Homework 4

Brandon Lucio

09/28/2021

Not from the text book

A data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The 17 variables are given in Table 1. An R data (CDI_data.rda) is included. You can put it into your homework file directory and use readRDS("CDI_data.rda") to load the data.

```
CDI <- readRDS("CDI_data.rda")
```

Problem 1

The number of active physicians (Y) is to be regressed against total population (X_1), total personal income (X_2), and geographic region (X_3, X_4, X_5) that will be defined next.

a. Fit a first-order regression model. Let $X_3 = 1$ if **NE** and 0 otherwise. $X_4 = 1$ if **NC** and 0 otherwise, and $X_5 = 1$ if S and 0 otherwise. Comment on your results.

```
# Creating dummy variables
CDI$X3 <- ifelse(CDI$Region == 1, 1 ,0) # NE
CDI$X4 <- ifelse(CDI$Region == 2, 1 ,0) # NC
CDI$X5 <- ifelse(CDI$Region == 3, 1 ,0) # S
```

```
Physicians.Model <- lm(NumPhysicians ~ TotalPop + TotalPIncome + X3 + X4 + X5,
                       data = CDI)
summary(Physicians.Model)
```

```
##
## Call:
## lm(formula = NumPhysicians ~ TotalPop + TotalPIncome + X3 + X4 +
##     X5, data = CDI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1866.8  -207.7   -81.5    72.4  3721.7
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.075e+02  7.028e+01  -2.952  0.00332 **
## TotalPop      5.515e-04  2.835e-04   1.945  0.05243 .
## TotalPIncome  1.070e-01  1.325e-02   8.073  6.8e-15 ***
## X3            1.490e+02  8.683e+01   1.716  0.08685 .
## X4            1.455e+02  8.515e+01   1.709  0.08817 .
## X5            1.912e+02  8.003e+01   2.389  0.01731 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.1 on 434 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8999
## F-statistic: 790.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

Thanks to our dummy variables we can see that we have 4 different models for each region. Given by:

Region	Model(y_i)
North East	-207.496 + 0.000551(TotalPop) +0.107(TotalPIncome) + 149.0196
North Central	-207.496 + 0.000551(TotalPop) +0.107(TotalPIncome) + 145.5264
South	-207.496 + 0.000551(TotalPop) +0.107(TotalPIncome) + 191.2163
West	-207.496 + 0.000551(TotalPop) +0.107(TotalPIncome)

This shows that the West region has the lowest active physicians compared to the other regions. The South has 191 more, North Central with 146 more, and North East with 149 more. We also see that the overall test of a relationship gives a p-value of essentially 0. This means that at least one of the predictors. The adjusted R^2 is 0.8999 this means that about 90% of the variability in Y can be explained by the predictors.

b. Use a level of significance $\alpha = 0.10$ to test whether any geographic effects are present. State the alternatives, decision rule, and conclusion. What is the p-value of the test?

Regressor	Alternative	P-Value	Decision Rule	Conclusion
North East	$\beta_3 \neq 0$	0.08685	P-value < 0.10	X3 is significant if all other variables are constant
North Central	$\beta_4 \neq 0$	0.08817	P-value < 0.10	X4 is significant if all other variables are constant
South	$\beta_5 \neq 0$	0.01731	P-value < 0.10	X5 is significant if all other variables are constant

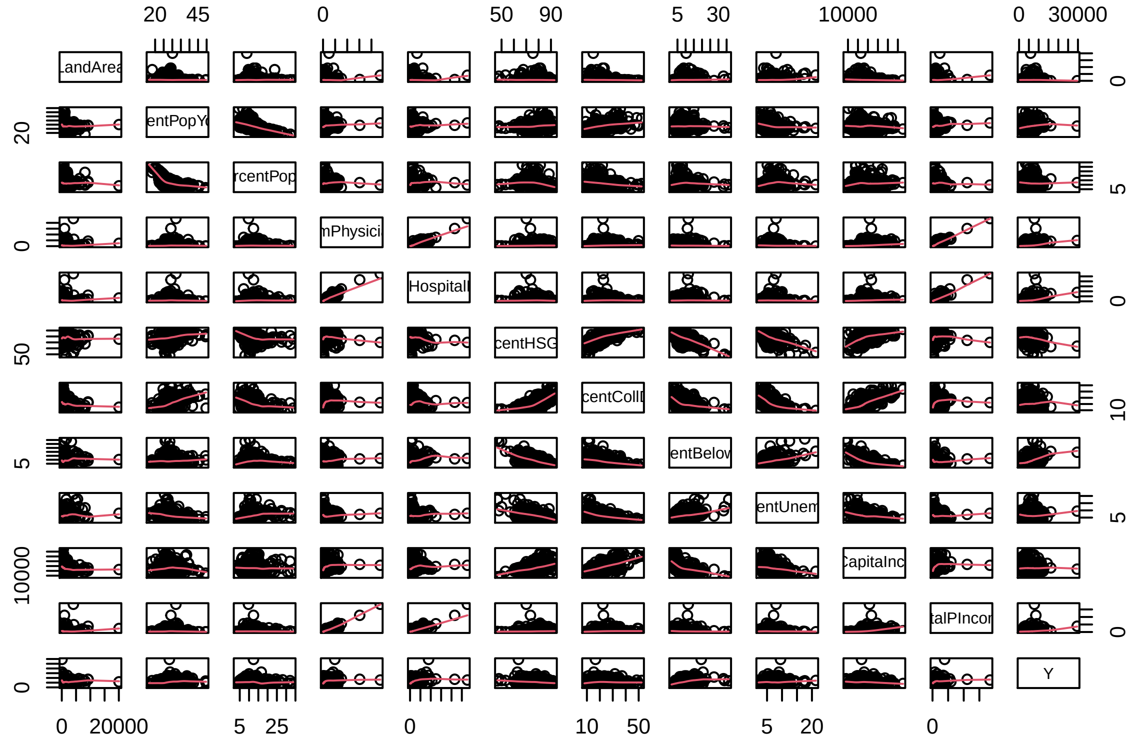
Problem 2

A public safety official wishes to predict the rate of serious crimes (Y , total number of serious crimes per 100,000 population, i.e., you need to recalculate the response here). The pool of potential predictor variables includes all other variables in the data set except total population, total serious crimes, county, state, and region. It is believed that a model with predictor variables in first-order terms with no interaction terms will be appropriate. Consider the even-numbered cases to constitute the model-building data set to be used for the following analyses.

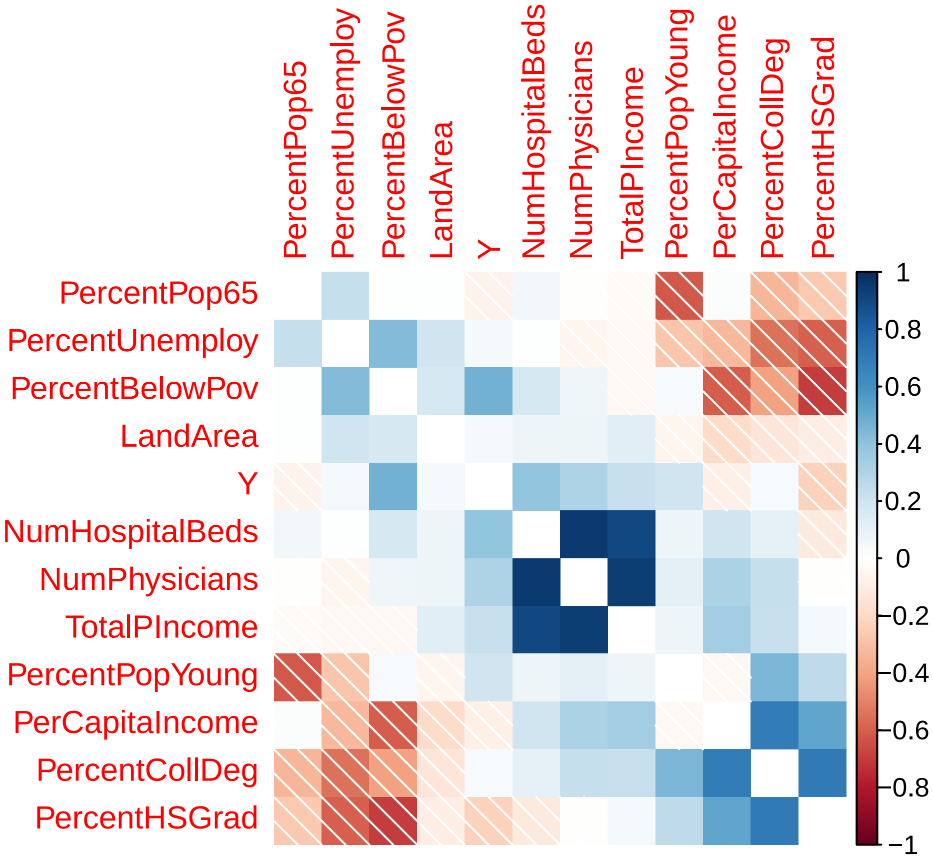
```
CDI2 <- readRDS("CDI_data.rda") %>% mutate(Y = TotCrimes/(TotalPop/100000))
CDI2<- CDI2[,-c(1,2,3,5,10,17)]
model <- lm(Y ~., data = CDI2)
```

a. Obtain the scatter plot matrix. Also obtain the correlation matrix of the X variables. Is there evidence of strong linear pairwise associations among the predictor variables here?

