STA108 Project 2

Changhyun Park, Makenna Elise Stever, Seok Hyun Kim

05/03/2022

Introduction: Multiple linear regression can help us build models that are more complex and informative than one's built on a single variable. Using the same CDI dataset we used in Project 1, we will attempt to make conclusions based on this model. This process will include the observation of our predictor variables using the stem-and-leaf plot and calculation of the matrices(scatter plot and correlation) of models along with their coefficients of determination.We will conduct F-tests to determine whether the addition of a particular predictor is useful to the model. Then, we will circle back to our initial objective and determine which model is strongest and what can be done to improve it.

Part I: Multiple linear regression I. This part consists of Project 6.28 in the book, with the following additional part:

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 (XI), land area (X2), and total personal income (X3). Proposed model II includes as predictor variables population density (XI, total population divided by land area), percent of population greater than 64 years old (X2), and total personal income (X3).

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

```
CDI <- read.table("~/Downloads/Datasets_export/CDI.txt")
colnames(CDI) <- c("Idenitifation number", "County", "State", "Land area",
"Total population", "Percent of population 18~34", "Percent of population 65
or older", "Number of active physicians", "Number of hospital beds", "Total
serious crimes", "Percent high school graduates", "Percent bachelor's
degrees", "Percent below poverty level", "Percent unemployment", "per capita
income", "total person income", "geographic region")
dim(CDI)
```

```
## [1] 440  17
```

```
Y <- CDI$`Number of active physicians`

X11 <- CDI$`Total population`
X12 <- CDI$`Land area`
X13 <- CDI$`total person income`

X21 <- CDI$`Total population`/CDI$`Land area`
X22 <- CDI$`Percent of population 65 or older`
```

```
X23 <- CDI$`total person income`

stem(X11)

##
##   The decimal point is 6 digit(s) to the right of the |
##
##   0 |
11111111111111111111111111111111111111111111111111111111111111111111111111+254
##   0 | 55555555555555555555555566666666666666677777777777777777778888888888
##   1 | 000000122233333444
##   1 | 55699
##   2 | 1134
##   2 | 58
##   3 |
##   3 |
##   4 |
##   4 |
##   5 | 1
##   5 |
##   6 |
##   6 |
##   7 |
##   7 |
##   8 |
##   8 | 9

stem(X12)

##
##   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

stem(X13)

##
##   The decimal point is 4 digit(s) to the right of the |
##
##    0 |
1111111111112222222222222222222222222222222222222222222222222222+263
##    1 | 000000000000111111111222223333344444455555555567788888888999
##    2 | 001111233344477788899
##    3 | 0255678899
##    4 | 19
##    5 | 59
##    6 |
##    7 |
##    8 |
##    9 |
##   10 |
##   11 | 1
##   12 |
##   13 |
##   14 |
##   15 |
##   16 |
##   17 |
##   18 | 4

stem(X21)

##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 |
000000000000000011111111111111111111111111111111111111111111111111+321
##    2 | 00001112233456700111145
##    4 | 05884
##    6 | 2464
##    8 | 19
##   10 | 378
##   12 |
##   14 | 4
##   16 |
##   18 |
##   20 |
##   22 |
##   24 |
```

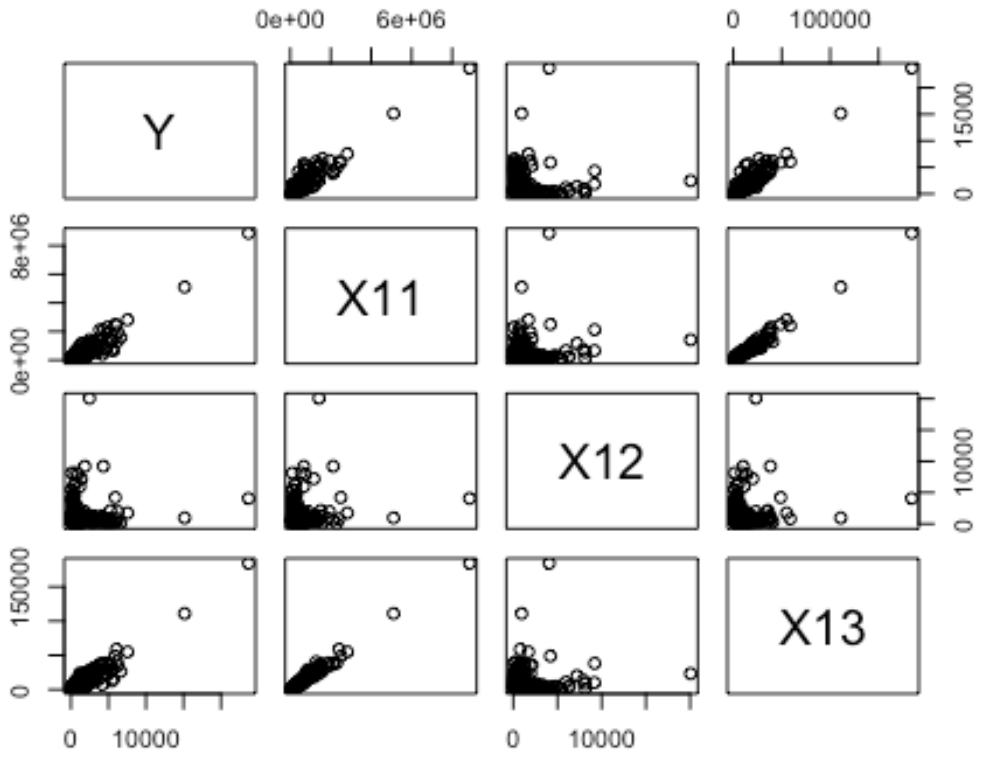
```
##   26 |
##   28 |
##   30 |
##   32 | 4

stem(X22)

##
##   The decimal point is at the |
##
##    2 | 0
##    4 | 47890389
##    6 | 112345567799013456667899
##    8 |
00112222333344445556667777788888999900022223333344444445555666677
##   10 |
000111111222222222233333344444455555556666666677777778888888888999999+36
##   12 |
000000001111122223333333333444455555556666667777777778888999000000000+36
##   14 | 000011111112233344444555677889000000111122223455667778
##   16 | 12556699901122345
##   18 | 06778
##   20 | 070
##   22 | 018828
##   24 | 47
##   26 | 055
##   28 | 1
##   30 | 7
##   32 | 138

stem(X23)

##
##   The decimal point is 4 digit(s) to the right of the |
##
##    0 |
11111111111122222222222222222222222222222222222222222222222222222+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
```

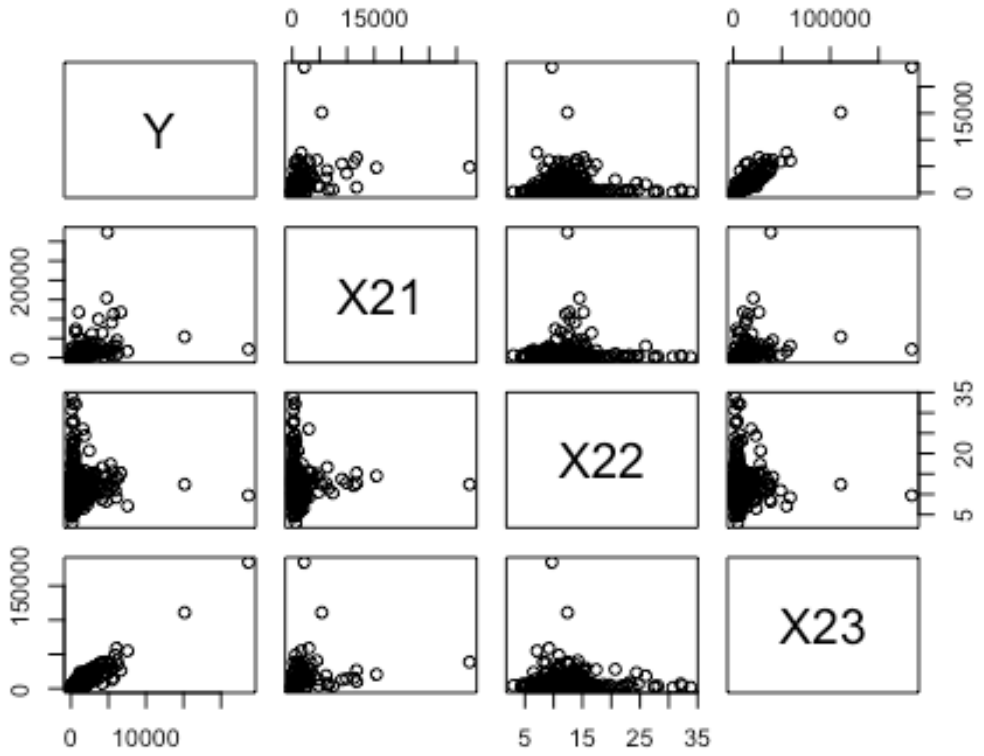
All stem-and-leaf plots display similar patterns except for the one for the percent of population older than 64 years old(X22). The predictor variables except for X22 are mostly skewed to the right, thus have a few outliers in their rightmost outskirts. However, the stem-and-leaf plot for X22 looks relatively symmetrical, which means that the predictor variable values are relatively more spread out throughout the interval between the minimum and maximum percentage of population older than 64 years old. From these results, it is plausible to infer that the demographic feature such as age structure is quite universal across these 440 counties, whereas a few of the counties can stand out in terms of other features regarding the total population or income.

