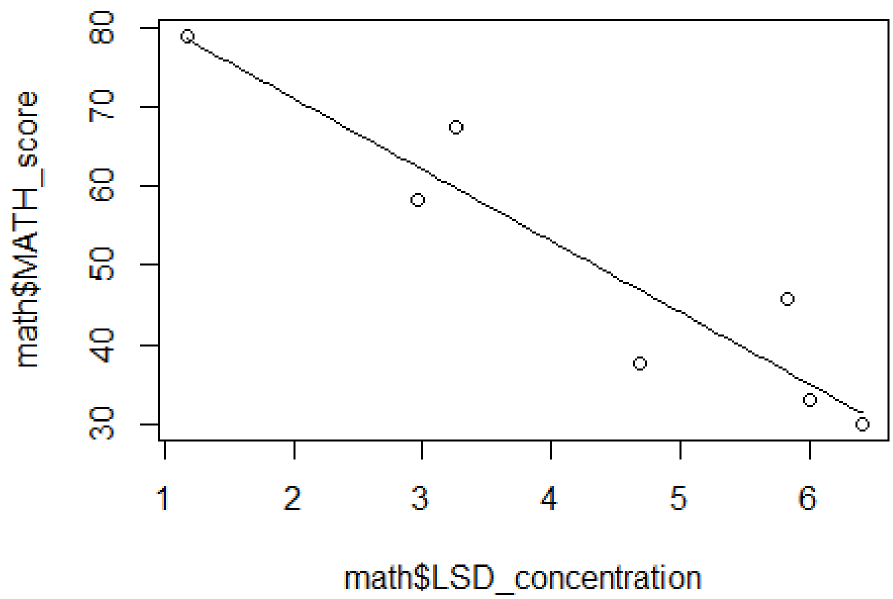
Jamie Burke
Ecostats_hw1.R

```
####Q1####

math<-read.csv("math_scores_hw1.csv")
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
## 7              6.41      29.97

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 ...

summary(math)

##  LSD_concentration   MATH_score
##  Min.   :1.170     Min.   :29.97
##  1st Qu.:3.115     1st Qu.:35.20
##  Median :4.690     Median :45.65
##  Mean   :4.333     Mean   :50.09
##  3rd Qu.:5.915     3rd Qu.:62.84
##  Max.   :6.410     Max.   :78.93

mathlm<-lm(math$MATH_score~math$LSD_concentration)

plot(math$MATH_score~math$LSD_concentration)
coef(mathlm)

##            (Intercept) math$LSD_concentration
##              89.123874              -9.009466

curve(89.123874+x*-9.009466, add=TRUE)
```




```r
confint(mathlm)

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

a<-89.123874
b<--9.009466
yhatmath<-a+b*math$LSD_concentration
yhatmath

## [1] 78.58280 62.36576 59.75301 46.86948 36.59869 35.06708 31.37320

r2<-function(yhat,y){
  RSS<-sum((((yhat))-(y))^2)
  TSS<-sum(((y)-(mean(y)))^2)
  return(1-RSS/TSS)}
r2(yhatmath, math$MATH_score)

## [1] 0.877835

rmse=function(yhat,y){return(sqrt(mean((y-yhat)^2)))}
rmse(yhatmath, math$MATH_score)

## [1] 6.022355
```

