Assignment 1

Due: 6/17/2018

Student name: Lavinia Wang #1473704

Problem 3 (10 points) Perform in R, the following calculations from linear algebra. For the following matrices and vectors. Submit both R code and the solution for credit.

a. $v \cdot w$ (dot product)

```
     [,1]
[1,]   10
```

b. $-3 * w$

```
     [,1]
[1,]   -9
[2,]   21
[3,]  -15
```

c. $M * v$

```
     [,1]
[1,]    8
[2,]  130
[3,]  100
```

d. $M+N$

```
     [,1] [,2] [,3]
[1,]  -19    0    0
[2,]   30   70   20
[3,]   20   30   30
```

e. $M-N$

```
     [,1] [,2] [,3]
[1,]   21   10    0
[2,]   30   30   10
[3,]  -20   10  -10
```

f. $Z^T Z$ (Make sure you get the right dimensions on this matrix)

```
     [,1] [,2]
[1,]    4    9
[2,]    9   65
```

g. $(Z^T Z)^{-1}$ (You do not have to perform Gaussian-Elimination here … i.e. this is a hint on what the dimensions of the matrix should be ☺)

```
             [,1]        [,2]
[1,]  0.36312849 -0.05027933
[2,] -0.05027933  0.02234637
```

Assignment 1
Due: 6/17/2018
Student name: Lavinia Wang #1473704

h. Z^TY

```
     [,1]
[1,]   15
[2,]   -3
```

i. $\beta=(Z^TZ)^{-1}Z^TY$

```
            [,1]
[1,]  5.5977654
[2,] -0.8212291
```

j. det(Z^TZ)

```
[1] 179
```

R code:

```
#Problem 1：
Z = matrix(c(1, 4, 1, 6, 1, -3, 1, 2), nrow = 4, ncol = 2, byrow = T)
Z
Y = matrix(c(-1, 1, 7, 8), nrow = 4, ncol = 1, byrow = F)
Y
M = matrix(c(1, 5, 0, 30, 50, 15, 0, 20, 10), nrow = 3, ncol = 3, byrow = T)
M
N = matrix(c(-20, -5, 0, 0, 20, 5, 20, 10, 20), nrow = 3, ncol = 3, byrow = T)
N
v = matrix(c(-2, 2, 6), nrow = 3, ncol = 1, byrow = F)
v
w = matrix(c(3, -7, 5), nrow = 3, ncol = 1, byrow = F)
w

#a. v·w(dot product)
dim(v)
dim(w)
prod = t(v) %*% w
prod

#b. −3*w
prod1 = -3* w
prod1

#c. M*v
dim(M)
dim(v)

#d. M+N
sum = M + N
sum

#e. M-N
subtr = M - N
subtr
```

Assignment 1
Due: 6/17/2018
Student name: Lavinia Wang #1473704

```
#f.Z^T * Z
ZTZ = t(Z) %*% Z
ZTZ

#g. (Z^T * Z) ^-1
ZTZI = solve(ZTZ)
ZTZI

#h.Z^T * Y
ZTY = t(Z) %*% Y
ZTY

#i.(Z ^T * Z) ^ -1 * Z^T * Y
beta = ZTZI %*% ZTY
beta

#j.det(Z ^T * Z)
det= det(ZTZ)
det
```

Problem 4 (10 points –other types of regression models): There are other types of regression models outside of linear and logistic regression. **Using Google Scholar**, locate **a journal article**, which utilizes **one** of the **types of regressions** listed below or another regression outside of linear/logistic that interests you. **Write a summary** of the journal article and how it utilizes the regression model in **two paragraphs**. **Cite the paper in APA format.**

Plant breeding based on grain yield (GY) is expensive and time-consuming. In this article, the authors discussed about an alternative estimation technique using multivariate ridge regression models to evaluate the performance of crops in the three water regimes and phenological stages. Regression techniques based on penalizing regression parameters and linear combination from the predictor variables are more appropriate for spectral assessment in dealing with collinearity. The model was built to evaluate the ability of canopy reflectance to predict GY in a wide range of bread wheat genotypes.

