STA 108 Project Part 2

Ishita Dutta, Devika Sunil Kumar, Fernanda Serna Godoy

5/13/2021

Introdution to Part 2:

This project focuses on taking in the county demographic information, also known as CDI, for 440 of the most populous counties in the United States between the years 1990 and 1992 using 17 different measuring parameters.

The first part of the project (not here), was an analysis through simple linear regression. We took regression analysis and determined best fit, then further analyzed by taking LS estimates, residuals, confidence intervals, etc. We concluded which one of 3 indicators best fits our data.

For the second part of the project, what you are viewing now, we split the data into 2 proposed models from 6 parameters of data. Model 1 is consists of total population, land area, and total personal income. Model 2 will use the percent of the population ages 65+, population density, and total personal income again. Both models will use the number of active physicians as a response parameter. This we will analyze in two ways:

In the first way, we take stem plots of each predictor variable, as well as the scatterplot and correlation matrix for both models. We then find the first order regression model, Rˆ2, and residuals for each model. These we plot against Y, each of the predictor variables, and each of the 2-factor interaction terms. We expand the two models by adding all the two factor interactions and find Rˆ2.

In the second way, we will add 3 new predictor variables and find which one is the best when other variables have already been included. First, we find Rˆ2 and calculate the coefficient of partial determination given total population and personal income are already included. Then we use the F* statistic to determine whether the variable is actually helpful in the regression model some variables have been added through testing. Finally, we do the same with each pair of the 3 predictors to find and confirm the best pair.

Part I - Multiple Linear Regressions

6.28

Refer to the CDI data set in Appendix C.2. You have been asked to evaluate two alternative models for predicting the number of active physicians (Y) in a CDI. Proposed model I includes as predictor variables total population (X1), land area (X2), and total personal income (X3).Proposed model II includes as predictor variables population density (X1, total population divided by land area), percent of population greater than 64 years old (X2), and total personal income (X3).

6.28 a)

Prepare a stem-and-leaf plot for each of the predictor variables. What noteworthy information is provided by your plots?

```
## Model 1

## Population

##
##   The decimal point is 6 digit(s) to the right of the |
##
##   0 | 11111111111111111111111111111111111111111111111111111111111111111111+254
##   0 | 55555555555555555555555566666666666666777777777777777777778888888888
##   1 | 000000122233333444
##   1 | 55699
##   2 | 1134
##   2 | 58
##   3 |
##   3 |
##   4 |
##   4 |
##   5 | 1
##   5 |
##   6 |
##   6 |
##   7 |
##   7 |
##   8 |
##   8 | 9

## Land Area

##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 00001111111111112222222222222222222233333333333333333333333333444444+252
##    1 | 00000000000000011111111111111222222222223333334444555666677778889999
##    2 | 0001111466778
##    3 | 3344688
##    4 | 00122368
##    5 | 45
##    6 | 023
##    7 | 29
##    8 | 11
##    9 | 22
##   10 |
##   11 |
##   12 |
##   13 |
##   14 |
##   15 |
##   16 |
##   17 |
##   18 |
##   19 |
##   20 | 1

## Total Personal Income
```

```
##
##   The decimal point is 4 digit(s) to the right of the |
##
##    0 | 11111111111112222222222222222222222222222222222222222222222222222222+263
##    1 | 000000000000111111111222223333344444445555555567788888888999
##    2 | 001111233344477788899
##    3 | 0255678899
##    4 | 19
##    5 | 59
##    6 |
##    7 |
##    8 |
##    9 |
##   10 |
##   11 | 1
##   12 |
##   13 |
##   14 |
##   15 |
##   16 |
##   17 |
##   18 | 4
```

```
## Model 2
```

```
## Population Density
```

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 00000000000000001111111111111111111111111111111111111111111111111111+321
##    2 | 00001112233456700111145
##    4 | 05884
##    6 | 2464
##    8 | 19
##   10 | 378
##   12 |
##   14 | 4
##   16 |
##   18 |
##   20 |
##   22 |
##   24 |
##   26 |
##   28 |
##   30 |
##   32 | 4
```

```
## Percent Older than 64
```

```
##
##   The decimal point is at the |
##
```

```
##    2 | 0
##    4 | 47890389
##    6 | 1123455677990134566678899
##    8 | 00112222233334444555666777778888899990002222333333444444445555666677
##   10 | 00011111122222222223333334444445555555666666666777777888888888899999+36
##   12 | 00000000111112222333333333344445555555666666777777777888899900000000+36
##   14 | 000011111112233344444555677889000000111122223455667778
##   16 | 12556699901122345
##   18 | 06778
##   20 | 070
##   22 | 018828
##   24 | 47
##   26 | 055
##   28 | 1
##   30 | 7
##   32 | 138
```

