HW4 R Markdown

Santanu Mukherjee

10/12/2021

R Markdown

Problem 1

Reading the file *"CDI_data.rda"*

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = df1)
##
## 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 **
## x1           5.515e-04  2.835e-04   1.945  0.05243 .
## x2           1.070e-01  1.325e-02   8.073  6.8e-15 ***
## x31          1.490e+02  8.683e+01   1.716  0.08685 .
## x41          1.455e+02  8.515e+01   1.709  0.08817 .
## x51          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
```

Answer 1a

If our coefficients were β_3, β_4, β_5 for x_3 , x_4 and x_5, then the null hypothesis is H_0:β_i=0 for all i = 3,4,5

Based on the linear regression output data. we see that using a significance level of $\alpha = 0.05$, we can reject the null hypothesis as one of the p=values for one of the β values are significant. In other words, geographic effects are present for this response variable.

```
##
## Call:
## lm(formula = y ~ x3 + x4 + x5, data = df1)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1292.9  -733.6  -552.8    80.3 22323.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1353.9      203.4   6.657 8.43e-11 ***
## x31           -259.5      268.9  -0.965   0.3351
## x41           -493.5      266.2  -1.854   0.0644 .
## x51           -532.8      249.6  -2.134   0.0334 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1785 on 436 degrees of freedom
## Multiple R-squared:  0.01242,    Adjusted R-squared:  0.005627
## F-statistic: 1.828 on 3 and 436 DF,  p-value: 0.1413
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1849.1  -198.3   -71.4    39.7  3755.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.444e+01  3.283e+01  -1.963   0.0503 .
## x1           5.310e-04  2.775e-04   1.914   0.0563 .
## x2           1.072e-01  1.297e-02   8.269 1.64e-15 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 568 on 437 degrees of freedom
## Multiple R-squared:  0.8998, Adjusted R-squared:  0.8993
## F-statistic:  1961 on 2 and 437 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2
## Model 2: y ~ x1 + x2 + x3 + x4 + x5
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    437 140967081
## 2    434 139093455  3   1873626 1.9487  0.121
```

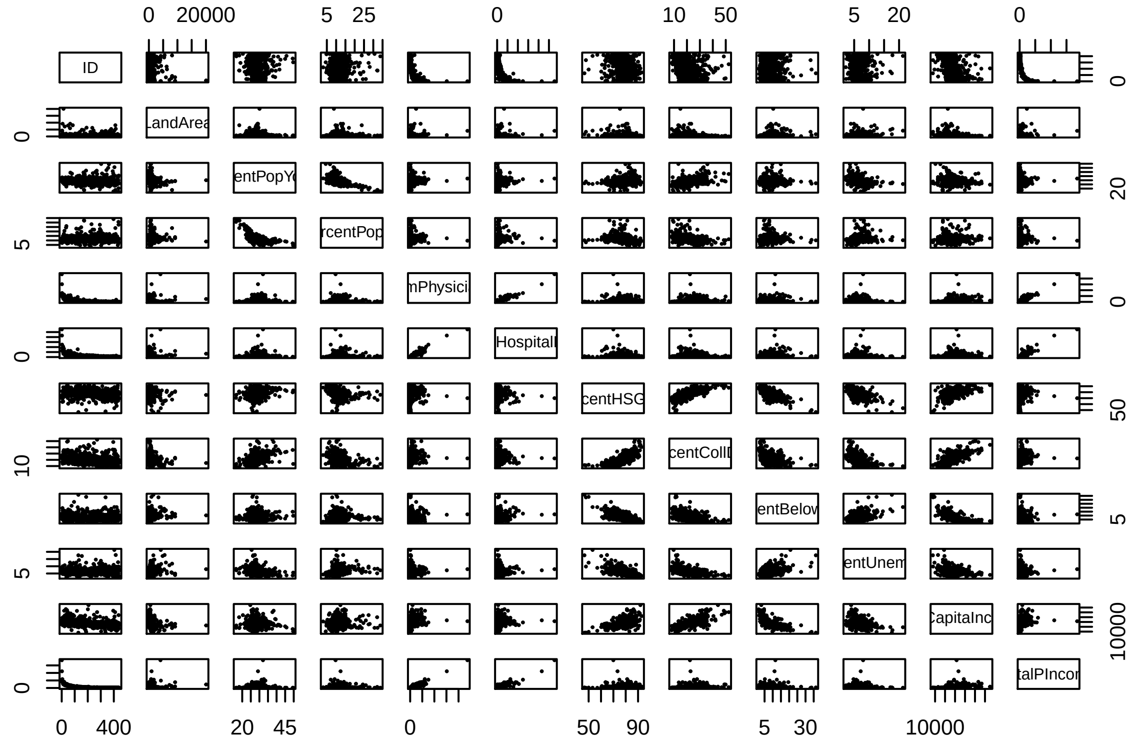
Answer 1b

Based on the results of the ANOVA, it can be said that one or more geographic factors are significant.

Problem 2

Answer 2a

Scatter Plot for Problem 2 (a) AND Correlation Matrix *(rounded to nearest 3 decimal places)* **for Problem 2 (a) below**



