P8130 Final Project

Abstract

Introduction (brief context and background of the problem)

Methods (data description and statistical methods)

Results

Conclusions/Discussion

```
library(tidyverse)
library(ggplot2)
library(GGally)
library(PerformanceAnalytics)
library(performance)
library(MASS)
```

Read in dataset

```
cdi = read_csv("./cdi.csv") %>%
  janitor::clean_names()
```

```
## no missing value
cdi %>%
  dplyr::select(everything()) %>%
  summarise_all(funs(sum(is.na(.)))) %>%
  knitr::kable()
```

id	cty	state	area	pop	pop18	pop65	docs	beds	crimes	hsgrad	bagrad	poverty	unemp	pcincome	totalinc	region
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Data cleaning

```
# some normalization for better comparison
cdi =
  cdi %>%
  mutate(crm_1000 = crimes/pop*1000,  # as indicated by the project prompt
         docs_1000 = docs/pop*1000,  # every 1000 people how many doctors
         beds_1000 = beds/pop*1000,
         pop_density = pop/area,  # how many people per square miles
         northeast = ifelse(region==1, 1, 0),
         northcentral = ifelse(region==2, 1, 0),
         south = ifelse(region==3, 1, 0)) %>%
  dplyr::select(-id, -crimes,-area, -docs, -beds, -totalinc, -region)
```

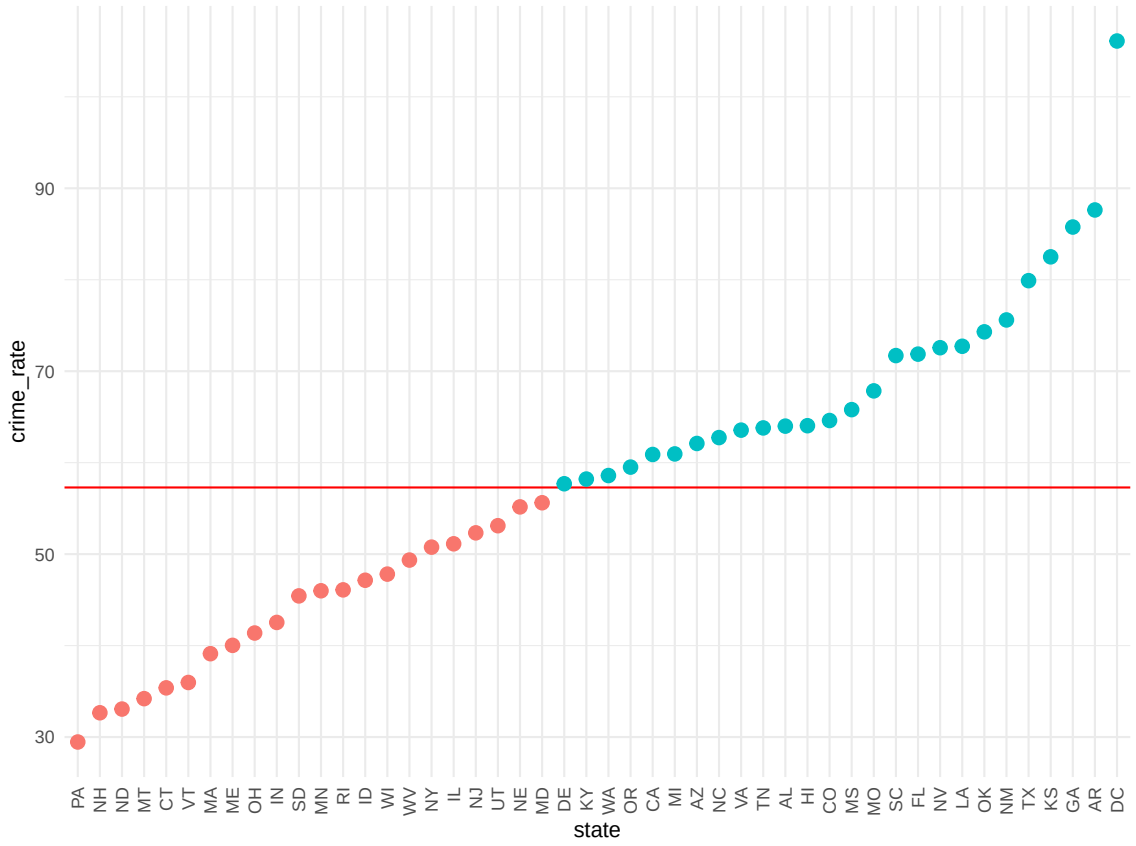
Data Exploration

```
## summary statistics, tentative, NOT FINAL
sum_cdi =
  cdi %>%
  dplyr::select(-c(cty, state))
summary(sum_cdi)
```

```
##       pop              pop18           pop65            hsgrad
##  Min.   : 100043   Min.   :16.40   Min.   : 3.000   Min.   :46.60
##  1st Qu.: 139027   1st Qu.:26.20   1st Qu.: 9.875   1st Qu.:73.88
##  Median : 217280   Median :28.10   Median :11.750   Median :77.70
##  Mean   : 393011   Mean   :28.57   Mean   :12.170   Mean   :77.56
##  3rd Qu.: 436064   3rd Qu.:30.02   3rd Qu.:13.625   3rd Qu.:82.40
##  Max.   :8863164   Max.   :49.70   Max.   :33.800   Max.   :92.90
##      bagrad          poverty           unemp           pcincome
##  Min.   : 8.10   Min.   : 1.400   Min.   : 2.200   Min.   : 8899
##  1st Qu.:15.28   1st Qu.: 5.300   1st Qu.: 5.100   1st Qu.:16118
##  Median :19.70   Median : 7.900   Median : 6.200   Median :17759
##  Mean   :21.08   Mean   : 8.721   Mean   : 6.597   Mean   :18561
##  3rd Qu.:25.32   3rd Qu.:10.900   3rd Qu.: 7.500   3rd Qu.:20270
##  Max.   :52.30   Max.   :36.300   Max.   :21.300   Max.   :37541
##     crm_1000         docs_1000         beds_1000        pop_density
##  Min.   :  4.601   Min.   : 0.3559   Min.   : 0.1649   Min.   :   13.26
##  1st Qu.: 38.102   1st Qu.: 1.2127   1st Qu.: 2.1972   1st Qu.:  192.34
##  Median : 52.429   Median : 1.7509   Median : 3.3287   Median :  335.91
##  Mean   : 57.286   Mean   : 2.1230   Mean   : 3.6493   Mean   :  888.44
##  3rd Qu.: 72.597   3rd Qu.: 2.4915   3rd Qu.: 4.5649   3rd Qu.:  756.55
##  Max.   :295.987   Max.   :17.0377   Max.   :19.6982   Max.   :32403.72
##    northeast       northcentral        south
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.2341   Mean   :0.2455   Mean   :0.3455
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

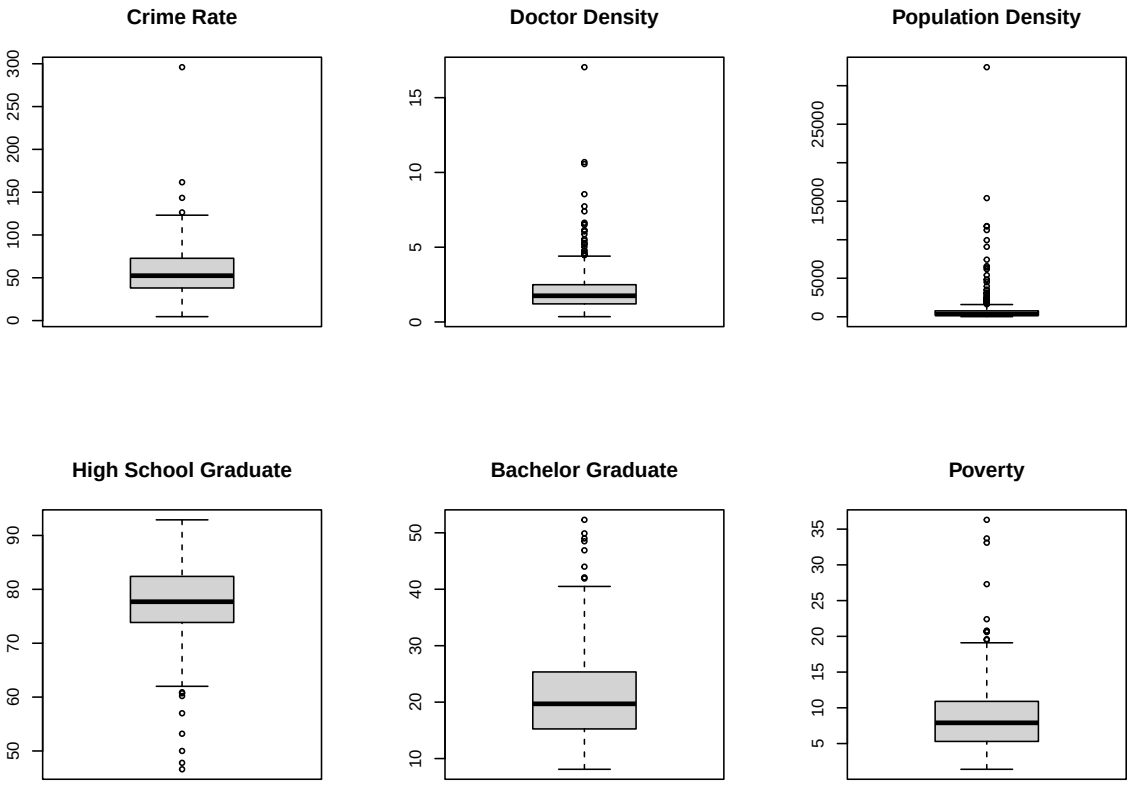
```
mean_crm = mean(sum_cdi$crm_1000)
cdi_state = cdi %>%
  group_by(state) %>%
  summarize(crime_rate = mean(crm_1000)) %>%
  mutate(low_high = ifelse(crime_rate>mean_crm, TRUE,FALSE))


