Lab 9 | Jon Lee

Code

Lab 9 | BINF 6310 | Spring 2020 | Jon Lee

Send your answers together with code to afodor@uncc.edu (mailto:afodor@uncc.edu) by Thursday, April 2nd.

1. This question uses data from this paper: https://science.sciencemag.org/content/347/6217/78 (https://science.sciencemag.org/content/347/6217/78) Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science 02 Jan 2015: Vol. 347, Issue 6217, pp. 78-81

(1A): Download the data from here examining the relationship between the number of cell divisions and cancer risk: https://fodorclasses.github.io/classes/stats2020/cancerRisk.txt (https://fodorclasses.github.io/classes/stats2020/cancerRisk.txt)

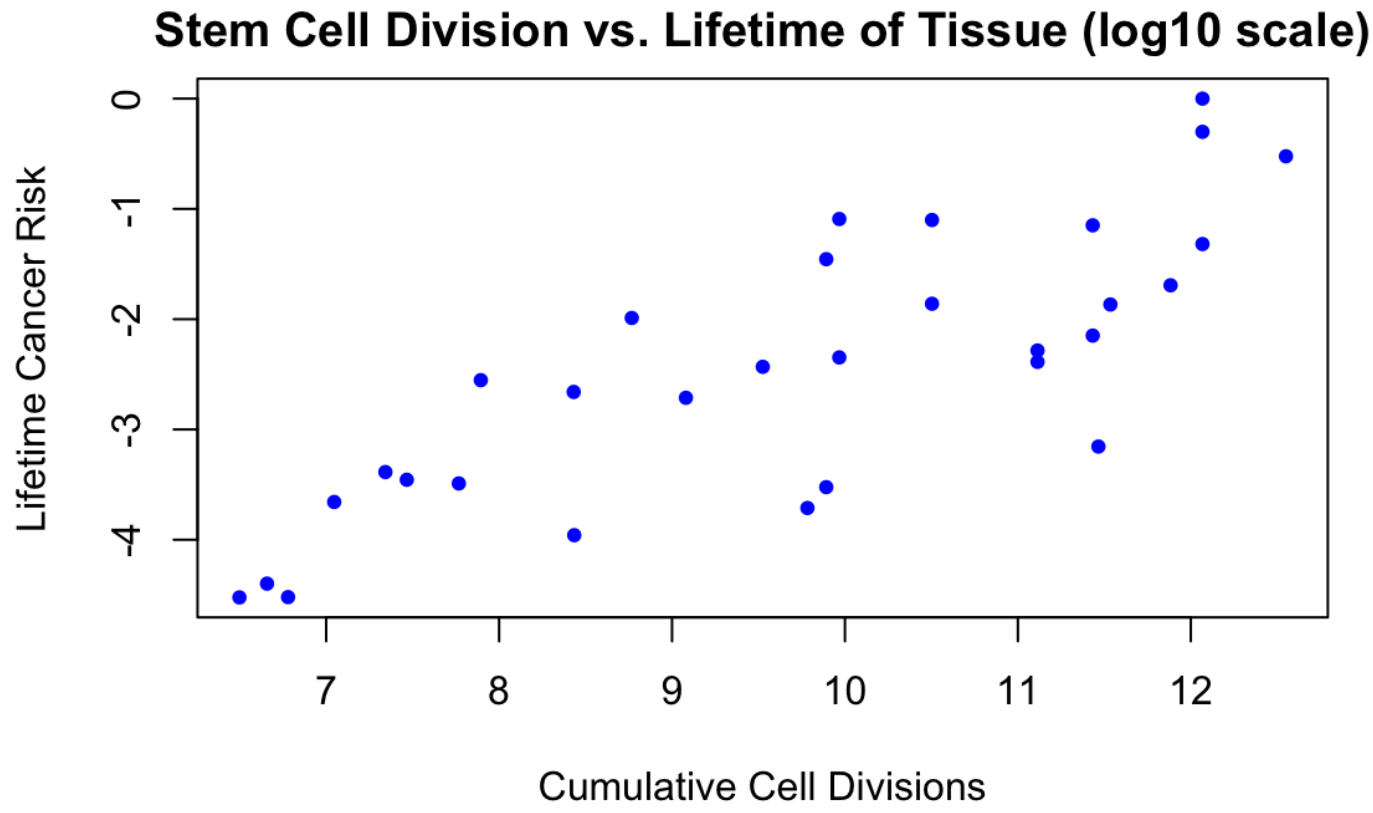
Hide

```
#read in data file
cancerRisk <- read.table("cancerRisk.txt", sep = "\t", header = TRUE)
```

