HW 1

Kristina Parker

1/22/2019

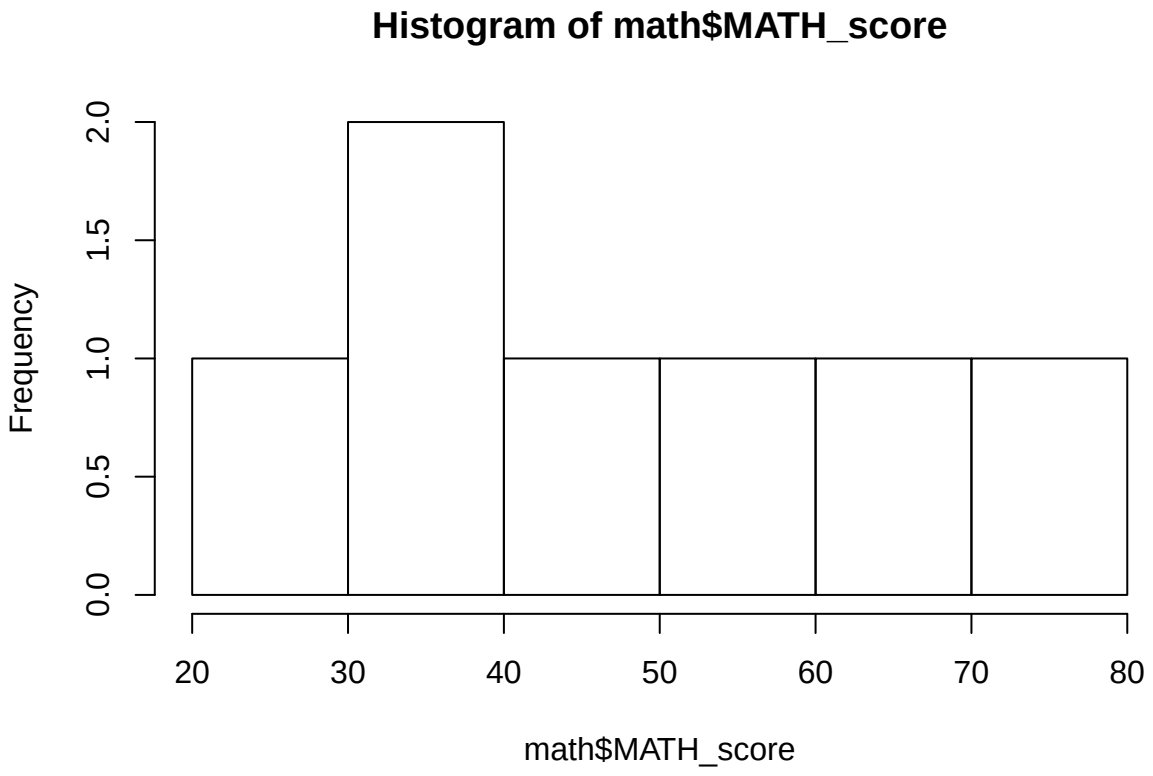
```
#r^2 function
r2 <- function(y_hat, y) {
  RSS <- sum((((y_hat)) - (y))^2)
  TSS <- sum(((y) - (mean(y)))^2)
  return(1 - RSS / TSS)
  }
#RMSE function
rmse = function(y_hat, y) {
  return(sqrt(mean((y - y_hat)^2)))
}
```

1.

