Multiple Linear Regression

Introduction.

Our data set consists of county demographic information (CDI) for 440 of the most populous counties in the United States. Each of our observations has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. The information generally pertains to the years 1990 and 1992. The variables that we are interested in are: total population, number of active physicians, number of hospital beds, percent bachelor's degrees, total personal income, region and per capita income.

There were three main objectives in our study. First we needed to create a simple linear model between Y variable and a few X variables . Our Y variable was the number of active physicians in a CDI, and X's were total population, number of hospital beds, and total personal income. The goal was to find the relationship between Y and X, and come up with intercepts, the coefficient of determination(R2) and graph our linear model. The second objective was to find linear relationship for each geographic region and regress per capita income in a CDI (Y) against the percentage of individuals in a county having at least a bachelor's degree (X). Second objective also included finding confidence intervals for B1(slope) of our models. The most important objective of this project was to learn how to work with a data, computing tool and applying things we have learned in the class in practice.

Main tool that was used in this study in statistical computing language R. The next pages will go in a following format. First I give a chunk of the code, a plot(if needed), below there is a brief elaboration of the code and answers to the problem questions. Furthermore on white space I'm planning to clarify my work by pen.

Part 1: Multiple linear regression 1.

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 1 includes as predictor variables total population (X 1), land area (X 2), and total personal income (X 3). Proposed model 2 includes as predictor variables population density (X 1, total population divided by land area), percent of population greater than 65 years old (X 2), and total personal income (X 3).

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

- Stem-and-leaf (stemplots) plots help us to identify skew/symmetry and possible outliers. Also they help us to determine the "shape" of a data set, and locate the center.
- Stemplots are created by dividing the range of the data into equal units to be used on the

stem. Then we attach “leafs” to represent each data point.

- Most of our stemplots are right skewed (positive skewness) with some obvious outliers. Stemplot for "the percentage population greater than 65” is the only bell shaped curve, in which data points are centered in the mid 10’s and tail off in both directions.

```
> stem(dat$pop)

  The decimal point is 6 digit(s) to the right of the |

  0 | 11111111111111111111111111111111111111111111111111111111111111111111111111111111+254
  0 | 55555555555555555555556666666666666677777777777777777778888888888
  1 | 000000122233333444
  1 | 55699
  2 | 1134
  2 | 58
  3 |
  3 |
  4 |
  4 |
  5 | 1
  5 |
  6 |
  6 |
  7 |
  7 |
  8 |
  8 | 9
> stem(dat$land)

  The decimal point is 3 digit(s) to the right of the |

   0 | 000011111111111222222222222222222233333333333333333333333444444+252
   1 | 000000000000001111111111111222222222233333344445556666777788899999
   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
> stem(dat$income)

  The decimal point is 4 digit(s) to the right of the |

   0 | 11111111111122222222222222222222222222222222222222222222222222222222+263
   1 | 000000000001111111122222333334444444555555556778888888999
   2 | 001111233344477788899
   3 | 0255678899
   4 | 19
   5 | 59
   6 |
   7 |
   8 |
   9 |
  10 |
  11 | 1
  12 |
  13 |
  14 |
  15 |
  16 |
  17 |
  18 | 4
```