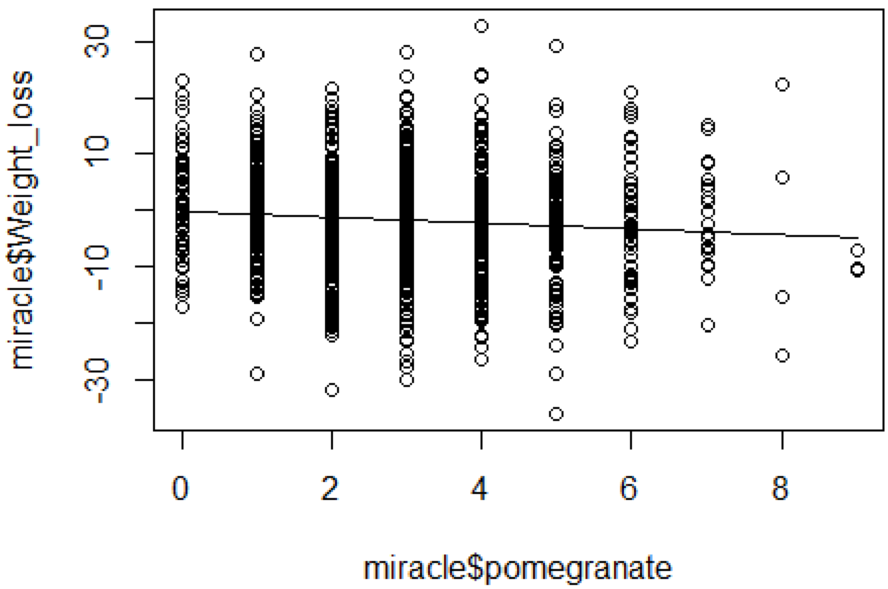
A. With an intercept of 89.123, a score of 85 would need a concentration of between zero and one just by looking at the graph. Using the equation, we can determine that a score of 85= 89.123874+ -9.00946*X so x= 0.4582. So a concentration of 0.4582 or less will ensure a score of 85.

B. With an r2 value of 0.8777835, LSD concentration does predict math scores well.

C. The normal distribution is probably inappropriate for this data because you have a small sample size so it is unlikely you actually have normally distributed data.

```
#####Q2####

miracle<-read.csv("miracle_food_hw1.csv")
head(miracle)

##   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(miracle)

## '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 ...

summary(miracle)

##   Weight_loss       pomegranate
##  Min.   :-36.240   Min.   :0.000
##  1st Qu.: -8.570   1st Qu.:2.000
##  Median : -1.650   Median :3.000
##  Mean   : -1.724   Mean   :2.942
##  3rd Qu.:  5.037   3rd Qu.:4.000
##  Max.   : 32.890   Max.   :9.000

miraclelm<-lm(miracle$Weight_loss~miracle$pomegranate)

coef(miraclelm)

##         (Intercept) miracle$pomegranate
##          -0.1789802          -0.5251053

confint(miraclelm)

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

```r
plot(miracle$Weight_loss~miracle$pomegranate)
curve(-0.1789802+-0.5251053*x, add=TRUE)
```




```r
c<--0.1789802
d<--0.5251053
yhatmiracle<-c+d*miracle$pomegranate


r2<-function(yhat,y){
  RSS<-sum((((yhat))-(y))^2)
  TSS<-sum(((y)-(mean(y)))^2)
  return(1-RSS/TSS)}
r2(yhatmiracle, miracle$Weight_loss)

## [1] 0.008083812

rmse=function(yhat,y){return(sqrt(mean((y-yhat)^2)))}
rmse(yhatmiracle, miracle$Weight_loss)

## [1] 9.961044
```

The r2 value us very low so number of pomegranates eaten does not predict weight loss very well and from this data, it would be hard to sell pomegranates as a miracle weight loss food.

```
####Q3####

MAE<-function(yhat,y){return(mean(abs(y-yhat)))}
  MAE(yhatmath,math$MATH_score)
```

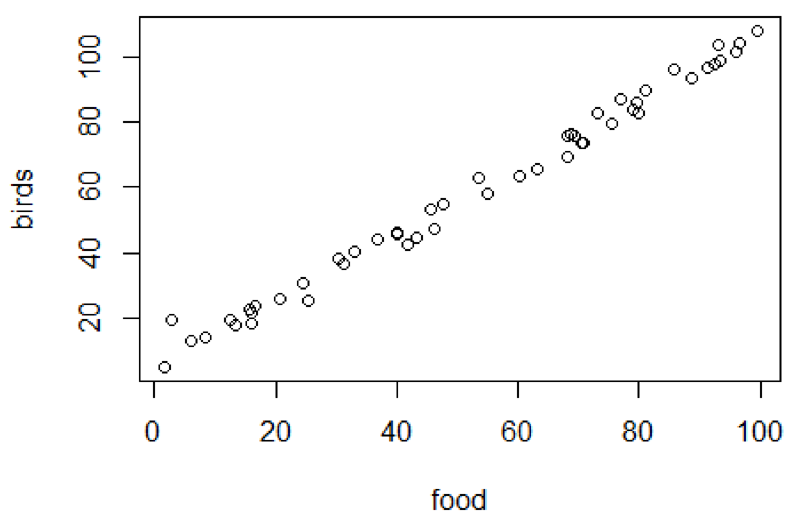
```
## [1] 4.890145
```

