Keiland Pullen DSC 424 | Fall 2021
Homework 1

1.)

While reviewing the 2012 London Olympic data, I observed that there could be a possible relationship between the number of female athletes and the number of Gold medals. The dataset consists of data for 20 countries, 14 of those countries won Gold medals, 6 of which had more female athletes than male athletes or an equal number of female and male athletes. Of those 6 countries, 4 of them won more Gold medals than any other medal.

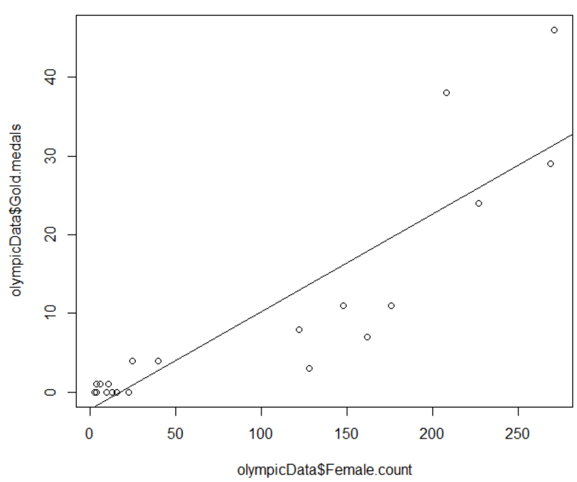
To examine this relationship, I created several new variables based on the Olympic data. The variables are:

MedalCount : the total number of Olympic medals for each country
AthleteCount : the total number Olympic athletes for each country
FemPercentage : the percentage of female Olympic athletes representing each country
AthPercentage : the percentage of Olympic athletes for each county
AthFemCtry : the percentage of female Olympic athletes in the country's population

olympicData$MedalCount <- olympicData$Gold.medals + olympicData$Silver.medals + olympicData$Bronze.medals
olympicData$AthleteCount <- olympicData$Female.count + olympicData$Male.count
olympicData$FemPercentage <- olympicData$Female.count / olympicData$AthleteCount
olympicData$AthPercentage <- olympicData$AthleteCount / olympicData$X2010.population
olympicData$AthFemCtry <- olympicData$Female.count / olympicData$X2010.population

After exploring several models constructed with the new variables, the best model seems to contain a single variable. Unfortunately, the new variables had no positive impact. That single variable with the most impact is "Female.count". It is defined as the number of female Olympic athletes representing each respective country.

modelFinal <- lm(Gold.medals ~ Female.count, data=olympicData)
summary(modelFinal)



The Gold.medals – Female.count plot tells us that there is a relationship between the two. The obvious observation is that the more female Olympic athletes on a country's team, then the more Gold medals that team may win. There are a number of outliers which could be explained by additional variables such as the number of Olympic events, team vs. individual events, Female events and maybe mixed (male and female) events.

Keiland Pullen DSC 424 | Fall 2021
Homework 1

2.a.)

```
> model <- lm(housingData$MEDV ~ . , data=housingData)
> summary(model)

Call:
lm(formula = housingData$MEDV ~ ., data = housingData)

Residuals:
    Min      1Q  Median      3Q     Max
-15.1304  -2.7673  -0.5814   1.9414  26.2526
```

Coefficients:

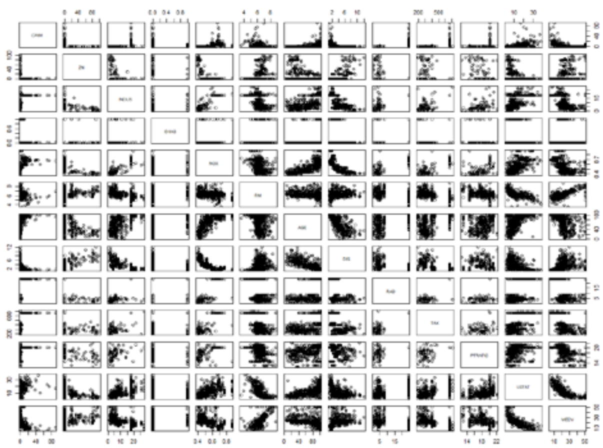
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	41.617270	4.936039	8.431	3.79e-16	***
CRIM	-0.121389	0.033000	-3.678	0.000261	***
ZN	0.046963	0.013879	3.384	0.000772	***
INDUS	0.013468	0.062145	0.217	0.828520	
CHAS	2.839993	0.870007	3.264	0.001173	**
NOX	-18.758022	3.851355	-4.870	1.50e-06	***
RM	3.658119	0.420246	8.705	< 2e-16	***
AGE	0.003611	0.013329	0.271	0.786595	
DIS	-1.490754	0.201623	-7.394	6.17e-13	***
RAD	0.289405	0.066908	4.325	1.84e-05	***
TAX	-0.012682	0.003801	-3.337	0.000912	***
PTRATIO	-0.937533	0.132206	-7.091	4.63e-12	***
LSTAT	-0.552019	0.050659	-10.897	< 2e-16	***