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

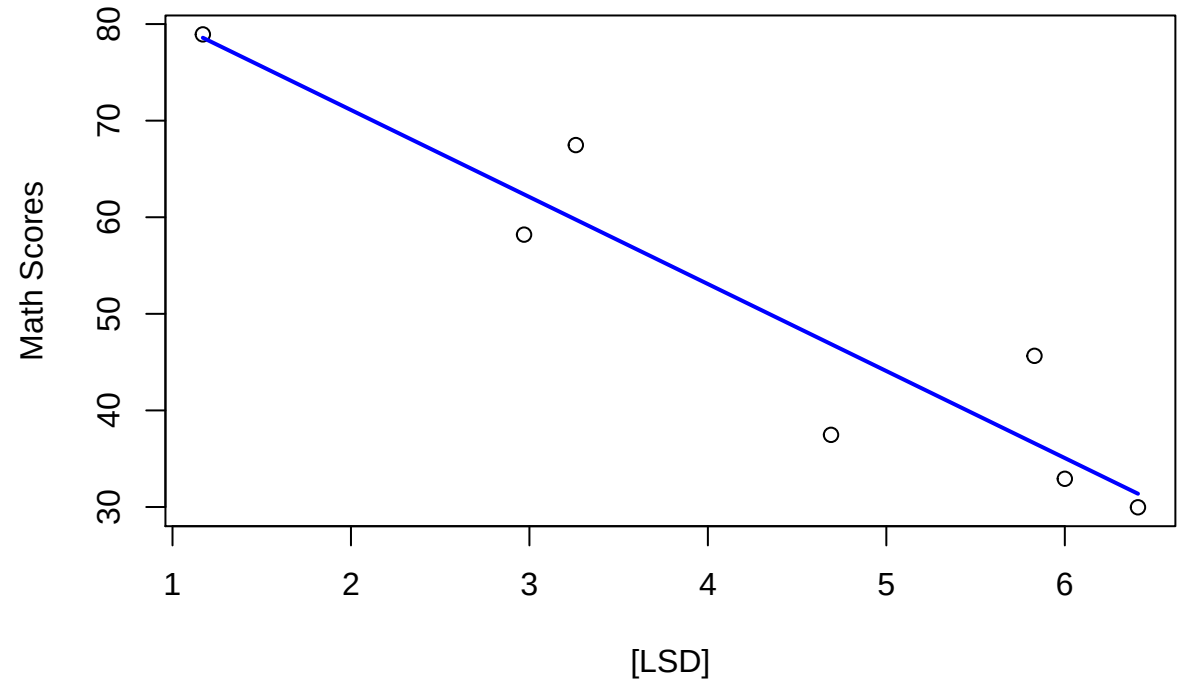
```
##   LSD_concentration MATH_score
## 1              1.17      78.93
## 2              2.97      58.20
## 3              3.26      67.47
## 4              4.69      37.47
## 5              5.83      45.65
## 6              6.00      32.92
```

```
hist(math$MATH_score)
```



a)

```
plot(math$MATH_score ~ math$LSD_concentration,
     xlab = "[LSD]", ylab = "Math Scores")
curve(89.123874 - 9.009466 * x, add = T, col = "blue", lwd = 2)
```



d)

```
modmath <- lm(math$MATH_score ~ math$LSD_concentration)
y_hat = 89.123874 - 9.009466 * math$LSD_concentration
```

```
r2_m = r2(y_hat, math$MATH_score)
# line of best fit suggests model predicts the data well
rmse_m = rmse(y_hat, math$MATH_score)
# residual mean square error suggest that the model does not fit the data well due to having high varia
```

c)

```
#parameter estimates
coef(modmath)
```

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

```
#95% CI
confint(modmath)
```

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

A)

```
LSDconc <- (85 - 89.123874)/9.009466
LSDconc
```

```
## [1] -0.4577268
```

Based on the model the concentration of LSD in tissue has to be <-0.4577268 for a math score >85%.

B) The level of LSD tissue concentration does not predict math test score well. Although, the linear model has an r^2 value of 0.878, inferring that LSD concentration predicts test performance, the RMSE = 6.022 shows that this model does not fit the data well becuase there is a lot of residual error (variance) between the data and predicted model.