cdi_state %>%
  mutate(state = fct_reorder(state, crime_rate)) %>%
  ggplot(aes(x = state, y = crime_rate))+
  geom_hline(yintercept = mean_crm, color = "red")+
  geom_point(aes(color = low_high),size = 3)+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust= 1),
        legend.position = "none")
```

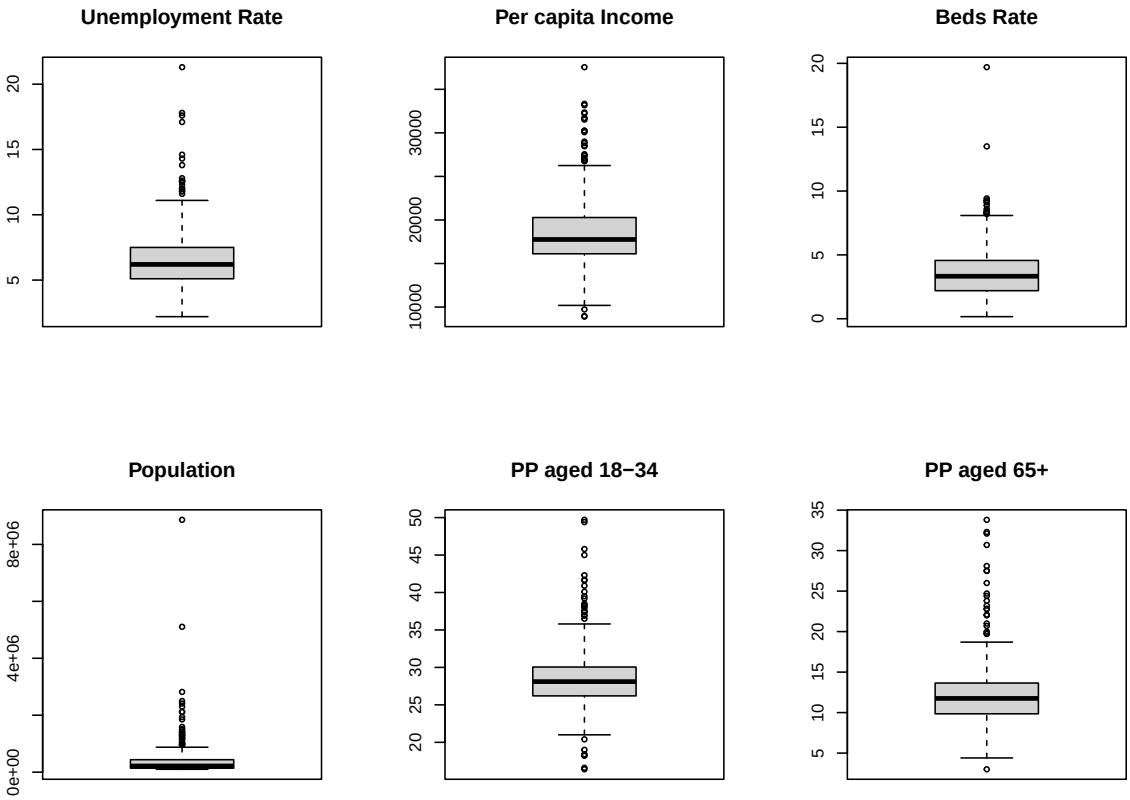


boxplot for each variable

```
par(mfrow=c(2,3))
boxplot(sum_cdi$crm_1000, main='Crime Rate')
boxplot(sum_cdi$docs_1000, main='Doctor Density')
boxplot(sum_cdi$pop_density,main='Population Density' )
boxplot(sum_cdi$hsgrad, main='High School Graduate')
boxplot(sum_cdi$bagrad, main='Bachelor Graduate')
boxplot(sum_cdi$poverty, main='Poverty')
```

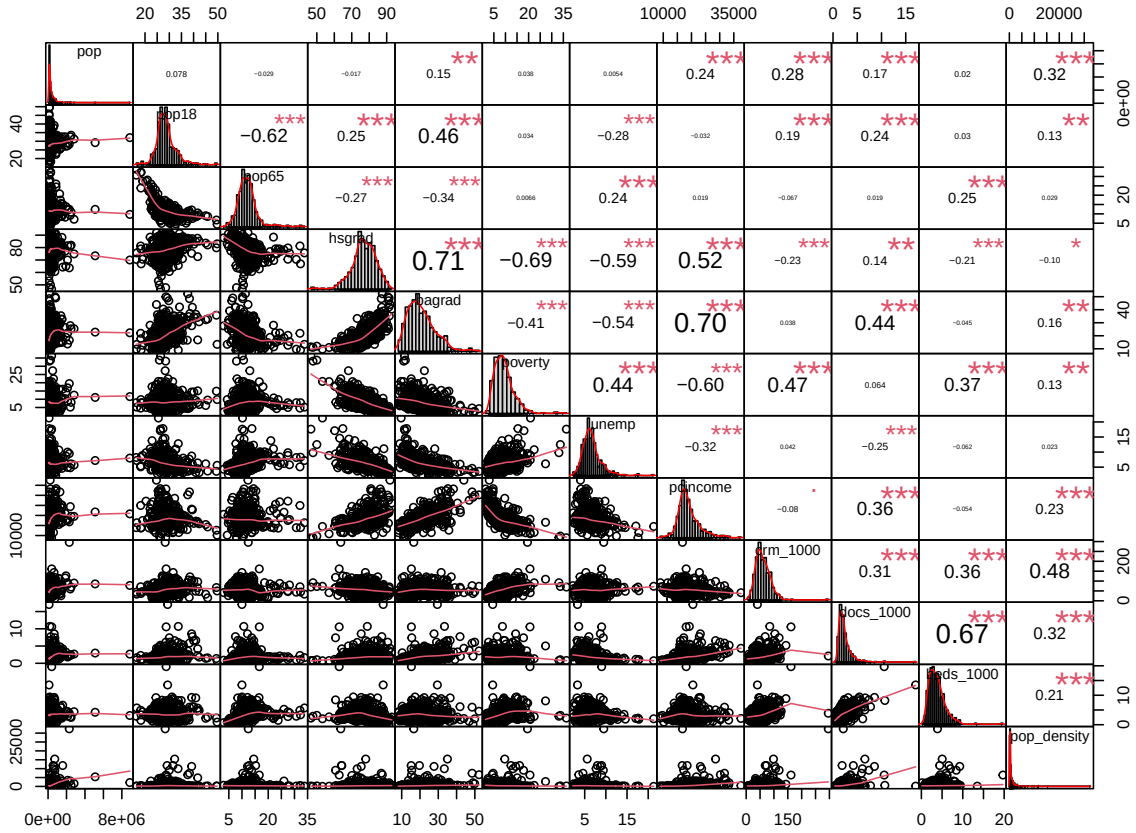



