

HW5 Solutions (24 points)

February 13, 2018

(8 pts) PROBLEM #1: Voltage and Insulating fluid (Exercise 22 on p80)

(a) From the times at 26kV (Y_1 is the notation used by your book) and the times at 28kV (Y_2 is the notation used by your book), two new variables were formed by taking the log-transforms, $Z_{26} = \log Y_1$ and $Z_{28} = \log Y_2$.

(b) The difference in means is $\bar{Z}_{26} - \bar{Z}_{28} = 0.295$. In words, the mean log-time at 26kV is 0.295 log(kV) LARGER than the mean log-time at 28kV!

(c) The antilog of the difference is $\exp(\bar{Z}_{26} - \bar{Z}_{28}) = \exp(0.295) = 1.34$. Because

$$\begin{aligned}\exp(\bar{Z}_{26} - \bar{Z}_{28}) &= \exp(\text{median}Z_{26} - \text{median}Z_{28}) \\ &= \exp(\text{median} \log(Y_1) - \text{median} \log Y_2) \\ &= \exp(\log(\text{median}Y_1) - \log \text{median}Y_2) \\ &= \exp(\log(\text{median}Y_1)/(\text{median}Y_2)) \\ &= (\text{median}Y_1)/(\text{median}Y_2),\end{aligned}$$

then our estimate for the true RATIO in median times is 1.34. In other words, the median time to breakdown is 34% longer for the 26kV group compared to the 28kV group.

(d) The SD = 3.36 for the breakdown times in the 26kV group which is more than a factor of 2 more than the SD = 1.14 for the breakdown times in the 28kV group. This suggests that we should use an unpooled 2-sample 95% t -CI for the true difference in mean log-transformed times:

$$0.295 \pm t_{0.975, df=2.2836} \sqrt{\text{Var}(Z_{26})/3 + \text{Var}(Z_{28})/5} = [-7.38, 7.97].$$

The $df = 2.2836$ is the Satterthwaite degrees of freedom (formula from Chapter 2 notes, see Appendix for calculation). Hence a 95% CI for the true ratio of median times is the back-transformed interval $\exp([-7.38, 7.97]) = [6.2 \times 10^{-4}, 2.88 \times 10^3]$. In other words, the evidence fails to suggest that there is a difference in the median breakdown times. We cannot say with confidence which of the two voltages yield a higher median breakdown time.

(e) A scatterplot of breakdown times vs. voltage is given in Figure 1 in the Appendix.

(f) The untransformed times did not appear normal (according to Box-Cox) which is why we transformed the times. We must assume normality of the log-transformed breakdown times because we only have a few data in each group, much less than 30.

(g) Box-Cox transformations of the data are of the form

$$Z = \begin{cases} Y^\lambda & \text{for } \lambda \neq 0 \\ \log(Y) & \text{for } \lambda = 0 \end{cases} .$$

Figure 2 shows a 95% CI for λ . Because this CI contains $\lambda = 0$, then the book's choice to use a log-transform of the data (that corresponds to $\lambda = 0$) is consistent with Box-Cox.

(16 pts) PROBLEM #2: A report investigating income differences associated with education levels (Exercise 25 on p147)

1 Introduction

A random sample of $n = 2584$ Americans with paying jobs in 2005 were selected from the 1979 National Longitudinal Survey of Youth (NLSY). They were asked about their income and the number of years of education in 2006. Education was simplified into 5 distinct levels: less than 12 years, 12 years (high school diploma), 13-15 years (some college), 16 years (bachelors degree) or more than 16 years. The question of interest is whether there is an association between income and educational level.

2 Statistical Methods Used

The data are summarized by medians, IQRs, means and SDs in Table 1. Side-by-side boxplots were used to compare the (untransformed) incomes of individuals based on their education level in Figure 3. Due to a few individuals with large incomes, the incomes are severely right-skewed with variability increasing as the median and mean income increase. Most individuals have incomes less than \$100,000. Figure 3 makes it hard to see the spread in these majority of "common" incomes. The \log_{10} -transformed incomes are displayed in a boxplot in Figure 4. As suggested by Display 3.8 in the textbook, the latter figure shows that on the log scale, the data are much more symmetric and also have about the same spread. Also, the spread in the majority of incomes that are less than $\$100,000 = 10^5$ are better represented.

