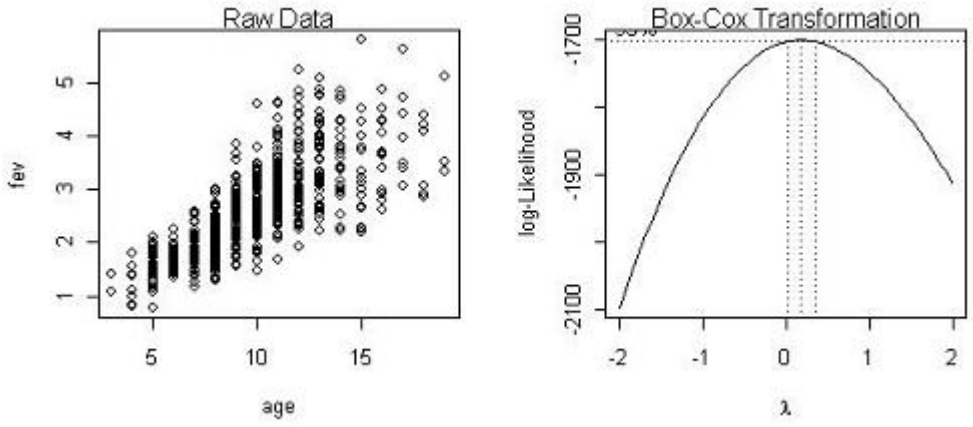
1. Problems 1 and 2 reconsider the data in the file fevdata.txt on the course web page. The first line contains the variable names which are age, fev, ht, sex, and smoke. In this assignment, we will consider models that use age (age of children, measured in years), ht (height of children in inches), and fev (forced expiratory volume, a measure of lung capacity, measured in liters).

   (a) Fit a linear model to predict FEV from age. Use the R function boxcox to find a simple power transformation (-1 = reciprocal, 0 = log, 0.5 = square root) close to the Box-Cox maximum likelihood estimate. Which simple transformation seems best?

```
> fevdata = read.table("fevdata.txt", header = T)
> str(fevdata)
'data.frame':   654 obs. of  5 variables:
 $ age  : int  9 8 7 9 9 8 6 6 8 9 ...
 $ fev  : num  1.71 1.72 1.72 1.56 1.90 ...
 $ ht   : num  57 67.5 54.5 53 57 61 58 56 58.5 60 ...
 $ sex  : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 1 1 1 ...
 $ smoke: Factor w/ 2 levels "false","true": 1 1 1 1 1 1 1 1 1 1 ...

> attach(fevdata)
> lm.age = lm(fev~age)
> par(mfrow = c(1,2))
> plot(age, fev)
> mtext("Raw Data")
> library(MASS)
> boxcox(lm.age)
> mtext("Box-Cox Transformation")
```



   For a simple transformation, the log of the response variable (FEV) appears to be the most appropriate. According to the scatter plot, as the age variable increase, so does the variance in the response variable (fev). Furthermore, according to the box-cox diagram, the 95% confidence interval for λ is closer to 0 than $\frac{1}{2}$, and therefore a log transformation on the response variable fev was chosen.

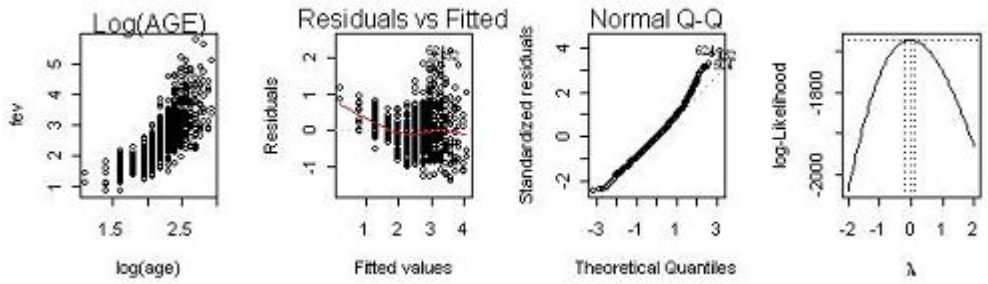
(b) Fit a linear model with the best simple transformed response predicted by age and examine the residual plot of the fit. Has the transformation improved adherence to the constant variance assumption? Is this linear model acceptable? Briefly explain why or why not.

```
> lm.age.new1 = lm(log(fev)~age)
> summary(lm.age.new1)
> par(mfrow = c(2,2))
> plot(age, log(fev))
> mtext("Transformed Data")
> plot(lm.age.new1, which = 1:2)
> plot(lm.age.new1, which = 4)
```

The constant variance assumption is better met under the log transformation of the response variable. However, there now appears to be a curve in the scatter diagram that was not originally there (see figure in part c). Although I believe this model is acceptable, it may not be optimal.