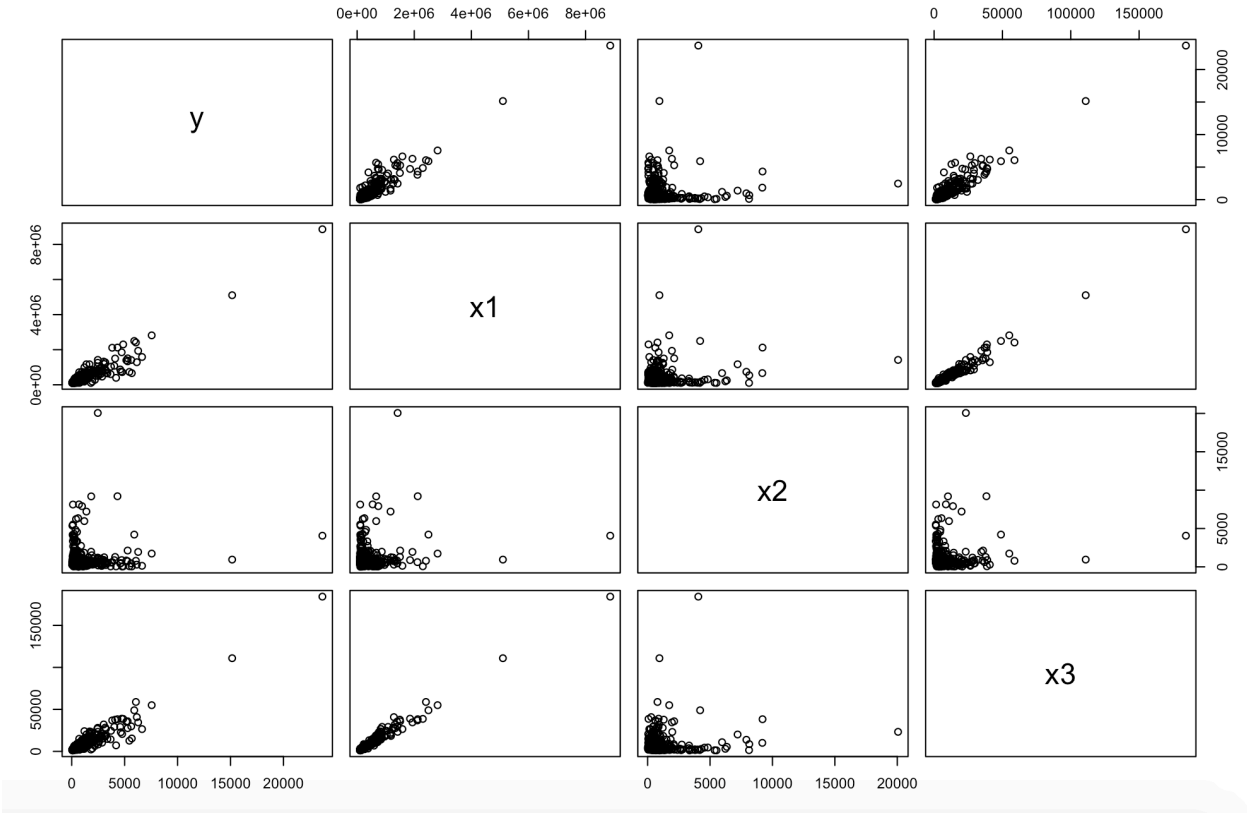
```
> stem(dat$`pop65+`)

  The decimal point is at the |

   2 | 0
   4 | 47890389
   6 | 112345567799013456667899
   8 | 00112222233334444555666777778888899990002222333333444444445555666677
  10 | 000111111222222222233333344444455555556666666677777778888888888899999+36
  12 | 000000001111122223333333333444445555555666666677777777788889990000000+36
  14 | 000011111112233344444555677889000000111122223455667778
  16 | 12556699901122345
  18 | 06778
  20 | 070
  22 | 018828
  24 | 47
  26 | 055
  28 | 1
  30 | 7
  32 | 138

> stem(popden)

  The decimal point is 3 digit(s) to the right of the |

   0 | 000000000000000011111111111111111111111111111111111111111111111111111111111+321
   2 | 00001112233456700111145
   4 | 05884
   6 | 2464
   8 | 19
  10 | 378
  12 |
  14 | 4
  16 |
  18 |
  20 |
  22 |
  24 |
  26 |
  28 |
  30 |
  32 | 4
```

b. Obtain the scatter plot matrix and the correlation matrix for each proposed model. Summarize the information provided.
First I created a dataset of variables that are in our interest for model1, and called it model1. The same steps have been done for model2.

- In Model 1 we can observe Y has a positive linear association with X1(total population) and X3(total personal income).

Model1

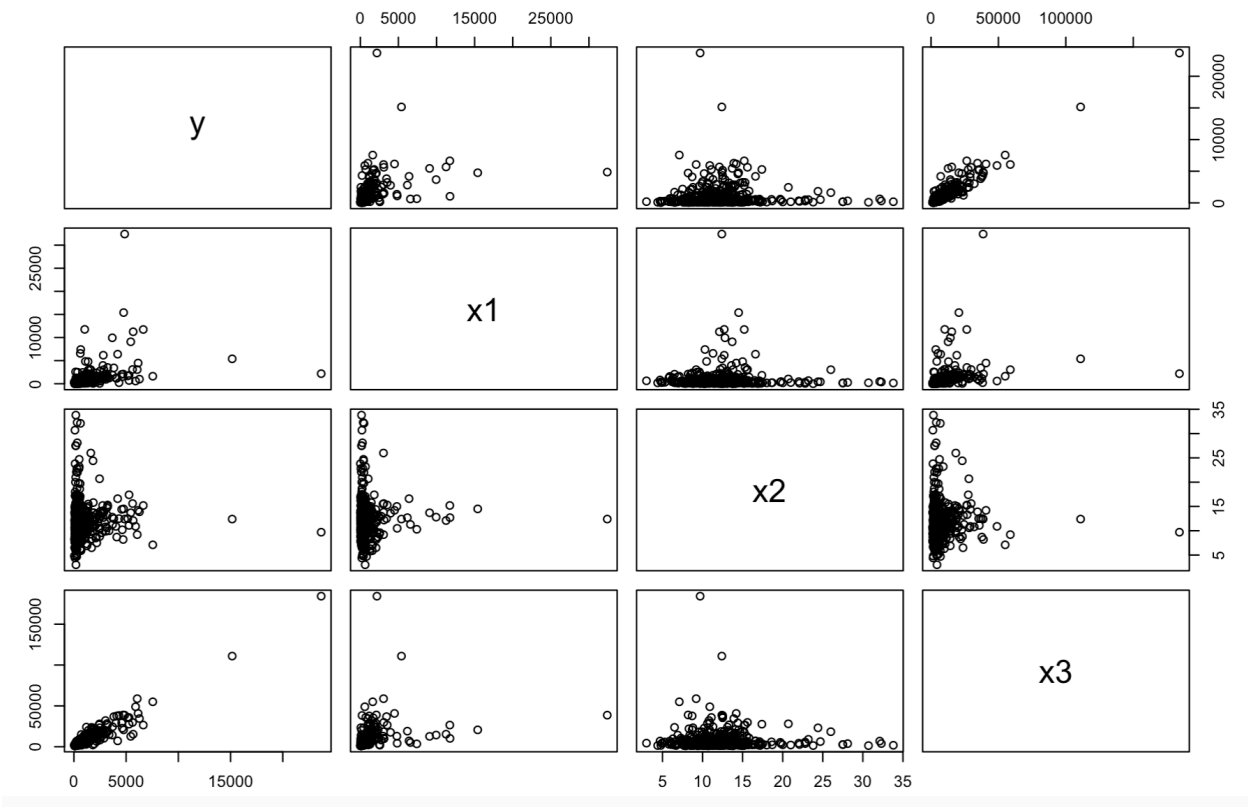
```
> y<-c(dat$physic)
> x1<-c(dat$pop)
> x2<-c(dat$land)
> x3<-c(dat$income)
> model1<-data.frame()
> model1<-data.frame(y,x1,x2,x3)
> model1
       y      x1    x2     x3
1  23677 8863164  4060 184230
2  15153 5105067   946 110928
3   7553 2818199  1729  55003
> cor(model1)
            y        x1         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
```



