Juliette HW 1

Juliette J. Rubin

January 23, 2019

Question 1

Math scores and LSD

```
math <- read.csv("math_scores.csv")
model1 <- lm(math$MATH_score ~ math$LSD_concentration)
coef(model1)
```

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

```
confint(model1)
```

```
                        2.5 %     97.5 %
```

(Intercept) 71.00758 107.240169 math$LSD_concentration -12.87325 -5.145685

```
# model fit with the given functions
yhat <- 89.12 + (-9.009) * math$LSD_concentration
y <- math$MATH_score

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

r2(yhat, y)
```

[1] 0.877835

```
# result: 0.878

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

[1] 6.022355

```
# result: 6.02
```

We find that the equation of this line is comprised of the following values: Intercept = 89.12, Slope=-9.009 and the metric of model fit as determined by our functions above is R^2=0.878. Additionally, our confidence interval output indicates that there is a sig effect of LSD concentration on math scores, as 95% of the time we are going to get an effect of LSD on math scores.

1B) Although our math score data seems to be tightly associated with LSD intake (r2: 0.878), we can't say much about its predictive qualities because r^2 says little about prediction error, and thus we could be wrong

1C) The normal distribution might not be acceptable for these data because grades do not go to positive and negative infinity. There are set bookends.

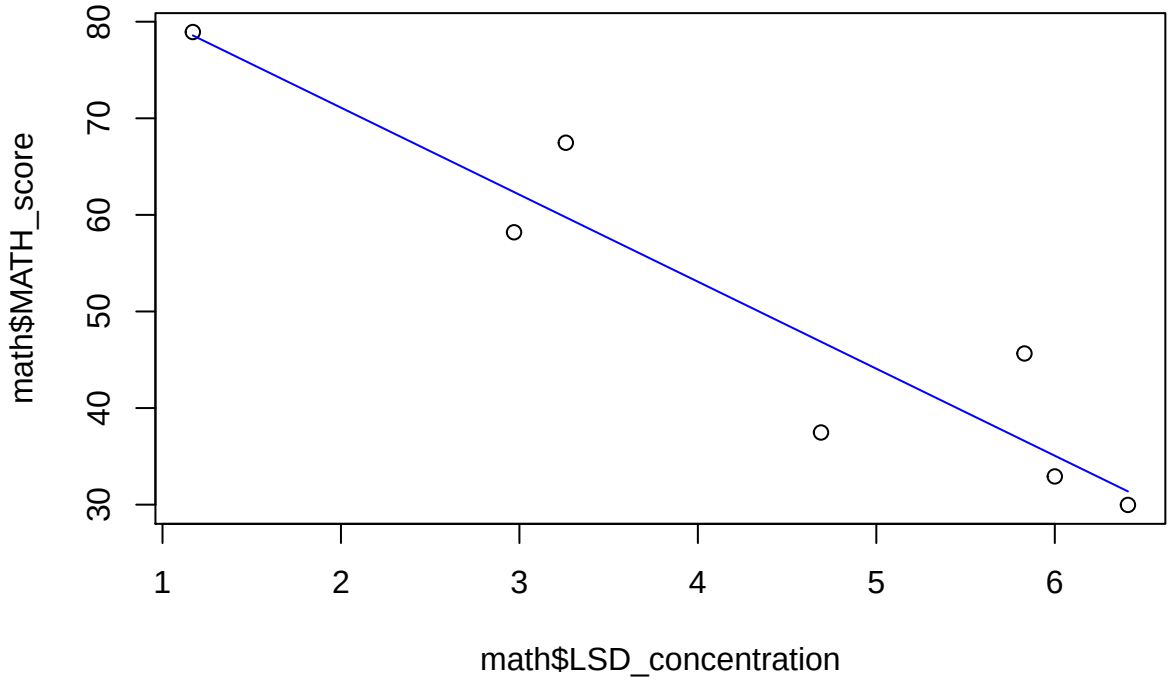
```r
# y=mx+b
(85 - 89.12)/-9.009
```

```
## [1] 0.4573205
```

```r
# result = 0.46 LSD concentration
```

Math score plot

```r
plot(math$MATH_score ~ math$LSD_concentration)
curve(89.12 + (-9.009) * x, add = T, col = "blue")
```



Question 2

Pomegranate miracle food

```r
miracle <- read.csv("miracle_food.csv")
model2 <- lm(miracle$Weight_loss ~ miracle$pomegranate)
coef(model2)
```

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