On a log10-log10 scale graph Lifetime_cancer_risk (on the y-axis) vs. CumulativeCellDivisions (on the x-axis). (This reproduces Fig. 1 from the paper). (You can read in the file with read.table("cancerRisk.txt", header=TRUE, sep=")

Hide

```
#log10 transform data and assign to vectors
x <- log10(cancerRisk$CumulativeCellDivisions)
y <- log10(cancerRisk$Lifetime_cancer_risk)

#plot data
plot(x,y, pch = 20, col = "blue", xlab = "Cumulative Cell Divisions", ylab = "Lifetime C
ancer Risk", main = "Stem Cell Division vs. Lifetime of Tissue (log10 scale)")
```



(1B): Using the lm function, fit a linear model with Lifetime_cancer_risk as the Y variable and CumulativeCellDivisions as the x-data. Add the regression line to the plot using the function abline(myLm) (where myLm is the linear model you created).

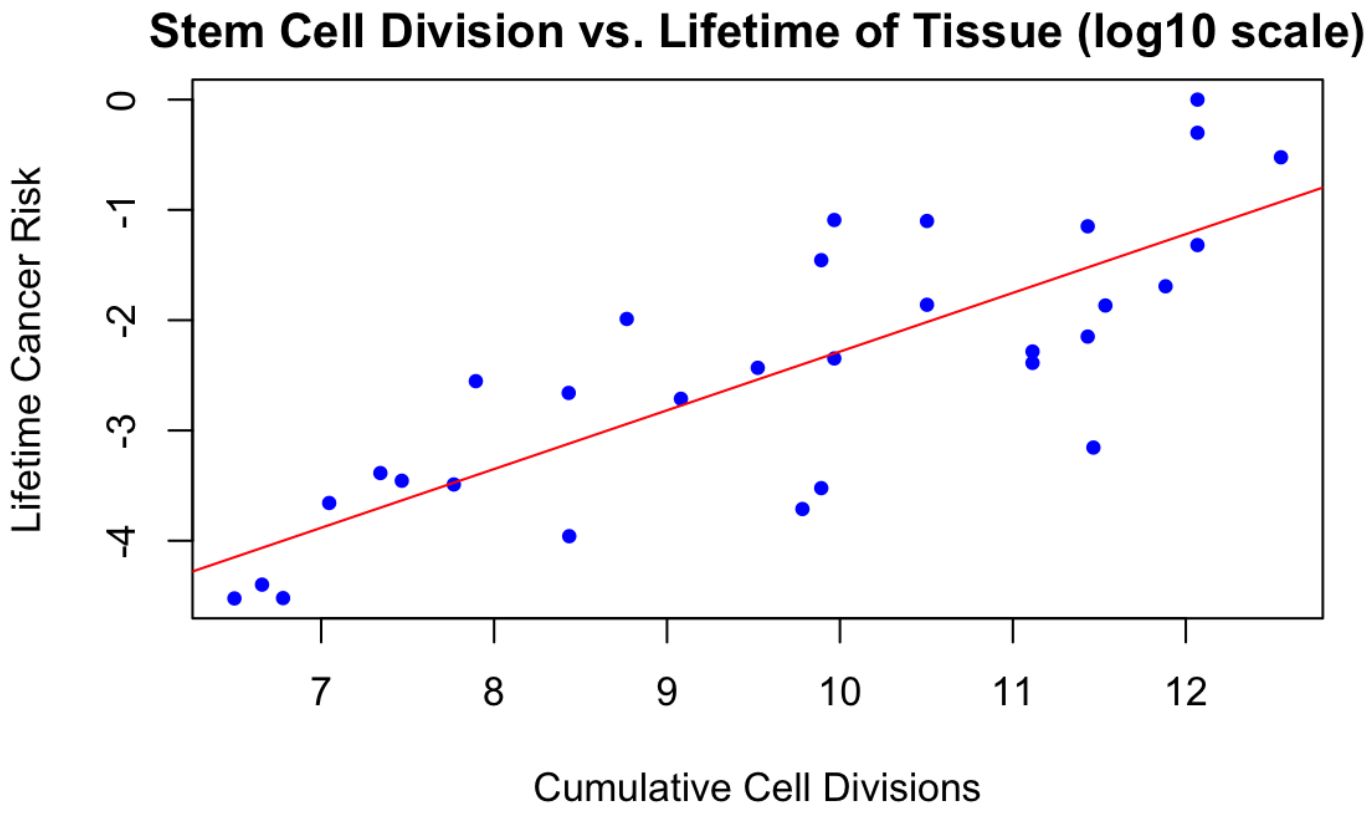
Hide

```
#perform linear model and disply summary
cancerRiskLM <- lm(y ~ x)
summary(cancerRiskLM)
```

```

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.65115 -0.46564  0.06167  0.43180  1.21040 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.61081    0.72292 -10.528 2.03e-11 ***
x            0.53264    0.07317   7.279 5.12e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7491 on 29 degrees of freedom
Multiple R-squared:  0.6463,    Adjusted R-squared:  0.6341 
F-statistic: 52.99 on 1 and 29 DF,  p-value: 5.124e-08
```