```
##                 ID     LandArea PercentPopYoung PercentPop65 NumPhysicians
## ID              "High" ""       ""              ""           ""
## LandArea        ""     "High"   ""              ""           ""
## PercentPopYoung ""     ""       "High"          ""           ""
## PercentPop65    ""     ""       ""              "High"       ""
## NumPhysicians   ""     ""       ""              ""           "High"
## NumHospitalBeds ""     ""       ""              ""           "High"
## PercentHSGrad   ""     ""       ""              ""           ""
## PercentCollDeg  ""     ""       ""              ""           ""
## PercentBelowPov ""     ""       ""              ""           ""
## PercentUnemploy ""     ""       ""              ""           ""
## PerCapitaIncome ""     ""       ""              ""           ""
## TotalPIncome    ""     ""       ""              ""           "High"
##                 NumHospitalBeds PercentHSGrad PercentCollDeg PercentBelowPov
## ID              ""              ""            ""             ""
## LandArea        ""              ""            ""             ""
## PercentPopYoung ""              ""            ""             ""
## PercentPop65    ""              ""            ""             ""
## NumPhysicians   "High"          ""            ""             ""
## NumHospitalBeds "High"          ""            ""             ""
## PercentHSGrad   ""              "High"        ""             ""
## PercentCollDeg  ""              ""            "High"         ""
## PercentBelowPov ""              ""            ""             "High"
## PercentUnemploy ""              ""            ""             ""
## PerCapitaIncome ""              ""            ""             ""
## TotalPIncome    "High"          ""            ""             ""
##                 PercentUnemploy PerCapitaIncome TotalPIncome
## ID              ""              ""              ""
## LandArea        ""              ""              ""
## PercentPopYoung ""              ""              ""
## PercentPop65    ""              ""              ""
## NumPhysicians   ""              ""              "High"
## NumHospitalBeds ""              ""              "High"
## PercentHSGrad   ""              ""              ""
## PercentCollDeg  ""              ""              ""
## PercentBelowPov ""              ""              ""
## PercentUnemploy "High"          ""              ""
## PerCapitaIncome ""              "High"          ""
## TotalPIncome    ""              ""              "High"
```

Yes there is evidence of strong linear pairwise associations among some predictor variables (NumPhysicians, TotalPIncome) and (NumHospitalBeds , TotalPIncome)

Answer 2b

There are several model selection techniques like Forward Selection and Backward Elimination.Using sequential selection process

```r
lmp2b.all <- lm(y ~ ., data = train)
lmp2b.f = ols_step_forward_p(lmp2b.all, penter = 0.05)
lmp2b.b = ols_step_backward_p(lmp2b.all, prem = 0.1)
lmp2b.s = ols_step_both_p(lmp2b.all, pent = 0.05, prem = 0.1)

summary(lmp2b.all)
```

```
## 
## Call:
## lm(formula = y ~ ., data = train)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5344.2 -1145.1  -173.1   951.8 16715.2 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      621.61239 5199.26494   0.120 0.904949    
## ID                -5.53348    1.53904  -3.595 0.000405 ***
## LandArea          -0.02867    0.09473  -0.303 0.762464    
## PercentPopYoung  131.77241   60.61873   2.174 0.030855 *  
## PercentPop65       1.93557   48.86136   0.040 0.968439    
## NumPhysicians     -0.48427    0.35865  -1.350 0.178414    
## NumHospitalBeds    0.90002    0.24454   3.680 0.000297 ***
## PercentHSGrad    -13.63472   45.46541  -0.300 0.764560    
## PercentCollDeg    -9.71262   51.25681  -0.189 0.849895    
## PercentBelowPov  315.93128   63.80595   4.951 1.53e-06 ***
## PercentUnemploy -232.99628   90.03896  -2.588 0.010345 *  
## PerCapitaIncome    0.11413    0.08387   1.361 0.175029    
## TotalPIncome      -0.05135    0.04644  -1.106 0.270102    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 2117 on 207 degrees of freedom
## Multiple R-squared:  0.5192, Adjusted R-squared:  0.4913
## F-statistic: 18.62 on 12 and 207 DF,  p-value: < 2.2e-16
```

```
summary(lmp2b.f$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5697.2 -1156.8  -127.2  1050.3 17196.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2460.47116 1374.37596   1.790  0.07483 .
## NumHospitalBeds    0.74811    0.18252   4.099 5.90e-05 ***
## PercentBelowPov  270.90562   37.56727   7.211 9.54e-12 ***
## ID                -6.21781    1.46942  -4.231 3.45e-05 ***
## PercentUnemploy -203.31810   75.33827  -2.699  0.00752 **
## PercentPopYoung  109.20657   39.88602   2.738  0.00671 **
## TotalPIncome      -0.08155    0.03581  -2.277  0.02378 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2113 on 213 degrees of freedom
## Multiple R-squared:  0.5068, Adjusted R-squared:  0.4929
## F-statistic: 36.48 on 6 and 213 DF,  p-value: < 2.2e-16
```

