hw1_lm

Anna Roser

January 22, 2019

```
knitr::opts_chunk$set(echo = TRUE)
library(fitdistrplus)
```

```
## Warning: package 'fitdistrplus' was built under R version 3.4.4
```

```
## Loading required package: MASS
```

```
## Loading required package: survival
```

```
## Loading required package: npsurv
```

```
## Loading required package: lsei
```

```
library(grDevices)
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.4.4
```

Question 1: Math and LSD

A) What level of LSD tissue concentration do you need to ensure a test score of >85%?

Solve 85 = 89.12-9.01x

You need a LSD concentration of 0.45g to get >85% score

B)How well does LSD tissue concentration predict test performance?

For every one gram of LSD ingested, the participants score drops by 9.01%

C)Why might the normal distribution be inappropriate to

model these data?

Because we have a small sample size and high variance in the data (6.05)

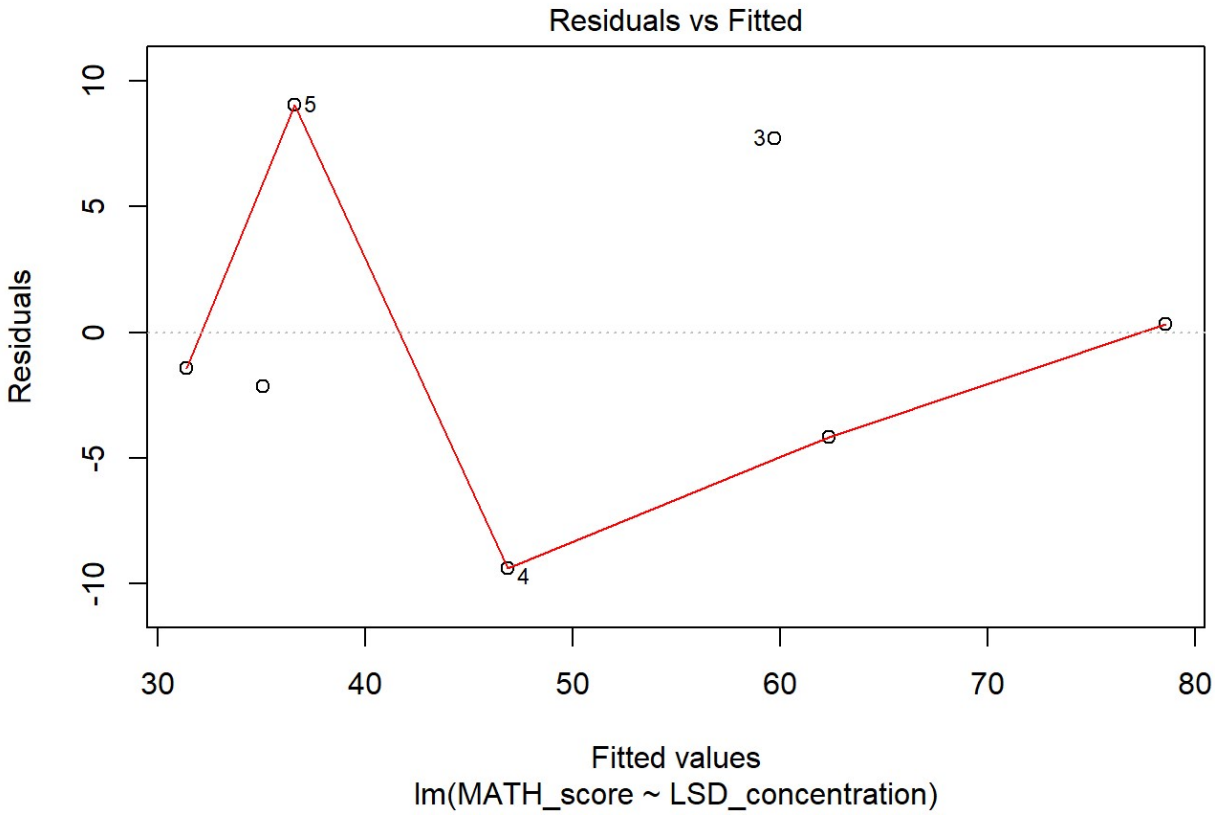
```
math<-read.csv("math_scores.csv")
head(math)
```

