MA 4710 Homework 5

Benjamin Hendrick

February 25, 2016

Problem 3.4

Load the data into R and rename the variables.

```
CH03PR03 <- read.table("~/GitHub/MA-4710/Homework 5/CH03PR03.txt", quote="\"", comment.char="")
names(CH03PR03) <- c("gpa","act","intel","rank")
```

Part F

Obtain and residuals from the linear model between Y and X_1.

```
gpa.lm <- lm(gpa~act, data= CH03PR03)
gpa.resid <- resid(gpa.lm)
```

Plot the residuals against the intelligence score X_2.

```
plot(x = CH03PR03$intel, y = gpa.resid,
     xlab = "Intelligence Score",
     ylab = "Residuals",
     main = "Residuals against Intelligence Score")
```

Figure 1 suggests that there is a correlation between the error terms and the intelligence score. Therefore, the model wouldn't be improved by this correlation.

Plot the residuals against the class rank X_3.

```
plot(x = CH03PR03$rank, y = gpa.resid,
     xlab = "Class Rank",
     ylab = "Residuals",
     main = "Residuals of Rank")
```

Figure 2 suggests that there is no correlation between the error terms and the class rank. Therefore, the model may benefit by including the class rank variable.

Problem 3.15

Load the data into R and rename the variables.

```
CH03PR15 <- read.table("~/GitHub/MA-4710/Homework 5/CH03PR15.txt", quote="\"", comment.char="")
names(CH03PR15) <- c("conc", "time")
```

