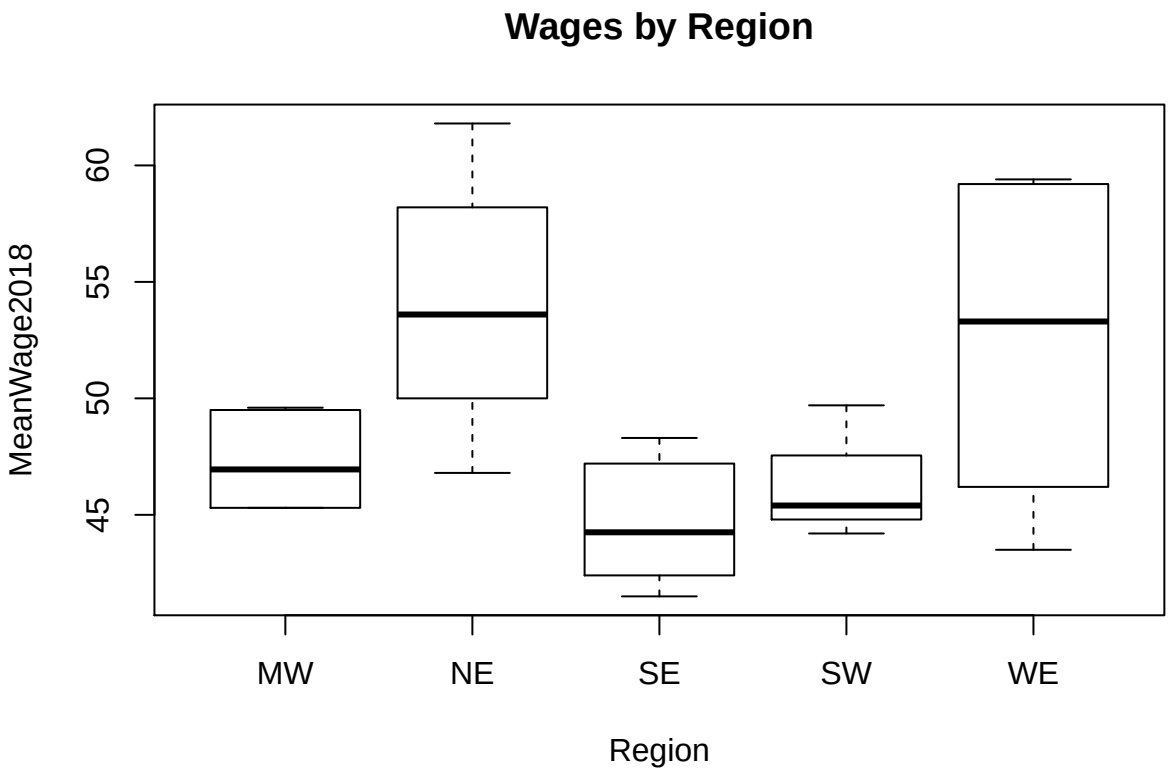
Fall 2022 Stat 407/507 Exam 1 solutions

Problem 1a

```
StateWages <- read.table("https://webpages.uidaho.edu/~chrisw/stat507live/WageStatesF22Exam1.txt",heade

StateWages$Region <- as.factor(StateWages$Region)

boxplot(MeanWage2018 ~ Region, data=StateWages,main="Wages by Region")
```