```r
confint(model2)
```

```
##                          2.5 %     97.5 %
```

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

```
yhat <- -0.179 + (-0.525) * miracle$pomegranate
y <- miracle$Weight_loss

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

r2(yhat, y)
```

```
## [1] 0.008083811
```

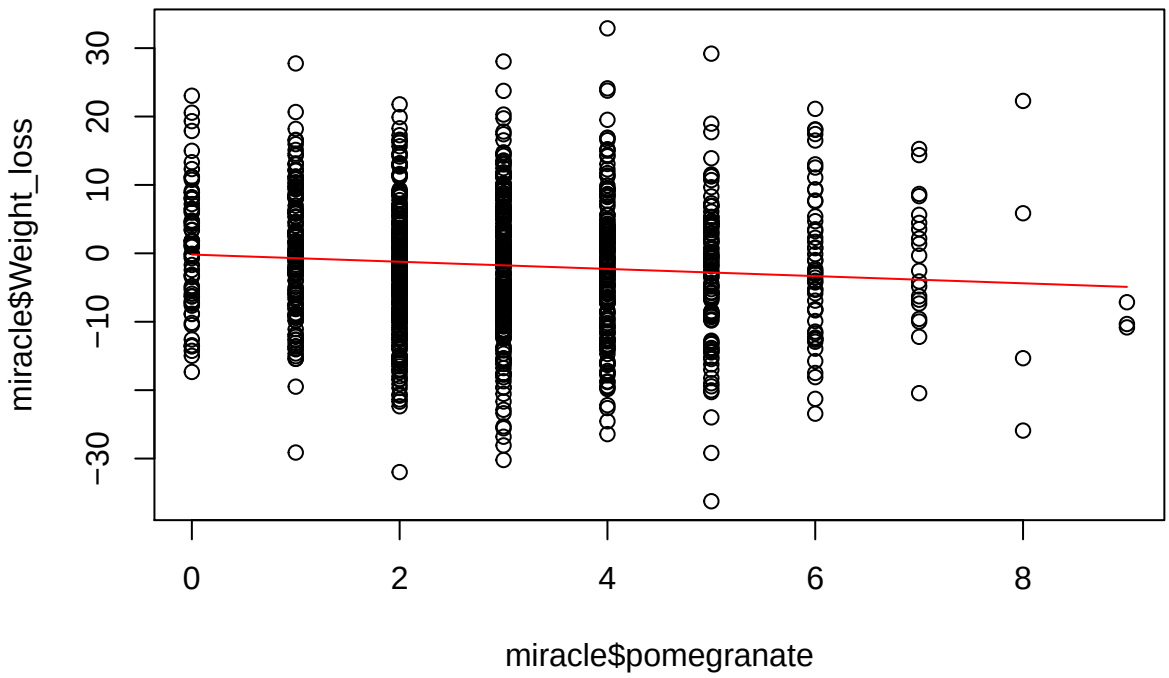
```
# r^2 = 0.008

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

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

```
# result:9.96
```

We find that the equation of this line is comprised of the following values: intercept=-0.179, slope=-0.525. Our confidence intervals do not indicate that this is a significant effect and our r^2 value is quite low, meaning that our model is worse than just assessing mean weight loss. I therefore cannot agree with the farmer's claim (though I don't dismiss its efficacy as an ad campaign).

```
plot(miracle$Weight_loss ~ miracle$pomegranate)
curve(-0.179 + (-0.525 * x), add = T, col = "red")
```



Question 3

```
# MAE of math data
yhat <- 89.12 + -9.009 * math$LSD_concentration
y <- math$MATH_score

MAE <- function(yhat, y) {
    return(mean(abs(y - yhat)))
}

MAE(yhat, y)
```

```
## [1] 4.890244
```

```
# result: 4.89


# MAE of pomegranate data
yhat <- -0.179 + -0.525 * miracle$pomegranate
y <- miracle$Weight_loss

MAE <- function(yhat, y) {
    return(mean(abs(y - yhat)))
}
```

```r
MAE(yhat, y)
```

```
## [1] 7.981464
```

```r
# result: 7.98
```

Comparing the measures of model fit for math score data: r^2: 0.88 RMSE:6.02 MAE: 4.89

These measures provide quite different results in terms of our model fit. It seems like r^2 might make it appear as though LSD concentration is a better predictor of math scores than it actually is.

