Lab 3: Regression Competition

Ben Thomas
Olek Wojcik

9/25/2019

Our final model and functions

Our data wrangling function

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
group_J_process <- function(training_data) {
  training_data <- training_data %>%
    mutate(sqrInvInc = (100*pctWInvInc)^2,
           sqrPop = population^2)
}
```

Our fit function

```
library(tidyverse)
group_J_fit <- function(training_data) {
  m1 <- lm(ViolentCrimesPerPop ~ population + sqrPop + log(medIncome) + PctHousOccup +
             NumInShelters + PctKids2Par + pctWInvInc + sqrInvInc + PctPersDenseHous +
             racePctWhite + PctWorkMomYoungKids, data = training_data)
  m1
}
```

Our MSE function

```
group_J_MSE <- function(model, data) {
  MSE <- mean((data$ViolentCrimesPerPop - predict.lm(model, data)) ^ 2)
  return(MSE)
}
```

Running the MSE function on our data/fit

```
d <- read.csv("http://andrewpbray.github.io/data/crime-train.csv")
d_new <- group_J_process(d)
group_J_MSE(group_J_fit(d_new), d_new)
```

```
## [1] 0.01927444
```

Preliminary work

Exercise 1

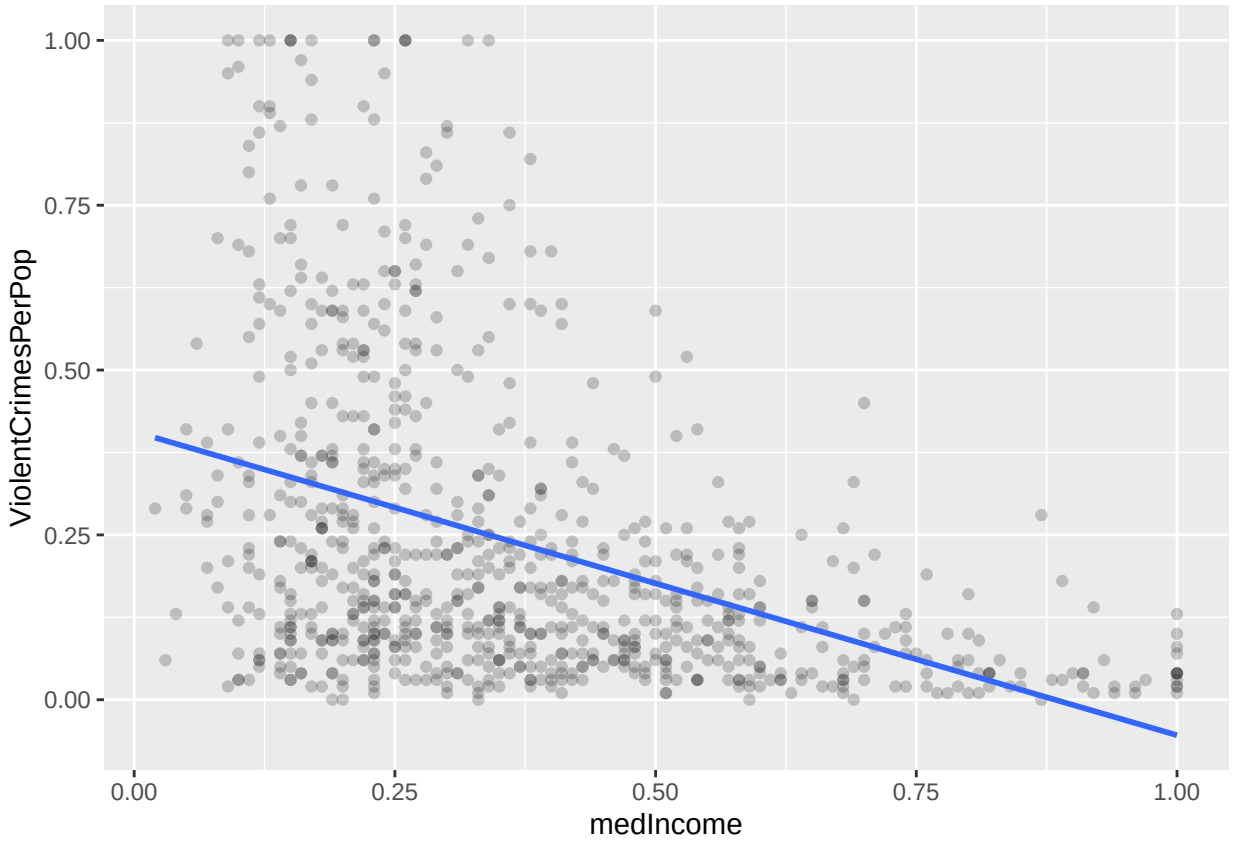
What we expect to have a significant effect:

- medIncome
- PctNotHSGrad
- PctUnemployed
- PctPersDenseHous
- RacialMatchCommPol
- PopDens
- PctUnemployed

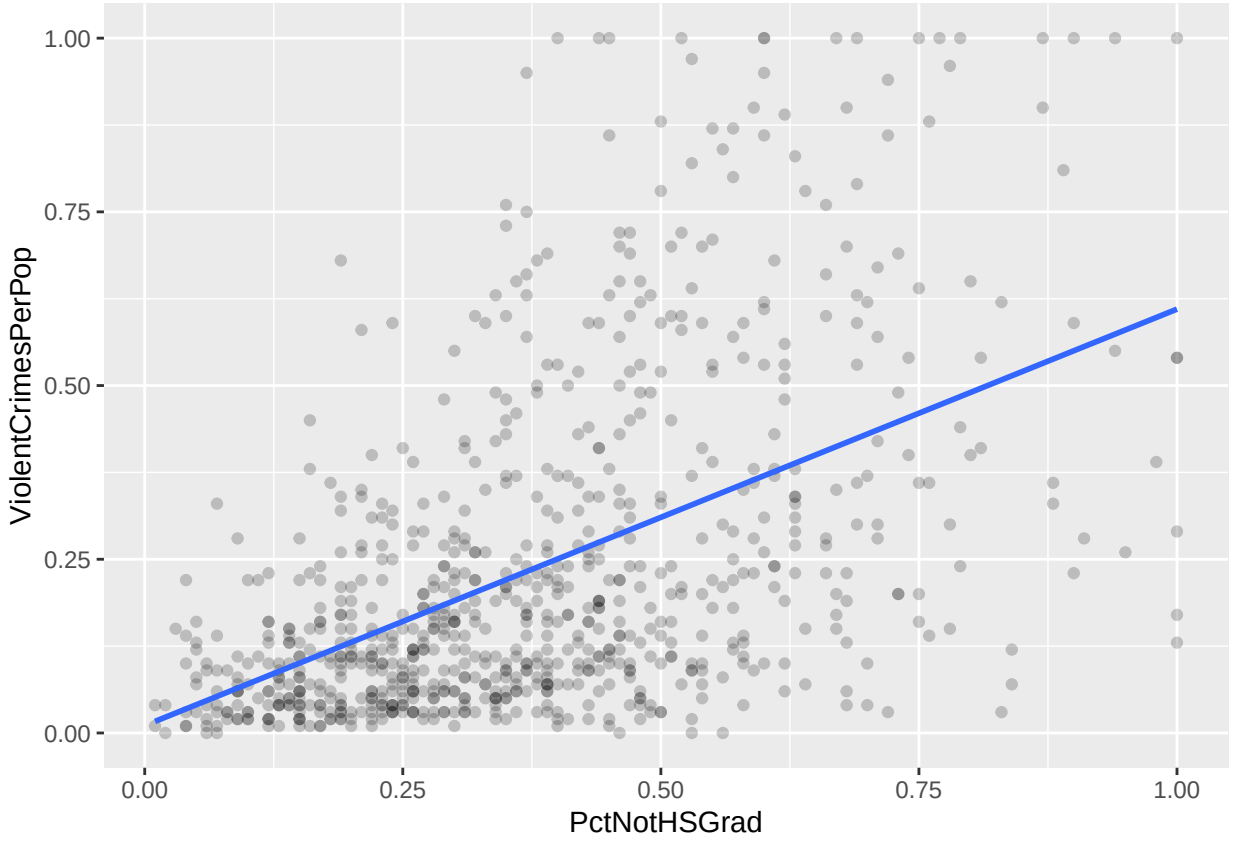
Exercise 2

We've graphed a number of scatterplots with ViolentCrimesPerPop and variables of interest to determine what these relationships look like.

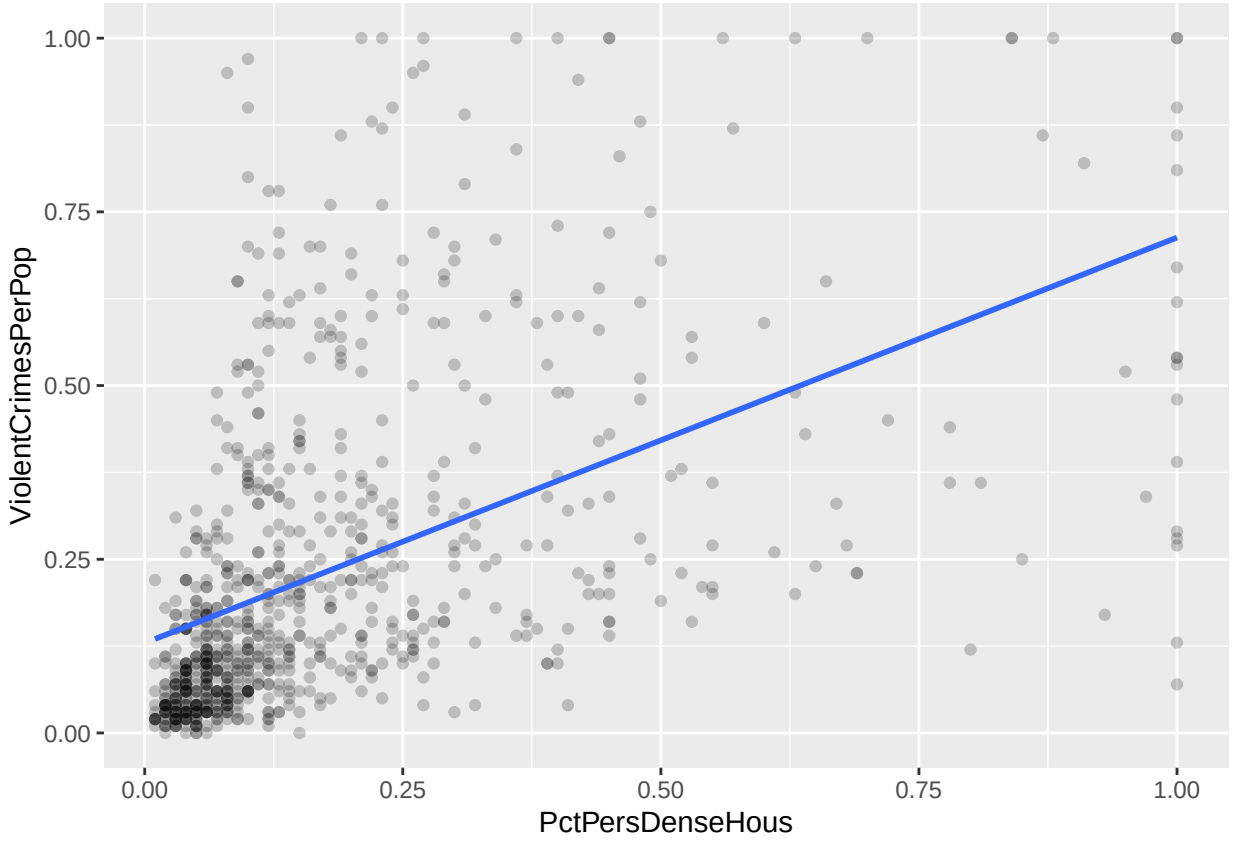
```
library(tidyverse)
d <- read.csv("http://andrewpbray.github.io/data/crime-train.csv")
ggplot(data = d, mapping = aes(y = ViolentCrimesPerPop, x = medIncome)) +
geom_point(alpha = 0.2) +
geom_smooth(method = "lm", se = FALSE)
```



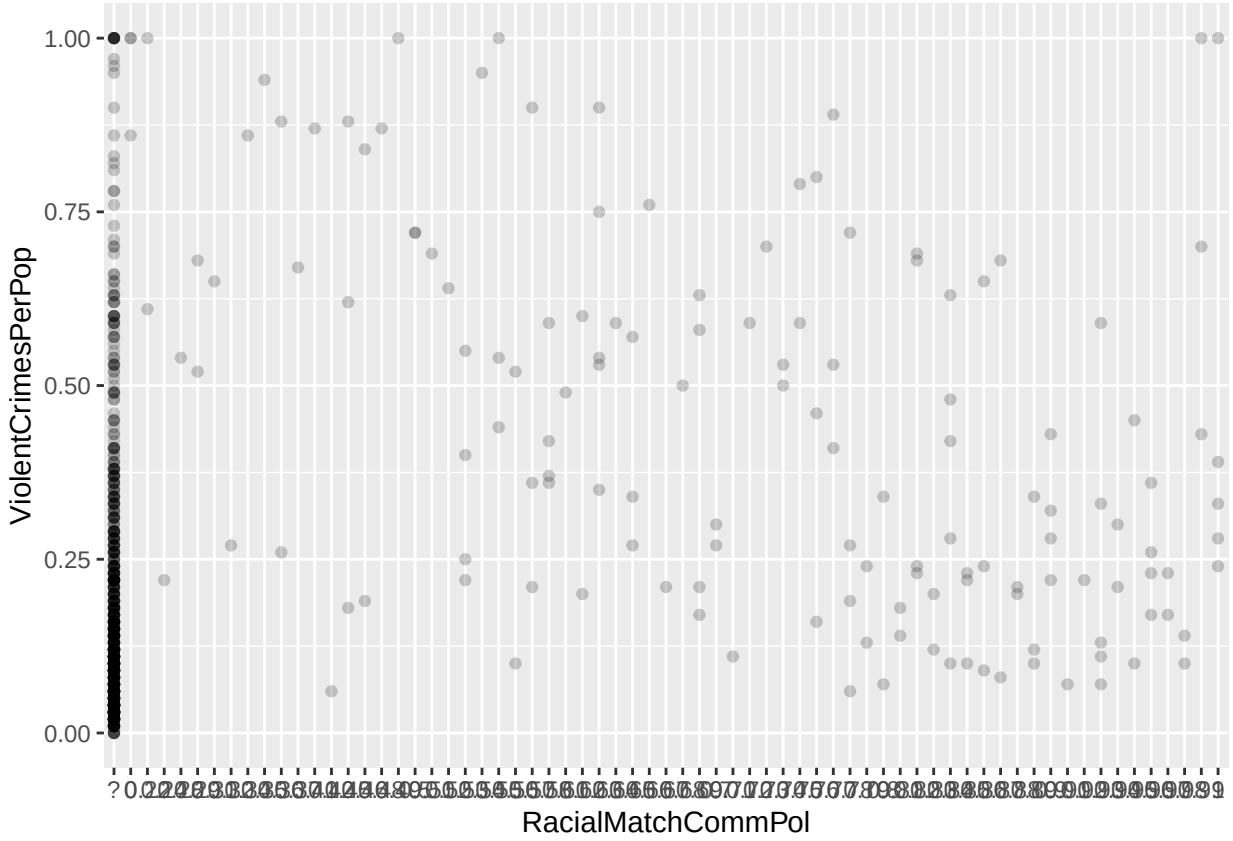
```
ggplot(data = d, mapping = aes(y = ViolentCrimesPerPop, x = PctNotHSGrad)) +
geom_point(alpha = 0.2) +
geom_smooth(method = "lm", se = FALSE)
```



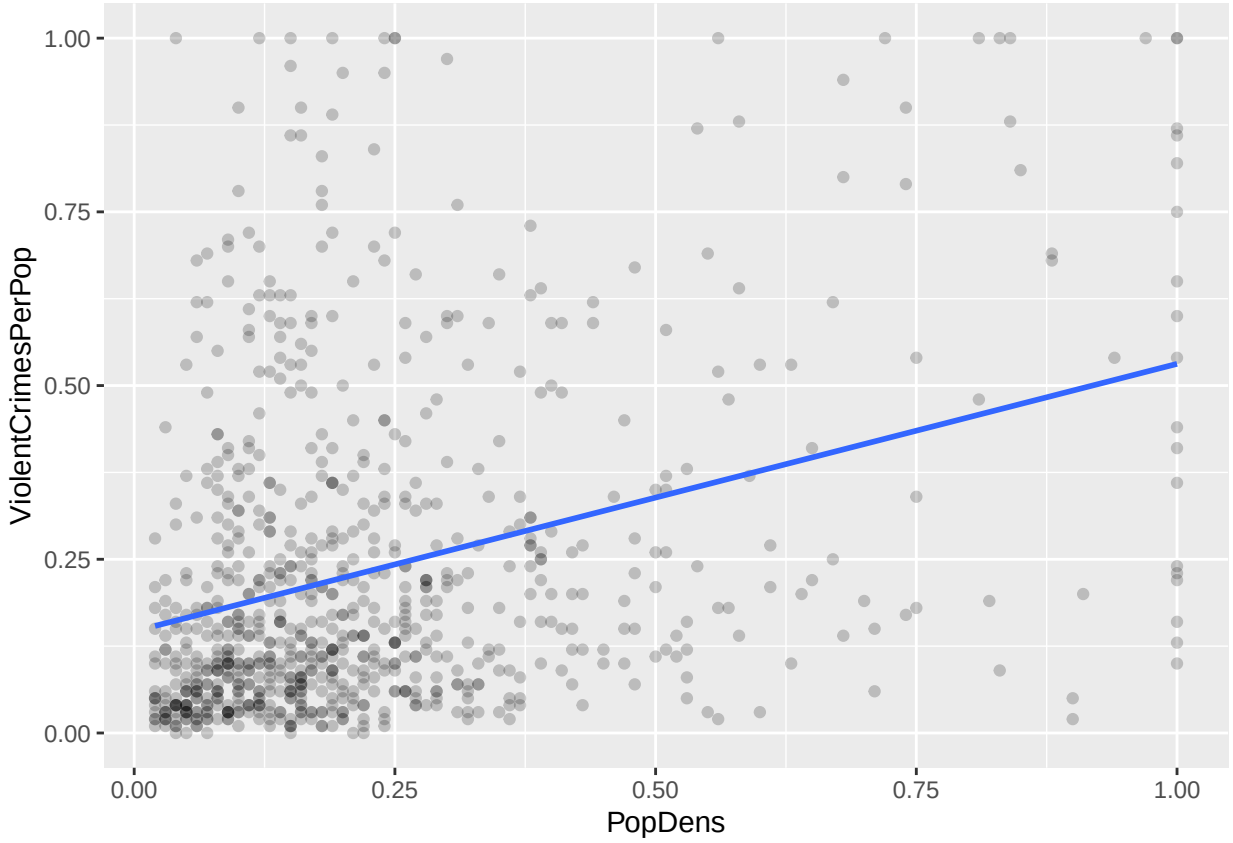
```
ggplot(data = d, mapping = aes(y = ViolentCrimesPerPop, x = PctPersDenseHous)) +
geom_point(alpha = 0.2) +
geom_smooth(method = "lm", se = FALSE)
```



```
ggplot(data = d, mapping = aes(y = ViolentCrimesPerPop, x = RacialMatchCommPol)) +
geom_point(alpha = 0.2) +
geom_smooth(method = "lm", se = FALSE)
```