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

```
model1 <- data.frame(Y, X11, X12, X13)
model2 <- data.frame(Y, X21, X22, X23)

pairs(model1)
```



```
cor(model1)

##               Y       X11         X12       X13
## Y   1.00000000 0.9402486 0.07807466 0.9481106
## X11 0.94024859 1.0000000 0.17308335 0.9867476
## X12 0.07807466 0.1730834 1.00000000 0.1270743
## X13 0.94811057 0.9867476 0.12707426 1.0000000

pairs(model2)
```




```
cor(model2)

##                Y        X21         X22         X23
## Y    1.00000000 0.40643863 -0.00312863  0.94811057
## X21  0.40643863 1.00000000  0.02918445  0.31620475
## X22 -0.00312863 0.02918445  1.00000000 -0.02273315
## X23  0.94811057 0.31620475 -0.02273315  1.00000000
```

For model 1, there exist clear positive correlations between the response variable, which is the number of active physicians in each county, and two predictor variables, X11 - the total population - and X13 - total personal income. Both the scatter plot matrix and correlation matrix show that it is hard to conclude a presence of any relationship between Y and X12, the land area. Furthermore, there also exists a positive correlation between two predictor variables, X11 and X13.

For model 2, only X23, the total personal income, has a clear, positive relationship with Y. The correlation matrix reveals that X21, the population density, is positively correlated with Y as well, but the value of the corresponding correlation coefficient, which is 0.40643863, and the scatter plot of Y against X22 show that the correlation is not as strong as the one between Y and X23. X22, the percent of population older than 64 years old, does not seem to be related with Y at all, and there also exists a positive correlation between X21 and X23 to some extent.

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

```
fit1 <- lm(Y ~ X11 + X12 + X13, data = model1)
fit1

##
## Call:
## lm(formula = Y ~ X11 + X12 + X13, data = model1)
##
## Coefficients:
## (Intercept)          X11          X12          X13
##  -1.332e+01    8.366e-04   -6.552e-02    9.413e-02

fit2 <- lm(Y ~ X21 + X22 + X23, data = model2)
fit2

##
## Call:
## lm(formula = Y ~ X21 + X22 + X23, data = model2)
##
## Coefficients:
## (Intercept)          X21          X22          X23
##  -170.57422      0.09616      6.33984      0.12657
```

Model 1.

$$Y_i = -13.32 + 0.0008366X_{i11} - 0.06552X_{i12} + 0.09413X_{i13} + \epsilon_i,\ i = 1, \cdots, 440$$

Model 2.

$$Y_i = -170.57422 + 0.09616X_{i21} + 6.33984X_{i22} + 0.12657X_{i23} + \epsilon_i,\ i = 1, \cdots, 440$$

d. Calculate R^2 for each model. Is one model clearly preferable in terms of this measure?

```
summary(fit1)

##
## Call:
## lm(formula = Y ~ X11 + X12 + X13, data = model1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1855.6  -215.2   -74.6    79.0  3689.0
##
```

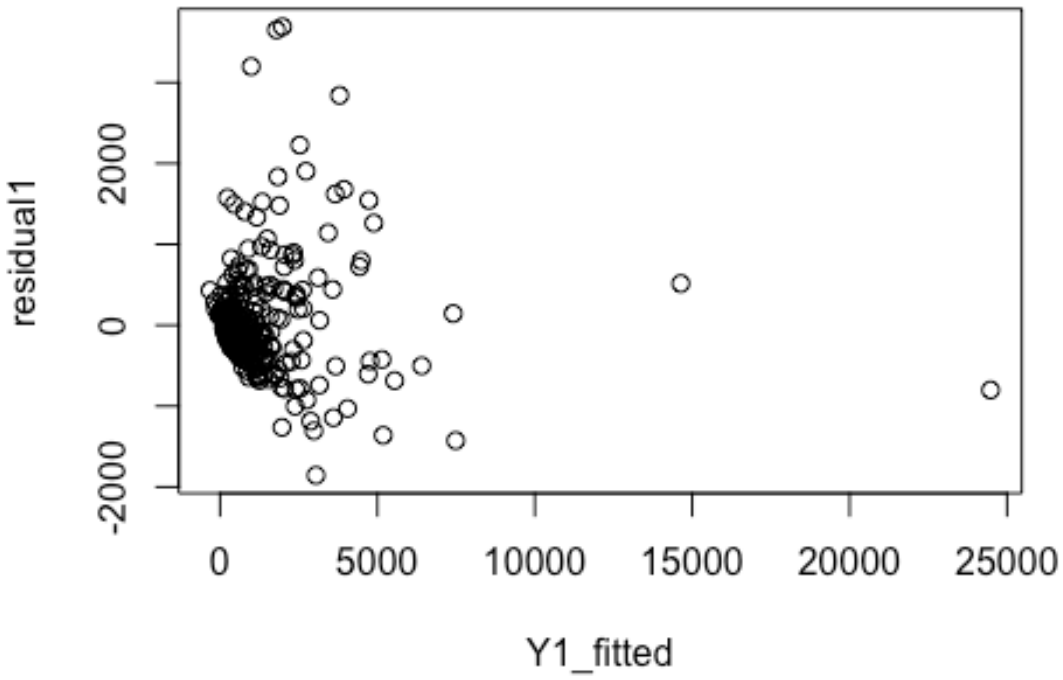
```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.332e+01  3.537e+01  -0.377 0.706719
## X11          8.366e-04  2.867e-04   2.918 0.003701 **
## X12         -6.552e-02  1.821e-02  -3.597 0.000358 ***
## X13          9.413e-02  1.330e-02   7.078 5.89e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 560.4 on 436 degrees of freedom
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.902
## F-statistic:  1347 on 3 and 436 DF,  p-value: < 2.2e-16

summary(fit2)

##
## Call:
## lm(formula = Y ~ X21 + X22 + X23, data = model2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3055.75  -175.30   -38.05    72.88  3045.81
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.706e+02  8.353e+01  -2.042   0.0418 *
## X21          9.616e-02  1.224e-02   7.857  3.1e-14 ***
## X22          6.340e+00  6.384e+00   0.993   0.3212
## X23          1.266e-01  2.084e-03  60.723  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 533.5 on 436 degrees of freedom
## Multiple R-squared:  0.9117, Adjusted R-squared:  0.9111
## F-statistic:  1501 on 3 and 436 DF,  p-value: < 2.2e-16
```

The multiple R-squared for Model 1 obtained by R is 0.9026, and R-squared for Model 2 is 0.9117. Since Model 2 yields a greater coefficient of determination, it seems more preferable than Model 1.

