Lab 4

Shan He, Joanna Huang, Tiffany Jaya

17 December 2017

Introduction

The purpose of this report is to generate policy suggestions based on our understanding of the determinants of crime in North Carolina in 1987. We will list out the limitations of our analysis, including any estimates that suffer from endogeneity bias.

Exploratory Data Analysis

```
# load the data
data <- read.csv("crime.csv")
# verify that it only contains data from 1987
unique(data$year)
```

```
## [1] 87
```

```
# list number of counties
length(unique(data$county))
```

```
## [1] 90
```

```
# list number of western, central, and urban counties
c(sum(data$west == 1), sum(data$central == 1), sum(data$urban == 1))
```

```
## [1] 21 34  8
```

```
# list number of western & urban counties and central & urban counties
c(sum(data$west == 1 & data$urban == 1), sum(data$central == 1 & data$urban == 1))
```

```
## [1] 1 5
```

```
# verify number of missing values
colSums(sapply(data, is.na))
```

```
##        X   county     year   crmrte   prbarr  prbconv  prbpris   avgsen
##        0        0        0        0        0        0        0        0
##    polpc  density    taxpc     west  central    urban pctmin80     wcon
##        0        0        0        0        0        0        0        0
##     wtuc     wtrd     wfir     wser     wmfg     wfed     wsta     wloc
##        0        0        0        0        0        0        0        0
##      mix  pctymle
##        0        0
```

The dataset contains 90 counties from North Carolina, all of which is collected in 1987. Out of the 90 counties, 21 are from western NC (out of which 1 is also urban), 34 are from central NC (out of which 5 is also urban), and 8 are considered urban counties. There are no missing values which will make our analysis easier.

For now, we will not take into consideration probabilities that are greater than 1 or less than 0 as well as percentages that are greater than 1 or less than 0. The assumption is that probabilities are in the range [0, 1]

and percentages are in the range [0, 100]. Until we know the reason why the values are outside their range, we will not employ datapoints that do not conform to this assumption.

```
# list number of probabilities (prbarr, prbconv, prbpris, mix) that are not in range [0, 1]
c(sum(data$prbarr < 0 | 1 < data$prbarr), sum(data$prbconv < 0 | 1 < data$prbconv),
  sum(data$prbpris < 0 | 1 < data$prbpris), sum(data$mix < 0 | 1 < data$mix))
```

```
## [1]  1 10  0  0
```

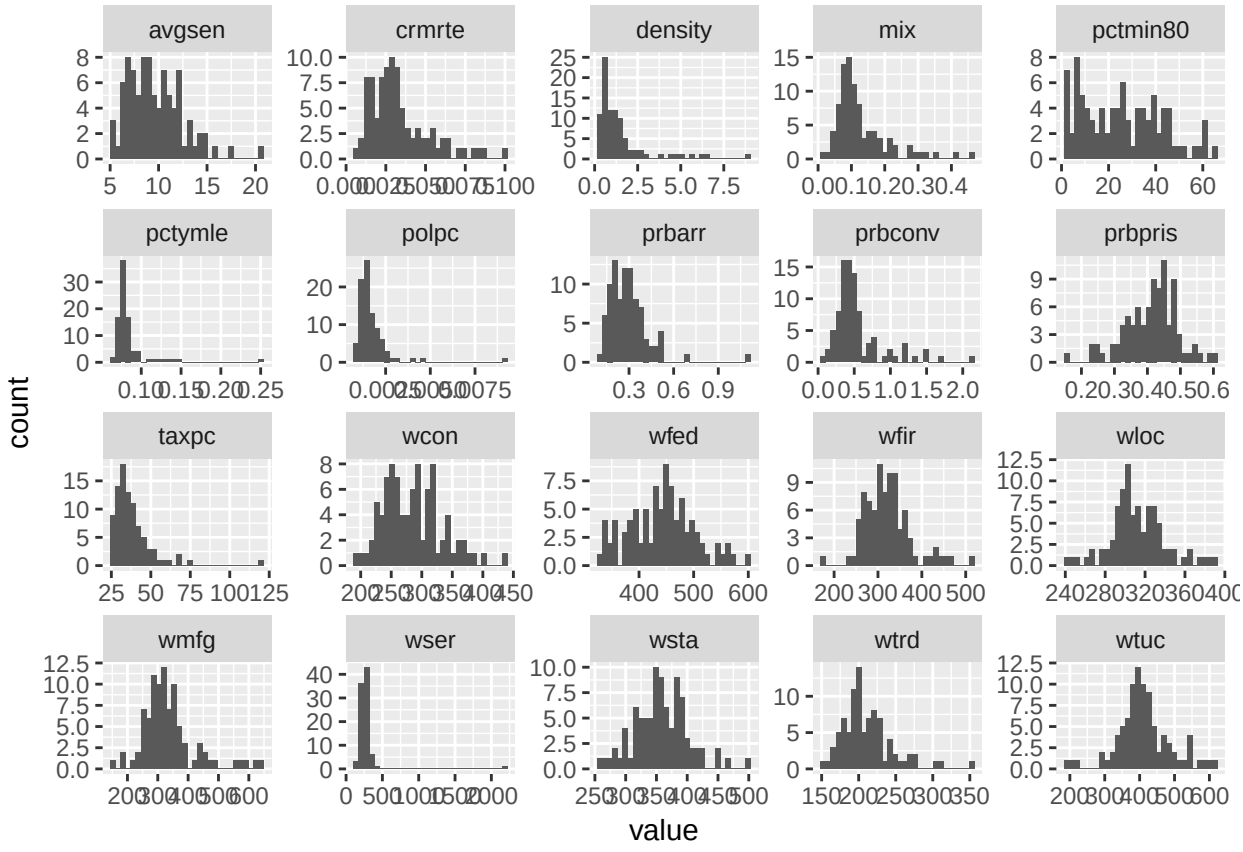
```
# list number of percentages (pctymle, pctmin80) that are not in range [0, 100]
c(sum(data$pctymle < 0 | 100 < data$pctymle), sum(data$pctmin80 < 0 | 100 < data$pctmin80))
```

```
## [1] 0 0
```

prbarr and *prbconv* contain 1 and 10 datapoints respectively that do not conform to the probability assumption.

We then plot each numeric variable in a histogram to see its sample distribution.

```
# plot every variable except X, county, year, west, central, urban
library(ggplot2) # ggplot
library(tidyr) # gather
num.data <- data[!(names(data) %in% c("X", "county", "year", "west", "central", "urban"))]
ggplot(gather(num.data), aes(value)) +
       facet_wrap(~key, scales="free") +
       geom_histogram()
```




```
library(moments) # skewness
skewness(num.data)
```

```
##      crmrte      prbarr     prbconv     prbpris      avgsen       polpc
##  1.28174888  2.52529596  2.03950599 -0.45254022  1.00116340  4.98348795
##     density       taxpc    pctmin80        wcon        wtuc        wtrd
##  2.65301071  3.29057447  0.36566169  0.60680223  0.06819768  1.46120657
##        wfir        wser        wmfg        wfed        wsta        wloc
##  0.82063146  8.69918165  1.42253166  0.13223761  0.36236826  0.29513808
##         mix     pctymle
##  1.91657046  4.56069073
```

Most of the sample distributions appear to be positively skewed. We will take into consideration of logarithmic transformations, depending whether the interpretations make sense, when it is time to include the variables into the regression model.