First, I tried to test the following hypotheses (where μ_i is the true MEAN income for NLSY people with an education level of i):

$$H_0 : \mu_{<12} = \mu_{12} = \mu_{13-15} = \mu_{16} = \mu_{>16} \text{ vs } H_a : \mu_i \neq \mu_j \text{ for some } i \text{ and } j.$$

An ANOVA was fit to the (untransformed) incomes and the residual plots in Figure 5 were used to assess the constant variance assumption of the ANOVA model. There is a clear increasing trend in the spread of the residuals as the education level increases in the residual vs. fits plot. Hence constant variance appears to be violated. There are 2 more assumptions: we know that the data within each group

are independent because the data are from a RS; and we know that the groups are independent because the random sample was taken from the NLSY. We do not need to worry about normality even though there is severe right skew evident in the histogram, boxplot, and normal probability plots of the residuals that indicate that the data in each group are not normal. This is because we have many data ($n > 30$) in each education group, hence the data need not be normal. \end{enumerate}

To address the non-constant variance, the incomes were \log_{10} -transformed. Now the hypotheses to be tested are (where $\tilde{\mu}_i$ is the true MEDIAN income for NLSY people with an education level of i):

$$H_0 : \tilde{\mu}_{<12} = \tilde{\mu}_{12} = \tilde{\mu}_{13-15} = \tilde{\mu}_{16} = \tilde{\mu}_{>16} \text{ vs } H_a : \tilde{\mu}_i \neq \tilde{\mu}_j \text{ for some } i \text{ and } j.$$

An ANOVA was fit to these \log_{10} -transformed incomes and the residual plots in Figure 7 were used to assess the constant variance assumption of the ANOVA model. There is equal spread in the residuals as the education level increases in the residual vs. fits plot so the constant variance assumption appears to be satisfied. The ANOVA table for this analysis is presented in Table 2.

Box-Cox transformations of the data are of the form

$$Z = \begin{cases} Income^\lambda & \text{for } \lambda \neq 0 \\ \log(Income) & \text{for } \lambda = 0 \end{cases} .$$

A 95% CI for λ (Figure 6) suggested a CI close to zero (although it did not contain 0), so Box-Cox did not suggest a log-transform but instead suggested a transform $Income^{1/3}$. Fitting an ANOVA to these third-root incomes give results (see R Output in Appendix if interested in this) similar to the ANOVA fit to the log-incomes. This model does appear to fit the $Income^{1/3}$ data better than the fit to the log-Income data (Figure 8).

3 Summary of Statistical Findings

The data suggest that there is a difference in the median income depending on educational level ($F = 62.9$, p -value < 0.00005). It appears median income increases as the number of years of education increase. We will assess this statement with statistical significance in HW6.

4 Scope of Inference

Because these data are from a random sample from Americans who took the NLSY in 1979, then these results suggest that for all Americans who took the NLSY in 1979, median income is associated with education. These observational study data do not suggest that education caused the difference in mean income.

5 Appendix

5.1 Tables

	Median	IQR	Mean	SD	n
<12	23500.00	23000.00	28301.45	21021.90	136.00
12	31000.00	28025.00	36864.90	29369.73	1020.00
13-15	38000.00	34000.00	44875.96	33913.54	648.00
16	56500.00	57000.00	69996.97	64256.80	406.00
>16	60500.00	56000.00	76855.46	65428.29	374.00

Table 1: Summary of incomes by 5 education levels

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Educ	4	41.05	10.26	62.87	0.0000
Residuals	2579	421.00	0.16		

Table 2: Results of ANOVA applied to log10-transformed incomes as a function of Education level