1187 samples with bands between the 350 and 2500 nm wavelengths were used to perform the analyses. When training the ridge regression model, the authors repetitively created sets of randomly selected data that corresponded to 30% of the total data to ensure a correct selection of the lambda parameter. The results showed that models generated by ridge regression explained between 77% and 91% of yield variability and good correlations with GY and integrate morpho-physiological information that determines GY.

Javier Hernandez, Gustavo A. Lobos, Iván Matus, Alejandro del Pozo, Paola Silva nd Mauricio Galleguillos. *Using Ridge Regression Models to Estimate Grain Yield from Field Spectral Data in Bread Wheat (Triticum Aestivum L.) Grown under Three Water Regimes*. Retrieved from http://www.mdpi.com/2072-4292/7/2/2109/htm

Problem 5: (15 pts – regression analysis, visualization, and interpretation): The data in the file *housedata.csv* are collected from 1,000 homes being sold in King County, Washington from May 2014 through May 2015. There are 21 columns in the data file but not all are relevant here.

Assignment 1

Due: 6/17/2018

Student name: Lavinia Wang #1473704

We are interested in which predictors are significant variables of determining price values of houses.

a. (5 points) Before running any regressions make sure to check for multicollinearity. How did you check for multicollinearity? If there is multicollinearity, how do you plan to resolve it? Are there any other issues with the dataset we have to consider before running the regressions?

   To check multicollinearity, we could compute the VIF scores. If the VIF score is greater or equal to 10, we can conclude that there exists multicollinearity in the model. If that were the case, we will compute the correlation matrix then figure out which variables (absolute value greater than 0.7) should be removed from the regression model with domain knowledge. Before running the regression model, we need to make sure the dataset is pre-processed, i.e. data cleaning (outlier identification, missing value), data integration and data transformation.

b. Run a multiple regression of price on the variables listed above.

   i. (5 points) Run the model using an automatic method (i.e. stepwise, forward, backward). Explain why you chose the method. Comment on the overall significance of the regression fit. Which predictors have coefficients that are significantly different from zero at the .05 level?

   After running and comparing stepwise, forward and backward methods, I will go with backward. This method goes through each independent variable and would "throw away" those are not statistically significant. The final model has R^2 equal to 0.72 and it almost the same with adjusted R^2.

   sqft_living (the living space area), lat (latitude coordinate of the home), waterfront (is it near the waterfront-values are higher near the waterfront), view (the number of times the house was viewed), grade (provided by the King County auditor), yr_built (year the house was built), sqft_living15 (area of the living room in 2015-could be different from above if the house was renovated, also affecting the lotsize area), zipcode (zipcode of the home) and yr_renovated (the year the house was renovated) are significantly different from zero at the .05 level.

```
> summary(houseBackward)

Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + floors +
    waterfront + view + condition + grade + yr_built + yr_renovated +
    zipcode + lat + long + sqft_living15, data = houseNum)

Residuals:
    Min      1Q  Median      3Q     Max
-843328  -87093   -1136   71560 1364242

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.313e+07  1.246e+07   1.054 0.292262
bedrooms      -1.648e+04  8.385e+03  -1.966 0.049584 *
bathrooms      2.587e+04  1.327e+04   1.950 0.051428 .
sqft_living    1.228e+02  1.427e+01   8.602  < 2e-16 ***
floors         2.125e+04  1.317e+04   1.614 0.106920
```