```
ggplot(data = d, mapping = aes(y = ViolentCrimesPerPop, x = PopDens)) +
geom_point(alpha = 0.2) +
geom_smooth(method = "lm", se = FALSE)
```

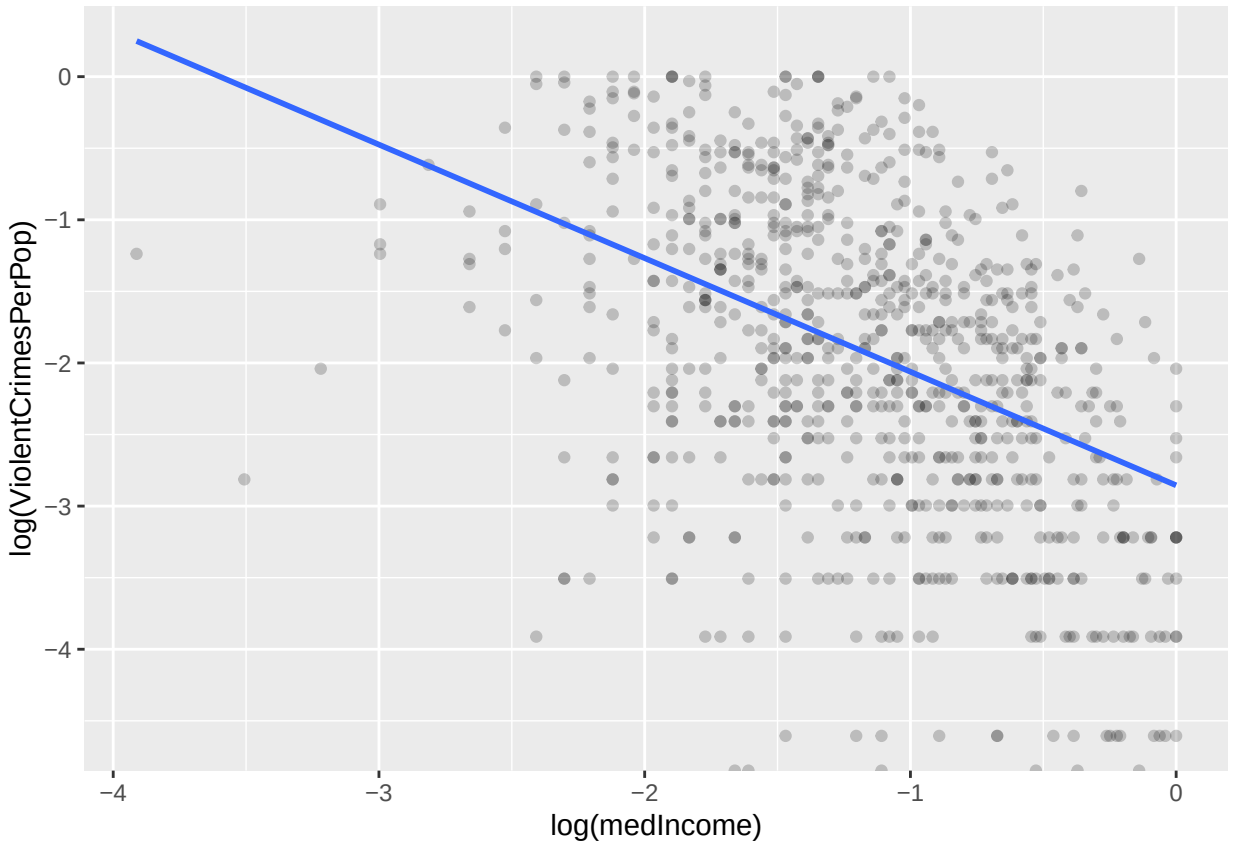


```
ggplot(data = d, mapping = aes(y = ViolentCrimesPerPop, x = PctUnemployed)) +
geom_point(alpha = 0.2) +
geom_smooth(method = "lm", se = FALSE)
```

After going through most of the police variables, it seems like the missing data makes them questionable variables at best. We therefore decided to stick to the other variables to make our model. We decided to log some variables to make them neater and more linear, but later on we discovered that due to an error, ViolentCrimesPerPop can't be logged without an error that breaks the regression.

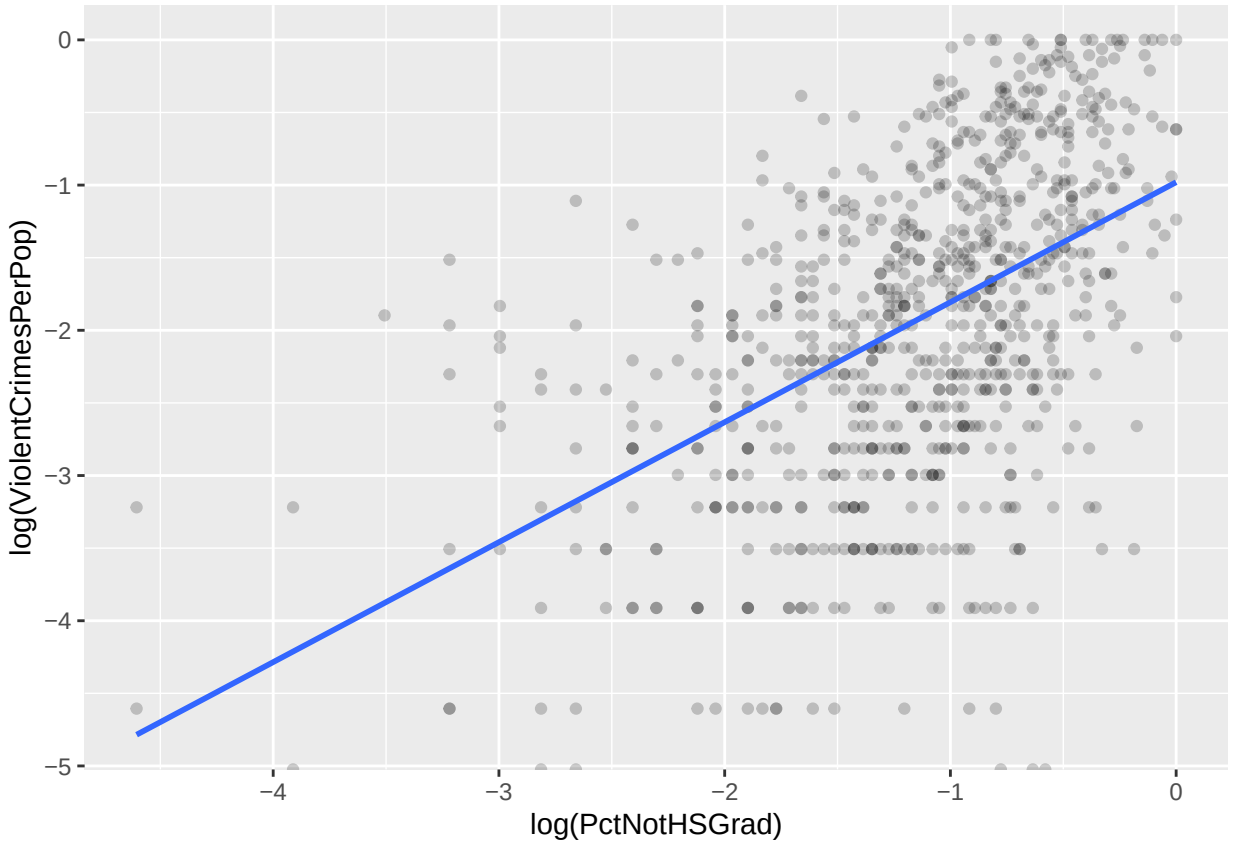
```
ggplot(data = d, mapping = aes(y = log(ViolentCrimesPerPop), x = log(medIncome))) +
geom_point(alpha = 0.2) +
geom_smooth(method = "lm", se = FALSE)
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```


log(ViolentCrimesPerPop)
0
−1
−2
−3
−4
−4
−3
−2
−1
0
log(medIncome)

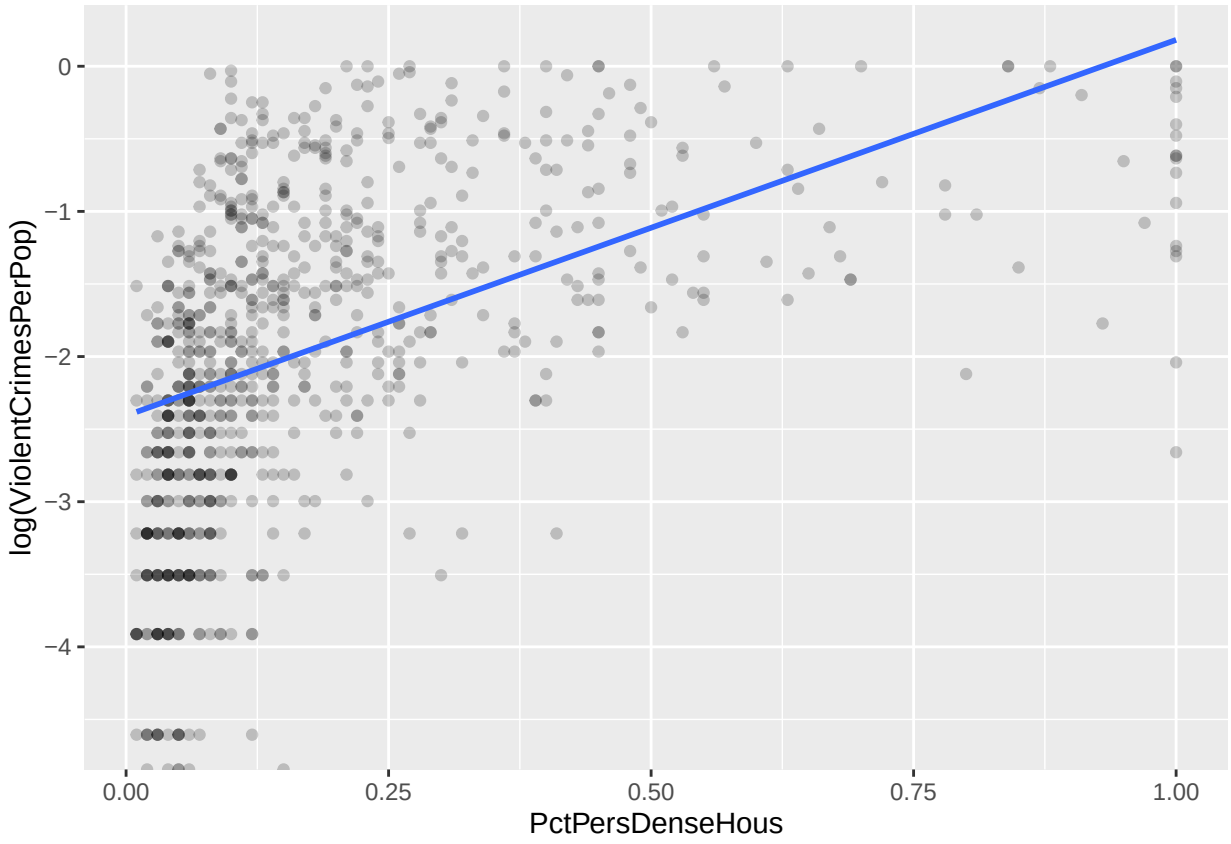
```r
ggplot(data = d, mapping = aes(y = log(ViolentCrimesPerPop), x = log(PctNotHSGrad))) +
geom_point(alpha = 0.2) +
geom_smooth(method = "lm", se = FALSE)
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```



