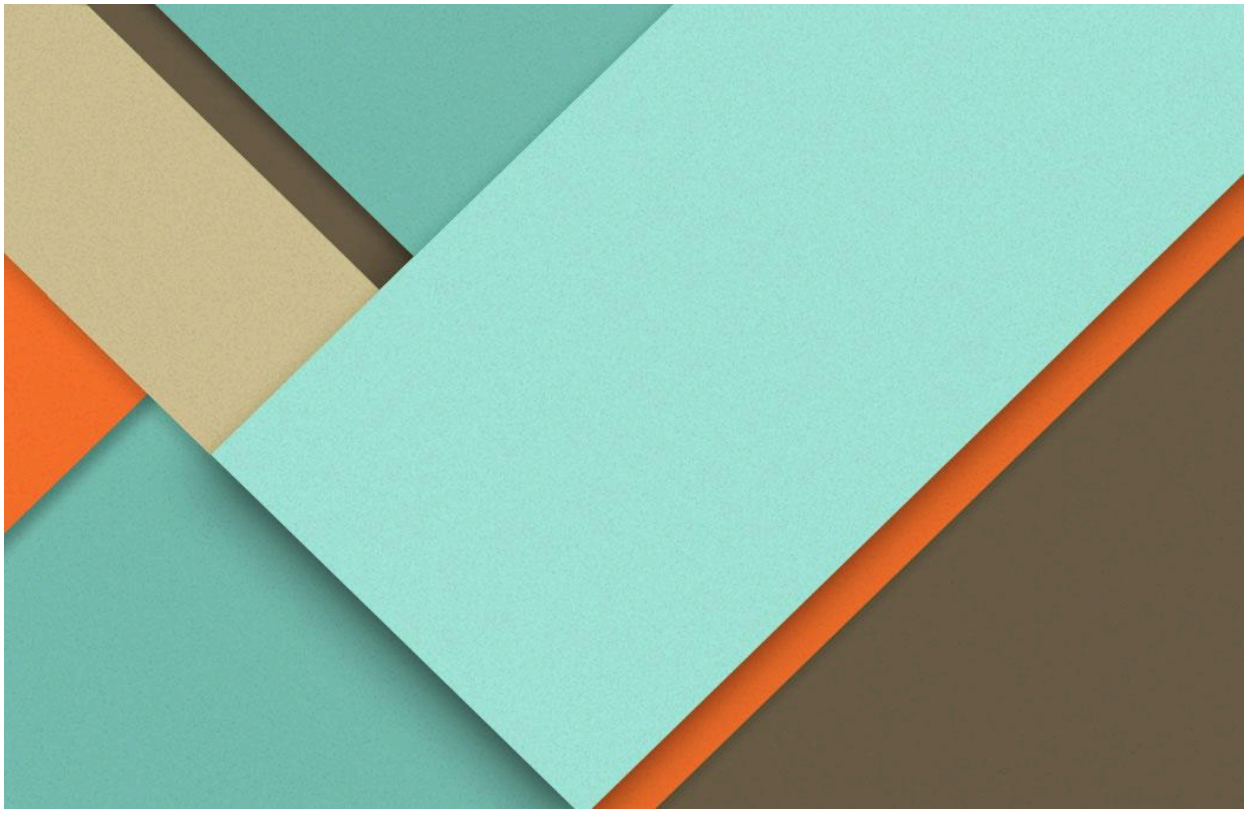


County Demographic Info: Case Study II

03.09.2020

—

Radhika Kulkarni & Sameerah Helal
STA 108 A01
Professor Jiming Jiang

Introduction

In the previous project we determined the relationship between the number of active physicians for the 440 most populous counties' demographics in the United States and the predictor variables of total population, number of hospital beds, and the total personal income. We also determined whether the per capita income has any relationship with the percentage of the population with bachelor's degrees in the counties.

In this project, with the same dataset, we would like to determine :
Part I
For two proposed models:

1) Model 1: the relationships, if any, between three predictor variables of total population (X_1), land area (X_2), and total personal income (X_3) and the number of active physicians in a US county (Y) , and
2) Model 2: the relationships, if any, between three predictor variables of population density (X_1), percent of population greater than 64 years (X_2), and total personal income (X_3) and the number of active physicians in a US county (Y).

The variables in 1) are the total population, land area, and the total personal income. We hypothesized that each of these variables would have no relationship with the number of physicians in a county. We will determine whether or not this statement is true for each of the predictor variables.

We plan to test our hypotheses by comparing stem and leaf plots, regression models, residual plots, and normal probability plots. We will also use R squared calculations to determine how much of the variability in the number of active physicians in the county is explained by the regression functions involving, respectively, total population, number of hospital beds, and total personal income. Additionally we shall do this for two factor interactions of X_1X_2, X_1X_3, and X_2X_3. We realize, however, that we cannot definitively prove whether or not our hypothesis is true because we will not have performed the standard method of hypothesis testing to use as sufficient evidence to support or reject our hypothesis.

The variables in 2) are population density, percent of population greater than 64 years, and the total personal income. We hypothesized that each of these variables would have no relationship with the number of physicians in a county. We will determine whether or not this statement is true for each of the predictor variables.