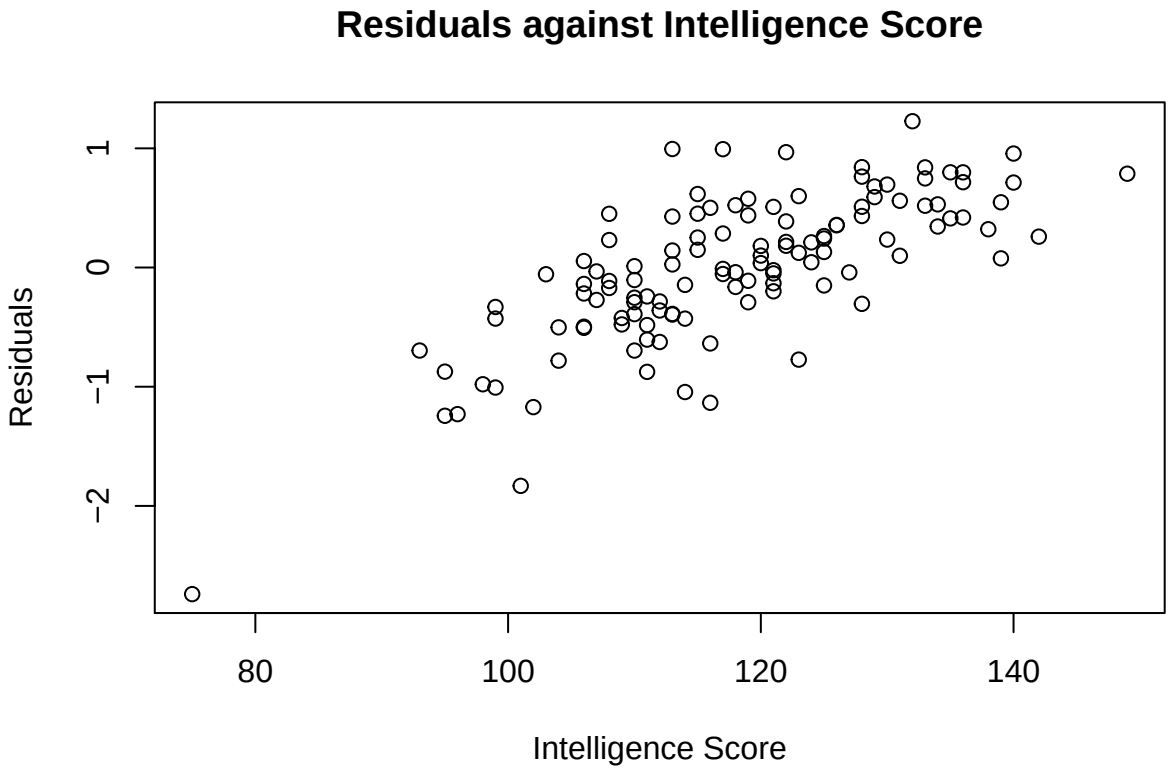

Figure 1: Scatter plot of residuals against the intelligence score X_2

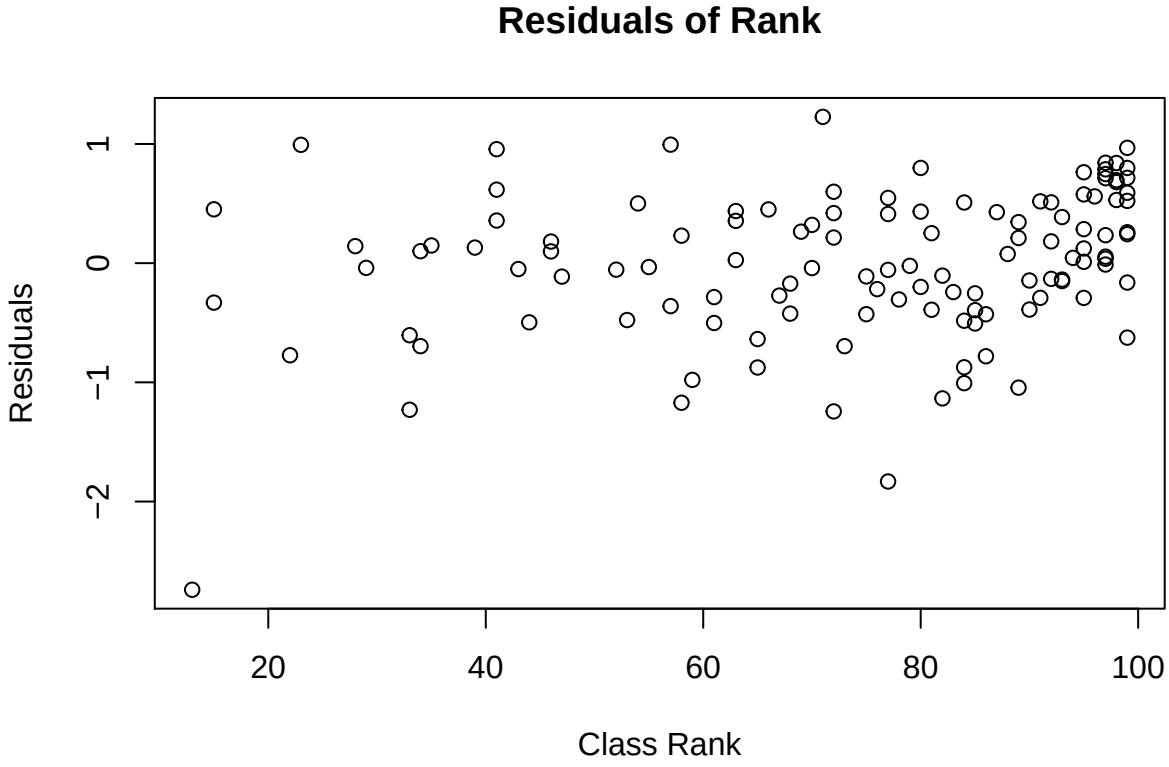

Figure 2: Scatter plot of residuals against the class rank X_3

Part A

Fit the linear regression function using the `lm` function.

```
chem.lm <- lm(conc~time, data = CH03PR15)
```

Part B

Use the `anova` function to perform the F test to determine wheter or not the lack of fit of the linear regression.

```
anova(chem.lm)
```

```
## Analysis of Variance Table
##
## Response: conc
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## time       1 12.5971  12.597  55.994 4.611e-06 ***
## Residuals 13  2.9247   0.225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let $H_0 : \beta = 0$ and $H_1 : \beta \neq 0$.

Because the p-value is significantly smaller than $\alpha = 0.025$, we reject H_0 and conclude that the slope β is not zero and the model is a good fit for the data.

Part C

The test in Part B does not indicate what regression function is appropriate when it leads to the conclusion that lack of fit of a linear regression function exists.

Problem 3.16

Use the same `CH03PR15` from Problem 3.15.

Part A

Plot the data in a scatter plot.