- In Model 2, only X3(total personal income) has a strong positive linear association with Y. Y has a weak positive association with X1(population density).

Model 2

```
> x1=c(popden)
> x2=c(dat$`pop65+`)
> x3=c(dat$income)
> model2=data.frame(y,x1,x2,x3)
> model2
      y          x1   x2     x3
1 23677  2183.04532  9.7 184230
2 15153  5396.47674 12.4 110928
3  7553  1629.95894  7.1  55003
> plot(model2)
> cor(model2)
            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
```



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

```
> model1

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
(Intercept)           x1           x2           x3
 -1.332e+01    8.366e-04   -6.552e-02    9.413e-02

> model2=lm(y~x1+x2+x3)
> model2

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
(Intercept)          x1          x2          x3
 -170.57422     0.09616     6.33984     0.12657
```

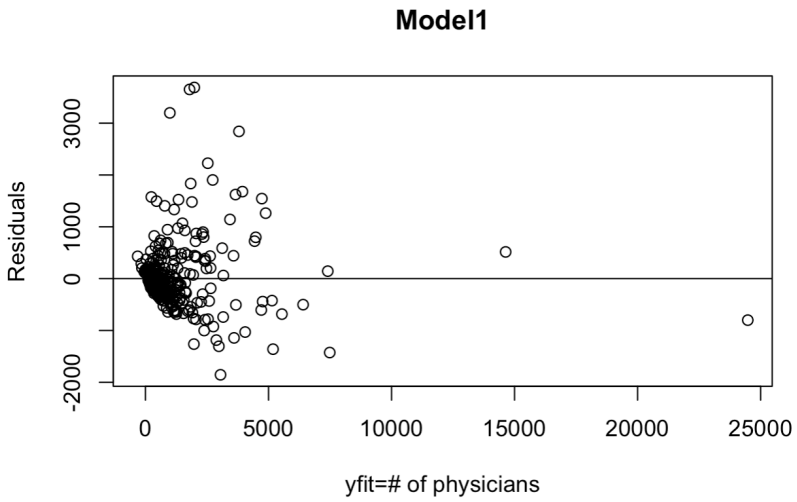
d. Calculate R2 for each model. Is one model clearly preferable in terms of this measure?

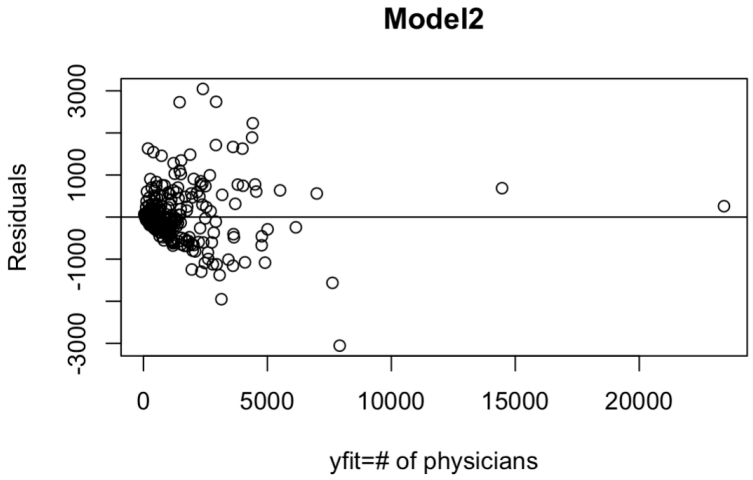
- There is no clearly preferable model, because R2 for both models are very close to each other. But model 2 has a slightly higher R2.

```
> summary(model2)$r.squared
[1] 0.9117491
> summary(model1)$r.squared
[1] 0.9026432
```

e. For each model, obtain the residuals and plot them against Y-fitted, 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?

- The residuals against fitted values plot is useful for detecting nonlinearity, heteroskedasticity, and outliers.
- Both Model 1 and Model 2 plots of the residuals against Y-fitted look normal, there are no patterns. Both models have some outliers that is why points(observations) are mostly on the left side of the graphs.