```
## Total Personal Income
```

```
##
##   The decimal point is 4 digit(s) to the right of the |
##
##    0 | 11111111111112222222222222222222222222222222222222222222222222222222+263
##    1 | 000000000000111111112222223333344444445555555567788888888999
##    2 | 001111233344477788899
##    3 | 0255678899
##    4 | 19
##    5 | 59
##    6 |
##    7 |
##    8 |
##    9 |
##   10 |
##   11 | 1
##   12 |
##   13 |
##   14 |
##   15 |
##   16 |
##   17 |
##   18 | 4
```

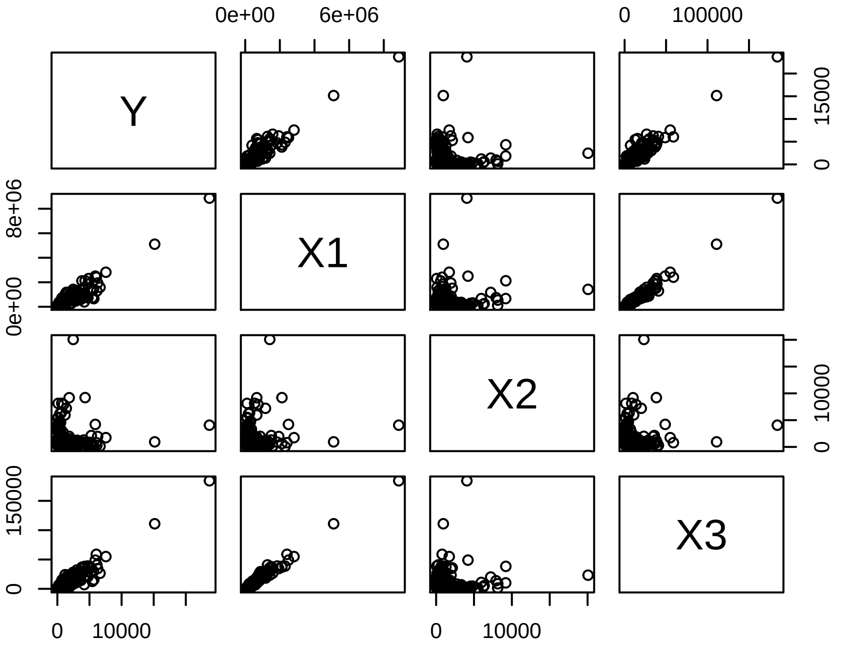
All the plots are skewed right with some outliers present in the data. The percent of population greater than 64 years old stemplot has the most even distribution compared to the other variables.

6.28 b)

Obtain the scatter plot matrix and the correlation matrix for each proposed model. Summarize the information provided.

```
## Model 1:
```

```
## Scatter Plot Matrix Model 1:
```

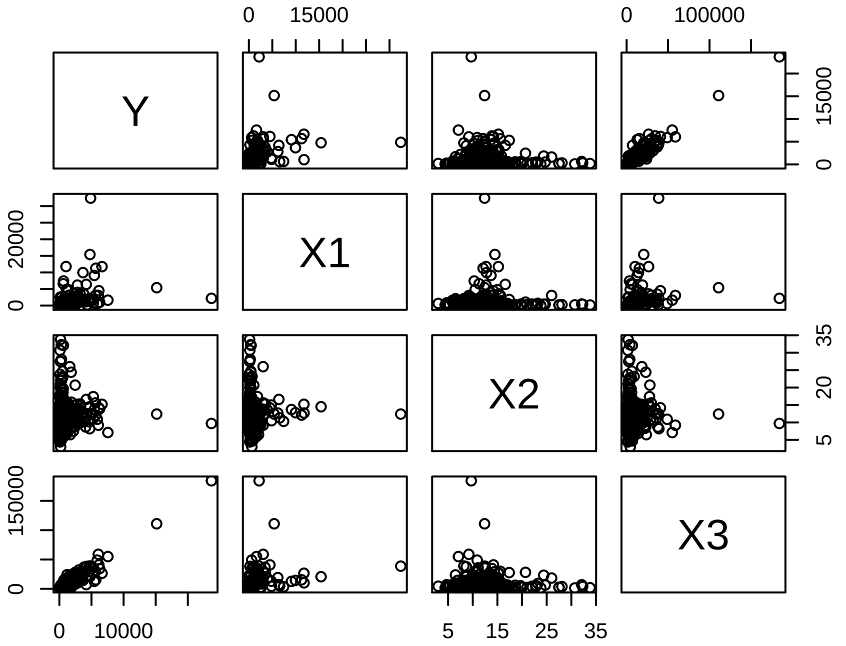


```
## Correlation Matrix Model 1:
```

```
##             Y        X1         X2        X3
## Y  1.00000000 0.9402486 0.07807466 0.9481106
## X1 0.94024859 1.0000000 0.17308335 0.9867476
## X2 0.07807466 0.1730834 1.00000000 0.1270743
## X3 0.94811057 0.9867476 0.12707426 1.0000000
```

```
## Model 2:
```

```
## Scatter Plot Matrix Model 2:
```



```
## Correlation Matrix Model 2:
```

```
##               Y         X1          X2          X3
## Y   1.00000000 0.40643863 -0.00312863  0.94811057
## X1  0.40643863 1.00000000  0.02918445  0.31620475
## X2 -0.00312863 0.02918445  1.00000000 -0.02273315
## X3  0.94811057 0.31620475 -0.02273315  1.00000000
```

For Model 1, the scatterplot matrix and correlation matrix obtained above show a strong relationship between Y and X1, as well as between Y and X3, where there is a very small difference between the two and a weaker relationship between Y and X2.

For Model 2, the scatterplot matrix and correlation matrix obtained above show a strong relationship between Y and X3 and a weaker relationship when considering Y and X1 or X2.

6.28 c)

For each proposed model, fit the first-order regression model (6.5) with three predictor variables.