From the histograms, we also see several notable outliers. We are under the impression that a county which has outlier in one variable will likely have outlier in another variable. For this reason, we have listed counties which have repeated outliers when we iterate through the entire numeric variables.

```
# iterate through each numeric variable and list the outlier counties and their respective frequency
county.ids <- c()
for(var in num.data) {
  var.out <- boxplot.stats(var)$out
  county.ids <- c(county.ids, data[var %in% var.out, ]$county)
}
table(county.ids)
```

```
## county.ids
##   1   3   5   7  11  19  35  39  49  51  53  55  63  67  69  71  79  81
##   1   1   1   1   2   4   2   2   1   3   1   3   5   1   3   2   1   2
##  85  87  93  99 105 111 113 115 119 123 127 129 131 133 135 137 139 143
##   1   1   1   2   1   1   1   5  10   1   2   3   1   1   2   2   1   2
## 147 149 169 173 175 181 183 185 187 189 195 197
##   1   1   1   4   1   2   4   2   1   1   2   1
```

```
# list the most extreme outlier
library(outliers) # outlier
outlier(num.data)
```

```
##        crmrte        prbarr       prbconv       prbpris        avgsen
##    0.09896590    1.09090996    2.12121010    0.15000001   20.70000076
##         polpc       density         taxpc      pctmin80          wcon
##    0.00905433    8.82765198  119.76145172   64.34819794  436.76663208
##          wtuc          wtrd          wfir          wser          wmfg
##  187.61726379  354.67611694  509.46551514 2177.06811523  646.84997559
##          wfed          wsta          wloc           mix       pctymle
##  597.95001221  499.58999634  388.08999634    0.46511629    0.24871162
```

One outlier that is interesting to note is that the weekly wage in the service industry for county with id 185 is $2177.10, which is approximately eight times higher than the median. We do not know if the value is inputted incorrectly or if the county in general is making a weekly wage of $2177.10 in the service industry.

```
summary(data$wser)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.3   253.1   275.3   277.6  2177.1
```

Research Question

James Q. Wilson and George Kelling's "broken windows theory" in 1982 led to a nation-wide movement for stricter crime-fighting policies between the 1980s and 1990s. The theory states:

> *if the first broken window in a building is not repaired, then people who like breaking windows will assume that no one cares about the building and more windows will be broken. Soon the building will have no windows. . . .*

The belief was that by adopting a zero tolerance approach that enforced even the lowest level offenses, crime rates would subsequently go down. While New York City notably enforced this more stringent approach, San Francisco went the opposite direction of less strident law enforcement policies that reduced arrests, prosecutions and incarceration rates. Both sides experienced considerable declines in crime rates. Thus we hope to test the "broken windows theory" for the counties of South Carolina in 1987 and answer the question: Does the conservative approach of deterrence through arrests, incapacitation through imprisonment, harsh sentencing and higher police per capita lead to lower crime rates?

Model 1: only the explanatory variables of key interest

Based on the exploratory data analysis and goal of testing the "broken windows theory", our initial proposed model will include all variables related to stricter law enforcement policies: *prbarr*, *prbconv*, *prbpris*, *avgsen*, and *polpc*. Assuming the "broken windows theory" is valid, we expect generally negative coefficients for all variables.

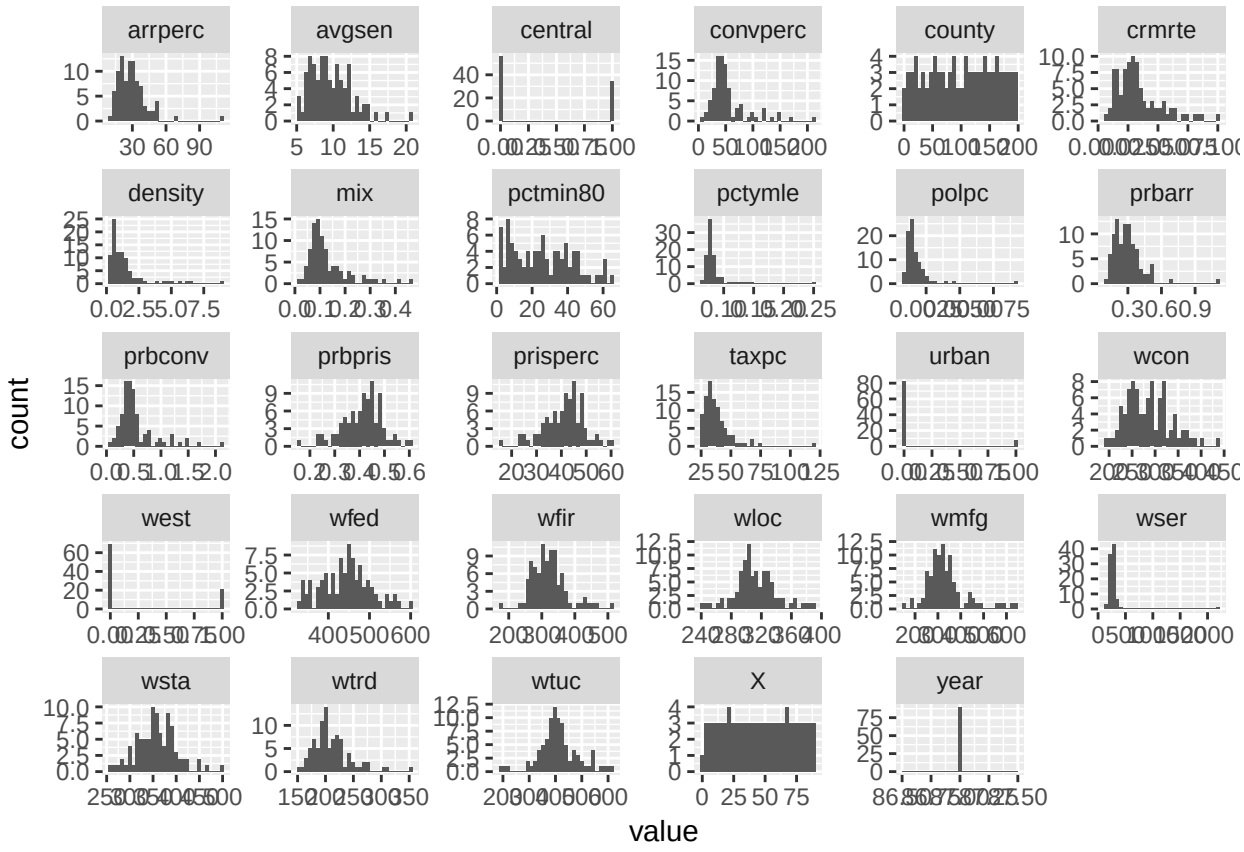
To make *prbarr*, *prbconv* and *prbpris* interpretation more intuitive to understand, we decided to transform them into percentages, and will use these new variables moving forward.

```
data$arrperc = data$prbarr * 100
data$convperc = data$prbconv * 100
data$prisperc = data$prbpris * 100
```

In our EDA, we noted that crmrte, prbarr, prbconv, avgsen, and polpc are positively skewed, thus we should consider taking the log. Since the values cannot be negative and have meaningful zero-points, this decision is valid. It is important to remember moving forward that the interpretation of arrperc, and convperc will be in terms of percentage point changes rather than percentage changes.

```
plot.data <- na.omit(data[, sapply(data, is.numeric)])
 ggplot(gather(plot.data), aes(value)) +
        facet_wrap(~key, scales="free") +
        geom_histogram()
```