```
par(mfrow=c(2,3))
boxplot(sum_cdi$unemp, main='Unemployment Rate')
boxplot(sum_cdi$pcincome, main='Per capita Income')
boxplot(sum_cdi$beds_1000, main='Beds Rate')
boxplot(sum_cdi$pop, main='Population')
boxplot(sum_cdi$pop18, main='PP aged 18-34')
boxplot(sum_cdi$pop65, main='PP aged 65+')
```



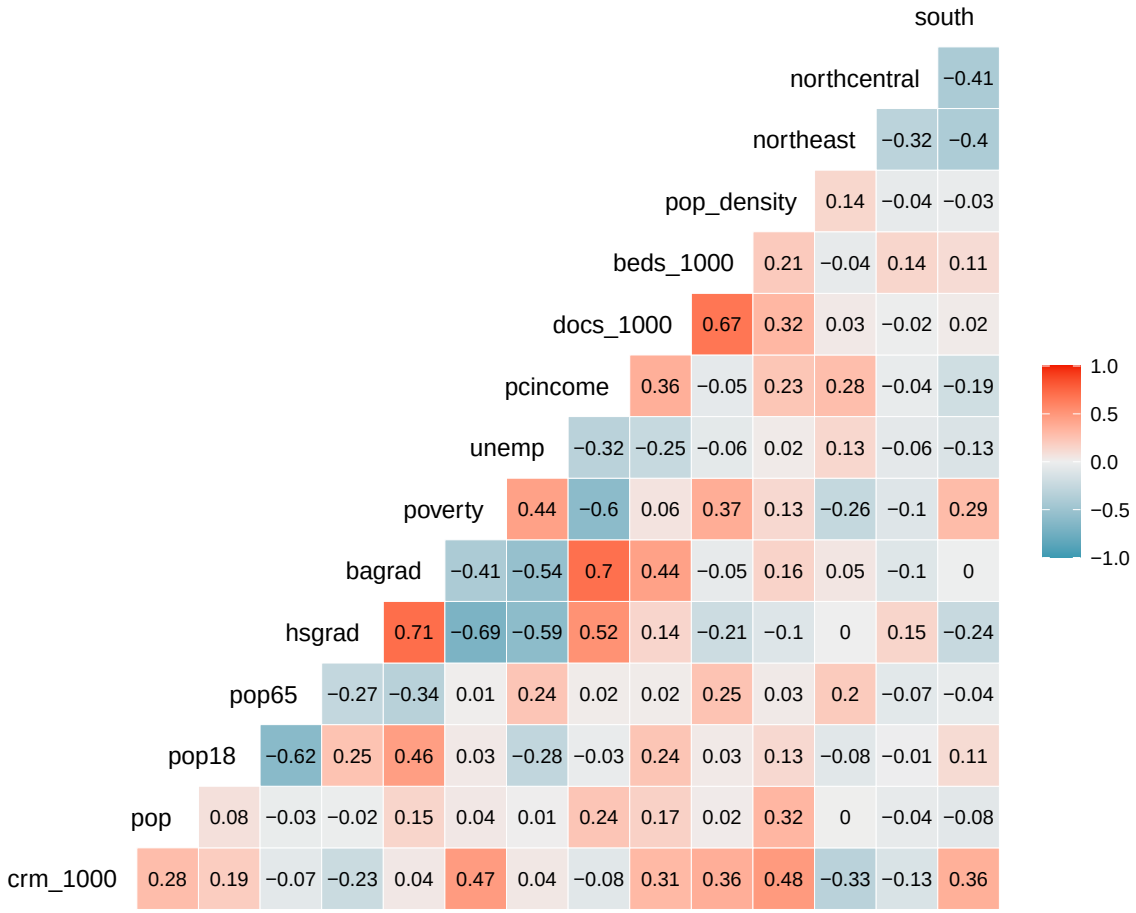
Marginal Correlation and Correlation martix

```
corr_matrix =
  cdi %>%
  dplyr::select(-state, -cty, -northeast, -northcentral, -south) %>%
  #sum_cdi %>%
  chart.Correlation(histogram = TRUE, method = "pearson")
```



Correlation Heatmap

```
cdi %>%
  dplyr::select(-state, -cty) %>%
  dplyr::select(crm_1000, everything()) %>%
  ggcorr(label=TRUE, hjust = 0.9, layout.exp = 2, label_size = 3, label_round = 2)
```




```
#corrplot(cor(cdi_1), type = "upper", diag = FALSE)
```

Build Model

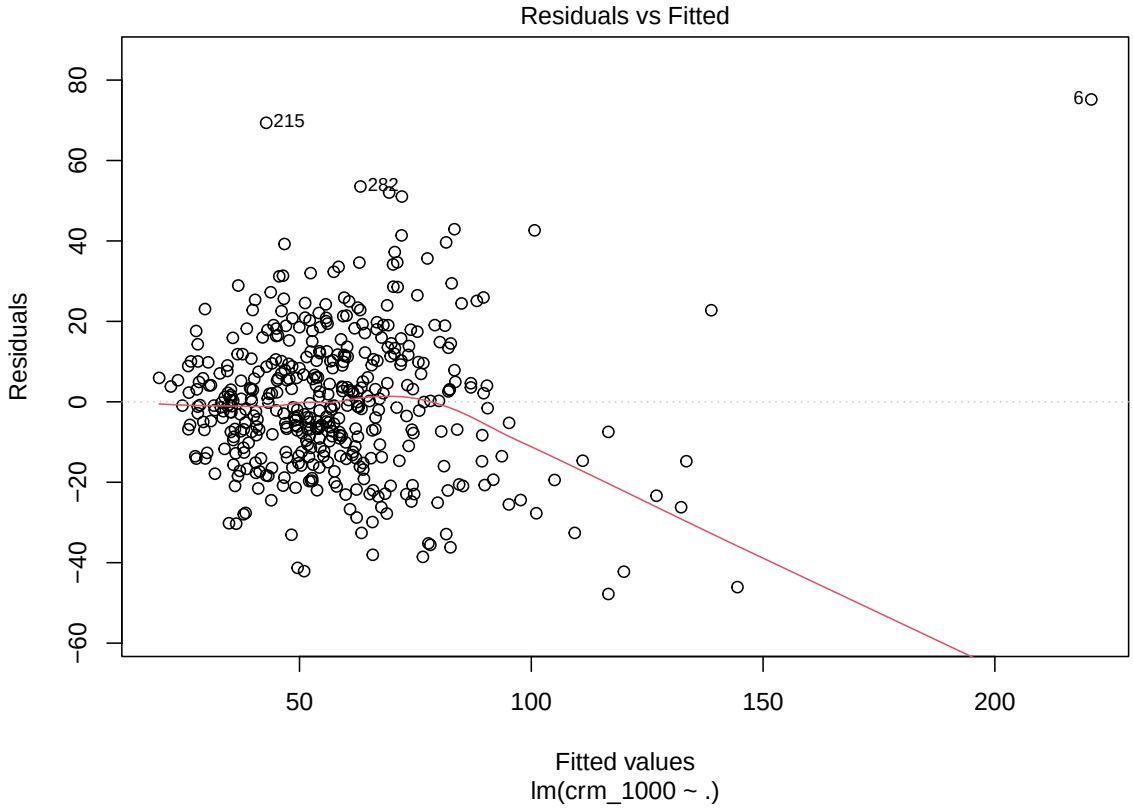
```
mult_fit = lm(crm_1000 ~ ., data = sum_cdi)
summary(mult_fit)
```