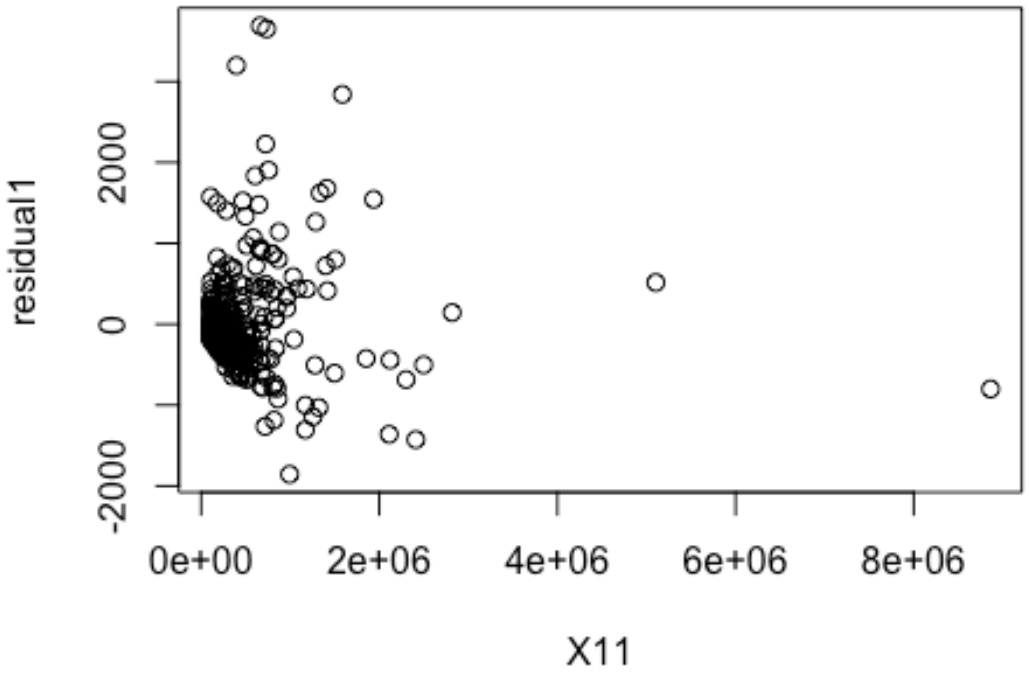
e. For each model, obtain the residuals and plot them against Y_hat, each of the three predictor variables, and each of the two-factor interaction terms. Also prepare a normal probability plot for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly preferable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

```
residual1 <- fit1$residuals
Y1_fitted <- fit1$fitted.values
X11X12 <- X11 * X12

## Warning in X11 * X12: NAs produced by integer overflow
```

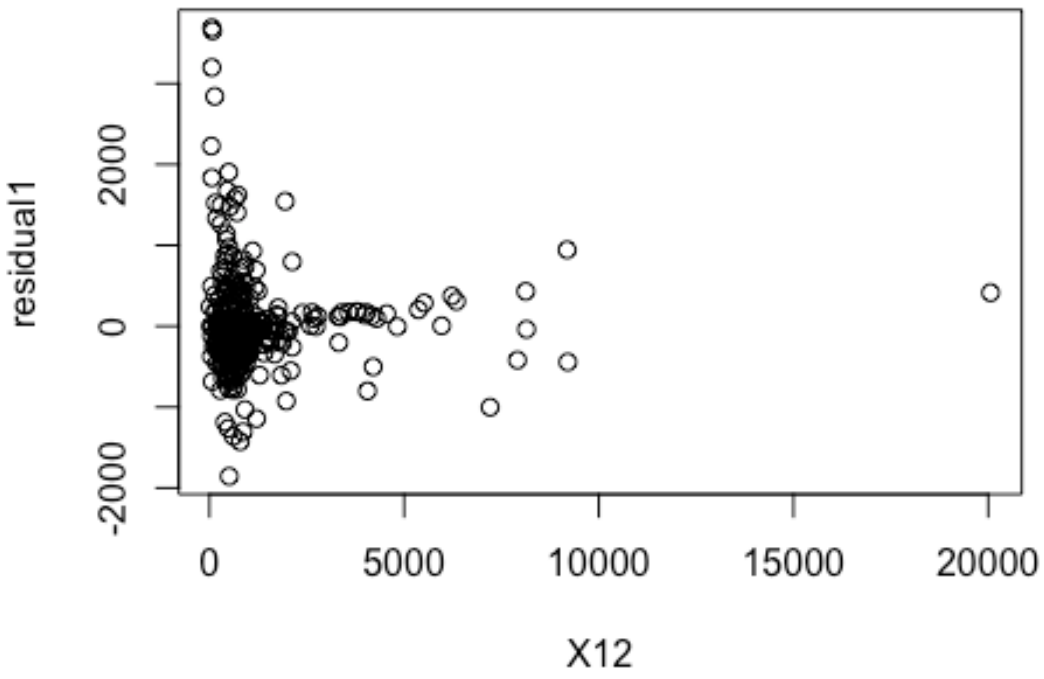
```
X11X13 <- X11 * X13

## Warning in X11 * X13: NAs produced by integer overflow

X12X13 <- X12 * X13

plot(Y1_fitted, residual1)
```