```
ggplot(data = d, mapping = aes(y = log(ViolentCrimesPerPop), x = PctPersDenseHous)) +
geom_point(alpha = 0.2) +
geom_smooth(method = "lm", se = FALSE)
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```



Exercise 3

Here is Olek's original model:

```
d <- d %>%
mutate( log_ViolentCrimesPerPop = log(ViolentCrimesPerPop), log_medIncome = log(medIncome), log_PctNotHS
  drop_na(log_ViolentCrimesPerPop, log_medIncome, log_PctNotHSGrad, log_PctPersDenseHous)  %>%
   mutate(sqrInvInc = (100*pctWInvInc)^2,
        sqrPop = population^2,
        invPop = 1/population)

olek1 = lm(ViolentCrimesPerPop ~ log_medIncome + log_PctNotHSGrad + PctPersDenseHous, data = d)

summary(olek1)
```

```
##
## Call:
## lm(formula = ViolentCrimesPerPop ~ log_medIncome + log_PctNotHSGrad +
##     PctPersDenseHous, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69800 -0.11287 -0.03701  0.06834  0.69017
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.13647    0.03529   3.867 0.000119 ***
## log_medIncome    -0.07459    0.01569  -4.754 2.37e-06 ***
## log_PctNotHSGrad  0.06060    0.01500   4.040 5.87e-05 ***
## PctPersDenseHous  0.45144    0.03587  12.587  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.189 on 796 degrees of freedom
## Multiple R-squared:  0.355,  Adjusted R-squared:  0.3525
## F-statistic:   146 on 3 and 796 DF,  p-value: < 2.2e-16
```

Here is Ben's original model:

```
ben1 <- lm(ViolentCrimesPerPop ~ PctPopUnderPov + PctLess9thGrade + PctUnemployed + PersPerFam + PctHous

summary(ben1)
```

```
##
## Call:
## lm(formula = ViolentCrimesPerPop ~ PctPopUnderPov + PctLess9thGrade +
##     PctUnemployed + PersPerFam + PctHousOccup + NumInShelters,
##     data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65072 -0.10087 -0.03115  0.06721  0.73235
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.07705    0.03432   2.245 0.025058 *
## PctPopUnderPov   0.20533    0.05420   3.788 0.000163 ***
## PctLess9thGrade  0.14054    0.04527   3.104 0.001974 **
## PctUnemployed    0.22254    0.05605   3.970 7.83e-05 ***
## PersPerFam       0.12143    0.04652   2.610 0.009221 **
## PctHousOccup    -0.14258    0.03900  -3.655 0.000274 ***
## NumInShelters    0.58721    0.05720  10.266  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1799 on 793 degrees of freedom
## Multiple R-squared:  0.4181, Adjusted R-squared:  0.4137
## F-statistic: 94.95 on 6 and 793 DF,  p-value: < 2.2e-16
```

These were our initial models. We decided to experiment with different variables, and we saw that many variables captured others: for example, if we added PctKids2Par to the model, PctNotHSGrad would become insignificant. This led us to transform some variables as well: we looked at graphs of our different variables to check their linearity, and found that population looked quite bent, and so we chose to squre it, and the term was significant. We also looked at the diagnostics charts. We need to have a problem with homoskedasticity, but couldn't find a way around it. The models looked quite similar in the diagnostics, so we decided to choose the one with the higher adjusted R squared, which was Ben's.

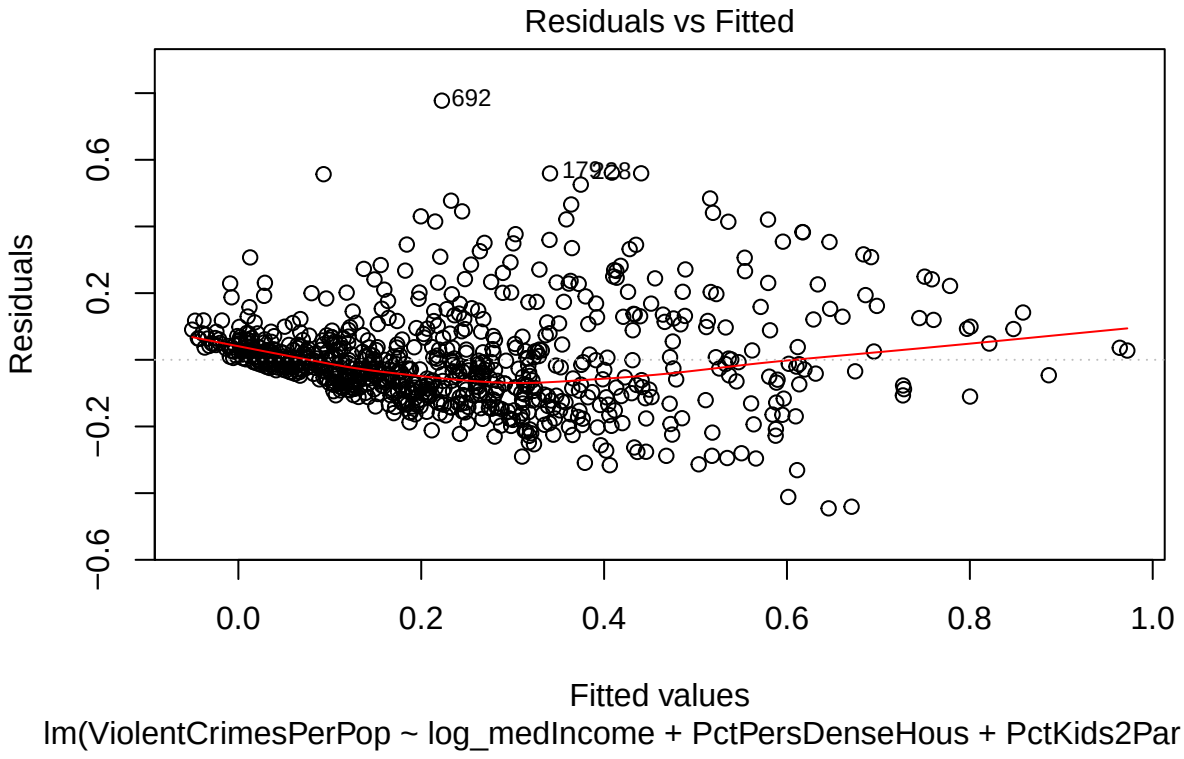
```r
olek2 = lm(ViolentCrimesPerPop ~ log_medIncome + PctPersDenseHous + PctKids2Par + NumUnderPov,  data = d

summary(olek2)
```

```
##
## Call:
## lm(formula = ViolentCrimesPerPop ~ log_medIncome + PctPersDenseHous +
##     PctKids2Par + NumUnderPov, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44555 -0.08363 -0.01473  0.05172  0.77743
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.81989    0.03845  21.325  < 2e-16 ***
## log_medIncome     0.07942    0.01202   6.605 7.28e-11 ***
## PctPersDenseHous  0.22744    0.02805   8.109 1.93e-15 ***
## PctKids2Par      -0.87515    0.04012 -21.815  < 2e-16 ***
## NumUnderPov       0.22113    0.03952   5.595 3.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1434 on 795 degrees of freedom
## Multiple R-squared:  0.6294, Adjusted R-squared:  0.6276
## F-statistic: 337.6 on 4 and 795 DF,  p-value: < 2.2e-16
```

