Lab 4: Reducing Crime

W 203: Statistics for Data Science

Morris Michael Burkhardt, Venu Reddy, K Iwasaki

08/16/2017

Introduction

The purpose of this report is to explore crime data from 90 counties collected for the year 1988 and to summarize some key decisive factors that influence crime through statistical modeling and inference techniques. Additionally, the report is designed to help generate policy suggestions that could be applied to local government as a part of a political campaign research.

The following sections of the report detail our explorative data analysis (EDA), statistical modeling techniques, followed by some concluding policy suggestions.

Regression models

Initial exploratory analysis

```
library(car)
library(lmtest)
library(sandwich)
library(stargazer)
Crime_data = read.csv("crime_v2.csv")
```

The crime data frame has 90 data points for each variable.

We will first take a look at the summary.

```
summary(Crime_data)
```

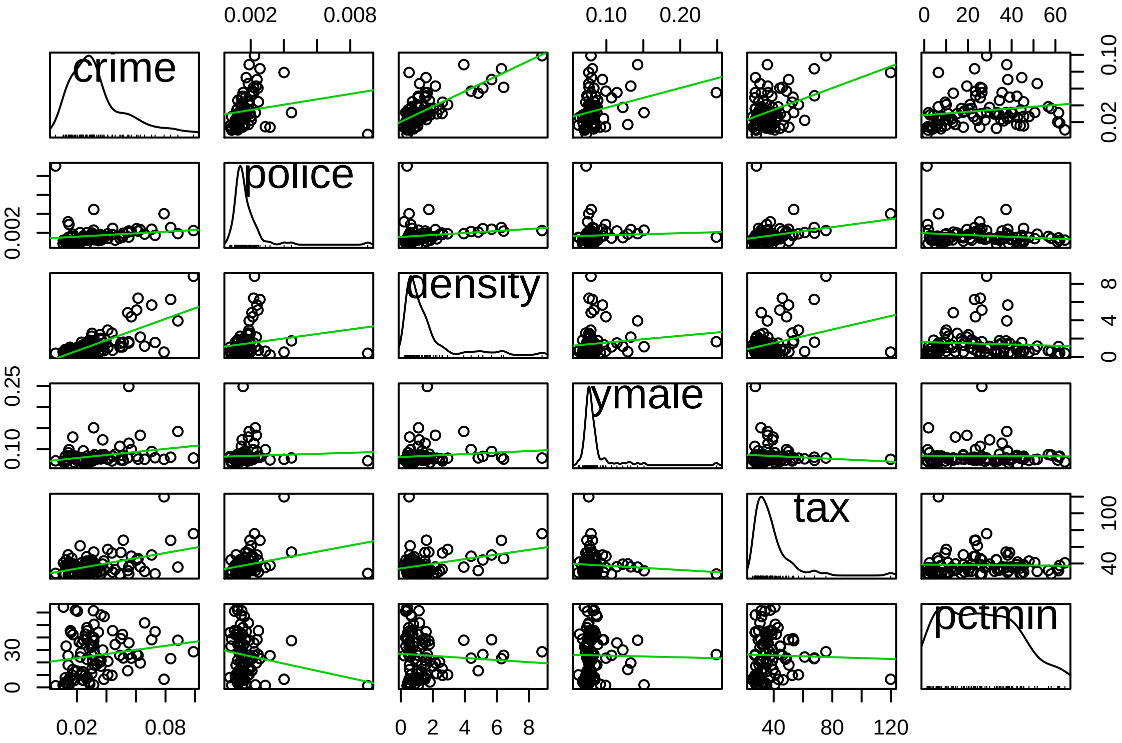
```
##        X             county           year        crime
##  Min.   : 1.00   Min.   :  1.0   Min.   :88   Min.   :0.005533
##  1st Qu.:23.25   1st Qu.: 51.5   1st Qu.:88   1st Qu.:0.020604
##  Median :45.50   Median :103.0   Median :88   Median :0.030002
##  Mean   :45.50   Mean   :100.6   Mean   :88   Mean   :0.033510
##  3rd Qu.:67.75   3rd Qu.:150.5   3rd Qu.:88   3rd Qu.:0.040249
##  Max.   :90.00   Max.   :197.0   Max.   :88   Max.   :0.098966
##     probarr          probconv          probsen            avgsen
##  Min.   :0.1500   Min.   :0.09277   Min.   :0.06838   Min.   : 5.380
##  1st Qu.:0.3642   1st Qu.:0.20495   1st Qu.:0.34422   1st Qu.: 7.375
##  Median :0.4222   Median :0.27146   Median :0.45170   Median : 9.110
##  Mean   :0.4106   Mean   :0.29524   Mean   :0.55086   Mean   : 9.689
##  3rd Qu.:0.4576   3rd Qu.:0.34487   3rd Qu.:0.58513   3rd Qu.:11.465
##  Max.   :0.6000   Max.   :1.09091   Max.   :2.12121   Max.   :20.700
##      police             density            tax              west
##  Min.   :0.0007459   Min.   :0.2034   Min.   : 25.69   Min.   :0.0000
##  1st Qu.:0.0012378   1st Qu.:0.5472   1st Qu.: 30.73   1st Qu.:0.0000
##  Median :0.0014897   Median :0.9792   Median : 34.92   Median :0.0000
##  Mean   :0.0017080   Mean   :1.4379   Mean   : 38.16   Mean   :0.3778  
##  3rd Qu.:0.0018856   3rd Qu.:1.5693   3rd Qu.: 41.01   3rd Qu.:1.0000  
##  Max.   :0.0090543   Max.   :8.8277   Max.   :119.76   Max.   :1.0000  
##     central           urban              pctmin          wagecon     
##  Min.   :0.0000   Min.   :0.00000   Min.   : 1.284   Min.   :193.6  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:10.024   1st Qu.:250.8  
##  Median :0.0000   Median :0.00000   Median :24.852   Median :281.2  
##  Mean   :0.2333   Mean   :0.08889   Mean   :25.713   Mean   :285.4  
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:38.183   3rd Qu.:315.0  
##  Max.   :1.0000   Max.   :1.00000   Max.   :64.348   Max.   :436.8  
##     wagetuc         wagetrd         wagefir         wageser      
##  Min.   :187.6   Min.   :154.2   Min.   :170.9   Min.   : 133.0  
##  1st Qu.:374.3   1st Qu.:190.7   1st Qu.:285.6   1st Qu.: 229.3  
##  Median :404.8   Median :203.0   Median :317.1   Median : 253.1  
##  Mean   :410.9   Mean   :210.9   Mean   :321.6   Mean   : 275.3  
##  3rd Qu.:440.7   3rd Qu.:224.3   3rd Qu.:342.6   3rd Qu.: 277.6  
##  Max.   :613.2   Max.   :354.7   Max.   :509.5   Max.   :2177.1  
##     wagemfg         wagefed         wagesta         wageloc     
##  Min.   :157.4   Min.   :326.1   Min.   :258.3   Min.   :239.2  
##  1st Qu.:288.6   1st Qu.:398.8   1st Qu.:329.3   1st Qu.:297.2  
##  Median :321.1   Median :448.9   Median :358.4   Median :307.6  
##  Mean   :336.0   Mean   :442.6   Mean   :357.7   Mean   :312.3  
##  3rd Qu.:359.9   3rd Qu.:478.3   3rd Qu.:383.2   3rd Qu.:328.8  
##  Max.   :646.9   Max.   :598.0   Max.   :499.6   Max.   :388.1  
##       mix               ymale        
##  Min.   :0.01961   Min.   :0.06216  
##  1st Qu.:0.08060   1st Qu.:0.07437  
##  Median :0.10095   Median :0.07770  
##  Mean   :0.12905   Mean   :0.08403  
##  3rd Qu.:0.15206   3rd Qu.:0.08352  
##  Max.   :0.46512   Max.   :0.24871  
```