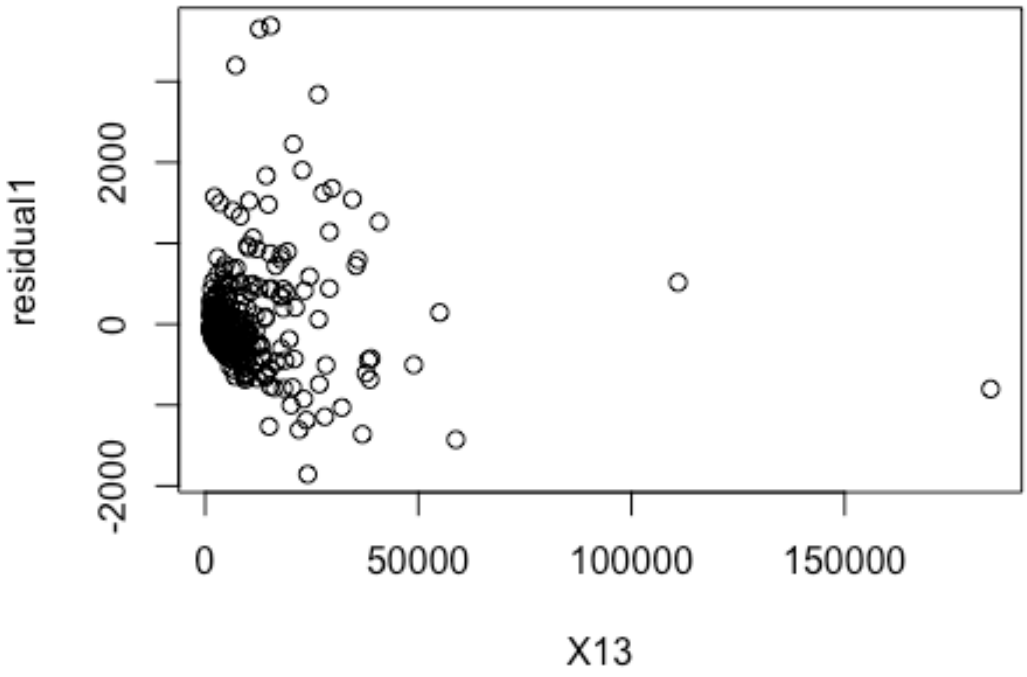

```
plot(X11, residual1)
```



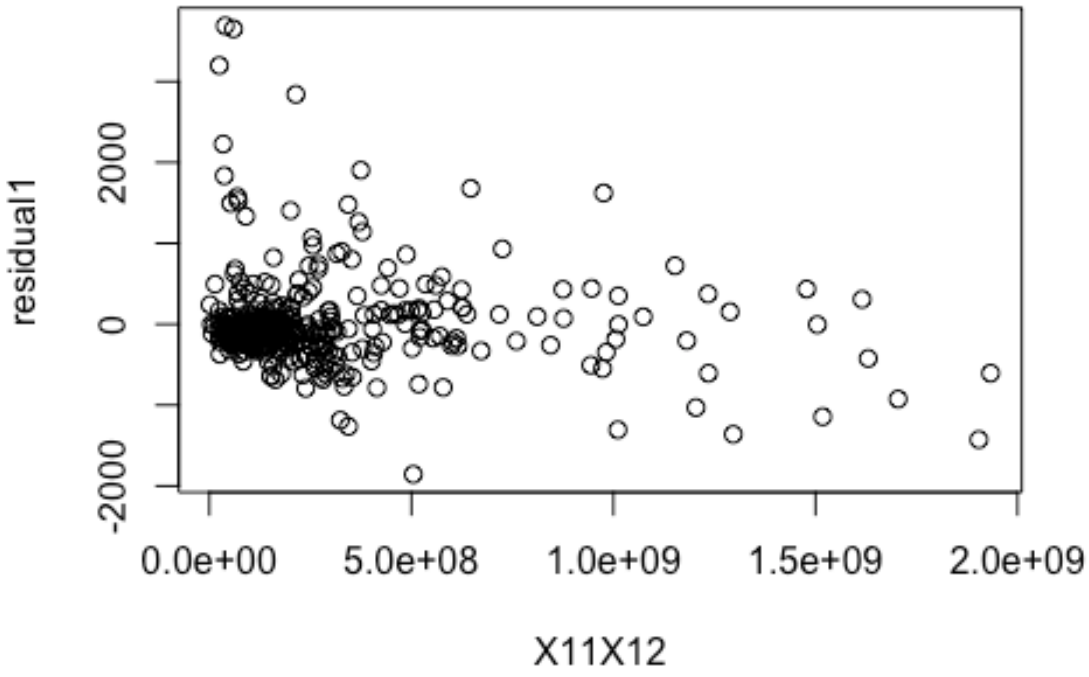

```
plot(X12, residual1)
```



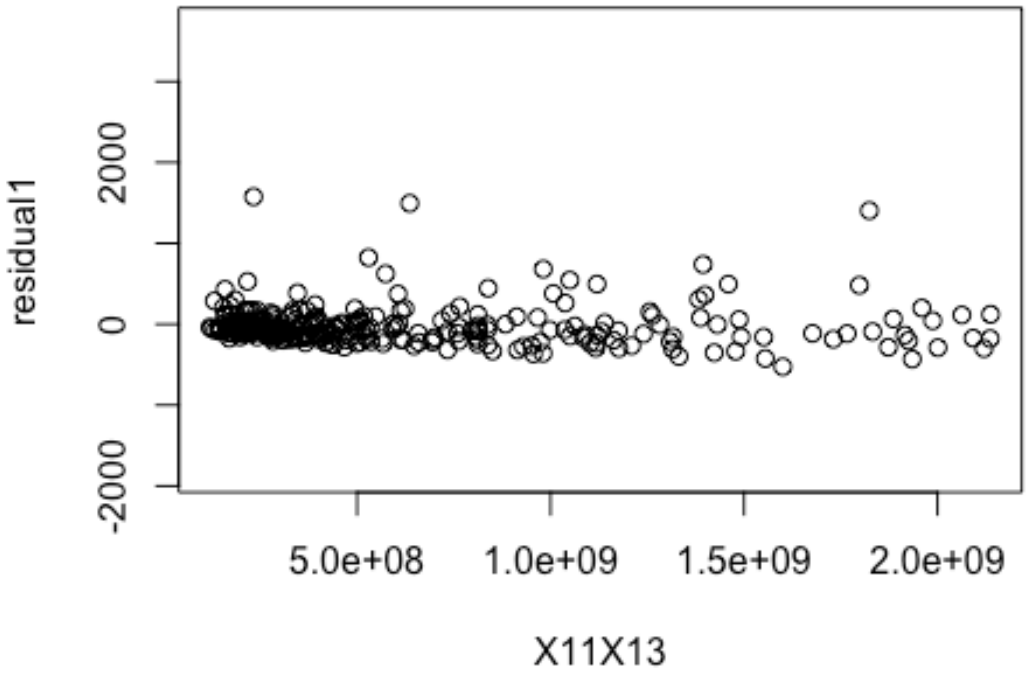

```
plot(X13, residual1)
```



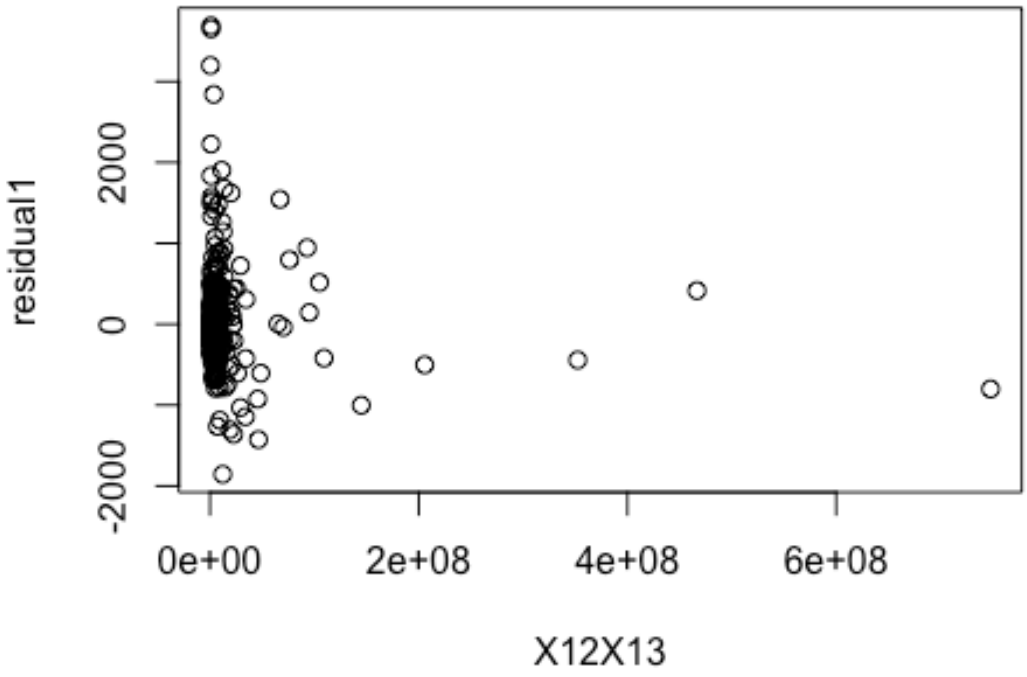

```
plot(X11X12, residual1)
```



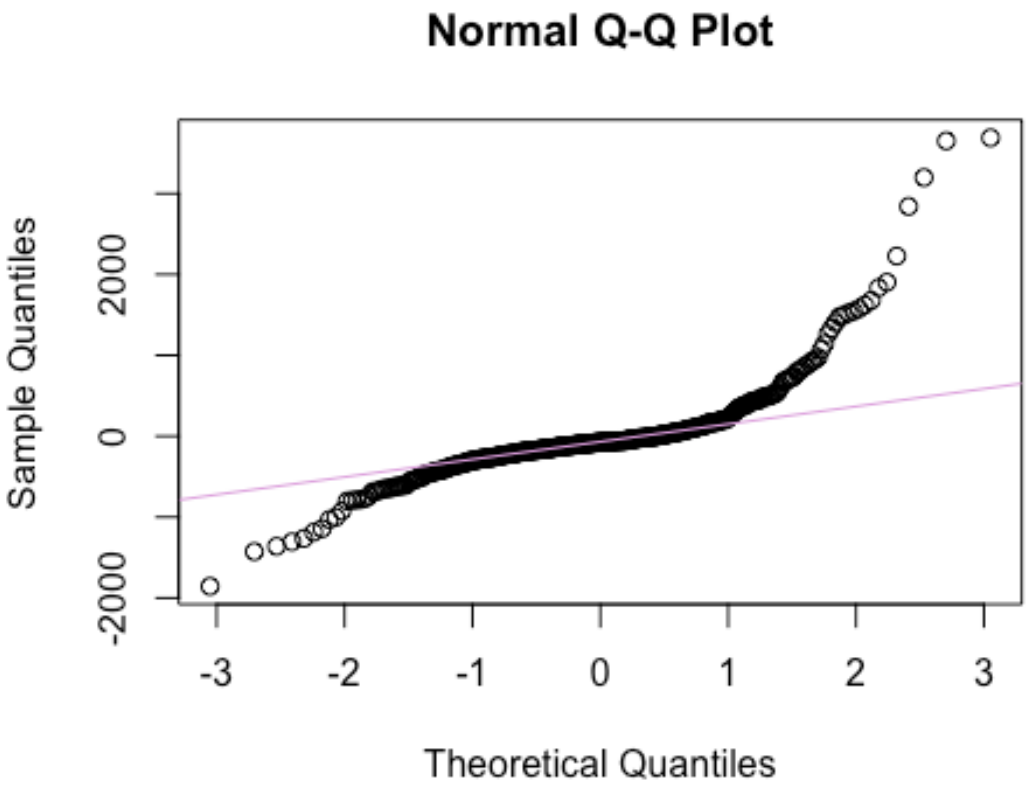

```
plot(X11X13, residual1)
```




```
plot(X12X13, residual1)
```