```
plot(x = CH03PR15$time, y = CH03PR15$conc,
     xlab = "Time",
     ylab = "Concentration of Solution",
     main = "Concentration of Solution over Time") # try log transformation
```

Figure 3 suggests that a *log* transformation is necessary because the data has a negative exponential trend and is heteroscedastic.

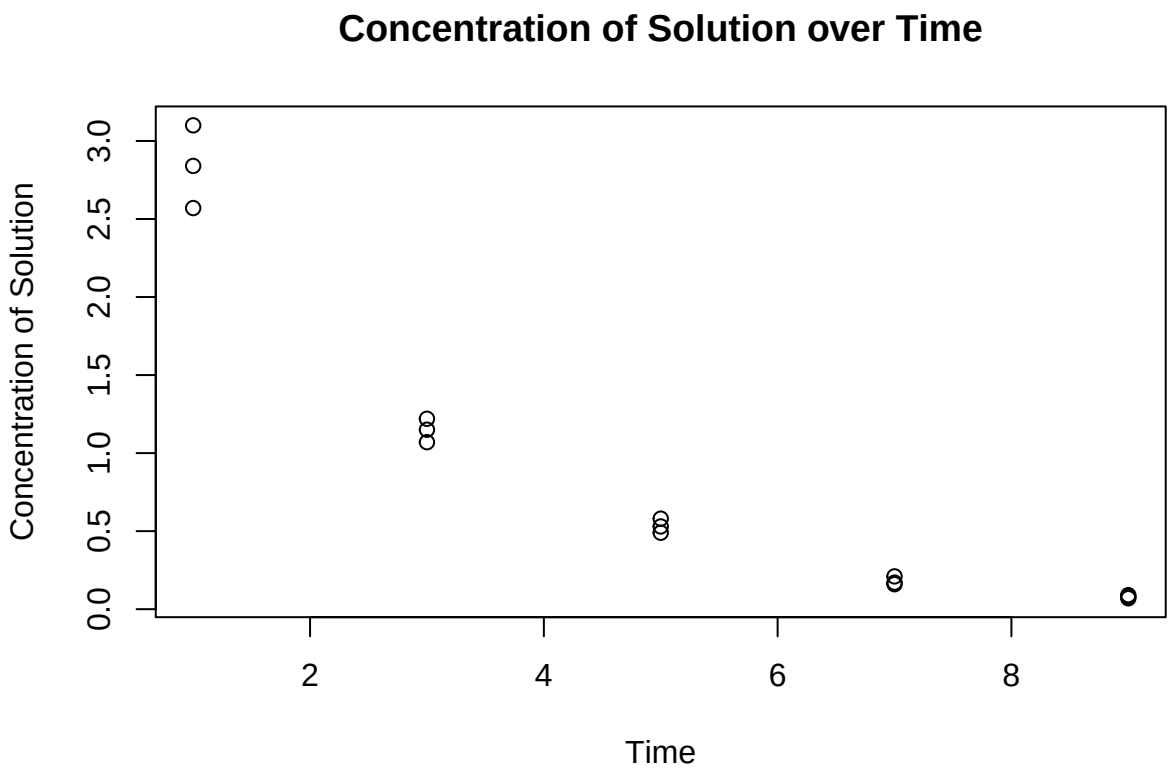

Figure 3: Scatter plot of concentration of solution over time.

Part B

Conduct a Box-Cox transformation on the data where $\lambda = -.2, -.1, 0, .1, .2$.

```
gmean <- exp(mean(log(CH03PR15$conc)))
sse <- NULL
lambda <- NULL
i <- 1
for (lam in seq(-0.2,0.2,0.1)){
  if (lam != 0){
  tY <- (CH03PR15$conc^lam - 1) / (lam*gmean^(lam-1))
} else {
  tY <- log(CH03PR15$conc)*gmean
}
  test <- anova(lm(tY~CH03PR15$time))
  sse[i] <- test['Residuals','Sum Sq']
  lambda[i] <- lam
  i <- i+1
}
```

The SEE values for each λ value are:

λ	SSE
-.2	0.1235305
-.1	0.0650507
0	0.038973

λ	SSE
.1	0.0439606
.2	0.0813179

Select $\lambda = 0$ because it has the smallest SSE. This can also be confirmed in R.