To test our hypothesis we will compare stem and leaf plots, regression function data, residual plots, and normal probability plots. We will also use R squared calculations to determine how much of the variability in the number of active physicians in the county is explained by the regression functions involving, respectively, population density, percent of population greater than 64 years, and total personal income. Additionally we shall do this

for two factor interactions of X_1X_2, X_1X_3, and X_2X_3. Again, we cannot definitively prove whether or not our hypothesis is true because we will not have performed the standard method of hypothesis testing to use as sufficient evidence to support or reject our hypothesis.

We will use this to determine whether or not there is a relationship between population density, percent of population greater than 64 years, and the total personal income and the number of active physicians.

We also hypothesize that both models are equally preferable in terms of appropriateness. We will determine whether or not these statements are true for each of the predictor variables and models, respectively.

Part II

The coefficient of partial determination for the variables of Land Area (X_3), Percent of the Population 65 or older (X_4), and number of Hospital Beds (X_5) when the variables X_1 and X_2 were already included in the model. For whichever variable yields a coefficient that indicates the highest correlation, we will conduct an F-test to see whether its coefficient in the model is really nonzero. We will do the same for every pair combination of these variables: $X_3, X_4; X_3, X_5;$ and X_4, X_5.

We predict that the variables of percent of population over 65 and the number of hospital beds will have the highest correlation, both individually and in combination. We also hypothesize that the F-tests will cause us to reject the null hypotheses in both cases.

The main tool we will use to generate our graphs and output is R-Studio. Screenshots of code and output are given at the end of the document.

Part I: Multiple Regression Part 1

6.28

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

PROPOSED MODEL 1 PLOTS:

```
> stem(totalPop)

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

  0 | 11111111111111111111111111111111111111111111111111111111111111111111111111111+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
```

The stem plot for predictor variable 1:
x1 = Total Population is heavily skewed right with most of the values clustering around 0, with at least two possible outliers.

```
> stem(landArea)

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

   0 | 000011111111111122222222222222222223333333333333333333333333333444444+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
```

The stem plot for predictor variable 2:
x2 = Land Area is skewed right with most of the values clustering around 0, with at least 1 possible outlier.

```
> stem(totalPersonalIncome)

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

   0 | 1111111111111222222222222222222222222222222222222222222222222222222222222222+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
```

The stem plot for predictor variable 3:
x3 = Total Personal Income is heavily skewed right with most of the values clustering around 0, with at least two possible outliers.

PROPOSED MODEL 2 PLOTS:

```
> stem(popDens)

  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
```

The stem plot for predictor variable 1:
x1 = Population Density is heavily skewed right with most of the values clustering around 0, with at least one possible outlier.

```
> stem(perPopG64)

  The decimal point is at the |

   2 | 0
   4 | 47890389
   6 | 1123455677990134566678899
   8 | 00112222333344445556667777788888999900022223333334444444455556666 77
  10 | 000111111222222222233333344444455555556666666677777778888888888999999+36
  12 | 000000001111122223333333334444555555566666667777777788889990000000+36
  14 | 00001111111223334444455567788900000011112222345566 7778
  16 | 12556699901122345
  18 | 06778
  20 | 070
  22 | 018828
  24 | 47
  26 | 055
  28 | 1
  30 | 7
  32 | 138
```

The stem plot for predictor variable 2:

x2 = Percent of Population Greater than 64 is approximately symmetric around x=10 excluding the right tail, with at least two possible outliers. Including the tail we can say that the stem plot is slightly right skewed with at least two possible outliers.

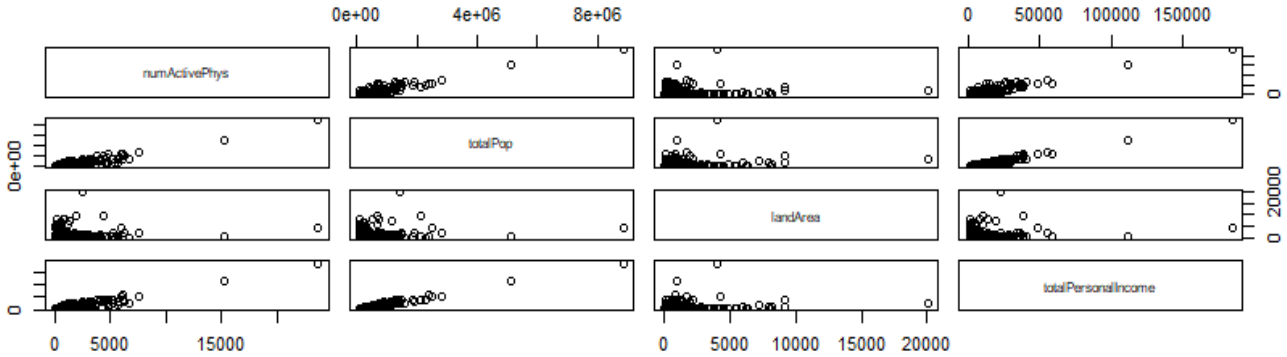
```
> stem(totalPersonalIncome)

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

   0 | 1111111111111222222222222222222222222222222222222222222222222222222222+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
```

The stem plot for predictor variable 3:
x3 = Total Personal Income is heavily skewed right with most of the values clustering around 0, with at least two possible outliers.