```
pairs(CDI2, panel = panel.smooth)
```



```
corrplot(cor(CDI2), method = 'shade', order = 'AOE', diag = FALSE)
```



- From the scatterplot we can see some linear relationships between PercentHSGrad: PercentColDeg, PercentBelowPov, PercentUnemploy. This relationship is given more strength when looking at the correlation matrix, with correlation coefficients of 0.707, -0.692, -0.594. NumPhysicians and NumHospitalBeds also have a high correlation at 0.950.

b. Use several model selection techniques to pick up models and comment on them. If there are multiple candidates, pick one you deem attractive.

```
#Forward Selection
ols_step_forward_p(model, penter = 0.05)
```

```
##
##                                    Selection Summary
## ---------------------------------------------------------------------------------------
##         Variable                        Adj.
## Step        Entered          R-Square    R-Square      C(p)         AIC         RMSE
## ---------------------------------------------------------------------------------------
##    1    PercentBelowPov        0.2226      0.2209    125.4425    8106.3547    2412.1841
##    2    NumHospitalBeds        0.3217      0.3186     55.9061    8048.3846    2255.8518
##    3    PercentCollDeg         0.3604      0.3560     29.9273    8024.5098    2192.9989
##    4    NumPhysicians          0.3867      0.3811     12.9156    8008.0174    2149.8627
## ---------------------------------------------------------------------------------------
```

```r
#Backward Selection
ols_step_backward_p(model, prem = 0.1)
```

```
##
##
##                                Elimination Summary
## -----------------------------------------------------------------------------------
##         Variable                         Adj.
## Step       Removed          R-Square    R-Square     C(p)        AIC         RMSE
## -----------------------------------------------------------------------------------
##    1     PercentPop65         0.4074      0.3936    10.0041    8004.9490    2128.0910
##    2     PercentHSGrad        0.4074       0.395     8.0156    8002.9608    2125.6435
##    3     LandArea             0.4069      0.3959     6.3307    8001.2846    2123.9575
##    4     PercentCollDeg       0.4062      0.3966     4.8555    7999.8233    2122.7970
##    5     TotalPIncome         0.4049      0.3966     3.8389    7998.8312    2122.7741
## -----------------------------------------------------------------------------------
```