5.2 Figures

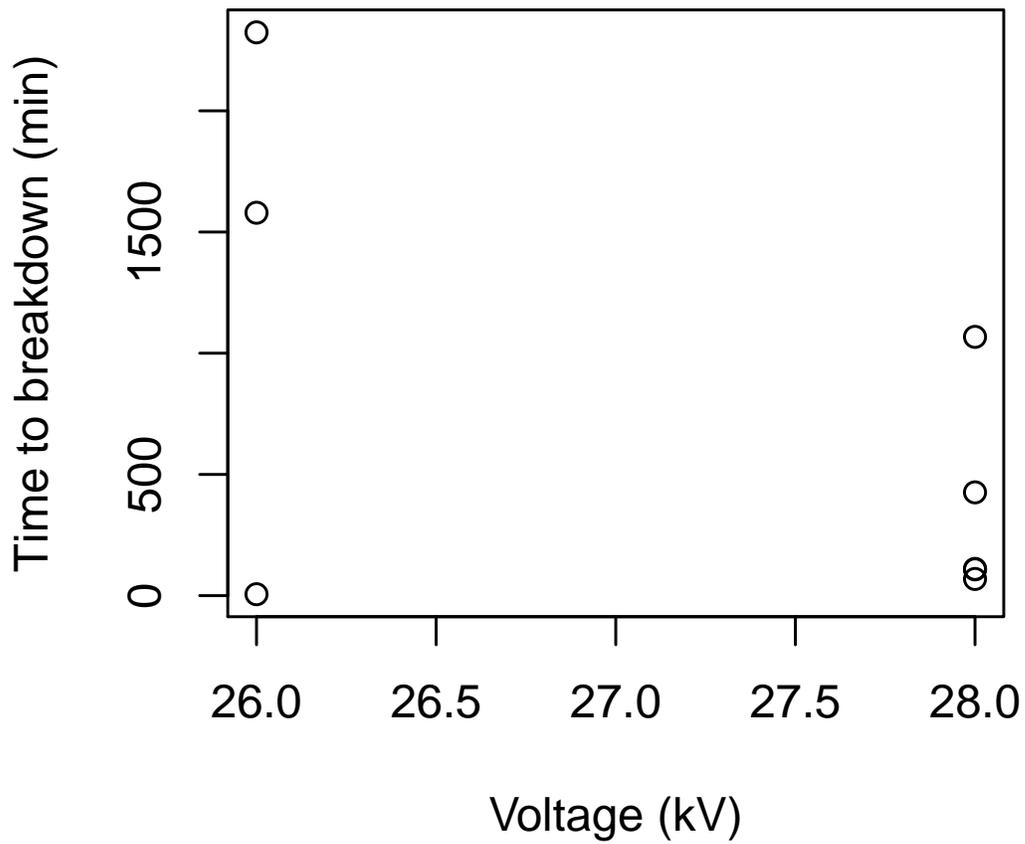


Figure 1: Breakdown times by voltage. This plot shows that the mean and median breakdown time is lower for the 28kV group in this study. It also shows that the 26kV group has more variability, suggesting that you should use an unpooled t -tool for data analysis.

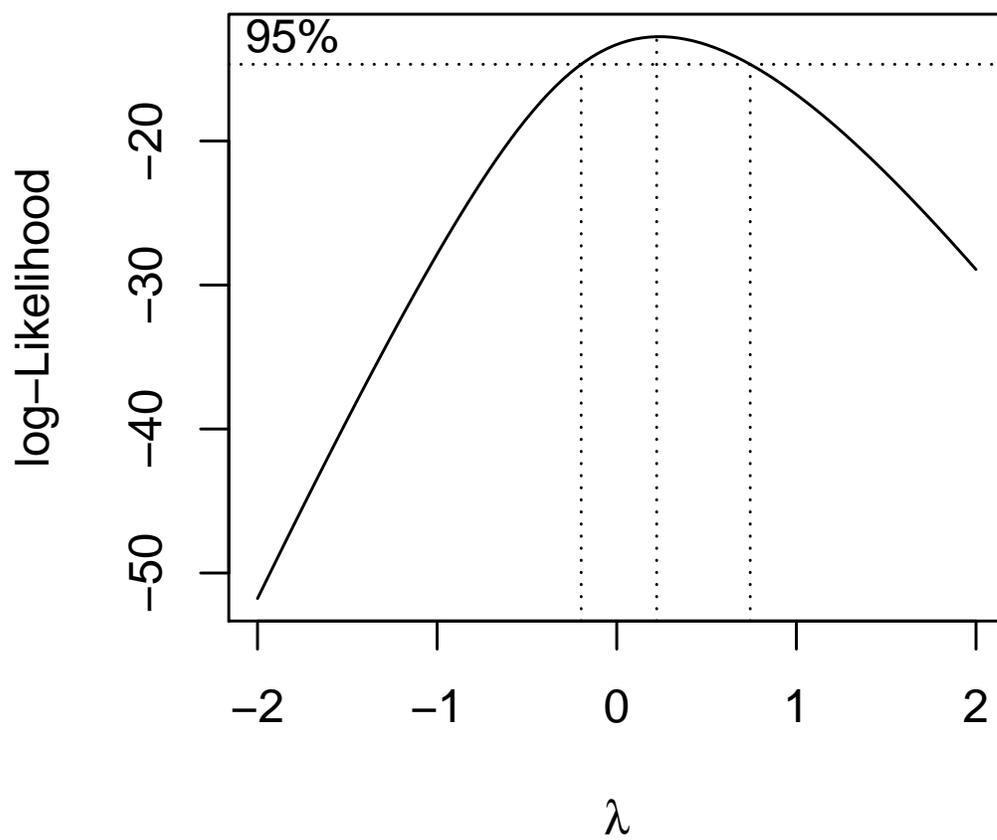


Figure 2: Determining the Box-Cox transform for problem 1. This supports your book's choice to log-transform the data.