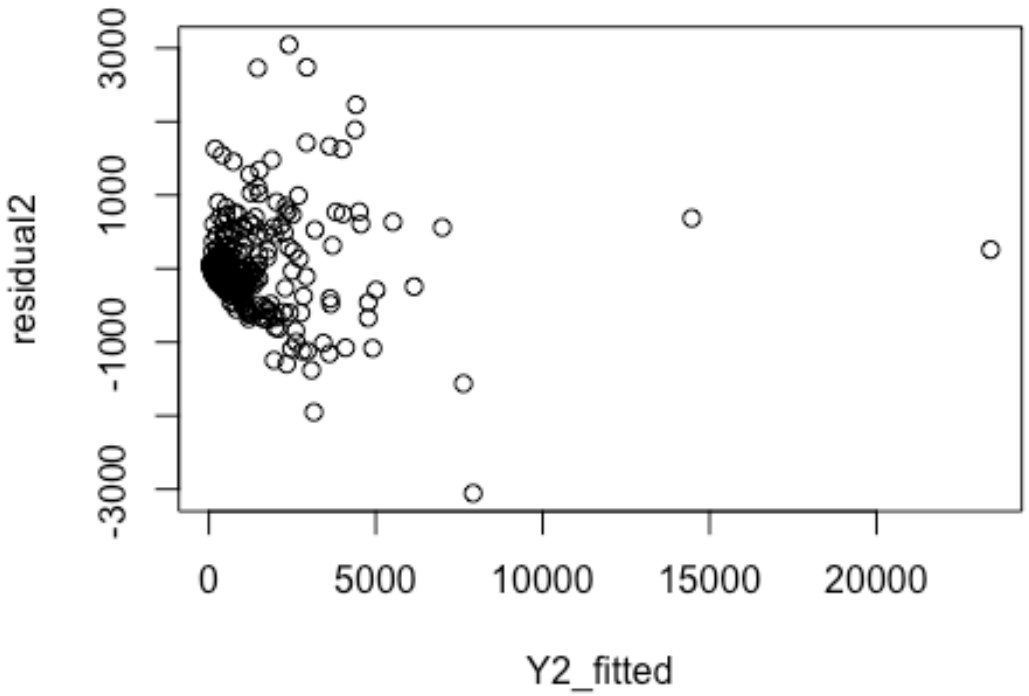



```
qqnorm(residual1)
qqline(residual1, col = "plum")
```

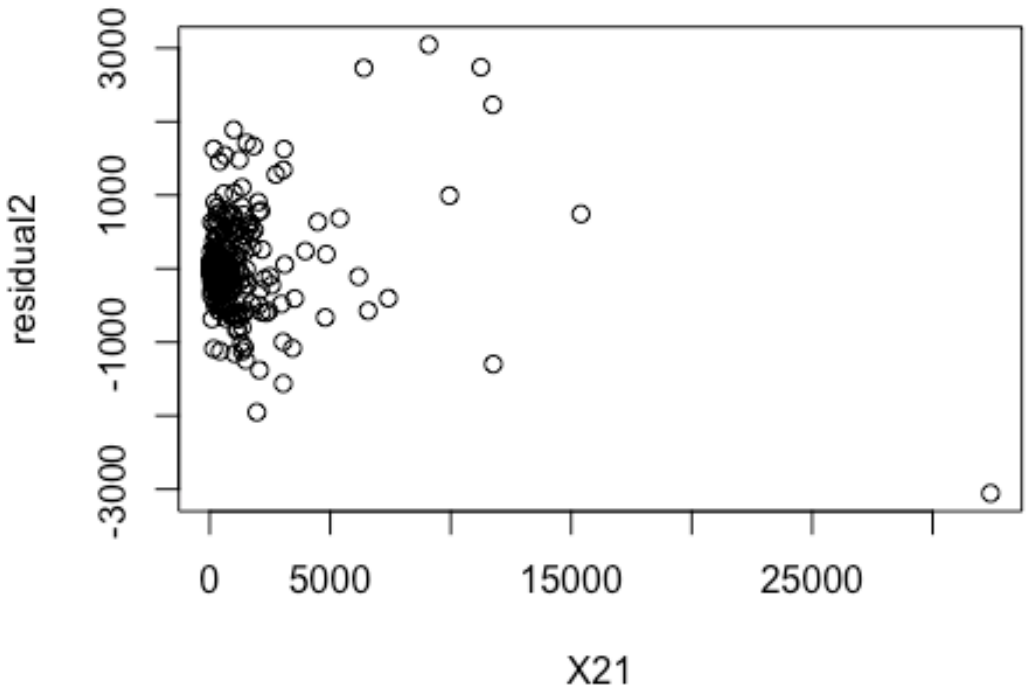


Model 1. Taking into account the contextual discovery made from #6.28(a), the general shape of all seven residual plots is understandable. The residual points in the rightmost region of each residual plot represent the outliers of each predictor variable in Model 1. Beside that, all residual plots except for the one against X11X13 hint at the violation of equal-variance assumption. This is generally because of the outliers in each predictor variable values, yet the variance of residuals seems to decrease as the value on the X-axis increases. However, all residual plots display that the residuals are distributed mostly symmetric around 0, which confirms the linearity assumption. Furthermore, there does not seem to exist any clear interaction effects as the residual plots against the interaction terms show no systematic pattern. The normal probability plot for Model 1 also raises the possibility of the normality assumption being violated since it has heavy tails. Although most data lie on the straight reference line, the presence of heavy tails at both ends suggest that the distribution of the residuals has higher probabilities in the tails than a normal distribution.

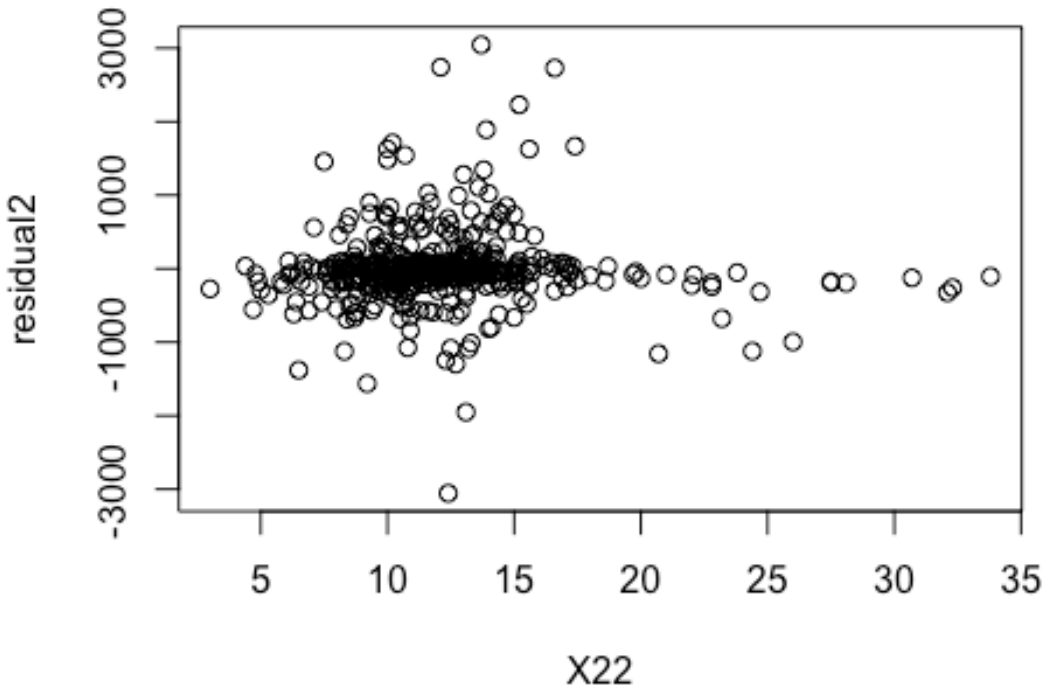
```
residual2 <- fit2$residuals
Y2_fitted <- fit2$fitted.values
X21X22 <- X21 * X22
X21X23 <- X21 * X23
X22X23 <- X22 * X23

plot(Y2_fitted, residual2)
```