```r
# Both
ols_step_both_p(model, pent = 0.05, prem = 0.1 )
```

```
##
##                                   Stepwise Selection Summary
## -------------------------------------------------------------------------------------------------
##                               Added/                     Adj.
## Step       Variable          Removed     R-Square      R-Square      C(p)         AIC         RMSE
## -------------------------------------------------------------------------------------------------
##    1     PercentBelowPov    addition        0.223         0.221    125.4430    8106.3547    2412.1841
##    2     NumHospitalBeds    addition        0.322         0.319     55.9060    8048.3846    2255.8518
##    3     PercentCollDeg     addition        0.360         0.356     29.9270    8024.5098    2192.9989
##    4      NumPhysicians     addition        0.387         0.381     12.9160    8008.0174    2149.8627
## -------------------------------------------------------------------------------------------------
```

As the backward selection model has a more predictors I picked this model to move forward with. This should in general lead to a higher R^2

c. Use the model you picked in part (b) above and run a regression of that model for the odd-numbered cases (test data). Compare the estimated regression coefficients and their estimated standard deviations of odd-numbered case data with those in the model you obtained in part (b). In addition, compare the error mean squares and coefficients of multiple determination. Does the model fitted to the test data set yield similar estimates as the model fitted to the model-building data set?

```r
# Getting only odd numbered cases as test data
test_CDI2 <- CDI2[c(TRUE, FALSE),-c(1,3,6,7,11)]
train_CDI2 <- CDI2[!c(TRUE, FALSE),-c(1,3,6,7,11)]

model.test <- lm(Y ~., data = test_CDI2)
summary(model.test)
```

```
##
## Call:
## lm(formula = Y ~ ., data = test_CDI2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5184.1 -1377.7   -81.2  1295.0  5540.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.586e+03  1.766e+03  -1.464  0.14466
## PercentPopYoung  7.240e+01  3.451e+01   2.098  0.03707 *
## NumPhysicians   -7.929e-01  3.053e-01  -2.597  0.01005 *
## NumHospitalBeds  7.431e-01  2.415e-01   3.076  0.00237 **
## PercentBelowPov  3.393e+02  4.555e+01   7.448 2.33e-12 ***
## PercentUnemploy -8.641e+01  6.864e+01  -1.259  0.20945
## PerCapitaIncome  1.985e-01  4.923e-02   4.032 7.71e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2042 on 213 degrees of freedom
## Multiple R-squared:  0.3408, Adjusted R-squared:  0.3222
## F-statistic: 18.35 on 6 and 213 DF,  p-value: < 2.2e-16
```

```r
model.full <- lm(Y ~., data = CDI2[,-c(1,3,6,7,11)])
summary(model.full)
```

```
##
## Call:
## lm(formula = Y ~ ., data = CDI2[, -c(1, 3, 6, 7, 11)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5500.4 -1305.0  -164.9  1132.4 17403.0
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.134e+03  1.255e+03  -2.498  0.01287 *
## PercentPopYoung  1.031e+02  2.619e+01   3.936 9.64e-05 ***
## NumPhysicians   -8.854e-01  2.063e-01  -4.292 2.18e-05 ***
## NumHospitalBeds  9.297e-01  1.572e-01   5.914 6.77e-09 ***
## PercentBelowPov  3.415e+02  3.114e+01  10.967  < 2e-16 ***
## PercentUnemploy -1.367e+02  5.123e+01  -2.668  0.00791 **
## PerCapitaIncome  1.810e-01  3.574e-02   5.065 6.07e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2123 on 433 degrees of freedom
## Multiple R-squared:  0.4049, Adjusted R-squared:  0.3966
## F-statistic: 49.09 on 6 and 433 DF,  p-value: < 2.2e-16
```

From the summary outputs we can see that test model has similar results for the coefficients for the predictors except for PercentPopYoung and PercentUnemploy

d. Now use the odd-numbered cases as test (validate) data and fit your model part (b) with training data set (even-numbered cases). Calculate the MSE of the predictions to the test data and compare it to

MSE of the predictions to training data set. Is there evidence of a substantial bias problem in MSE here?

```
# Fitting model with training data set
model.train <- lm(Y ~., data = train_CDI2)

# Computing the training MSE
MSE.Training <- mean((model.train$residuals)^2)

# Testing MSE
data <- data.frame(pred = predict(model.train, , newdata = test_CDI2),actual = test_CDI2$Y )

MSE.Test <- mean((data$actual-data$pred)^2)

sprintf("Training MSE: %.f",MSE.Training)
```

```
## [1] "Training MSE: 4486552"
```

```
sprintf("Test MSE: %.f",MSE.Test)
```

```
## [1] "Test MSE: 4926209"
```

The test MSE is larger than the Training as to be expected. As such there does not seem to a substantial bias problem.

e. Finally, fit the selected regression model to the combined training and test data sets. Are the estimated regression coefficients and their estimated standard deviations appreciably different from those for the model fitted to the training data set? Should you expect any differences in the estimates? Explain.

```
summary(model.full)
```

```
##
## Call:
## lm(formula = Y ~ ., data = CDI2[, -c(1, 3, 6, 7, 11)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5500.4 -1305.0  -164.9  1132.4 17403.0
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.134e+03  1.255e+03  -2.498  0.01287 *
## PercentPopYoung  1.031e+02  2.619e+01   3.936 9.64e-05 ***
## NumPhysicians   -8.854e-01  2.063e-01  -4.292 2.18e-05 ***
## NumHospitalBeds  9.297e-01  1.572e-01   5.914 6.77e-09 ***
## PercentBelowPov  3.415e+02  3.114e+01  10.967  < 2e-16 ***
## PercentUnemploy -1.367e+02  5.123e+01  -2.668  0.00791 **
## PerCapitaIncome  1.810e-01  3.574e-02   5.065 6.07e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2123 on 433 degrees of freedom
## Multiple R-squared:  0.4049, Adjusted R-squared:  0.3966
## F-statistic: 49.09 on 6 and 433 DF,  p-value: < 2.2e-16
```