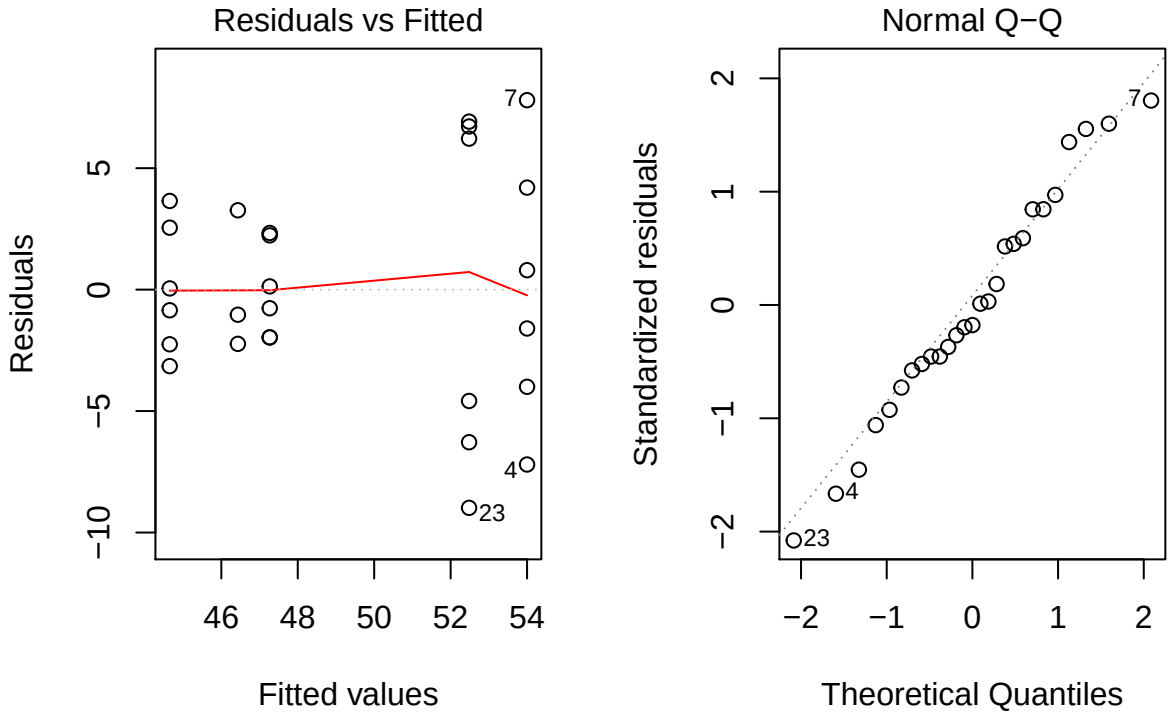
The boxplot shows that the medians of NE and WE are clearly greater than for the other regions, so we will likely reject H_0. The within-group variability also seems different among groups, suggesting a violation of the equal-variance assumption.

Problem 1b, c

```
StateWages.lm1 <- lm(MeanWage2018 ~ Region, data=StateWages)

anova(StateWages.lm1)
```

```
## Analysis of Variance Table
##
## Response: MeanWage2018
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Region     4 372.46  93.116  4.1524 0.01178 *
## Residuals 22 493.34  22.425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#summary(StateWages.lm1)

par(mfrow=c(1,2))
plot(StateWages.lm1, which=c(1,2))
```