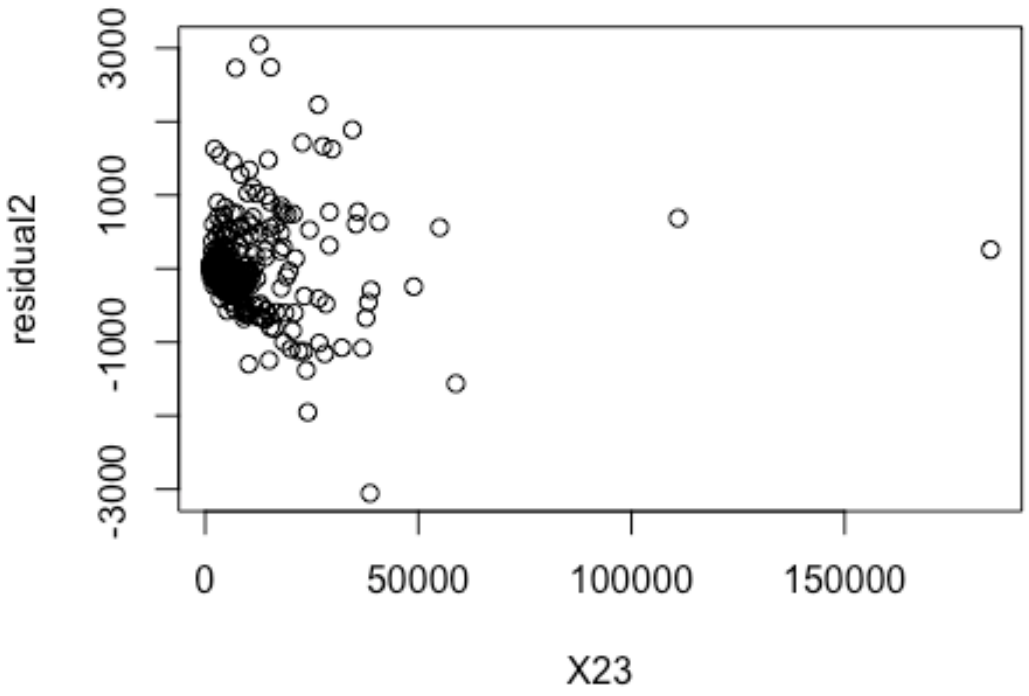

```
plot(X21, residual2)
```



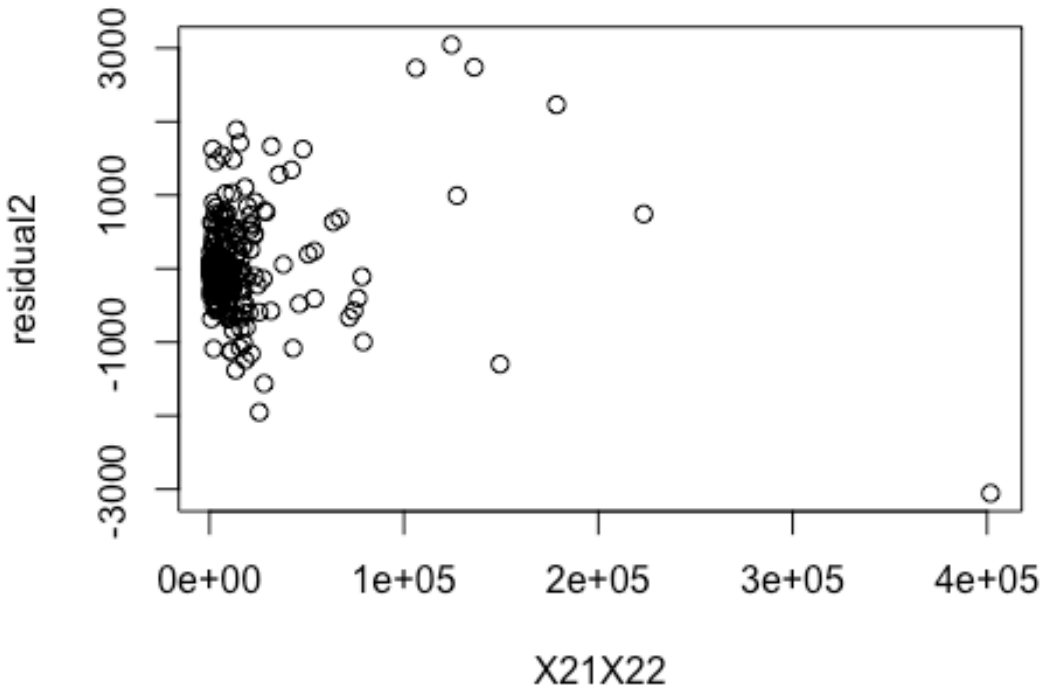

```
plot(X22, residual2)
```



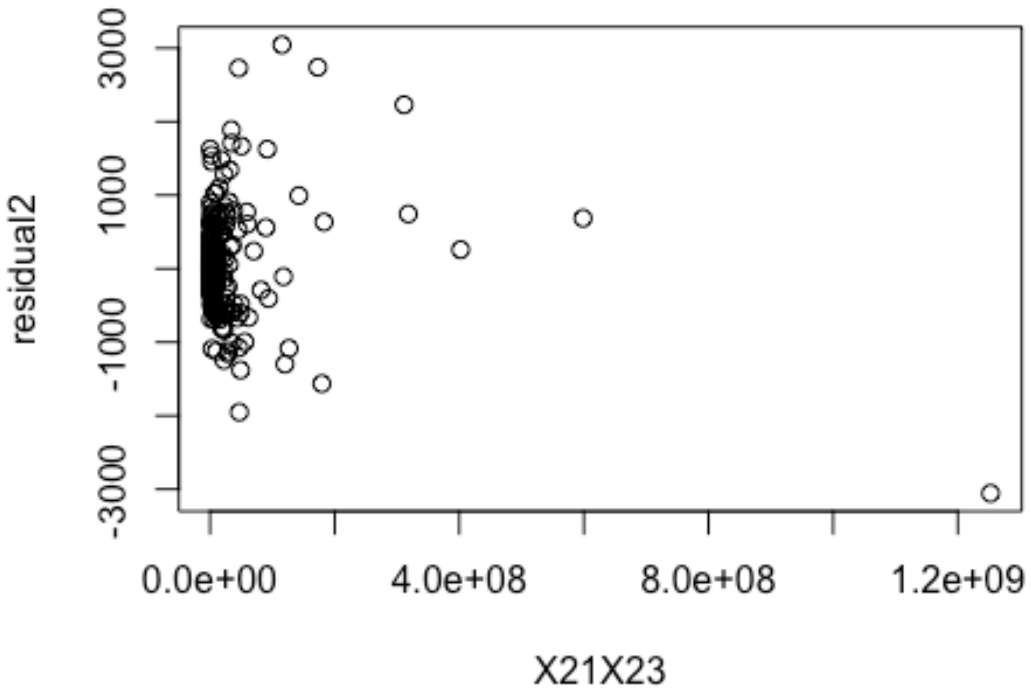

```
plot(X23, residual2)
```



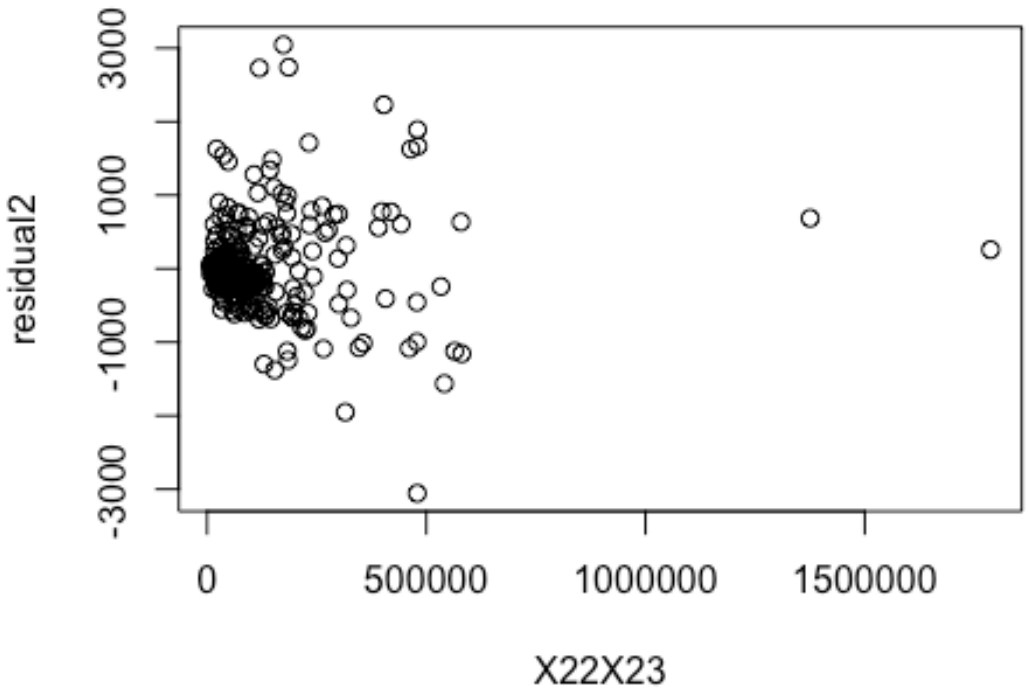

```
plot(X21X22, residual2)
```



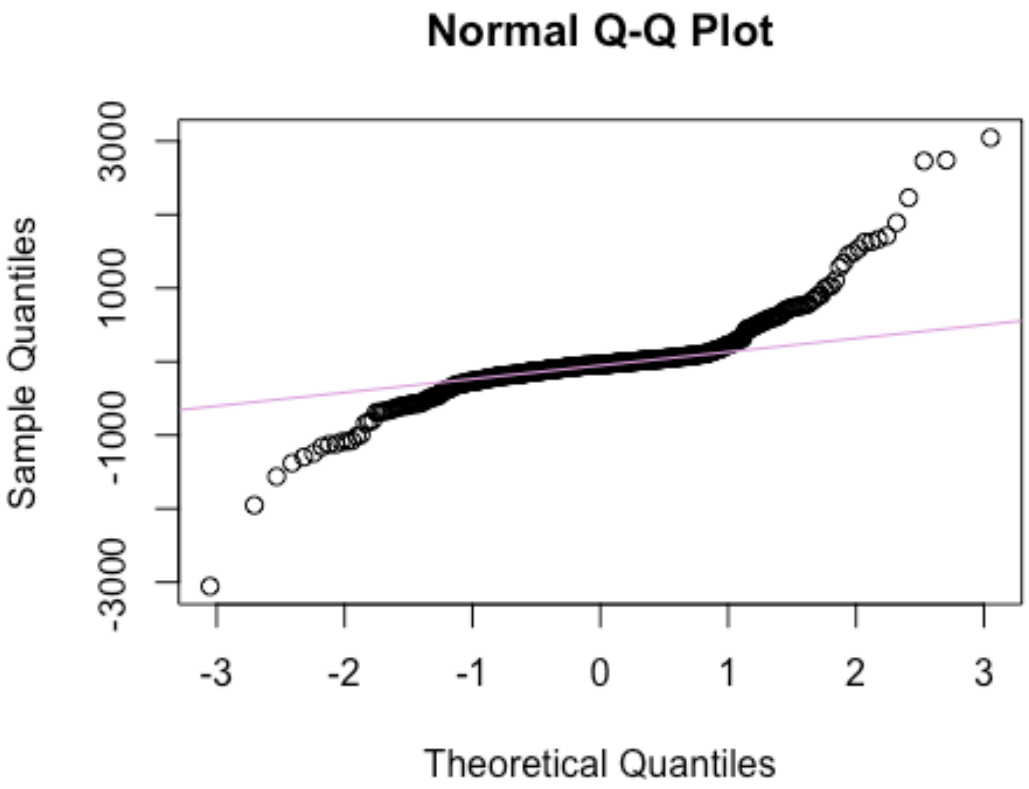

```
plot(X21X23, residual2)
```




```
plot(X22X23, residual2)
```




```
qqnorm(residual2)
qqline(residual2, col = "plum")
```



Model 2. The seven residual plots for Model 2 can be interpreted in the same way as the residual plots for Model 1. All of them suggest the violation of the equal-variance assumption as the variance of residuals seems to decrease as the value on the X-axis increases. There is only one exception to this, which is the residual plot against X22. Overall, all residual plots are symmetric about 0, which again confirms the linearity assumption. Furthermore, there does not seem to exist any clear interaction effects as the residual plots against the interaction terms show no systematic pattern. The normal probability plot for Model 2 also raises the same concern as in Model 1 that the normality assumption might have been violated. The heavy tails suggest that the distribution of the residuals might have higher probabilities in the tails than a normal distribution.

It is hard to tell which model would be better in terms of appropriateness by the residual plots and normal probability plots.

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.

```
fit1_1 <- lm(Y ~ X11+ X12 + X13 + X11X12 + X11X13 + X12X13)
fit1_1

##
## Call:
```

```
## lm(formula = Y ~ X11 + X12 + X13 + X11X12 + X11X13 + X12X13)
##
## Coefficients:
## (Intercept)          X11          X12          X13       X11X12
X11X13
##  -7.035e+01    5.109e-04    1.725e-02    9.797e-02   -1.604e-07    3.954e-
08
##      X12X13
##  -2.169e-07
```

Model 1.

$$Y_i = -70.35 + 0.0005109X_{i11} + 0.01725X_{i12} + 0.0979X_{i13} - 0.0000001604X_{i11}X_{i12} + 0.00000003954X_{i11}X_{i13} - 0.0000002169X_{i12}X_{i13} + \epsilon_i,\ i = 1, \cdots, 440$$

```
summary(fit1_1)

##
## Call:
## lm(formula = Y ~ X11 + X12 + X13 + X11X12 + X11X13 + X12X13)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -486.09 -112.23  -40.54   43.31 1606.99
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.035e+01  1.247e+02  -0.564   0.5730