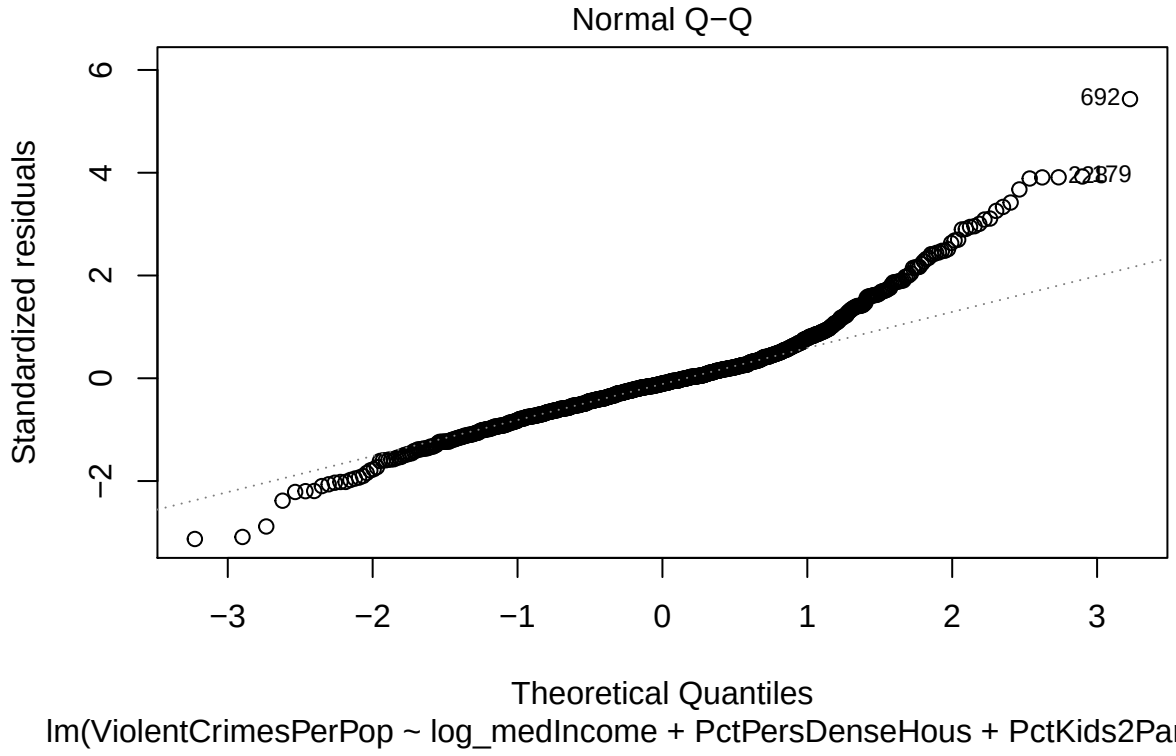
```r
ben2 <- lm(ViolentCrimesPerPop ~ population + sqrPop + log(medIncome) + PctHousOccup + NumInShelters + I

summary(ben2)
```

```
##
## Call:
## lm(formula = ViolentCrimesPerPop ~ population + sqrPop + log(medIncome) +
##     PctHousOccup + NumInShelters + PctKids2Par + pctWInvInc +
##     sqrInvInc + PctPersDenseHous + racePctWhite + PctWorkMomYoungKids,
##     data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46658 -0.07734 -0.01559  0.04443  0.75172
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.067e+00  6.689e-02  15.958  < 2e-16 ***
## population           3.186e-01  1.157e-01   2.753 0.006036 **
## sqrPop              -3.289e-01  1.253e-01  -2.624 0.008851 **
## log(medIncome)       7.466e-02  1.518e-02   4.920 1.05e-06 ***
## PctHousOccup        -8.128e-02  2.954e-02  -2.751 0.006069 **
## NumInShelters        2.460e-01  8.033e-02   3.062 0.002271 **
## PctKids2Par         -6.548e-01  5.614e-02 -11.662  < 2e-16 ***
## pctWInvInc          -6.507e-01  1.794e-01  -3.627 0.000305 ***
## sqrInvInc            4.540e-05  1.507e-05   3.013 0.002668 **
## PctPersDenseHous     7.940e-02  3.870e-02   2.051 0.040554 *
## racePctWhite        -1.380e-01  4.067e-02  -3.394 0.000723 ***
## PctWorkMomYoungKids -2.153e-02  3.226e-02  -0.667 0.504650
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1399 on 788 degrees of freedom
## Multiple R-squared:  0.6504, Adjusted R-squared:  0.6455
## F-statistic: 133.2 on 11 and 788 DF,  p-value: < 2.2e-16
```