It looks like there are no missing values (NA or NaN) and no values used to code missing values (such as for instance -1). The west, central and urban variables are all dummy variables.

The lowest and highest value for 'probarr', the 'probability' of arrest are very smooth values and therefore suspicious. This variable might be top- and bottom-coded. Since we have no further information on the collection method, we will just take leave the data here as it is.

We will furthermore take a brief look at a scatterplot matrix of our dependent variable with some of the variables we may consider to be key variables.

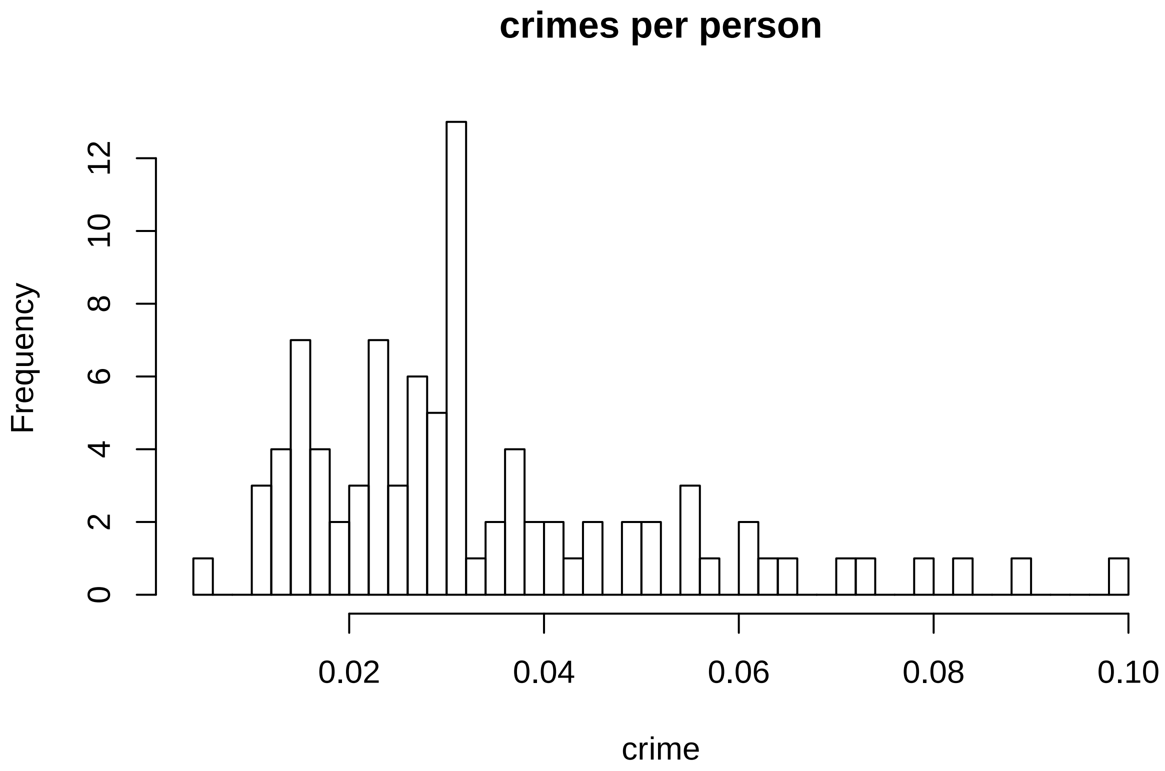
```
scatterplotMatrix(Crime_data[, c("crime", "police", "density", "ymale", "tax", "pctmin")],
                  smoother = FALSE)
```



Especially the police, the density and the tax variable seem to have a substantial correlation with the crime variable.

We will now take a closer look at our dependent variable, crime.

```
hist(Crime_data$crime, breaks = 50, main = "crimes per person", xlab = "crime")
```



The crime variable has a minimum value of 0.0055332 and a maximum value of 0.0989659. The variable has a positive skew with quite a few outliers on the right and one outlier towards the left.

We should probably log-transform the crime variable, but before we do that, let us take a closer look at the outlier on the low end.

```
head(Crime_data[order(Crime_data$crime), ],1)
```

```
##     X county year     crime probarr probconv probsen avgsen     police
## 51 51    115   88 0.0055332     0.5  1.09091     1.5   20.7 0.00905433
##      density     tax west central urban  pctmin  wagecon  wagetuc  wagetrd
## 51 0.3858093 28.1931    0       1     0 1.28365 204.2206 503.2351 217.4908
##     wagefir  wageser wagemfg wagefed wagesta wageloc mix      ymale
## 51 342.4658 245.2061  448.42   442.2  340.39  386.12 0.1 0.07253495
```