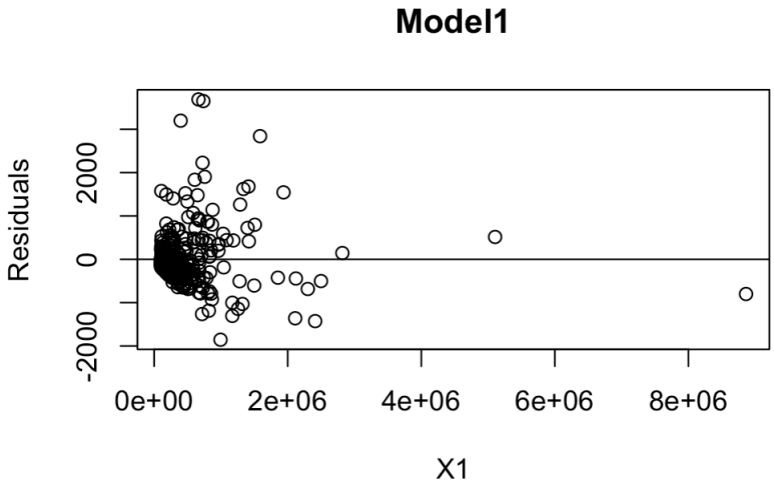


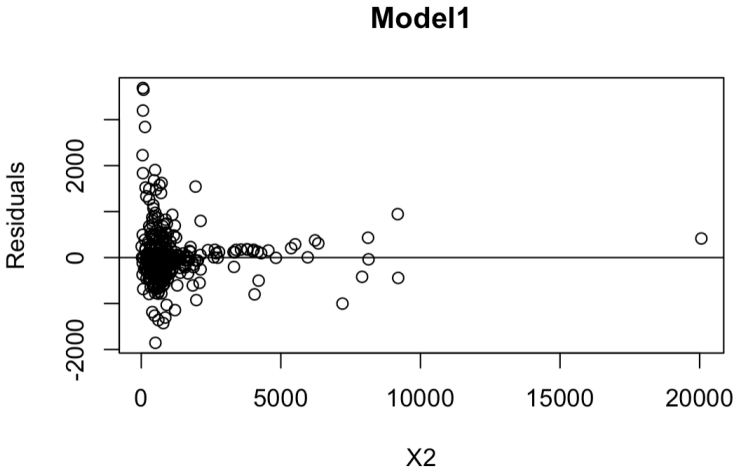


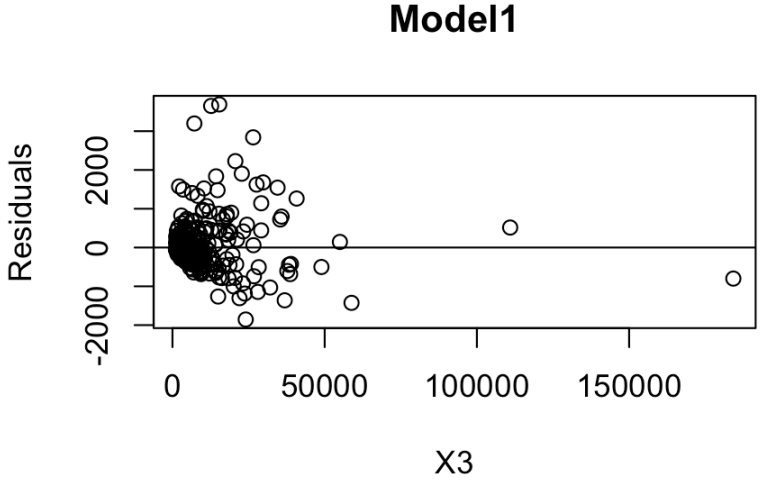

```
> yfit1=model1$fitted.values
> yfit2=model2$fitted.values
> plot(yfit1,model1$residuals, xlab="yfit=# of physicians", ylab="Residuals",main="Model1")
> abline(h=0)
> plot(yfit2,model2$residuals, xlab="yfit=# of physicians", ylab="Residuals",main="Model2")
> abline(h=0)
```

- The residuals against predictors plots are also useful for detecting nonlinearity, heteroskedasticity, and outliers. In simple linear regression there is just one plot of residuals against X. In multiple regression there are multiple predictors, and we need to make a plot for each one.