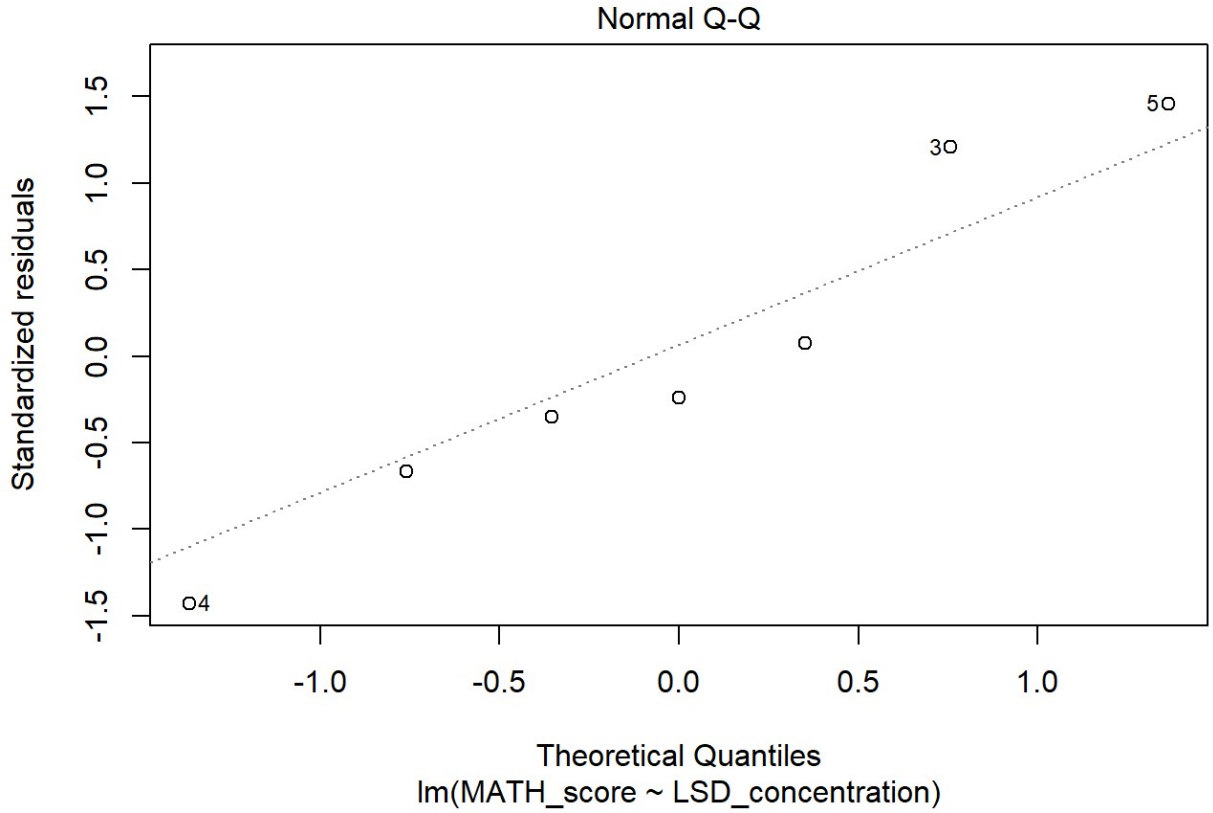
```
##   LSD_concentration MATH_score
## 1              1.17      78.93
## 2              2.97      58.20
## 3              3.26      67.47
## 4              4.69      37.47
## 5              5.83      45.65
## 6              6.00      32.92
```