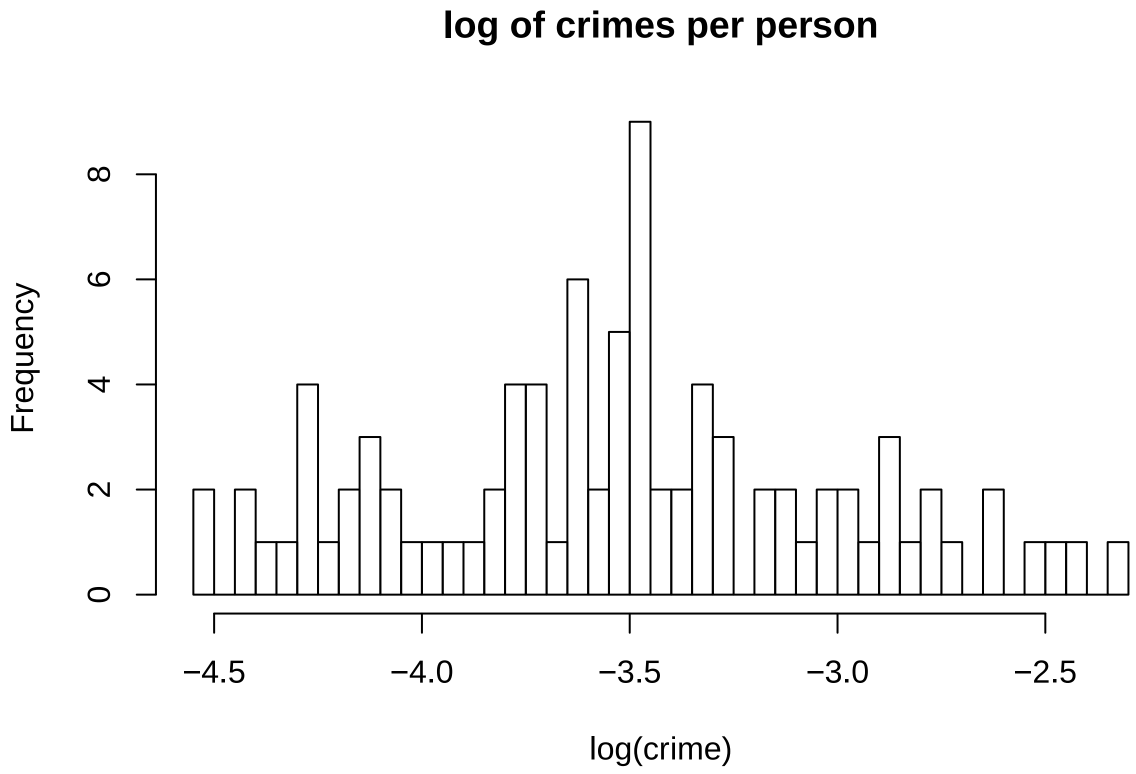
This outlying data point is county number 115. Most of the data for county 115 seem extreme or otherwise suspicious: There are a few very smooth values, such as probarr = 0.5, probsen = 1.5, probconv = 12/11 and mix = 0.1000. Furthermore, county number 115 has some extreme values: The avgsen variable takes on the largest value accross all counties (20.70) and crime takes on the lowest value accross all counties (0.0055332). All these smooth and extreme values indicate that it is likely that there are only very few crimes recorded in county 115. This data point therefore adds a lot of extra variation to our data.

We will therefore remove this data point.

```
Crime_data = Crime_data[Crime_data$county != 115 , ]
```

Next we will log transform the crime variable.

```
hist(log(Crime_data$crime), breaks = 50, main = "log of crimes per person",
     xlab = "log(crime)")
```



The highest value of the tax variable is about three times as high as the mean tax value and around 60 % higher than the second highest tax value. This is either an error in the data, or this county might be very different from other counties in many aspects. Maybe it is a very small county with just one big company, or one rich person lives there who boosts up the mean income tax.

```
head(Crime_data[order(-Crime_data$tax), ],2)
```

```
##     X county year      crime  probarr probconv  probsen avgsen     police
## 25 25     55   88 0.0790163 0.304348 0.224628 0.207831  13.57 0.00400962
## 53 53    119   88 0.0989659 0.486183 0.149094 0.347800   7.13 0.00223135
##      density       tax west central urban   pctmin  wagecon  wagetuc
## 25 0.5115089 119.76145    0       0     0  6.49622 309.5238 445.2762
## 53 8.8276520  75.67243    1       0     1 28.54600 436.7666 548.3239
##     wagetrd  wagefir  wageser wagemfg wagefed wagesta wageloc        mix
## 25 189.7436 284.5933 221.3903  319.21  338.91  361.68  326.08 0.08437271
## 53 354.6761 509.4655 354.3007  494.30  568.40  329.22  379.77 0.16869897
##         ymale
## 25 0.07613807
## 53 0.07916495
```

We will therefore remove this data point.

```
Crime_data = Crime_data[Crime_data$tax < 110, ]
```

The highest service wage value is almost eight times as high as the mean service wage and over 5.5 times higher than the second highest service wage value. This looks like an error in the data. It is unlikely that the average service wage in one county is so much higher than in all the other counties.

```
head(Crime_data[order(-Crime_data$wageser), ],2)
```

```
##     X county year     crime  probarr probconv  probsen avgsen     police
## 84 84    185   88 0.0108703 0.442857 0.195266 2.121210   5.38 0.00122210
## 29 29     63   88 0.0706599 0.363636 0.133225 0.459216  11.51 0.00237609
##      density      tax west central urban  pctmin  wagecon  wagetuc
## 84 0.3887588 40.82454    1       0     0 64.3482 226.8245 331.5650
## 29 5.6744967 50.19918    1       0     1 38.2230 349.3267 548.9865
##     wagetrd  wagefir   wageser wagemfg wagefed wagesta wageloc        mix
## 84 167.3726 264.4231 2177.0681  247.72  381.33  367.25  300.13 0.04968944
## 29 238.9154 435.1107  391.3081  646.85  563.77  415.51  362.58 0.07585382
##         ymale
## 84 0.07008217
## 29 0.09468981
```

We will therefore remove this datapoint.

```
Crime_data = Crime_data[Crime_data$wageser < 2100, ]
```

We are aware that it is not mathematically correct to just calculate the mean of all average wages, as we do not know the weights that each business sector has in every county. The average value of all wages of different business sectors within a county should however still be a good indicator for the average wage in that county. We will therefore calculate an average wage variable, avgwage.

```
Crime_data$avgwage = (Crime_data$wagecon + Crime_data$wagefed + Crime_data$wagefir +
                        Crime_data$wageloc + Crime_data$wagemfg + Crime_data$wageser +
                        Crime_data$wagesta + Crime_data$wagetrd + Crime_data$wagetuc) / 9
```

