Di Zhen (1717719)

dizhen

2020.3.6

Question 1

```
OLD <- c(2041.7,257.0,524.8,257.0,128.8,128.8,1023.3,134.9,257.0,125.9,257.0,123.0,1000.0,120.2,128.8)
NEW <- c(12302.7,6918.3,4466.8,1584.9,933.3,1659.6,9120.1,575.4,2630.3,2398.8,1905.5,851.1,3467.4,1380.4
mydata <- data.frame(OLD, NEW)
```

```
if(!require(car)){install.packages("car")}
```

```
## Loading required package: car
```

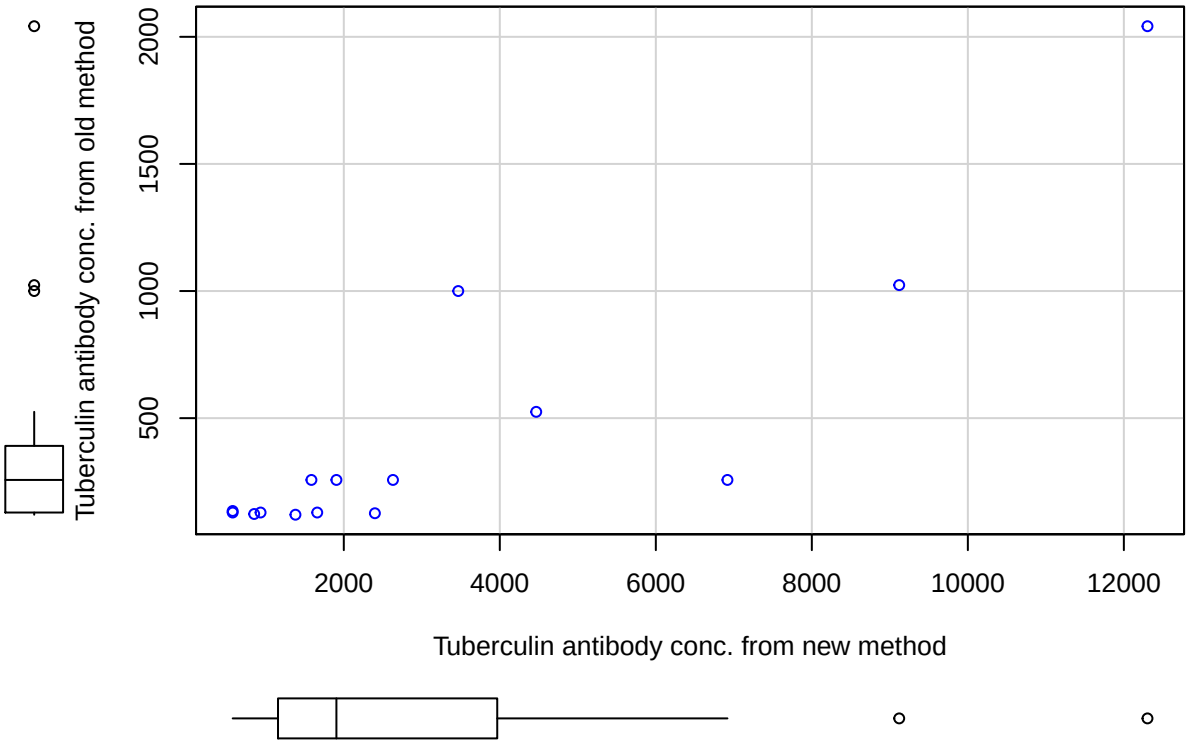
```
## Warning: package 'car' was built under R version 3.6.3
```

```
## Loading required package: carData
```

```
library(car)

scatterplot(OLD ~ NEW, data = mydata, smooth = FALSE, regLine = FALSE,
     main = "Figure 1. Tuberculin antibody concentrations estimated from new and old laboratory methods
     ylab = "Tuberculin antibody conc. from old method",
     xlab = "Tuberculin antibody conc. from new method",
     cex.main = 0.94)
```

Figure 1. Tuberculin antibody concentrations estimated from new and old laboratory methods



Question 2

```
mydata.LM <- lm(OLD~NEW,data = mydata)
summary(mydata.LM)
```

```
##
## Call:
## lm(formula = OLD ~ NEW, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -652.41  -74.15   22.18   69.15  554.99
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.60568  101.69092  -0.212    0.835
## NEW           0.13457    0.02136   6.301 2.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 277 on 13 degrees of freedom
## Multiple R-squared:  0.7534, Adjusted R-squared:  0.7344
## F-statistic: 39.71 on 1 and 13 DF,  p-value: 2.739e-05
```

The estimated intercept is around -21.61, the estimated slope is around 0.13.