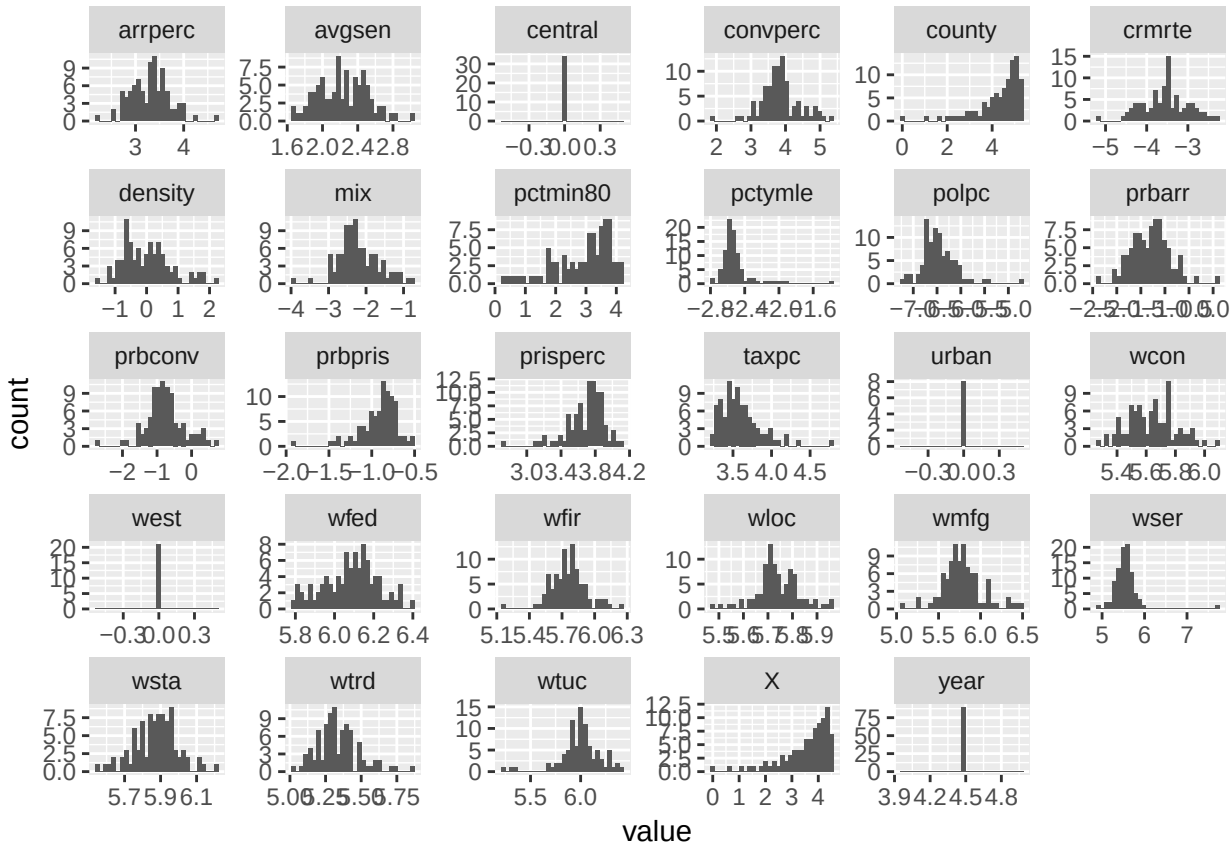
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