```r
summary(model.train)
```

```
##
## Call:
## lm(formula = Y ~ ., data = train_CDI2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8189.2 -1128.6  -148.9   934.6 16424.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.300e+03  1.769e+03  -1.866  0.06345 .
## PercentPopYoung  1.360e+02  4.097e+01   3.320  0.00106 **
## NumPhysicians   -7.062e-01  2.890e-01  -2.443  0.01537 *
## NumHospitalBeds  9.877e-01  2.061e-01   4.792 3.10e-06 ***
## PercentBelowPov  3.520e+02  4.305e+01   8.178 2.58e-14 ***
## PercentUnemploy -2.198e+02  7.574e+01  -2.902  0.00409 **
## PerCapitaIncome  1.458e-01  5.159e-02   2.826  0.00516 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2153 on 213 degrees of freedom
## Multiple R-squared:  0.4882, Adjusted R-squared:  0.4738
## F-statistic: 33.87 on 6 and 213 DF,  p-value: < 2.2e-16
```

The two different models do have noticeable differences when it comes to the regression coefficients and their estimated standard deviations. I do think we would expect differences if this was done with different data as we are adding more data to the model which would give different values.

Problem 3

Now consider the regression model, with total crime (per 100,000) as response and the variables 6, 8, 9, 13, 14, and 15 as predictors, in first-order terms is to be evaluated in detail based on the training data set (even-numbered cases).

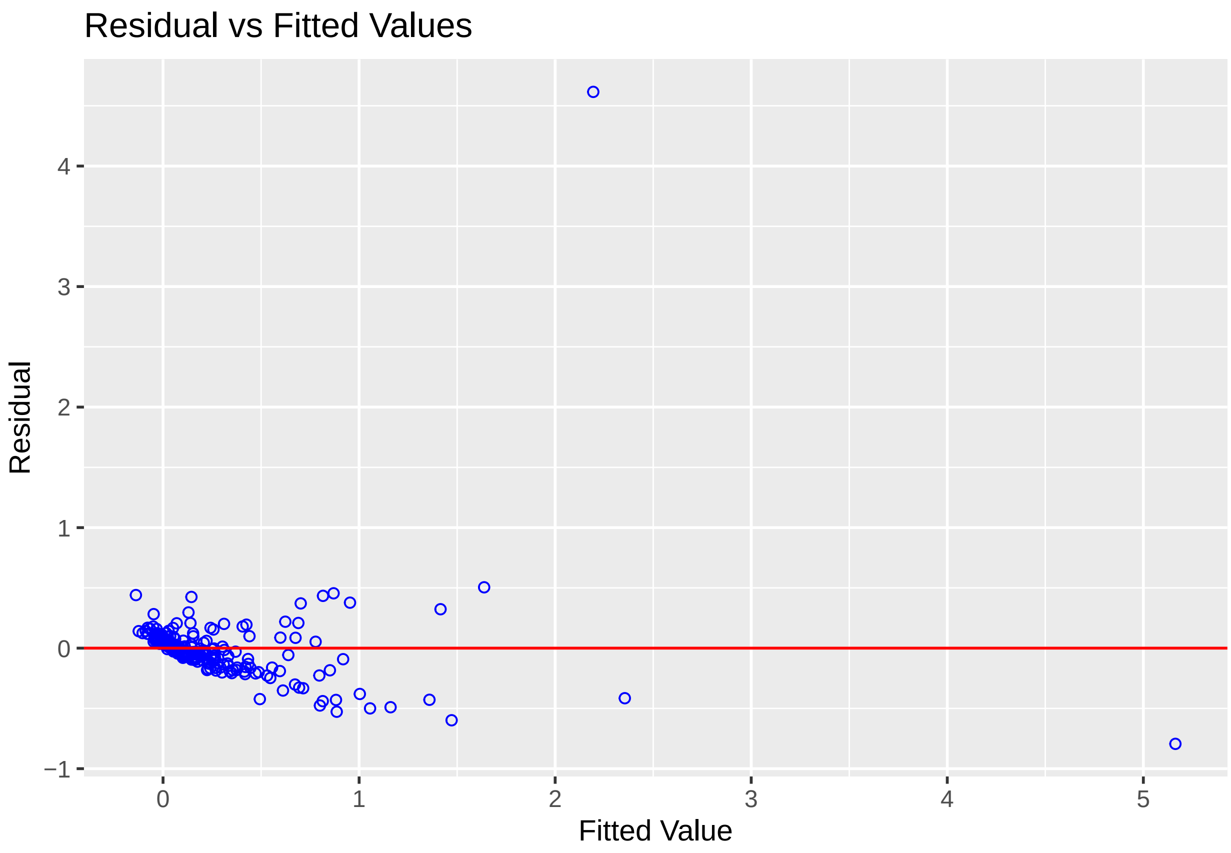
```r
CDI3 <- readRDS("CDI_data.rda") %>% mutate(Y = TotCrimes/(100000))
CDI3<- CDI3[,c(6,8,9,13,14,15,18)]
train_CDI3 <- CDI3[!c(TRUE, FALSE),] # Selecting Even observations for training