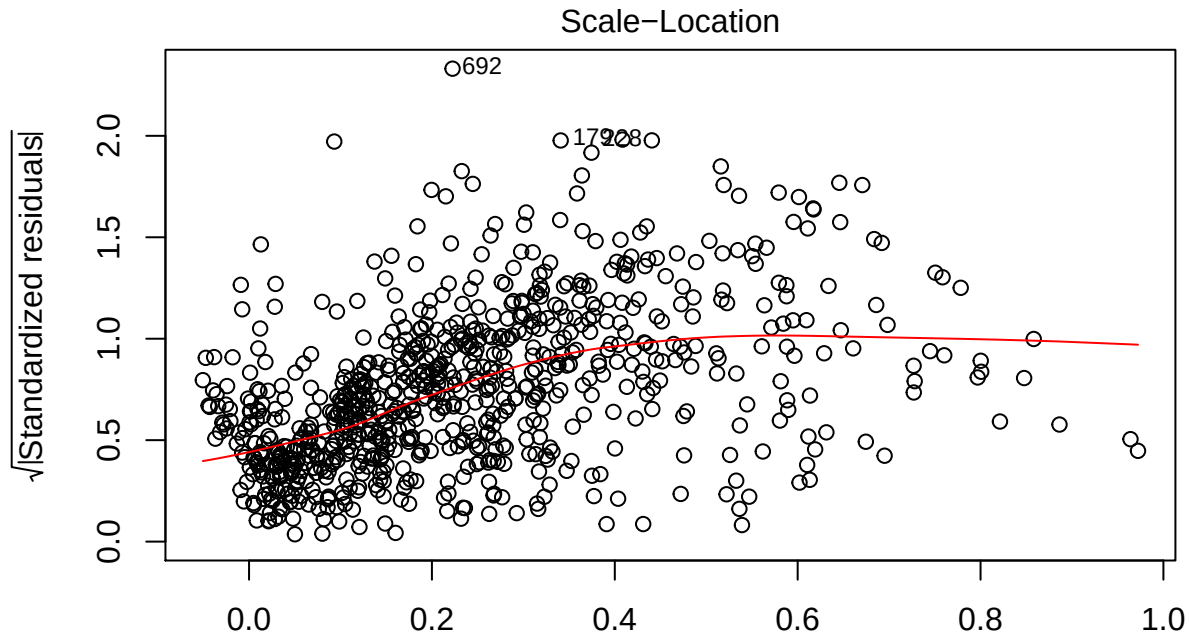
```
plot(olek2, 1)
```



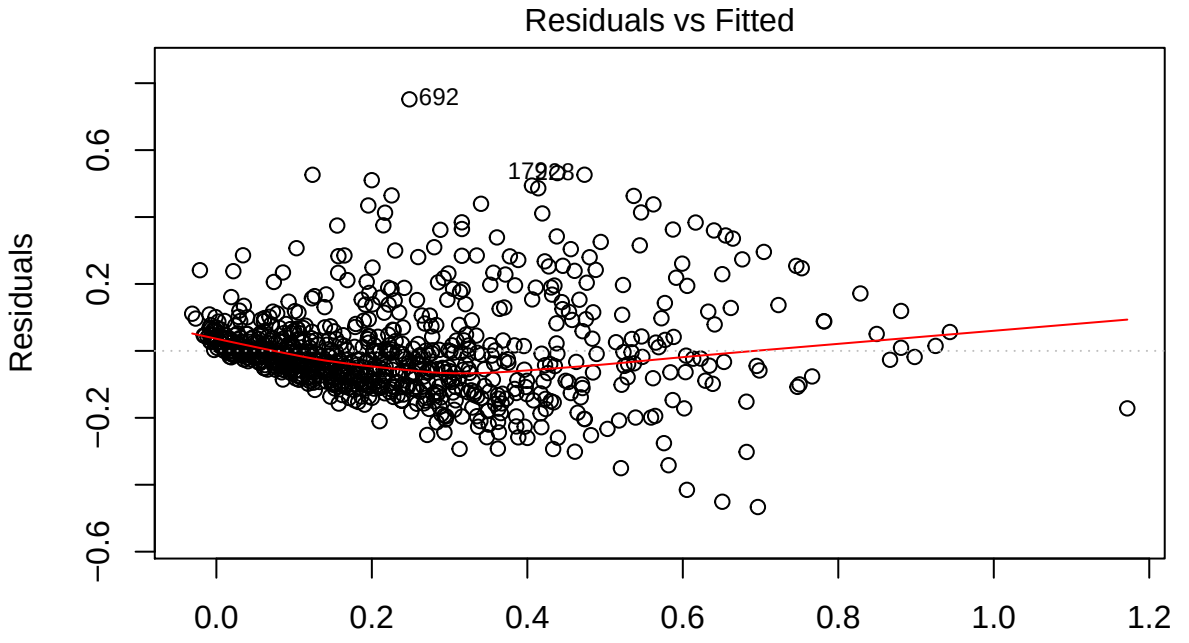

```
plot(olek2, 2)
```



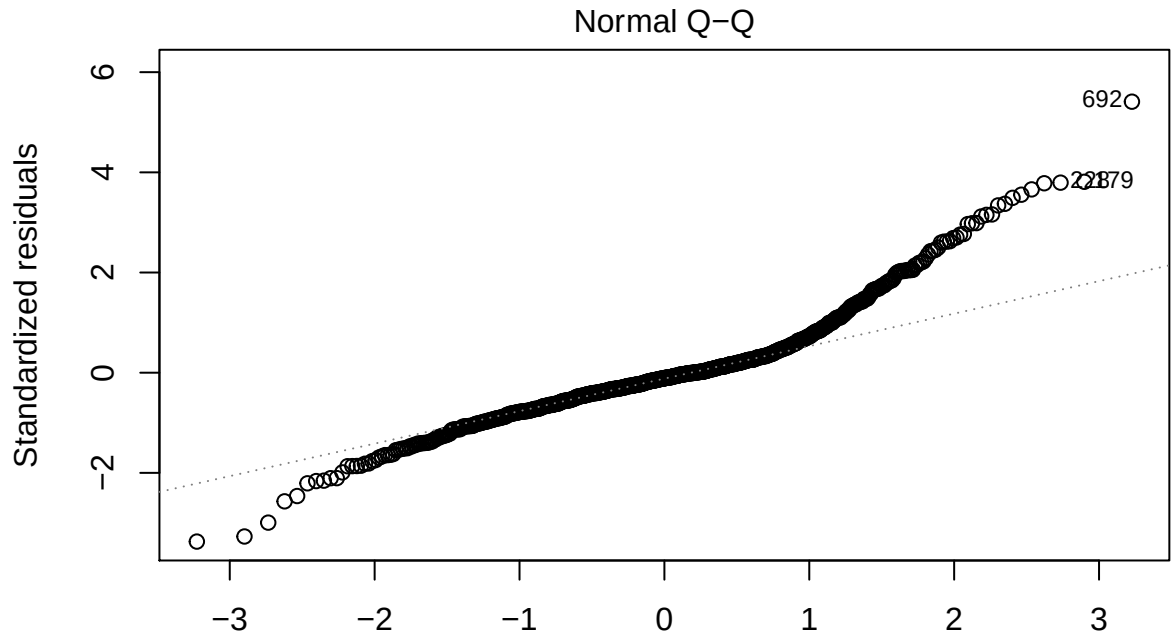

```
plot(olek2, 3)
```



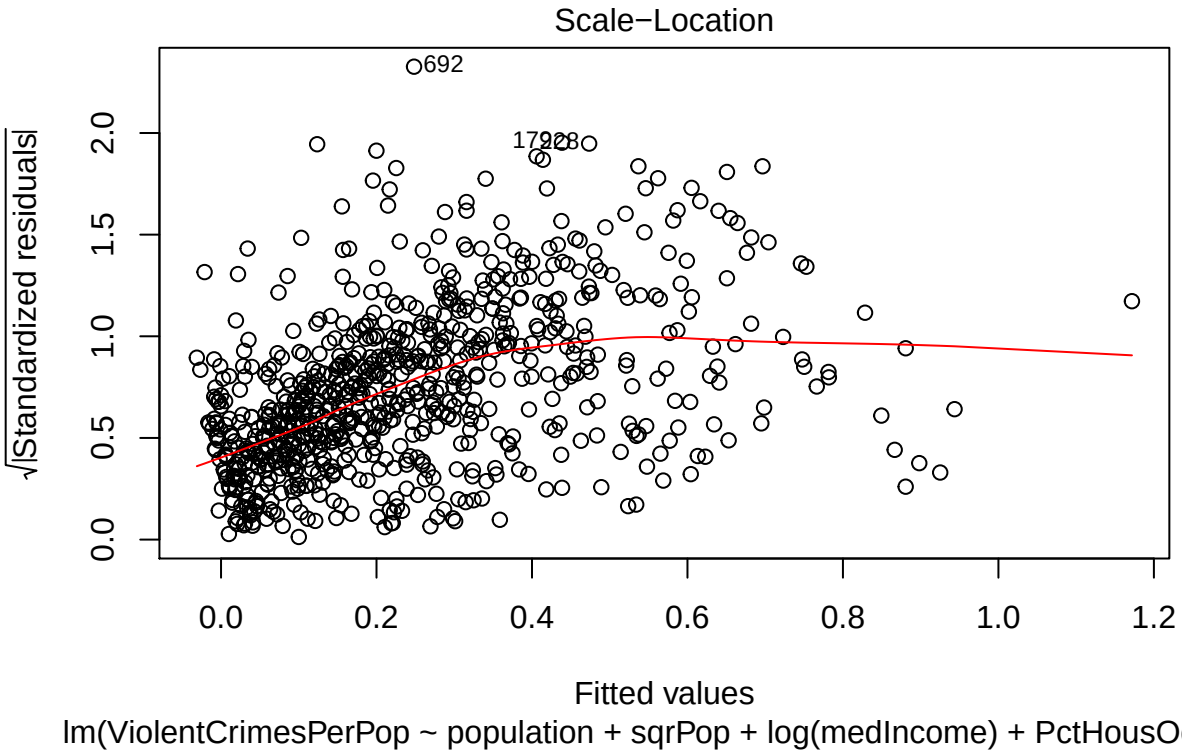

```
plot(ben2, 1)
```




```
plot(ben2, 2)
```


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(ViolentCrimesPerPop ~ population + sqrPop + log(medIncome) + PctHousOccu .

```
plot(ben2, 3)
```


Scale−Location
√|Standardized residuals|
Fitted values
lm(ViolentCrimesPerPop ~ population + sqrPop + log(medIncome) + PctHousOccu .

Additional Exercise

Let's try an automated selection method, to see what model pops out.

```
library(leaps)
d <- read.csv("http://andrewpbray.github.io/data/crime-train.csv")

# First let's trim the data
d_new <- d %>%
  select(population:PctSameState85,
         ViolentCrimesPerPop)

# Let's run the automated selection model
reg_mod <- regsubsets(ViolentCrimesPerPop ~ .,
                      data = d_new,
                      nvmax = 25,
                      method = "forward")

# Let's find the model which maximizes adjusted R2
summary(reg_mod)$which[which.max(summary(reg_mod)$adjr2),]
```

```
library(leaps)
d <- read.csv("http://andrewpbray.github.io/data/crime-train.csv")

# Let's trim the data
d_new <- d %>%
  select(population:PctSameState85,
         ViolentCrimesPerPop)

# The backward model
reg_mod2 <- regsubsets(ViolentCrimesPerPop ~ .,
                       data = d_new,
                       nvmax = 25,
                       method = "backward")

# Let's maximize our adjusted R2
summary(reg_mod2)$which[which.max(summary(reg_mod2)$adjr2),]
```

Let's fit the model with the variables which were in the best forward model

```
forward = lm(ViolentCrimesPerPop ~
               racePctHisp +
               PctNotHSGrad +
               PctEmplProfServ +
               FemalePctDiv +
               PctIlleg +
               PctImmigRec10 +
               PctVacMore6Mos +
               MedRent +
               NumStreet +
               pctUrban +
               pctWSocSec +
               MedRentPctHousInc +
               PctHousOccup +
               PctHousNoPhone +
               RentLowQ  +
               racePctWhite +
               indianPerCap +
               PctEmploy +
               MalePctDivorce +
               PctWorkMom +
               PctImmigRec5 +
               MedOwnCostPctIncNoMtg,
             data = d)

summary(forward)
```

```
##
## Call:
## lm(formula = ViolentCrimesPerPop ~ racePctHisp + PctNotHSGrad +
##     PctEmplProfServ + FemalePctDiv + PctIlleg + PctImmigRec10 +
##     PctVacMore6Mos + MedRent + NumStreet + pctUrban + pctWSocSec +
##     MedRentPctHousInc + PctHousOccup + PctHousNoPhone + RentLowQ +
##     racePctWhite + indianPerCap + PctEmploy + MalePctDivorce +
##     PctWorkMom + PctImmigRec5 + MedOwnCostPctIncNoMtg, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49444 -0.07353 -0.01442  0.05175  0.69495
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.05159    0.09228   0.559 0.576283
## racePctHisp            0.05404    0.03332   1.622 0.105258
## PctNotHSGrad           0.16876    0.06159   2.740 0.006286 **
## PctEmplProfServ        0.06530    0.04008   1.629 0.103681
## FemalePctDiv          -0.04272    0.08619  -0.496 0.620259
## PctIlleg               0.34559    0.04622   7.477 2.06e-13 ***
## PctImmigRec10          0.05763    0.05495   1.049 0.294665
## PctVacMore6Mos        -0.02876    0.03384  -0.850 0.395526
## MedRent                0.43753    0.09179   4.767 2.23e-06 ***
## NumStreet              0.18914    0.04044   4.677 3.44e-06 ***
## pctUrban               0.05205    0.01338   3.890 0.000109 ***
```

```
## pctWSocSec              0.13808    0.06273   2.201 0.028024 *
## MedRentPctHousInc       0.07703    0.03848   2.002 0.045647 *
## PctHousOccup           -0.10273    0.03307  -3.106 0.001964 **
## PctHousNoPhone         -0.07016    0.04419  -1.588 0.112777
## RentLowQ               -0.46576    0.09122  -5.106 4.14e-07 ***
## racePctWhite           -0.22903    0.04071  -5.625 2.58e-08 ***
## indianPerCap           -0.06421    0.03250  -1.976 0.048546 *
## PctEmploy               0.17195    0.08284   2.076 0.038245 *
## MalePctDivorce          0.35614    0.08284   4.299 1.93e-05 ***
## PctWorkMom             -0.10952    0.03973  -2.757 0.005977 **
## PctImmigRec5           -0.03489    0.04821  -0.724 0.469410
## MedOwnCostPctIncNoMtg -0.11068    0.03107  -3.562 0.000390 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1337 on 777 degrees of freedom
## Multiple R-squared:  0.685,  Adjusted R-squared:  0.6761
## F-statistic: 76.82 on 22 and 777 DF,  p-value: < 2.2e-16
```