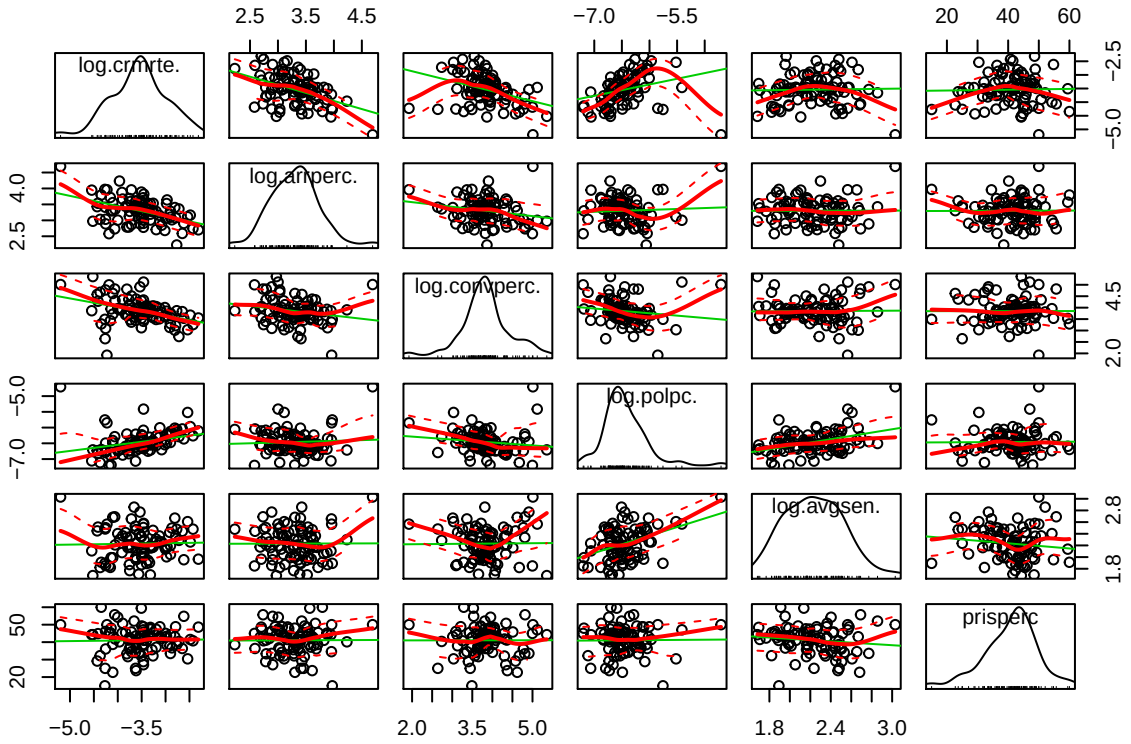

```
plot.data <- na.omit(log(data[, sapply(data, is.numeric)]))
 ggplot(gather(plot.data), aes(value)) +
        facet_wrap(~key, scales="free") +
        geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 207 rows containing non-finite values (stat_bin).
```




```
library(car)
scatterplotMatrix(~log(crmrte) + log(arrperc) + log(convperc) + log(polpc) + log(avgsen) + prisperc, da
```




```
 + cor(data[, c("crmrte", "arrperc", "convperc", "polpc", "avgsen", "prisperc")])
```

```
##               crmrte     arrperc    convperc      polpc      avgsen
## crmrte     1.00000000 -0.39528302 -0.38596559 0.16728162  0.01979653
## arrperc   -0.39528302  1.00000000 -0.05579621 0.42596481  0.17869425
## convperc  -0.38596559 -0.05579621  1.00000000 0.17186516  0.15585232
## polpc      0.16728162  0.42596481  0.17186516 1.00000000  0.48815230
## avgsen     0.01979653  0.17869425  0.15585232 0.48815230  1.00000000
## prisperc   0.04799540  0.04583324  0.01102265 0.04820783 -0.09468083
##              prisperc
## crmrte     0.04799540
## arrperc    0.04583324
## convperc   0.01102265
## polpc      0.04820783
## avgsen    -0.09468083
## prisperc   1.00000000
```

```
 + cor(data[!(names(data) %in% c("X", "county", "year"))], use = "complete.obs")
```

```
##               crmrte      prbarr     prbconv      prbpris      avgsen
## crmrte     1.00000000 -0.39528302 -0.38596559  0.047995395  0.01979653
## prbarr    -0.39528302  1.00000000 -0.05579621  0.045833245  0.17869425
## prbconv   -0.38596559 -0.05579621  1.00000000  0.011022645  0.15585232
## prbpris    0.04799540  0.04583324  0.01102265  1.000000000 -0.09468083
## avgsen     0.01979653  0.17869425  0.15585232 -0.094680833  1.00000000
## polpc      0.16728162  0.42596481  0.17186516  0.048207825  0.48815230
## density    0.72777835 -0.30053317 -0.22791204  0.072609846  0.07159560
```

```
## taxpc     0.44871512 -0.13719105 -0.12738963 -0.092360509  0.08654323
## west     -0.38033874  0.18649897  0.07198455 -0.035044526  0.09855151
## central   0.16588032 -0.16888612 -0.04640007  0.164520114 -0.15816897
## urban     0.61506307 -0.20856276 -0.19709186  0.050354117  0.14391388
## pctmin80  0.18165059  0.04907002  0.06249824  0.106136091 -0.16633664
## wcon      0.39296155 -0.25183650 -0.11745577 -0.059611223 -0.03030263
## wtuc      0.23599574 -0.07035781 -0.00716159  0.124730237  0.23116592
## wtrd      0.42722262 -0.09948428 -0.13454762  0.139338689  0.10822274
## wfir      0.33602609 -0.17253501  0.03217747  0.032777974  0.17792907
## wser     -0.05206995 -0.13133303  0.45666832  0.038011073 -0.15103677
## wmfg      0.35256117 -0.15316974  0.01757978  0.009408759  0.11045461
## wfed      0.48991634 -0.20792619 -0.06085923  0.084965065  0.15240383
## wsta      0.19984674 -0.16253921 -0.12843449 -0.031213974  0.12840868
## wloc      0.35982934 -0.02447781  0.05060548  0.081193439  0.14575388
## mix      -0.13200035  0.41289804 -0.30425124  0.116588825 -0.14170497
## pctymle   0.29033966 -0.18096201 -0.16222602 -0.082759753  0.07099989
## arrperc  -0.39528302  1.00000000 -0.05579621  0.045833245  0.17869425
## convperc -0.38596559 -0.05579621  1.00000000  0.011022645  0.15585232
## prisperc  0.04799540  0.04583324  0.01102265  1.000000000 -0.09468083
##                polpc     density       taxpc          west     central
## crmrte    0.16728162  0.72777835  0.44871512 -0.3803387441  0.16588032
## prbarr    0.42596481 -0.30053317 -0.13719105  0.1864989678 -0.16888612
## prbconv   0.17186516 -0.22791204 -0.12738963  0.0719845544 -0.04640007
## prbpris   0.04820783  0.07260985 -0.09236051 -0.0350445258  0.16452011
## avgsen    0.48815230  0.07159560  0.08654323  0.0985515064 -0.15816897
## polpc     1.00000000  0.16152857  0.28055315  0.1441740336 -0.04600949
## density   0.16152857  1.00000000  0.32047367 -0.1945888906  0.35682850
## taxpc     0.28055315  0.32047367  1.00000000 -0.1738761174  0.03361974
## west      0.14417403 -0.19458889 -0.17387612  1.0000000000 -0.42986348
## central  -0.04600949  0.35682850  0.03361974 -0.4298634768  1.00000000
## urban     0.15770869  0.82068254  0.34574617 -0.0800034085  0.15927023
## pctmin80 -0.16911752 -0.07470698 -0.02797739 -0.6245144292 -0.05487554
## wcon     -0.02379236  0.45134939  0.26395677 -0.1923563276  0.39786886
## wtuc      0.17277373  0.33119447  0.17129001  0.0217295819  0.18855844
## wtrd      0.12384123  0.59414742  0.18392144 -0.1913375272  0.38668510
## wfir      0.19522607  0.54597415  0.13094363 -0.0528935846  0.29060049
## wser     -0.01638582  0.04344734  0.07594777 -0.0633567781  0.19261249
## wmfg      0.27043619  0.43766213  0.25860844 -0.0106669263  0.17368746
## wfed      0.16187035  0.58693219  0.06207230 -0.2106600886  0.34923553
## wsta      0.04891417  0.22310548 -0.03498830 -0.0785093734  0.08527707
## wloc      0.38698768  0.46001747  0.21990116 -0.1429070937  0.33323127
## mix       0.02411189 -0.13172771 -0.04355958  0.0008159465 -0.09210923
## pctymle   0.05022177  0.11478144 -0.09154375 -0.0362124738 -0.10371067
## arrperc   0.42596481 -0.30053317 -0.13719105  0.1864989678 -0.16888612
## convperc  0.17186516 -0.22791204 -0.12738963  0.0719845544 -0.04640007
## prisperc  0.04820783  0.07260985 -0.09236051 -0.0350445258  0.16452011
##                urban    pctmin80        wcon        wtuc         wtrd
## crmrte    0.61506307  0.18165059  0.39296155  0.23599574  0.427222622
## prbarr   -0.20856276  0.04907002 -0.25183650 -0.07035781 -0.099484278
## prbconv  -0.19709186  0.06249824 -0.11745577 -0.00716159 -0.134547618
## prbpris   0.05035412  0.10613609 -0.05961122  0.12473024  0.139338689
## avgsen    0.14391388 -0.16633664 -0.03030263  0.23116592  0.108222741
## polpc     0.15770869 -0.16911752 -0.02379236  0.17277373  0.123841229
## density   0.82068254 -0.07470698  0.45134939  0.33119447  0.594147416
```

```
## taxpc     0.34574617 -0.02797739  0.26395677  0.17129001  0.183921439
## west     -0.08000341 -0.62451443 -0.19235633  0.02172958 -0.191337527
## central   0.15927023 -0.05487554  0.39786886  0.18855844  0.386685101
## urban     1.00000000  0.01619569  0.31926691  0.22632785  0.431728441
## pctmin80  0.01619569  1.00000000 -0.10793251 -0.18913279 -0.064824402
## wcon      0.31926691 -0.10793251  1.00000000  0.40937889  0.564058565
## wtuc      0.22632785 -0.18913279  0.40937889  1.00000000  0.351683658
## wtrd      0.43172844 -0.06482440  0.56405857  0.35168366  1.000000000
## wfir      0.40171167 -0.07717356  0.48893774  0.32761956  0.668154525
## wser      0.05589097  0.19672114 -0.01316438 -0.01924404 -0.020741268
## wmfg      0.40362299 -0.11688213  0.34739282  0.46892658  0.371487416
## wfed      0.42602595  0.03081152  0.50666394  0.39866915  0.640521866
## wsta      0.30194045  0.09274887 -0.01885609 -0.15340397  0.007267295
## wloc      0.33835635 -0.10590108  0.51704129  0.33301976  0.581463886
## mix      -0.06417238  0.20123542 -0.19587213 -0.25346871 -0.125754703
## pctymle   0.09396449 -0.01925657 -0.02214779 -0.10249879 -0.109277017
## arrperc  -0.20856276  0.04907002 -0.25183650 -0.07035781 -0.099484278
## convperc -0.19709186  0.06249824 -0.11745577 -0.00716159 -0.134547618
## prisperc  0.05035412  0.10613609 -0.05961122  0.12473024  0.139338689
##                wfir         wser         wmfg        wfed         wsta
## crmrte    0.33602609 -0.052069955  0.352561171  0.48991634  0.199846744
## prbarr   -0.17253501 -0.131333029 -0.153169735 -0.20792619 -0.162539207
## prbconv   0.03217747  0.456668322  0.017579785 -0.06085923 -0.128434487
## prbpris   0.03277797  0.038011073  0.009408759  0.08496507 -0.031213974
## avgsen    0.17792907 -0.151036772  0.110454606  0.15240383  0.128408685
## polpc     0.19522607 -0.016385818  0.270436194  0.16187035  0.048914168
## density   0.54597415  0.043447343  0.437662125  0.58693219  0.223105478
## taxpc     0.13094363  0.075947768  0.258608438  0.06207230 -0.034988299
## west     -0.05289358 -0.063356778 -0.010666926 -0.21066009 -0.078509373
## central   0.29060049  0.192612486  0.173687456  0.34923553  0.085277071
## urban     0.40171167  0.055890972  0.403622994  0.42602595  0.301940450
## pctmin80 -0.07717356  0.196721137 -0.116882126  0.03081152  0.092748871
## wcon      0.48893774 -0.013164375  0.347392817  0.50666394 -0.018856088
## wtuc      0.32761956 -0.019244042  0.468926579  0.39866915 -0.153403974
## wtrd      0.66815452 -0.020741268  0.371487416  0.64052187  0.007267295
## wfir      1.00000000  0.013716140  0.497583408  0.62317882  0.240700059
## wser      0.01371614  1.000000000  0.008986754  0.02067471  0.037471156
## wmfg      0.49758341  0.008986754  1.000000000  0.51823047  0.052336590
## wfed      0.62317882  0.020674709  0.518230474  1.00000000  0.188250660
## wsta      0.24070006  0.037471156  0.052336590  0.18825066  1.000000000
## wloc      0.55443563  0.076971337  0.450453501  0.51941357  0.164641269
## mix      -0.21232339 -0.173562869 -0.344125134 -0.31220529 -0.075726032
## pctymle   0.01075555 -0.043107714  0.024179451 -0.06046726  0.218316221
## arrperc  -0.17253501 -0.131333029 -0.153169735 -0.20792619 -0.162539207
## convperc  0.03217747  0.456668322  0.017579785 -0.06085923 -0.128434487
## prisperc  0.03277797  0.038011073  0.009408759  0.08496507 -0.031213974
##                wloc           mix      pctymle     arrperc    convperc
## crmrte    0.359829341 -0.1320003539  0.290339658 -0.39528302 -0.38596559
## prbarr   -0.024477813  0.4128980444 -0.180962011  1.00000000 -0.05579621
## prbconv   0.050605485 -0.3042512443 -0.162226023 -0.05579621  1.00000000
## prbpris   0.081193439  0.1165888249 -0.082759753  0.04583324  0.01102265
## avgsen    0.145753884 -0.1417049658  0.070999887  0.17869425  0.15585232
## polpc     0.386987678  0.0241118925  0.050221768  0.42596481  0.17186516
## density   0.460017473 -0.1317277105  0.114781444 -0.30053317 -0.22791204
```

```
## taxpc     0.219901156 -0.0435595792 -0.091543750 -0.13719105 -0.12738963
## west     -0.142907094  0.0008159465 -0.036212474  0.18649897  0.07198455
## central   0.333231267 -0.0921092281 -0.103710667 -0.16888612 -0.04640007
## urban     0.338356350 -0.0641723765  0.093964486 -0.20856276 -0.19709186
## pctmin80 -0.105901082  0.2012354175 -0.019256570  0.04907002  0.06249824
## wcon      0.517041291 -0.1958721285 -0.022147787 -0.25183650 -0.11745577
## wtuc      0.333019765 -0.2534687080 -0.102498785 -0.07035781 -0.00716159
## wtrd      0.581463886 -0.1257547028 -0.109277017 -0.09948428 -0.13454762
## wfir      0.554435635 -0.2123233861  0.010755553 -0.17253501  0.03217747
## wser      0.076971337 -0.1735628695 -0.043107714 -0.13133303  0.45666832
## wmfg      0.450453501 -0.3441251344  0.024179451 -0.15316974  0.01757978
## wfed      0.519413570 -0.3122052928 -0.060467265 -0.20792619 -0.06085923
## wsta      0.164641269 -0.0757260320  0.218316221 -0.16253921 -0.12843449
## wloc      1.000000000 -0.2535193780 -0.001651489 -0.02447781  0.05060548
## mix      -0.253519378  1.0000000000 -0.092856609  0.41289804 -0.30425124
## pctymle  -0.001651489 -0.0928566094  1.000000000 -0.18096201 -0.16222602
## arrperc  -0.024477813  0.4128980444 -0.180962011  1.00000000 -0.05579621
## convperc  0.050605485 -0.3042512443 -0.162226023 -0.05579621  1.00000000
## prisperc  0.081193439  0.1165888249 -0.082759753  0.04583324  0.01102265
##              prisperc
## crmrte    0.047995395
## prbarr    0.045833245
## prbconv   0.011022645
## prbpris   1.000000000
## avgsen   -0.094680833
## polpc     0.048207825
## density   0.072609846
## taxpc    -0.092360509
## west     -0.035044526
## central   0.164520114
## urban     0.050354117
## pctmin80  0.106136091
## wcon     -0.059611223
## wtuc      0.124730237
## wtrd      0.139338689
## wfir      0.032777974
## wser      0.038011073
## wmfg      0.009408759
## wfed      0.084965065
## wsta     -0.031213974
## wloc      0.081193439
## mix       0.116588825
## pctymle  -0.082759753
## arrperc   0.045833245
## convperc  0.011022645
## prisperc  1.000000000
```