Model1

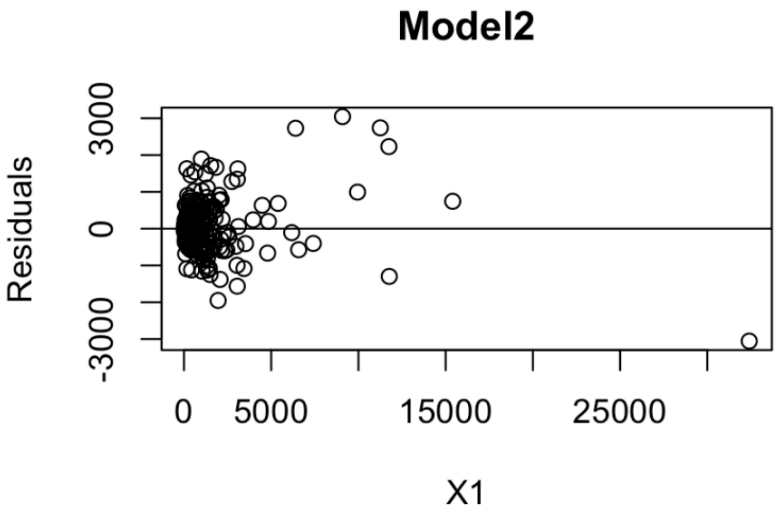


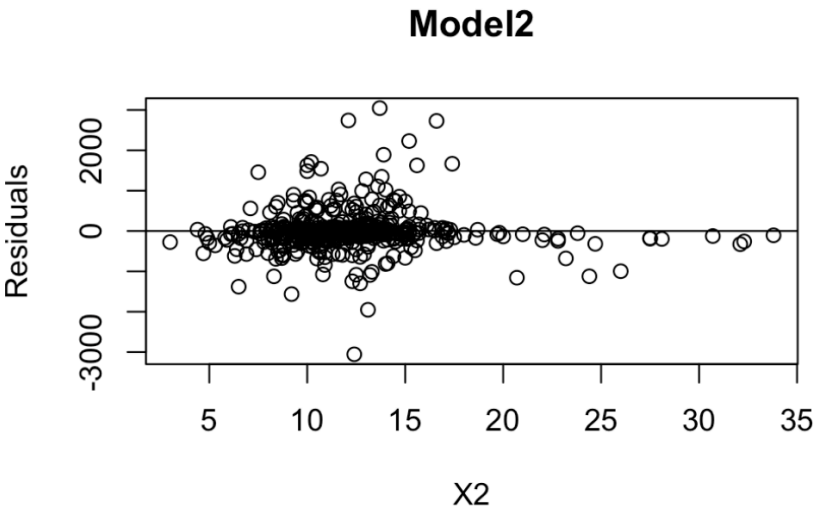


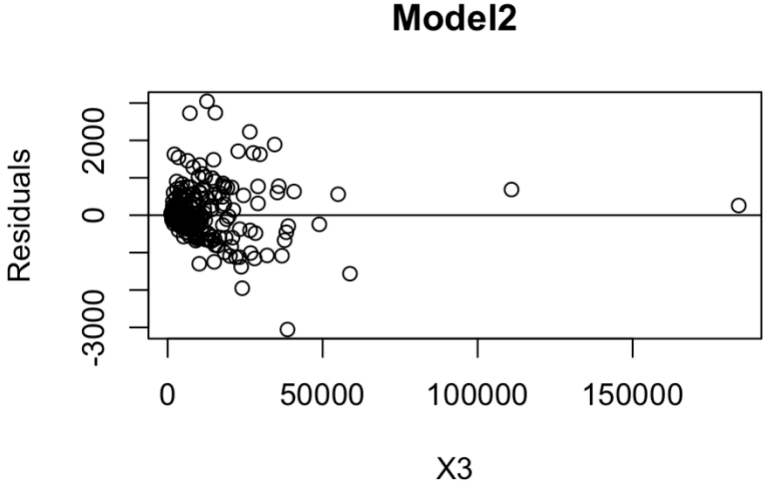



```
> plot(x1, model1$residuals, xlab = 'X1', ylab = 'Residuals', main="Model1")
> plot(x2, model1$residuals, xlab = 'X2', ylab = 'Residuals', main="Model1")
> plot(x3, model1$residuals, xlab = 'X3', ylab = 'Residuals', main="Model1")
> abline(h=0)
```