(c) Repeat parts (a) and (b) using log(age) as the explanatory variable. Which set of transformations of fev and age seems to be best to match linear model assumptions?

i.
```
> lm.age.new2 = lm(fev~log(age))
> summary(lm.age.new2)
> par(mfrow = c(1,4))
> plot(log(age), fev)
> mtext("Log(AGE)")
> plot(lm.age.new2, which = 1:2)
> boxcox(lm.age.new2)
```

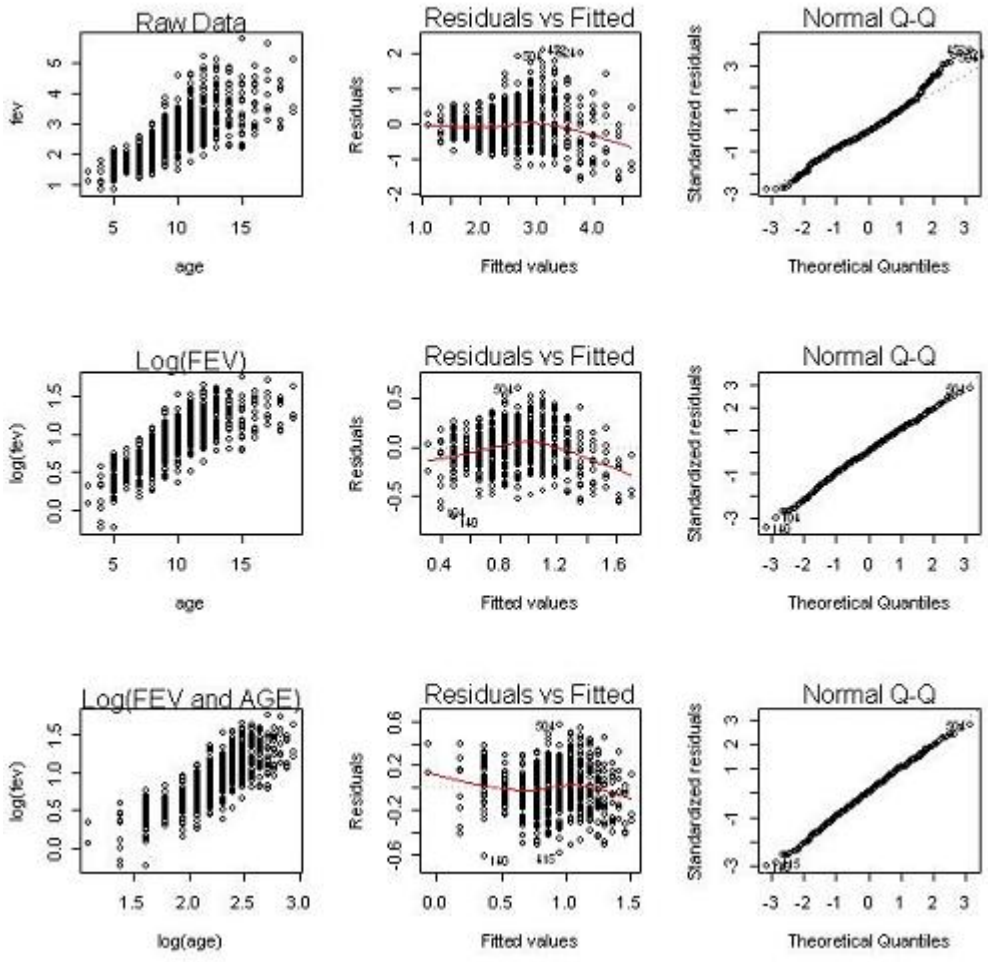


According to the Box-Cox diagram, a log transformation should be performed on the response variable (FEV). As it stands, the model is in violation of the constant variance and normality assumptions.

ii.
```
> lm.age.new3 = lm(log(fev)~log(age))
> summary(lm.age.new3)
> plot(log(age), log(fev))
> mtext("Transformed Data")
> plot(lm.age.new3, which = 1:2)
> plot(lm.age.new3, which = 4)
```

