Brian Holliday

Professor Li

Intro to Data Mining

3 May 2020

Project 6a

1.

Find the best line of linear regression using formulas (1) and (2).

Figure 1: Regression Code

```
#Number 1
x <- c(1,2,0,3)
y <- c(1,2,0,2)

k <- sum((x - mean(x))*(y - mean(y)))/sum((x - mean(x))^2)
k

b <- mean(y) - k*mean(x)
b

#rline = 0.2*x + 0.7
```

From the formulas for the regression line in the one-dimensional case, we get a regression line of:

Y = 0.2x + 0.7.

This is a coefficient of 0.2 on our variable and an intercept of 0.7. In this case k represents the intercept and b represents the coefficient of our variable.

2.

Find the best line of linear regression using the R function lm( ).

Figure 2: Regression using lm()

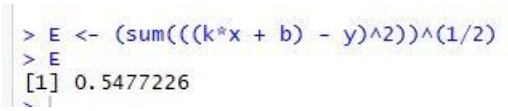
```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
        0.2          0.7

>
```

Just like with the formula, we arrive at the same answer using lm().

3.

Compute the L^2 -error for the least square solution.

Figure 3: L^2 Error

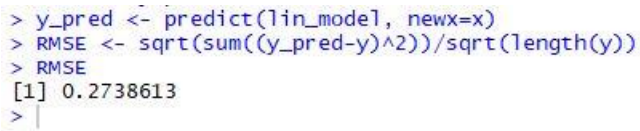


This is a measure of the accuracy of our model. Since the smaller the error, the better our model is at predicting our data points, we want to minimize this number.

4.

Compute the root mean square error (RMSE) krk2 √ n , where r is the residual and n is the number of observations.

Figure 4: RMSE



This one of the tests for the accuracy of our model. It tests on average how far our data points are from our regression line, so we will want to minimize this number as well.