Hide

```
#replot data with linear model line
plot(x,y, pch = 20, col = "blue", xlab = "Cumulative Cell Divisions", ylab = "Lifetime C
ancer Risk", main = "Stem Cell Division vs. Lifetime of Tissue (log10 scale)")
abline(cancerRiskLM, col = "red")
```



(1C): What is the p-value for the null hypothesis that the slope of the regression between these two variables is zero? What is the r-squared value of the model?

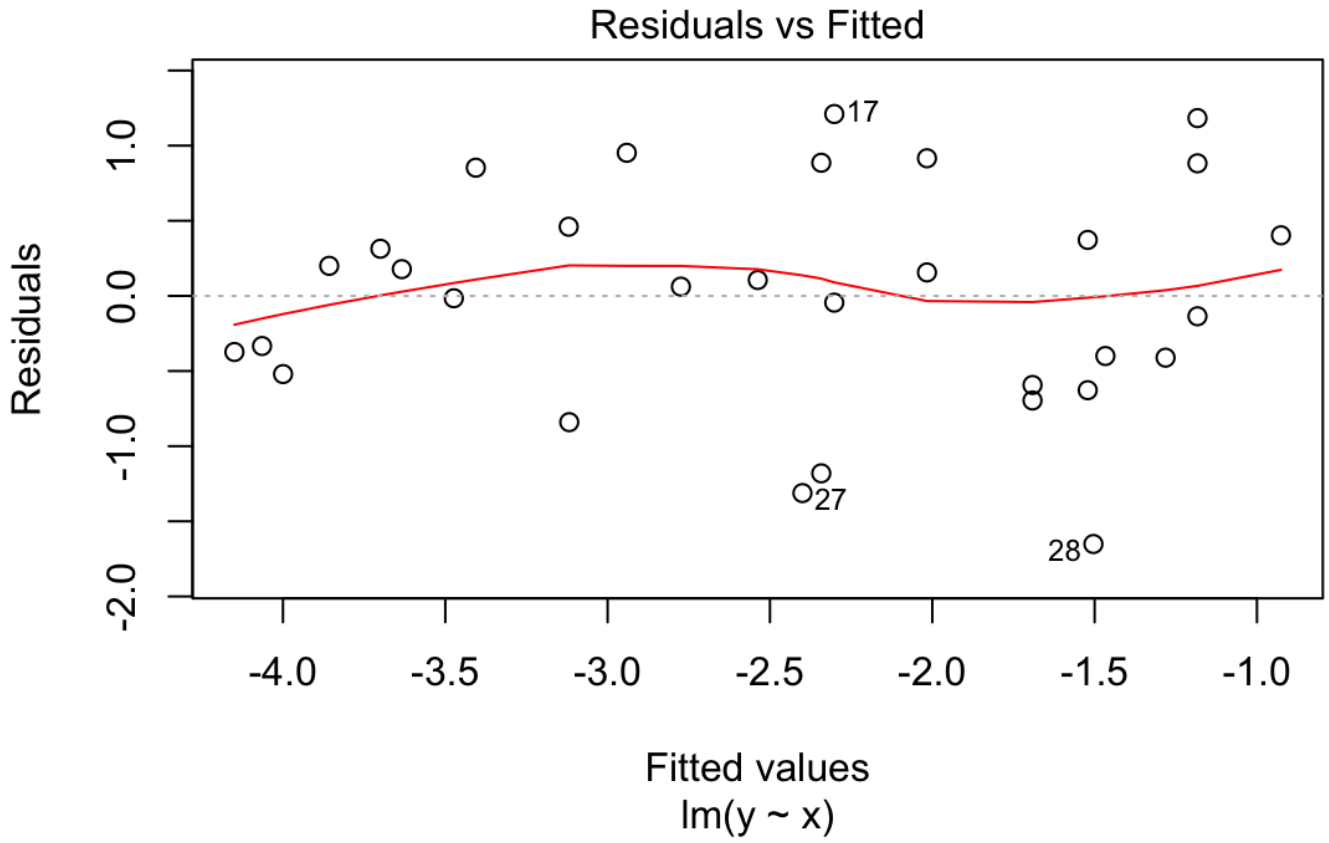
Answer:

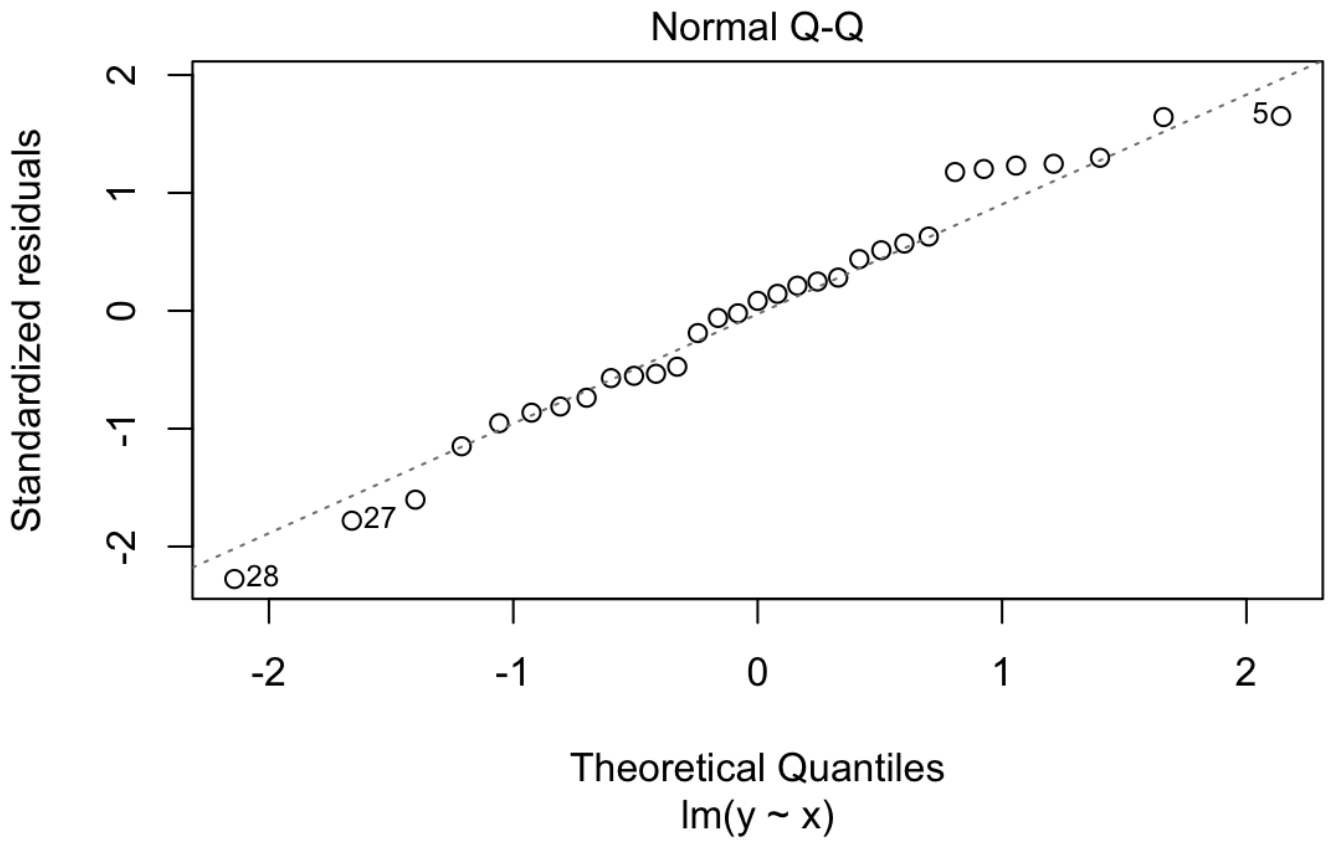
The p-value for the null that the slope is equal to 0 is 5.12e-08, and the r-squared value of the model is 0.6463. So therefore we reject the null that the slope is equal to 0 and the model is arelatively good fit to the data, but its not fantastic.