```
##
## Call:
## lm(formula = crm_1000 ~ ., data = sum_cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.786 -11.422  -0.934  10.200  75.180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.805e+01  2.770e+01  -1.734 0.083592 .
## pop           5.486e-06  1.579e-06   3.474 0.000566 ***
## pop18         6.947e-01  3.305e-01   2.102 0.036150 *
## pop65        -1.998e-01  3.055e-01  -0.654 0.513410
## hsgrad        6.143e-01  2.690e-01   2.284 0.022864 *
## bagrad       -4.835e-01  2.971e-01  -1.628 0.104327
## poverty       1.856e+00  3.864e-01   4.803 2.17e-06 ***
## unemp         6.111e-01  5.314e-01   1.150 0.250812
```

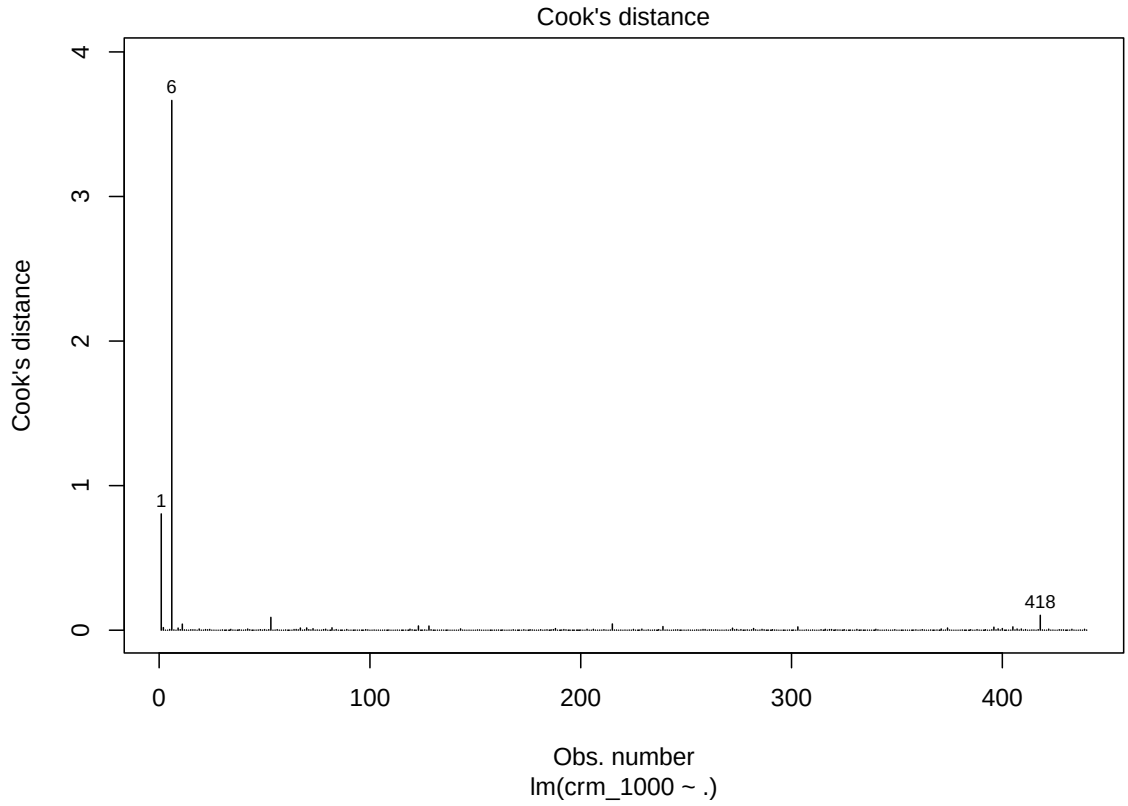
```
## pcincome      1.039e-03  4.734e-04   2.195 0.028670 *
## docs_1000    -6.634e-01  1.019e+00  -0.651 0.515556
## beds_1000     3.157e+00  7.939e-01   3.977 8.21e-05 ***
## pop_density   4.901e-03  4.537e-04  10.802  < 2e-16 ***
## northeast    -2.118e+01  3.125e+00  -6.778 4.09e-11 ***
## northcentral -1.220e+01  2.984e+00  -4.089 5.18e-05 ***
## south         6.614e+00  2.863e+00   2.310 0.021353 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.81 on 425 degrees of freedom
## Multiple R-squared:  0.589,  Adjusted R-squared:  0.5755
## F-statistic: 43.51 on 14 and 425 DF,  p-value: < 2.2e-16
```

some model diagnostics

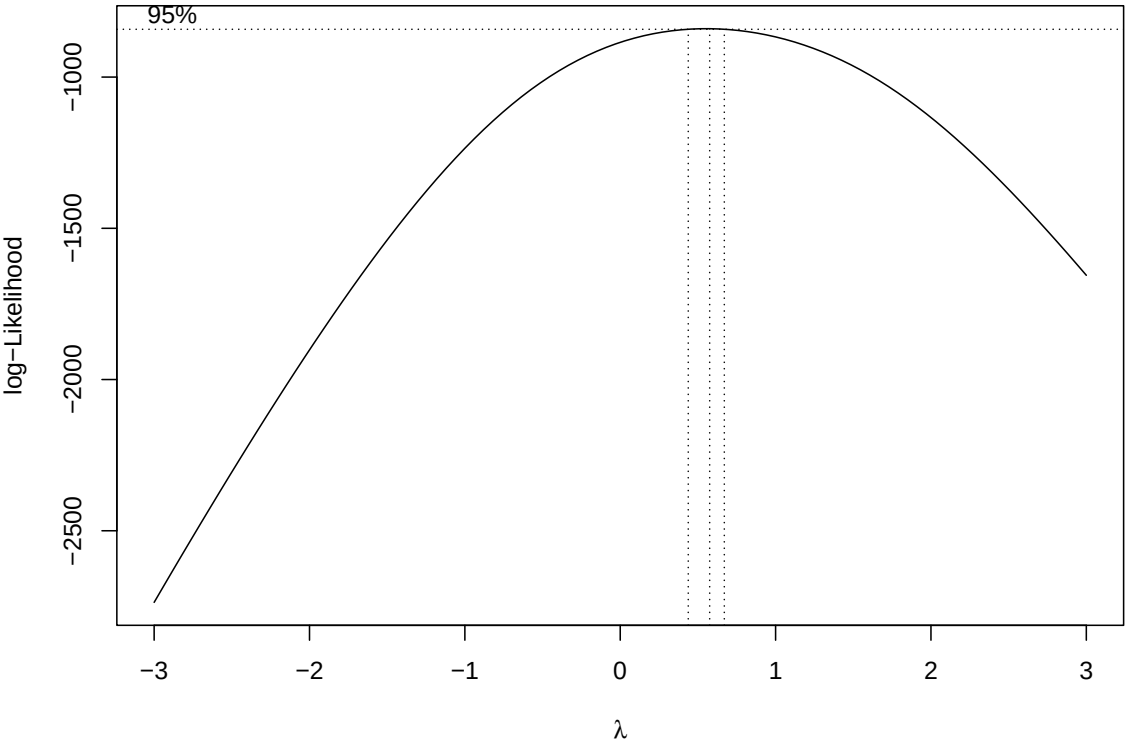
```
plot(mult_fit, which = 1)
```




```
plot(mult_fit, which = 4)
```




```
bc_model = boxcox(mult_fit, lambda = seq(-3, 3, by = 0.25))
```




```
lamb = bc_model$x[which.max(bc_model$y)]
lamb
```

```
## [1] 0.5757576
```

~0.5, thus applied square root to the Y

```
sum_cdi_mod = sum_cdi[-c(1,6),]
full_trans_fit = lm(sqrt(crm_1000) ~.,data = sum_cdi_mod)
summary(full_trans_fit)
```

```
##
## Call:
## lm(formula = sqrt(crm_1000) ~ ., data = sum_cdi_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0654 -0.6625  0.0540  0.7183  3.9085
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.644e-02  1.786e+00   0.043 0.965879
## pop           7.281e-07  1.425e-07   5.111 4.87e-07 ***
## pop18         7.584e-02  2.159e-02   3.513 0.000491 ***
## pop65        -2.316e-04  1.965e-02  -0.012 0.990601
## hsgrad        2.583e-02  1.733e-02   1.491 0.136820
## bagrad       -3.462e-02  1.911e-02  -1.812 0.070658 .
## poverty       1.111e-01  2.492e-02   4.457 1.07e-05 ***
## unemp         4.736e-02  3.407e-02   1.390 0.165214
## pcincome      1.058e-04  3.141e-05   3.367 0.000828 ***
## docs_1000    -2.102e-02  6.581e-02  -0.319 0.749576
## beds_1000     2.286e-01  5.101e-02   4.481 9.59e-06 ***
## pop_density   8.083e-05  4.359e-05   1.854 0.064417 .
## northeast    -1.719e+00  2.008e-01  -8.565  < 2e-16 ***
## northcentral -9.851e-01  1.912e-01  -5.151 3.97e-07 ***
## south         3.042e-01  1.835e-01   1.658 0.098155 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.141 on 423 degrees of freedom
## Multiple R-squared:  0.551,  Adjusted R-squared:  0.5361
## F-statistic: 37.08 on 14 and 423 DF,  p-value: < 2.2e-16
```

```
check_collinearity(full_trans_fit)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##          Term  VIF Increased SE Tolerance
##           pop 1.00         1.00      1.00
##         pop18 2.65         1.63      0.38
##         pop65 2.07         1.44      0.48
##        hsgrad 3.28         1.81      0.31
##        bagrad 3.74         1.93      0.27
##       poverty 2.43         1.56      0.41
##         unemp 1.89         1.37      0.53
##      pcincome 1.02         1.01      0.98
##     docs_1000 2.62         1.62      0.38
```

```
##     beds_1000 3.16         1.78      0.32
##   pop_density 1.01         1.01      0.99
##     northeast 2.21         1.49      0.45
##  northcentral 2.28         1.51      0.44
##         south 2.46         1.57      0.41
```

Backward Elimination