```
summary(lmp2b.b$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5856.2 -1073.3  -176.3   962.2 16583.3
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -648.20276 1895.72260  -0.342  0.73274
## ID                -4.91120    1.44816  -3.391  0.00083 ***
## PercentPopYoung  124.10404   40.14828   3.091  0.00226 **
## NumPhysicians     -0.76359    0.28269  -2.701  0.00747 **
## NumHospitalBeds    0.86825    0.20429   4.250 3.20e-05 ***
## PercentBelowPov  339.19772   42.19184   8.039 6.27e-14 ***
## PercentUnemploy -230.45089   74.01000  -3.114  0.00210 **
## PerCapitaIncome    0.10140    0.05204   1.948  0.05268 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2101 on 212 degrees of freedom
## Multiple R-squared:  0.5146, Adjusted R-squared:  0.4986
## F-statistic: 32.11 on 7 and 212 DF,  p-value: < 2.2e-16
```

```
summary(lmp2b.s$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5697.2 -1156.8  -127.2  1050.3 17196.6
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2460.47116 1374.37596   1.790  0.07483 .
## NumHospitalBeds    0.74811    0.18252   4.099 5.90e-05 ***
## PercentBelowPov  270.90562   37.56727   7.211 9.54e-12 ***
## ID                -6.21781    1.46942  -4.231 3.45e-05 ***
## PercentUnemploy -203.31810   75.33827  -2.699  0.00752 **
## PercentPopYoung  109.20657   39.88602   2.738  0.00671 **
## TotalPIncome      -0.08155    0.03581  -2.277  0.02378 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2113 on 213 degrees of freedom
## Multiple R-squared:  0.5068, Adjusted R-squared:  0.4929
## F-statistic: 36.48 on 6 and 213 DF,  p-value: < 2.2e-16
```

Based on the models above, the attractive one to me is the first one (Selection Summary)

Answer 2c

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4498.4 -1306.1  -195.9  1231.6  6313.7
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3328.752    740.873   4.493 1.14e-05 ***
## PercentBelowPov  321.933     34.123   9.435  < 2e-16 ***
## ID                -6.548      1.138  -5.754 2.97e-08 ***
## PercentCollDeg    56.333     21.767   2.588   0.0103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1965 on 216 degrees of freedom
## Multiple R-squared:  0.3805, Adjusted R-squared:  0.3719
## F-statistic: 44.23 on 3 and 216 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5697.2 -1156.8  -127.2  1050.3 17196.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2460.47116 1374.37596   1.790  0.07483 .
## NumHospitalBeds    0.74811    0.18252   4.099 5.90e-05 ***
## PercentBelowPov  270.90562   37.56727   7.211 9.54e-12 ***
## ID                -6.21781    1.46942  -4.231 3.45e-05 ***
## PercentUnemploy -203.31810   75.33827  -2.699  0.00752 **
## PercentPopYoung  109.20657   39.88602   2.738  0.00671 **
## TotalPIncome      -0.08155    0.03581  -2.277  0.02378 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2113 on 213 degrees of freedom
## Multiple R-squared:  0.5068, Adjusted R-squared:  0.4929
## F-statistic: 36.48 on 6 and 213 DF,  p-value: < 2.2e-16
```

```
## [1] 4323641
```

```
## [1] 3792582
```

Based on the regression results, the model fitted to the test data set does not yield similar estimates as the model fitted to the model-building data set.

Answer 2d

```
## [1] "The MSE for training data set is : 4215612.91117253"
```

```
## [1] "The MSE for test data set is : 8529195.12819667"
```

So, the test MSE (**MSEte**) is greater than the training MSE (**MSEtr**). This can happen because of outliers. There is evidence of Bias here in the model.