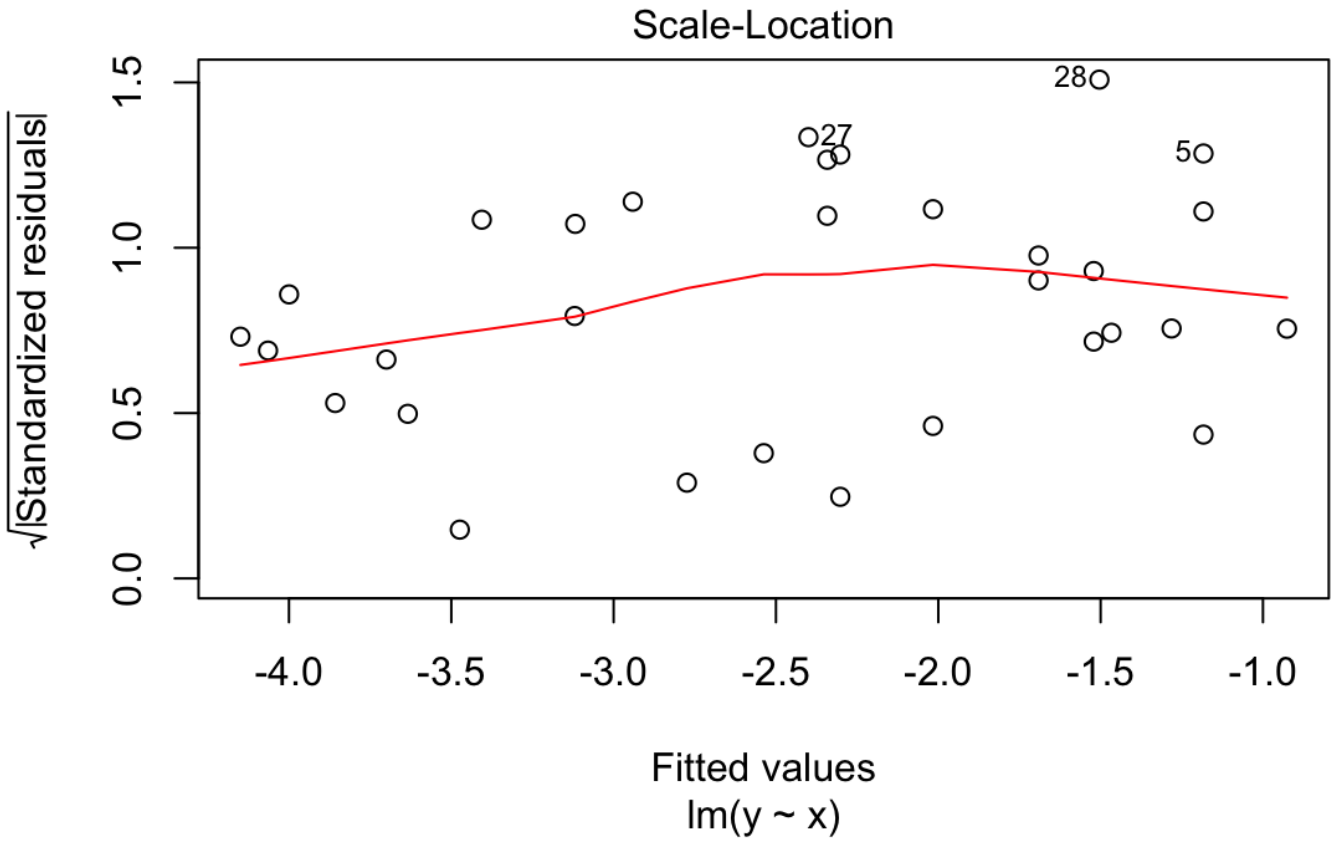
(1D): Are the assumptions of constant variance and normal distribution of the residues reasonable for this model? Justify your answer.