```
## Model 1:
```

```
## Y =  -13.31615  +  0.0008366178 X1 -0.06552296 X2  +  0.09413199 X3
```

```
## Model 2:
```

```
## Y =  -170.5742  +  0.09615889 X1  +  6.339841 X2  +  0.1265665 X3
```

6.28 d)

Calculate R2 for each model. Is one model clearly preferable in terms of this measure?

```
## Model 1 R2:
```

```
## [1] 0.9026432
```

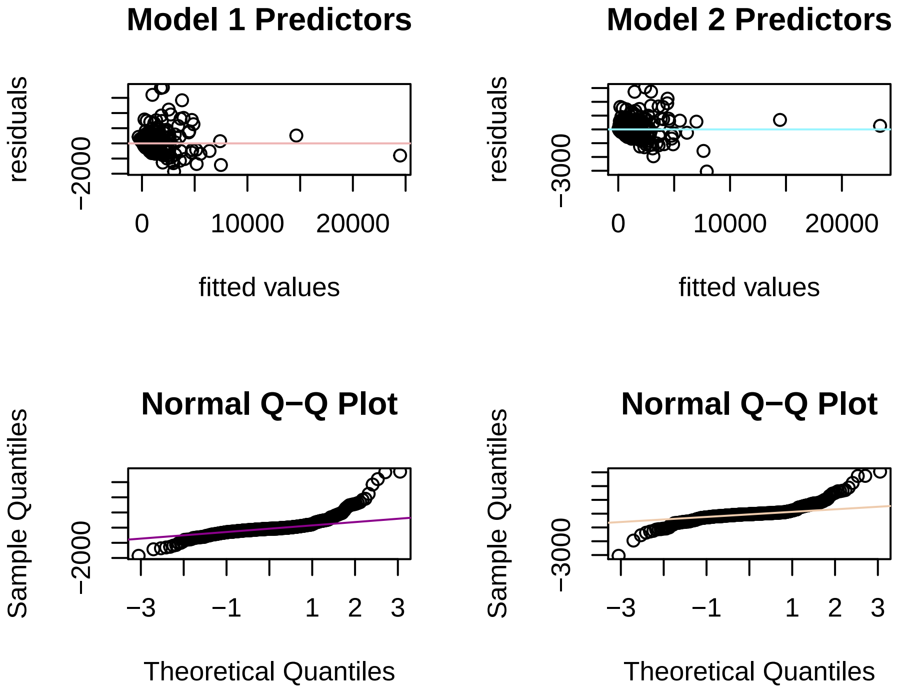
```
## Model 2 R2:
```

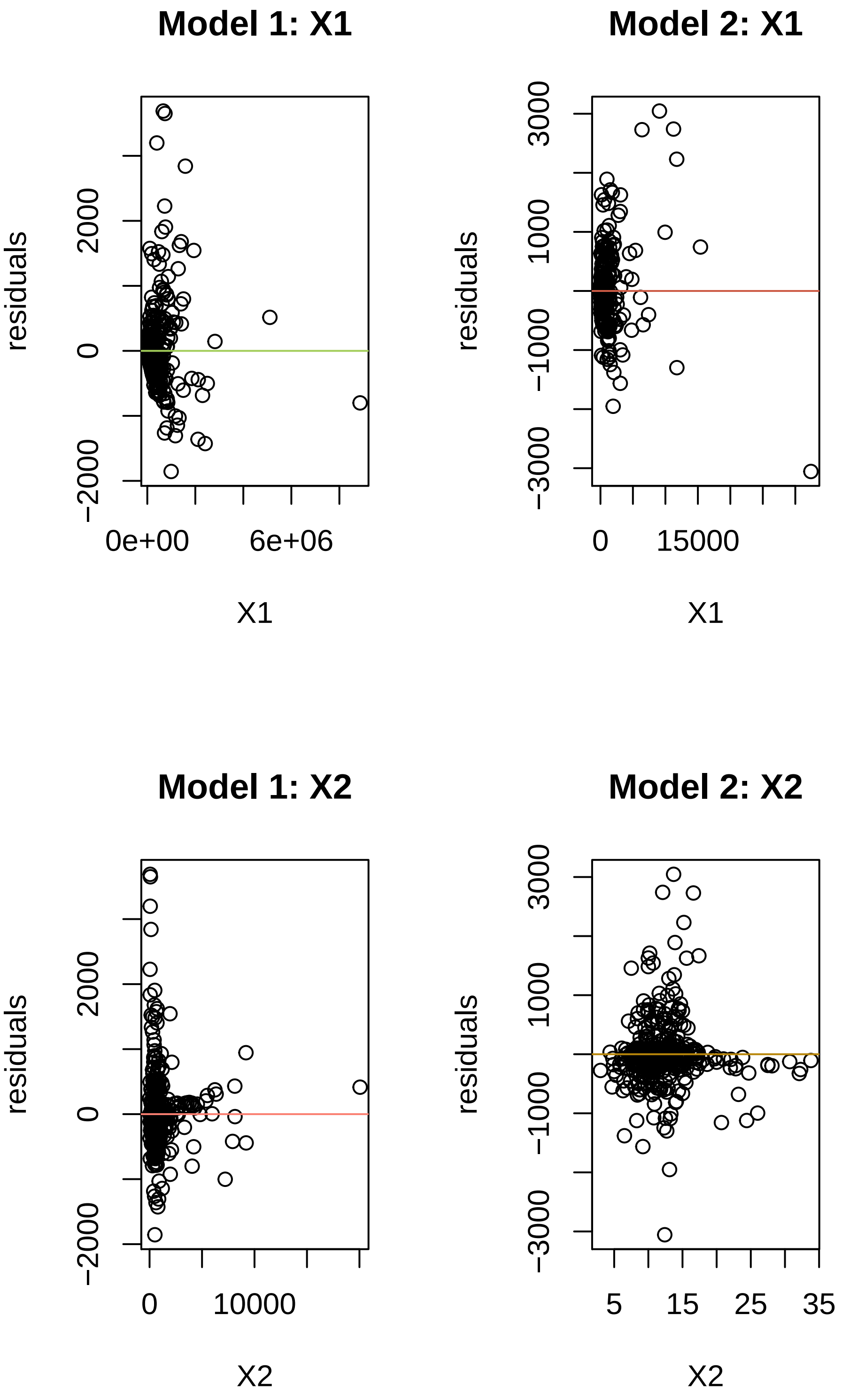
```
## [1] 0.9117491
```

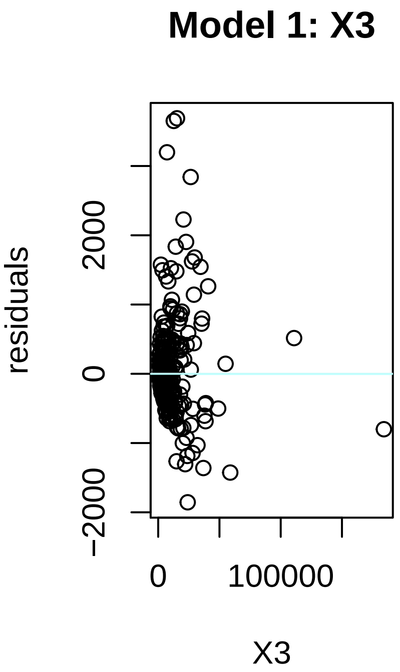
Based on the coefficients of determination (Rˆ2) obtained above for each of the models, model 2 seems to be preferable in terms of this measure (because the value obtained is greater than the one for model 1).

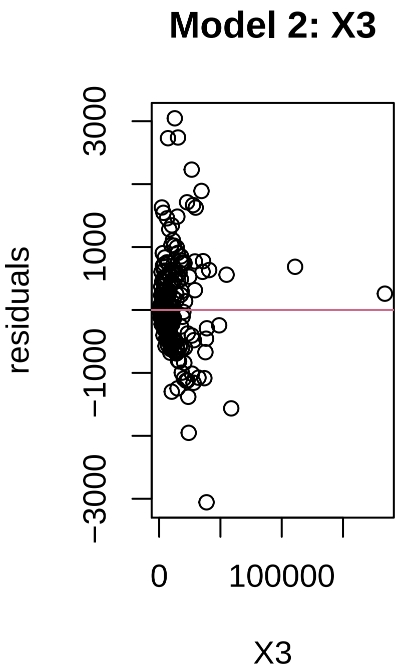
6.28 e)

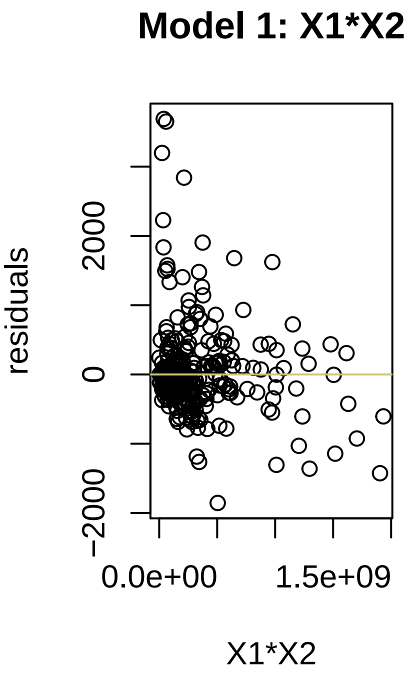
For each model, obtain the residuals and plot them against Y, each of the three predictor variables, and each of the two-factor interaction terms. Also prepare a normal probability plot for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly preferable in terms of appropriateness?