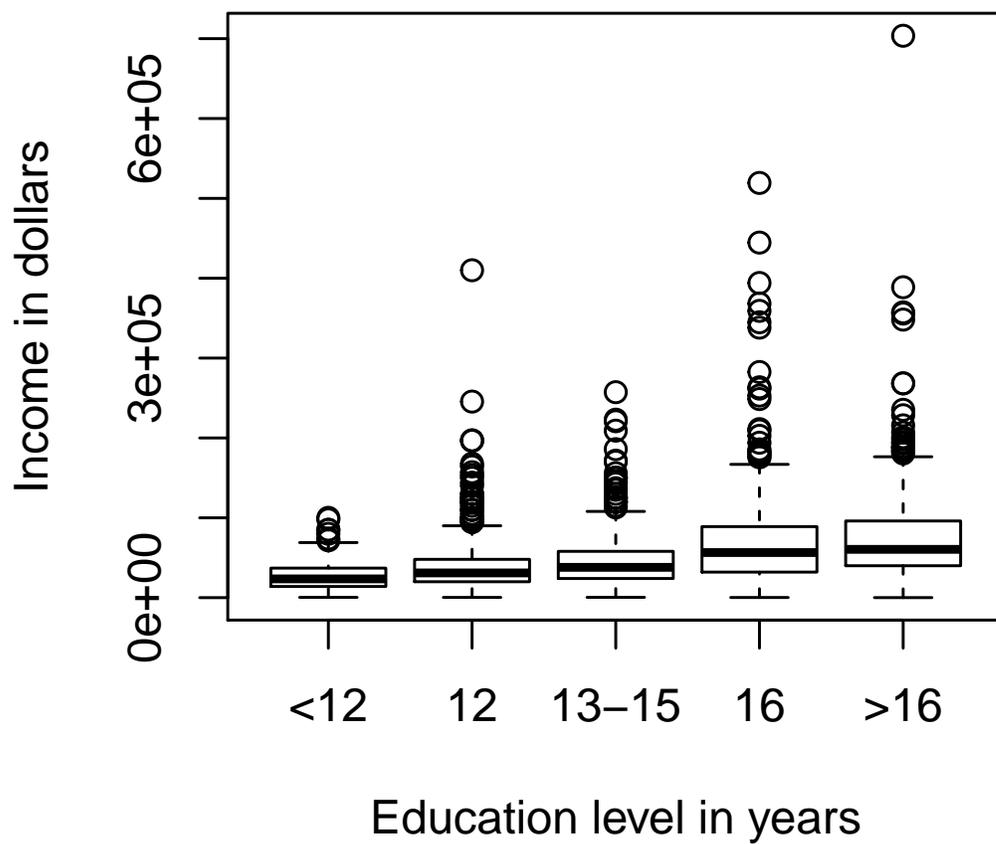


Figure 3: Incomes by Education level. The data have differing variance for the different education levels.

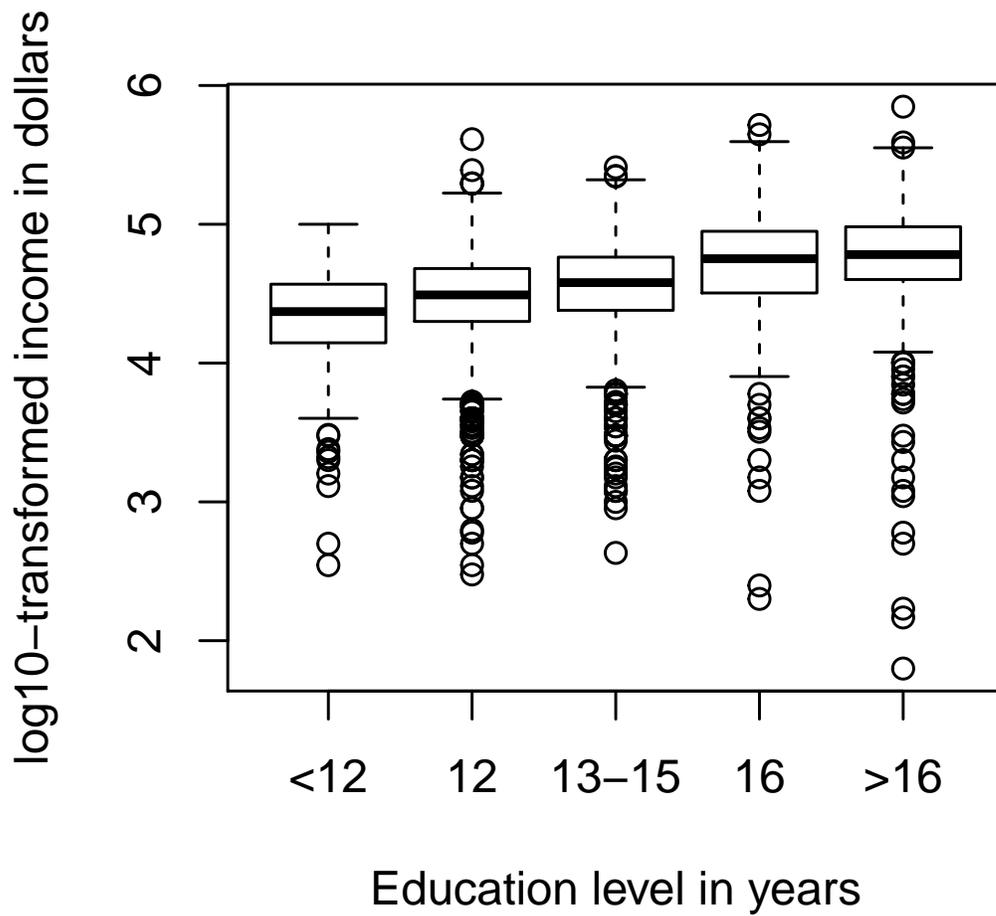


Figure 4: Incomes by Education level. The data appear to have about the same variance on the log scale.