Model2


Model2
Residuals
3000
0
-3000
0
5000
15000
25000
X1

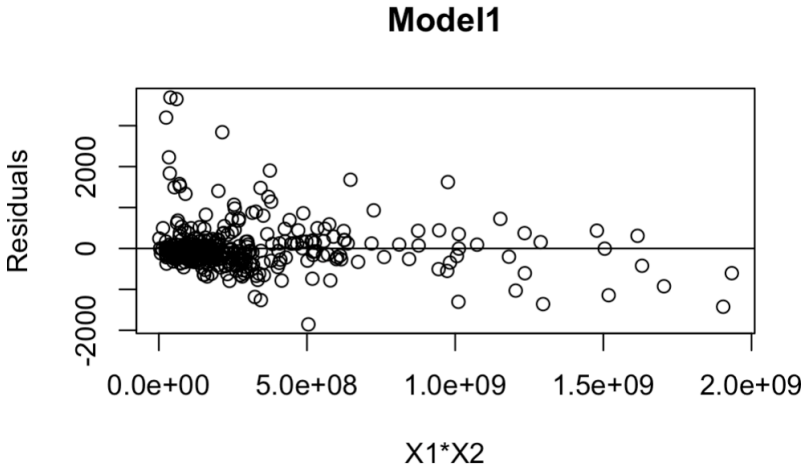

Model2
Residuals
2000
0
-3000
5
10
15
20
25
30
35
X2

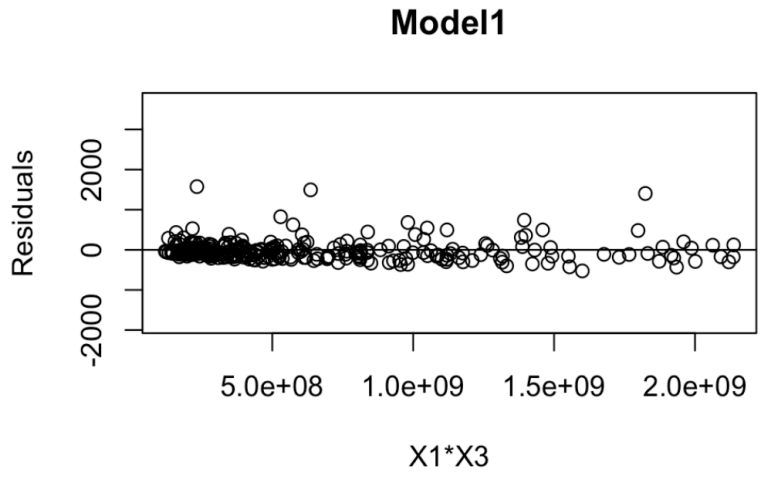

Model2
Residuals
2000
0
-3000
0
50000
100000
150000
X3

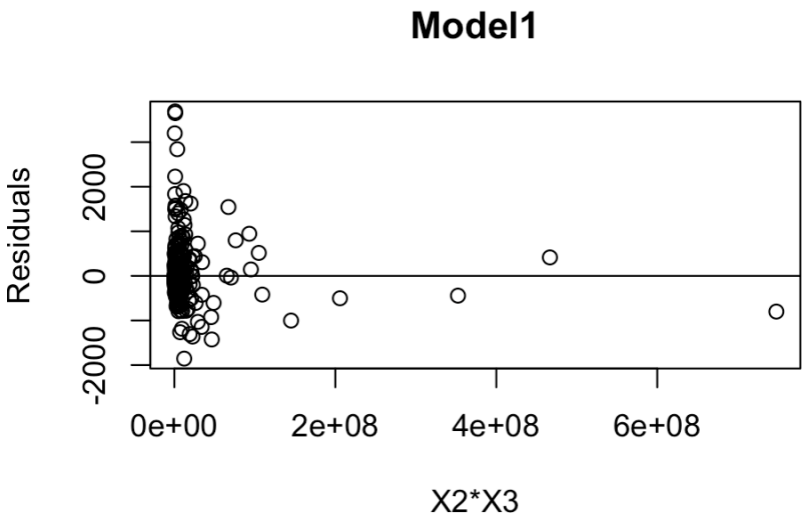
```
x1=popden
x2=dat$`pop65+`
x3=dat$income
plot(x1, model2$residuals, xlab = 'X1', ylab = 'Residuals', main="Model2")
abline(h=0)
plot(x2, model2$residuals, xlab = 'X2', ylab = 'Residuals', main="Model2")
abline(h=0)
plot(x3, model2$residuals, xlab = 'X3', ylab = 'Residuals', main="Model2")
abline(h=0)
```

The residuals against two-factor interaction terms(X1X2, X1X3, X2X3) for each of our models.