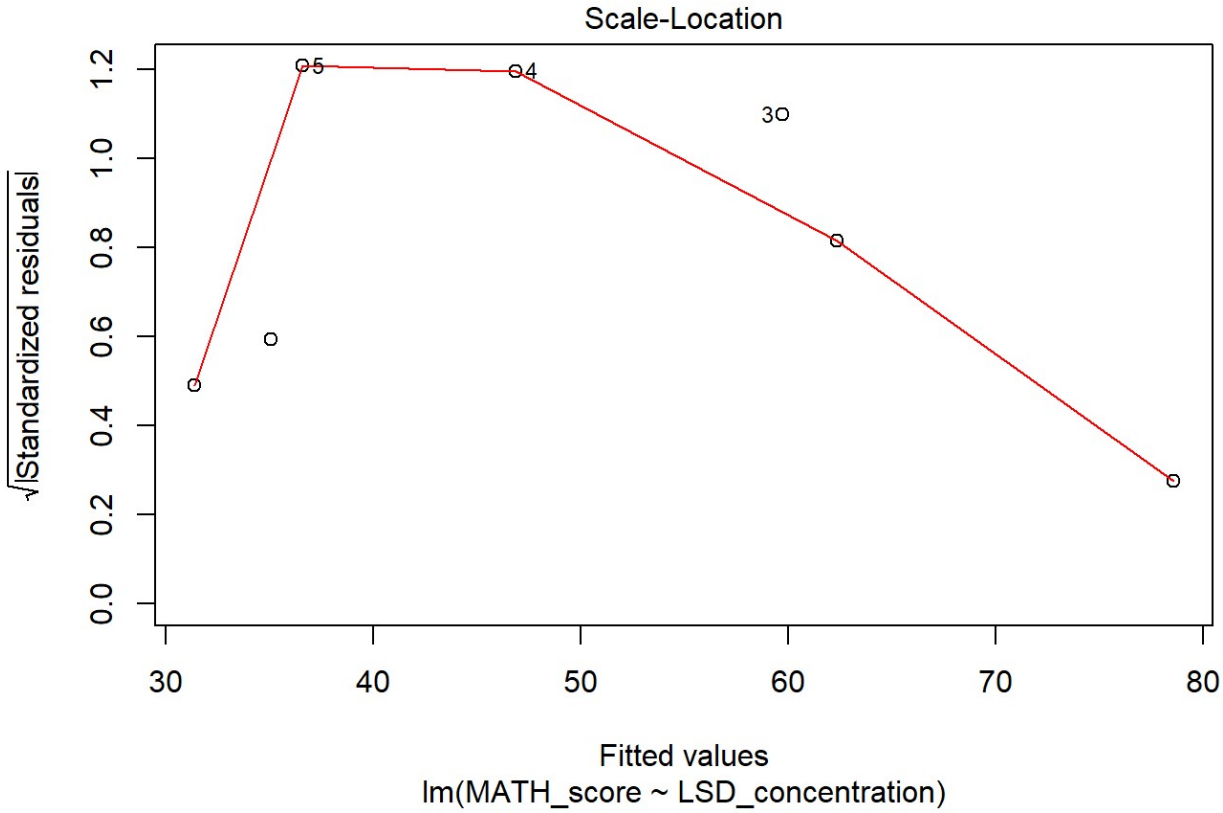
```
str(math)
```

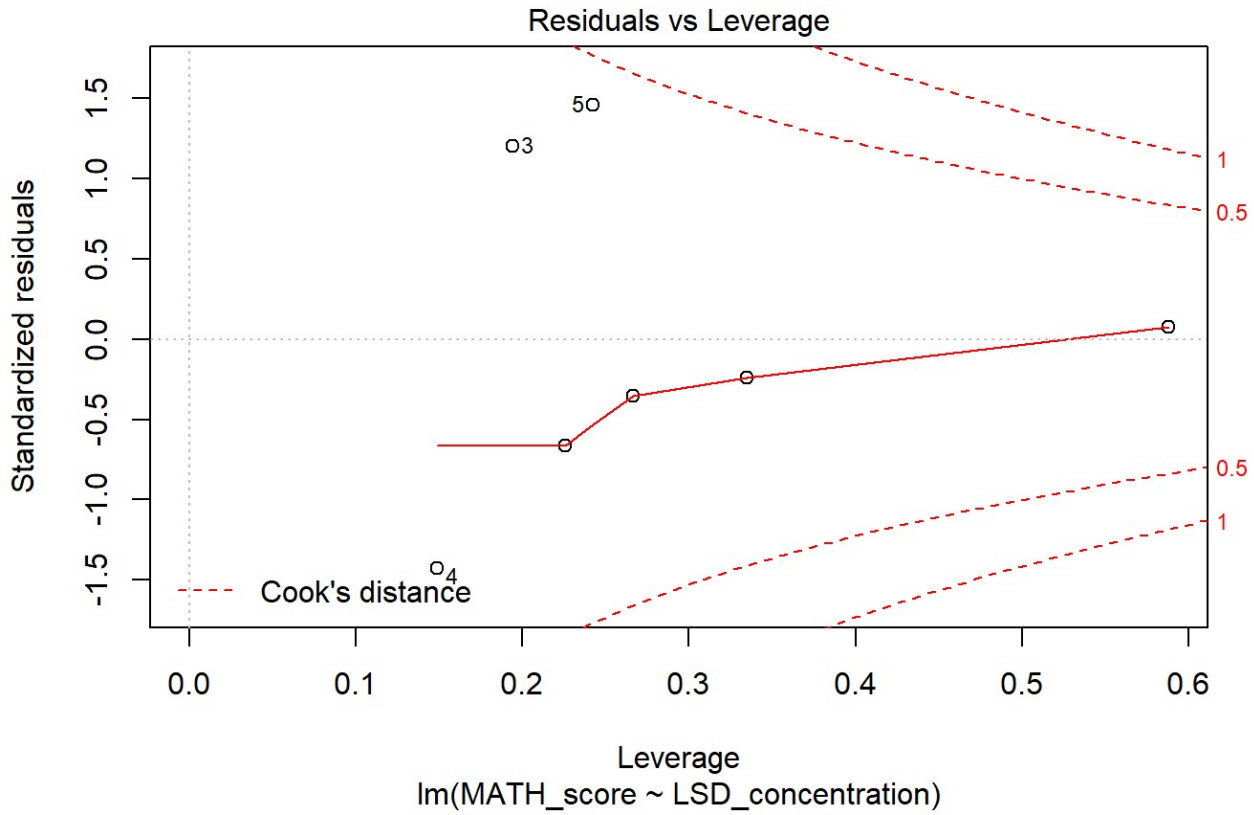
```
## 'data.frame':    7 obs. of  2 variables:
##  $ LSD_concentration: num  1.17 2.97 3.26 4.69 5.83 6 6.41
##  $ MATH_score       : num  78.9 58.2 67.5 37.5 45.6 ...
```

```
#make a model
modmath<-lm(MATH_score~LSD_concentration, data = math)
plot(modmath)
```


Residuals vs Fitted
Residuals
10
5
0
-5
-10
5
3
4
30
40
50
60
70
80
Fitted values
lm(MATH_score ~ LSD_concentration)


Normal Q-Q
Standardized residuals
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
5
3
4
-1.0
-0.5
0.0
0.5
1.0
Theoretical Quantiles
lm(MATH_score ~ LSD_concentration)


Scale-Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
5
4
3
30
40
50
60
70
80
Fitted values
lm(MATH_score ~ LSD_concentration)