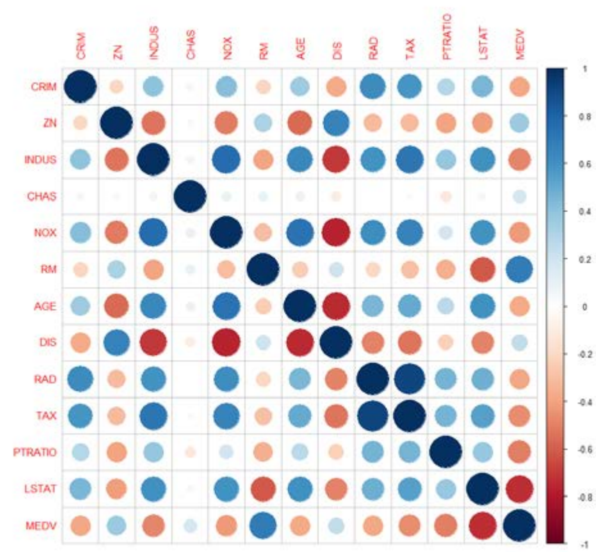
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.798 on 493 degrees of freedom
Multiple R-squared:  0.7343,	Adjusted R-squared:  0.7278
F-statistic: 113.5 on 12 and 493 DF,  p-value: < 2.2e-16
```

For this model, the $R^2 = 0.7343$ and the Adj-$R^2 = 73\%$. The p-value looks good and the t-tests are good. There are 2 variables that have no impact and those variables are “AGE” and “INDUS”.

2.b.)

```
# Plot the data
plot(housingData)

# Correlation Plot with circles
library(corrplot)
c = cor(housingData)
corrplot(c)
```





There are strong linear relationships between MEDV with RM (positive) and LSTAT (negative). There are weak relationships with the INDUS variable and the DIS variable. For those two variables, a LOG transformation was performed.

```
housingData$INDUs_LOG <- log(housingData$INDUS)
housingData$DIS_LOG <-log(housingData$DIS)

### Linear model for MEDV
model_2 <- lm(housingData$MEDV ~ CRIM + ZN + INDUs_LOG + CHAS + NOX + RM + AGE + DIS_LOG + RAD + TAX + PTRATIO + LSTAT , data=housingData)
summary(model_2)
```

The resulting model increases the $R^2 = 0.758$ and the Adj-$R^2 = 75\%$.

While the R^2 and Adj- R^2 changed to more favorable values, the significance of the “ZN” value was flipped by the introduction of the transformed “INDUS” variable. The “AGE” variable remained insignificant.

Keiland Pullen DSC 424 | Fall 2021
Homework 1

Call:
lm(formula = housingData$MEDV ~ CRIM + ZN + INDUs_LOG + CHAS +
NOX + RM + AGE + DIS_LOG + RAD + TAX + PTRATIO + LSTAT, data = housingData)

Residuals:

Min	1Q	Median	3Q	Max
-15.6141	-2.5796	-0.4811	2.0424	23.5575

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	52.684191	4.968726	10.603	< 2e-16	***
CRIM	-0.168037	0.032069	-5.240	2.39e-07	***
ZN	0.024202	0.013389	1.808	0.071285	.
INDUs_LOG	-1.170654	0.514068	-2.277	0.023200	*
CHAS	3.056188	0.833876	3.665	0.000274	***
NOX	-24.343941	3.773470	-6.451	2.65e-10	***
RM	3.645342	0.401190	9.086	< 2e-16	***
AGE	-0.014656	0.012982	-1.129	0.259452	
DIS_LOG	-8.787635	0.831636	-10.567	< 2e-16	***
RAD	0.307574	0.062193	4.945	1.04e-06	***
TAX	-0.013057	0.003442	-3.793	0.000167	***
PTRATIO	-0.876562	0.126248	-6.943	1.21e-11	***
LSTAT	-0.544701	0.048330	-11.270	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.579 on 493 degrees of freedom
Multiple R-squared: 0.758, Adjusted R-squared: 0.7521
F-statistic: 128.7 on 12 and 493 DF, p-value: < 2.2e-16

2.c.)

To perform stepwise selection, the following models were created:

Model containing Log Transformations:
model_trans <- lm(housingData$MEDV ~ CRIM - ZN + INDUs_LOG + CHAS + NOX + RM - AGE + DIS_LOG + RAD + TAX + PTRATIO + LSTAT , data=housingData)
summary(model_trans)

Model containing forward selection:
forwardModel <- step(model1, direction = 'forward', scope = formula(model_trans))
summary(forwardModel)

Model containing backward selection:
backwardModel <- step(model1, direction = 'backward', scope = formula(model_trans))
summary(backwardModel)

Model combing both, backward and forward selection:
bothModel <- step(model1, direction='both', scope = formula(model_trans))
summary(bothModel)