Answer 2e

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5595.4 -1196.5  -130.5  1199.4 18287.4
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1687.59133 1302.79426   1.295  0.19589
## PercentBelowPov  291.01179   31.78758   9.155  < 2e-16 ***
## ID                -6.44314    1.01307  -6.360 5.14e-10 ***
## PercentUnemploy -123.39038   50.08794  -2.463  0.01415 *
## PercentPopYoung   78.76746   24.88860   3.165  0.00166 **
## NumHospitalBeds    0.54473    0.11583   4.703 3.46e-06 ***
## TotalPIncome      -0.08143    0.02034  -4.003 7.38e-05 ***
## PerCapitaIncome    0.07186    0.03557   2.020  0.04398 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2047 on 432 degrees of freedom
## Multiple R-squared:  0.4477, Adjusted R-squared:  0.4388
## F-statistic: 50.03 on 7 and 432 DF,  p-value: < 2.2e-16
```

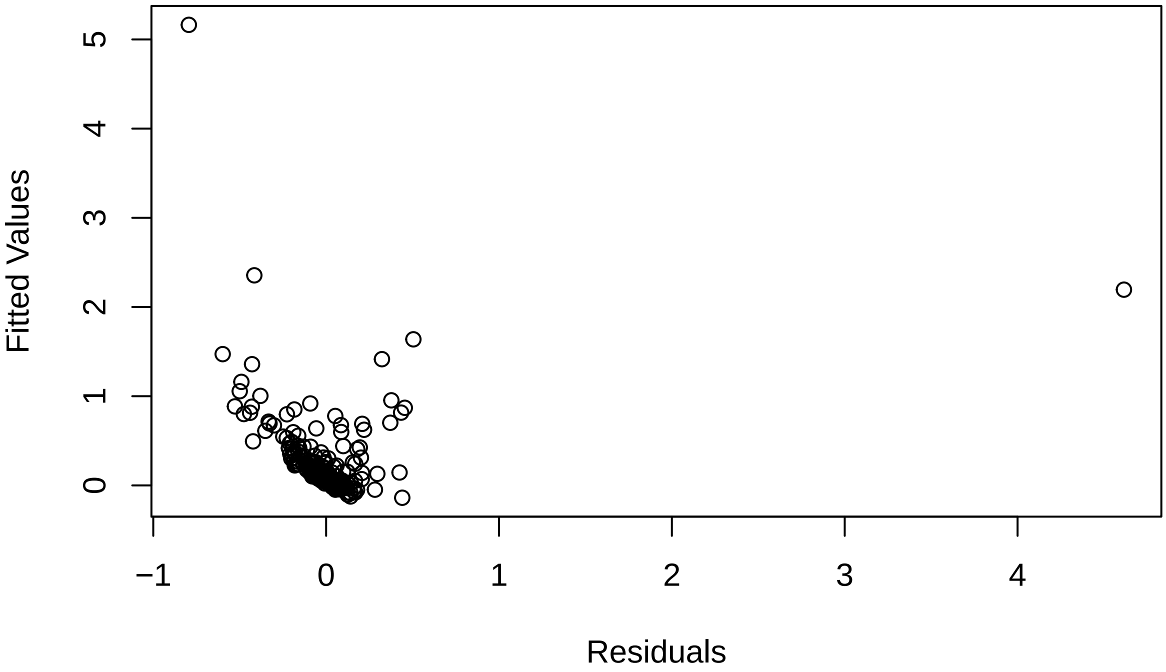
```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5697.2 -1156.8  -127.2  1050.3 17196.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2460.47116 1374.37596   1.790  0.07483 .