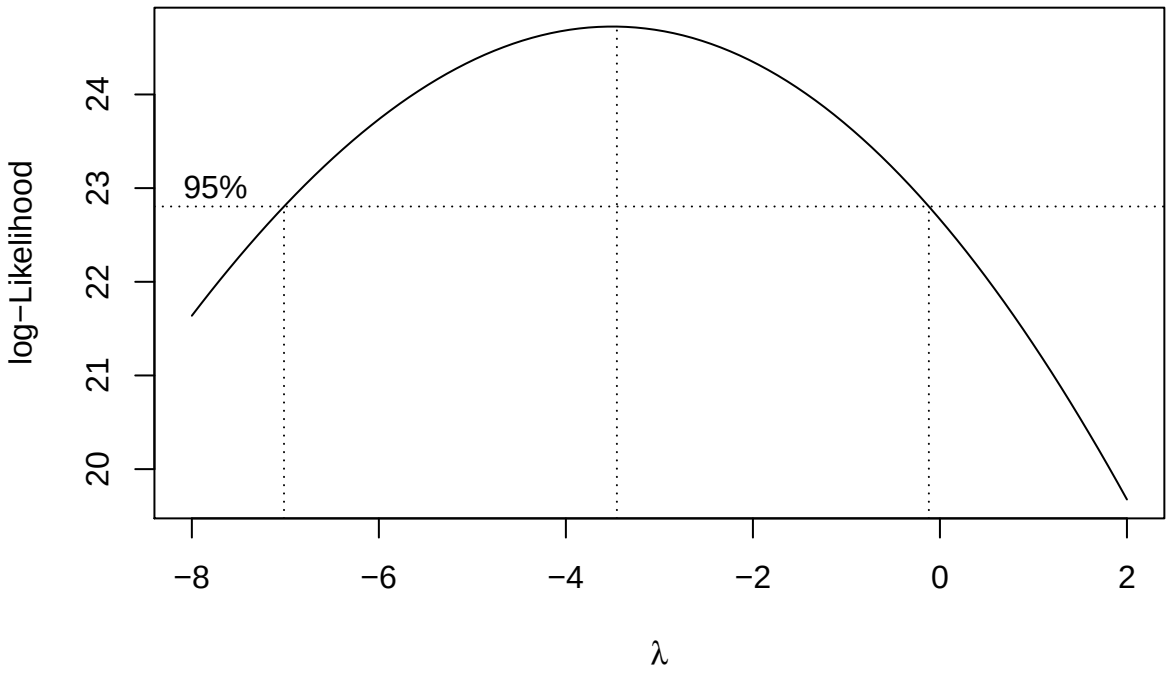

```
par(mfrow=c(1,1))
```

The total df are $4 + 22 = 26$ and the total SS are $372.46 + 493.34 = 865.8$. With an F value of $F = 4.15$ on 4 and 22 degrees of freedom with a P value of $P = .012$, we reject the null hypothesis of $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ and conclude that the mean wages differ by region, if the model assumptions are valid.

However, the residual plots show a problem. Although the normal plot shows points mostly close to the reference line, the residual by predicted plot shows a megaphone pattern indicating that the homogeneous variance assumption is doubtful.

Problem 1d

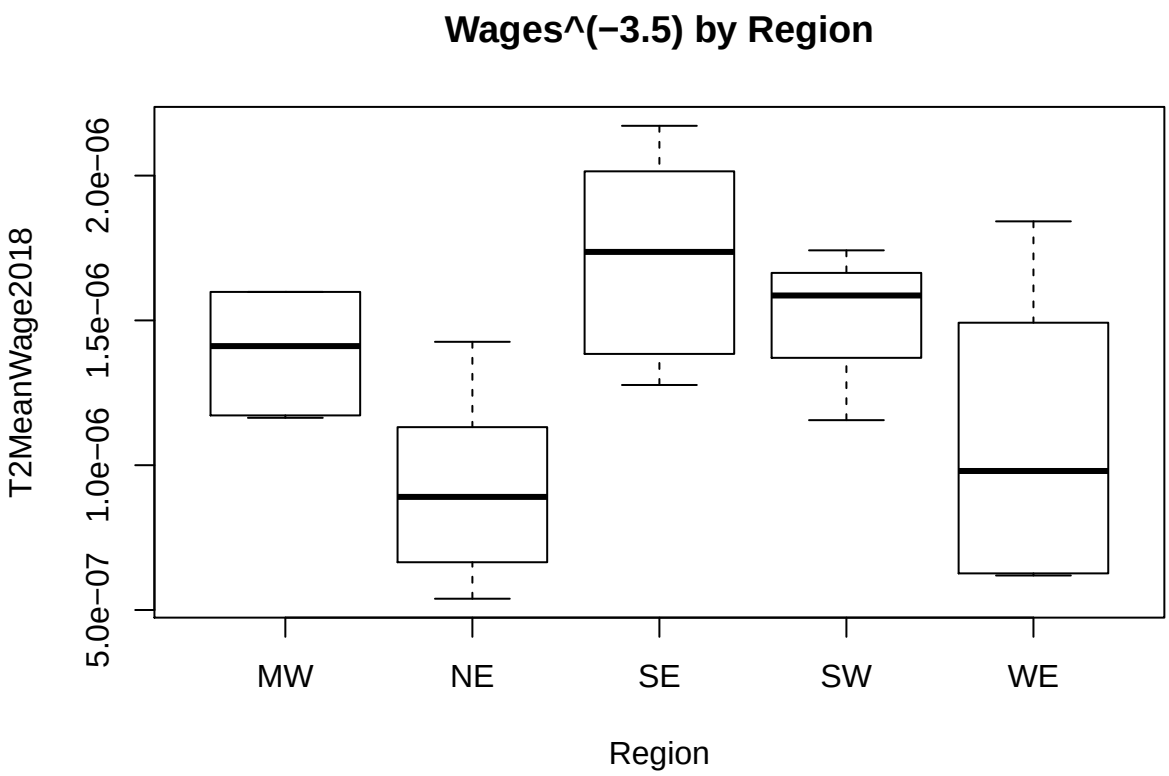
```
library(MASS)

# boxcox(MeanWage2018 ~ Region, data=StateWages,lambda = seq(-2.00, 2.00, length = 50))

boxcox(MeanWage2018 ~ Region, data=StateWages,lambda = seq(-8.00, 2.00, length = 50))
```



After a few tries to make sure that the entire 95% confidence interval is visible, we see that the confidence interval for the power transformation parameter λ is approximately from -7 to just below 0. Such a wide range for the confidence interval indicates that there is not strong evidence for a particular transformation. Since the max/min ratio for the data is not large, we might expect that a power transformation would not be very effective. Since the value 1 is not in the interval we can reject the null hypothesis of $H_0 : \lambda = 1$. The optimal value of λ is -3.5 so we will use that as our transformation.