Question 3

In the test for H0: intercept = 0, t-value = -0.212, p-value = 0.8350. This means that there is 83.50% probability that we would observe at least as extreme a result as this for the intercept estimate under the null hypothesis. Because 0.835 > 0.05, we fail to reject the null hypothesis and conclude that intercept = 0.

Question 4

In the test for H0: slope = 0, t-value = 6.301, p-value = 2.74e-05. This means that there is < 0.01 % probability that we would observe at least as extreme a result as this for the slope estimate under the null hypothesis. Because 2.74e-05 < 0.05, we reject the null hypothesis and conclude that the slope unequals to 0.

Question 5

```
confint(mydata.LM, level = 0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept) -241.2955542 198.0841887
## NEW            0.0884362   0.1807103
```

The 95% confidence interval for slope is from 0.09 to 0.18.

Question 6

We are 95% confident that the slope estimate lies between the interval 0.09 and 0.18, because on repeated sampling, 95% of intervals constructed in ths manner will contain the true slope

Question 7

```
mydata$OLDt <- log10(mydata$OLD)
mydata$NEWt <- log10(mydata$NEW)
```

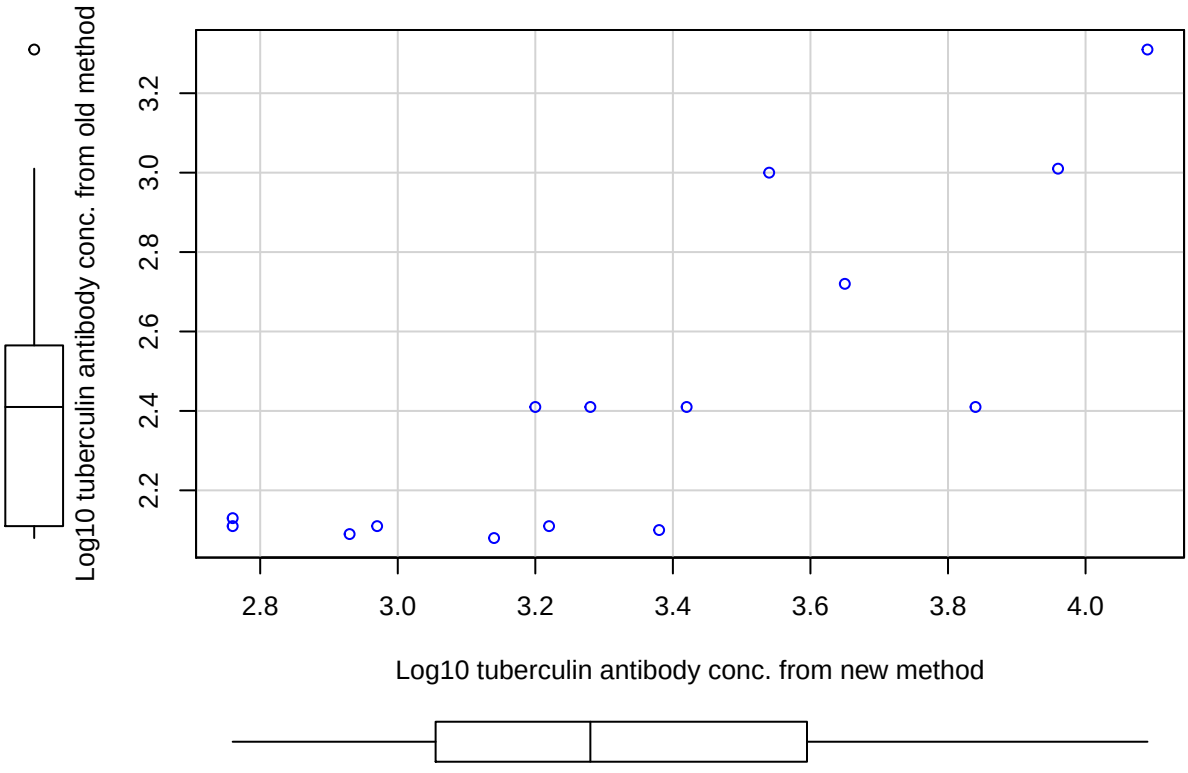
OLD	NEW
3.309992	4.090000
2.409933	3.839999
2.719994	3.649997
2.409933	3.200002
2.109916	2.970021
2.109916	3.220003
3.010003	3.960000
2.130012	2.759970
2.409933	3.420005

OLD	NEW
2.100026	3.379994
2.409933	3.280009
2.089905	2.929981
3.000000	3.540004
2.079904	3.140005
2.109916	2.759970

Question 8

```
scatterplot(OLDt ~ NEWt, data = mydata, smooth = FALSE, regLine = FALSE,
     main = "Figure 2. Tuberculin antibody concentrations estimated from new and old laboratory methods
     ylab = "Log10 tuberculin antibody conc. from old method",
     xlab = "Log10 tuberculin antibody conc. from new method",
     cex.main = 0.94)
```