## NumHospitalBeds    0.74811    0.18252   4.099 5.90e-05 ***
## PercentBelowPov  270.90562   37.56727   7.211 9.54e-12 ***
## ID                -6.21781    1.46942  -4.231 3.45e-05 ***
## PercentUnemploy -203.31810   75.33827  -2.699  0.00752 **
## PercentPopYoung  109.20657   39.88602   2.738  0.00671 **
## TotalPIncome      -0.08155    0.03581  -2.277  0.02378 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2113 on 213 degrees of freedom
## Multiple R-squared:  0.5068, Adjusted R-squared:  0.4929
## F-statistic: 36.48 on 6 and 213 DF,  p-value: < 2.2e-16
```

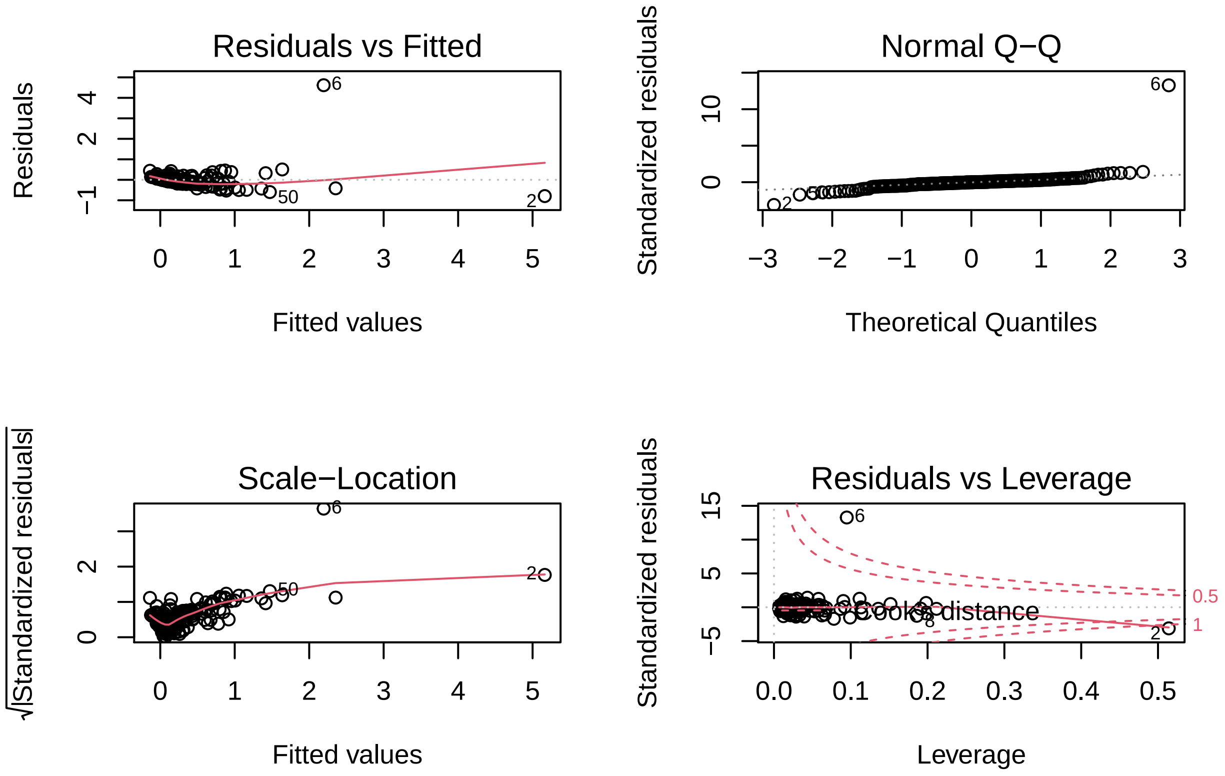
The estimated coefficients and their standard deviations for the combined training and test data sets is appreciably different from the estimated coefficients and their standard deviations of the model fitted to the training data set. Ideally I would expect as there are many factors here, the 2 datasets have different total number of rows, number of predictors.

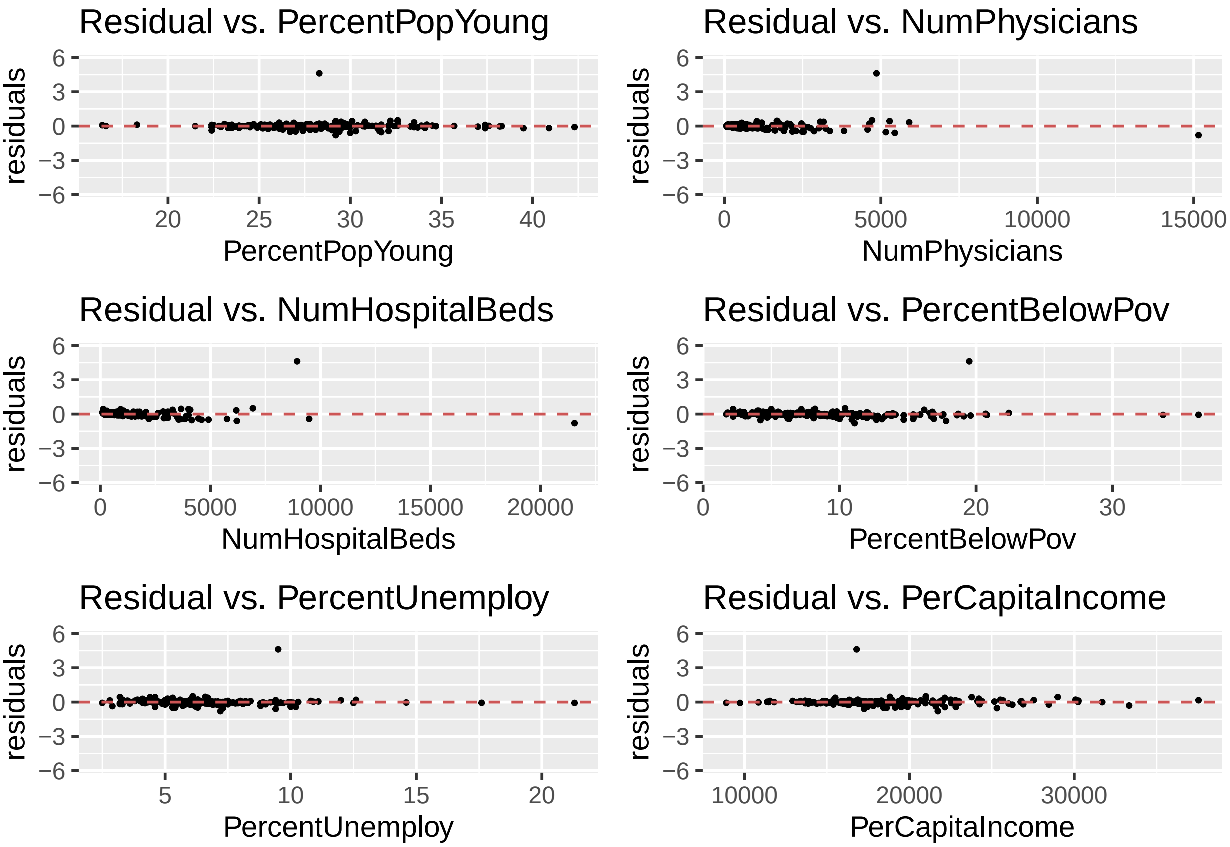
Problem 3

Answer 3a