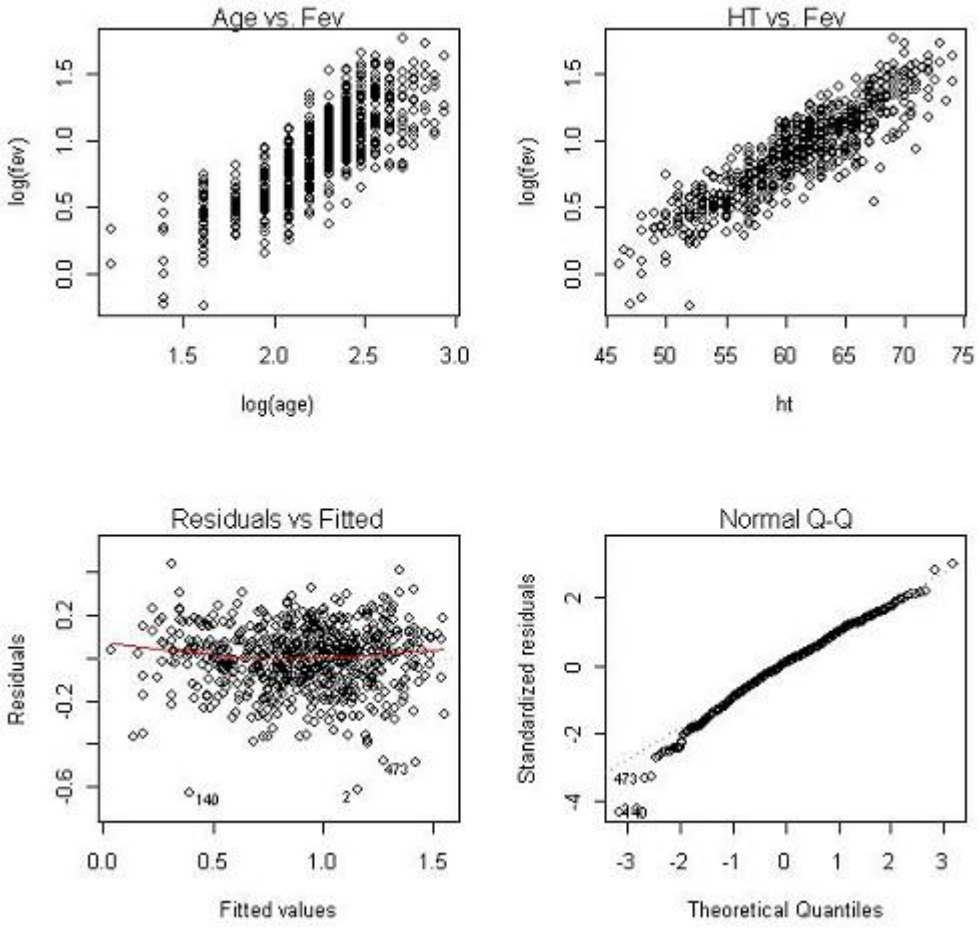
The assumptions are better met with the response variable and explanatory variable being log-transformed. According to the diagram below, the raw data is not satisfactory for satisfying the constant variance nor the normality assumption. The log of the response transform better meets the criteria, but appears to have an asymptotic curve as age increases. Therefore the final model is chosen as the best because it satisfies the model assumptions and appears to be linear with constant variance over most of the interval.

```
> par(mfrow = c(3,3))
> plot(age, fev)
> mtext("Raw Data")
> plot(lm.age, which = 1:2)
> plot(age, log(fev))
> mtext("Log(FEV)")
> plot(lm.age.new1, which = 1:2)
> plot(log(age), log(fev))
> mtext("Log(FEV and AGE)")
> plot(lm.age.new3, which = 1:2)
```



2. Add a second explanatory variable, ht, or a transformation of it, to the model found in Problem 1c. Write an equation to express the model. Find 95% confidence intervals for each model parameter (one intercept and two slopes) in the (possibly) transformed scale. As best as you can, interpret these confidence intervals in the scale of the initial measurements. (Note: There is no single set of transformations that is unambiguously best. Use your judgment!)

```
> lm.age.ht = lm(log(fev)~log(age)+ht)
> summary(lm.age.ht)
> par(mfrow = c(2,2))
> plot(log(age), log(fev))
> mtext("Age vs. Fev")
> plot(ht, log(fev))
> mtext("HT vs. Fev")
> plot(lm.age.ht, which = 1:2)
```



According to the residual plots, the model does appear to have constant variance. The normality assumption is questionable. There are a few outliers on the bottom left of the graph that may be influencing the fit. Therefore the model is found to be:

$$fev = -2.17 + 0.19 * log(age) + 0.04 * ht$$

The t-critical value is found to be:

```
> qt(0.975, 651)
[1] 1.963615
```

Therefore the 95% confidence intervals are:

- **Intercept:**

```
> -2.170548 + 1.963615*0.064153 [1] -2.044576
> -2.170548 - 1.963615*0.064153
[1] -2.29652
```