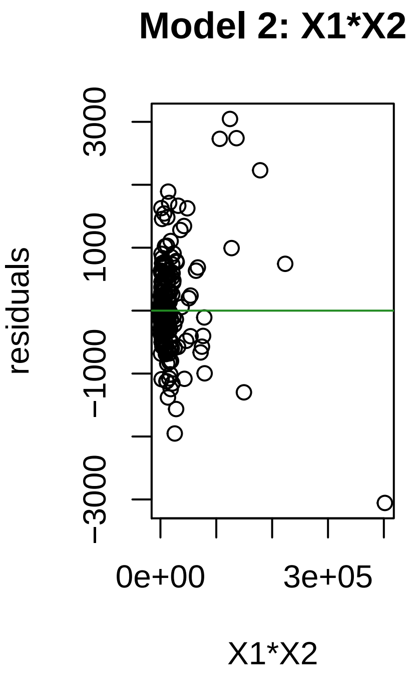


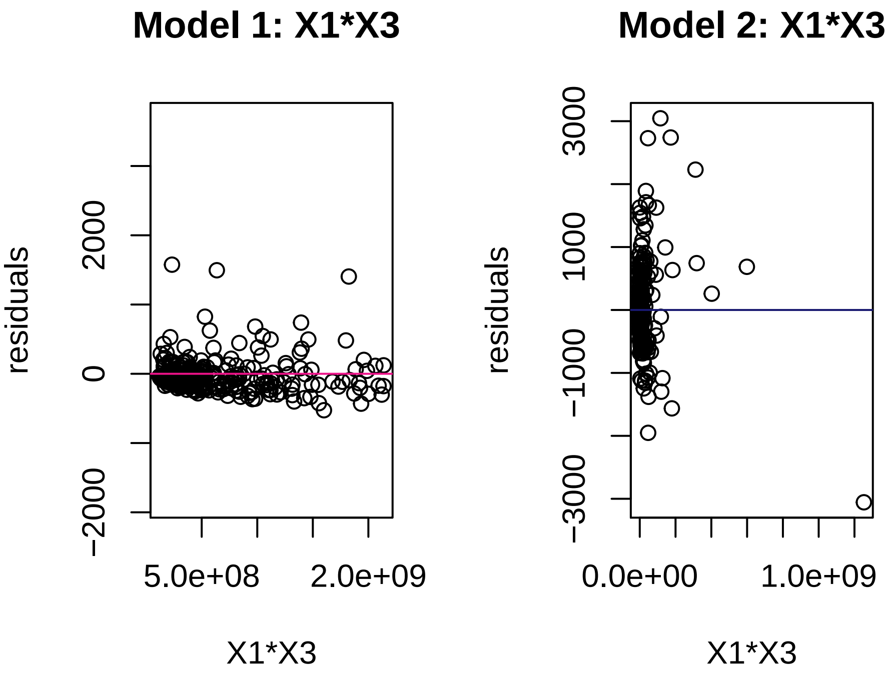

Model 1: X1
residuals
−2000
0
2000
0e+00
6e+06
X1
Model 2: X1
residuals
−3000
−1000
1000
3000
0
15000
X1
Model 1: X2
residuals
−2000
0
2000
0
10000
X2
Model 2: X2
residuals
−3000
−1000
1000
3000
5
15
25
35
X2

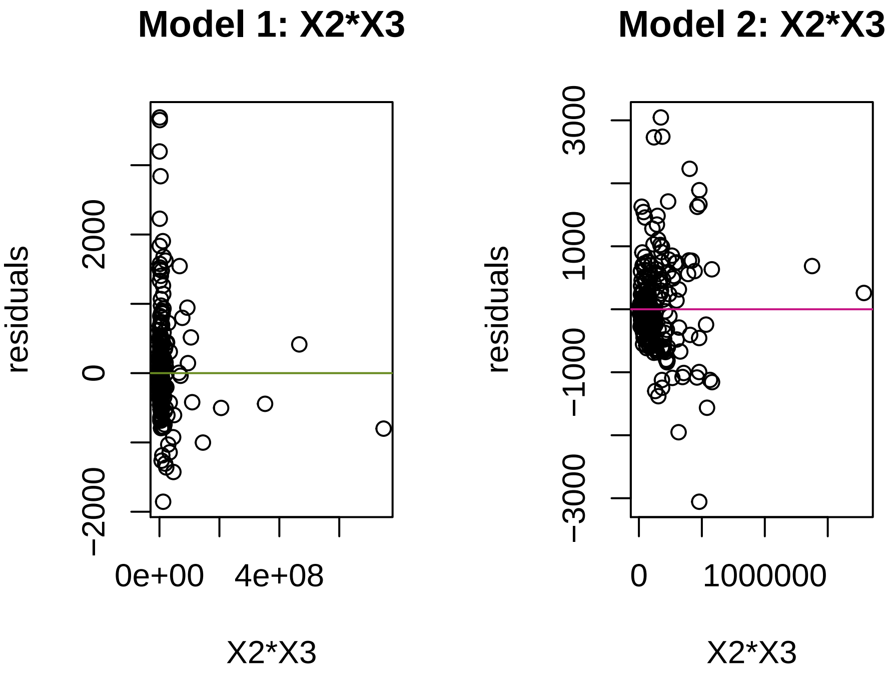

Model 1: X3
residuals
−2000
0
2000
0
100000
X3


Model 2: X3
residuals
−3000
−1000
1000
3000
0
100000
X3


Model 1: X1*X2
residuals
−2000
0
2000
0.0e+00
1.5e+09
X1*X2


Model 2: X1*X2
residuals
−3000
−1000
1000
3000
0e+00
3e+05
X1*X2





Model 1 graphs –> Best predictor is from Model 1 is X2 if we are to look at individual graphs. When combining graphs, it seems the plots are relatively equal in terms of accuracy on the regression model, as they adopt similar graphs. This is the same for the individual graphs. If one of the combined graphs had to be chosen, the closest one is that of X1 * X3.

Model 2 graphs –> Best predictor is from Model 2 is X2 if we are to look at individual graphs. When combining graphs, it seems the plots are relatively equal in terms of accuracy on the regression model, as they adopt similar graphs. This is the same for the individual graphs. If one of the combined graphs had to be chosen, the closest one is that of X2 * X3.

Normal Probability Plots –> The plots both show that the residuals of both models are following a normal distribution, however model 2 is more strictly following the distribution, as it fits the normality line more than model 1's residual normal probability plot

Overall –> Overall, the models are relatively equal in terms of their accuracy to their respective regressions. The best predictors are from Model 2 as that residual plot is more normal oriented than Model 1 is. Choosing a single best model from there is not advisable as they are all relatively the same in terms of accuracy.

6.28 f)

Now expand both models proposed above by adding all possible two-factor interactions. Note that, for a model with X1, X2, X3 as the predictors, the two-factor interactions are X1X2, X1X3, X2X3. Repeat part d for the two expanded models.