model.3 <- lm(Y ~., data = train_CDI3 )
summary(model.3)
```

```
##
## Call:
## lm(formula = Y ~ ., data = train_CDI3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7942 -0.0917 -0.0044  0.0771  4.6153
##
```

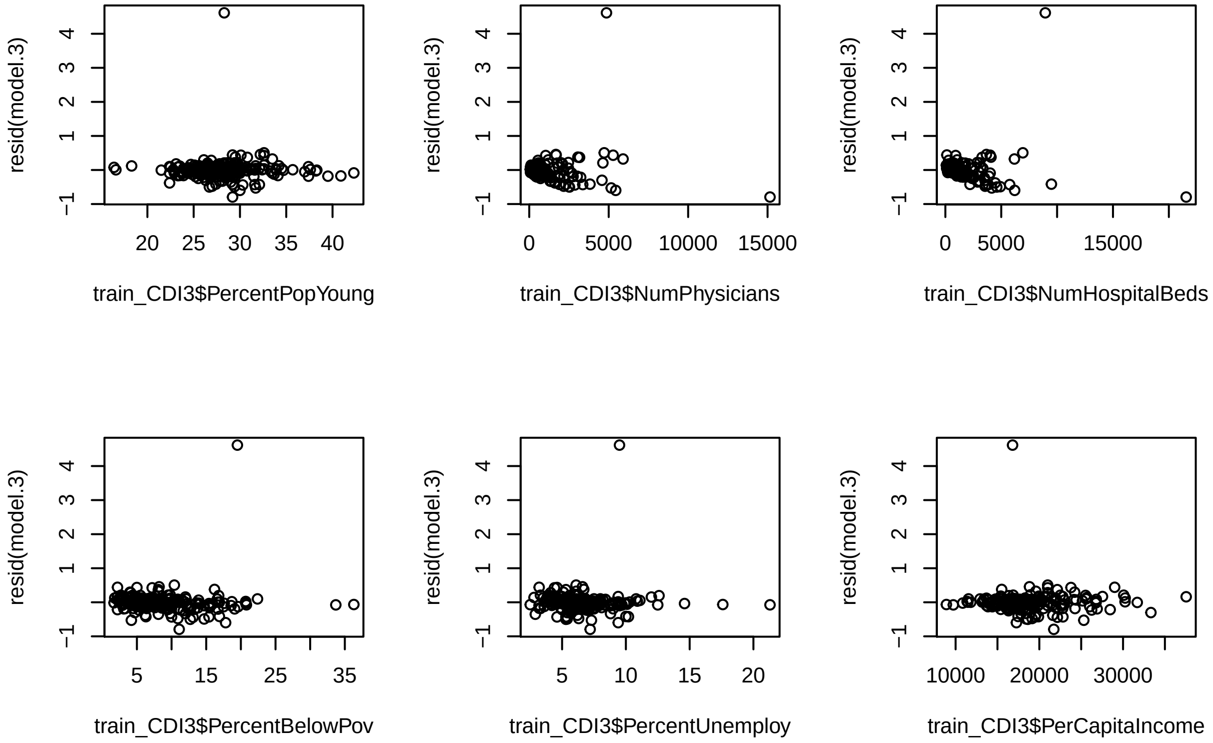
```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.638e-01  3.004e-01  -0.878    0.381
## PercentPopYoung   5.791e-03  6.957e-03   0.832    0.406
## NumPhysicians    -2.522e-05  4.909e-05  -0.514    0.608
## NumHospitalBeds   2.599e-04  3.501e-05   7.424 2.69e-12 ***
## PercentBelowPov   2.185e-03  7.310e-03   0.299    0.765
## PercentUnemploy   9.696e-03  1.286e-02   0.754    0.452
## PerCapitaIncome  -2.480e-06  8.762e-06  -0.283    0.777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3656 on 213 degrees of freedom
## Multiple R-squared:  0.6499, Adjusted R-squared:  0.6401
## F-statistic: 65.91 on 6 and 213 DF,  p-value: < 2.2e-16
```

a. Obtain the residuals and plot them separately against Y , each predictor variable in the model, On the basis of these plots, should any modifications in the model be made?

```
ols_plot_resid_fit(model.3)
```



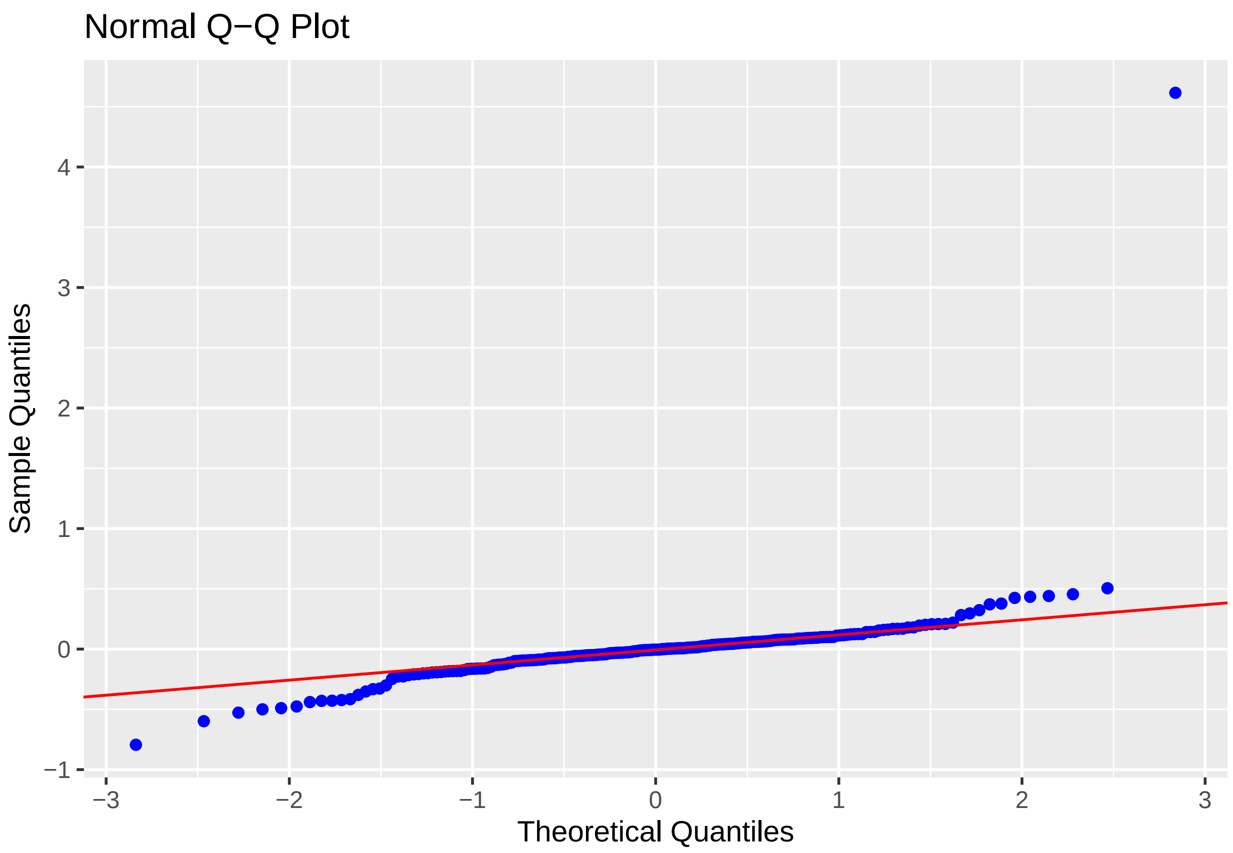
```
par(mfrow = c(2,3))
plot(y = resid(model.3), x = train_CDI3$PercentPopYoung )
plot(y = resid(model.3), x = train_CDI3$NumPhysicians )
plot(y = resid(model.3), x = train_CDI3$NumHospitalBeds)
plot(y = resid(model.3), x = train_CDI3$PercentBelowPov )
plot(y = resid(model.3), x = train_CDI3$PercentUnemploy )
plot(y = resid(model.3), x = train_CDI3$PerCapitaIncome )
```



- There do seem to be some outliers but the biggest issue seems to be from NumPhysicians and NumHospitalBeds. They seem to have nonconstant variance and similar residuals. More investigation should be done to rule out multicollinearity.

b. Do a normal probability plot of the residuals and test for normality at a significance level of 0.10.