```
> summary(bothModel)
Call:
lm(formula = housingData$MEDV ~ LSTAT + RM + PTRATIO + DIS_LOG +
    NOX + CHAS + CRIM + INDUs_LOG + RAD + TAX, data = housingData)

Residuals:
    Min      1Q  Median      3Q     Max
-15.404  -2.714  -0.520   2.068  23.421

Coefficients:
              Estimate     Std. Error    t value  Pr(>|t|)
(Intercept)   53.956186    4.946293      10.908  < 2e-16 ***
LSTAT         -0.560675    0.045731      -12.260  < 2e-16 ***
RM            3.587707     0.393967      9.107  < 2e-16 ***
PTRATIO       -0.961234    0.121144      -7.935 1.42e-14 ***
DIS_LOG       -8.031130    0.751994      -10.680  < 2e-16 ***
NOX           -25.685345   3.699879      -6.942 1.22e-11 ***
CHAS          3.068444     0.834572      3.677 0.000262 ***
CRIM          -0.161072    0.032040      -5.027 6.97e-07 ***
INDUs_LOG     -1.536306    0.475052      -3.234 0.001302 **
RAD           0.296215     0.061754      4.797 2.14e-06 ***
TAX           -0.010993    0.003289      -3.342 0.000895 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.595 on 495 degrees of freedom
Multiple R-squared:  0.7553,	Adjusted R-squared:  0.7503
F-statistic: 152.8 on 10 and 495 DF,  p-value: < 2.2e-16
```

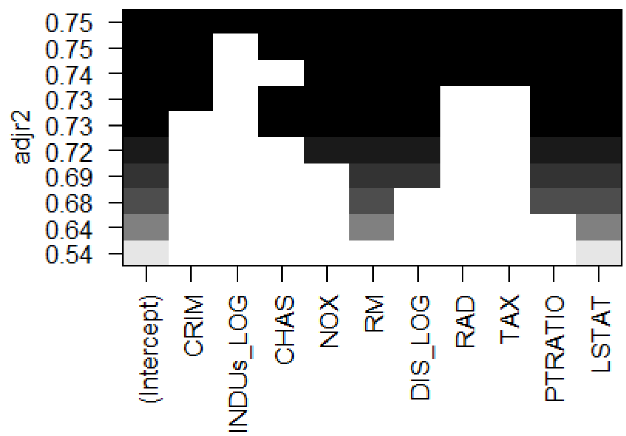
Keiland Pullen DSC 424 | Fall 2021
Homework 1

The resulting model dropped the variables “AGE” and “ZN”. The Adjusted-R-squared value decreased slightly from a value of 75.21% to a value of 75.03%.

2.d.)

housingSubsets = regsubsets(housingData$MEDV ~ CRIM - ZN + INDUs_LOG + CHAS + NOX + RM - AGE + DIS_LOG + RAD + TAX + PTRATIO + LSTAT, data=housingData, nvmax=20)
summary(housingSubsets)