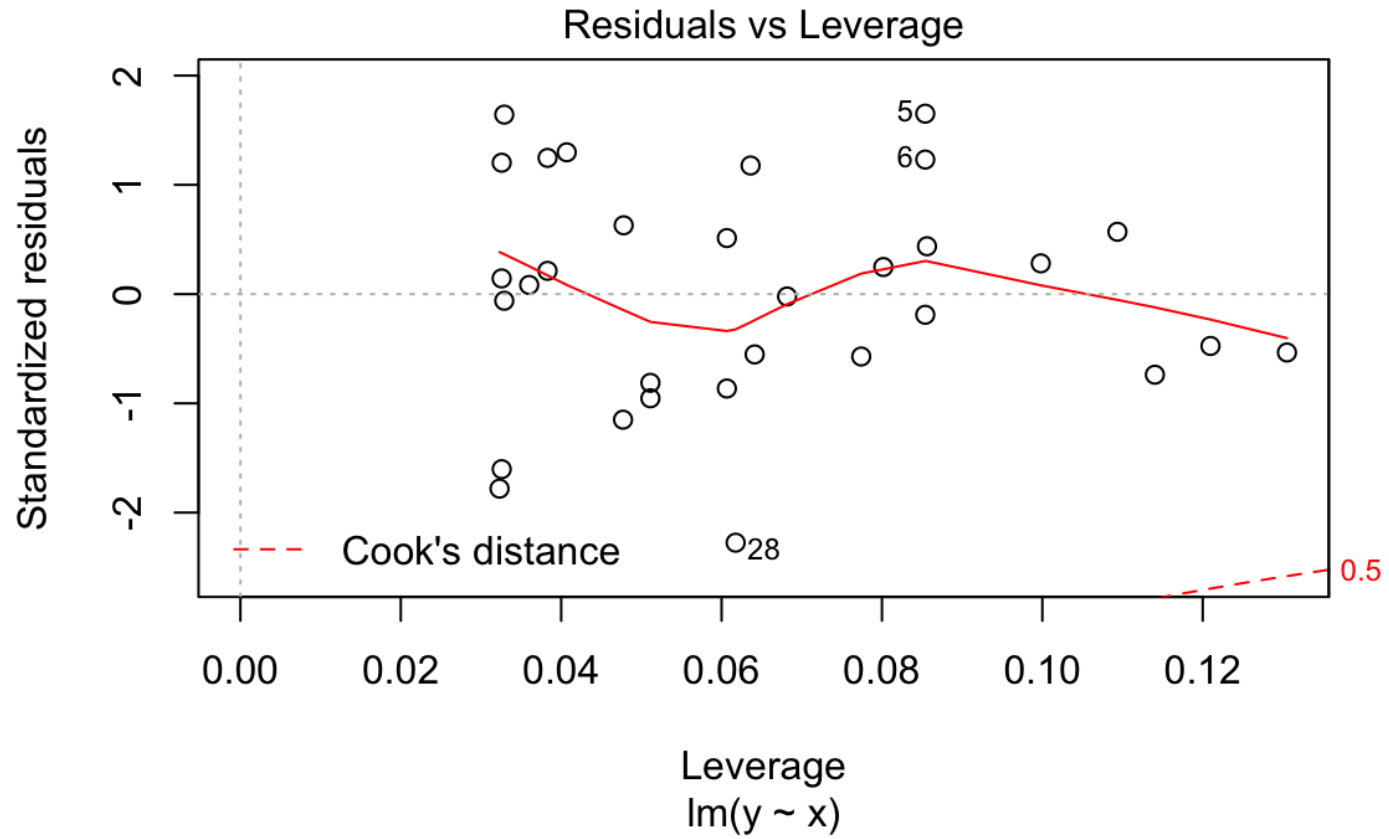
Hide

```
#plot residual graphs for assumption constraints
plot(cancerRiskLM)
```