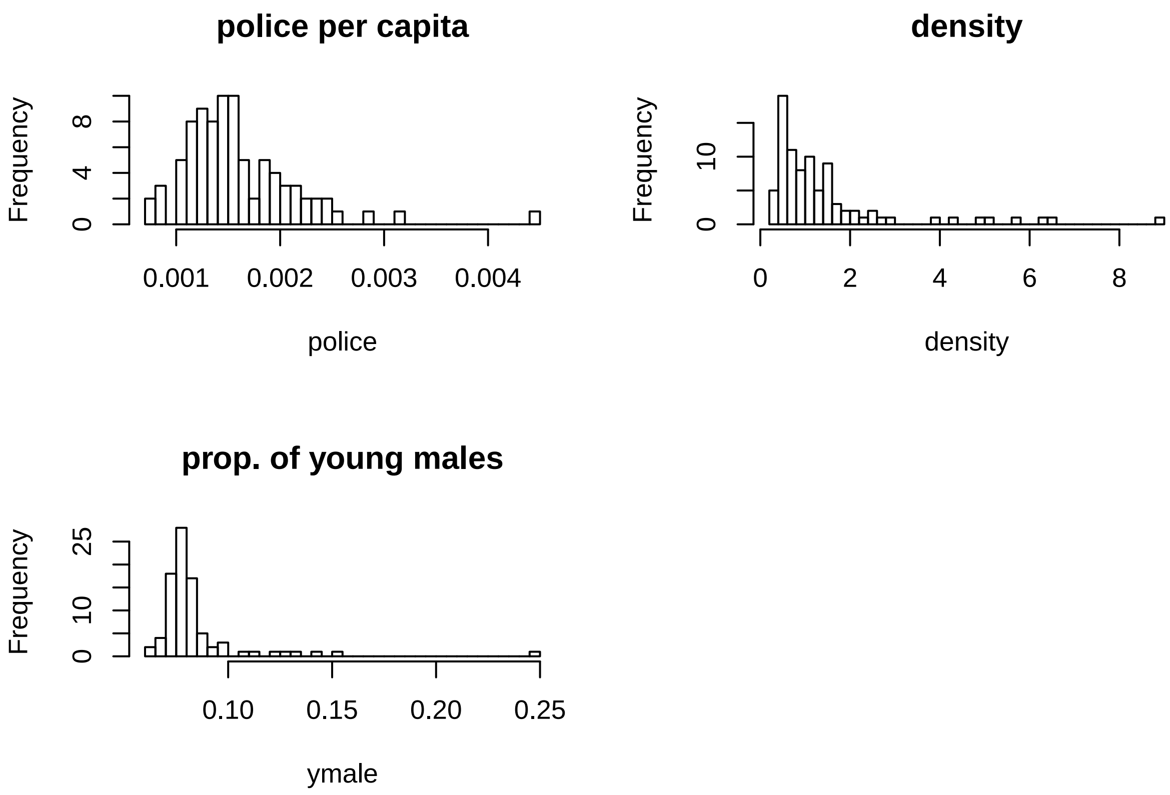
We will take a closer look at the other variables, as we include them into our model.

Model with key explanatory variables

We believe that police per capita (‘police’ variable), density (‘density’ variable) and the proportion of males between the ages of 15 and 24 (‘ymale’ variable) all have a key effect on crime.

We will first take a look at those independent variables:

```
par(mfrow=c(2,2))
hist(Crime_data$police, breaks = 50, main = "police per capita", xlab = "police")
hist(Crime_data$density, breaks = 50, main = "density", xlab = "density")
hist(Crime_data$ymale, breaks = 50, main = "prop. of young males", xlab = "ymale")
```



All three variables have a positive skew and large outliers to the right.

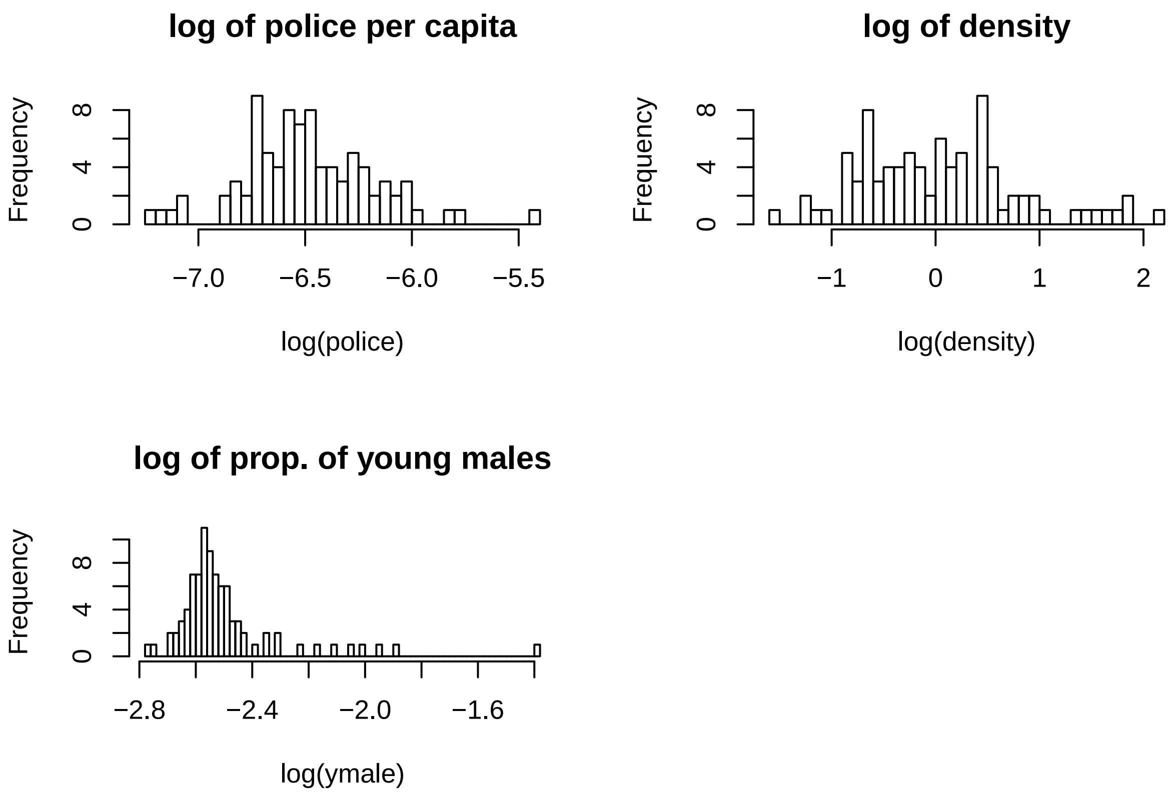
Let us take a closer look at the large values in the ‘ymale’ variable. The greatest value is 0.2487116. Next we will look at the five largest values.

```
head(Crime_data[order(-Crime_data$ymale), "ymale"])
```

```
## [1] 0.2487116 0.1509264 0.1422378 0.1330291 0.1289471 0.1222448
```

The largest value is significantly higher than the second largest value. It is very unlikely that almost 25 % of a county’ popluation are males between the ages of 15 and 24 years. There may however be an explanation for this high percentage, such as an all men’s college in that county. We will therefore keep this data point in our analysis.

```
par(mfrow=c(2,2))
hist(log(Crime_data$police), breaks = 50,
     main = "log of police per capita", xlab = "log(police)")
hist(log(Crime_data$density), breaks = 50,
     main = "log of density", xlab = "log(density)")
hist(log(Crime_data$ymale), breaks = 50,
     main = "log of prop. of young males", xlab = "log(ymale)")
```



Before we include these variables into our regression model, let us look at the correlation amongst our independent variables.

```r
d = data.frame(log(Crime_data$police), log(Crime_data$density), log(Crime_data$ymale))
colnames(d) = c("log(police)", "log(density)", "log(ymale)")
cor(d)
```

```
##              log(police) log(density) log(ymale)
## log(police)    1.0000000    0.4651711  0.2643834
## log(density)   0.4651711    1.0000000  0.2271323
## log(ymale)     0.2643834    0.2271323  1.0000000
```

None of the correlations is extremely high, so we can exclude multicollinearity.

