003011 **
GlueD         35.625      8.782   4.057 0.000265 ***
GlueE         23.750      8.782   2.704 0.010494 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.56 on 35 degrees of freedom
Multiple R-Squared: 0.5087,     Adjusted R-squared: 0.4525
F-statistic:  9.06 on 4 and 35 DF,  p-value: 3.93e-05


##PART B - ANOVA
```

```
> anova(lm.glue)
Analysis of Variance Table

Response: Strength
          Df  Sum Sq Mean Sq F value   Pr(>F)
Glue       4 11179.6  2794.9  9.0597 3.93e-05 ***
Residuals 35 10797.4   308.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



##############################
## Assignment 5 - Problem 2 ##
##############################

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

##Review structure of data
> str(oat)
'data.frame':   30 obs. of  2 variables:
 $ age  : int  4 9 14 18 22 26 29 33 34 37 ...
 $ ratio: num  0.32 0.41 0.79 0.86 1.28 1.28 1.48 1.15 1.47 1.67 ...

##Attach data set
> attach(oat)

##Plot data - appears to be quadratic
> plot(oat)


##PART A - Define linear models
> lm0 = lm(ratio~1)
> lm1 = lm(ratio~age)
> lm2 = lm(ratio~age + I(age^2))
> lm3 = lm(ratio~age + I(age^2) + I(age^3))

> summary(lm1)
> summary(lm2)
> summary(lm3)

##Plots to check assumptions
> par(mfrow = c(3,2))
> plot(lm1, which = 1)
> plot(lm1, which = 2)
```

```
> plot(lm2, which = 1)
> plot(lm2, which = 2)
> plot(lm3, which = 1)
> plot(lm3, which = 2)


##PART B - Model Selection
##AIC
> AIC(lm0)
[1] 33.57922
> AIC(lm1)
[1] 33.95334
> AIC(lm2)
[1] 6.374188
> AIC(lm3)
[1] 8.095754

##BIC
> AIC(lm0,k=log(nrow(oat)))
[1] 36.38162
> AIC(lm1,k=log(nrow(oat)))
[1] 38.15693
> AIC(lm2,k=log(nrow(oat)))
[1] 11.97898
> AIC(lm3,k=log(nrow(oat)))
[1] 15.10174


##PART C - Define linear models
> plot(age,log(ratio))

> log.lm0 = lm(log(ratio)~1)
> log.lm1 = lm(log(ratio)~age)
> log.lm2 = lm(log(ratio)~age + I(age^2))
> log.lm3 = lm(log(ratio)~age + I(age^2) + I(age^3))

> summary(log.lm1)
> summary(log.lm2)
> summary(log.lm3)

> par(mfrow = c(3,2))
> plot(log.lm1, which = 1)
> plot(log.lm1, which = 2)
> plot(log.lm2, which = 1)
> plot(log.lm2, which = 2)
> plot(log.lm3, which = 1)
```

```
> plot(log.lm3, which = 2)


##PART D - Model Selection
##AIC
> AIC(log.lm0)
[1] 34.41431
> AIC(log.lm1)
[1] 32.50785
> AIC(log.lm2)
[1] -6.656999
> AIC(log.lm3)
[1] -9.829357

##BIC
> AIC(log.lm0,k=log(nrow(oat)))
[1] 37.21671
> AIC(log.lm1,k=log(nrow(oat)))
[1] 36.71145
> AIC(log.lm2,k=log(nrow(oat)))
[1] -1.052210
> AIC(log.lm3,k=log(nrow(oat)))
[1] -2.823370
```