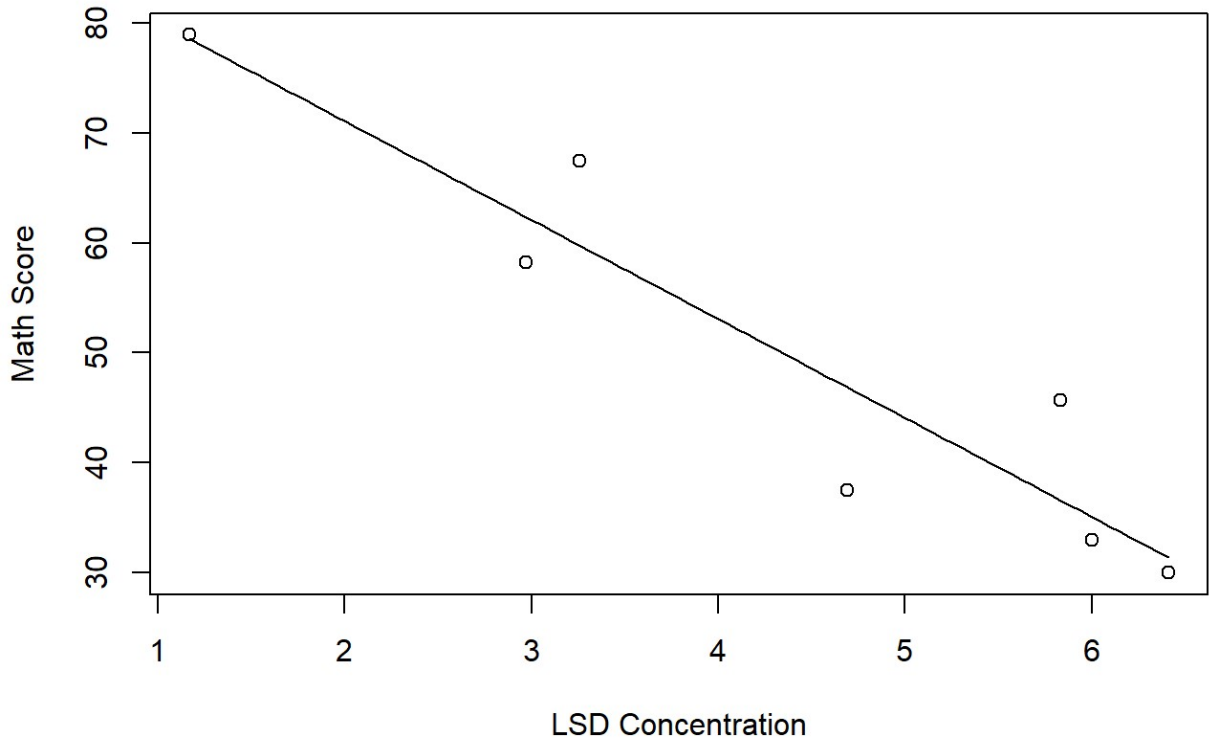



```
#lm is a function to fit linear models, regression, single stratum analysis of varianc
e and analysis of covariance
coef(modmath)
```

```
##       (Intercept) LSD_concentration
##         89.123874         -9.009466
```

a) Make a scatterplot

```
plot(math$MATH_score~math$LSD_concentration, xlab = "LSD Concentration", ylab = "Math
Score")
curve(89.12-9.01*x, add = T)
```



b)parameter estimates and 95%CI

```
coef(modmath)
```

```
##       (Intercept) LSD_concentration
##         89.123874         -9.009466
```

```
confint(modmath)
```

```
##                       2.5 %     97.5 %
## (Intercept)        71.00758 107.240169
## LSD_concentration -12.87325  -5.145685
```

c)calculate R2 or RMSE

```
r2<-function(y_hat, y) {
  TSS<-sum((y-mean(y))^2)
  RSS<-sum((y-y_hat)^2)
  return(1-RSS/TSS)
}

y_hat<-89.12-9.01*math$LSD_concentration

r2(y_hat, math$MATH_score)
```

```
## [1] 0.8778348
```

```
rmse=function(y_hat,y){
  return(sqrt(mean((y-y_hat)^2)))
}

rmse(y_hat, math$MATH_score)
```

```
## [1] 6.022358
```

Question 2: POM miracle food

With the least squares output in mind, do you agree or disagree with the miracle claim?

No, I do not agree with this miracle cure because the confidence interval for weight change crosses zero, which means that the amount of poms per day cannot be said to significantly increase weight loss.

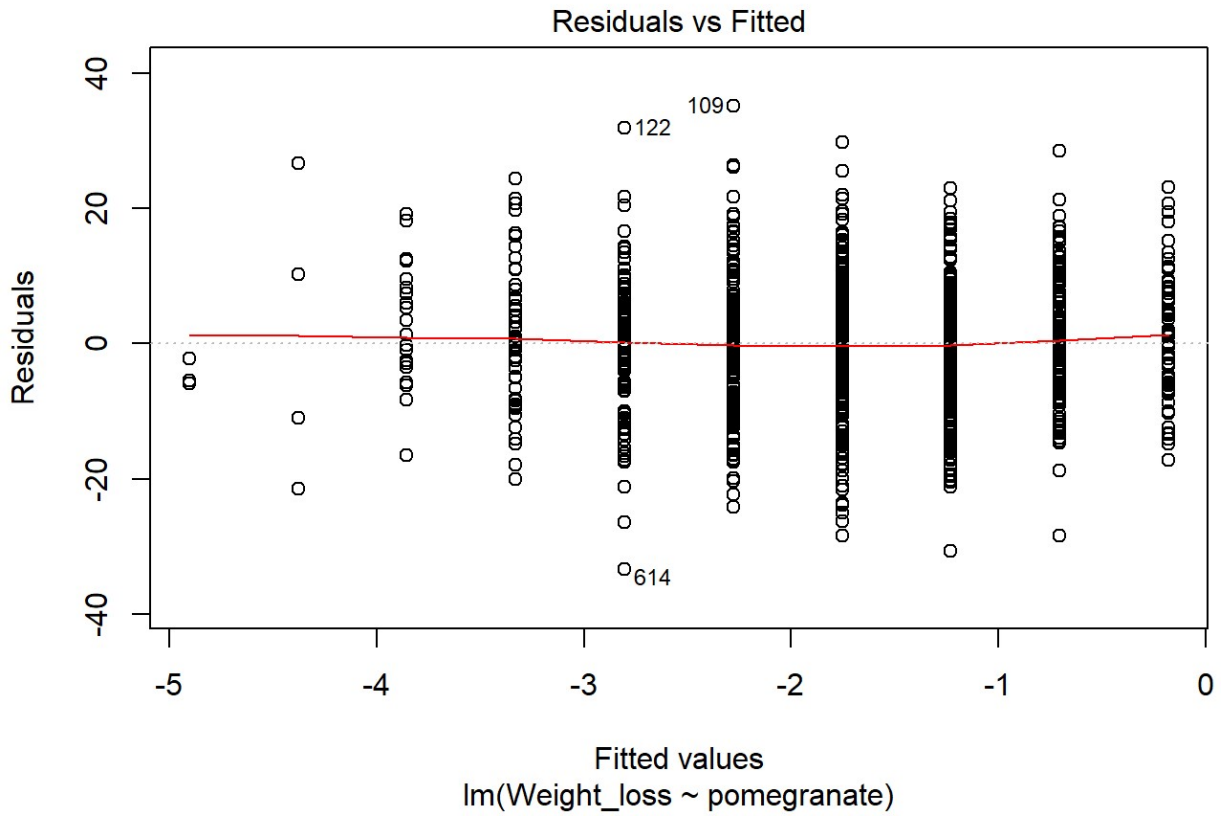
```
food<-read.csv("miracle_food.csv")
head(food)
```