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

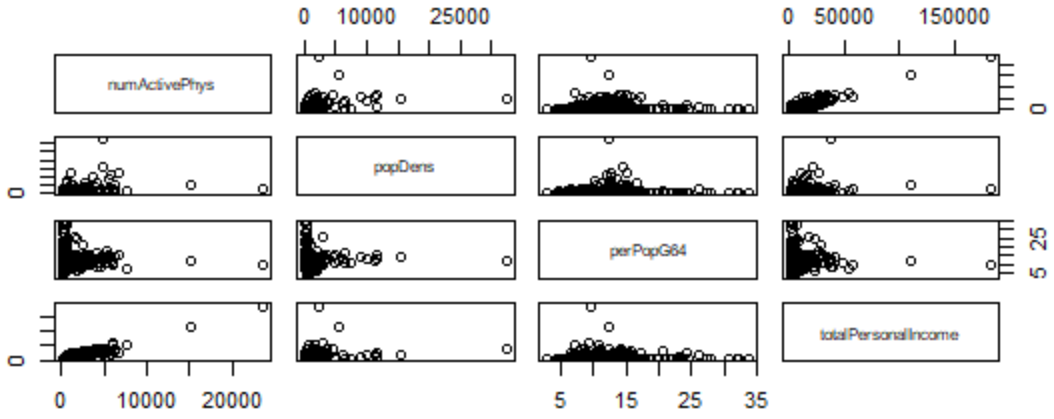
PROPOSED MODEL 1 Correlation Matrix and Scatter Plot

```
> cor(data1)
                    numActivePhys  totalPop   landArea totalPersonalIncome
numActivePhys          1.00000000 0.9402486 0.07807466           0.9481106
totalPop               0.94024859 1.0000000 0.17308335           0.9867476
landArea               0.07807466 0.1730834 1.00000000           0.1270743
totalPersonalIncome    0.94811057 0.9867476 0.12707426           1.0000000
```



For our first proposed model we can see from the scatter plot and correlation matrix the strong correlation (~0.940) between total population and number of active physicians in the population, and total personal income with number of active physicians (~0.948). Land area is very weakly correlated (~0.078) with the number of active physicians in the population. All the predictor variables are poorly correlated to each other.
PROPOSED MODEL 2 Correlation Matrix and Scatter Plot

```
> corr(data2)
                    numActivePhys     popDens   perPopG64 totalPersonalIncome
numActivePhys          1.00000000 0.40643863 -0.00312863          0.94811057
popDens                0.40643863 1.00000000  0.02918445          0.31620475
perPopG64             -0.00312863 0.02918445  1.00000000         -0.02273315
totalPersonalIncome    0.94811057 0.31620475 -0.02273315          1.00000000
>
```



For our second proposed model we can see from the scatter plot and correlation matrix the middling correlation (~0.406) between population density and number of active physicians in the population, and strong correlation between total personal income with

number of active physicians (~0.948). Land area is negatively correlated(~-.003) or even uncorrelated -due to its closeness to zero -with the number of active physicians in the population. All the predictor variables are poorly correlated or uncorrelated with each other.

c. For each proposed model, fit the first-order regression model (6.5) with three predictor variables. (see end of project for outputs)

PROPOSED MODEL 1 REGRESSION FUNCTION:

$$\hat{Y} = -13.31615 + 0.0008366178X_1 - 0.06552296X_2 + 0.09413199X_3$$

PROPOSED MODEL 2 REGRESSION FUNCTION:

$$\hat{Y} = -170.57422325 + 0.09615889X_1 + 6.33984064X_2 + 0.12656649X_3$$

d. Calculate R^2 for each model. Is one model clearly preferable in terms of this measure?(see end of project for outputs)

Model 1:

$R^2 = 0.9026$

Approximately 90% of the variation in the number of active physicians (Y) can be explained by the regression function involving total population (X_1), land area (X_2), and total personal income (X_3).

Model 2:

$R^2 = 0.9117$

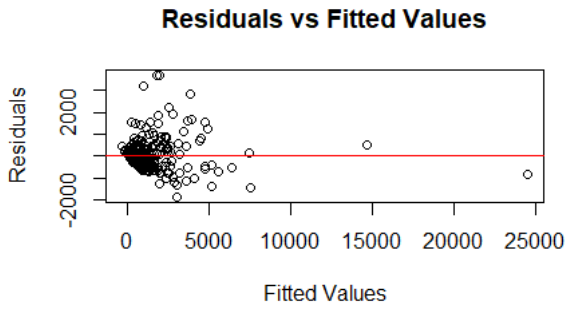
Approximately 91% of the variation in the number of active physicians (Y) can be explained by the regression function involving population density (X_1), percent of population greater than 64 years of age (X_2), and total personal income (X_3).

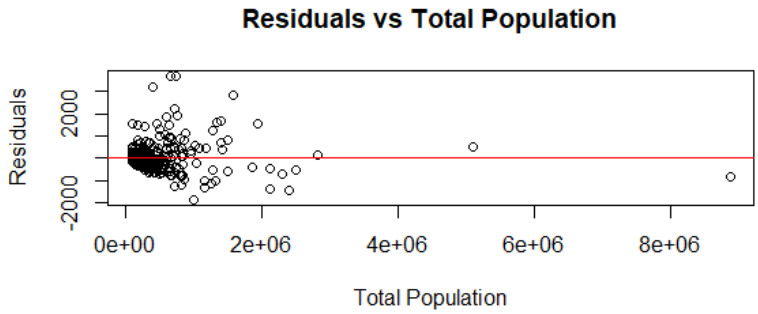
In terms of variability in number of active physicians, proposed model 2 --which involves population density, percent of population greater than 64 years of age, and total personal income-- is clearly slightly more preferable in terms of R^2, with approximately 91% explained compared to 90% explained by the predictor variables in proposed model 1.