```
multi_back = step(full_trans_fit, direction='backward')
```

```
## Start:  AIC=130.27
## sqrt(crm_1000) ~ pop + pop18 + pop65 + hsgrad + bagrad + poverty +
##     unemp + pcincome + docs_1000 + beds_1000 + pop_density +
##     northeast + northcentral + south
##
##                Df Sum of Sq    RSS    AIC
## - pop65         1     0.000 550.67 128.27
## - docs_1000     1     0.133 550.81 128.37
## - unemp         1     2.516 553.19 130.26
## <none>                      550.67 130.27
## - hsgrad        1     2.892 553.56 130.56
## - south         1     3.577 554.25 131.10
## - bagrad        1     4.275 554.95 131.66
## - pop_density   1     4.475 555.15 131.81
## - pcincome      1    14.762 565.43 139.85
## - pop18         1    16.064 566.74 140.86
## - poverty       1    25.858 576.53 148.37
## - beds_1000     1    26.137 576.81 148.58
## - pop           1    34.004 584.68 154.51
## - northcentral  1    34.547 585.22 154.92
## - northeast     1    95.493 646.17 198.31
##
## Step:  AIC=128.27
## sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad + poverty + unemp +
##     pcincome + docs_1000 + beds_1000 + pop_density + northeast +
##     northcentral + south
##
##                Df Sum of Sq    RSS    AIC
## - docs_1000     1     0.133 550.81 126.37
## <none>                      550.67 128.27
## - unemp         1     2.550 553.22 128.29
## - hsgrad        1     2.903 553.58 128.57
## - south         1     3.583 554.26 129.11
## - bagrad        1     4.277 554.95 129.66
## - pop_density   1     4.515 555.19 129.84
## - pcincome      1    14.879 565.55 137.94
## - pop18         1    21.617 572.29 143.13
## - poverty       1    27.010 577.68 147.24
## - beds_1000     1    28.382 579.05 148.28
## - pop           1    34.067 584.74 152.56
## - northcentral  1    34.747 585.42 153.07
## - northeast     1    96.401 647.07 196.93
##
## Step:  AIC=126.37
```

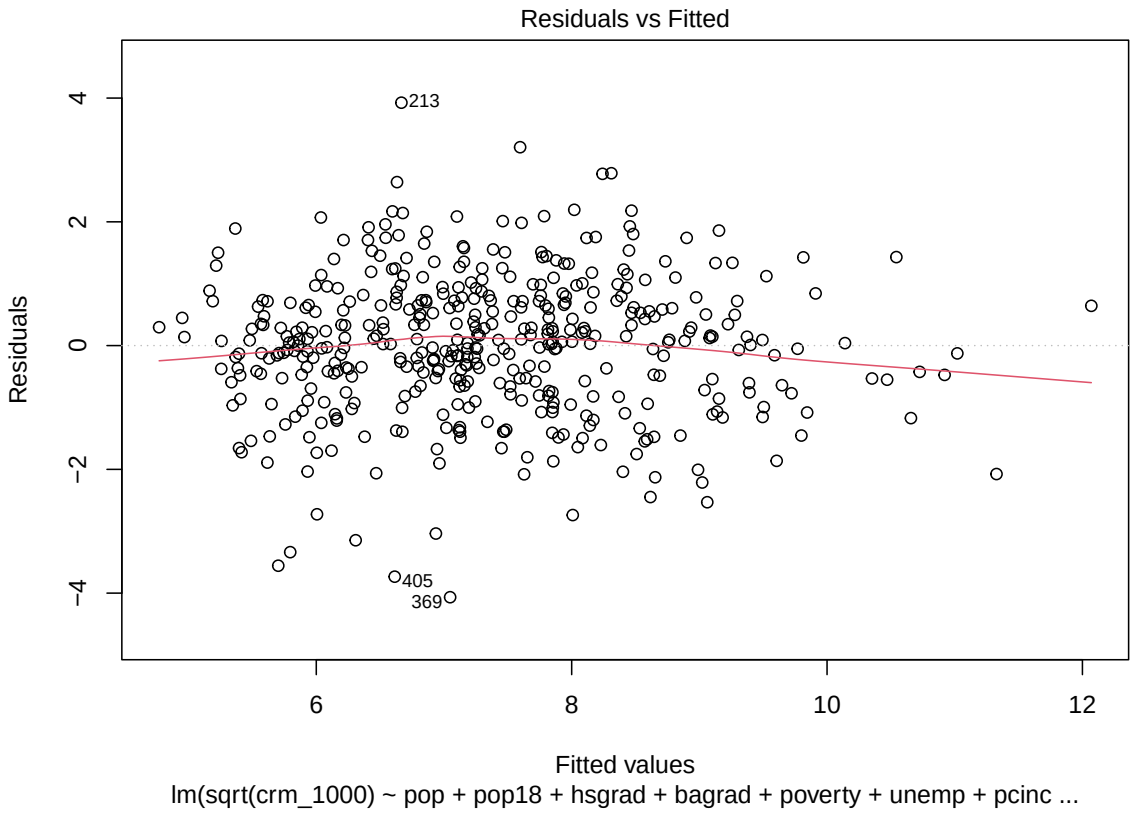
```
## sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad + poverty + unemp +
##     pcincome + beds_1000 + pop_density + northeast + northcentral +
##     south
##
##                Df Sum of Sq    RSS    AIC
## <none>                      550.81 126.37
## - unemp         1     2.533 553.34 126.38
## - hsgrad        1     3.010 553.82 126.76
## - south         1     3.944 554.75 127.50
## - pop_density   1     4.387 555.19 127.85
## - bagrad        1     4.988 555.79 128.32
## - pcincome      1    14.747 565.55 135.94
## - pop18         1    21.486 572.29 141.13
## - poverty       1    27.234 578.04 145.51
## - pop           1    33.948 584.75 150.57
## - northcentral  1    35.244 586.05 151.54
## - beds_1000     1    52.476 603.28 164.23
## - northeast     1    97.351 648.16 195.66
```

```
multi_back
```

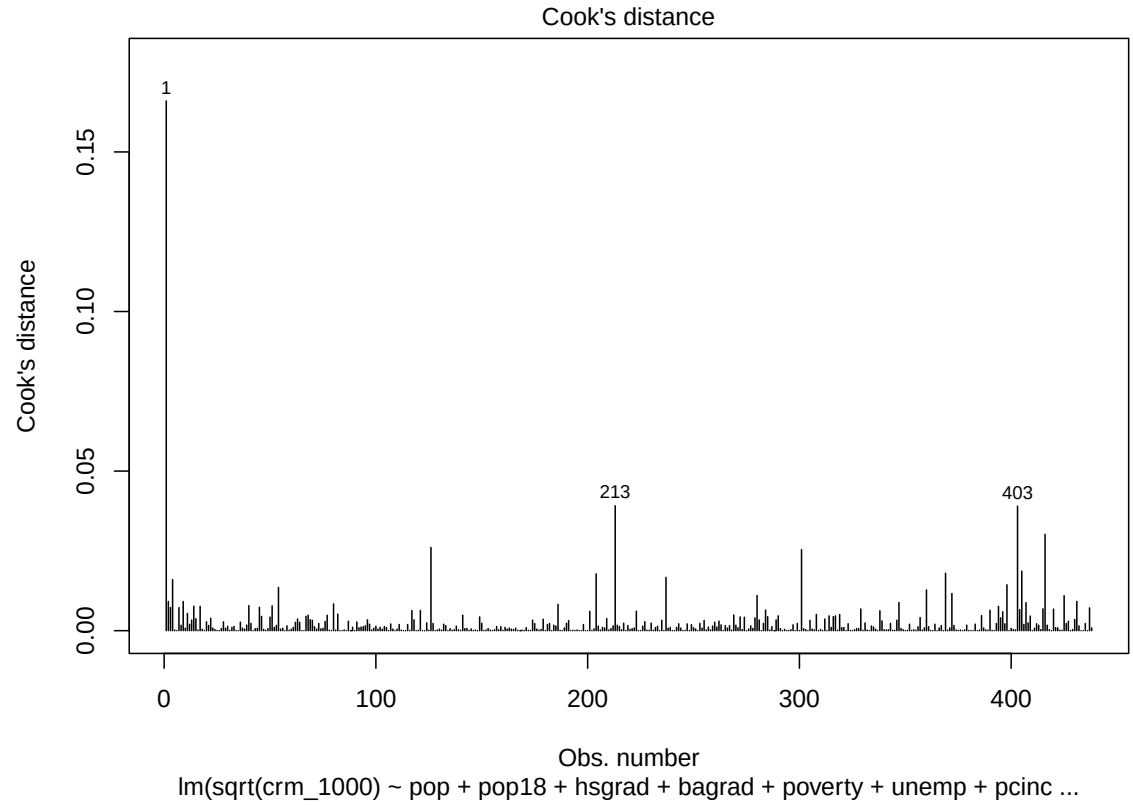
```
##
## Call:
## lm(formula = sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad +
##     poverty + unemp + pcincome + beds_1000 + pop_density + northeast +
##     northcentral + south, data = sum_cdi_mod)
##
## Coefficients:
##  (Intercept)           pop         pop18        hsgrad        bagrad
##    9.096e-02     7.261e-07     7.546e-02     2.624e-02    -3.617e-02
##      poverty         unemp      pcincome     beds_1000   pop_density
##    1.115e-01     4.714e-02     1.048e-04     2.172e-01     7.880e-05
##    northeast  northcentral         south
##   -1.711e+00    -9.731e-01     3.142e-01
```

sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad + poverty + unemp+ pcincome + beds_1000 + pop_density + northeast + northcentral + south, data = sum_cdi

```
plot(multi_back, which = 1)
```




```
plot(multi_back, which = 4)
```



```r
back_without = sum_cdi[-c(1,213,403),]

with = lm(sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad + poverty + unemp+
    pcincome + beds_1000 + pop_density + northeast + northcentral +
    south, data = sum_cdi)
without = lm(sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad + poverty + unemp+
    pcincome + beds_1000 + pop_density + northeast + northcentral +
    south, data = back_without)
summary(with); summary(without)
```

```
##
## Call:
## lm(formula = sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad +
##     poverty + unemp + pcincome + beds_1000 + pop_density + northeast +
##     northcentral + south, data = sum_cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0525 -0.7474  0.0681  0.7419  4.0605
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.607e-02  1.705e+00  -0.021 0.983127
## pop           3.602e-07  1.036e-07   3.476 0.000560 ***
## pop18         6.399e-02  1.869e-02   3.423 0.000679 ***
## hsgrad        3.579e-02  1.757e-02   2.037 0.042260 *
## bagrad       -3.749e-02  1.890e-02  -1.984 0.047915 *
## poverty       1.207e-01  2.483e-02   4.859 1.66e-06 ***
## unemp         4.193e-02  3.461e-02   1.211 0.226387
## pcincome      9.396e-05  3.076e-05   3.055 0.002393 **
## beds_1000     1.886e-01  3.441e-02   5.483 7.17e-08 ***
## pop_density   2.129e-04  2.957e-05   7.201 2.72e-12 ***
## northeast    -1.659e+00  2.020e-01  -8.215 2.56e-15 ***
## northcentral -9.426e-01  1.914e-01  -4.924 1.22e-06 ***
## south         3.399e-01  1.848e-01   1.839 0.066567 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.169 on 427 degrees of freedom
## Multiple R-squared:  0.5598, Adjusted R-squared:  0.5474
## F-statistic: 45.25 on 12 and 427 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad +
##     poverty + unemp + pcincome + beds_1000 + pop_density + northeast +
##     northcentral + south, data = back_without)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0071 -0.7513  0.0820  0.7086  4.1358
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   3.714e-01  1.693e+00   0.219 0.826445
## pop           7.086e-07  1.441e-07   4.917 1.26e-06 ***
## pop18         6.348e-02  1.851e-02   3.429 0.000664 ***
## hsgrad        3.245e-02  1.743e-02   1.862 0.063317 .
## bagrad       -3.448e-02  1.877e-02  -1.837 0.066912 .
## poverty       1.116e-01  2.474e-02   4.512 8.31e-06 ***
## unemp         4.426e-02  3.425e-02   1.292 0.196948
## pcincome      7.972e-05  3.076e-05   2.592 0.009885 **
## beds_1000     1.962e-01  3.429e-02   5.722 2.00e-08 ***
## pop_density   1.898e-04  2.999e-05   6.330 6.24e-10 ***
## northeast    -1.655e+00  1.999e-01  -8.279 1.63e-15 ***
## northcentral -9.430e-01  1.896e-01  -4.972 9.62e-07 ***
## south         3.419e-01  1.828e-01   1.870 0.062140 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.156 on 424 degrees of freedom
## Multiple R-squared:  0.5708, Adjusted R-squared:  0.5586
## F-statistic: 46.98 on 12 and 424 DF,  p-value: < 2.2e-16
```