Residuals vs Fitted
Residuals
1.0
0.0
-1.0
-2.0
17
27
28
-4.0
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
Fitted values
lm(y ~ x)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
5
27
28
-2
-1
0
1
2
Theoretical Quantiles
lm(y ~ x)


Scale-Location
√|Standardized residuals|
28
27
5
0.0
0.5
1.0
1.5
-4.0
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
Fitted values
lm(y ~ x)


Residuals vs Leverage
Standardized residuals
5
6
28
Cook's distance
0.5
-2
-1
0
1
2
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Leverage
lm(y ~ x)

Answer:

The assumptions here would be reasonable, becauase when looking at the normality of the residuals, they follow the normality fairly well, and the outliers are not too extreme. Constance variance also hold fairly well when looking at the residuals too because they follow a constant line with outliers that are again not too extreme.

2. Consider the case-control file for the colorectal adenomas data set that is here:
   http://afodor.github.io/classes/stats2015/caseControlData.txt
   (http://afodor.github.io/classes/stats2015/caseControlData.txt)

A separate file gives obesity (BMI) data for these same subjects:
http://afodor.github.io/classes/stats2015/BMI_Data.txt (http://afodor.github.io/classes/stats2015/BMI_Data.txt)

Hide

```
#read in data files
caseControlData <- read.table("caseControlData.txt", sep = "\t", header = TRUE)
BMI_Data <- read.table("BMI_Data.txt", sep = "\t", header = TRUE)

#remove case and control from sample caseControlData
caseControlData$sample <- sub("case", "", caseControlData$sample)
caseControlData$sample <- sub("control", "", caseControlData$sample)

#remove extraneous information from the suffix from sample caseControlData
caseControlData$sample <- substring(caseControlData$sample, 1, 10)

#import dplyr package to library
library(dplyr)
```

```
Registered S3 method overwritten by 'dplyr':
  method           from
  print.rowwise_df

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

Hide

```
#rename BMI_data studyid column to sample
colnames(BMI_Data)[1] <- "sample"

#inner join BMI_Data and caseControlData to form one data Table
data <- inner_join(BMI_Data, caseControlData)
```