We will regress log of police, log of density and log of ymale on log of crime.

$$\log(crime) = \beta_0 + \beta_1 \cdot \log(police) + \beta_2 \cdot \log(density) + \beta_3 \cdot \log(ymale)$$

```r
model1 = lm(log(crime) ~ log(police) + log(density) + log(ymale), data = Crime_data)
coeftest(model1, vcov = vcovHC)
```

```
## 
## t test of coefficients:
## 
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -0.669450   1.217628 -0.5498  0.58393    
## log(police)   0.323074   0.180188  1.7930  0.07662 .  
## log(density)  0.393143   0.057079  6.8877 1.01e-09 ***
## log(ymale)    0.310913   0.162265  1.9161  0.05880 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

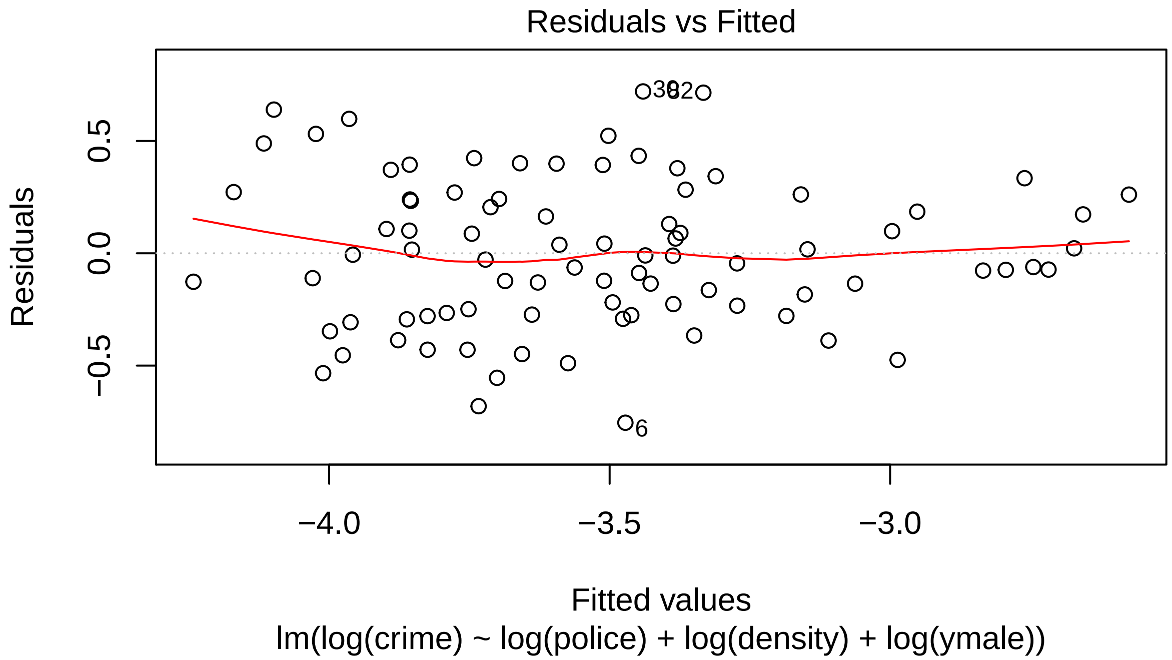
We will calculate the Akaike Information Criterion for this model, so that we can later compare it with our other models.

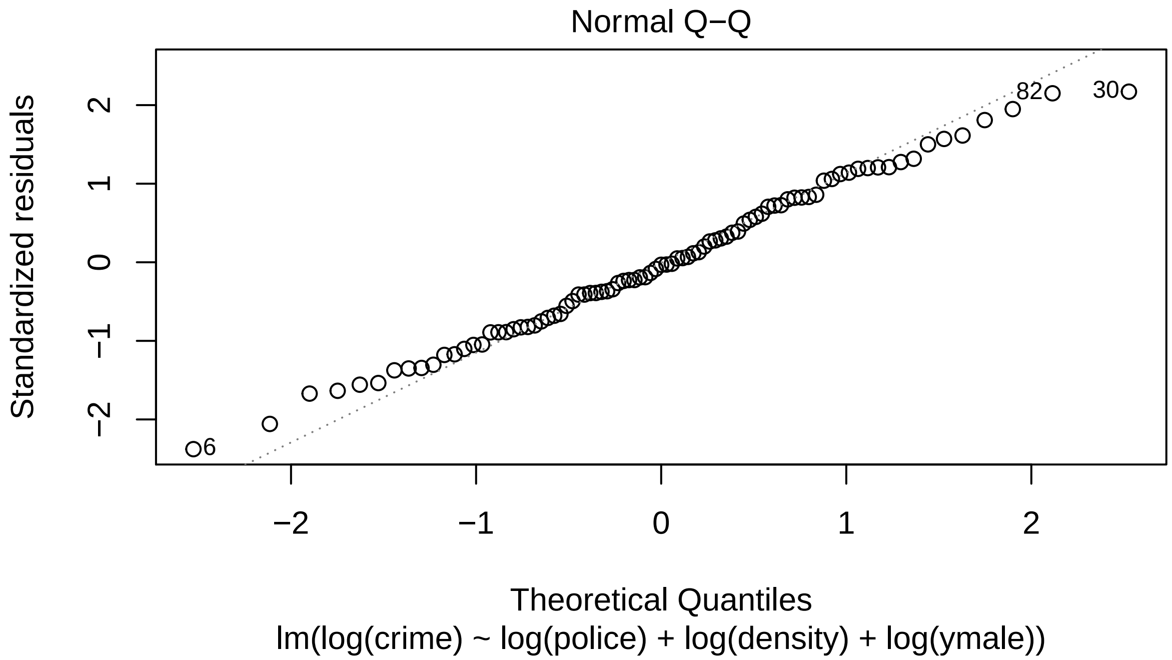
```
AIC(model1)
```

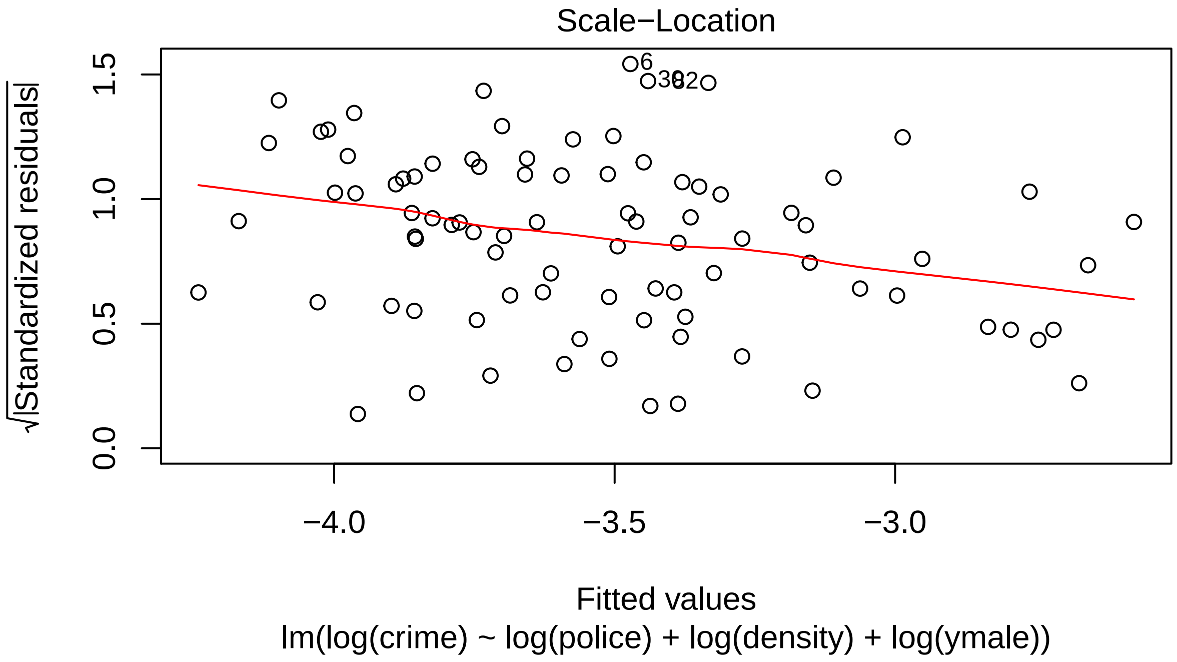
```
## [1] 62.63199
```

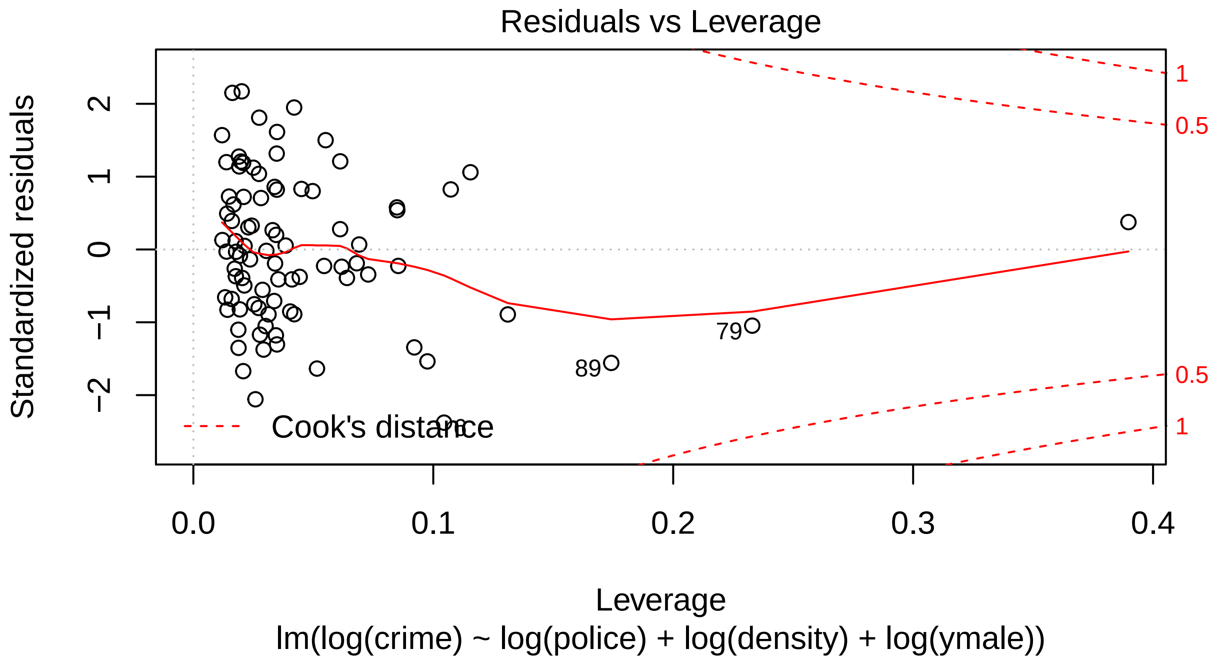
To check if the model fulfills the classical linear model assumptions, we will first draw the diagnostic plots.

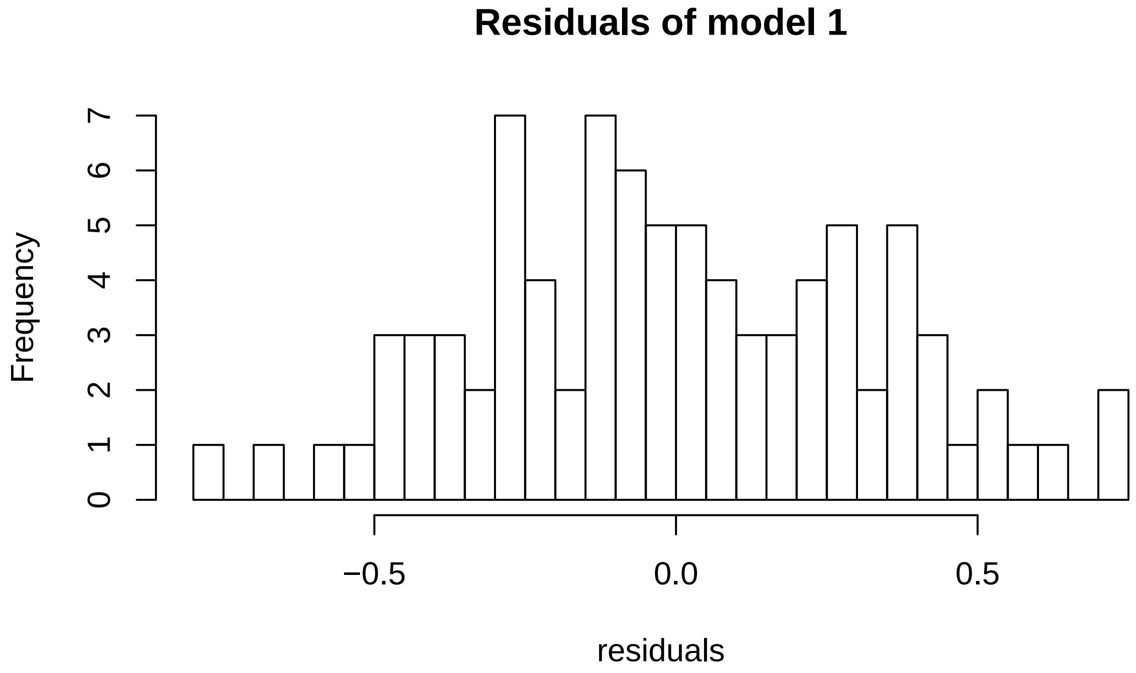
```
plot(model1)
hist(model1$residuals, breaks = 30,
      main = "Residuals of model 1", xlab = "residuals")
```




Normal Q−Q
Standardized residuals
−2
−1
0
1
2
Theoretical Quantiles
lm(log(crime) ~ log(police) + log(density) + log(ymale))
6
82
30


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−4.0
−3.5
−3.0
Fitted values
lm(log(crime) ~ log(police) + log(density) + log(ymale))
6
30
82





We will now check all of the classical linear model assumptions:

- Linearity assumption (MLR1): Since we have not restricted the error, our population model is linear in parameters. The linearity assumption is therefore met.
- Random Sampling (MLR2): It is uncertain if the selection of counties is a true random sample, as we

have no information on how our sample was collected. We are therefore unable to determine, whether the random sampling assumption is met.

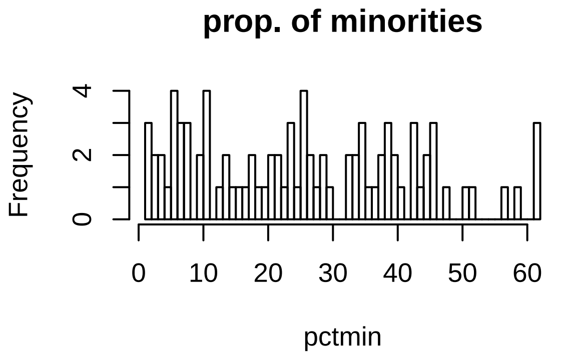
- No Perfect collinearity (MLR3): We can see from R's data summaries, that none of the independent variables is constant - aside of the 'year' variable, which we will of course not be using in any of our models. We also looked at a correlation matrix (see above) and were not able to identify any exact linear relationships among the independent variables. R would furthermore throw an error, if we were trying to create a model with perfect collinearity amongst the independent variables. The 'no perfect collinearity' assumption is therefore met.