## X11          5.109e-04  8.812e-04   0.580   0.5625
## X12          1.725e-02  3.906e-02   0.442   0.6591
## X13          9.797e-02  4.607e-02   2.127   0.0343 *
## X11X12      -1.604e-07  5.145e-07  -0.312   0.7555
## X11X13       3.954e-08  1.728e-07   0.229   0.8191
## X12X13      -2.169e-07  2.417e-05  -0.009   0.9928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 238.9 on 288 degrees of freedom
##   (145 observations deleted due to missingness)
## Multiple R-squared:  0.3572, Adjusted R-squared:  0.3438
## F-statistic: 26.67 on 6 and 288 DF,  p-value: < 2.2e-16
```

Multiple R_squared = 0.3572.

```
fit2_1 <- lm(Y ~ X21 + X22 + X23 + X21X22 + X21X23 + X22X23)
fit2_1

##
## Call:
## lm(formula = Y ~ X21 + X22 + X23 + X21X22 + X21X23 + X22X23)
```

```
##
## Coefficients:
## (Intercept)          X21          X22          X23       X21X22
X21X23
##  -9.367e+00   -4.179e-01   -1.106e+01    1.477e-01    4.652e-02   -3.276e-
06
##      X22X23
##  -1.289e-03
```

Model 2.

$$Y_i = -9.367 - 0.4179X_{i21} - 0.1106X_{i22} + 0.1477X_{i23} + 0.04652X_{i21}X_{i22} - 0.000003276X_{i21}X_{i23} - 0.001289X_{i22}X_{i23} + \epsilon_i,\ i = 1, \cdots, 440$$

```
summary(fit2_1)

##
## Call:
## lm(formula = Y ~ X21 + X22 + X23 + X21X22 + X21X23 + X22X23)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2409.57  -163.91   -12.32   103.25  2721.84
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.367e+00  9.928e+01  -0.094    0.925
## X21         -4.179e-01  1.055e-01  -3.960 8.76e-05 ***
## X22         -1.106e+01  7.792e+00  -1.419    0.157
## X23          1.477e-01  9.739e-03  15.168  < 2e-16 ***
## X21X22       4.652e-02  7.925e-03   5.870 8.67e-09 ***
## X21X23      -3.276e-06  7.439e-07  -4.404 1.34e-05 ***
## X22X23      -1.289e-03  8.743e-04  -1.474    0.141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 500 on 433 degrees of freedom
## Multiple R-squared:  0.923,  Adjusted R-squared:  0.922
## F-statistic: 865.4 on 6 and 433 DF,  p-value: < 2.2e-16
```

Multiple R-squared = 0.923.

Since Model 2 with all interaction factors included yields much greater coefficient of determination than Model 1 with all of its interaction factors included, model 2 is more preferable.