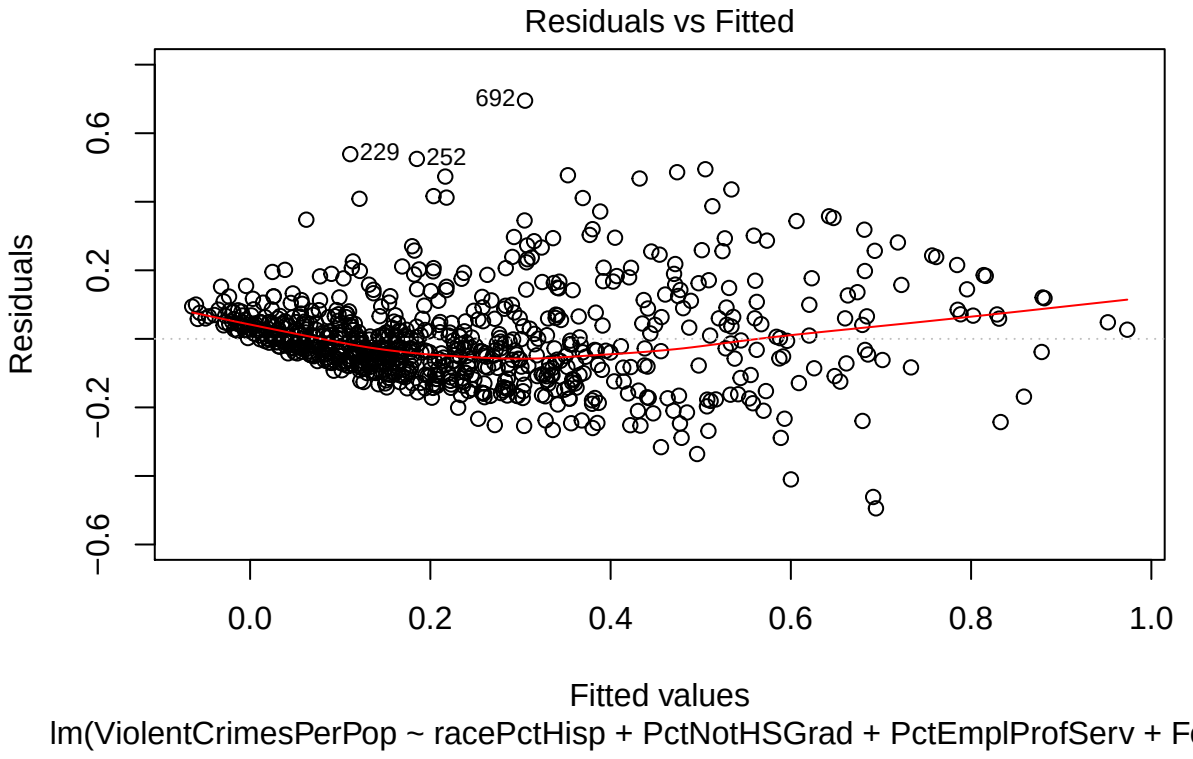
Now let's use the variables that were in the best backwards model

```
backward = lm(ViolentCrimesPerPop ~
                medFamInc +
                MalePctNevMarr +
                PctImmigRec8 +
                PersPerOccupHous +
                PctHousLess3BR +
                PctVacantBoarded +
                population +
                numbUrban +
                PctIlleg +
                PctVacMore6Mos +
                MedRent +
                NumStreet +
                pctWSocSec +
                agePct12t29 +
                medIncome +
                RentLowQ +
                racePctWhite +
                PctEmploy +
                MalePctDivorce +
                PctWorkMom +
                PctImmigRec5 +
                PctLargHouseOccup +
                MedOwnCostPctIncNoMtg,
              data = d)

summary(backward)
```

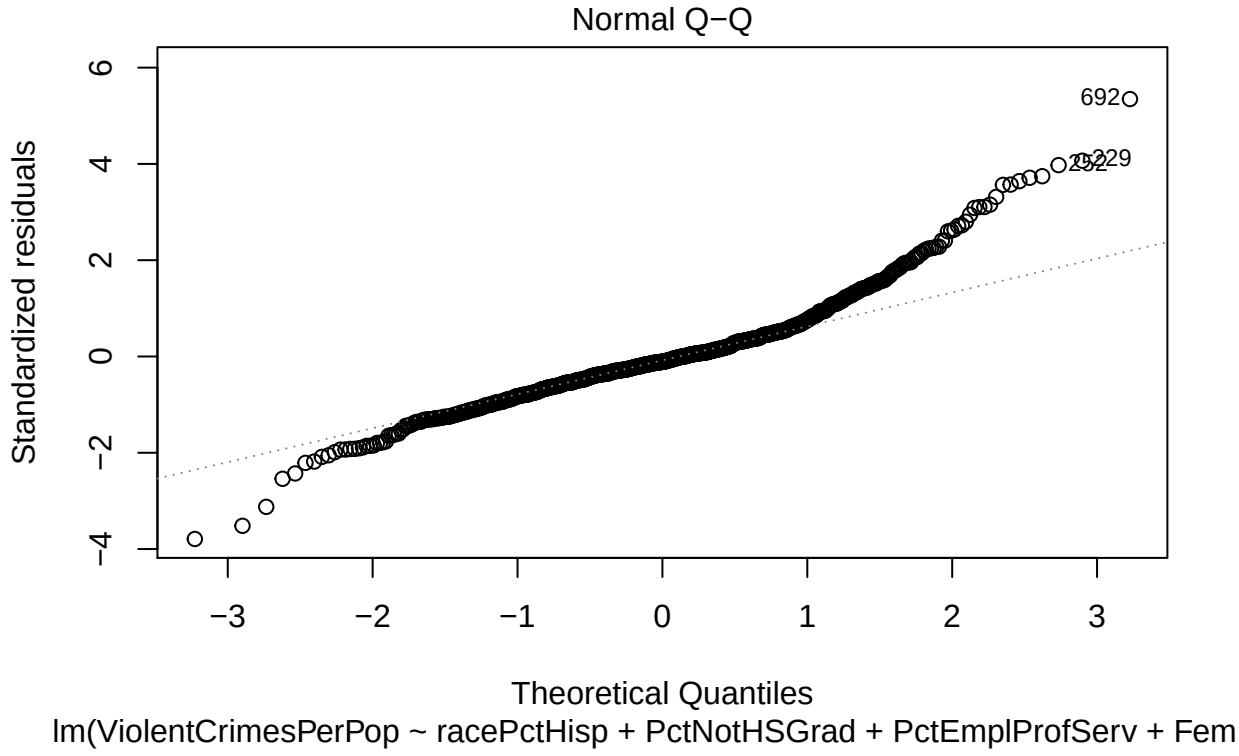
```
##
## Call:
## lm(formula = ViolentCrimesPerPop ~ medFamInc + MalePctNevMarr +
##     PctImmigRec8 + PersPerOccupHous + PctHousLess3BR + PctVacantBoarded +
##     population + numbUrban + PctIlleg + PctVacMore6Mos + MedRent +
##     NumStreet + pctWSocSec + agePct12t29 + medIncome + RentLowQ +
##     racePctWhite + PctEmploy + MalePctDivorce + PctWorkMom +
##     PctImmigRec5 + PctLargHouseOccup + MedOwnCostPctIncNoMtg,
##     data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54385 -0.07590 -0.01376  0.04941  0.67302
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.18669    0.12400  -1.506 0.132577
## medFamInc              0.44296    0.19165   2.311 0.021076 *
## MalePctNevMarr         0.11926    0.07017   1.700 0.089601 .
## PctImmigRec8           0.27606    0.07896   3.496 0.000499 ***
## PersPerOccupHous       0.49763    0.12298   4.047 5.72e-05 ***
## PctHousLess3BR         0.16578    0.05934   2.794 0.005340 **
## PctVacantBoarded       0.06903    0.03118   2.214 0.027135 *
## population            -1.22758    0.35971  -3.413 0.000676 ***
## numbUrban              1.28893    0.35825   3.598 0.000341 ***
```

```
## PctIlleg               0.32383    0.04882   6.634 6.14e-11 ***
## PctVacMore6Mos        -0.04129    0.03327  -1.241 0.214872
## MedRent                0.51233    0.10698   4.789 2.01e-06 ***
## NumStreet              0.16112    0.04901   3.288 0.001055 **
## pctWSocSec             0.22017    0.07676   2.868 0.004241 **
## agePct12t29           -0.13990    0.09021  -1.551 0.121345
## medIncome             -0.58086    0.22426  -2.590 0.009775 **
## RentLowQ              -0.44919    0.09177  -4.895 1.20e-06 ***
## racePctWhite          -0.21764    0.04061  -5.359 1.10e-07 ***
## PctEmploy              0.09871    0.07655   1.289 0.197613
## MalePctDivorce         0.34180    0.04981   6.862 1.39e-11 ***
## PctWorkMom            -0.08601    0.04090  -2.103 0.035793 *
## PctImmigRec5          -0.21657    0.07332  -2.954 0.003236 **
## PctLargHouseOccup     -0.21539    0.08422  -2.557 0.010736 *
## MedOwnCostPctIncNoMtg -0.10115    0.03033  -3.335 0.000894 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1327 on 776 degrees of freedom
## Multiple R-squared:  0.6903, Adjusted R-squared:  0.6812
## F-statistic: 75.21 on 23 and 776 DF,  p-value: < 2.2e-16
```