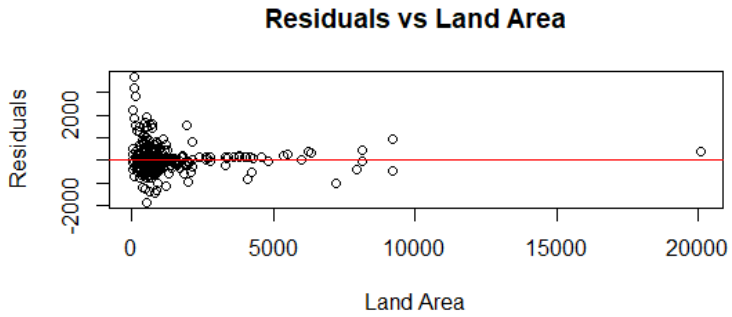
e. For each model, obtain the residuals and plot them against: $\hat{Y}$, each of the predictor variables, and each of the two factor interactive terms. Also prepare a normal probability plot for each of the two fitted models. Interpret your plots and

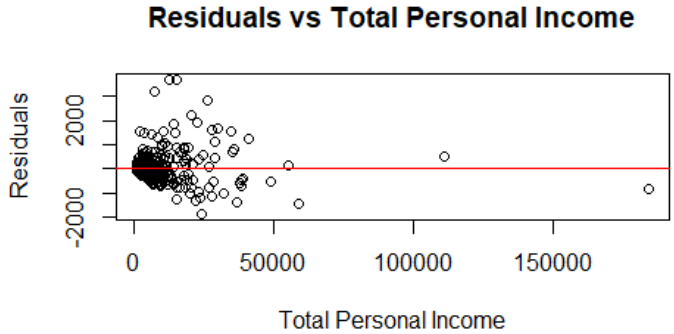
state your findings. Is one model clearly preferable in terms of appropriateness?

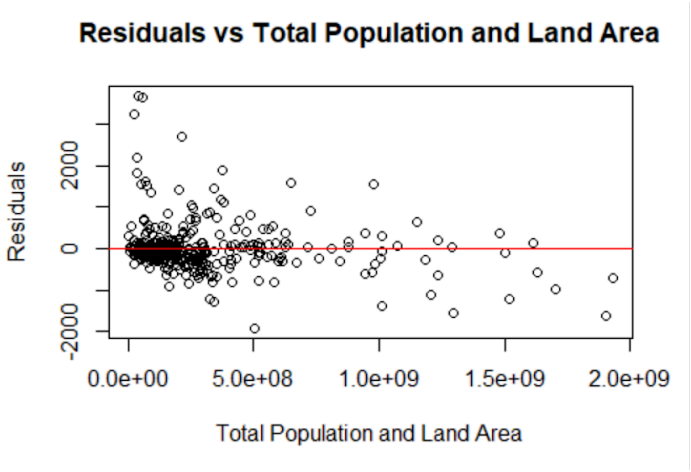
Model 1 Plots:

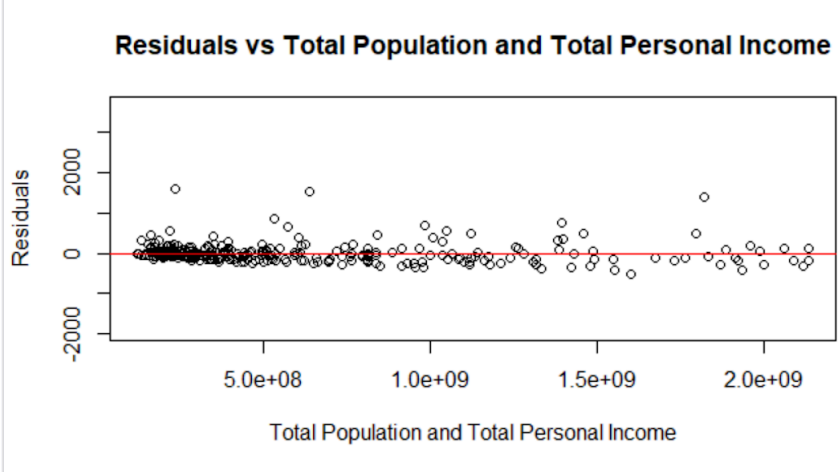


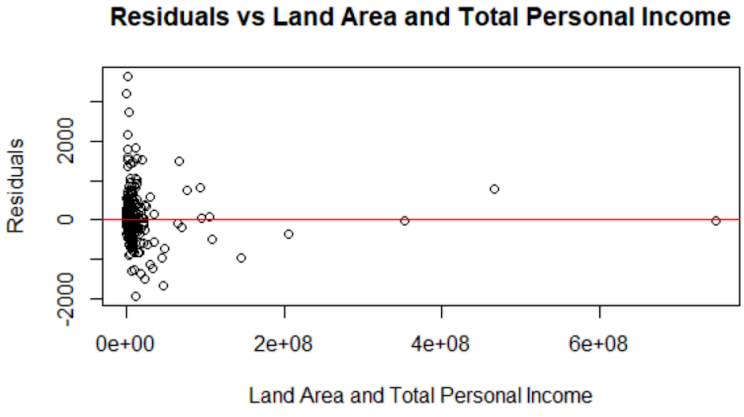


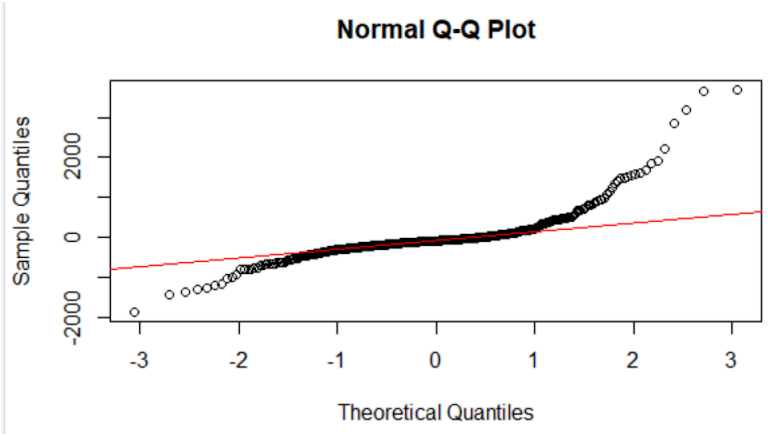





Residuals vs Total Population and Land Area
Residuals
2000
0
-2000
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
Total Population and Land Area

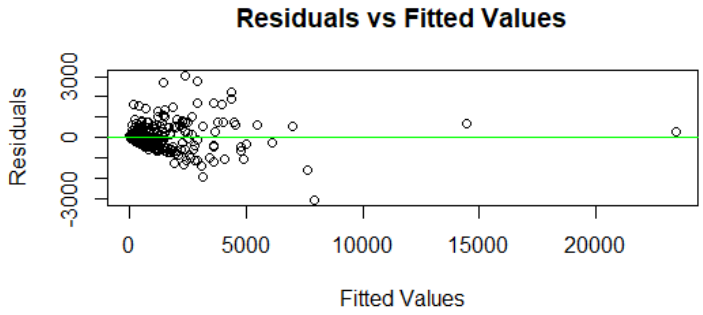

Residuals vs Total Population and Total Personal Income
Residuals
2000
0
-2000
5.0e+08
1.0e+09
1.5e+09
2.0e+09
Total Population and Total Personal Income

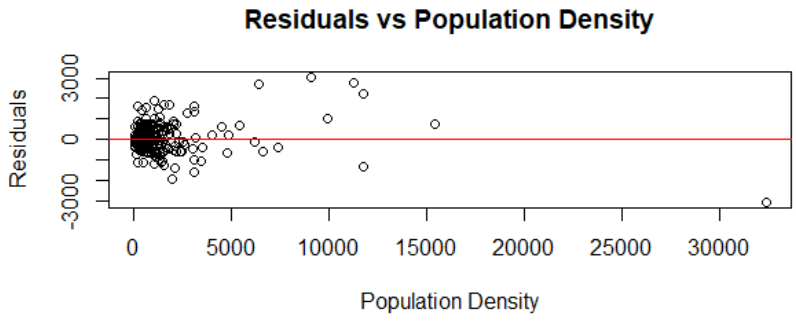

Residuals vs Land Area and Total Personal Income
Residuals
2000
0
-2000
0e+00
2e+08
4e+08
6e+08
Land Area and Total Personal Income

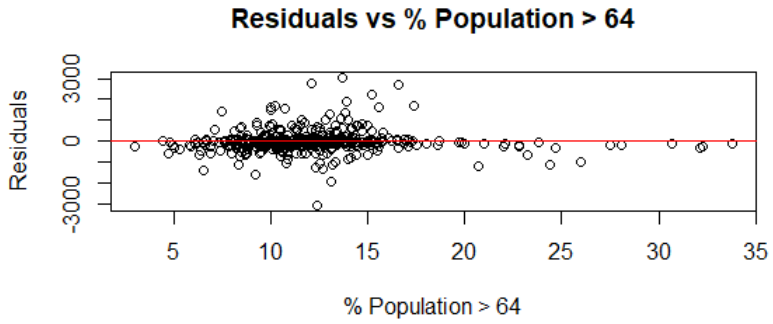


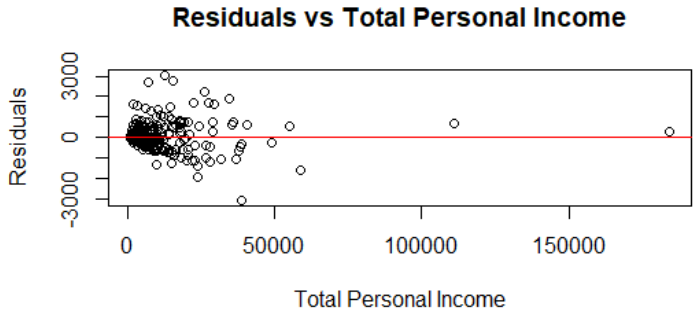
The residual plots have non-constant variance (except for X_2) but most of the data are clustered around the zero line, with no obvious pattern, indicating a linear relationship. There do seem to be a few outliers in each residual plot. The normal probability plot has long tails, indicating nonconstant variance in the model.