Part II: Mutiple linear regression II. This part consists of Project 7.37 in the book, with the following changes.

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), and number of hospital beds(X5).

```
X1 <- CDI$`Total population`
X2 <- CDI$`total person income`
X3 <- CDI$`Land area`
X4 <- CDI$`Percent of population 65 or older`
X5 <- CDI$`Number of hospital beds`

fit737_1 <- lm(Y ~ X1 + X2)
anova(fit737_1)

## 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

fit737_2 <- lm(Y ~ X1 + X2 + X3)
anova(fit737_2)

## 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 ***
## X3          1    4063370    4063370   12.941 0.0003583 ***
## Residuals 436  136903711     313999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

fit737_3 <- lm(Y ~ X1 + X2 + X4)
anova(fit737_3)

## 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
```

```
fit737_4 <- lm(Y ~ X1 + X2 + X5)
anova(fit737_4)
```

```
## 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
```

$$R^2_{3|1,2} = \frac{SSR(X_3|X_1,X_2)}{SSE(X_1,X_2)} = \frac{4063370}{140967081} = 0.02882496$$

$$R^2_{4|1,2} = \frac{SSR(X_4|X_1,X_2)}{SSE(X_1,X_2)} = \frac{541647}{140967081} = 0.003842365$$

$$R^2_{5|1,2} = \frac{SSR(X_5|X_1,X_2)}{SSE(X_1,X_2)} = \frac{78070132}{140967081} = 0.5538182$$

b. On the basis of the results in part (a), which of the three additional predictor variables is best? Is the extra sum of squares associated with this variable larger than those for the other two variables?

X5, the number of hospital beds is the best additional predictor variable since it yields the greatest coefficient of partial determination. Yes, the extra sum of squares associated with X5, which is $\frac{SSR(X_5|X_1,X_2)}{SSE(X_1,X_2)}$ is greater than $\frac{SSR(X_3|X_1,X_2)}{SSE(X_1,X_2)}$ and $\frac{SSR(X_4|X_1,X_2)}{SSE(X_1,X_2)}$ as shown in (a).

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 $\alpha = .01$. State the alternatives, decision rule, and conclusion. Would the F^* test statistics for the other two potential predictor variables be as large as the one here? Discuss.

```
anova(fit737_4)
```

```
## 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
```

```
pf(541.18, 1, 436, lower.tail=FALSE)
```

```
## [1] 2.009405e-78
```

$$H_0: (Reduced\ model)\ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$H_1: (Full\ model)\ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_5$$

Level of significance = 0.01

Decision rule = If the p-value is less than the level of the significance of 0.01, H0 is rejected in favor of H1.

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} / \frac{SSE(F)}{df_F} = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_5)}{437 - 436} / \frac{SSE(X_1, X_2, X_5)}{436}$$
$$= \frac{MSR(X_5|X_1, X_2)}{MSE(X_1, X_2, X_5)} = \frac{78070132}{144259} = 541.1803$$

$$p - value = 2.009405(10)^{-78}$$

Since the p-value is less than the level of significance of 0.01, H0 is rejected in favour of H1. In conclusion, the number of hospital beds(X5) is helpful in the regression model when total population(X1) and total personal income(X2) are already included in the model.

F^* test statistic for the other two variables, X3 and X4, will not be as large as F^* for X5. Thus, the corresponding p-values for those variables are likely to be much larger than the level of significance, leading us to accept the null hypothesis that X3 and X4 are not important predictor variables for the regression model which already includes X1 and X2.

d. Compute three additional coefficients of partial determination: $R^2_{Y,X3,X4|X1,X2}$, $R^2_{Y,X3,X5|X1,X2}$, and $R^2_{Y,X4,X5|X1,X2}$. 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.

```
fit737_5 <- lm(Y ~ X1 + X2 + X3 + X4)
anova(fit737_5)
```

```
## Analysis of Variance Table
##
## Response: Y
##             Df     Sum Sq    Mean Sq   F value    Pr(>F)
## X1           1 1243181164 1243181164 3967.7399 < 2.2e-16 ***
## X2           1   22058054   22058054   70.4005 6.842e-16 ***
```

```
## X3          1    4063370    4063370   12.9687 0.0003533 ***
## X4          1     608535     608535    1.9422 0.1641413
## Residuals 435  136295177     313322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit737_6 <- lm(Y ~ X1 + X2 + X3 + X5)
anova(fit737_6)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq  F value    Pr(>F)
## X1          1 1243181164 1243181164 8636.745 < 2.2e-16 ***
## X2          1   22058054   22058054  153.244 < 2.2e-16 ***
## X3          1    4063370    4063370   28.229 1.724e-07 ***
## X5          1   74289406   74289406  516.110 < 2.2e-16 ***
## Residuals 435   62614306     143941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit737_7 <- lm(Y ~ X1 + X2 + X4 + X5)
anova(fit737_7)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq  F value Pr(>F)
## X1          1 1243181164 1243181164 8804.285 <2e-16 ***
## X2          1   22058054   22058054  156.216 <2e-16 ***
## X4          1     541647     541647    3.836 0.0508 .
## X5          1   79002640   79002640  559.502 <2e-16 ***
## Residuals 435   61422794     141202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
39772144/141202
```

```
## [1] 281.6684
```

$$R^2_{Y,X3,X4|X1,X2} = \frac{SSR(X_3, X_4|X_1, X_2)}{SSE(X_1, X_2)} = \frac{SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2)}{SSE(X_1, X_2)}$$
$$= \frac{1269911123 - 1265239218}{140967081} = \frac{4671905}{140967081} = 0.03314182$$

$$R^2_{Y,X3,X5|X1,X2} = \frac{SSR(X_3, X_5|X_1, X_2)}{SSE(X_1, X_2)} = \frac{SSR(X_1, X_2, X_3, X_5) - SSR(X_1, X_2)}{SSE(X_1, X_2)}$$
$$= \frac{1343591994 - 1265239218}{140967081} = \frac{78352776}{140967081} = 0.5558232$$

$$R^2_{Y,X4,X5|X1,X2} = \frac{SSR(X_4, X_5|X_1, X_2)}{SSE(X_1, X_2)} = \frac{SSR(X_1, X_2, X_4, X_5) - SSR(X_1, X_2)}{SSE(X_1, X_2)}$$
$$= \frac{1344783505 - 1265239218}{140967081} = \frac{79544287}{140967081} = 0.5642756$$

Since $R^2_{Y,X4,X5|X1,X2}$ has the greatest value, (X4, X5) is relatively more important than other pairs of variables.

Let's conduct the F-test to check if adding the pair of X4 and X5 is helpful for the regression model that already has X1 and X2 as predictor variables:

$$H_0: (Reduced\ model)\ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$H_1: (Full\ model)\ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 X_5$$

Level of significance = 0.05

Decision rule = If the p-value is less than the level of the significance of 0.05, H0 is rejected in favour of H1.

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} / \frac{SSE(F)}{df_F} = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_4, X_5)}{437 - 435} / \frac{SSE(X_1, X_2, X_4, X_5)}{435}$$
$$= \frac{MSR(X_4, X_5|X_1, X_2)}{MSE(X_1, X_2, X_4, X_5)} = \frac{39772144}{141202} = 281.6684$$

```
pf(281.6684, 2, 435, lower.tail = FALSE)

## [1] 3.377913e-79
```

$$p - value = 3.377913(10)^{-79}$$

Since the p-value is less than the level of significance of 0.01, H0 is rejected in favour of H1. In conclusion, adding percent of population 65 or older(X4) and the number of hospital beds(X5) as predictor variables is helpful in the regression model when total population(X1) and total personal income(X2) are already included in the model.

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?

Multiple linear regression is a culmination of all topics covered by this course. Without proper understanding of the simple linear regression model, it is impossible to understand the complexities of a multiple linear regression model. In this particular project, we've used various applications we've learned over the course of the quarter, including F-tests with the reduced/full model, interpretation of the ANOVA table, and the formulaic derivation of the extra SS. It is clear to see that land area has little to do with the number of active physicians in a county. The more relevant predictors include total population, total personal income and number of hospital beds. If we were to have to choose a model between model 1 and model 2, model 1 seems more favorable, but it also seems possible to build a better one. If our objective was to build a model to predict the number of active

physicians, surely the most robust model would be one build on the three aforementioned variables.