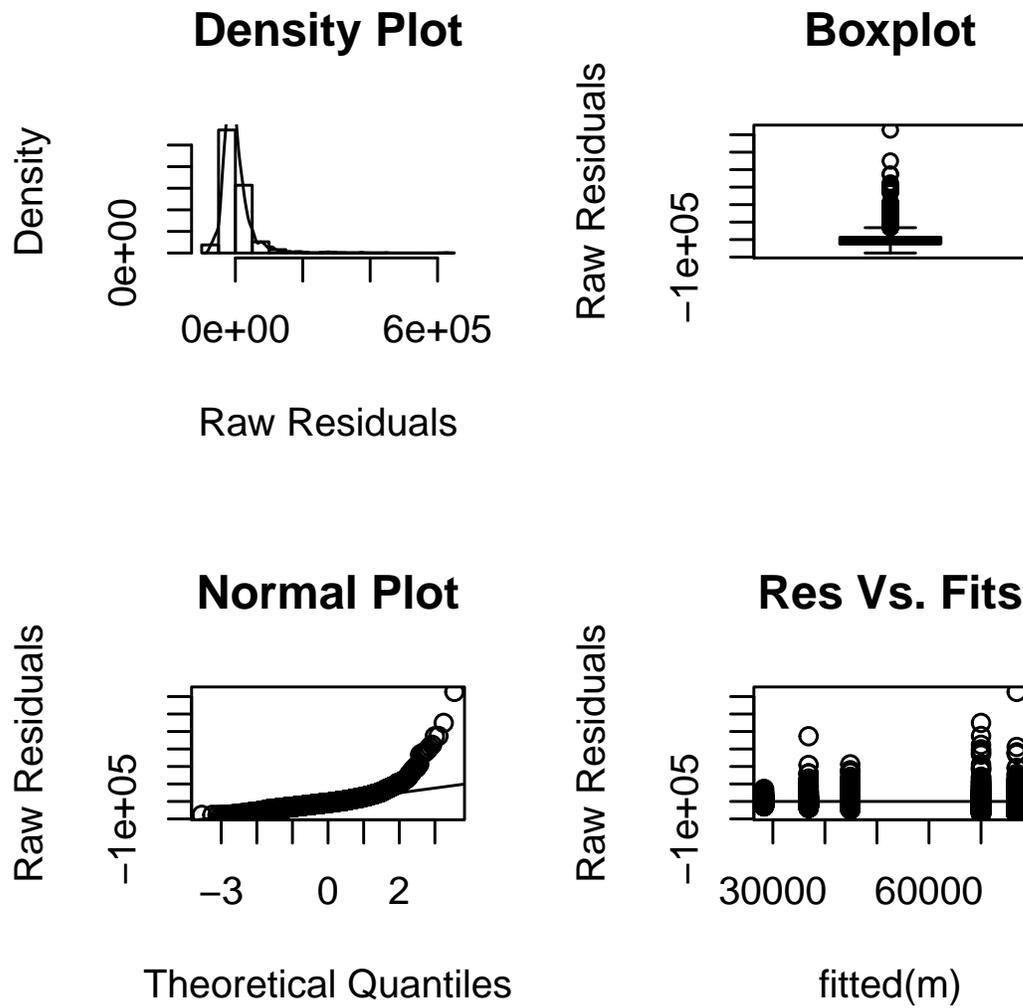


Figure 5: Assessing fit of the ANOVA to the untransformed incomes. The variance increases as the mean income does!