```
lambda[which.min(sse)]
```

```
## [1] 0
```

```
plot(lambda,sse,type="o",
     main="Box-Cox Transform",
     xlab = "Lambda",
     ylab = "SSE")
```

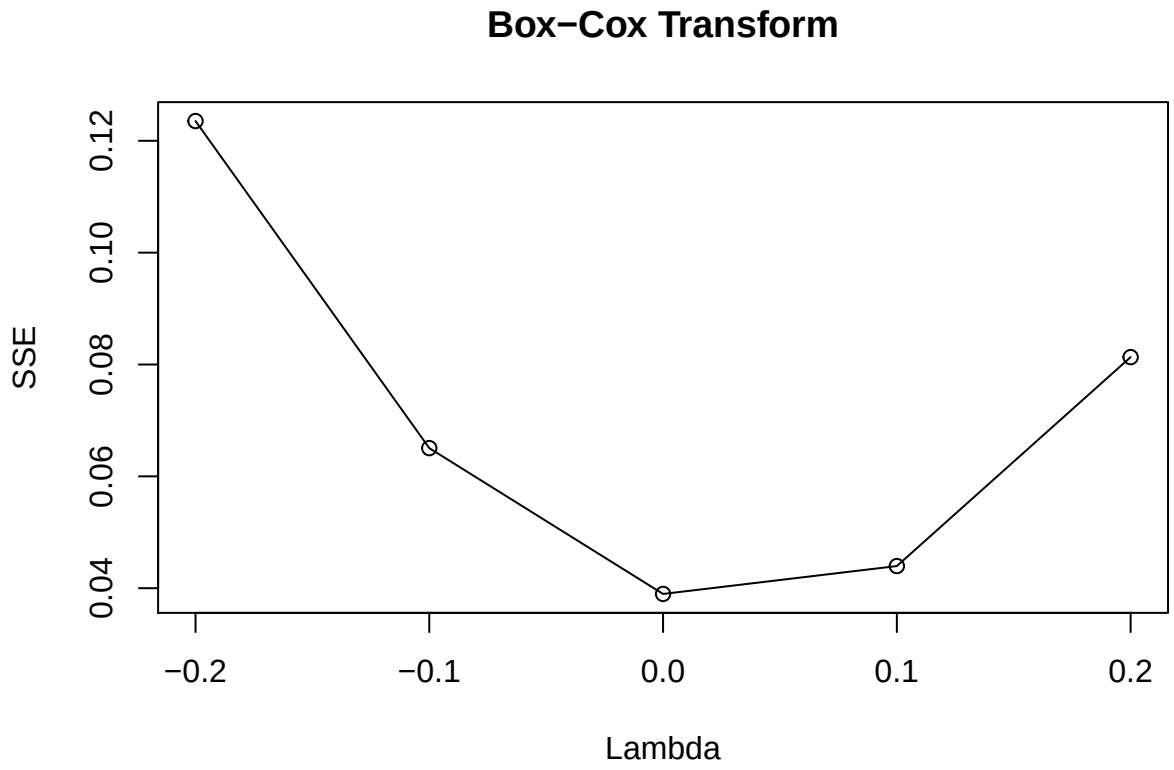

Figure 4: SSE values for different lambda values of the Box-Cox transformation

Based on the table above and Figure 4, we should select $\lambda = 0$ and therefore use a *log* transformation on the data. This supports the decision in Part A.

Part C

Create a new column of *log* transformed values.

```
CH03PR15$logConc <- log(CH03PR15$conc)
```

Obtain the estimated linear regression function for the transformed data.

```
logConc.lm <- lm(logConc~time, data = CH03PR15)
```

Part D

Plot the transformed data against its linear model found in Part C.