Looking at the relationships between crmrte (y) and the newly transformed x variables, it appears that the relationships are linear and we can continue with our proposed multiregression model.

The model:

$$log(crmrte) = \beta_0 + \beta_1 log(arrperc) + \beta_2 log(convperc) + \beta_3 (prisperc) + \beta_4 log(avgsen) + \beta_5 log(polpc) + \mu$$

We will now run the model and test the validity of the 6 CLM assumptions:

```
m1 = lm(log(crmrte) ~ log(arrperc) + log(convperc) + prisperc + log(avgsen) + log(polpc), data=data)
```

CLM 1 - A linear model

The model is specified such that the dependent variable is a linear function of the explanatory variables.

Is the assumption valid? **Yes**

Response: No response required.

CLM 2 - Random Sampling

Is the assumption valid?

Response:

CLM 3 - Multicollinearity

As a quick test of the multicollinearity condition, we check the correlation of the two explanatory variables and their Variance Inflation Factors (VIF):

```
data$logarrperc = log(data$arrperc)
data$logconvperc = log(data$convperc)
data$logpolpc = log(data$polpc)
data$logavgsen = log(data$avgsen)
X = data.matrix(subset(data, select=c("logarrperc", "logconvperc", "logpolpc", "logavgsen", "prisperc"))
(Cor = cor(X))
```

```
##               logarrperc  logconvperc    logpolpc    logavgsen
## logarrperc   1.000000000 -0.202355412  0.05424285  0.003832922
## logconvperc -0.202355412  1.000000000 -0.13863931  0.011629953
## logpolpc     0.054242853 -0.138639312  1.00000000  0.395079636
## logavgsen    0.003832922  0.011629953  0.39507964  1.000000000
## prisperc     0.005955257  0.001618678  0.01041348 -0.127111767
##                 prisperc
## logarrperc   0.005955257
## logconvperc  0.001618678
## logpolpc     0.010413481
## logavgsen   -0.127111767
## prisperc     1.000000000
```

```
vif(m1)
```

```
##  log(arrperc) log(convperc)      prisperc   log(avgsen)    log(polpc)
##      1.043513      1.066315      1.021166      1.216359      1.220948
```

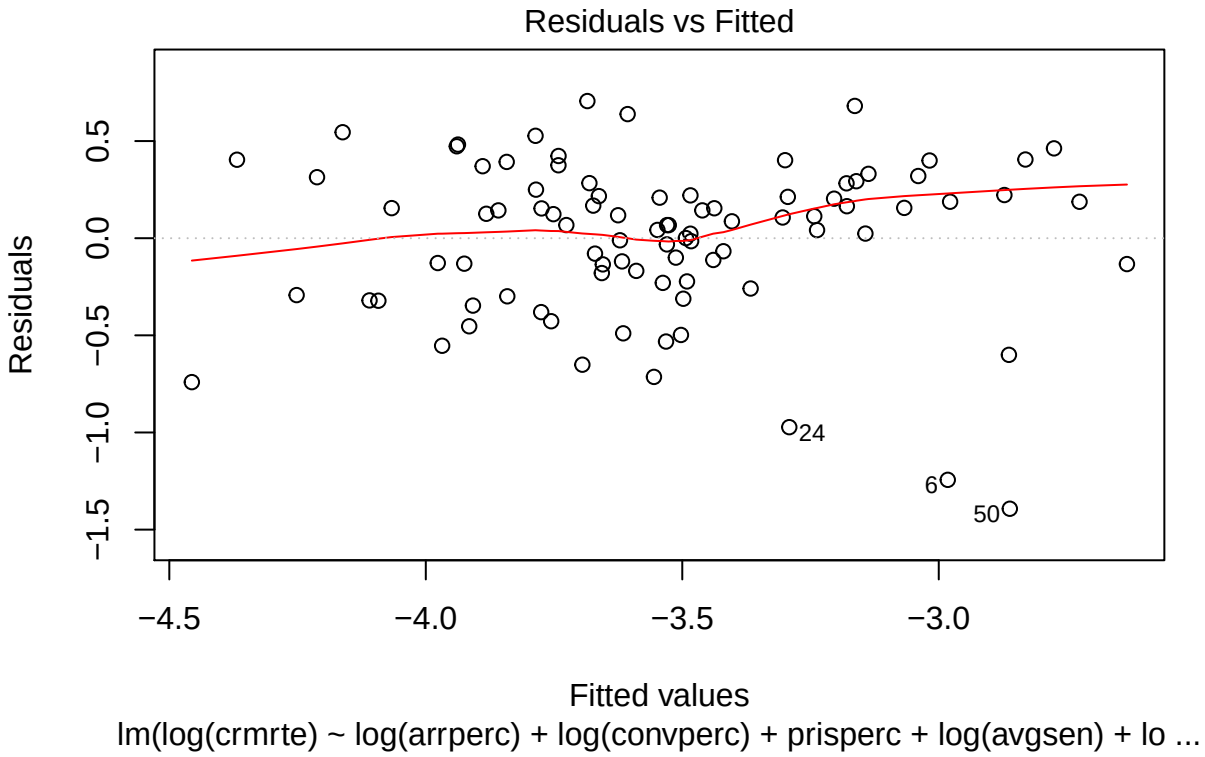
The explanatory variables (logarrperc, logconvperc, logpolpc, logavgsen and prisperc) are not perfectly correlated and the VIFs are low (i.e. less than 10), so there is no perfect multicollinearity of the independent variables.