```
check_collinearity(without)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##          Term  VIF Increased SE Tolerance
##           pop 1.00         1.00      1.00
##         pop18 1.92         1.39      0.52
##        hsgrad 3.28         1.81      0.30
##        bagrad 3.49         1.87      0.29
##       poverty 2.40         1.55      0.42
##         unemp 1.85         1.36      0.54
##      pcincome 1.03         1.01      0.98
##     beds_1000 1.44         1.20      0.69
##   pop_density 1.00         1.00      1.00
##     northeast 2.14         1.46      0.47
##  northcentral 2.17         1.47      0.46
##         south 2.38         1.54      0.42
```

Forward Selection

```
multi_forward = step(full_trans_fit, direction = 'forward')
```

```
## Start:  AIC=130.27
## sqrt(crm_1000) ~ pop + pop18 + pop65 + hsgrad + bagrad + poverty +
##     unemp + pcincome + docs_1000 + beds_1000 + pop_density +
##     northeast + northcentral + south
```

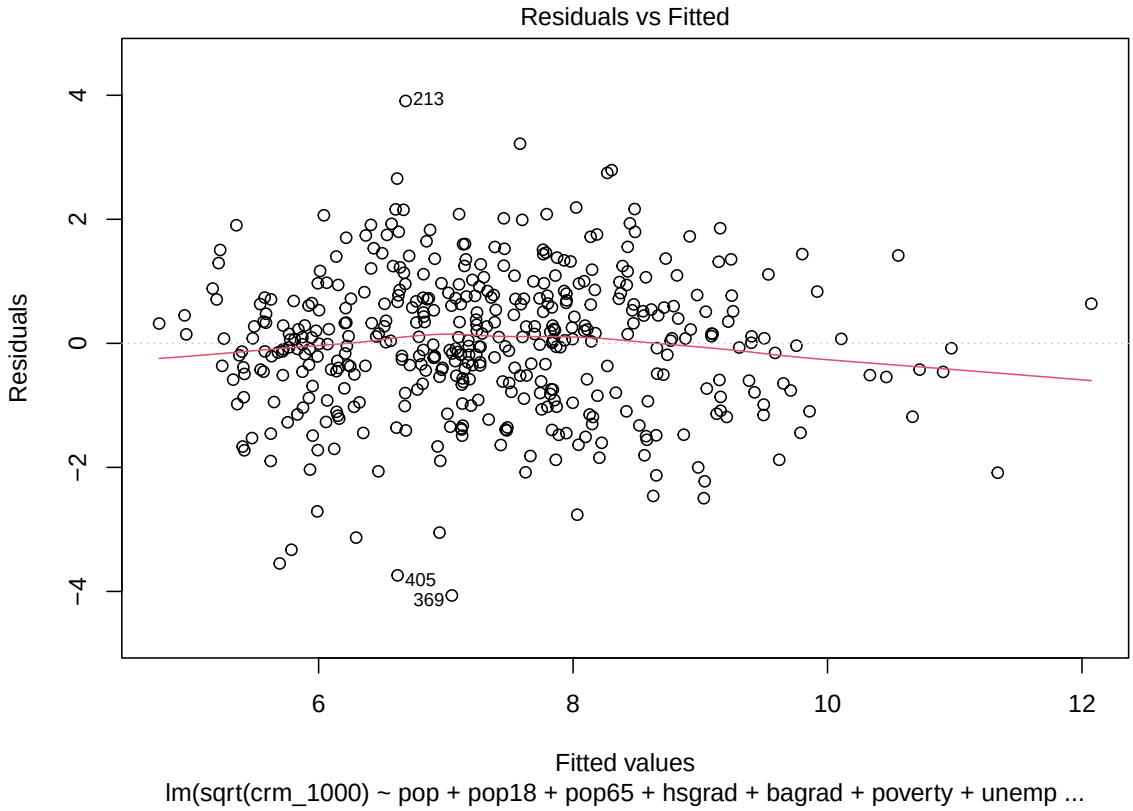
```
multi_forward
```

```
##
## Call:
## lm(formula = sqrt(crm_1000) ~ pop + pop18 + pop65 + hsgrad +
##     bagrad + poverty + unemp + pcincome + docs_1000 + beds_1000 +
##     pop_density + northeast + northcentral + south, data = sum_cdi_mod)
```

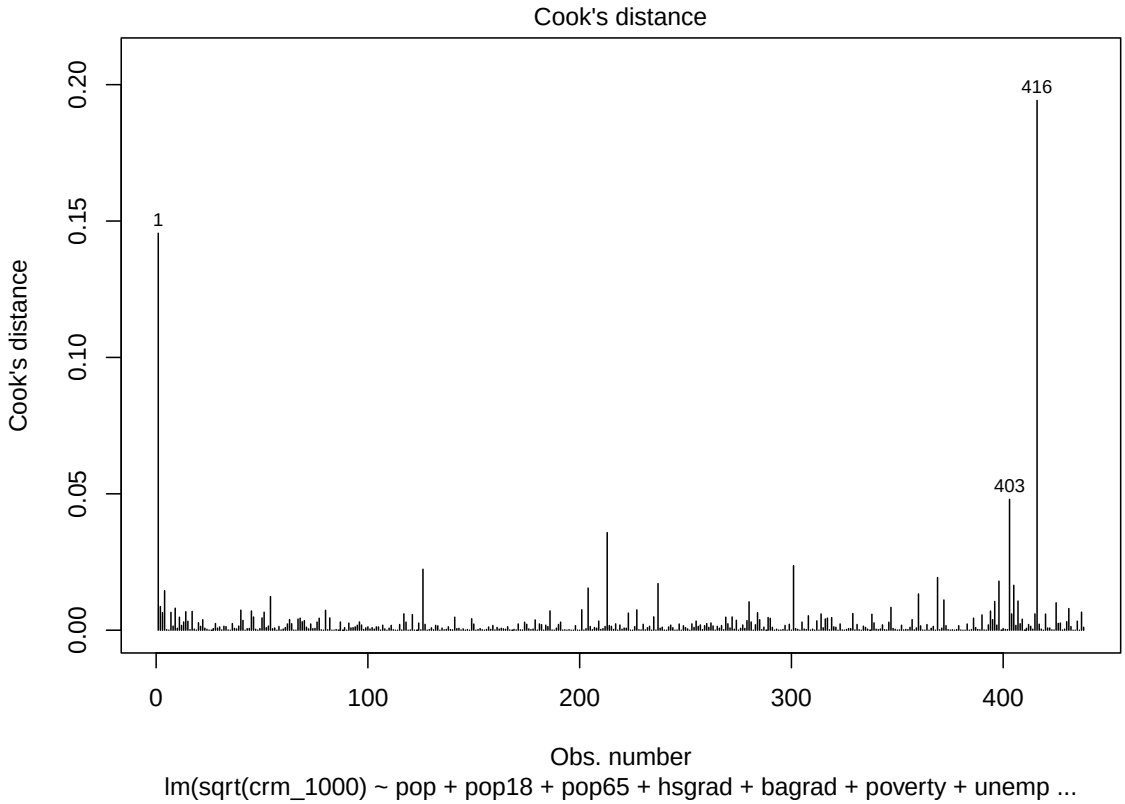
```
##
## Coefficients:
##  (Intercept)           pop         pop18         pop65        hsgrad
##    7.644e-02     7.281e-07     7.584e-02    -2.316e-04     2.583e-02
##       bagrad       poverty         unemp      pcincome     docs_1000
##   -3.462e-02     1.111e-01     4.736e-02     1.058e-04    -2.102e-02
##    beds_1000   pop_density     northeast  northcentral         south
##    2.286e-01     8.083e-05    -1.719e+00    -9.851e-01     3.042e-01
```

sqrt(crm_1000) ~ pop + pop18 + pop65 + hsgrad + bagrad + poverty + unemp + pcincome + docs_1000 + beds_1000 + pop_density + northeast + northcentral + south, data = sum_cdi_mod