```
## Model 1- 2 Factor R2:
```

```
## [1] 0.9063789
```

```
## Model 2- 2 Factor R2:
```

```
## [1] 0.9230238
```

Part II - Multiple Linear Regressions 2

7.37

Refer to the CDI data set in Appendix C.2. For predicting the number of active physicians(Y) in a country, it has been decided to include total population(X1) and the total personal income(X2) as predictor variables. The question now is whether an additional predictor variable would be helpful in the model and, if so, which variable would be the most helpful. Assume that a first-order multiple regression model is appropriate.

7.37 a)

For each of the following variables, calculate the coefficient of partial determination given that XI and X2 are included in the model: land area (X3), percent of population 65 or older (X4), number of hospital beds (X5), and total serious crimes (X6).

```
## Coefficient of Partial Determination for X3:  0.02882495
```

```
## Coefficient of Partial Determination for X4:  0.003842367
```

```
## Coefficient of Partial Determination for X5:  0.5538182
```

7.37 b)

On the basis of the results in part (a), which of the four additional predictor variables is best? Is the extra sum of squares associated with this variable larger than those for the other three variables? Based on the coefficients of partial determination obtained above for each of our variables, X5 (number of hospital beds) is the best predictor variable. There is a clear difference between the value obtained for this variable compared to the other two.

7.37 c)

Using the F* test statistic, test whether or not the variable determined to be best in part (b) is helpful in the regression model when XI and X2 are included in the model; use alpha = 0.Ol. State the alternatives, decision rule, and conclusion. Would the F* test statistics for the other three potential predictor variables be as large as the one here? Discuss. Ho –> B5 = 0 Ha –> B5 != 0

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1243181164	1243181163.8	8617.6991	0
X2	1	22058054	22058054.1	152.9059	0
X5	1	78070132	78070131.6	541.1801	0
Residuals	436	62896949	144259.1		

```
## F* -->  541.1801
```

```
## F(0.99, 1, n-4) -->  6.693358
```

Decision Rule –> Reject null hypothesis if F-statistic is greater than critical value at level 0.01. Conclusion–> Reject the null Hypothesis, conclude B5 != 0.

7.37 d)

Compute three additional coefficients of partial determination:(CHANGE NOTATION HERE AT THE END) R2Y,X3,X4|X1,X2,R2Y,X3,X5|X1,X2, and R2Y,X4,X5|X1,X2(STOP CHANGING THE NOTATION HERE AT THE END). Which pair of predictors is relatively more important than other pairs? Use the F test to find out whether adding the best pair to the model is helpful given that X1, X2 are already included.

```
## Coefficient of Partial Determination for X3, X4:  0.03314181
```

```
## Coefficient of Partial Determination for X3, X5:  0.5558232
```

```
## Coefficient of Partial Determination for X4, X5:  0.5642756
```

In this case, the pairing of X4, X5 is the best, as it has the highest coefficient of partial determination.

Ho –> B4 = B5 = 0 Ha –> B4 != 0 or B5 != 0

```
## F* -->  281.6688
```

```
## F(0.99, 1, n-4) -->  6.693358
```

Decision –> Reject if F* > F(0.99, 1, n-4) Conclusion –> Reject null. B4, B5, or both is not 0. In this case, adding the pair is beneficial as it has an impact on the data because one or both of B4 and B5 is not 0.

Part III - Discussion

Discuss about your results from a practical standpoint. What particular parts of the course material do you find most relevant to your analysis in this project (try to be as specific as possible)? Any suggestions on how to improve the linear regression models?

Part1: For the first part of this project we considered to models composed of three predictor variables each and compared them to analyze which model would be better at predicting the number of active physicians (Y). On the first section we focused on understanding the distribution of our data as well as the relationship between the response and predictor variables for each of the proposed models. Based on our results, model 2 appears to be the best model alternative (which includes the predictors population density, percent of population greater than 64 years old, and total personal income).

Our analysis for this part was greatly influenced by both the coefficient of partial determination and the correlation and scatterplot matrices since they allowed us to visualize and to obtain a clearer idea of the fit of each model when comparing and analyzing them.

A suggestion that could improve the linear regression function for this part would be considering the distribution of the data (by looking at the graphs obtained for this part) and paying attention to the outliers.

Part II: On the second section of this project we focused on making predictions and building our analysis towards finding the best additional variable when starting with a model that was already in progress. We were able to compare three different variables: X3 (land area), X4 (percent of population 65 or older), and X5 (number of hospital beds). Based on our results, the number of hospital beds represents the best variable to be added to the model and, based on the F* test statistic, adding this variable to the model would indeed be helpful.

For this part, the coefficient of partial determination and the F* test statistic guided our analysis and overall conclusions and allowed us to understand the impact and considerations that must be followed when trying to improve a model.

A suggestion that could improve the linear regression model for this part would be considering other predictor variables outside the ones proposed in this section in order to make sure that the additional variable was the best one among all variables in the data.

Appendix