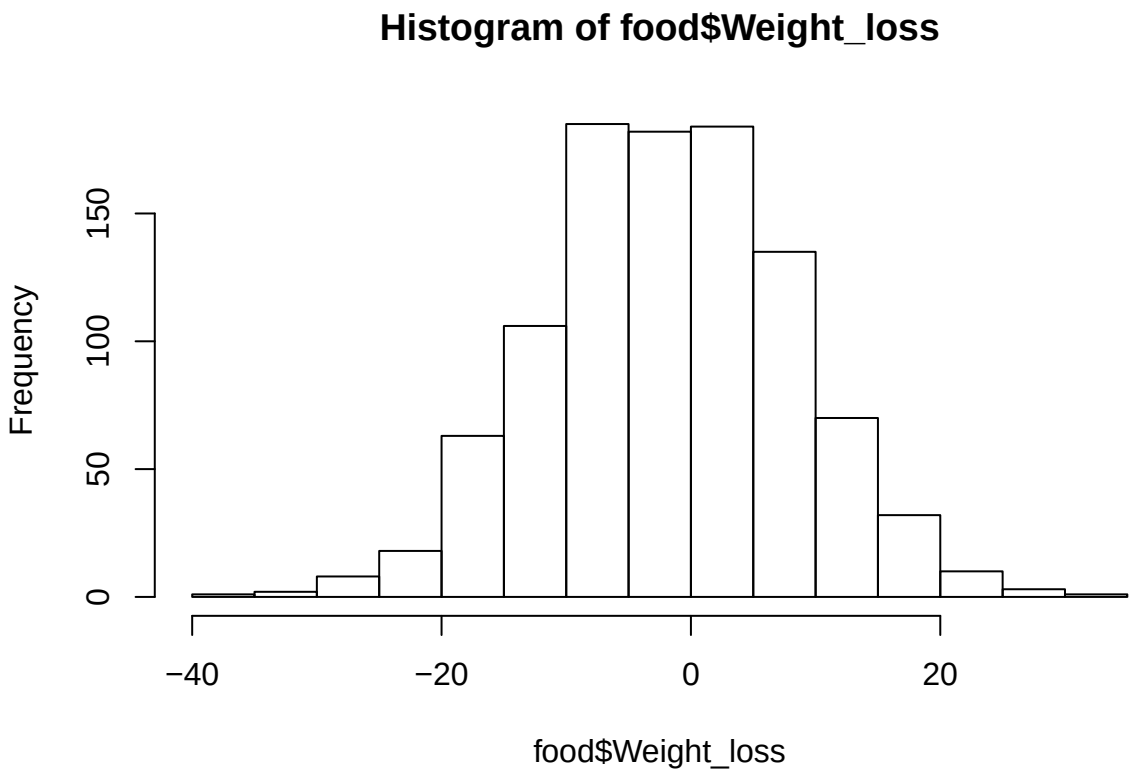
C) Normal distribution model is an inappropriate model because the distribution of the data isn't normal, as it is in many natural experiments. The small sample size is skewed causing the model to poorly predict test performances based on LSD tissue concentration.

2.