```
Subset selection object
Call: regsubsets.formula(housingData$MEDV ~ CRIM - ZN + INDUs_LOG +
   CHAS + NOX + RM - AGE + DIS_LOG + RAD + TAX + PTRATIO + LSTAT,
   data = housingData, nvmax = 20)
10 Variables  (and intercept)
           Forced in  Forced out
CRIM       FALSE      FALSE
INDUs_LOG  FALSE      FALSE
CHAS       FALSE      FALSE
NOX        FALSE      FALSE
RM         FALSE      FALSE
DIS_LOG    FALSE      FALSE
RAD        FALSE      FALSE
TAX        FALSE      FALSE
PTRATIO    FALSE      FALSE
LSTAT      FALSE      FALSE
1 subsets of each size up to 10
Selection Algorithm: exhaustive
          CRIM INDUs_LOG CHAS NOX RM  DIS_LOG RAD TAX PTRATIO LSTAT
1  ( 1 )  " "  " "       " "  " " " " " "     " " " " " "     "*"
2  ( 1 )  " "  " "       " "  " " "*" " "     " " " " " "     "*"
3  ( 1 )  " "  " "       " "  " " "*" " "     " " " " "*"     "*"
4  ( 1 )  " "  " "       " "  " " "*" "*"     " " " " "*"     "*"
5  ( 1 )  " "  " "       " "  "*" "*" "*"     " " " " "*"     "*"
6  ( 1 )  " "  " "       "*"  "*" "*" "*"     " " " " "*"     "*"
7  ( 1 )  "*"  " "       "*"  "*" "*" "*"     " " " " "*"     "*"
8  ( 1 )  "*"  " "       " "  "*" "*" "*"     "*" "*" "*"     "*"
9  ( 1 )  "*"  " "       "*"  "*" "*" "*"     "*" "*" "*"     "*"
10  ( 1 ) "*"  "*"       "*"  "*" "*" "*"     "*" "*" "*"     "*"
```



Similar to the previous models, the variables that were dropped are “AGE” and “ZN”.

2.e.)

Observing the plot, the variables “INDUs_LOG” and “CHAS” could be dropped and have a small effect on the Adjusted-RSquare value:

```
finalModel <- lm(housingData$MEDV ~ CRIM - ZN - INDUs_LOG - CHAS + NOX + RM - AGE + DIS_LOG + RAD + TAX + PTRATIO + LSTAT ,data=housingData)
summary(finalModel)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.3188	-2.7440	-0.6223	2.2746	23.5462

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	53.030459	5.014360	10.576	< 2e-16 ***
CRIM	-0.157660	0.032416	-4.864	1.55e-06 ***
NOX	-27.383803	3.679984	-7.441	4.42e-13 ***
RM	3.857049	0.394708	9.772	< 2e-16 ***
DIS_LOG	-7.454013	0.728458	-10.233	< 2e-16 ***
RAD	0.330525	0.062480	5.290	1.84e-07 ***
TAX	-0.014159	0.003268	-4.333	1.78e-05 ***
PTRATIO	-1.095041	0.118675	-9.227	< 2e-16 ***
LSTAT	-0.576107	0.046483	-12.394	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.684 on 497 degrees of freedom
Multiple R-squared: 0.7448, Adjusted R-squared: 0.7407
F-statistic: 181.3 on 8 and 497 DF, p-value: < 2.2e-16

Keiland Pullen DSC 424 | Fall 2021
Homework 1

lm.beta(bothModel)
lm.beta(finalModel)

```
> lm.beta(bothModel)

Call:
lm(formula = housingData$MEDV ~ LSTAT + RM + PTRATIO + DIS_LOG + 
    NOX + CHAS + CRIM + INDUs_LOG + RAD + TAX, data = housingData)

Standardized Coefficients::
(Intercept)       LSTAT          RM     PTRATIO     DIS_LOG         NOX        CHAS        CRIM   INDUs_LOG         RAD         TAX 
 0.00000000 -0.43533390  0.27408457 -0.22626900 -0.47114484 -0.32361904  0.08474041 -0.15064202 -0.12978977  0.28043836 -0.20145180 
```

```
> lm.beta(finalModel)

Call:
lm(formula = housingData$MEDV ~ CRIM - ZN - INDUs_LOG - CHAS + 
    NOX + RM - AGE + DIS_LOG + RAD + TAX + PTRATIO + LSTAT, data = housingData)

Standardized Coefficients::
(Intercept)        CRIM         NOX          RM     DIS_LOG         RAD         TAX     PTRATIO       LSTAT 
  0.0000000  -0.1474511  -0.3450185   0.2946611  -0.4372884   0.3129212  -0.2594559  -0.2577663  -0.4473168 
```

3.)