Problem 1e

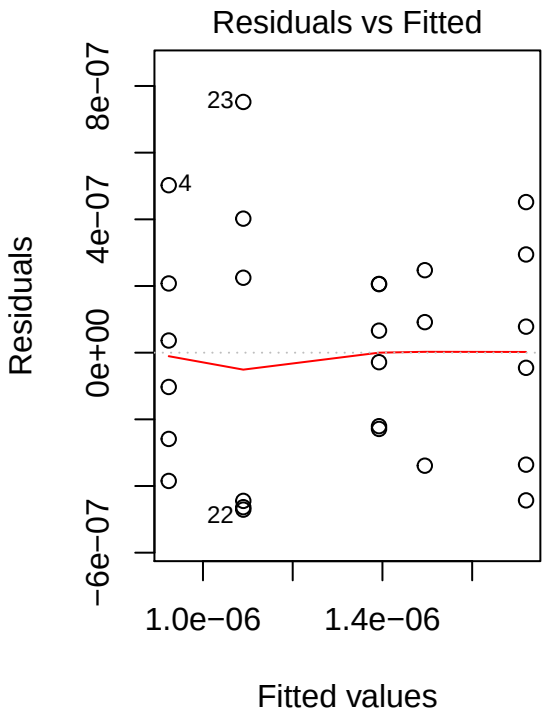
```r
StateWages$T2MeanWage2018 <- StateWages$MeanWage2018^(-3.5)

boxplot(T2MeanWage2018 ~ Region, data=StateWages,main="Wages^(-3.5) by Region")
```

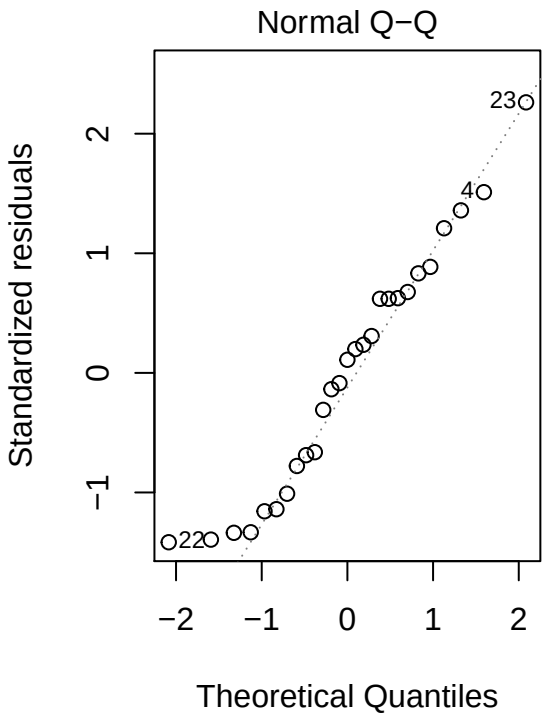



```r
StateWages.lm4 <- lm(T2MeanWage2018 ~ Region, data=StateWages)

anova(StateWages.lm4)
```

```
## Analysis of Variance Table
##
## Response: T2MeanWage2018
##           Df     Sum Sq    Mean Sq F value   Pr(>F)
## Region     4 2.3373e-12 5.8431e-13  4.4053 0.009114 **
## Residuals 22 2.9181e-12 1.3264e-13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow=c(1,2))
plot(StateWages.lm4, which=c(1,2))
```






```
par(mfrow=c(1,1))
```

After transformation, the boxplot still shows group differences (now inverted because of the inverse transformation). The null hypothesis is still rejected with fairly similar F and P values. The residual by predicted plot looks a bit better as the megaphone shape is gone but the normal plot looks worse. An argument could be made to use the transformation because of a better variance pattern, but an argument can also be given to present the data on the original scale since the results do not change.