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

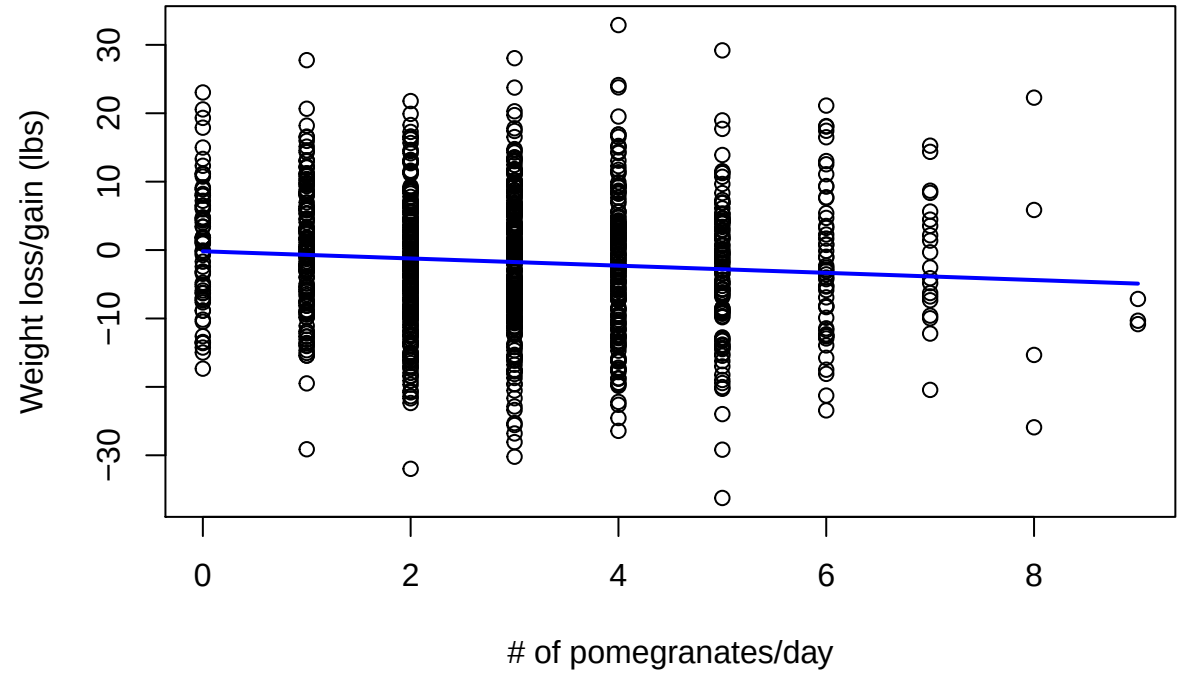
```
##   Weight_loss pomegranate
## 1       -0.89           2
## 2        6.31           2
## 3      -30.21           3
## 4       -6.28           7
## 5       11.38           4
## 6        1.67           2
```

```
hist(food$Weight_loss)
```



a)

```
plot(food$Weight_loss ~ food$pomegranate,
     xlab = "# of pomegranates/day", ylab = "Weight loss/gain (lbs)")
curve( -0.1789802 - 0.5251053 * x, add = T, col = "blue", lwd = 2)
```



b)

```
modfood <- lm(food$Weight_loss ~ food$pomegranate)
summary(modfood)
```

```
##
## Call:
## lm(formula = food$Weight_loss ~ food$pomegranate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.435  -6.780  -0.041   6.807  35.169
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.1790     0.6268  -0.286  0.77528
## food$pomegranate  -0.5251     0.1841  -2.852  0.00444 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.971 on 998 degrees of freedom
## Multiple R-squared:  0.008084,   Adjusted R-squared:  0.00709
## F-statistic: 8.133 on 1 and 998 DF,  p-value: 0.004435
```

```
Y_hat <- -0.1789802 - 0.5251053 * food$pomegranate
r2_f = r2(Y_hat, food$Weight_loss)
rmse_f =rmse(Y_hat, food$Weight_loss)
```

c)

```
#parameter estimates
coef(modfood)
```

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

```
# 95% CI
confint(modfood)
```

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

A) I disagree with the miracle claim of pomegranate promotes weight loss. The effect of pomegranate on weight is significant but a linear model isn't the right model to show effect of pomegranate consumption. Although, the data is normally distributed the data has high variance (RMSE = 9.961044) and the data is non-linear so there is no linear relationship predicted (r^2 = 0.008).

3.

A.

```
# Function for calculating MAE
mae <- function(y_hat, y) {
 return(mean(abs(y - y_hat)))
}
#calculation of MAE for question 1 & 2
mae_m = mae(y_hat, math$MATH_score)
mae_f = mae(Y_hat, food$Weight_loss)
```

B.

Table 1: Comparison of different metrics to assess model fit.