Figure 2. Tuberculin antibody concentrations estimated from new and old laboratory methods



Question 9

```
mydata.LMt <- lm(OLDt~NEWt,data = mydata)
summary(mydata.LMt)
```

```
##
## Call:
## lm(formula = OLDt ~ NEWt, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41764 -0.13194  0.03307  0.12454  0.41388
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2631     0.5094  -0.516 0.614217
## NEWt          0.8049     0.1513   5.319 0.000139 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2343 on 13 degrees of freedom
## Multiple R-squared:  0.6851, Adjusted R-squared:  0.6609
## F-statistic: 28.29 on 1 and 13 DF,  p-value: 0.0001393
```

The regression equation: Log10 tuberculin antibody conc. from old method = -0.2631 + 0.8049 * (Log10 tuberculin antibody conc. from new method)

Question 10

In the test for H0: intercept = 0, t-value = -0.516, p-value = 0.6142. This means that there is 61.42% probability that we would observe at least as extreme a result as this for the intercept estimate under the null hypothesis. Because $0.6142 > 0.05$, we fail to reject the null hypothesis and conclude that intercept = 0.

Question 11

In the test for H0: slope = 0, t-value = 5.319, p-value = 0.0001. This means that there is 0.01 % probability that we would observe at least as extreme a result as this for the slope estimate under the null hypothesis. Because $0.0001 < 0.05$, we reject the null hypothesis and conclude that the slope unequals to 0.

Question 12

The tests of hypotheses in (3) and (4) consistent with your tests in (9) and (10). In both case, we fail to reject the null hypothesis of intercept estimate but reject the null hypothesis of slope estimate. After log transformation of both X and Y, the intercept is still meaningless while the slope is still useful.

Question 13

```
confint(mydata.LMt, level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) -1.3636673 0.8374771
## NEWt         0.4779407 1.1317830
```

The 95% confidence interval for slope is from 0.4779407 to 1.1317830.

Question 14

We are 95% confident that the slope estimate lies between the interval 0.4779407 and 1.1317830, because on repeated sampling, 95% of intervals constructed in ths manner will contain the true intercept.

Question 15

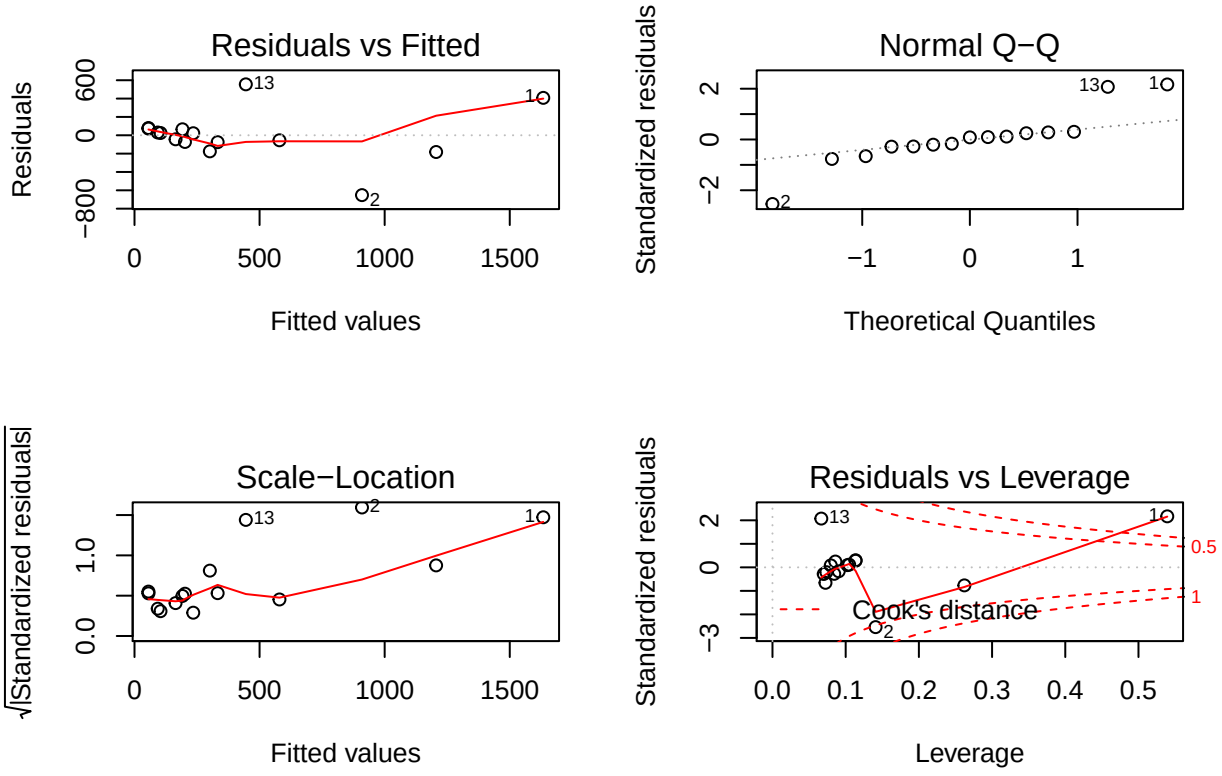
There are three assumptions have to be tested before fit a linear regression model:

1. Linearity between outcome (Y) and predictor (X) variables;
2. The outcome variable is normally distributed across predictor values;
3. The variance of outcome is the same across predictor values.

If these hold, then residuals are normally distributed with a mean of 0 and a constant variance across the predictor values: Residuals ~ Normal $(0, \sigma^2)$

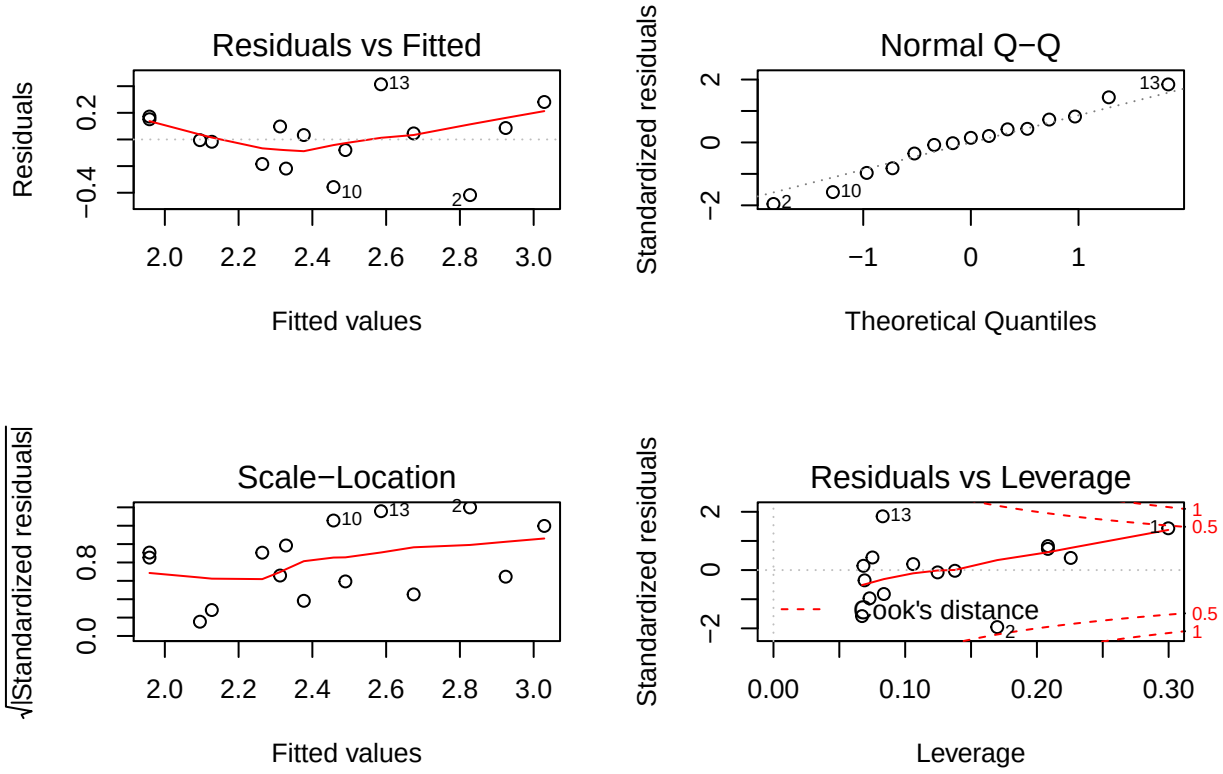
These four plots are used to test the assumptions for linear regression of original data:

```
par(mfrow=c(2,2))
plot(mydata.LM)
```



These four plots are used to test the assumptions for linear regression of log transformed data:

```
par(mfrow=c(2,2))
plot(mydata.LMt)
```



The first plot, a constant variance plot, checks for the homogeneity of the variance and the linear relation. If there is no pattern in this graph, then the assumptions are met. We can see that the plot of data after log transformation is better.

The second plot, a Q-Q plot, checks that the residuals follow a normal distribution. The points should fall on a line if the normality assumption is met. We can see that the plot of data after log transformation is better.

In conclusion, data after log transformation meet the assumption of linear regression better, so the second regression equation is more appropriate.

THE END