Model1

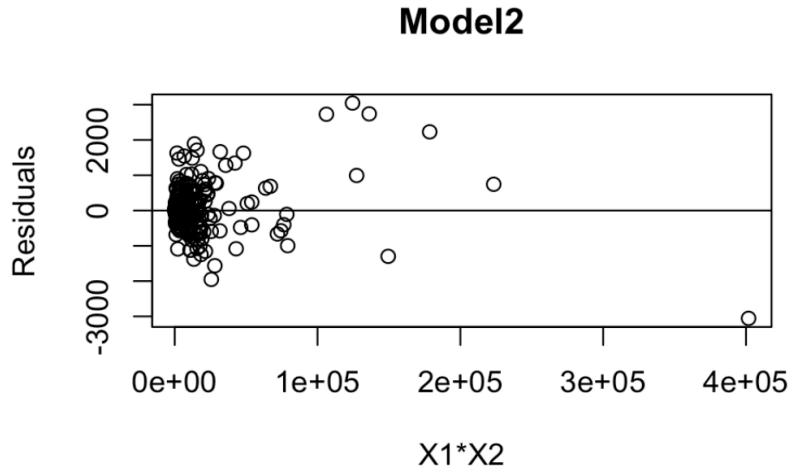


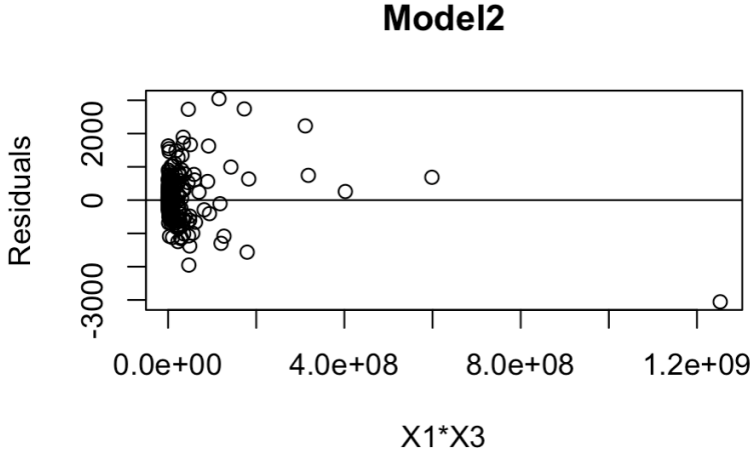


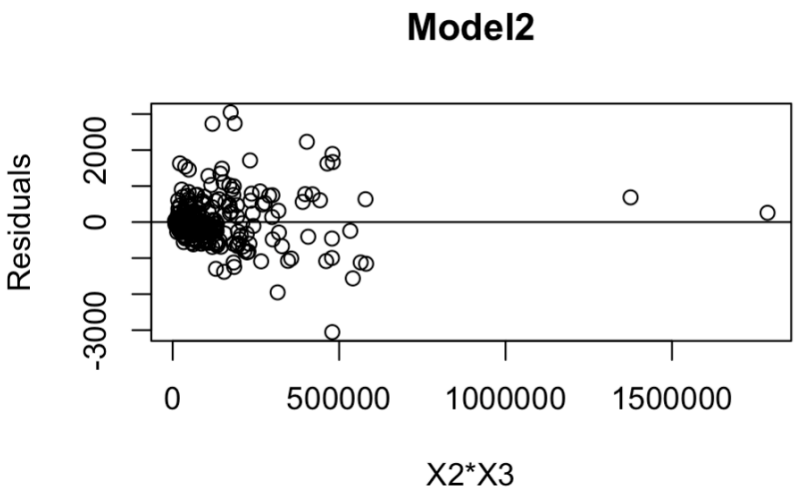


If there is no clear pattern in any of these two-factor interaction terms, so there is no indication that the new predictor(two-factor interaction term) should be included in the model.

Model2







- If there is no clear pattern in any of these two-factor interaction terms, so there is no indication that the new predictor(two-factor interaction term) should be included in the model.

f. Now expand both models proposed above by adding two-factor interactions. Repeat part d.(calculate R2) for the two expanded models.

Model1

```
> y=dat$physic
> x1=dat$pop
> x2=dat$land
> x3=dat$income
> model1t=lm(y~x1*x2*x3-x1:x2:x3)
> model1t

Call:
lm(formula = y ~ x1 * x2 * x3 - x1:x2:x3)

Coefficients:
(Intercept)           x1           x2           x3        x1:x2        x1:x3        x2:x3
 -5.826e+01    7.252e-04   -6.421e-02    1.087e-01    6.173e-07    1.696e-09   -3.706e-05
```

```
> summary(model1t)$r.sq
[1] 0.9063789
```

Model2

```
> x1=popden
> x2=dat$pop65
> head(x3)
[1] 184230 110928  55003  48931  58818  38658
```

```
> model2t=lm(y~x1*x2*x3-x1:x2:x3)
> model2t

Call:
lm(formula = y ~ x1 * x2 * x3 - x1:x2:x3)

Coefficients:
(Intercept)           x1           x2           x3        x1:x2        x1:x3        x2:x3
 -9.367e+00   -4.179e-01   -1.106e+01    1.477e-01    4.652e-02   -3.276e-06   -1.289e-03

> summary(model2t)$r.sq
[1] 0.9230238
```

Part 2: Multiple Linear Regression 2.

7.37. Refer to the CDI data set in Appendix C.2. For predicting the number of active physicians (Y) in a county, it has been decided to include total population (X1) and 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 most helpful. Assume that a first-order multiple regression model is appropriate.

a. For each of the following variables, calculate the coefficient of partial determination given that X1 and X2 are included in the model: land area (X3), percent of population 65 or older (X4), number of hospital beds (X5).

- The error sum of squares SSE for the model that contains total population (X1) and the total personal income (X2) is reduced by 2.88% when land area (X3) is added to the model.

```
> modelp2=lm(y~x1+x2)
> modelp2

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
 -64.438213     0.000531     0.107219

> anova(modelp2)
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq F value    Pr(>F)
x1          1 1243181164 1243181164 3853.88 < 2.2e-16 ***
x2          1   22058054   22058054   68.38 1.638e-15 ***
Residuals 437  140967081     322579
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> modelp2x1=lm(y~x1+x2+dat$land)
> anova(modelp2x1)
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq  F value    Pr(>F)
x1          1 1243181164 1243181164 3959.184 < 2.2e-16 ***
x2          1   22058054   22058054   70.249 7.271e-16 ***
dat$land    1    4063370    4063370   12.941 0.0003583 ***
Residuals 436  136903711     313999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> SSRnum=4063370
> SSEden=140967081
> R2=SSRnum/SSEden
> R2
[1] 0.02882496
```

- The error sum of squares SSE for the model that contains total population (X1), the total personal income (X2) is reduced by 0.38% when percent of population 65 or older (X4) is added to the model.

```
> anova(fit3)
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq   F value    Pr(>F)
x1          1 1243181164 1243181164 3859.8919 < 2.2e-16 ***
x2          1   22058054   22058054   68.4870 1.571e-15 ***
x4          1     541647     541647    1.6817    0.1954
Residuals 436  140425434     322077
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 541647/140967081
[1] 0.003842365
```

- The error sum of squares SSE for the model that contains total population (X1), the total personal income (X2), is reduced by 55.38% when number of hospital beds (X5) is added to the model.

```
> fit4=lm(y~x1+x2+x5)
> anova(fit4)
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq F value    Pr(>F)
x1          1 1243181164 1243181164 8617.70 < 2.2e-16 ***
x2          1   22058054   22058054  152.91 < 2.2e-16 ***
x5          1   78070132   78070132  541.18 < 2.2e-16 ***
Residuals 436   62896949     144259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 78070132/140967081
[1] 0.5538182
```

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?

- Since the error sum of squares SSE for the model that contains total population (X1), the total personal income (X2), has been reduced by 55.38% when number of hospital beds (X5) is added to the model. I conclude that X5 is the best additional predictor variable, since its marginal contribution is the greatest among other Xs we have tried.