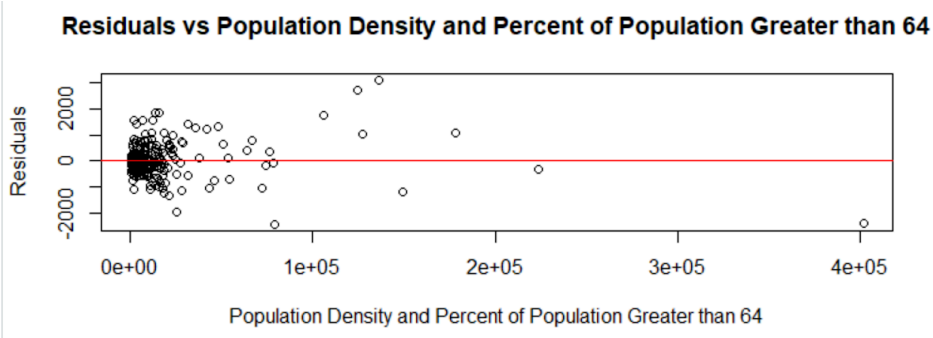
Model 2 Plots:

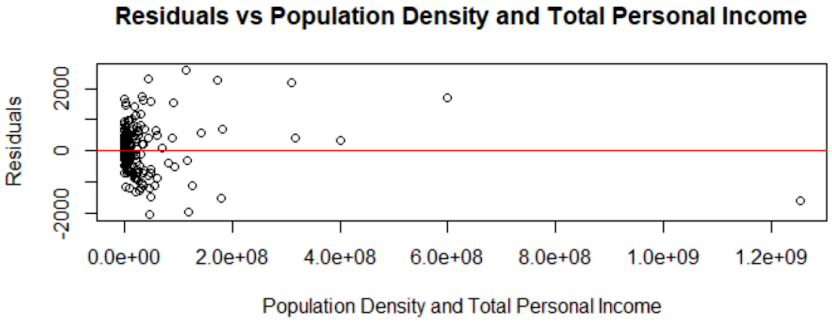


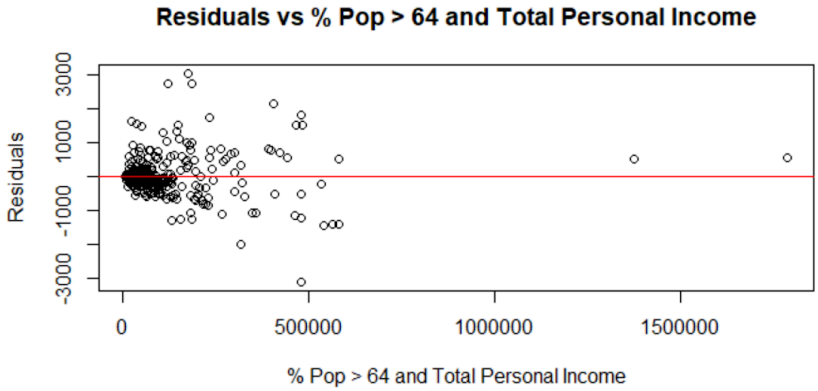


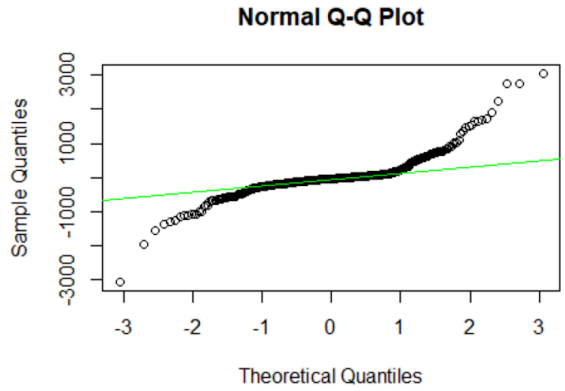

Residuals vs % Population > 64
Residuals
3000
0
-3000
5
10
15
20
25
30
35
% Population > 64


Residuals vs Total Personal Income
Residuals
3000
0
-3000
0
50000
100000
150000
Total Personal Income


Residuals vs Population Density and Percent of Population Greater than 64
Residuals
2000
0
-2000
0e+00
1e+05
2e+05
3e+05
4e+05
Population Density and Percent of Population Greater than 64


Residuals vs Population Density and Total Personal Income
Residuals
2000
0
-2000
0.0e+00
2.0e+08
4.0e+08
6.0e+08
8.0e+08
1.0e+09
1.2e+09
Population Density and Total Personal Income





In the residual plots for model 2 against fitted values and each of the predictor variables have non-constant variance (except for X_2) but most of the data are clustered around the zero line, with no obvious pattern, indicating a linear relationship. There do seem to be a few outliers in each residual plot. The normal probability plot has long tails, indicating nonconstant variance in the model. Model 2 seems to be preferable in terms of appropriateness because it has a higher R^2 value than model 1.

f. Now expand both models proposed above by adding all possible two-factor interactions. Note that, for a model with X_1, X_2, X_3 as the predictors, the two-factor interactions are X_1X_2, X_1X_3, X_2X_3. Repeat part d for the two expanded models. (See end of project for outputs)

EXPANDED MODEL 1:

$$Y = -5.825714e+1+7.252389e-4X_1-6.421359e-2X_2+1.086951e-1X_3+6.173047e-07X_1X_2+1.696289e-9X_1X_3-3.706225e-05X_2X_3$$

$R^2 = 0.9064$

Approximately 91% of the variation in the number of active physicians(Y) can be explained by the regression function involving total population(X_1), land area(X_2), total personal income (X_3), and two factor-interactive terms X_1X_2, X_1X_3, X_2X_3.