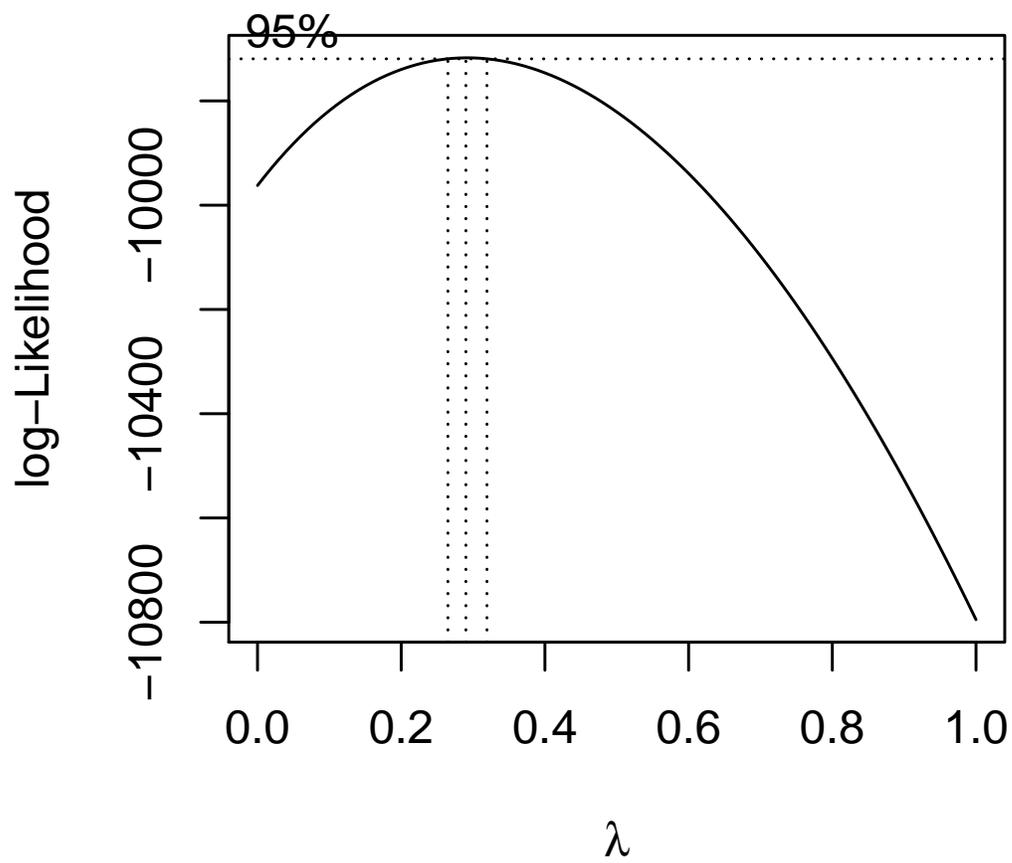


Figure 6: Determining the Box-Cox transform for problem 2

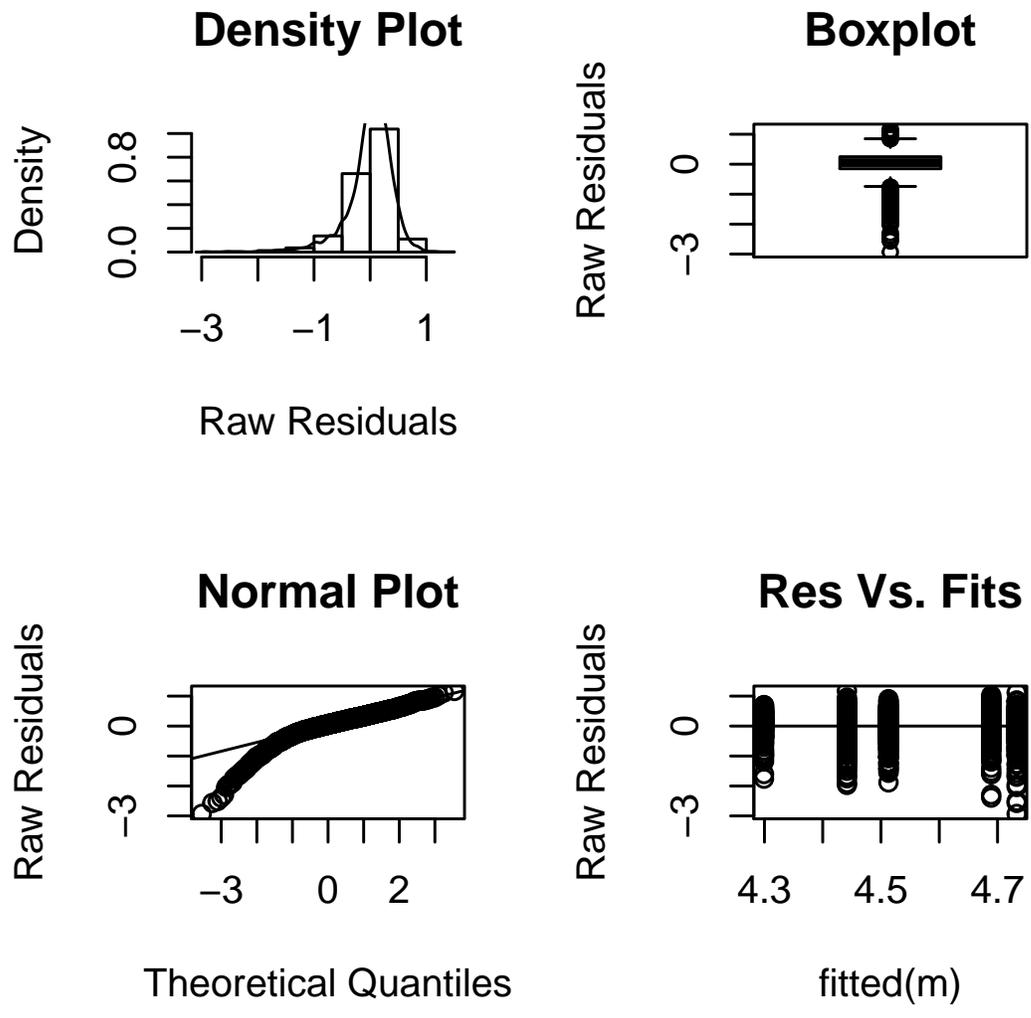


Figure 7: Assessing fit of the ANOVA to the log-transformed incomes

5.3 R-code

```
# PROBLEM 1
```

```
# The boxcox() function lives here  
library(MASS)
```

```
# Import data by hand  
V=c(26,26,26,28,28,28,28,28)  
V
```

```
## [1] 26 26 26 28 28 28 28 28
```

```
T=c(5.79,1579.52,2323.7,68.8,108.29,110.29,426.07,1067.6)  
T
```

```
## [1] 5.79 1579.52 2323.70 68.80 108.29 110.29 426.07 1067.60
```

```
# (a) log-transform the data
```

```
#The ln-transformed data at 26kV is Z26 and the ln-transformed data at 28kV is Z28  
Z26 = log(T[V==26]) # Yes, this is a natural log!  
data.frame(T[V==26],Z26) # Look at the data
```

```
## T.V....26. Z26  
## 1 5.79 1.756132  
## 2 1579.52 7.364876  
## 3 2323.70 7.750916
```

```
Z28 = log(T[V==28])  
data.frame(T[V==28],Z28) # Look at the data
```

```
## T.V....28. Z28  
## 1 68.80 4.231204  
## 2 108.29 4.684813  
## 3 110.29 4.703113  
## 4 426.07 6.054604  
## 5 1067.60 6.973168
```