```
plot(x = CH03PR15$time,
     y = CH03PR15$logConc,
     xlab = "Time",
     ylab = "Solution Concentration (Log Scale)",
     main = "Solution Concentration over Time")
abline(logConc.lm)
```

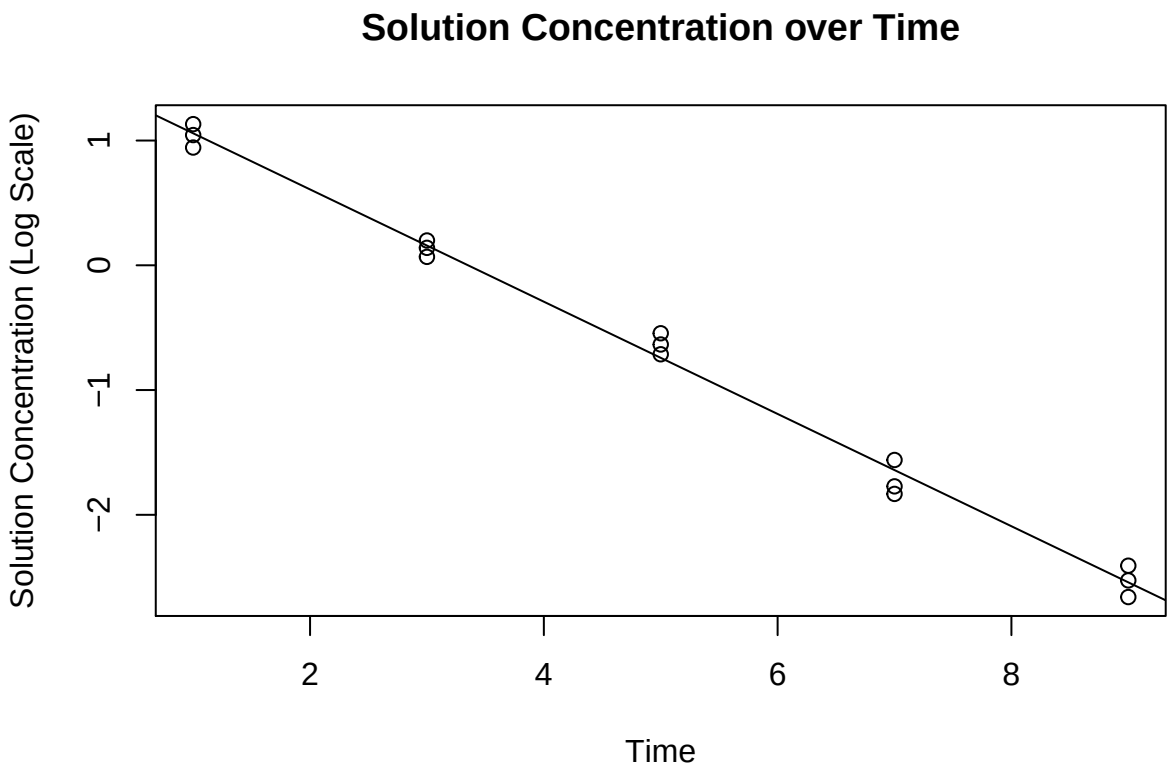

Figure 5: Plot of the transformed data against its linear model.

Figure 5 suggests that the transformed data fits the regression model very well.

Part E

Plot the residuals against their fitted values

```
plot(logConc.lm, which=1)
```

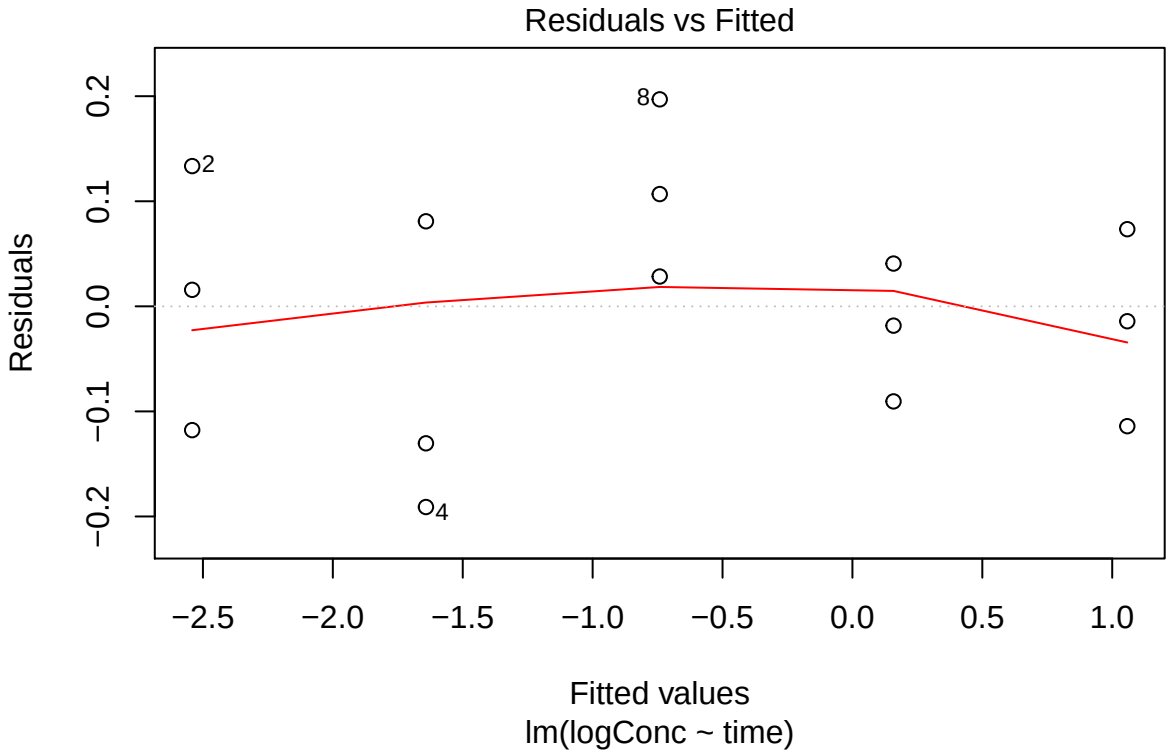

Figure 6: Residuals plotted against fitted values.

Plot the normal probability plot of the residuals.