```r
library(knitr)
knitr::opts_chunk$set(
    error = FALSE,
    message = FALSE,
    warning = FALSE,
    echo = FALSE, # hide all R codes!!
    fig.width=5, fig.height=4,#set figure size
    fig.align='center',#center plot
    options(knitr.kable.NA = ''), #do not print NA in knitr table
    tidy = FALSE #add line breaks in R codes
)

CDI = read.table("CDI.txt")
#"ID", "County", "State", "Area", "Population", "18_to_34", "65+", "Active_Physicians", "Beds", "Seriou
Y = CDI[,8] #Active Physicians

MIX1 = CDI[,5] #Total Population
```

```r
MIX2 = CDI[,4] #Land Area
MIX3 = CDI[,16] #Total Personal Income Model 1

MIIX1 = CDI[,5] / CDI[,4] #Population Density
MIIX2 = CDI[,7] #% Older than 64
MIIX3 = CDI[,16] #Total Personal Income Model 2

n = length(Y)

# Model 1
cat("Model 1\n")
cat("Population\n")
stem(MIX1)
cat("Land Area\n")
stem(MIX2)
cat("Total Personal Income\n")
stem(MIX3)

# Model 2
cat("Model 2\n")
cat("Population Density\n")
stem(MIIX1)
cat("Percent Older than 64\n")
stem(MIIX2)
cat("Total Personal Income\n")
stem(MIIX3)
cat("Model 1: \n")
cat("Scatter Plot Matrix Model 1: \n")
M1 = data.frame(Y, MIX1, MIX2, MIX3)
colnames(M1) = c("Y","X1","X2","X3")
pairs(M1)
cat("Correlation Matrix Model 1: \n")
cor(M1)

cat("\n\n")

cat("Model 2: \n")
cat("Scatter Plot Matrix Model 2: \n")
M2 = data.frame(Y, MIIX1, MIIX2, MIIX3)
colnames(M2) = c("Y","X1","X2","X3")
pairs(M2)
cat("Correlation Matrix Model 2: \n")
cor(M2)
cat("Model 1: \n")
fitM1 = lm(Y ~ MIX1+MIX2+MIX3)
betaM1=fitM1$coefficients
cat("Y = ", betaM1[1], " + ", betaM1[2], "X1" , betaM1[3], "X2", " + ", betaM1[4], "X3\n")

cat("\n\n")

cat("Model 2: \n")
fitM2 = lm(Y ~ MIIX1+MIIX2+MIIX3)
betaM2=fitM2$coefficients
```

```r
cat("Y = ", betaM2[1], " + ", betaM2[2], "X1", " + ", betaM2[3], "X2", " + ", betaM2[4], "X3\n")

cat("Model 1 R2: \n")
summary(fitM1)$r.squared

cat("\n")

cat("Model 2 R2: \n")
summary(fitM2)$r.squared
residualsM1 = fitM1$residuals
y_hatM1 = fitM1$fitted.values
residualsM2 = fitM2$residuals
y_hatM2 = fitM2$fitted.values

par(mfrow = c(2, 2))
plot(x=y_hatM1, y=residualsM1 ,xlab=’fitted values’, ylab=’residuals’, main = "Model 1 Predictors")
abline(h=0, col=’rosybrown2’)
plot(x=y_hatM2, y=residualsM2 ,xlab=’fitted values’, ylab=’residuals’, main = "Model 2 Predictors")
abline(h=0, col=’cadetblue1’)


qqnorm(residualsM1)
qqline(residualsM1, col=’darkmagenta’)
qqnorm(residualsM2)
qqline(residualsM2, col=’peachpuff2’)

par(mfrow = c(1, 2))
plot(MIX1, y=residualsM1, xlab=’X1’, ylab=’residuals’, main = "Model 1: X1")
abline(h=0, col=’darkolivegreen3’)
plot(MIIX1, y=residualsM2, xlab=’X1’, ylab=’residuals’, main = "Model 2: X1")
abline(h=0, col=’coral3’)

par(mfrow = c(1, 2))
plot(MIX2, y=residualsM1, xlab=’X2’, ylab=’residuals’, main = "Model 1: X2")
abline(h=0, col=’salmon’)
plot(MIIX2, y=residualsM2, xlab=’X2’, ylab=’residuals’, main = "Model 2: X2")
abline(h=0, col=’darkgoldenrod’)

par(mfrow = c(1, 2))
plot(MIX3, y=residualsM1, xlab=’X3’, ylab=’residuals’, main = "Model 1: X3")
abline(h=0, col=’paleturquoise1’)
plot(MIIX3, y=residualsM2, xlab=’X3’, ylab=’residuals’, main = "Model 2: X3")
abline(h=0, col=’palevioletred3’)

par(mfrow = c(1, 2))
plot(MIX1*MIX2, y=residualsM1, xlab=’X1*X2’, ylab=’residuals’, main = "Model 1: X1*X2")
abline(h=0, col=’khaki3’)
plot(MIIX1*MIIX2, y=residualsM2, xlab=’X1*X2’, ylab=’residuals’, main = "Model 2: X1*X2")
abline(h=0, col=’forestgreen’)

par(mfrow = c(1, 2))
plot(MIX3*MIX1, y=residualsM1, xlab=’X1*X3’, ylab=’residuals’, main = "Model 1: X1*X3 ")
abline(h=0, col=’deeppink2’)
```