```
  MAE(yhatmiracle,miracle$Weight_loss)
```

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

For both the miracle weight loss data, the MAE is high given the values for x and the slope or the increase in y for every 1 unit of x. For the math scores, the MAE is lower compared to values of x and the mean of the samples. The RMSE and the MAE give an measurement of how well the model fits the data, but in the original units. The r2 value measures how well the data fits the model but compares the data to the model or in this case the regression line and how much error is accounted for by the model.

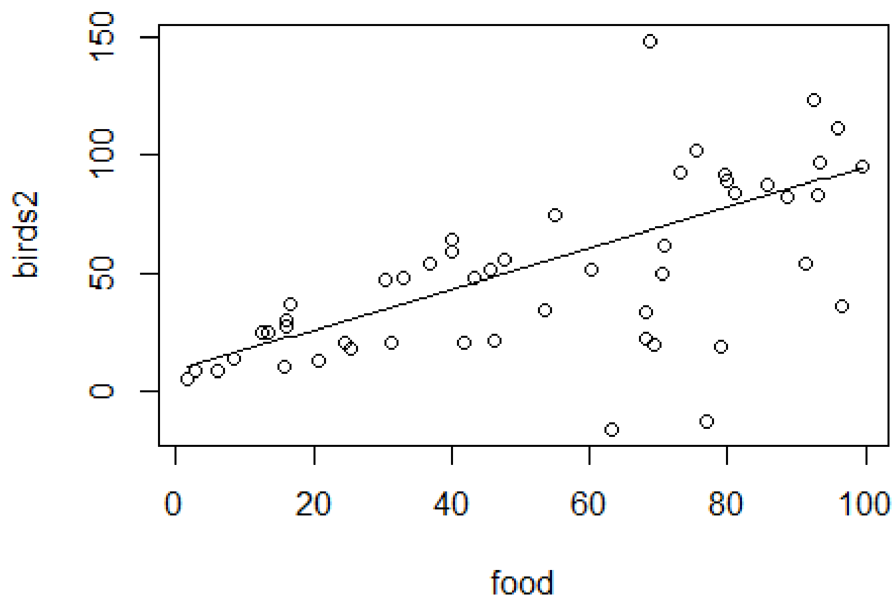
```
####Q4####

food<-runif(50, min=0, max=100)

slope_hw1Q4<-1
intercept_hw1Q4<-5

birds<-rnorm(n=50, mean=intercept_hw1Q4+slope_hw1Q4*food, sd=3)

plot(birds~food)
```



```
birdslm<-lm(birds~food)

coef(birdslm)

## (Intercept)        food
##    5.726237    1.000876

confint(birdslm)

##                 2.5 %   97.5 %
## (Intercept) 3.9530685 7.499405
## food        0.9716278 1.030124
```

The estimates from the model are very close to the original values set for the intercept and slope although the slope was a bit closer to the original than the intercept estimate.

```
####Q5####

sd_hw1Q4<-food*0.5

birds2<-rnorm(n=50, mean=intercept_hw1Q4+slope_hw1Q4*food, sd_hw1Q4)
birds2

plot(birds2~food)
birds2lm<-lm(birds2~food)
coef(birds2lm)

## (Intercept)        food
##   9.4104200   0.7483048

curve(8.9978391+0.8596975*x, add=TRUE)
```



As food increases, the number of birds increase as well but with a high supply of food, the number of birds varies more than with a lower amount of food. A in crease in SD as a function of the predictor may mean that an increase in the predictor can increase not just birds but other factors in the ecosystem which in turn could change the effect of food on number of birds.