EXPANDED MODEL 2:

$$Y = -9.367e+0 - 4.179e-1X_1 - 1.106e+1X_2 + 1.477e-1X_3 + 4.652e-2X_1X_2 - 3.276e-6X_1X_3 - 1.289e-3X_2X_3$$

$R^2 = 0.923$

Approximately 92% of the variation in the number of active physicians (Y) can be explained by the regression function involving population density (X_1), percent of population greater than 64 years (X_2), total personal income (X_3), and two factor-interactive terms X_1X_2, X_1X_3, X_2X_3.

In terms of variability in number of active physicians, proposed model 2-which involves population density, percent of population greater than 64 years of age, and total personal income as well as the two factor interactive terms -- X_1X_2, X_1X_3, X_2X_3 -- is clearly slightly more preferable in terms of R^2, with approximately 92% explained compared to approximately 91% explained by the predictor variables in proposed model 1.

Part II: Multiple Regression Part 2

7.37

a. For each of the following variables, calculate the coefficient of partial determination given that X_1 and X_2 are included in the model: Land Area (X_3), Percent of the Population 65 or older (X_4), number of Hospital Beds (X_5).

Our coefficients of partial determination are:

$$R^2_{Y,X_3|X_1,X_2} = 0.02882495$$
$$R^2_{Y,X_4|X_1,X_2} = 0.003842367$$
$$R^2_{Y,X_4|X_1,X_2} = 0.5538182$$

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 other variables?

The best predictor variable out of X_3, X_4, and X_5 is X_5 because it has the largest coefficient of partial determination given that X_1 and X_2 are included in the model. Yes, from the ANOVA table, the sum of squares for X_5 is larger than those for X_3 and X_4.

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 X_1 and X_2 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.

Our model is $Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \beta_5 * X_{i5}$. The alternatives are $H_0 : \beta_5 = 0; H_1 : \beta_5 \neq 0$.
If $F* \leq F(1-\alpha, 1, n-p)$, then we accept the null hypothesis. Otherwise, we reject the null hypothesis and accept the alternative hypothesis.
$F* = 541.1801 > 6.693358 = F(0.99, 1, 436)$ so we reject the null hypothesis and accept that β_5 is nonzero.
No, the other $F*$ test statistics would not be larger because all the other R^2 are less than that of X_5. Since $F* = R^2/(1-R^2) * ((n-p)/1)$ and $1 - R^2$ increases as R^2 decreases, $F*$ decreases as R^2 decreases.

d. Compute three additional coefficients of partial determination:

$R^2_{Y,X_3,X_4|X_1,X_2}, R^2_{Y,X_3,X_5|X_1,X_2}$, and $R^2_{Y}, X_4, X_5|X_1, X_2$.

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 X_1, X_2 are already included.

The additional coefficients of partial determination are:.

$R^{2}_{Y,X_3,X_4|X_1,X_2} = 0.03314181$

$R^{2}_{Y,X_3,X_5|X_1,X_2} = 0.5558232$

$R^{2}_{Y,X_4,X_5|X_1,X_2} = 0.5642756$

The pair X_4, X_5 is relatively more important than the other pairs because it has a higher coefficient of partial determination.

We conduct an F-test for the pair. Our model is
$Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \beta_4 * X_{i4} + \beta_5 * X_{i5}$. The alternatives are
$H_0 : \beta_4 = \beta_5 = 0; H_1 : \beta_4 \neq 0 or \beta_5 \neq 0$.
If $F* \leq F(1-\alpha, p-q, n-p)$, then we accept the null hypothesis. Otherwise, we reject the null hypothesis and accept the alternative hypothesis.
$F* = 281.6688 > 4.654269 = F(0.99, 2, 435)$ so we reject the null hypothesis and accept that either β_4 or β_5 is nonzero.

Part III: Discussion

In part (a) of Part I, we created stem plots for each of the predictor variables and found that the data is skewed right in the majority of cases with a cluster around 0, with a few outliers. We find that the second model is slightly more appropriate than the first as evidenced by the larger R^2 values for the unexpanded and expanded versions of the model in parts (b) and (d). The residual plots clearly have nonconstant variance (except for X_2 in both models) but most of the data are clustered around the zero line, indicating a linear relationship for both models. We also determined that X_1 and X_3 have a strong relationship with number of active physicians, while X_2 does not for the first model, while for model 2, population density had a middling correlation with number of active physicians while percent of population greater than 64 has almost no correlation and total personal income had a very strong relationship with number of active physicians. In part (f), the expanded models with two-factor interactions, we also found that

Model 2 is preferable in terms of appropriateness to Model 1 because its R^2 is larger than model 1's.