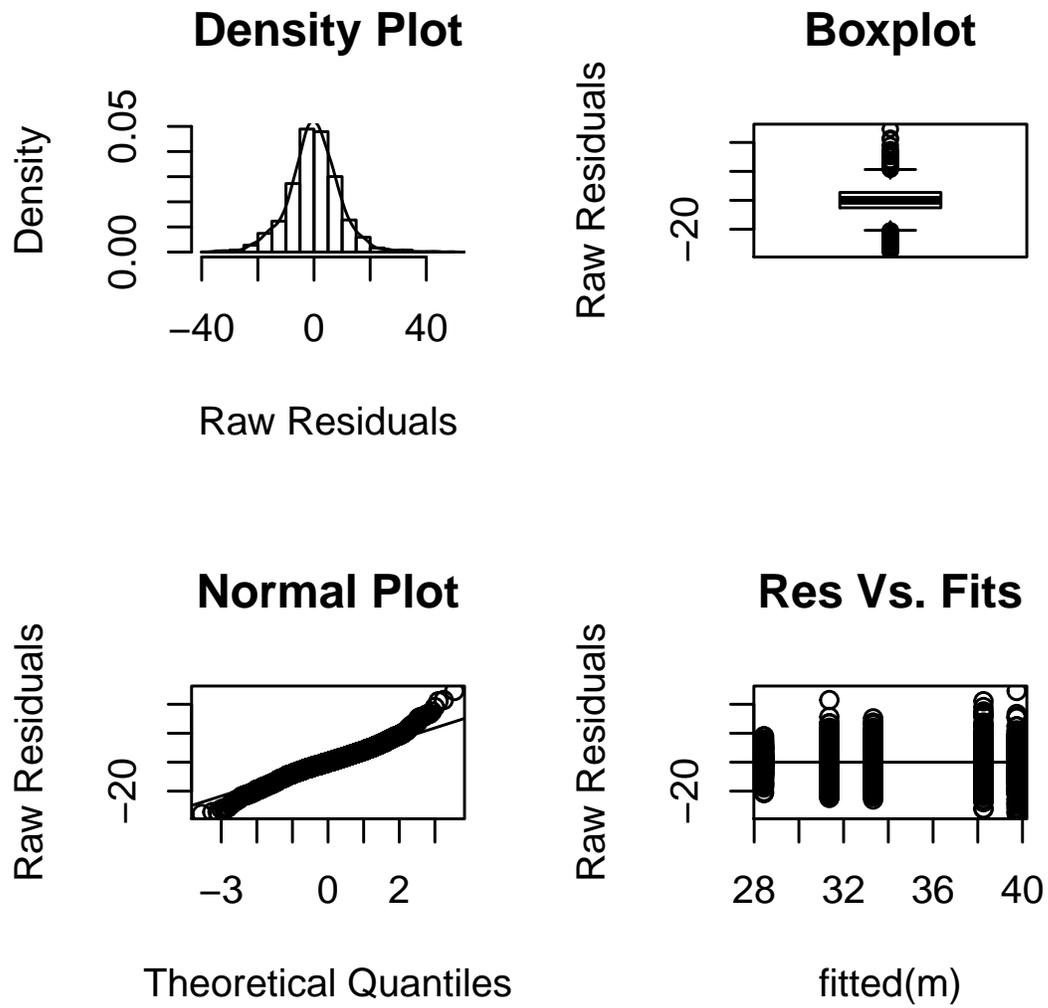


Figure 8: Assessing fit of the ANOVA to the Box-Cox transformed incomes

```
# (b) Take the difference in the means of the logs
```

```
mean(Z26) - mean(Z28)
```

```
## [1] 0.2945945
```

```
# (c) Back-transform to get the ratio of medians on the original time scale
```

```
exp(mean(Z26) - mean(Z28))
```

```
## [1] 1.342582
```

```
# (d) Build a t-CI for the ratio of means
```

```
# check the equal variance assumption
```

```
sd(Z26)
```

```
## [1] 3.355207
```

```
sd(Z28)
```

```
## [1] 1.144733
```

```
# One SD is more than 2 times the other so a 2-sample pooled t-test is not recommen
```

```
# Here's the Satterthwaite df for the unpooled test:
```

```
(var(Z26)/3 + var(Z28)/5)^2/((var(Z26)/3)^2/2 + (var(Z28)/5)^2/4)
```

```
## [1] 2.288904
```

```
# A unpooled t-CI for the differences in the mean log-transformed times
```

```
mean(Z26) - mean(Z28) + c(-1,1)*qt(.975,2.2836)*sqrt(var(Z26)/3 + var(Z28)/5)
```

```
## [1] -7.377441 7.966630
```

```
# A unpooled t-CI for the ratio of medians
```

```
exp(mean(Z26) - mean(Z28) + c(-1,1)*qt(.975,2.2836)*sqrt(var(Z26)/3 + var(Z28)/5))
```

```
## [1] 6.251990e-04 2.883124e+03
```

```
# Check my CI with with t.test()
```

```
t.test(Z26,Z28)
```

```
##
## Welch Two Sample t-test
##
## data: Z26 and Z28
## t = 0.14703, df = 2.2836, p-value = 0.8951
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -7.377556 7.966745
## sample estimates:
## mean of x mean of y
## 5.623975 5.329380
```

```
# (e) Scatterplot
plot(V,T,xlab="Voltage (kV)",ylab="Time to breakdown (min)")
```

```
# (g) Box-Cox transform
boxcox(T ~ V)
```

```
# PROBLEM 2
# Import the data
library(Sleuth3)
d = ex0525
summary(d)
```

```
##      Subject      Educ      Income2005
## Min.   :    2   12 :1020   Min.   :    63
## 1st Qu.: 1586  13-15: 648   1st Qu.: 23000
## Median : 3108   16  : 406   Median : 38231
## Mean   : 3494  <12 : 136   Mean   : 49417
## 3rd Qu.: 4636  >16 : 374   3rd Qu.: 61000
## Max.   :12140                Max.   :703637
```