- Zero Conditional Mean (MLR4): To check this assumption, we take a look at the above 'Residuals vs Fitted' plot. Zero Conditional mean, means that the expected value of the errors is zero and that the expted value is independent of the independent variables (x's) (or any linear combination of the x's, such as the fitted values). We are therefore looking for a horizontal spline curve (red curve) at zero in the 'Residuals vs Fitted' plot. Our plot shows a curve that is almost horizontal at zero. To the left and to the right end, the spline shows a slight upward trend, which might just be noise due to a low amount of data points in these regions. We therefore conclude that this assumption is sufficiently met.
- Homoskedasticity (MLR5): To check this assumption we can look at two of the diagnostic plots. Homoskedasticity means that the variance of the errors is indpendent of the independent variables (x's) (or any linear combination of the x's). At first, we will take a look at the 'Residuals vs Fitted' plot. To check the homoskedasticity assumption in this plot, we look for an even band of plotted data points. This assumption seems violated. It appears as if the band gets thinner towards the right hand side. Secondly, we will take a look at the Scale-Location plot. Here, the fitted values are plotted against the square root of the standardized residuals. We are therefore looking for a horizontal spline curve (red curve). This plot again suggests that the assumption of homoskedasticity is violated, as the spline curve declines. Since we have heteroskedasticity, we will be using heteroskedasticity-robust errors.
- Normality (MLR6): The normality assumption demands that the errors are normally distributed and also independent of the indpendent variables (x's). We check this assumption by looking at the qq-plot of the standardized residuals. The qq-plot indicates that the distribution of our residuals has heavier tails than a normal distribution. This assumption can also be checked by creating a histogram of the residuals and looking to see if the distribution has the shape of a normal distribution. The histogram also indicates a violation of the normality assumption. Since we have a sufficiently large sample, we should be fine to rely on asymptotics to get a normal sampling distribution of our coefficients. When heteroskedasticity robust standard errors are used, we only need to meet MLR1 through MLR4 for asymptotics to work. Under the premise, that the random sampling assumption is met (we do not have this information) we are therefore safe to rely on asymptotics for our coefficients to be normally distributed.

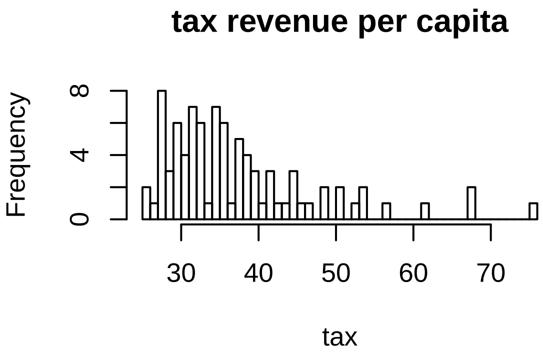
The 'Residuals vs Leverage' plot does not show any data points with very high influence.

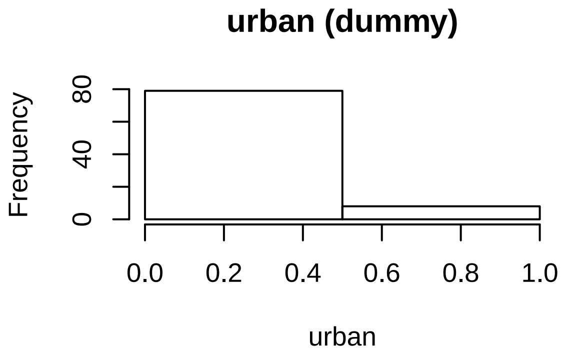
Model with key explanatory variables and covariates that increase accuracy without introducing bias

We believe that wealth (measured by tax revenue per capita), whether an area is urban or not, and the proportion of minorities and non-whites will increase the accuracy of the model. We will first take a look at these variables.