Comparing the measures of model fit for pomegranate data: r^2: 0.008 RMSE: 9.96 MAE: 7.98

Here, our r^2 value demonstrates that there is only a weak association (if any) between pomegranate eating and weight loss. Similarly, our RMSE and MAE values seem to indicate quite a weak model.

Question 4

```r
days_skiing <- runif(100)
slope <- -0.4
intercept <- 2
sigma <- 0.8
publications <- rnorm(n = 100, mean = intercept + slope * days_skiing, sd = 0.8)
publications
```

```
##   [1] 2.06405279 1.20847150 1.48438442 1.43623923 2.24452452 1.69558774
##   [7] 1.84299895 3.23186512 2.28595943 2.41090575 2.49700216 2.20868407
##  [13] 2.29914218 2.80302299 1.37254492 0.97831582 1.33952240 1.74869057
##  [19] 2.44756793 1.46288584 0.63078891 2.69755542 1.66716954 2.17430366
##  [25] 1.61029758 2.85635267 3.29192724 2.56623052 2.19336770 2.69979544
##  [31] 1.01001505 1.36127925 1.07550487 2.41330984 1.61904330 0.36878241
##  [37] 1.78092388 3.78046303 2.33988333 1.09733748 0.51737776 1.70426029
##  [43] 1.81133585 1.61079655 1.67301541 1.78127418 0.59358377 1.21521967
##  [49] 1.90969668 0.32866632 1.29991200 1.75850292 2.17701828 2.83752570
##  [55] 1.05238720 1.10339527 0.69856812 1.38696683 1.65512374 3.21889065
##  [61] 1.54233752 1.64173032 0.86121555 2.59037550 1.69744791 1.86760260
##  [67] 1.29193942 2.58877365 0.93530267 2.97741300 2.35325447 1.27687158
##  [73] 2.23430663 1.79805361 3.63928645 0.89072546 3.48791374 2.27566349
##  [79] 1.16161952 0.05330017 1.59002738 2.26988929 2.27565405 2.22985831
##  [85] 3.50823794 2.54354619 1.72081895 1.61398614 2.56741391 2.57623014
##  [91] 0.30414793 2.72318469 1.99044120 3.79192975 1.43043632 2.14946195
##  [97] 1.65085754 2.33874320 2.72484146 1.22475403
```

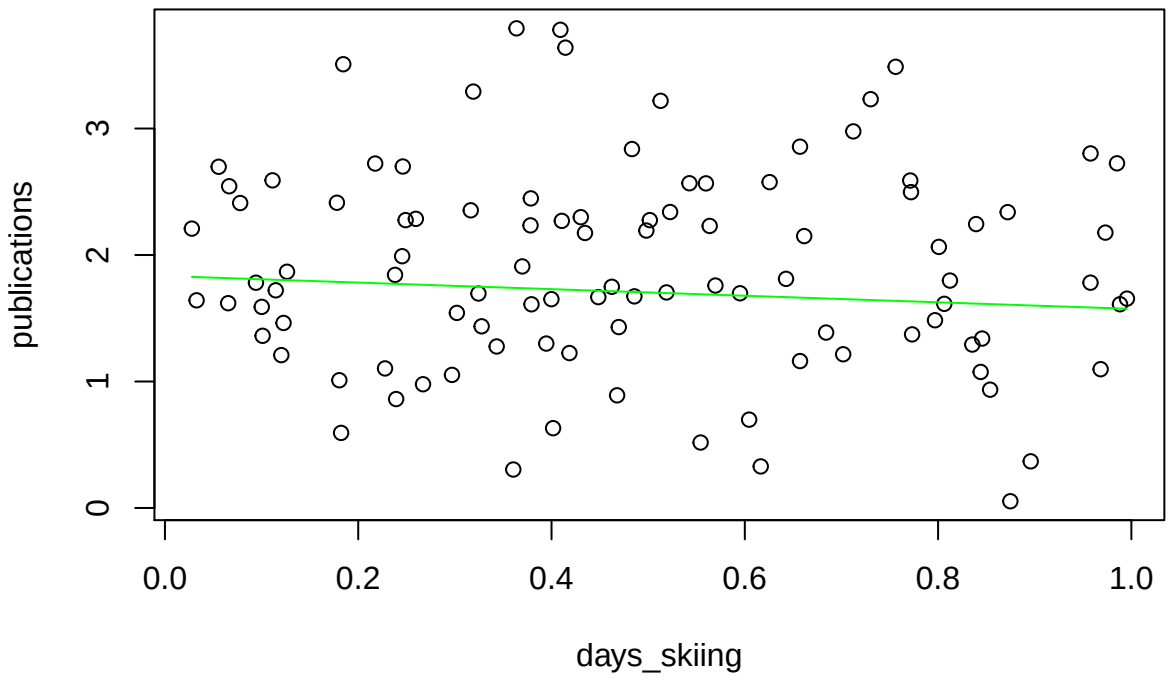
```r
model3 <- lm(publications ~ days_skiing)
coef(model3)
```

```
## (Intercept) days_skiing
##   1.9771959  -0.1800072
```

```r
confint(model3)
```

```
##                 2.5 %    97.5 %
## (Intercept)  1.6506819 2.3037100
## days_skiing -0.7688874 0.4088731
```