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 X1 and X2 are included in the model; use alpha=0.01. 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.

```
> model.f=lm(y~x1+x2+x3)
> model.r=lm(y~x1+x2)
> anova(model.f)
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq F value    Pr(>F)
x1          1 1243181164 1243181164 8617.70 < 2.2e-16 ***
x2          1   22058054   22058054  152.91 < 2.2e-16 ***
x3          1   78070132   78070132  541.18 < 2.2e-16 ***
Residuals 436   62896949     144259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(model.r)
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq F value    Pr(>F)
x1          1 1243181164 1243181164 3853.88 < 2.2e-16 ***
x2          1   22058054   22058054   68.38 1.638e-15 ***
Residuals 437  140967081     322579
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> f.star=(140967081-62896949)/(62896949/436)
> f.star
[1] 541.1801
> qf(1-0.01,1,436)
[1] 6.693358
```

- Ho: B3=0, Ha: B3≠0, F* is 541.18 and F(1-0.01, 1, 436) is 6.693, so we conclude Ha: B3≠0. Which means the number of hospital beds (X3) is relevant to our model.

d. Compute three additional coefficients of partial determination R2(X3,X4 | X1,X2), R2(X3,X5 | X1,X2), R2(X4,X5 | X1,X2). Total population (X1), the total personal income (X2), land area (X3) and percent of population 65 or older (X4), number of hospital beds (X5).
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 in the model?

```
> x1=dat$pop
> x2=dat$income
> x3=dat$land
> x4=dat$pop65
> x5=dat$beds
```

R2(X3,X4 | X1,X2)

```
> model.b=lm(y~x1+x2)
> model.a=lm(y~x1+x2+x3+x4)
> SSE.before = sum(model.b$residuals^2)
> SSE.after = sum(model.a$residuals^2)
```

```
> partial.R2 = (SSE.before - SSE.after)/(SSE.before)
> partial.R2
[1] 0.03314181
```

R2(X3,X5 | X1,X2)

```
> model.a=lm(y~x1+x2+x3+x5)
> SSE.after = sum(model.a$residuals^2)
> partial.R2 = (SSE.before - SSE.after)/(SSE.before)
> partial.R2
[1] 0.5558232
```

R2(X4,X5 | X1,X2)

```
> model.a=lm(y~x1+x2+x4+x5)
> SSE.after = sum(model.a$residuals^2)
> partial.R2 = (SSE.before - SSE.after)/(SSE.before)
> partial.R2
[1] 0.5642756
```

Hypothesis test of whether adding the best pair to the model is helpful given that X1 and X3 are already in the model.

```
> model.r=lm(y~x1+x2)
> model.f=lm(y~x1+x2+x4+x5)
> anova(model.r,model.f)
Analysis of Variance Table

Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + x4 + x5
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1    437 140967081
2    435  61422794  2  79544288 281.67 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Ho: B3=B4=0, Ha: B3≠0 or B4≠0. F* is 281.67 and F statistic is 0, hence we conclude Ha.

Part 3: Discussion.

In part one for problem set 6.28, we have two models to compare. Model 1 consists of collecting all the data for the number of active physicians(Y) and total population(X1), land area(X2), and total personal income(X3). Model 2 consists of collecting all the data for the number of active physicians(Y) and predictor variables are population density(X1), percent of population greater than 65 y.o.(X2), and total personal income(X3). In part a) we are preparing stem-and-leaf plot for all of our predictor variables, for both models. Stem-and-leaf plots help us to identify skew/symmetry and possible outliers. They also help to determine the 'shape' of a data set, and locate the center. Our data set consists of 440 observations, and most of

them located in the right side giving us positive skewness, with some outliers. In part b) we are obtaining a scatterplot matrix and the correlation matrix. This is done, to see the correlation of data btwn each other(predictor variables) and Y-variable(response variable). In here we conclude that in Model 1 the number of active physicians(Y) has a strong positive linear association with total population(X1) and total personal income(X3), and weak correlation with land area(X2). In Model 2 the number of active physicians(Y) has a strong positive linear association with only one predictor variable - total personal income(X3), other two predictors have weak correlation. In part c) we have obtained the first-order regression models for both of our models. In part d) we have obtained R2 for each of the models, in here we conclude that both models have high and positive R2 close to 0.90's, which means there is no clearly preferable model. Then in part e) we have obtained the residuals and plot them against Y-fitted, and each of the three predictor variables, and each of the two-factor interaction terms. This was done to detect nonlinearity, heteroskedasticity, and see if there are any outliers. We have concluded that all plots are normal with some outliers and don't have any specific patterns. In part f) we have added two-factor interactions, this can greatly expand understanding of the relationships among the variables in the model and allows more hypotheses to be tested.

In part two, problem 7.37 creating first-order multiple regression models and checking for the coefficients of partial determination for different predictor variables when added to our model (y~x1+x2). By doing this we are trying to find the variable that reduces the error sum of squares(SSE) to the greatest extend. We concluded that adding X5 (number of hospital beds) variable we reduce SSE by 55.38% and it's marginal contribution is the greatest among other variables. In part c) we support our finings in previous parts by completing F test of whether X5 is helpful to our regression model. The F-test has also confirmed that beta of X5 is not zero, which means X5 is relevant to our model.