```r
par(mfrow=c(2,2))
hist(Crime_data$pctmin, breaks = 50, main = "prop. of minorities", xlab = "pctmin")
hist(Crime_data$tax, breaks = 50, main = "tax revenue per capita", xlab = "tax")
hist(Crime_data$urban,breaks = 2, main = "urban (dummy)", xlab = "urban")
```

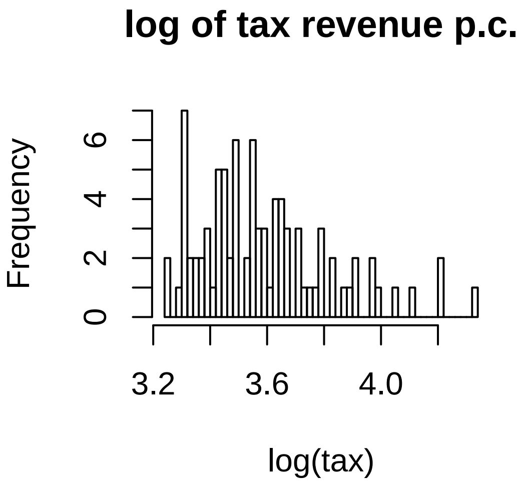






The tax variable is positively skewed. Since its minimum is a rather high value (25.6928654), it probably does not makes sense to perform a simple log transformation. To confirm this, we will take a quick look at the log transformation of the tax variable.