```
##
## Call:
## lm(formula = y ~ PercentPopYoung + NumPhysicians + NumHospitalBeds +
##     PercentBelowPov + PercentUnemploy + PerCapitaIncome, data = df3train)
##
## 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
```

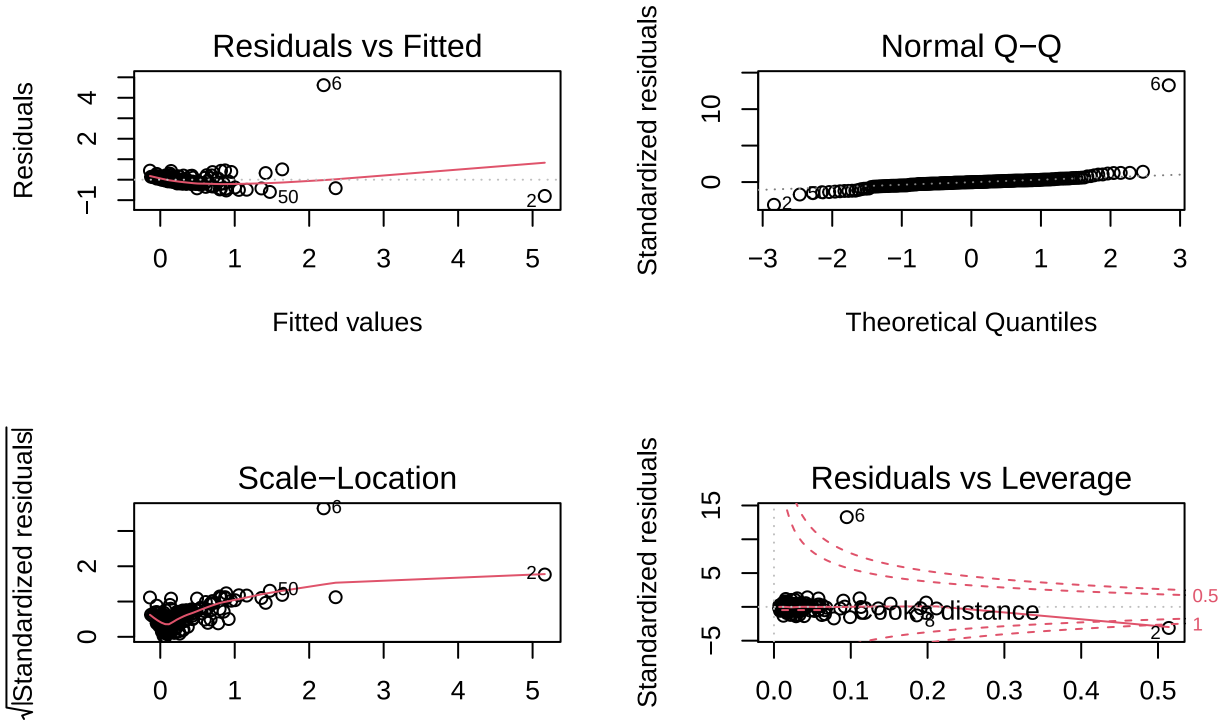

Fitted Values
0
1
2
3
4
5
−1
0
1
2
3
4
Residuals


Residuals vs Fitted
Residuals
4
2
−1
0
1
2
3
4
5
Fitted values
6
50
2
Normal Q−Q
Standardized residuals
10
0
−3
−2
−1
0
1
2
3
Theoretical Quantiles
6
2
5
Scale−Location
√|Standardized residuals|
2
0
0
1
2
3
4
5
Fitted values
6
50
2
Residuals vs Leverage
Standardized residuals
15
5
−5
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
6
Cook's distance
8
2
0.5
1



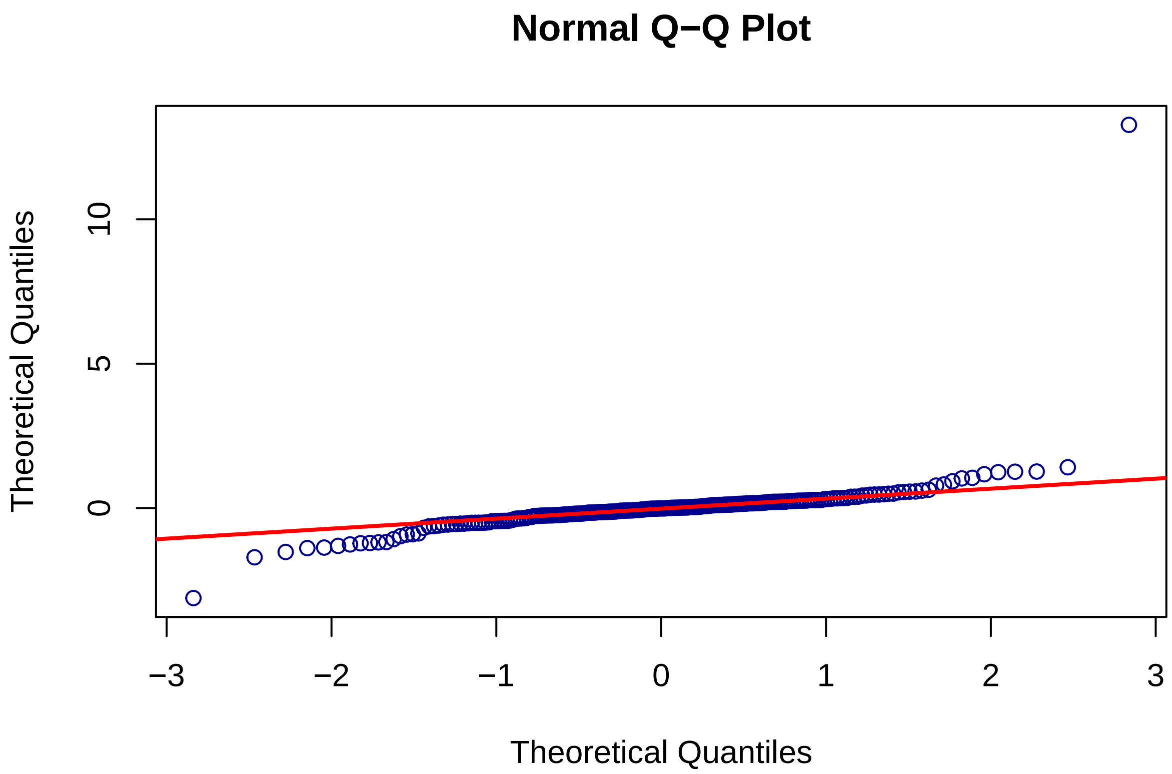
It is apparent that there are some outliers which are affecting the regression model. In addition to that, certain pattern cannot be explained in those residual plots.

Answer 3b