Problem 1e

```
# Some calculations:

# grand population mean
summary(c(47,57,45,46,55))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      45      46      47      50      55      57
```

```
# sum of alpha^2
4*var(c(47,57,45,46,55))
```

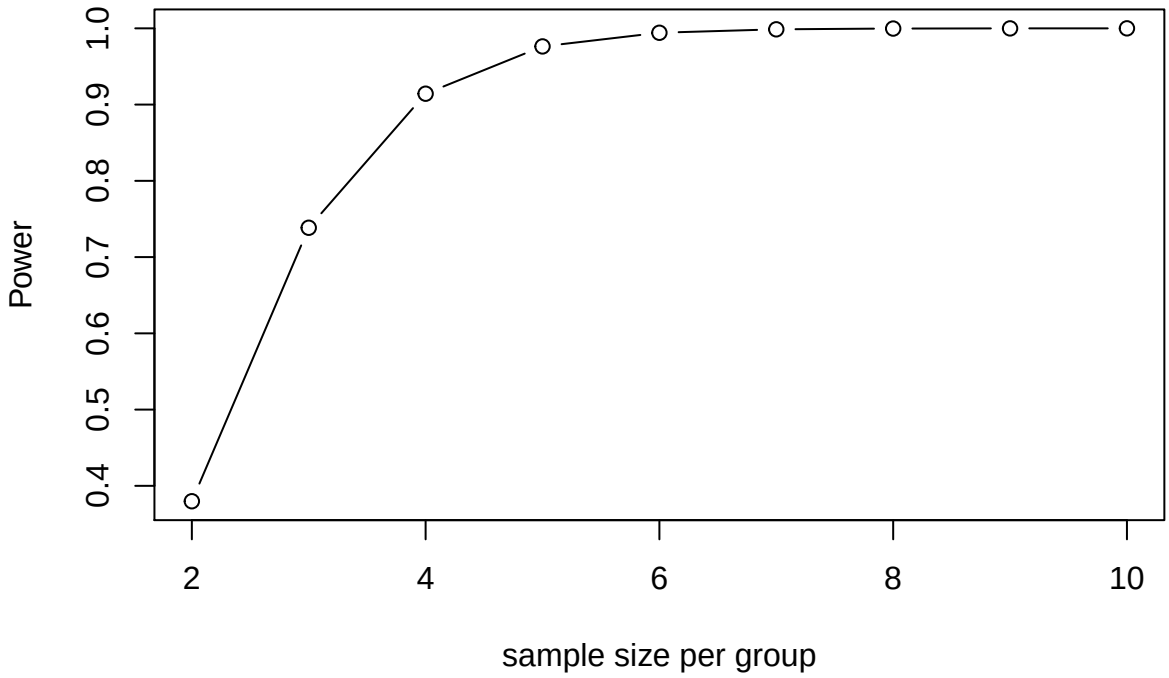
```
## [1] 124
```

```r
# MSE of original data = 22.4

# ratio value
4*var(c(47,57,45,46,55)) / 22.4
```

```
## [1] 5.535714
```

```r
powercr <- function(ngrp,rat,alpha,n1,nlast,ninc)
     {
        n       <- seq(n1,nlast,by=ninc)
        power   <- numeric(length(n))
        ndf     <-  ngrp -1
        ddf         <-  ngrp*(n-1)
        fcr         <-  qf(1-alpha,ndf,ddf)
        lambda      <-  n*rat
       # print(c(n,ddf,fcr,lambda))
        power    <-  1 - pf(fcr,ndf,ddf,lambda)
        print(cbind(n,ddf,lambda,fcr,power))
        plot(n,power,type="b",xlab="sample size per group",ylab="Power")
}

powercr(5,5.54,.05,2,10,1)
```

```
##        n ddf lambda      fcr     power
##  [1,]  2   5  11.08 5.192168 0.3798115
##  [2,]  3  10  16.62 3.478050 0.7384860
##  [3,]  4  15  22.16 3.055568 0.9142997
##  [4,]  5  20  27.70 2.866081 0.9762981
##  [5,]  6  25  33.24 2.758710 0.9942137
##  [6,]  7  30  38.78 2.689628 0.9987179
##  [7,]  8  35  44.32 2.641465 0.9997373
##  [8,]  9  40  49.86 2.605975 0.9999495
##  [9,] 10  45  55.40 2.578739 0.9999908
```



The third line of the printed values corresponds to $n = 4$ replicates per group, and it is the first line with power greater than 80% (power = .91). Thus we would need 4 states per region for a total sample size of 20 states.