```
plot(multi_forward, which = 1)
```




```
plot(multi_forward, which = 4)
```




```
forward_without = sum_cdi[-c(1,416,403),]

with_for = lm(sqrt(crm_1000) ~ pop + pop18 + pop65 + hsgrad +
    bagrad + poverty + unemp + pcincome + docs_1000 + beds_1000 +
    pop_density + northeast + northcentral + south, data = sum_cdi_mod)
without_for = lm(sqrt(crm_1000) ~ pop + pop18 + pop65 + hsgrad +
    bagrad + poverty + unemp + pcincome + docs_1000 + beds_1000 +
    pop_density + northeast + northcentral + south,data = forward_without)
summary(with_for); summary(without_for)
```

```
##
## Call:
## lm(formula = sqrt(crm_1000) ~ pop + pop18 + pop65 + hsgrad +
##     bagrad + poverty + unemp + pcincome + docs_1000 + beds_1000 +
##     pop_density + northeast + northcentral + south, data = sum_cdi_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0654 -0.6625  0.0540  0.7183  3.9085
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.644e-02  1.786e+00   0.043 0.965879
## pop           7.281e-07  1.425e-07   5.111 4.87e-07 ***
## pop18         7.584e-02  2.159e-02   3.513 0.000491 ***
## pop65        -2.316e-04  1.965e-02  -0.012 0.990601
## hsgrad        2.583e-02  1.733e-02   1.491 0.136820
## bagrad       -3.462e-02  1.911e-02  -1.812 0.070658 .
```

```
## poverty       1.111e-01  2.492e-02   4.457 1.07e-05 ***
## unemp         4.736e-02  3.407e-02   1.390 0.165214
## pcincome      1.058e-04  3.141e-05   3.367 0.000828 ***
## docs_1000    -2.102e-02  6.581e-02  -0.319 0.749576
## beds_1000     2.286e-01  5.101e-02   4.481 9.59e-06 ***
## pop_density   8.083e-05  4.359e-05   1.854 0.064417 .
## northeast    -1.719e+00  2.008e-01  -8.565  < 2e-16 ***
## northcentral -9.851e-01  1.912e-01  -5.151 3.97e-07 ***
## south         3.042e-01  1.835e-01   1.658 0.098155 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.141 on 423 degrees of freedom
## Multiple R-squared:  0.551,  Adjusted R-squared:  0.5361
## F-statistic: 37.08 on 14 and 423 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = sqrt(crm_1000) ~ pop + pop18 + pop65 + hsgrad +
##     bagrad + poverty + unemp + pcincome + docs_1000 + beds_1000 +
##     pop_density + northeast + northcentral + south, data = forward_without)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9895 -0.7426  0.0663  0.7331  4.0956
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.729e-01  1.810e+00   0.261  0.79407
## pop           7.179e-07  1.447e-07   4.963 1.01e-06 ***
## pop18         6.224e-02  2.156e-02   2.887  0.00409 **
## pop65        -5.357e-03  1.991e-02  -0.269  0.78798
## hsgrad        3.135e-02  1.755e-02   1.786  0.07475 .
## bagrad       -3.171e-02  1.942e-02  -1.633  0.10325
## poverty       1.092e-01  2.532e-02   4.314 2.00e-05 ***
## unemp         4.624e-02  3.458e-02   1.337  0.18189
## pcincome      8.192e-05  3.112e-05   2.632  0.00879 **
## docs_1000    -4.631e-02  6.643e-02  -0.697  0.48616
## beds_1000     2.244e-01  5.178e-02   4.333 1.84e-05 ***
## pop_density   1.926e-04  3.020e-05   6.377 4.74e-10 ***
## northeast    -1.671e+00  2.033e-01  -8.220 2.53e-15 ***
## northcentral -9.734e-01  1.942e-01  -5.011 7.97e-07 ***
## south         3.168e-01  1.863e-01   1.700  0.08990 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.158 on 422 degrees of freedom
## Multiple R-squared:  0.5718, Adjusted R-squared:  0.5576
## F-statistic: 40.25 on 14 and 422 DF,  p-value: < 2.2e-16
```

```
check_collinearity(without_for)
```

```
## # Check for Multicollinearity
##
## Low Correlation
```

```
##
##          Term  VIF Increased SE Tolerance
##           pop 1.00         1.00      1.00
##         pop18 2.61         1.62      0.38
##         pop65 2.06         1.43      0.49
##        hsgrad 3.32         1.82      0.30
##        bagrad 3.73         1.93      0.27
##       poverty 2.50         1.58      0.40
##         unemp 1.89         1.37      0.53
##      pcincome 1.02         1.01      0.98
##     docs_1000 2.76         1.66      0.36
##     beds_1000 3.32         1.82      0.30
##   pop_density 1.00         1.00      1.00
##     northeast 2.21         1.49      0.45
##  northcentral 2.27         1.51      0.44
##         south 2.46         1.57      0.41
```

Both direction

```
multi_both = step(full_trans_fit, direction = "both")
```

```
## Start:  AIC=130.27
## sqrt(crm_1000) ~ pop + pop18 + pop65 + hsgrad + bagrad + poverty +
##     unemp + pcincome + docs_1000 + beds_1000 + pop_density +
##     northeast + northcentral + south
##
##                Df Sum of Sq    RSS    AIC
## - pop65         1     0.000 550.67 128.27
## - docs_1000     1     0.133 550.81 128.37
## - unemp         1     2.516 553.19 130.26
## <none>                      550.67 130.27
## - hsgrad        1     2.892 553.56 130.56
## - south         1     3.577 554.25 131.10
## - bagrad        1     4.275 554.95 131.66
## - pop_density   1     4.475 555.15 131.81
## - pcincome      1    14.762 565.43 139.85
## - pop18         1    16.064 566.74 140.86
## - poverty       1    25.858 576.53 148.37
## - beds_1000     1    26.137 576.81 148.58
## - pop           1    34.004 584.68 154.51
## - northcentral  1    34.547 585.22 154.92
## - northeast     1    95.493 646.17 198.31
##
## Step:  AIC=128.27
## sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad + poverty + unemp +
##     pcincome + docs_1000 + beds_1000 + pop_density + northeast +
##     northcentral + south
##
##                Df Sum of Sq    RSS    AIC
## - docs_1000     1     0.133 550.81 126.37
## <none>                      550.67 128.27
## - unemp         1     2.550 553.22 128.29
## - hsgrad        1     2.903 553.58 128.57
## - south         1     3.583 554.26 129.11
```

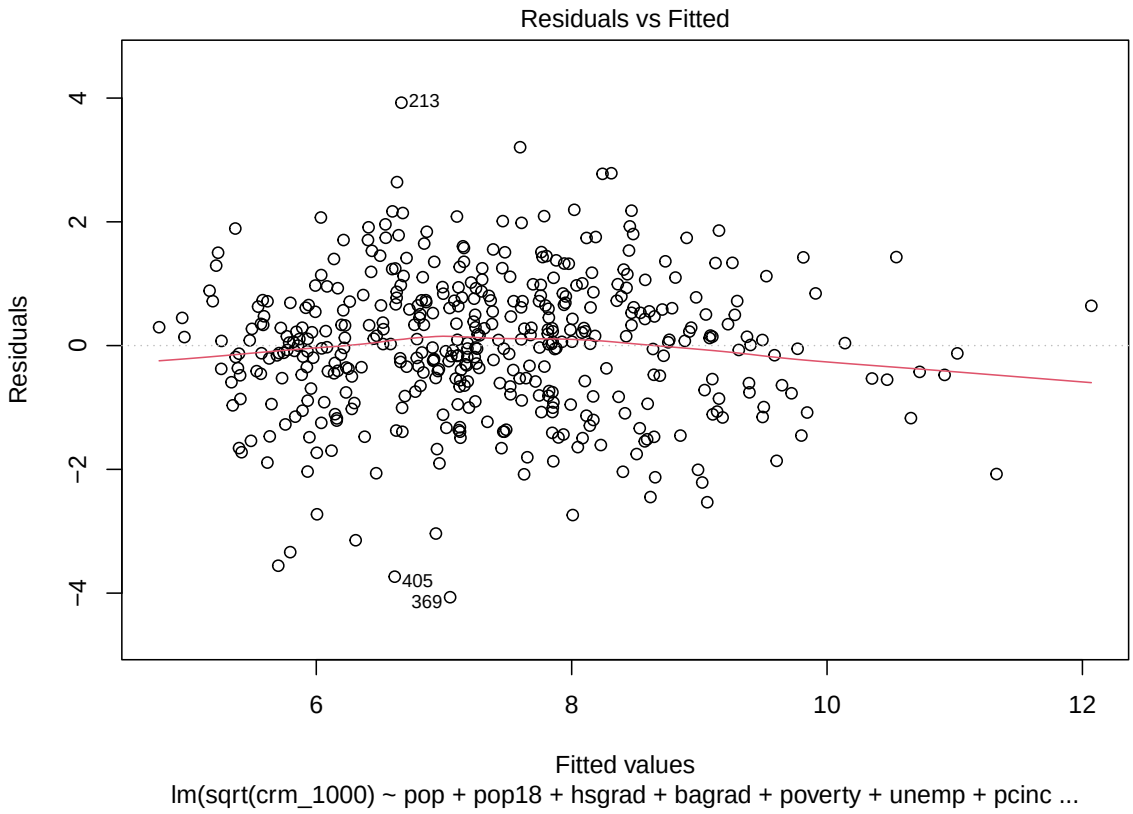
```
## - bagrad        1     4.277 554.95 129.66
## - pop_density   1     4.515 555.19 129.84
## + pop65         1     0.000 550.67 130.27
## - pcincome      1    14.879 565.55 137.94
## - pop18         1    21.617 572.29 143.13
## - poverty       1    27.010 577.68 147.24
## - beds_1000     1    28.382 579.05 148.28
## - pop           1    34.067 584.74 152.56
## - northcentral  1    34.747 585.42 153.07
## - northeast     1    96.401 647.07 196.93
##
## Step:  AIC=126.37
## sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad + poverty + unemp +
##     pcincome + beds_1000 + pop_density + northeast + northcentral +
##     south
##
##                Df Sum of Sq    RSS    AIC
## <none>                      550.81 126.37
## - unemp         1     2.533 553.34 126.38
## - hsgrad        1     3.010 553.82 126.76
## - south         1     3.944 554.75 127.50
## - pop_density   1     4.387 555.19 127.85
## + docs_1000     1     0.133 550.67 128.27
## - bagrad        1     4.988 555.79 128.32
## + pop65         1     0.000 550.81 128.37
## - pcincome      1    14.747 565.55 135.94
## - pop18         1    21.486 572.29 141.13
## - poverty       1    27.234 578.04 145.51
## - pop           1    33.948 584.75 150.57
## - northcentral  1    35.244 586.05 151.54
## - beds_1000     1    52.476 603.28 164.23
## - northeast     1    97.351 648.16 195.66
```