```
ols_plot_resid_qq(model.3)
```



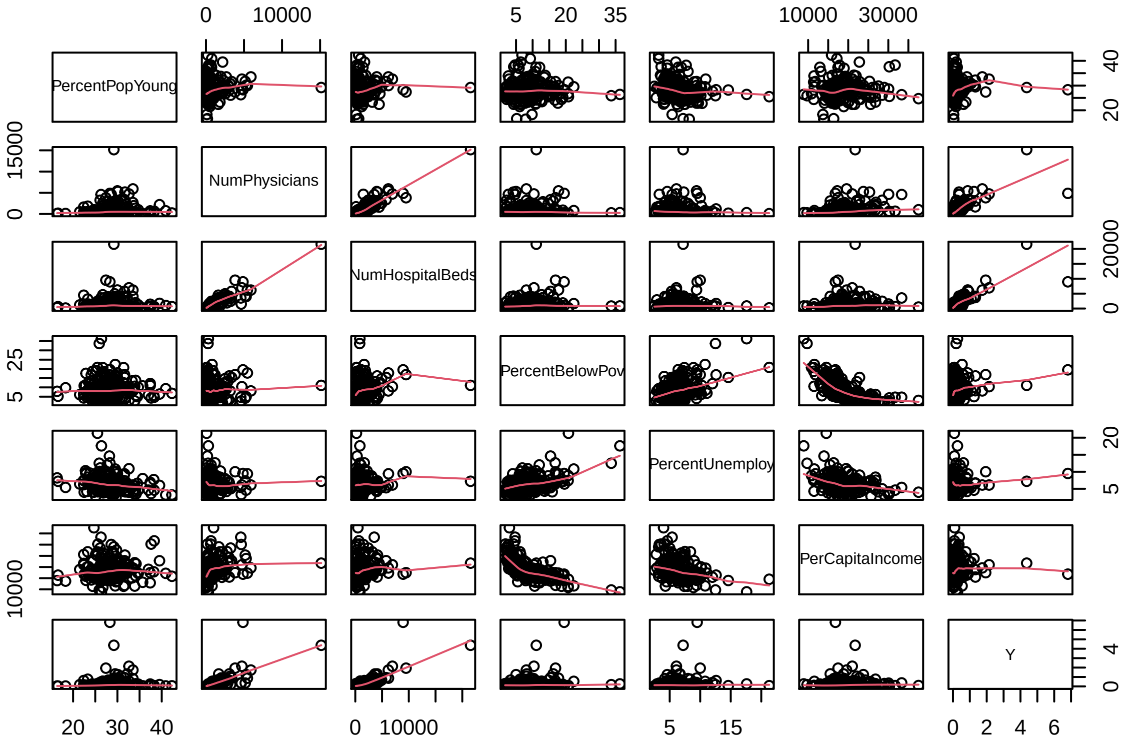
```
ols_test_normality(model.3)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue  
## -----------------------------------------------
## Shapiro-Wilk              0.4197         0.0000 
## Kolmogorov-Smirnov        0.2372         0.0000 
## Cramer-von Mises         54.9285         0.0000 
## Anderson-Darling         23.6723         0.0000 
## -----------------------------------------------
```

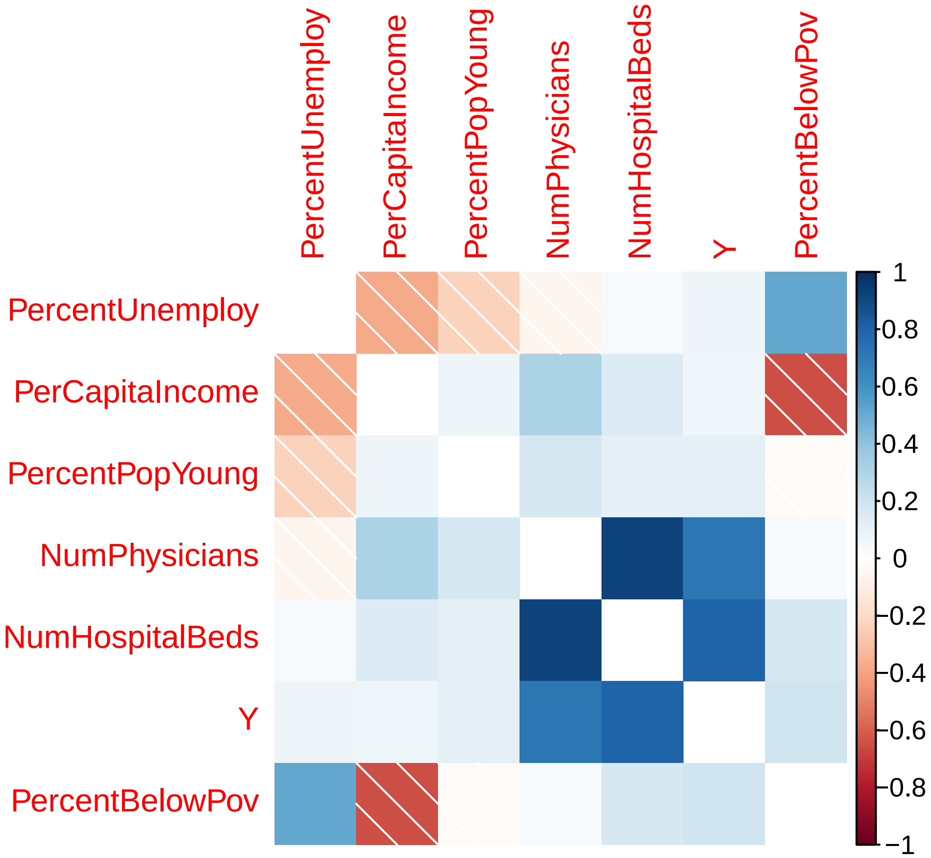
At $\alpha = 0.10$ we can see that the residuals are essentially normally distributed.

c. Obtain the scatter plot matrix, the correlation matrix of the X variables, and the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain.