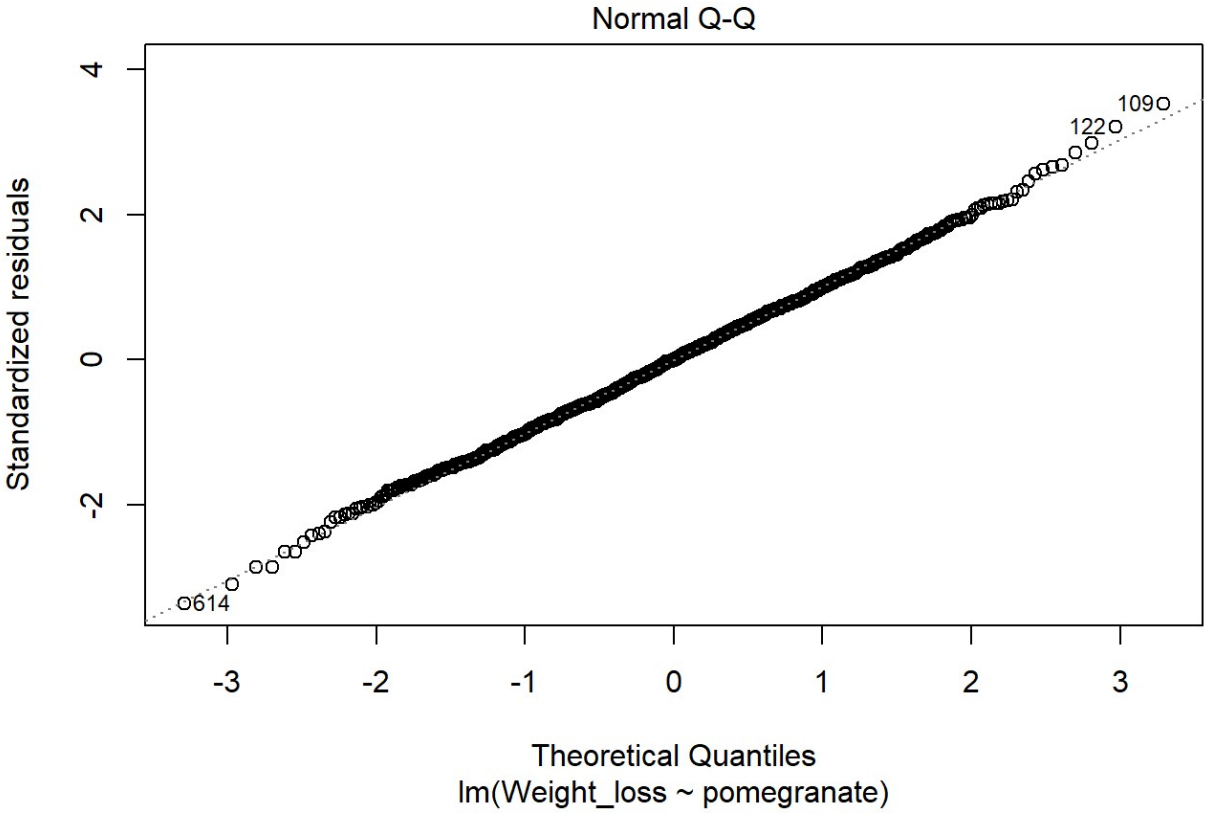
```
##   Weight_loss pomegranate
## 1       -0.89           2
## 2        6.31           2
## 3      -30.21           3
## 4       -6.28           7
## 5       11.38           4
## 6        1.67           2
```