```
multi_both
```

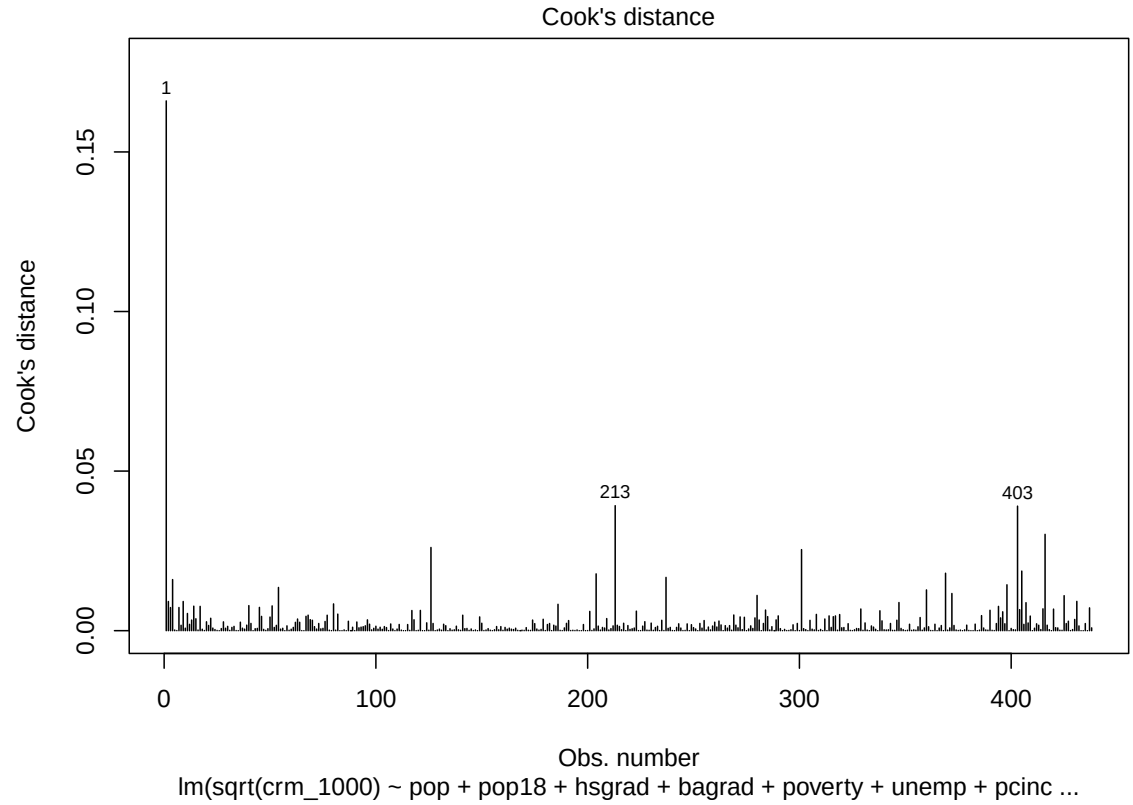
```
##
## Call:
## lm(formula = sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad +
##     poverty + unemp + pcincome + beds_1000 + pop_density + northeast +
##     northcentral + south, data = sum_cdi_mod)
##
## Coefficients:
##  (Intercept)           pop         pop18        hsgrad        bagrad
##    9.096e-02     7.261e-07     7.546e-02     2.624e-02    -3.617e-02
##      poverty         unemp      pcincome     beds_1000   pop_density
##    1.115e-01     4.714e-02     1.048e-04     2.172e-01     7.880e-05
##    northeast  northcentral         south
##   -1.711e+00    -9.731e-01     3.142e-01
```

sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad + poverty + unemp + pcincome + beds_1000 + pop_density + northeast + northcentral + south, data = sum_cdi_mod

```
plot(multi_both, which = 1)
```




```
plot(multi_both, which = 4)
```



```r
both_without = sum_cdi[-c(1,213,403),]

with_both = lm(sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad +
    poverty + unemp + pcincome + beds_1000 + pop_density + northeast +
    northcentral + south, data = sum_cdi_mod)
without_both = lm(sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad +
    poverty + unemp + pcincome + beds_1000 + pop_density + northeast +
    northcentral + south, data = both_without)
summary(with_both); summary(without_both)
```

```
##
## Call:
## lm(formula = sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad +
##     poverty + unemp + pcincome + beds_1000 + pop_density + northeast +
##     northcentral + south, data = sum_cdi_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0662 -0.6619  0.0502  0.7174  3.9254
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.096e-02  1.667e+00   0.055 0.956516
## pop           7.261e-07  1.419e-07   5.118 4.69e-07 ***
## pop18         7.546e-02  1.853e-02   4.072 5.57e-05 ***
## hsgrad        2.624e-02  1.722e-02   1.524 0.128270
## bagrad       -3.617e-02  1.844e-02  -1.962 0.050439 .
## poverty       1.115e-01  2.432e-02   4.584 6.01e-06 ***
## unemp         4.714e-02  3.372e-02   1.398 0.162867
## pcincome      1.048e-04  3.108e-05   3.373 0.000811 ***
## beds_1000     2.172e-01  3.414e-02   6.363 5.12e-10 ***
## pop_density   7.881e-05  4.283e-05   1.840 0.066502 .
## northeast    -1.711e+00  1.974e-01  -8.667  < 2e-16 ***
## northcentral -9.731e-01  1.866e-01  -5.215 2.88e-07 ***
## south         3.142e-01  1.801e-01   1.744 0.081807 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.138 on 425 degrees of freedom
## Multiple R-squared:  0.5509, Adjusted R-squared:  0.5382
## F-statistic: 43.45 on 12 and 425 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = sqrt(crm_1000) ~ pop + pop18 + hsgrad + bagrad +
##     poverty + unemp + pcincome + beds_1000 + pop_density + northeast +
##     northcentral + south, data = both_without)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0071 -0.7513  0.0820  0.7086  4.1358
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   3.714e-01  1.693e+00   0.219 0.826445
## pop           7.086e-07  1.441e-07   4.917 1.26e-06 ***
## pop18         6.348e-02  1.851e-02   3.429 0.000664 ***
## hsgrad        3.245e-02  1.743e-02   1.862 0.063317 .
## bagrad       -3.448e-02  1.877e-02  -1.837 0.066912 .
## poverty       1.116e-01  2.474e-02   4.512 8.31e-06 ***
## unemp         4.426e-02  3.425e-02   1.292 0.196948
## pcincome      7.972e-05  3.076e-05   2.592 0.009885 **
## beds_1000     1.962e-01  3.429e-02   5.722 2.00e-08 ***
## pop_density   1.898e-04  2.999e-05   6.330 6.24e-10 ***
## northeast    -1.655e+00  1.999e-01  -8.279 1.63e-15 ***
## northcentral -9.430e-01  1.896e-01  -4.972 9.62e-07 ***
## south         3.419e-01  1.828e-01   1.870 0.062140 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.156 on 424 degrees of freedom
## Multiple R-squared:  0.5708, Adjusted R-squared:  0.5586
## F-statistic: 46.98 on 12 and 424 DF,  p-value: < 2.2e-16
```

```
check_collinearity(without_both)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##          Term  VIF Increased SE Tolerance
##           pop 1.00         1.00      1.00
##         pop18 1.92         1.39      0.52
##        hsgrad 3.28         1.81      0.30
##        bagrad 3.49         1.87      0.29
##       poverty 2.40         1.55      0.42
##         unemp 1.85         1.36      0.54
##      pcincome 1.03         1.01      0.98
##     beds_1000 1.44         1.20      0.69
##   pop_density 1.00         1.00      1.00
##     northeast 2.14         1.46      0.47
##  northcentral 2.17         1.47      0.46
##         south 2.38         1.54      0.42
```