```
Joining, by = "sample"
Column `sample` joining factor and character vector, coercing into character vector
```

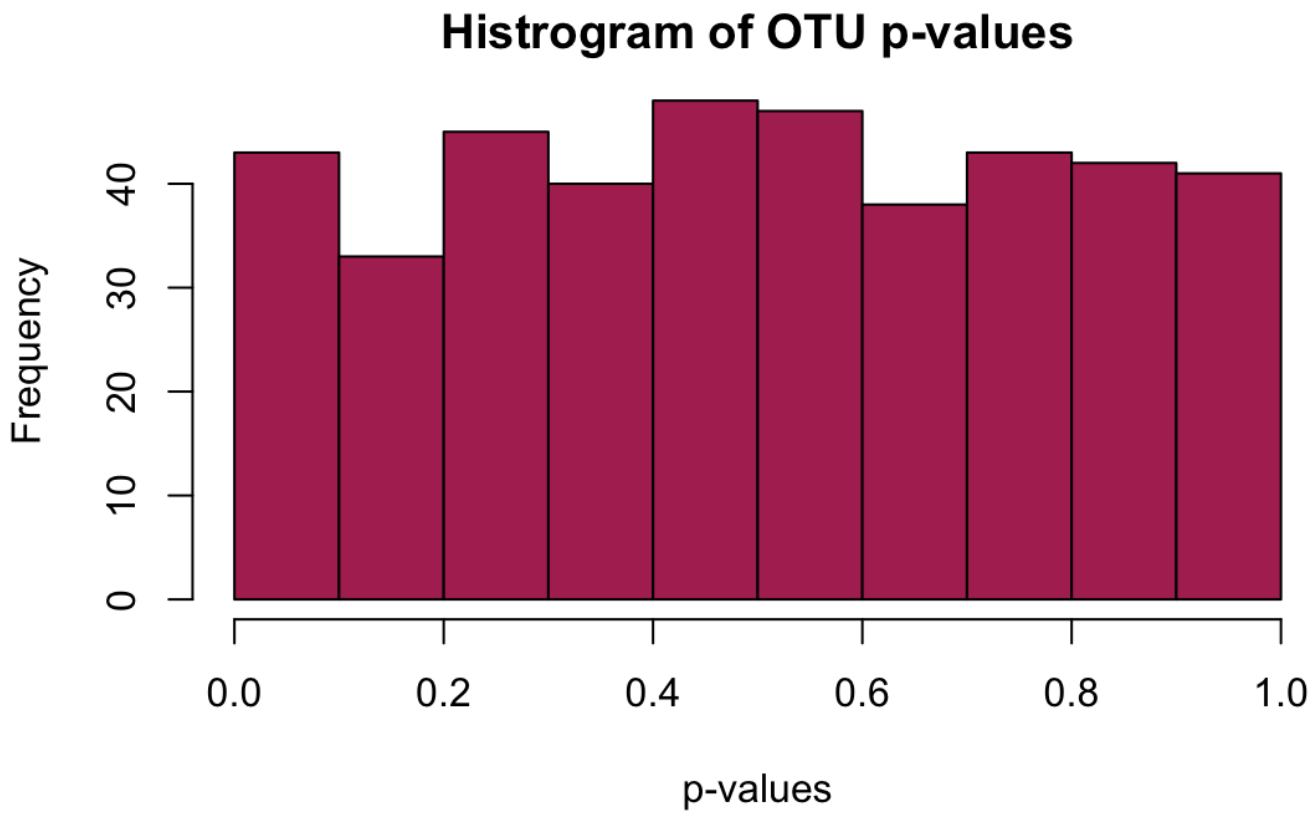
Hide

```
#remove NA row
data <- na.omit(data)
```

For each OTU in the spreadsheet, generate a p-value from linear regression comparing BMI to the relative abundance of each OTU. Graph out all the p-values.

Hide

```
#create vector of p-values
pVals <- vector()

#for loop to run a linear regression for each OTU
for(i in 1:(ncol(data)-2))
{
  j <- i+2
  X <- data[,2]
  Y <- data[,j]

  dataLM <- lm(Y ~ X)
  pVals[i] <- anova(dataLM)$"Pr(>F)"[1]
}

#plot p-values via histogram
hist(pVals, col = "maroon", xlab = "p-values", main = "Histrogram of OTU p-values")
```



Do they appear uniformly distributed? Does the microbial community appear to be influencing body weight in this cohort? Are any of these associations significant at a 10% false discovery rate?

Hide

```
#number of p-values at a 10% false discovery rate
count <- 0

for(i in 1:length(pVals))
{
  if(pVals[i] < 0.1)
  {
    count <- count + 1
  }
}

print(count)
```

```
[1] 43
```

Answer:

The p-values look to be uniformly distributed, which would indicate that the null hypothesis is true, that BMI does not influence the microbial community. For a 10% false discovery rate, we would see that 43 of the OTU are considered to be significant.

Hints: To lookup the ids in the BMI table, you will need to some processing on the “sample” column in the caseControl file. The following code will convert the a sampleID so that it will match the BMI file.

remove case and control key <- sub(“case”, "“, sampleID) key <- sub(”control“,”", key)

```
# remove extraneous information from the suffix
key <- strsplit( key, "_")[[1]][1]
```

Also, to get the p-value out of the linear model try:

```
 anova(myLm)$"Pr(>F)"[1]
```

We’ll see why that work shortly in future lectures.