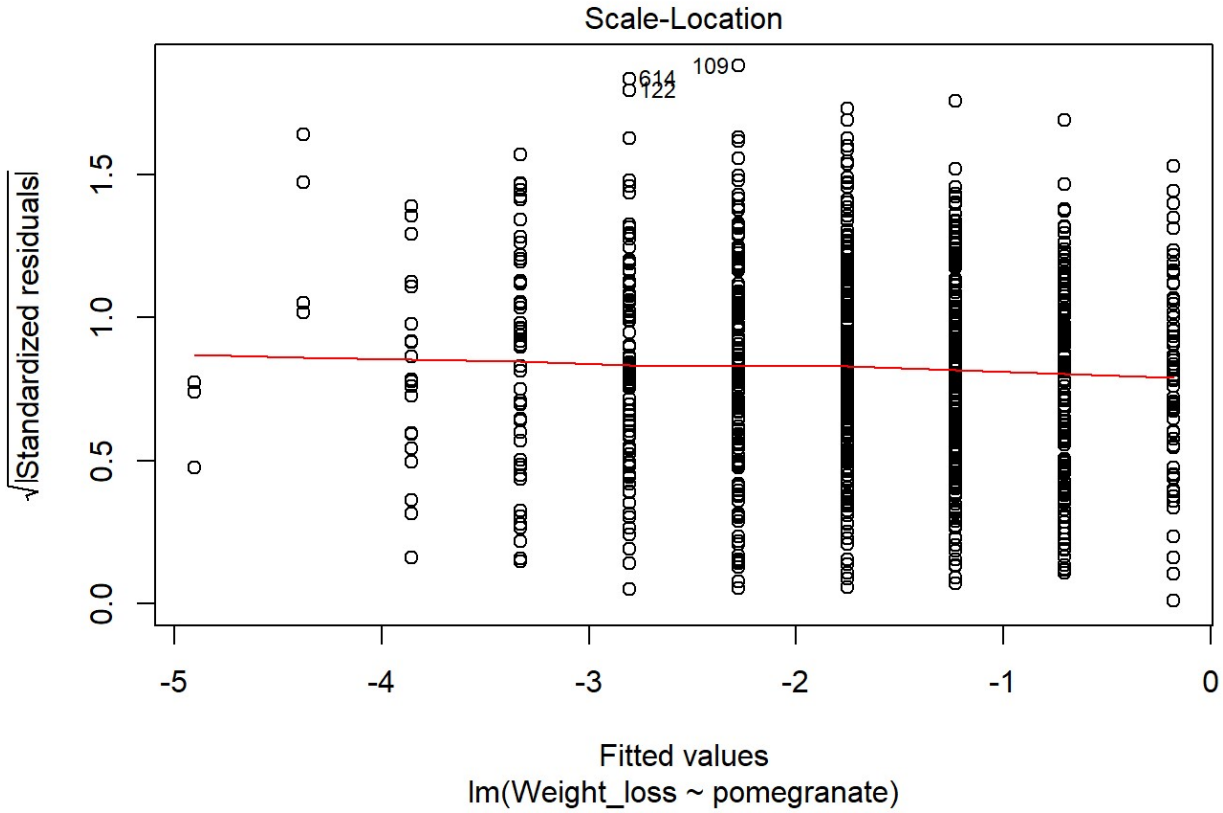
```
str(food)
```

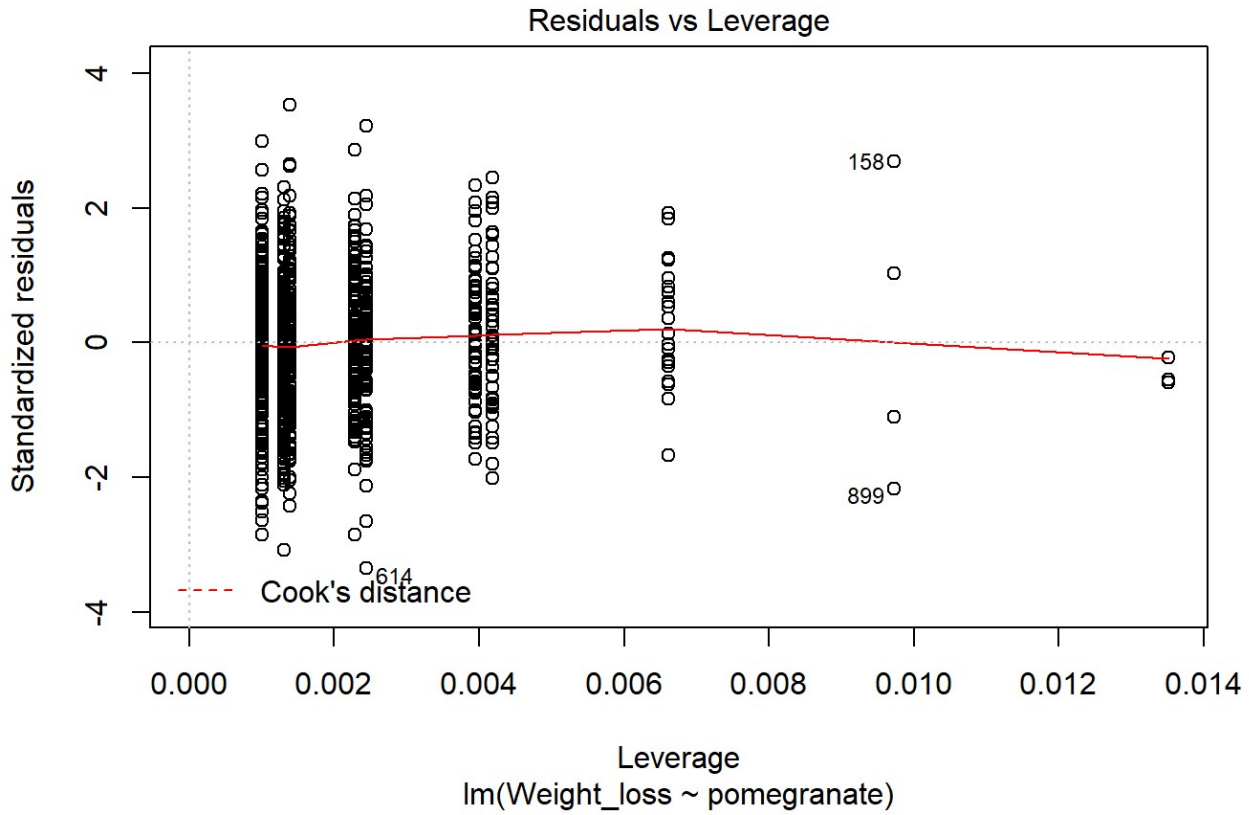
```
## 'data.frame':    1000 obs. of  2 variables:
##  $ Weight_loss: num  -0.89 6.31 -30.21 -6.28 11.38 ...
##  $ pomegranate: int  2 2 3 7 4 2 3 3 5 5 ...
```

```
#make a model
modfood<-lm(Weight_loss~pomegranate, data = food)
plot(modfood)
```




