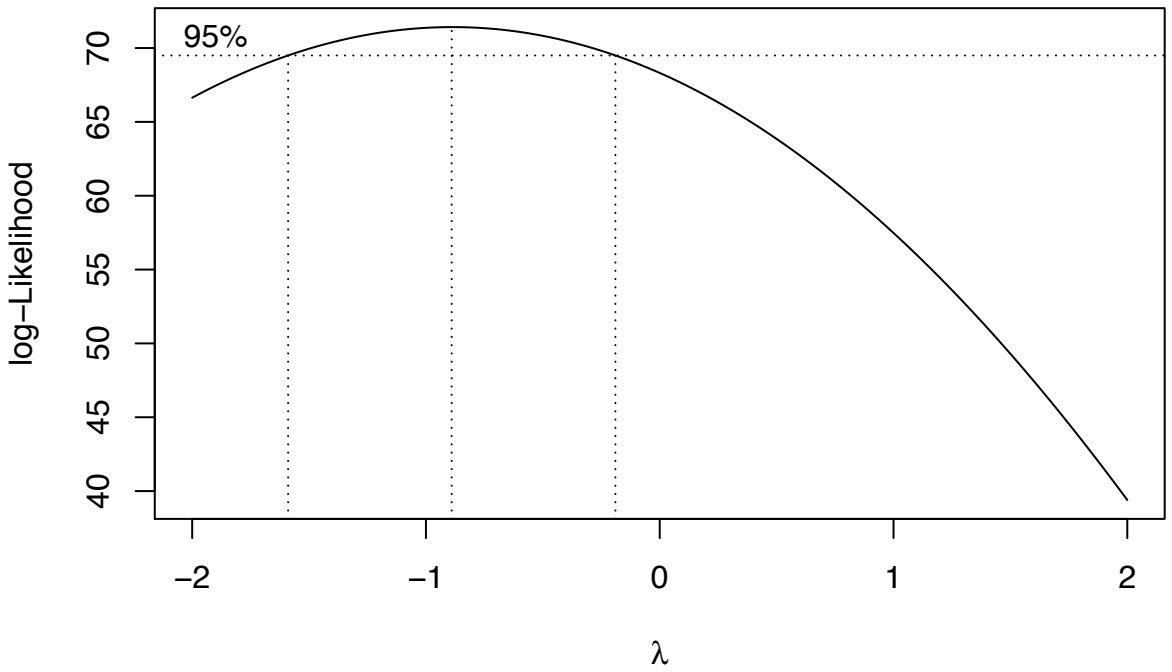
Q2 (a-d) - Data 1: working output

(a) To stabilizing the variance, we apply Box-cox power transformation, it suggests a simple variance stabilizing of the data. What is the simple transformation on Y suggested from boxcox()?

```
library(MASS)
bc=boxcox(y~workman,lambda=seq(-2,2,by=0.01))
```




```
bc$x[bc$y==max(bc$y)]
```

```
## [1] -0.89
```

Simple Transformation: Simple transformation on Y suggested from boxcox is 1/y.

(b) Examine the transformed Y (from Q2-a) variability for the ten workmen using the Bartlett test. What is your conclusion? Does it agree or disagree with Q1-d?

```
Yt = (1/work$y)
bartlett.test(Yt~workman, data=work)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Yt by workman
## Bartlett's K-squared = 3.5241, df = 9, p-value = 0.9399
```

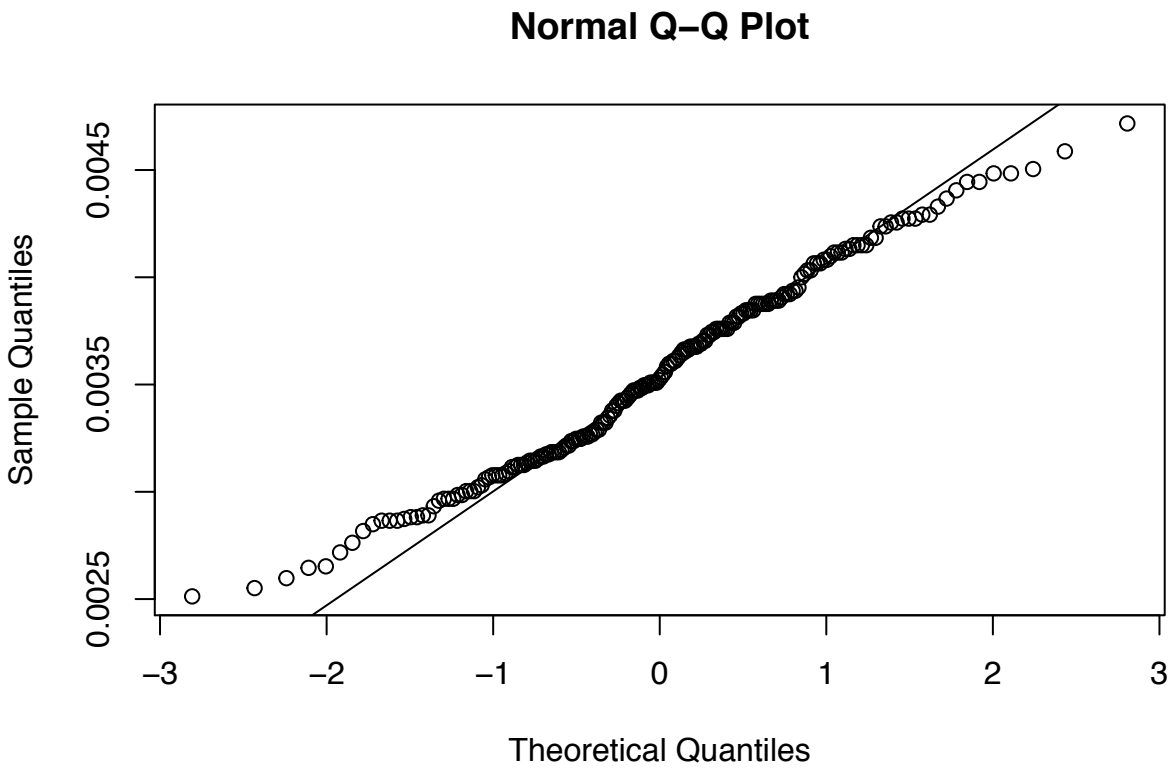
Conclusion: With 0.05 significance level, since the p-value is greater we fail to reject the fact that the variance is same for all groups. It's the opposite of the result we have found in Q1-d therefore it disagrees with it.

(c) Applying one-way ANOVA to this data, testing the equality of the output means for the ten workmen. How significant is the result? Does it agree with result you have in (Q1-b). Also repeat Q1-c to check the normality assumption for the transformed data, compare to Q1-c, what comment do you have ?

```
# one way anova
summary(aov(Yt~workman, data=work))
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## workman       9 3.829e-05 4.254e-06   132.9 <2e-16 ***
## Residuals   190 6.080e-06 3.200e-08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# normal qqplot
qqnorm(Yt); qqline(Yt)
```



Conclusion: Looking at the anova table, the p-value is extremely small(<2e-16). With 0.05 significance level, since pvalue < significance level, we strongly reject the null hypothesis and therefore can conclude that there are at least two groups with different means. It's the same result what we have found in Q1b. Looking at the normal qqplot, the residuals follow the line very well except at the upper quantile but even with that we can still say that the normality assumption is satisfied.

(d) Why would we want to prefer the second ANOVA over the first one, even though both give roughly the same significance? We prefer the second ANOVA over the first one because the residuals of the second one are normally distributed.