Model.Evaluation.Metrics	Question.1	Question.2
r^2	0.877835	0.0080838
RMSE	6.022355	9.9610444
MAE	4.890145	7.9814607

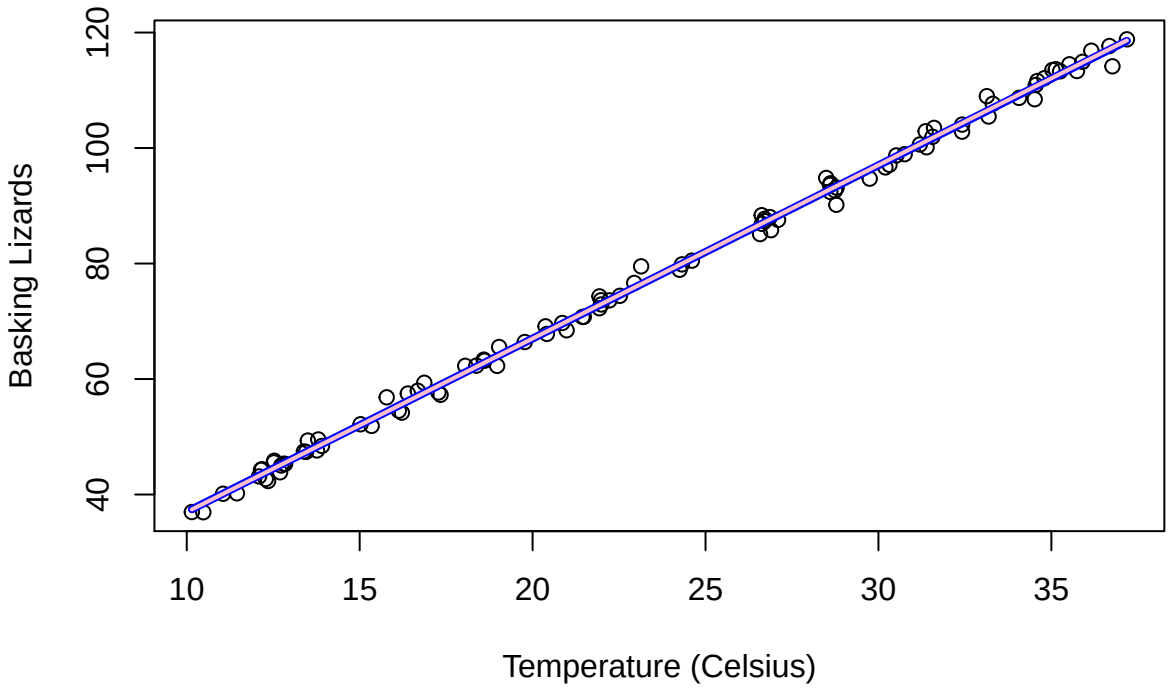
In question 1 the r^2 value is high, meaning the model predicts a lot of the variance in the total variance of the data. However, the model has high RMSE and MAE, meaning there is a high amount of error (large differences between the observed and predicted values). Although, the r^2 said the model predicts the model closely, there is actually a lot of error in the predictability of the model. The MAE is less than the RMSE because this is the overall error between the observed and predicted values. RMSE puts more weight to large errors (differences in values), meaning there is higher error to values further from the mean. In question 2 the linear model is shown to not be well suited for this data in all evaluation metrics. The r^2 is very low indicating that the model poorly predicts the data because little variance is explained. There is high amounts of error in both the RMSE and MAE; again higher RMSE because a heavier weight is placed on values with high differences.

4.

```r
#Simulate linear data
temperature <- runif(100, min = 10, max = 38)
intercept <- 7
slope <- 3
sigma <- 1.2
lizards <- rnorm(100, mean = intercept + temperature*slope, sd = sigma)
co <- coef(lm(lizards ~ temperature))
```

A.

```r
plot(lizards ~ temperature,
     xlab = "Temperature (Celsius)", ylab = "Basking Lizards")
curve(7 + 3*x, add = T, col = "blue", lwd = 4)
curve( co[1] + co[2] * x, add = T, col = "pink", lwd = 2)
```