```
plot(logConc.lm, which=2)
```

The plots (Figure 6 and Figure 7) suggest that the residuals have a high variance and are not normally distributed.

Part F

The transformed estimated regression function is expressed as

$$E(log(Y)) = 1.5079 - 0.4499X$$

The orignal estimated regression function can be expressed

$$E(Y) = 10^{1.5079} - 10^{0.4499^X} = 32.2033 - 2.81773^X$$

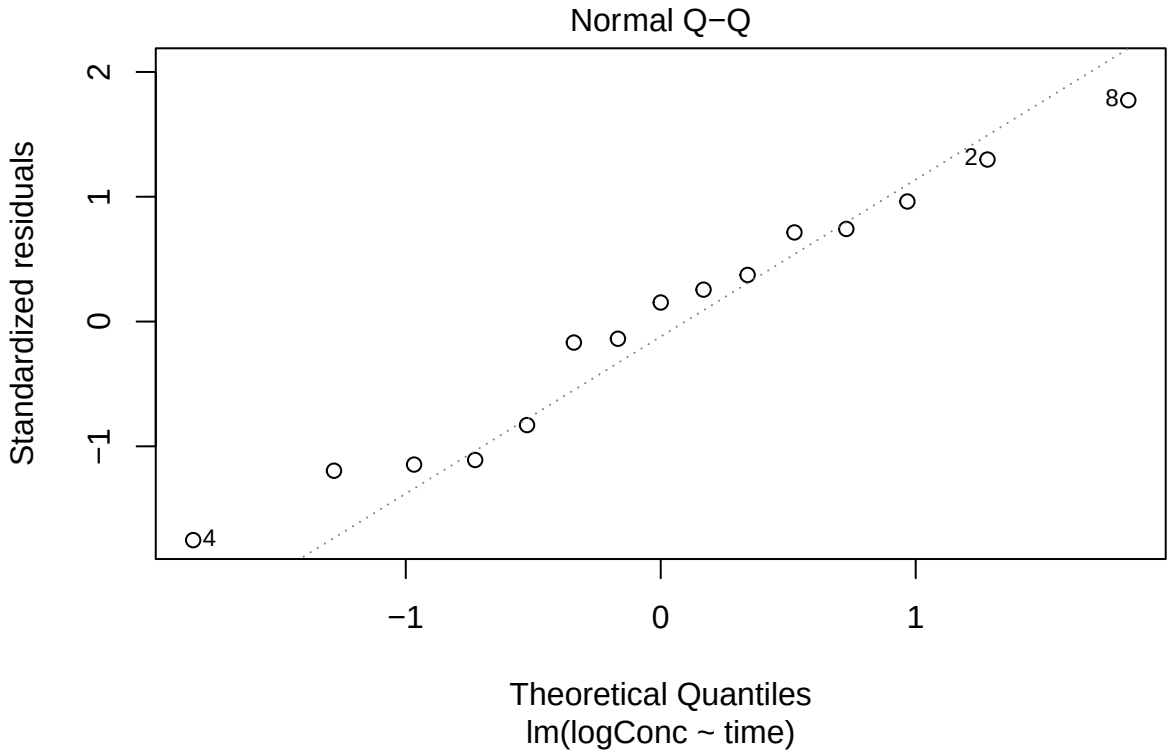

Figure 7: Normal probability plot of the residuals.