```
##
## Call:
## lm(formula = y ~ PercentPopYoung + NumPhysicians + NumHospitalBeds +
##     PercentBelowPov + PercentUnemploy + PerCapitaIncome, data = df3train)
##
## 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
```

```
##  Named num [1:220] -3.116 0.926 13.27 -1.26 1.412 ...
##  - attr(*, "names")= chr [1:220] "2" "4" "6" "8" ...
```



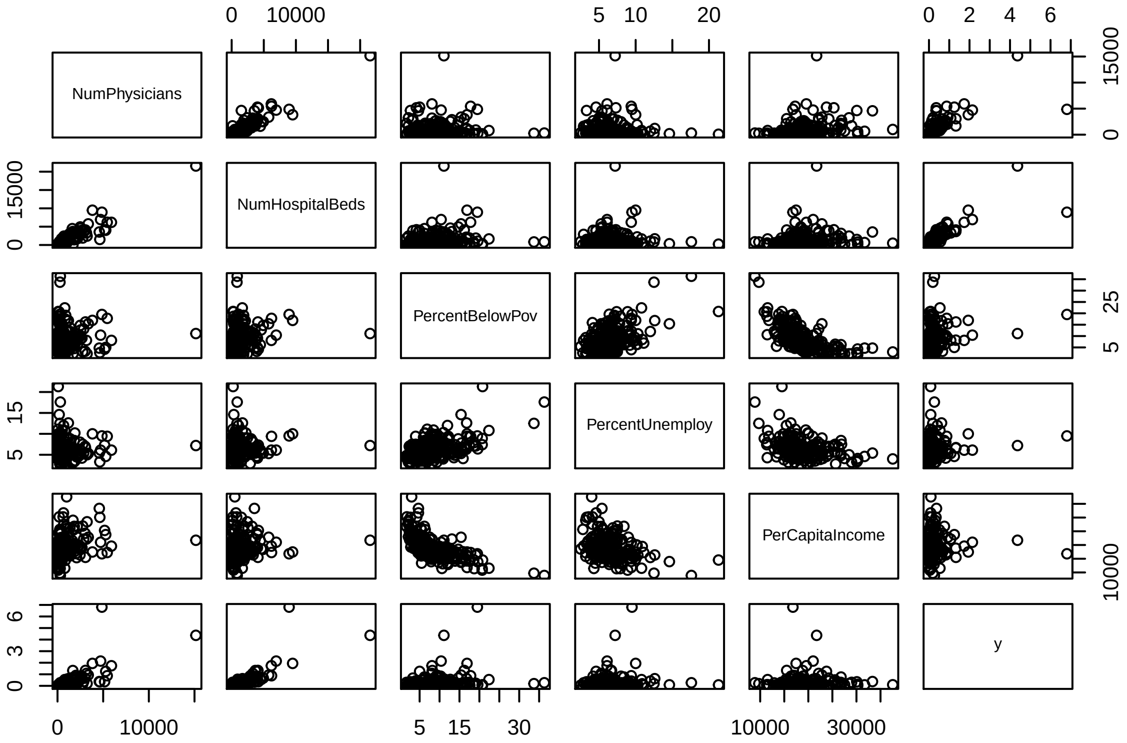
```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.4197         0.0000
## Kolmogorov-Smirnov        0.2372         0.0000
## Cramer-von Mises         54.9285       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

There are certain serious departure from normality in Q-Q Plot.
The normality test results in non-normality.

Answer 3c

Scatter Plot for Problem 3 (c) AND Correlation Matrix *(rounded to nearest 3 decimal places)* **for Problem 3 (c) below**

```
pairs(df3train[,-1])
```



```
corMatrix <- cor(subset(df3train, select=-c(y)))
corMatrix
```

```
##                  PercentPopYoung NumPhysicians NumHospitalBeds PercentBelowPov
## PercentPopYoung       1.00000000    0.17350201      0.11285330     -0.02532508
## NumPhysicians         0.17350201    1.00000000      0.92105026      0.03272668
## NumHospitalBeds       0.11285330    0.92105026      1.00000000      0.17297166
## PercentBelowPov      -0.02532508    0.03272668      0.17297166      1.00000000
## PercentUnemploy      -0.22212912   -0.05639102      0.03401543      0.51273477
## PerCapitaIncome       0.07819869    0.31211999      0.14487272     -0.64891566
##                  PercentUnemploy PerCapitaIncome
## PercentPopYoung      -0.22212912      0.07819869
## NumPhysicians        -0.05639102      0.31211999
## NumHospitalBeds       0.03401543      0.14487272
## PercentBelowPov       0.51273477     -0.64891566
## PercentUnemploy       1.00000000     -0.37887001
## PerCapitaIncome      -0.37887001      1.00000000
```

```
corMatrix[corMatrix >=.75] <- 'High'
corMatrix[corMatrix <.75] <- ''
corMatrix
```

```
##                  PercentPopYoung NumPhysicians NumHospitalBeds PercentBelowPov
## PercentPopYoung "High"          ""            ""              ""
## NumPhysicians   ""              "High"        "High"          ""
## NumHospitalBeds ""              "High"        "High"          ""
## PercentBelowPov ""              ""            ""              "High"
## PercentUnemploy ""              ""            ""              ""
## PerCapitaIncome ""              ""            ""              ""
##                 PercentUnemploy PerCapitaIncome
## PercentPopYoung ""              ""
## NumPhysicians   ""              ""
## NumHospitalBeds ""              ""
## PercentBelowPov ""              ""
## PercentUnemploy "High"          ""
## PerCapitaIncome ""              "High"
```

```
vif(lmp3a)
```