Is the assumption valid? **Yes Response**: No response required.

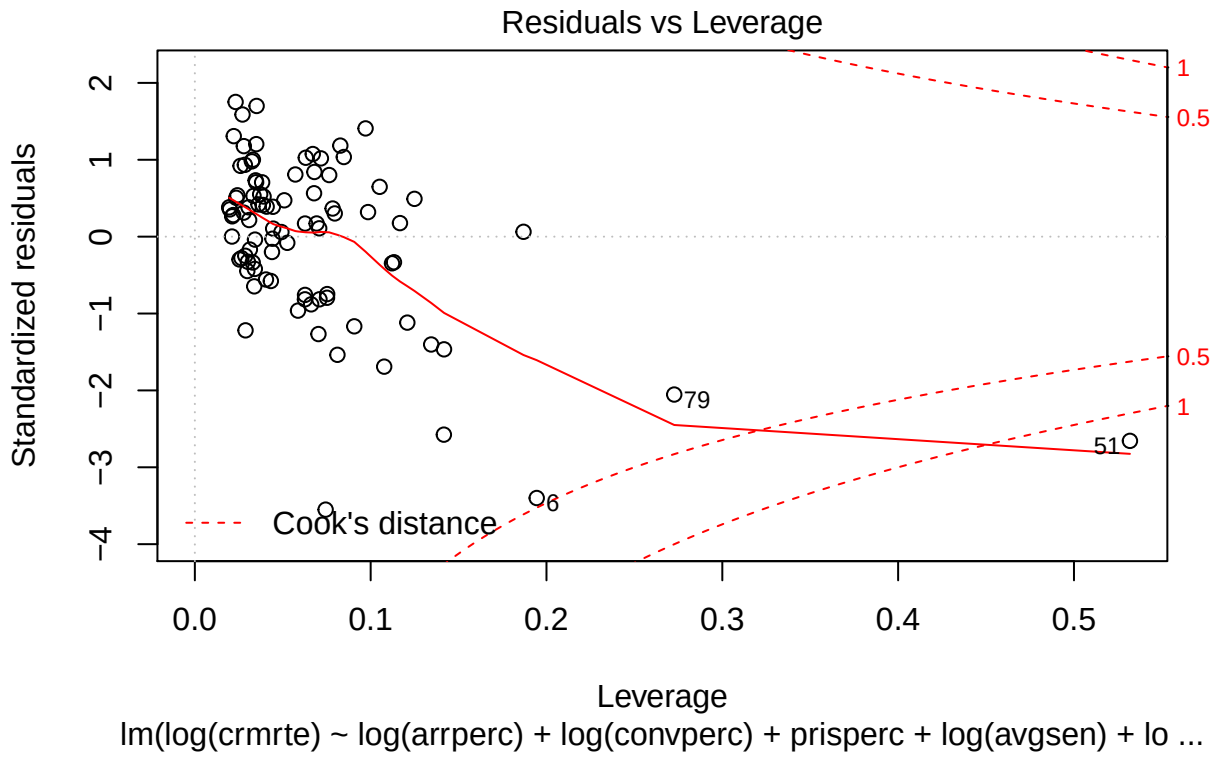
CLM 4 – Zero-Conditional Mean

To see whether there is a zero-conditional mean across all x's, we will plot the residuals against the fitted values.

```
plot(m1, which=1)
```




```
plot(m1, which=5)
```



Looking at the residuals vs. leverage plot, it appears there are a couple outliers that have considerable leverage on the regression, one of which also has a Cook's distance that's greater than 1.

```
data[data$X == 51 | data$X == 79,]
```

```
##     X county year    crmrte   prbarr  prbconv prbpris avgsen      polpc
## 51 51    115   87 0.0055332 1.090910 1.500000    0.50  20.70 0.00905433
## 79 79    173   87 0.0139937 0.530435 0.327869    0.15   6.64 0.00316379
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 51 0.3858093 28.19310    1       0     0  1.28365 204.2206 503.2351
## 79 0.2034221 37.72702    1       0     0 25.39140 231.6960 213.6752
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc       mix
## 51 217.4908 342.4658 245.2061 448.42 442.20 340.39 386.12 0.1000000
## 79 175.1604 267.0940 204.3792 193.01 334.44 414.68 304.32 0.4197531
##       pctymle  arrperc convperc prisperc logarrperc logconvperc  logpolpc
## 51 0.07253495 109.0910 150.0000       50   4.692182    5.010635 -4.704512
## 79 0.07462687  53.0435  32.7869       15   3.971112    3.490029 -5.755985
##    logavgsen
## 51  3.030134
## 79  1.893112
```

Upon closer inspection, the outlier, record 51, has percentages above 100% for both arrest and conviction. Since this is not actually possible, we should consider removing this outlier.

```
(cov(data$logarrperc,m1$residuals))
```

```
## [1] 0.00000000000000002436432
```

```
(cov(data$logconvperc,m1$residuals))
```

```
## [1] -0.00000000000000004325479
```

```
(cov(data$logavgsen,m1$residuals))
```

```
## [1] -0.000000000000000001992221
```

```
(cov(data$logpolpc,m1$residuals))
```

```
## [1] 0.000000000000000001992458
```

```
(cov(data$prisperc,m1$residuals))
```

```
## [1] 0.000000000000000004751364
```

The plots indicates little evidence that the zero-conditional mean assumption doesn't hold, as the red spline line remains close to zero despite its slight dip and rise at both ends due to less observations.

The covariances of the three independent variables with the residuals are very close to zero indicating they are likely exogenous.

One data point has a large Cook's distance and may have undue influence on the model fit.