```
pairs(train_CDI3, panel = panel.smooth)
```



```
corrplot(cor(train_CDI3) ,method = 'shade', order = 'AOE', diag = FALSE)
```



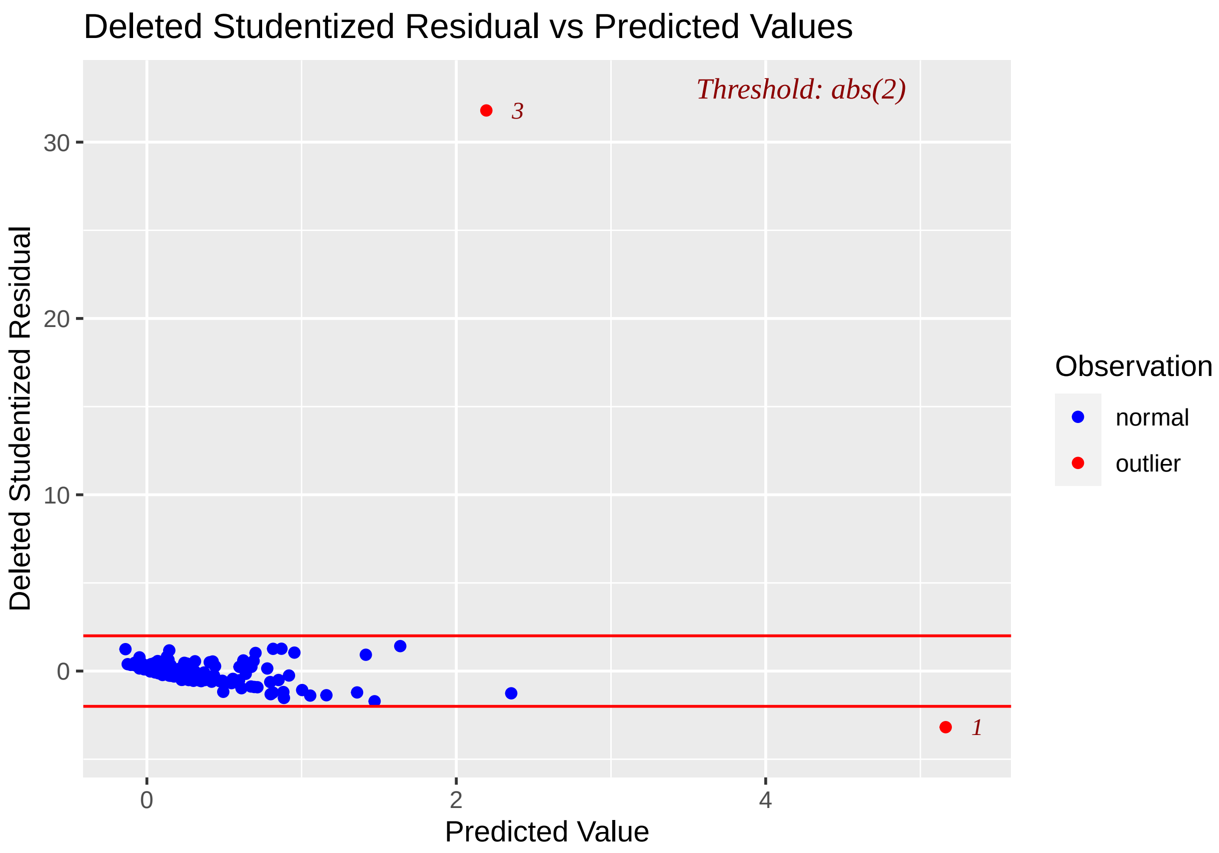
```
vif(model.3)
```

```
## PercentPopYoung   NumPhysicians NumHospitalBeds PercentBelowPov PercentUnemploy
##        1.106320        8.654011        8.038180        2.307348        1.447219
## PerCapitaIncome
##        2.226508
```

- From the vif, scatterplot, and correlation matrix we can see that NumPhyscians and NumHospitalBeds have high values of VIF (> 8) this indicates that these two values have collinearity.

d) Obtain the studentized deleted residuals and prepare a dot plot of these residuals. Are any outliers present? Use the Bonferroni outlier test (in R, use outlierTest in the *car* package) procedure with $\alpha = 0.05$. State the decision rule and conclusion.