KEILAND PULLEN - DSC 424 FALL 2021

#3)

a.) r·w (dot product) $\Rightarrow \begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \\ 3 \end{bmatrix}$

b.) $-3 * w \Rightarrow -3 * \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -6 \\ 3 \\ 3 \end{bmatrix}$

c.) $M * v \Rightarrow \begin{bmatrix} 20 & 5 & \emptyset \\ 5 & 25 & -10 \\ \emptyset & 10 & 5 \end{bmatrix} * \begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix} \Rightarrow \begin{bmatrix} (20*-1) + (5*1) + (\emptyset*3) \\ (5*-1) + (25*1) + (-10*3) \\ (\emptyset*-1) + (10*1) + (5*3) \end{bmatrix} \Rightarrow$

$$\begin{bmatrix} -20 + 5 + \emptyset \\ -5 + 25 - 30 \\ \emptyset + 10 + 15 \end{bmatrix} = \begin{bmatrix} -15 \\ -10 \\ 25 \end{bmatrix}$$

d.) $M + N \Rightarrow \begin{bmatrix} 20 & 5 & \emptyset \\ 5 & 25 & -10 \\ \emptyset & 10 & 5 \end{bmatrix} + \begin{bmatrix} -20 & \emptyset & 10 \\ 5 & 10 & 15 \\ 5 & 20 & -5 \end{bmatrix} = \begin{bmatrix} \emptyset & 5 & 10 \\ 10 & 35 & 5 \\ 5 & 30 & \emptyset \end{bmatrix}$

e.) $M - N \Rightarrow \begin{bmatrix} 20 & 5 & \emptyset \\ 5 & 25 & -10 \\ \emptyset & 10 & 5 \end{bmatrix} - \begin{bmatrix} -20 & \emptyset & 10 \\ 5 & 10 & 15 \\ 5 & 20 & -5 \end{bmatrix} = \begin{bmatrix} \emptyset & 5 & -10 \\ \emptyset & 15 & -25 \\ -5 & -10 & \emptyset \end{bmatrix}$

f.) $Z^T \Rightarrow \begin{bmatrix} 1 & 4 \\ 1 & 3 \\ 1 & 2 \\ 1 & -5 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 4 & 3 & 2 & -5 \end{bmatrix}$

$$\begin{bmatrix} 1 & 4 \\ 1 & 3 \\ 1 & 2 \\ 1 & -5 \end{bmatrix}$$

g.) $Z^T Z \Rightarrow \begin{bmatrix} 1 & 1 & 1 & 1 \\ 4 & 3 & 2 & -5 \end{bmatrix} * \begin{bmatrix} 1 & 4 \\ 1 & 3 \\ 1 & 2 \\ 1 & -5 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 1 & 1 & 1 \\ 4 & 3 & 2 & -5 \end{bmatrix} \begin{bmatrix} 4 & 4 \\ 4 & 54 \end{bmatrix} = \begin{bmatrix} 4 & 4 \\ 4 & 54 \end{bmatrix}$

Keiland Pullen DSC 424 | Fall 2021
Homework 1

4. a.)

```
v = matrix(c(-1,1,3))
v

w = matrix(c(2,-1,1))
w
v * w

 > v * w
      [,1]
[1,]   -2
[2,]   -1
[3,]    3
```

4. b.)

```
> -3*w
      [,1]
[1,]   -6
[2,]    3
[3,]   -3
```

4. c.)

```
> M = matrix(c(20, 5, 0, 5, 25, -10, 0, 10, 5), nrow=3, byrow=T)
> M
     [,1] [,2] [,3]
[1,]   20    5    0
[2,]    5   25  -10
[3,]    0   10    5

> # Multiply the vector v by M
> M %*% v
    [,1]
[1,]  -15
[2,]  -10
[3,]   25
```

4. d.)

```
> N = matrix(c(-20, 0, 10, 5, 10, 15, 5, 20, -5), nrow=3, byrow=T)
> N
     [,1] [,2] [,3]
[1,]  -20    0   10
[2,]    5   10   15
[3,]    5   20   -5
> # Now lets add the Matrices
> M + N
     [,1] [,2] [,3]
[1,]    0    5   10
[2,]   10   35    5
[3,]    5   30    0
```

4. e.)

```
> # Subtract them
> M - N
     [,1] [,2] [,3]
[1,]   40    5  -10
[2,]    0   15  -25
[3,]   -5  -10   10
```

4. f.)

```
> Z = matrix(c(1,1,1,1,4,3,2,-5), nrow=4, byrow=F)
> Z
     [,1] [,2]
[1,]    1    4
[2,]    1    3
[3,]    1    2
[4,]    1   -5
> t(Z)
     [,1] [,2] [,3] [,4]
[1,]    1    1    1    1
[2,]    4    3    2   -5
```

4. g.)

```
> t(Z) %*% Z
     [,1] [,2]
[1,]    4    4
[2,]    4   54
>
```