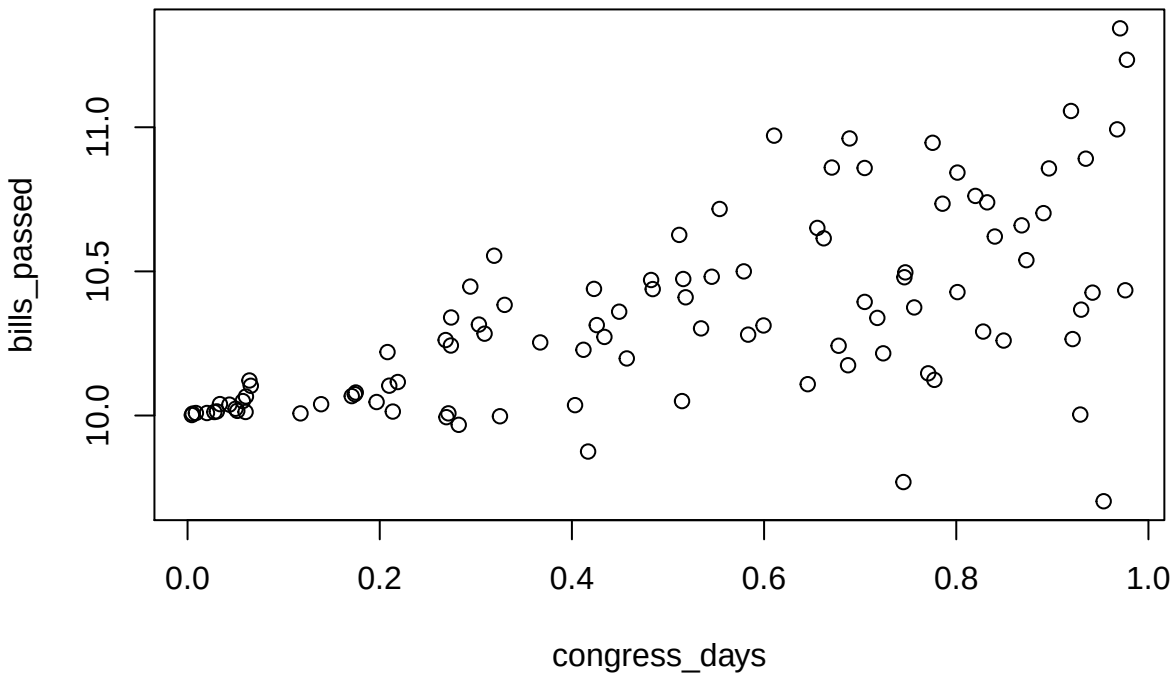
```r
plot(publications ~ days_skiing)
curve(1.834 + (-0.26 * x), add = T, col = "green")
```



When trying to assess how the number of days a person goes skiing in a year affects the number of publications the person submits that year, we find that the equations of the line is y=-0.26*days-skiing + 1.834. This is different from the true values that I came up with, presumably due to the variance introduced into this dataset with a sd of 0.8

Question 5

```
congress_days <- runif(100)
slope <- 0.7
intercept <- 10
sigma <- 0.5 * congress_days
bills_passed <- rnorm(n = 100, mean = intercept + slope * congress_days, sd = sigma)

plot(bills_passed ~ congress_days)
```



This would indicate that the more days congress spends in session the more variable it becomes whether they do their jobs and pass legislature (let alone what the content of those bills are), or whether the environment becomes too acerbic to get anything done.