In part (a) of Part II, we found the coefficient of partial determination for the variables of Land Area (X_3), Percent of the Population 65 or older (X_4), and number of Hospital Beds (X_5); measuring the proportionate reduction in variation in response to the addition of each variable. We did the same in part (d) for the pairs $X_3, X_4; X_3, X_5;$ and X_4, X_5. For both cases, we took as a given that X_1 and X_2 were already included in the model. In part (a), the variable X_5 had the highest coefficient of partial determination of the three variables and in part (d), the pair X_4, X_5 had the same quality of the three pairs; i.e. they had the closest R^2 to 1 so the highest correlation with Y under the given conditions. This means that the number of hospital beds is the best single variable to be added to the model if X_1 and X_2 are already included, the number of hospital beds together with the percent of population over 65 make the best pair to add to the model.

Both of our F-tests gave us very large p-values that let us reject the null hypothesis that the variable or variable pair we were testing had zero effect on the model.

The models might be improved if we expanded the model to include X_3X_4, etc. We could also check variables that we did not try to find correlation for in this project.

Outputs

Part I Screenshots

```
> summary(fit1)

Call:
lm(formula = V8 ~ (V5 + V4 + V16), data = CDI_dataset)

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
V5           8.366e-04  2.867e-04   2.918 0.003701 **
V4          -6.552e-02  1.821e-02  -3.597 0.000358 ***
V16          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

> coef(summary(fit1))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1950.2  -198.0   -61.1    76.6  3578.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.826e+01  4.727e+01  -1.232  0.21848
V5           7.252e-04  3.259e-04   2.225  0.02657 *
V4          -6.421e-02  3.014e-02  -2.131  0.03369 *
V16          1.087e-01  1.450e-02   7.496 3.76e-13 ***
V5:V4        6.173e-07  2.058e-07   2.999  0.00287 **
V5:V16       1.696e-09  1.041e-09   1.630  0.10392
V4:V16      -3.706e-05  1.152e-05  -3.217  0.00139 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 551.4 on 433 degrees of freedom
Multiple R-squared:  0.9064,    Adjusted R-squared:  0.9051
F-statistic: 698.7 on 6 and 433 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = numActivePhys ~ popDens + perPopG64 + totalPersonalIncome,
    data = data2)

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 *
popDens            9.616e-02  1.224e-02   7.857  3.1e-14 ***
perPopG64          6.340e+00  6.384e+00   0.993   0.3212
totalPersonalIncome  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

> coef(summary(fit2))
```

```
lm(formula = numActivePhys ~ (popDens * perPopG64 + popDens *
    totalPersonalIncome + perPopG64 * totalPersonalIncome), data =

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     0.92
popDens                         -4.179e-01  1.055e-01  -3.960 8.76e-0
perPopG64                       -1.106e+01  7.792e+00  -1.419     0.15
totalPersonalIncome              1.477e-01  9.739e-03  15.168   < 2e-1
popDens:perPopG64                4.652e-02  7.925e-03   5.870 8.67e-0
popDens:totalPersonalIncome     -3.276e-06  7.439e-07  -4.404 1.34e-0
perPopG64:totalPersonalIncome   -1.289e-03  8.743e-04  -1.474     0.14
---
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
```

Part II Screenshots

```
> y  <- cdi$V8
> x1 <- cdi$V5
> x2 <- cdi$V16
> x3 <- cdi$V4
> x4 <- cdi$V7
> x5 <- cdi$V9
```

```
> (a12 <- anova(lm(y~x1+x2)))
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
> sse12 <- a12[3,2]
> (a123 <- anova(lm(y~x1+x2+x3)))
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
> ssr312 <- a123[3,2]
> (rsq312 <- ssr312/sse12)
[1] 0.02882495
> (a124 <- anova(lm(y~x1+x2+x4)))
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
```

```
> ssr412 <- a124[3,2]
> (rsq412 <- ssr412/sse12)
[1] 0.003842367
> (a125 <- anova(lm(y~x1+x2+x5)))
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
> ssr512 <- a125[3,2]
> (rsq512 <- ssr512/sse12)
[1] 0.5538182
> sse125 <- a125[4,2]; n = nrow(cdi); p = 4
> (fstar = (ssr512)/(sse125/(n-p)))
[1] 541.1801
> (fval <- qf(.99,1,n-p))
[1] 6.693358
> (a1234 <- anova(lm(y~x1+x2+x3+x4)))
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
> ssr3412 <- a1234[3,2] + a1234[4,2]
> (rsq3412 <- ssr3412/sse12)
[1] 0.03314181
```

```
> (a1235 <- anova(lm(y~x1+x2+x3+x5)))
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
> ssr3512 <- a1235[3,2] + a1235[4,2]
> (rsq3512 <- ssr3512/sse12)
[1] 0.5558232
> (a1245 <- anova(lm(y~x1+x2+x4+x5)))
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
> ssr4512 <- a1245[3,2] + a1245[4,2]
> (rsq4512 <- ssr4512/sse12)
[1] 0.5642756
> sse1245 <- a1245[5,2]; p0 <- 5; q0 <- 3
> (fstar0 <- ((sse12-sse1245)/(p0-q0))/(sse1245/(n-p0)))
[1] 281.6688
> (fval0 <- qf(.99,p0-q0,n-p0))
[1] 4.654269
```