```r
plot(MIIX1*MIIX3, y=residualsM2, xlab='X1*X3', ylab='residuals', main = "Model 2: X1*X3")
abline(h=0, col='midnightblue')

par(mfrow = c(1, 2))
plot(MIX2*MIX3, y=residualsM1, xlab='X2*X3', ylab='residuals', main = "Model 1: X2*X3")
abline(h=0, col='olivedrab4')
plot(MIIX2*MIIX3, y=residualsM2, xlab='X2*X3', ylab='residuals', main = "Model 2: X2*X3")
abline(h=0, col='mediumvioletred')


cat("Model 1- 2 Factor R2: \n")
fit2M1 = lm(Y~ MIX1*MIX2+MIX1*MIX3+MIX2*MIX3)
summary(fit2M1)$r.squared

cat("\n")

cat("Model 2- 2 Factor R2: \n")
fit2M2 = lm(Y~ MIIX1*MIIX2+MIIX1*MIIX3+MIIX2*MIIX3)
summary(fit2M2)$r.squared

Y = CDI[,8]
X1 = CDI[,5]
X2 = CDI[,16]
X3 = CDI[,4]
X4 = CDI[,7]
X5 = CDI[,9]

#X3
full3 = lm(Y ~ X1+X2+X3)
Reduced = lm(Y ~ X1+X2)
yhat.f3 = full3$fitted.values
SSR.f3 = sum((yhat.f3 - mean(Y))^2) #SSR x1,x2,x3
yhat.r = Reduced$fitted.values
SSR.r = sum((yhat.r - mean(Y))^2) #SSR x1,x2
SSR3 = SSR.f3 - SSR.r #SSR(X1,X2.x3) - SSR(x1,x2)
SSE.r = sum((Y - yhat.r)^2) #SSE X1,X2
cat("Coefficient of Partial Determination for X3: ", SSR3/SSE.r, "\n")


#X4
full4 = lm(Y ~ X1+X2+X4)
yhat.f4 = full4$fitted.values
SSR.f4 = sum((yhat.f4 - mean(Y))^2) #SSR x1,x2,x4
SSR4 = SSR.f4 - SSR.r #SSR(X1,X2.x4) - SSR(x1,x2)
cat("Coefficient of Partial Determination for X4: ", SSR4/SSE.r, "\n")


#X5
full5 = lm(Y ~ X1+X2+X5)
yhat.f5 = full5$fitted.values
SSR.f5 = sum((yhat.f5 - mean(Y))^2) #SSR x1,x2,x4
SSR5 = SSR.f5 - SSR.r #SSR(X1,X2.x4) - SSR(x1,x2)
cat("Coefficient of Partial Determination for X5: ", SSR5/SSE.r, "\n")
```

```
n=length(Y)
kable(anova(full5))
cat("F* --> ",anova(full5)[3,4], "\n")
cat("F(0.99, 1, n-4) --> ",qf(0.99, 1,n - 4), "\n")
#R^2(3,4|1,2):
full34 = lm(Y ~ X1+X2+X3+X4)
yhat.f34 = full34$fitted.values
SSR.f34 = sum((yhat.f34 - mean(Y))^2) #SSR x1,x2,x3,x4
SSR34 = SSR.f34 - SSR.r #SSR(X1,X2.x3,x4) - SSR(x1,x2)
cat("Coefficient of Partial Determination for X3, X4: ", SSR34/SSE.r, "\n")

#R^2(3,4|1,2):
full35 = lm(Y ~ X1+X2+X3+X5)
yhat.f35 = full35$fitted.values
SSR.f35 = sum((yhat.f35 - mean(Y))^2) #SSR x1,x2,x3,x5
SSR35 = SSR.f35 - SSR.r #SSR(X1,X2.x3,x5) - SSR(x1,x2)
cat("Coefficient of Partial Determination for X3, X5: ", SSR35/SSE.r, "\n")

#R^2(3,4|1,2):
full45 = lm(Y ~ X1+X2+X4+X5)
yhat.f45 = full45$fitted.values
SSR.f45 = sum((yhat.f45 - mean(Y))^2) #SSR x1,x2,x4,x5
SSR45 = SSR.f45 - SSR.r #SSR(X1,X2.x4,x5) - SSR(x1,x2)
cat("Coefficient of Partial Determination for X4, X5: ", SSR45/SSE.r, "\n")

SSE.r = sum((Y - yhat.r)^2) #SSE X1,X2
yhat.45 = full45$fitted.values
SSE.45 = sum((Y - yhat.45)^2)
fstar = ((SSE.r - SSE.45)/(2)/((SSE.45)/(n - 5)))
cat("F* --> ",fstar, "\n")
cat("F(0.99, 1, n-4) --> ",qf(0.99, 1,n - 4), "\n")
```