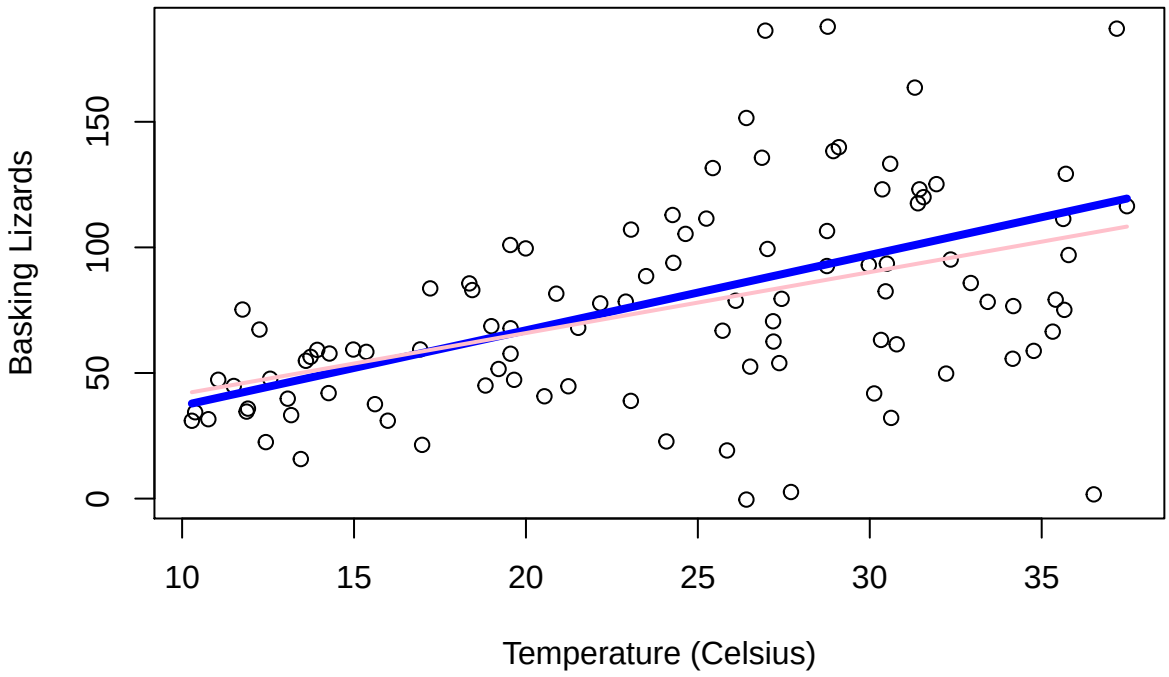


C. The model estimates are close to the values generated because the sigma value is low. There is little unexplain variance allowing for a line of best fit to model the data set well.

5.

A.

```
temperature <- runif(100, min = 10, max = 38)
intercept <- 7
slope <- 3
sigma <- 1.2 * temperature
lizards <- rnorm(100, mean = intercept + temperature*slope, sd = sigma)
plot(lizards ~ temperature,
     xlab = "Temperature (Celsius)", ylab = "Basking Lizards")
coef <- coef(lm(lizards ~ temperature))
curve(7 + 3 * x, add = T, col = "blue", lwd = 4)
curve( coef[1] + coef[2] * x, add = T, col = "pink", lwd = 2)
```




```
r2((7 + 3 * temperature), lizards)
```

```
## [1] 0.2035117
```

```
rmse((7 + 3 * temperature), lizards)
```

```
## [1] 35.81971
```

B. A potential biological explanation is that as temperatures increase there is higher variance in the number of basking lizards. As the sun rises lizards begin to bask to increase their body temperature so energy can be used towards foraging. As temperatures increase less lizards are found basking because their body temperature has reached capacity and left. Temperature and the number of basking lizards are correlated; however, this is not a linear relationship. The r^2 = 0.3457 and RMSE = 29.04 indicating that the linear model does not fit the data; there is a lot of unexplained variance and error in th model. The response variable increases in variance as the predictor variable increases, this is requires a different model than a linear regression.