Assignment 1
Due: 6/17/2018
Student name: Lavinia Wang #1473704

```
waterfront     7.086e+05  6.927e+04  10.229  < 2e-16 ***
view           6.429e+04  8.759e+03   7.340 4.46e-13 ***
condition      1.565e+04  9.066e+03   1.726 0.084660 .
grade          8.020e+04  8.883e+03   9.029  < 2e-16 ***
yr_built      -2.593e+03  3.134e+02  -8.273 4.20e-16 ***
yr_renovated   4.297e+01  1.554e+01   2.766 0.005785 **
zipcode       -4.698e+02  1.358e+02  -3.460 0.000564 ***
lat            5.849e+05  4.355e+04  13.431  < 2e-16 ***
long          -7.911e+04  5.220e+04  -1.516 0.129963
sqft_living15  6.394e+01  1.607e+01   3.979 7.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 178000 on 984 degrees of freedom
Multiple R-squared:  0.7294,      Adjusted R-squared:  0.7256
F-statistic: 189.5 on 14 and 984 DF,  p-value: < 2.2e-16
```

ii. ii. (5 points) Using the variables above, create a visualization, which will provide an interesting story or insight within this data and present it to the general public. Is there an important trend or lesson that you would like the public to understand about housing prices within King County?

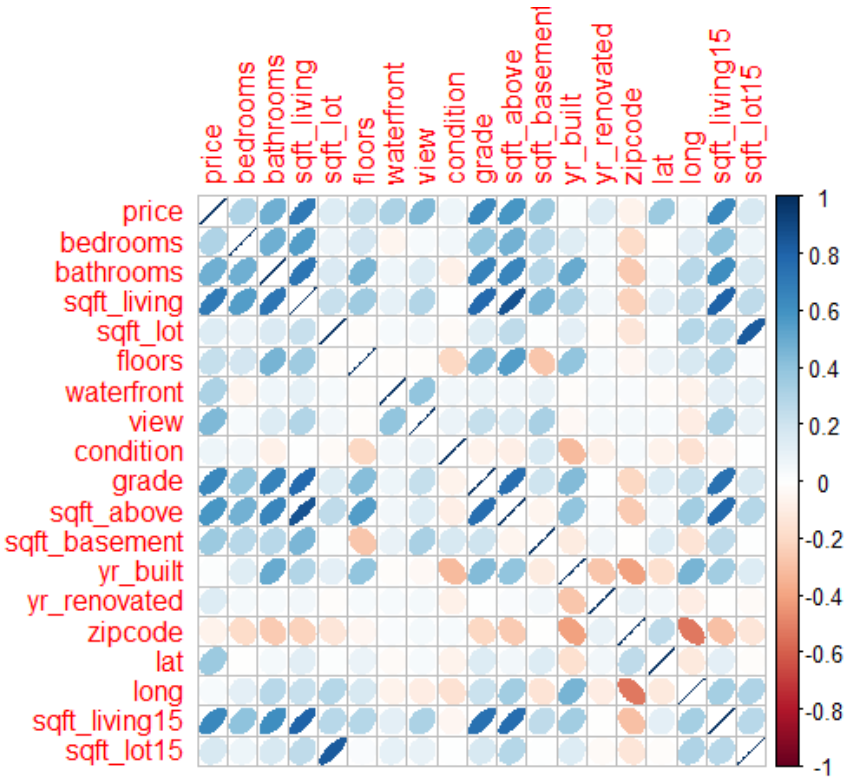

Figure 1: correlation matrix from RStudio

The housing prices in King County is significantly affected by grades provided by the King County auditor. 1 unit increase in grades will increase the housing price by 8 unit.

Assignment 1

Due: 6/17/2018

Student name: Lavinia Wang #1473704

Waterfront and the number of times the house was viewed will also increase price. Area of the living room in 2015-could be different from above if the house was renovated, also affecting the lotsize area and the year the house was renovated will cause the price to go up as the house is updated. Zipcode and latitude coordinate of the home matters as well because the larger latitude, the higher the price. But the smaller the zipcode, the higher the price. So the downtown area is more expensive than suburbs in King County. The newer the house, the higher the price. If the house is newly renewed, the price will also increase.

The data implies that the housing condition and location has a great impact on price. Newly renovated with waterfront in downtown area has the highest price in King County. Also, auditor plays an vital role in the real estate market.