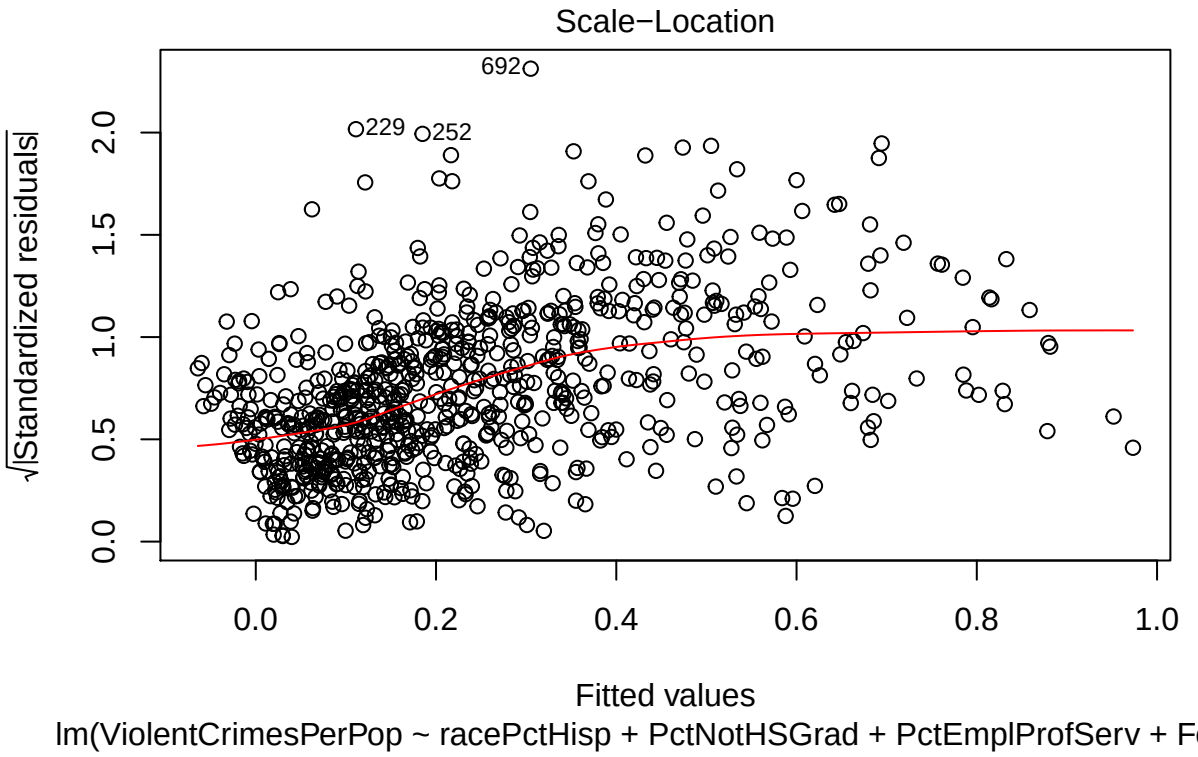
```
plot(forward, 1)
```



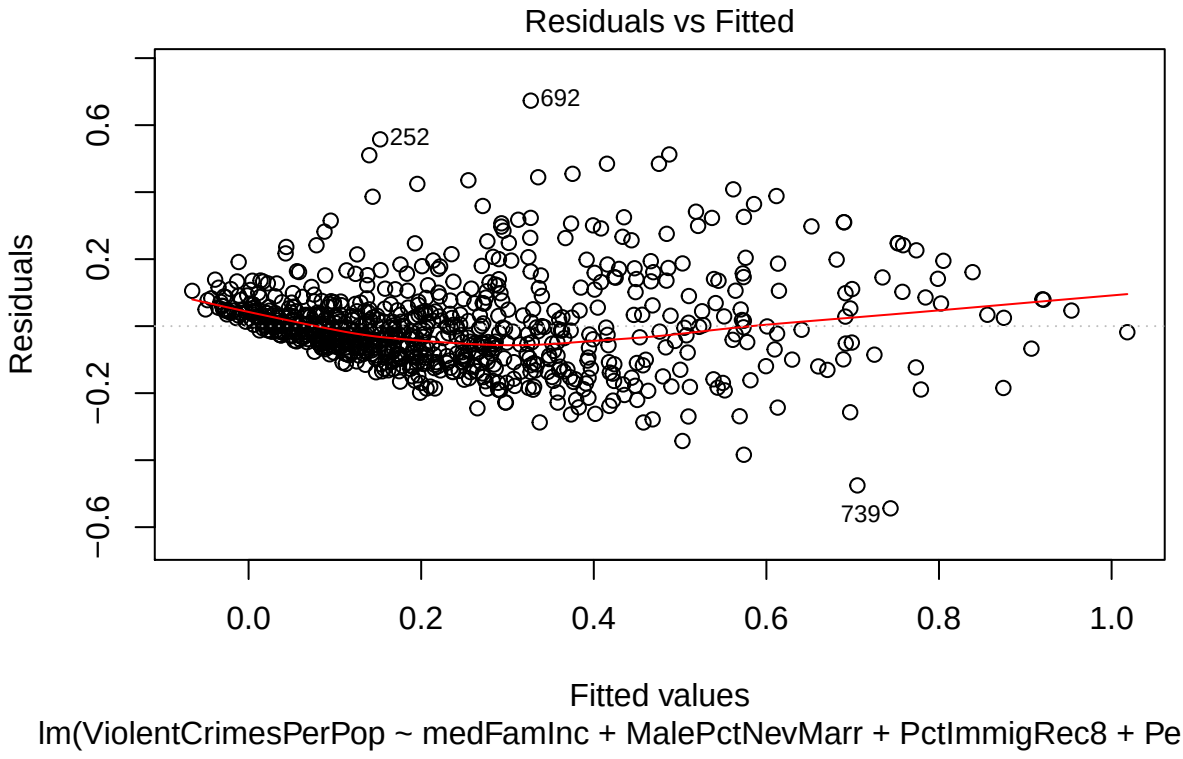

```
plot(forward, 2)
```



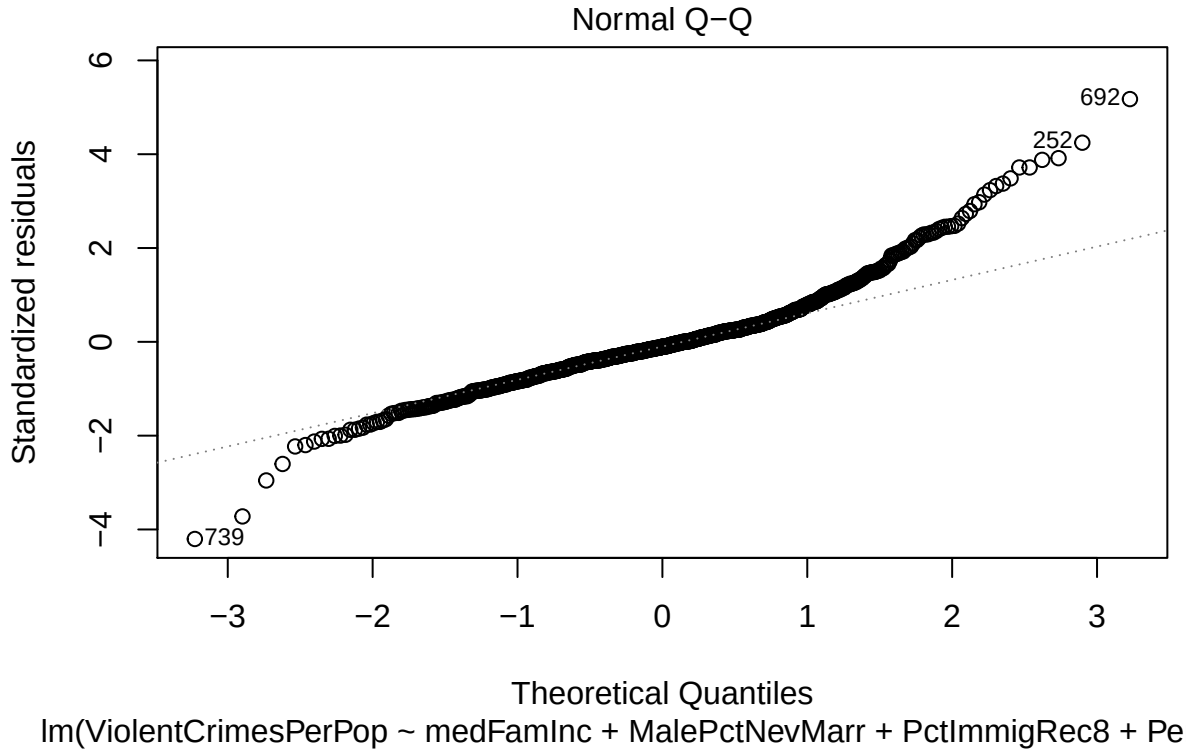

```
plot(forward, 3)
```


Scale−Location
692
229
252
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
Fitted values
lm(ViolentCrimesPerPop ~ racePctHisp + PctNotHSGrad + PctEmplProfServ + Fem

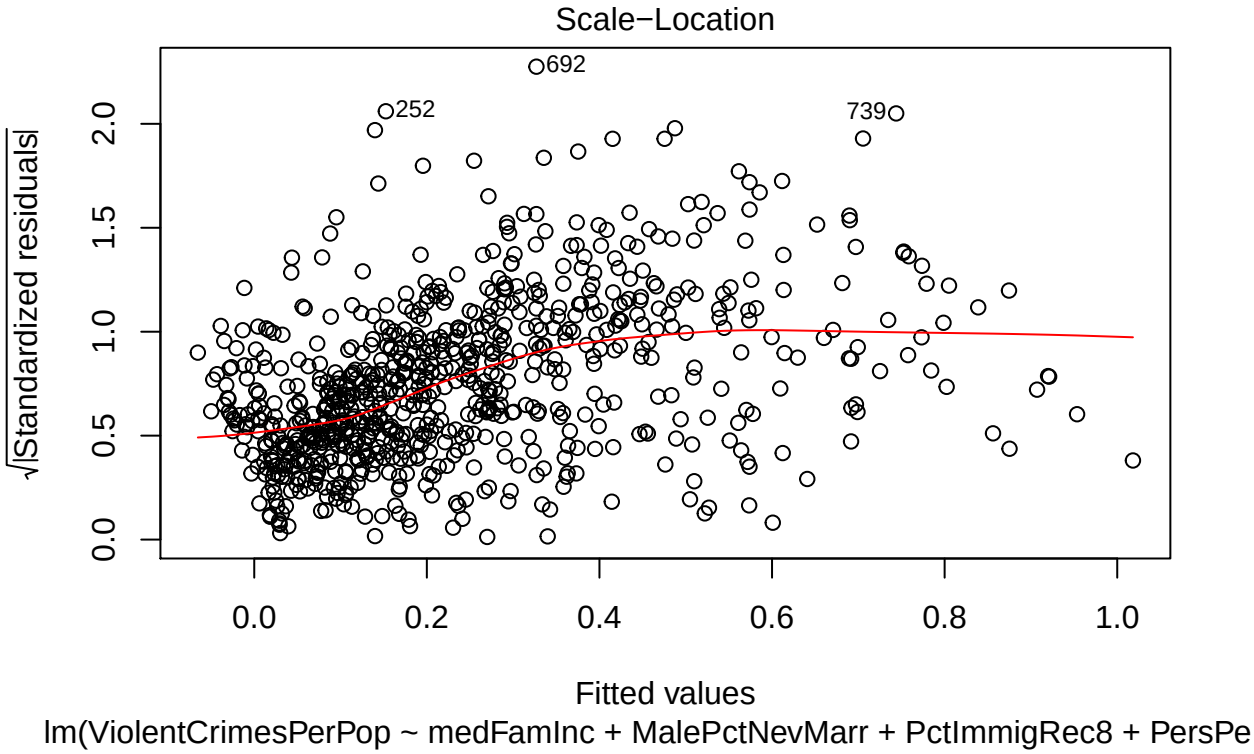
```
plot(backward, 1)
```




```
plot(backward, 2)
```




```
plot(backward, 3)
```



The backward method yielded a Adjusted R Squared which was higher than the forward method, as well as displying similar characteristics on the assessment plots, with heteroskedasticity remaining a problem.