Normal Q-Q
Standardized residuals
4
2
0
-2
109
122
614
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Weight_loss ~ pomegranate)

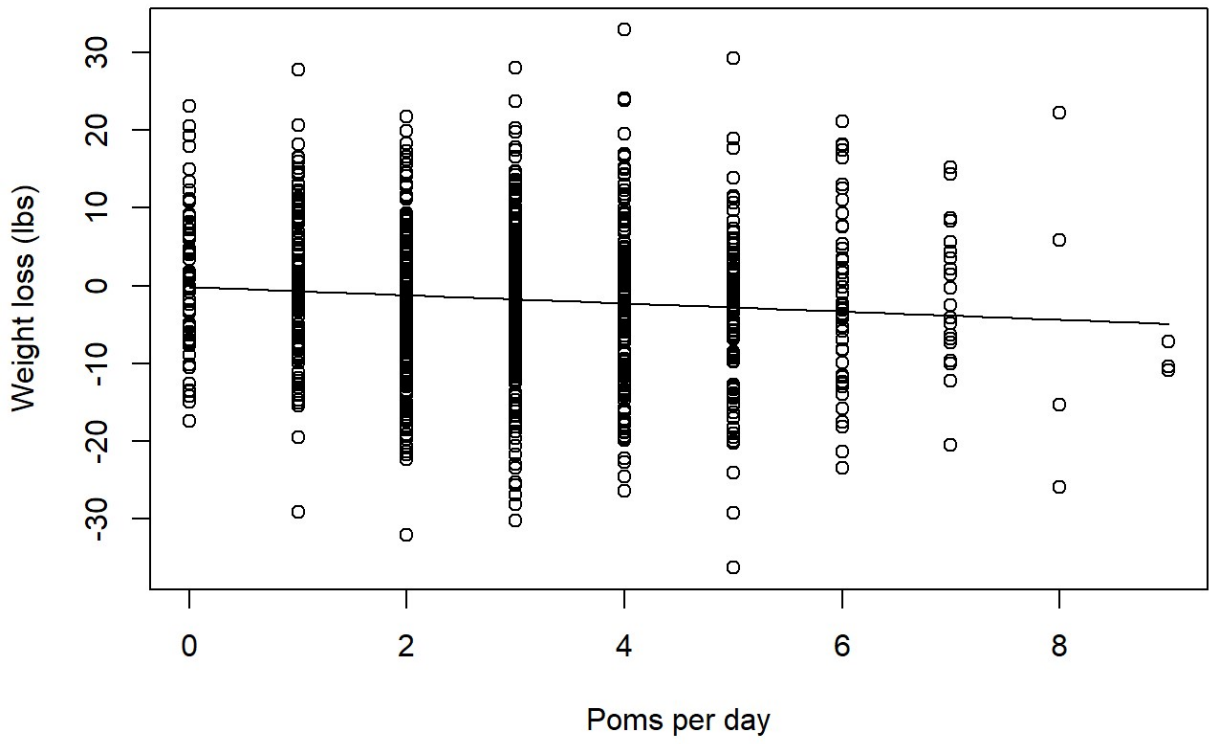

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
614
122
109
-5
-4
-3
-2
-1
0
Fitted values
lm(Weight_loss ~ pomegranate)




```
#lm is a function to fit linear models, regression, single stratum analysis of varianc
e and analysis of covariance
```

a)Scatterplot

```
plot(food$Weight_loss~food$pomegranate, xlab = "Poms per day", ylab = "Weight loss (lb
s)")
curve(-0.178-0.525*x, add = T)
```



b)parameter estimates and CI-95

```
coef(modfood)
```

```
## (Intercept) pomegranate
##  -0.1789802  -0.5251053
```

```
confint(modfood)
```

```
##                 2.5 %     97.5 %
## (Intercept) -1.408937  1.0509767
## pomegranate -0.886420 -0.1637906
```

c)R2 and RMSE

```
y_hat1<- -0.178-0.525*food$pomegranate

r2(y_hat1, food$Weight_loss)
```

```
## [1] 0.008083795
```

```
rmse(y_hat1, food$Weight_loss)
```

```
## [1] 9.961044
```

Question 3: Mean Absolute Error

A) Translate the mathematical equation for MAE into a function in R.

```
MAE<-function(y,x) {
 temp_lm<-lm(y~x)
 int <- coef(temp_lm)[1]
 slope <- coef(temp_lm)[2]
  y_hat<-int+slope*x
  n <- length(y)
  arss<-sum(abs(y-y_hat), na.rm = T)/n
  return(arss)
}
```

B) Compare RMSE, R2, and MAE for the linear models in questions 1 and 2. How do these metrics of model fit differ?