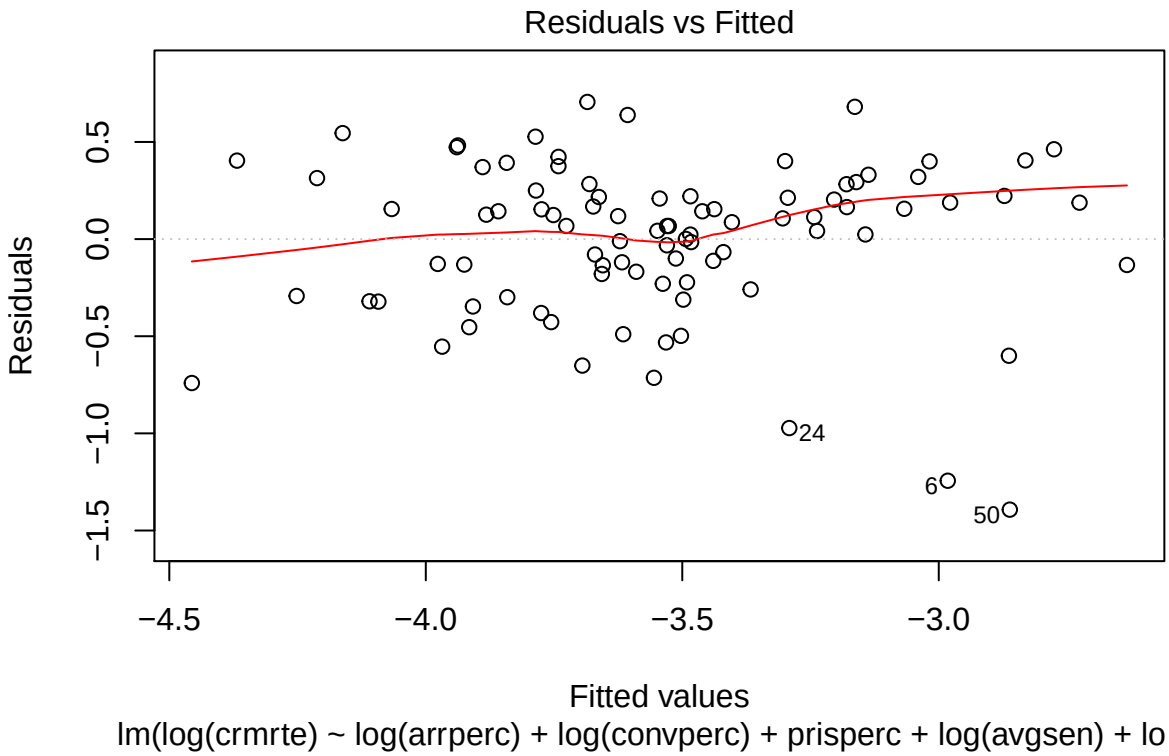
Is the assumption valid?

Response:

CLM 5 - Homoscedasticity

To determine whether the variance of μ is fixed for all x's, we will first take a look at the residuals plotted against the fitted values to see whether the variance of residuals is constant across the fitted values.

```
plot(m1, which=1)
```



The plot indicates no strong evidence of heteroskedasticity.

To further understand whether the model meets homoskedasticity, we will perform statistical tests Breusch-Pagan and the Score-test for non-constant error variance.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(m1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m1
## BP = 11.102, df = 5, p-value = 0.0494
```

With a p-value of 0.049, we cannot reject the null hypothesis of homoskedasticity at the 5% significance level.