```
# Simple summaries of the data
d$Educ = factor(as.character(d$Educ),levels = c("<12","12","13-15","16",>16"))
Mean = tapply(d$Income2005,d$Educ,mean)
SD = tapply(d$Income2005,d$Educ,sd)
Median = tapply(d$Income2005,d$Educ,median)
IQR = tapply(d$Income2005,d$Educ,IQR)
n = tapply(d$Income2005,d$Educ,length)
sum.tab = cbind(Median, IQR,Mean,SD,n)
```

```

# Make some plots of the data
boxplot(Income2005 ~ Educ,xlab="Education level in years",
        ylab="Income in dollars",data=d)
boxplot(log10(Income2005) ~ Educ,xlab="Education level in years",
        ylab="log10-transformed income in dollars",data=d)

# Get diagnostic checker
source("http://www.math.montana.edu/parker/courses/STAT411/diagANOVA.r")

# Conduct ANOVA on untransformed data
m = lm(Income2005 ~ Educ,data=d)
anova(m)

```

```

## Analysis of Variance Table
##
## Response: Income2005
##           Df      Sum Sq   Mean Sq F value    Pr(>F)
## Educ         4 6.8824e+11 1.7206e+11  89.613 < 2.2e-16 ***
## Residuals 2579 4.9517e+12 1.9200e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

# Check residuals
diagANOVA(m)

# See what transform that Box-Cox suggests
boxcox(Income2005 ~ Educ,lambda=seq(0,1,.01),data=d)

# log-transform the data and refit the model
m.log = lm(log10(Income2005) ~ Educ,data=d)
sum.aov = anova(m.log)
sum.aov

```

```

## Analysis of Variance Table
##
## Response: log10(Income2005)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Educ         4  41.05 10.2630   62.87 < 2.2e-16 ***
## Residuals 2579 421.00  0.1632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

# Check residuals of the new model
diagANOVA(m.log)

# Box-Cox transform the data and refit the model
m.bc = lm(Income2005^(1/3) ~ Educ,data=d)
anova(m.bc)

## Analysis of Variance Table
##
## Response: Income2005^(1/3)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Educ         4  31286   7821.4   93.133 < 2.2e-16 ***
## Residuals 2579 216586     84.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Check residuals of the new model
diagANOVA(m.bc)

```