R2 is the coefficient of determination which measures the proportion of total variation explained by the model. RMSE is the squared difference between values estimated by the model and observed values–it is a measure of error (it is scale dependent and effected by large numbers).Mean Absolute Error is the average absolute difference between predicted values and observed, true values and is reported in the same scale as the data was measured.

LSD and Math: R2: 0.877 RMSE: 6.022 MAE: 4.89

Poms and Weight: R2: 0.008 RMSE: 9.96 MAE: 7.98

Side note from Josh: if RMSE ~= MAE then you have many, small errors. If RMSE ~= MAE^2 then you have a few large errors

```
#MAE for food and math
MAE(food$Weight_loss,food$pomegranate)
```

```
## [1] 7.981461
```

```
MAE(math$MATH_score,math$LSD_concentration)
```

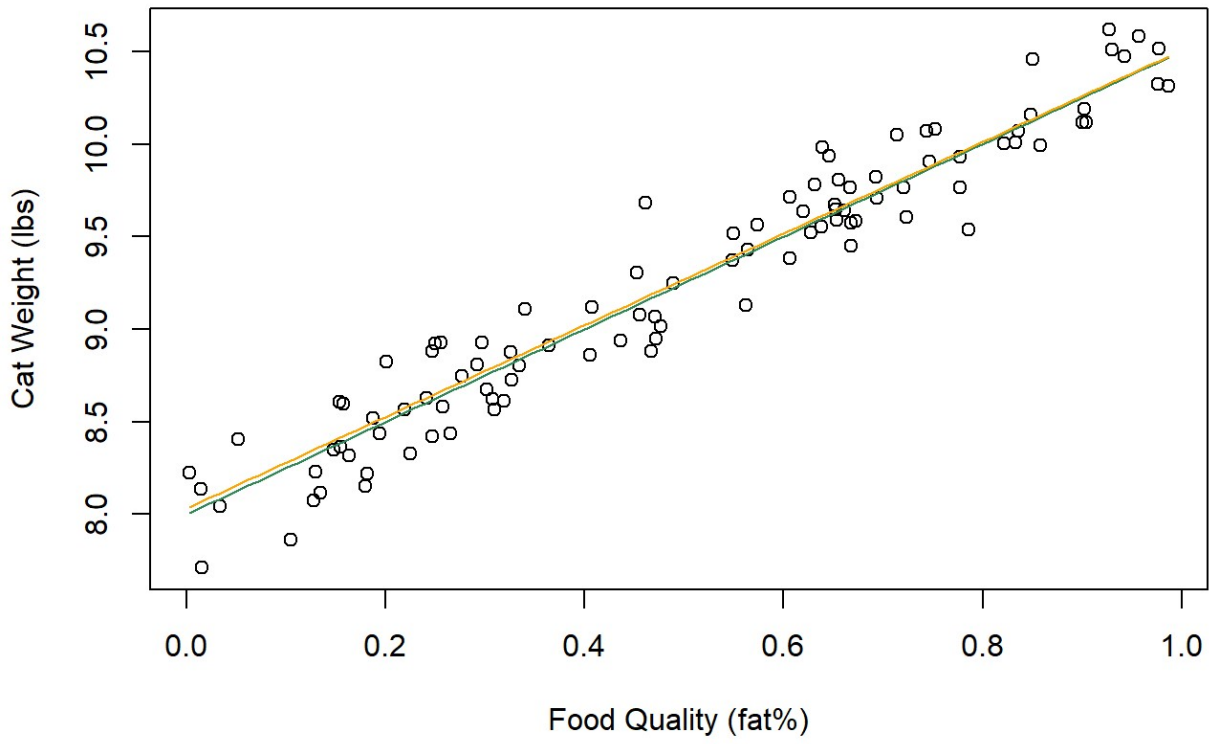
```
## [1] 4.890144
```

Question 4: Simulate Linear Data

```
###Step 1###
#generate predictor variable and data
food_qual<-runif(100)

###Step 2###
#create slope, intercept, and sigma
slope<-2.5
intercept<-8

###Step 3###
#use rnorm
cat_weight<-rnorm(n= 100, mean=intercept+slope*food_qual, sd=0.2)

#a
plot(cat_weight~food_qual, xlab = "Food Quality (fat%)", ylab = "Cat Weight (lbs)")

#b
catmod<-lm(cat_weight~food_qual)
coef(catmod)
```

```
## (Intercept)   food_qual
##    7.972716    2.556747
```

```
#c How do your estimates compare to the true values you came up with in step 2?
# The estimates were very close
curve(8+2.5*x, add = T, col= "seagreen")   #true values
curve(8.032+2.475*x, add = T, col= "orange")  #estimated parameters
```



Question 5: Simulate Unequal Variance

B

When turtles recieve larger doses of LSD, individual responses begin to vary more because the drug interacts with the brain chemical of each turtle differently. As a result, at higher doses, some turtles may be more excited and move faster while others feel more serene and move more slowly.

```
###Step 1###
#generate predictor variable and data
lsd_dose<-runif(100)

###Step 2###
#create slope, intercept, and sigma
slope1<-3
intercept1<-15

###Step 3###
#use rnorm
turtle_spd<-rnorm(n= 100, mean=intercept+slope*lsd_dose, sd=0.5*lsd_dose)

#a
plot(turtle_spd~lsd_dose, xlab = "LSD Dose (mg)", ylab = "Turtle Speed (cm/s)")
```