```
ncvTest(m1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3.60581    Df = 1     p = 0.05757803
```

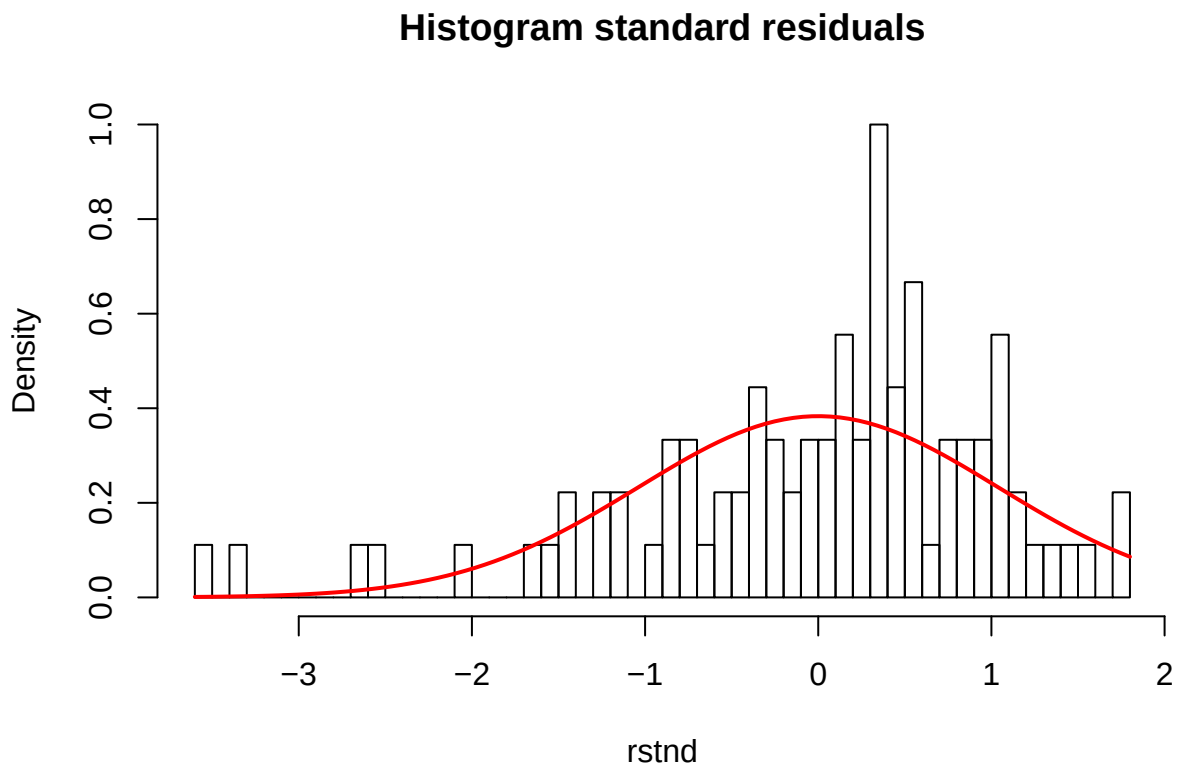
With a p-value of .057, we cannot reject the hypothesis of constant error variance.

Is the assumption valid? **Yes Response**: Though we can assume homoskedasticity, we will move forward with robust standard errors, as they provide more accurate p-values.

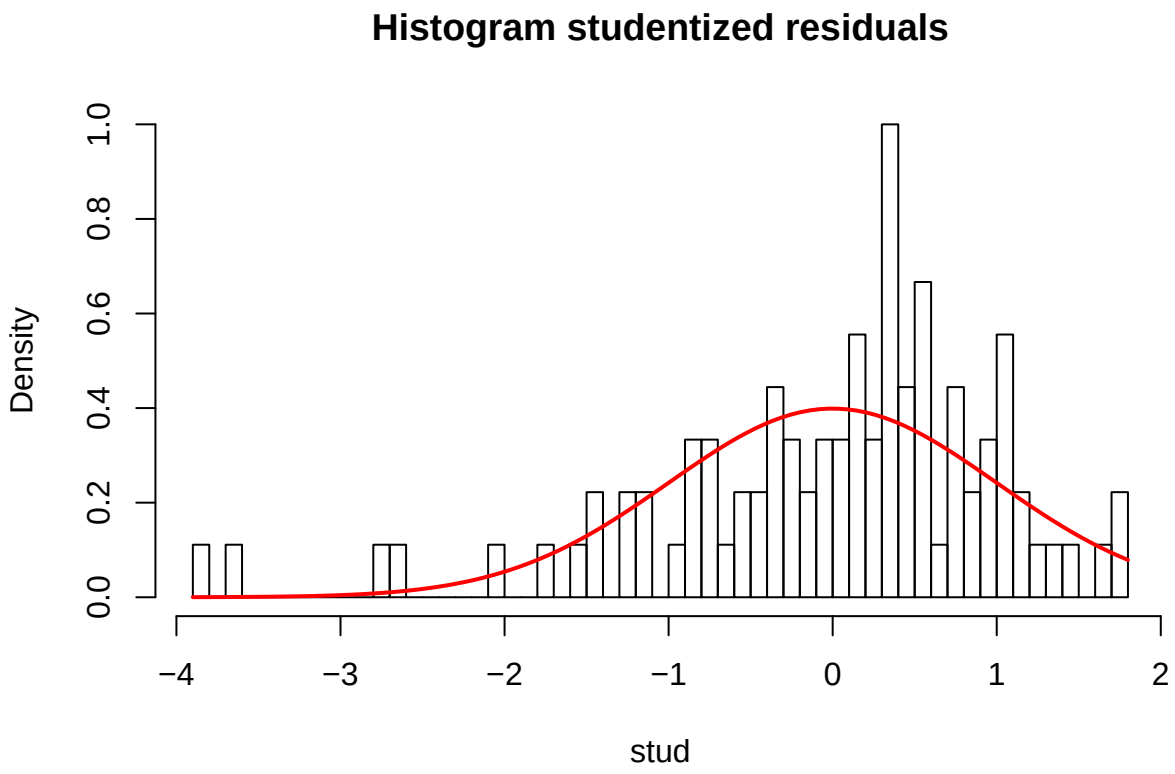
CLM 6 – Normality of residuals

To determine whether there is normality of the residuals, we will use a histogram or Q-Q plots of the residuals and visually observe whether there is normality.

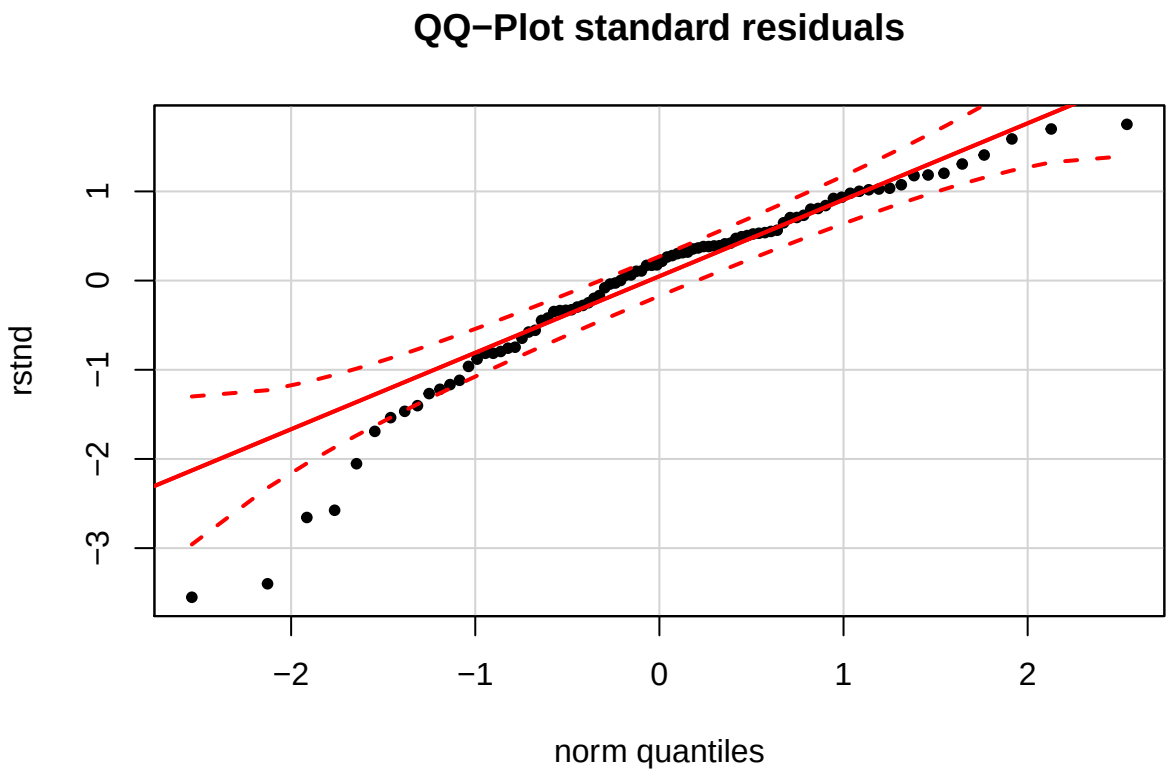
```
# normality of standard residuals
rstnd = rstandard(m1)
hist(rstnd, main="Histogram standard residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(rstnd)), col="red", lwd=2, add=TRUE)
```



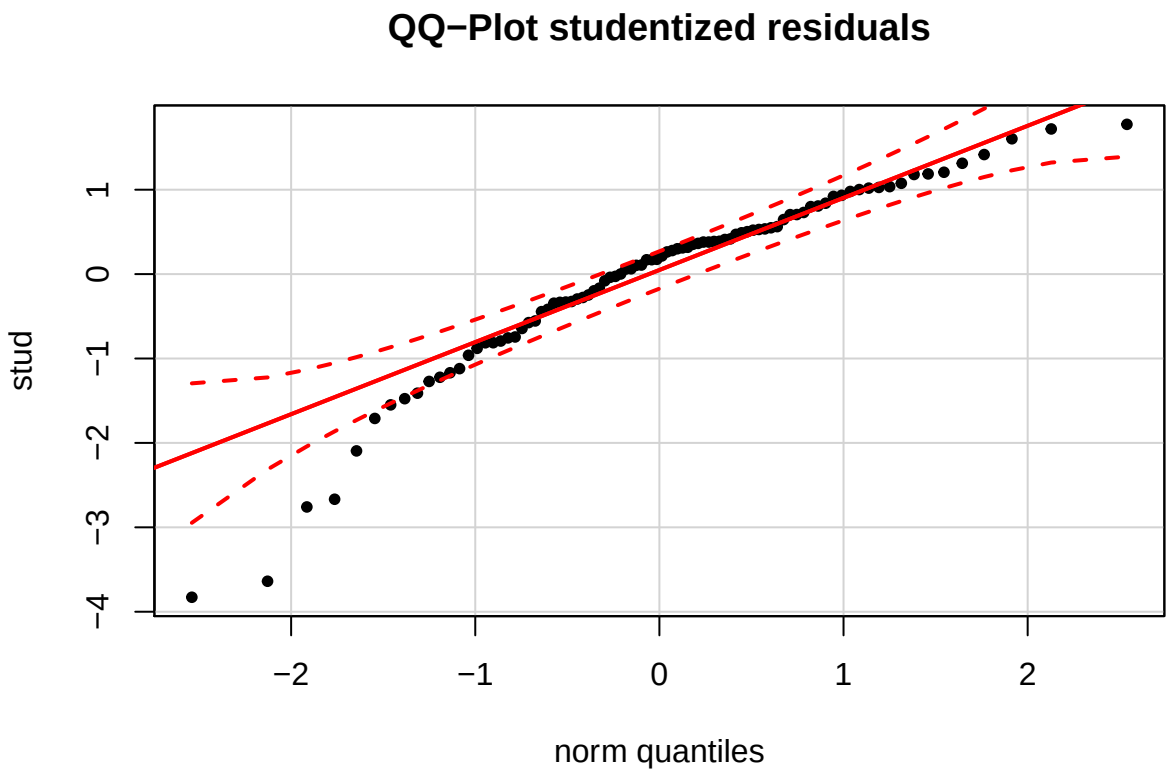

```
# normality of studentized residuals
stud = rstudent(m1)
hist(stud, main="Histogram studentized residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=1), col="red", lwd=2, add=TRUE)
```




```
# Q-Q plot standard residuals
qqPlot(rstnd, distribution="norm", pch=20, main="QQ-Plot standard residuals")
qqline(rstnd, col="red", lwd=2)
```




```
# Q-Q plot studentized residuals
qqPlot(stud, distribution="norm", pch=20, main="QQ-Plot studentized residuals")
qqline(stud, col="red", lwd=2)
```



The histograms of the residuals have a negative skew, and QQ-plots further demonstrate nonormality in the error distribution.

Is the assumption valid? **No** Response:

References:

"Shattering"Broken Windows": An Analysis of San Francisco's Alternative Crime Policies", CENTER ON JUVENILE AND CRIMINAL JUSTICE, October 1999 http://www.cjcj.org/uploads/cjcj/documents/shattering.pdf