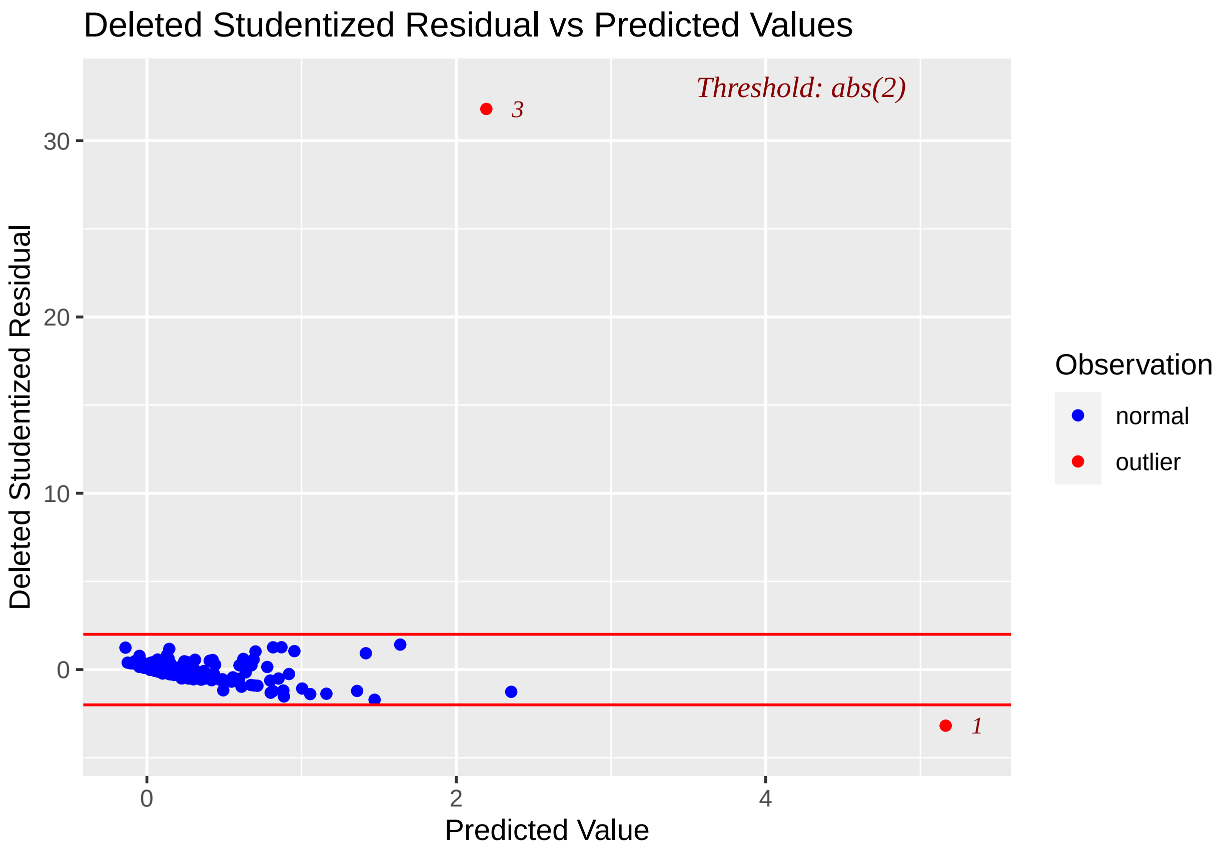
```
## PercentPopYoung   NumPhysicians NumHospitalBeds PercentBelowPov PercentUnemploy
##        1.106320        8.654011        8.038180        2.307348        1.447219
## PerCapitaIncome
##        2.226508
```

The VIF results does not indicate that there is multicollinearity effect.

Answer 3d

Scatter Plot for Problem 3 (d) - Studentized deleted residuals and a dot plot of these residuals

```
ols_plot_resid_stud_fit(lmp3a)
```



```
out1 <- outlierTest(lmp3a, cutoff=1)
out1
```

```
##    rstudent unadjusted p-value Bonferroni p
## 6 31.797435         1.2609e-82   2.7739e-80
## 2 -3.182395         1.6800e-03   3.6960e-01
```

Yes, according to the plot, the 1st observation and 3rd observation are outliers, which means # 2 and # 6 in the dataset are outliers.

Answer 3e

Diagonal elements for the HAT matrix

```
##          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
```

```
## [1] "The number of the elements of the HAT Matrix : 19"
```

Answer 3f

Check for Outliers

```
## [1] "The outliers produced by DFFITIS :"
```

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

```
## [1] "The outliers produced by Cooks Distance :"
```

```
##        2        6
## 1.467448 2.634136
```

```
## [1] "The outliers produced by DFBETAS :"
```

```
## [1] "2" "6"
```

```
## [1] -0.1705471 -0.2259768
```

Here I see that DFFITIS, DFBETAS and Cooks Distance - all of them says 2 and 6 are outliers.