- **Age:**

```
> 0.194570 - 1.963615*0.032529
[1] 0.1306956
> 0.194570 + 1.963615*0.032529
[1] 0.2584444
```

- **HT:**

```
> 0.043314 - 1.963615*0.001784
[1] 0.03981091
> 0.043314 + 1.963615*0.001784
[1] 0.04681709
```

3. In a study of genetic variation in sugar maple, seeds were collected from native trees in the eastern United States and Canada and planted in a nursery in Wooster, Ohio. The time of leafing out of these seedlings can be related to the latitude and mean July temperature of the place of origin of the seed. The variables are X_1 = latitude, X_2 = July mean temperature, and Y = weighted mean index of leafing out time. (Y is a measure of the degree to which the leafing out process has occurred. A high value is indicative that the leafing out process is well advanced.) The data is below and in the file maple.txt on the course web page. For the following problem, check assumptions and use transformations if warranted.

   (a) Find the regression of LeafIndex on Latitude. Is latitude a useful predictor of leaf index?

```
> maple = read.table("maple.txt", header = T)
> str(maple)
> attach(maple)
> lm1 = lm(LeafIndex~Latitude)
> summary(lm1)
```

   With a p-value of 1.03e-06, the coefficient in front of the latitude variable is significantly different than zero. Therefore we are confident that the latitude variable is statistically significant in predicting leaf index.

(b) Repeat part (a) for the regression of LeafIndex on JulyTemp.

```
> lm2 = lm(LeafIndex~JulyTemp)
> summary(lm2)
```

With a p-value of 8.23e-06, the coefficient in front of the JulyTemp variable is significantly different than zero. Therefore we are confident that the variable for temperature in July is statistically significant in predicting leaf index.

(c) Find the regression of LeafIndex on Latitude and JulyTemp. Compare the results of this analysis with your results from (a) and (b). How different are the slope coefficients in each case. What best explains the differences in their values?

```
> lm12 = lm(LeafIndex ~ Latitude + JulyTemp)
> summary(lm12)
```

According to the summary output, Latitude is the only variable that is significant (p = 0.0169). This contradicts what was found earlier in the smaller models - that each variable when regressed against the response variable individually is significant. The best explanation here is that latitude and July temperature are correlated in some way. Intuitively, placement on the earth (latitude) and the time of year (specifically July) will affect the temperature.

The coefficient estimates are given in the table below. The intercept term has changed drastically while the slope terms have only changed slightly. The sign on each term has stayed the same, but the estimate has gone closer to 0 while the standard error increased. This accounts for the increase in p-values seen in the combined model, showing that the coefficients are not necessarily significantly different than zero.

	Intercept	Latitude	JulyTemp
LeafIndex $\sim$ Latitude	-1.66716	0.45369	–
LeafIndex $\sim$ JulyTemp	40.74297	–	-0.33318
LeafIndex $\sim$ Latitude + JulyTemp	13.73184	0.31393	-0.13524

(d) Find ANOVA tables for the model in part (a) (LeafIndex $\sim$ Latitude) and the model in part (c) (LeafIndex $\sim$ Latitude + JulyTemp) using the function anova on the object created using lm. What parts of the row of the ANOVA table corresponding to Latitude are the same and what parts are different? To what formal hypothesis test does the p-value in the Latitude row of each ANOVA table correspond? Why are the p-values different?

```
> anova(lm1)
Response: LeafIndex
          Df  Sum Sq Mean Sq F value    Pr(>F)
Latitude   1 104.406 104.406   37.31 1.031e-06 ***
Residuals 30  83.949   2.798
```

```
> anova(lm12)
Response: LeafIndex
          Df  Sum Sq Mean Sq F value    Pr(>F)
Latitude   1 104.406 104.406 38.4959 9.103e-07 ***
JulyTemp   1   5.298   5.298  1.9534    0.1728
Residuals 29  78.652   2.712
```

The Treatment Sum of Squares and Mean Square Treatment corresponding to Latitude are the same in both tables. This is because the amount of variation explained by Latitude does not change in this fixed system. What does change is the amount of variation that is left up to the randomness (the residuals). In the ANOVA table for the first model, July Temperature is not singled out as a separate variable, and therefore this explained variation is combined with the residual variation (83.949 = 5.298 + 78.652). However in the second model, this variation is taken into account and the residual sum of square and mean square error are smaller. Therefore the F value, which takes the Mean Square Treatment divided by the Mean Square Error, is slightly larger in the second model. This is why the p-values are different.

The hypothesis being tested in both models is whether Latitude is significant or not ($H_o : \beta_{latitude} = 0$).