```
ols_plot_resid_stud_fit(model.3)
```



```
outlierTest(model.3, cutoff = 0.05)
```

```
##   rstudent unadjusted p-value Bonferroni p
## 6 31.79744         1.2609e-82   2.7739e-80
```

The Bonferroni Outlier Test uses a t-test to test whether the model's largest studentized residual value's outlier status is statistically different from the other observations in the model. If the Bonferroni p-value < 0.05 then we reject the null hypothesis: the point is not different from the other observations, and conclude that observation 6 is indeed an outlier of the rest of the data. The graph shows that observation 6 should also be considered an outlier.

d) Obtain the diagonal elements of the hat matrix. Using the rule of thumbs described in the notes, identify any outlying X observations.

- The rule of thumb is $h_{ii} > \frac{2(p+1)}{n}$

```
hat.model.3 <- hatvalues(model.3);

# Using the rule of thumb we can see that the following are potential outliers
hat.model.3[hat.model.3 > 2*(ncol(train_CDI3))/nrow(train_CDI3)]
```

```
##          2          4          6          8         12         16         32
## 0.51402810 0.09029419 0.09479256 0.18596015 0.11146392 0.09907790 0.11378334
##         36         48         50        128        188        206        262
## 0.06639403 0.19810133 0.07801920 0.19099167 0.13580327 0.15172126 0.06812158
##        272        344        392        396        404
## 0.09211415 0.08628643 0.06598204 0.11329602 0.21161220
```

d) Cases 2, 8, 48, 128, 206, and 404 are outlying with respect to their X values, and Cases 2 and 6 are reasonably far outlying with respect to their Y values. Obtain **DFFITS**, **DFBETAS**, and **Cook's distance** values for these cases to assess their influence. What do you conclude?

DFFITS: Measure of influence on a single fitted value. Given by the formula

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i,-i}}{\sqrt{MSE_{-i}h_{ii}}} = t_i\sqrt{\frac{h_{ii}}{1-hii}}$$

Decision ROT: $(DFFITS)_i > 1$ for small to medium data sets and $> 2\sqrt{(p+1)/n}$ for large data sets.

```
cases <- c('2','6','8','48','128','206','404')
# DFFITS
dffits(model.3)[cases]
```

```
##           2           6           8          48         128         206
## -3.27296924 10.28975791 -0.60285124  0.31483139 -0.09660473  0.20251420
##         404
## -0.11911519
```

DFBETAS: Measure of influence of case i on each $\hat{\beta}_j$. Given by

$$(DFBETAS)_{j,-i} = \frac{\hat{\beta}_j - \hat{\beta}_{j,-i}}{\sqrt{MSE_{-i}c_{jj}}}$$

where c_{jj} is the jth diagonal element of $(X^TX)^{-1}$

Decision ROT: $DFBETAS > 1$ for small to medium data and $> 2/\sqrt{n}$ for large data.

```
# DFBETAS
dfbeta(model.3)[cases,]
```

```
##      (Intercept) PercentPopYoung NumPhysicians NumHospitalBeds PercentBelowPov
## 2   -0.170547087    1.941163e-03 -2.198755e-05   -2.749610e-05    5.537637e-03
## 6   -0.225976848    5.618961e-04 -8.280215e-05    9.602833e-05    5.437026e-03
## 8    0.037692778   -2.560421e-04  2.135098e-05   -1.886063e-05    2.893052e-05
## 48   0.008533392   -3.897635e-04  1.358172e-05   -9.363473e-06    4.189950e-04
## 128  0.010823219   -1.660647e-05 -6.568463e-07    7.843378e-07   -4.440512e-04
## 206 -0.025385879   -2.739464e-04 -1.432882e-06    1.730647e-08    7.522162e-04
## 404  0.014032237   -9.393894e-05 -5.381388e-07    6.204144e-07   -5.561463e-06
##     PercentUnemploy PerCapitaIncome
## 2     -0.0016946189    6.916107e-06
## 6      0.0045462952    5.236780e-06
## 8     -0.0010223574   -1.058692e-06
## 48    -0.0004521879    1.806161e-07
## 128   -0.0004659840   -2.251551e-07
## 206   -0.0002989920    1.635733e-06
## 404   -0.0013745771   -1.660623e-07
```

Cook's Distance: Measure of influence on all fitted values, given by

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j,-1})^2}{(p+1)MSE} = \frac{e_i^2}{(p+1)MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

D_i is related to $F_{p+1,n-p-1}$

ROT: Alert when cdf of $F_{p+1,n-p-1}$ is larger than 50%

```r
# Cooks Distance
cooks.distance(model.3)[cases]
```

```
##           2           6           8          48         128         206 
## 1.467448313 2.634135550 0.051774880 0.014199746 0.001339249 0.005880134 
##         404 
## 0.002035972 
```

Now applying all the decision rules we get

```r
dffits(model.3)[which( abs(dffits(model.3)) > 1 )]
```

```
##         2         6 
## -3.272969 10.289758 
```

```r
dfbeta(model.3)[which( abs(dfbeta(model.3)) > min(1,2/sqrt(nrow(train_CDI3))) ),]
```

```
##   (Intercept) PercentPopYoung NumPhysicians NumHospitalBeds PercentBelowPov
## 2  -0.1705471    0.0019411631 -2.198755e-05   -2.749610e-05     0.005537637
## 6  -0.2259768    0.0005618961 -8.280215e-05    9.602833e-05     0.005437026
##   PercentUnemploy PerCapitaIncome
## 2    -0.001694619    6.916107e-06
## 6     0.004546295    5.236780e-06
```

Thus we can conclude that observations 2 and 6 are outliers.