```
hist(log(Crime_data$tax), breaks = 50,
     main = "log of tax revenue p.c.", xlab = "log(tax)")
```



Since the log transformation does not improve the distribution of our tax variable substantially, we will not use a transformation on it.

Next we will look at a correlation matrix of all our independent variables.

```
d = data.frame(log(Crime_data$police), log(Crime_data$density), log(Crime_data$ymale), Crime_data$pctmi
colnames(d) = c("log(police)", "log(density)", "log(ymale)", "pctmin", "tax", "urban")
cor(d)
```

```
##              log(police) log(density)  log(ymale)      pctmin         tax
## log(police)   1.00000000    0.4651711  0.26438341 -0.02655701  0.42984151
## log(density)  0.46517110    1.0000000  0.22713230 -0.18536290  0.33027321
## log(ymale)    0.26438341    0.2271323  1.00000000 -0.00723544 -0.07086104
## pctmin       -0.02655701   -0.1853629 -0.00723544  1.00000000  0.04787452
## tax           0.42984151    0.3302732 -0.07086104  0.04787452  1.00000000
## urban         0.37327480    0.6526433  0.12384738  0.01593864  0.49295817
##                    urban
## log(police)  0.37327480
## log(density) 0.65264330
## log(ymale)   0.12384738
## pctmin       0.01593864
## tax          0.49295817
## urban        1.00000000
```

Urban and log(density) are highly correlated, which intuitively makes sense. Including the urban variable in our model will prevent bias. The standard error however will likely increase. We will make sure to take a look at the variance inflation factor after modelling.

Furthermore, there is quite some correlation between the tax variable and each of the variables urban and log(police). The variables pctmin and log(ymale) are not highly correlated with any of the variables, while log(density) has some correlation with log(police).

We do not believe that people who commit crimes consider the probability of arrest, conviction or prison sentence or consider the average time it takes for a sentence. We will therefore regress log(police), log(density), log(ymale), pctmin, log(tax) and urban on log(crime).

$$\log(crime) = \beta_0 + \beta_1 \cdot \log(police) + \beta_2 \cdot \log(density) + \beta_3 \cdot \log(ymale) + \beta_4 \cdot pctmin + \beta_5 \cdot tax + \beta_6 \cdot urban$$

```
model2 = lm(log(crime) ~ urban + log(police) + log(density) + log(ymale) +
              pctmin + tax, data = Crime_data)
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -1.3913e+00  1.0128e+00 -1.3737   0.17337
## urban        -3.0075e-02  1.4031e-01 -0.2144   0.83082
## log(police)   2.7574e-01  1.6464e-01  1.6748   0.09787 .
## log(density)  4.6285e-01  5.0350e-02  9.1927 3.675e-14 ***
## log(ymale)    2.8238e-01  1.2265e-01  2.3024   0.02391 *
## pctmin        1.3268e-02  2.0369e-03  6.5139 5.992e-09 ***
## tax           3.0755e-05  3.5372e-03  0.0087   0.99308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will once again calculate the Akaike Information Criterion (AIC).

```
AIC(model2)
```

```
## [1] 21.73158
```

The AIC of our second model (21.7315838) indicates a substantially better 'ratio' between fit and parsimony, as it is only about a third of the AIC of our first model (62.63199). In other words, our second model (with three added variables) offers a better compromise between fit and parsimony than our first model. The relative quality of our second model is better than the relative quality of our first model.

To test whether the newly added coefficients are jointly statistically significant, we use the wald test, which generalizes the usual F-test of overall significance, but allows for a heteroskedasticity-robust covariance matrix.

```
wald1 = waldtest(model1, model2, vcov = vcovHC)
wald1
```

```
## Wald test
##
## Model 1: log(crime) ~ log(police) + log(density) + log(ymale)
## Model 2: log(crime) ~ urban + log(police) + log(density) + log(ymale) +
##     pctmin + tax
##   Res.Df Df      F    Pr(>F)
## 1     83
## 2     80  3 15.972 3.159e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The three added variables have joint statistical significance. We will examine the practical significance of all the coefficients later on.

Since we added highly correlated values to the model, we also want to take a look at the variance inflation factors (VIFs).

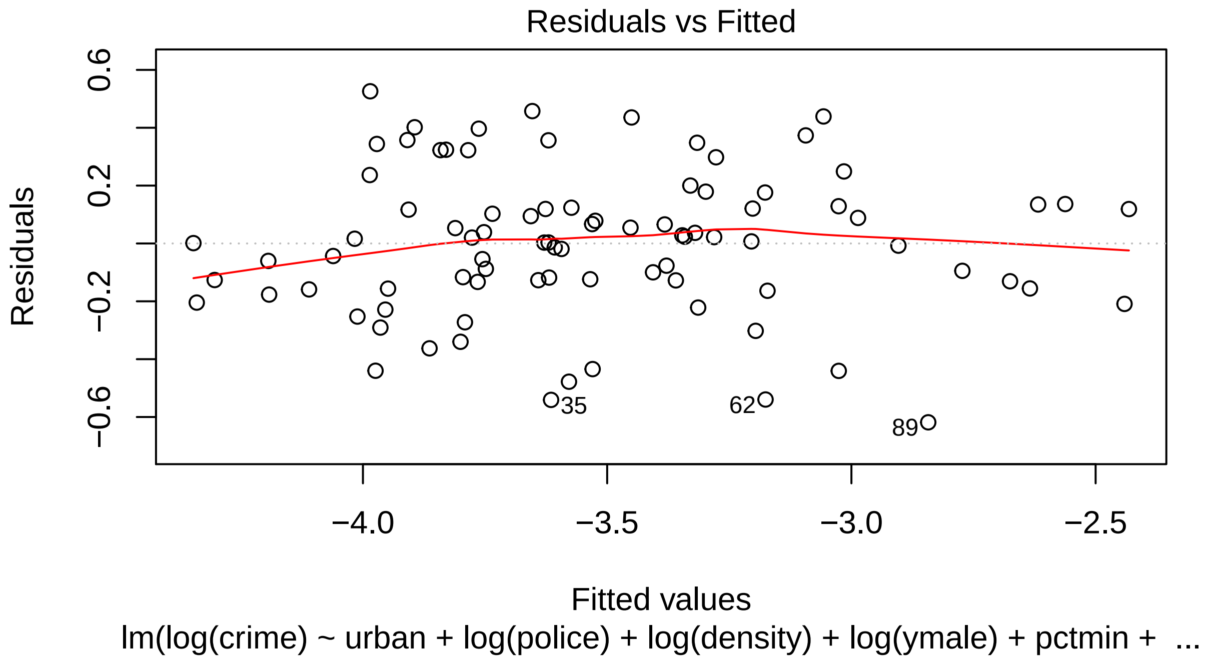
```
vif(model2)
```

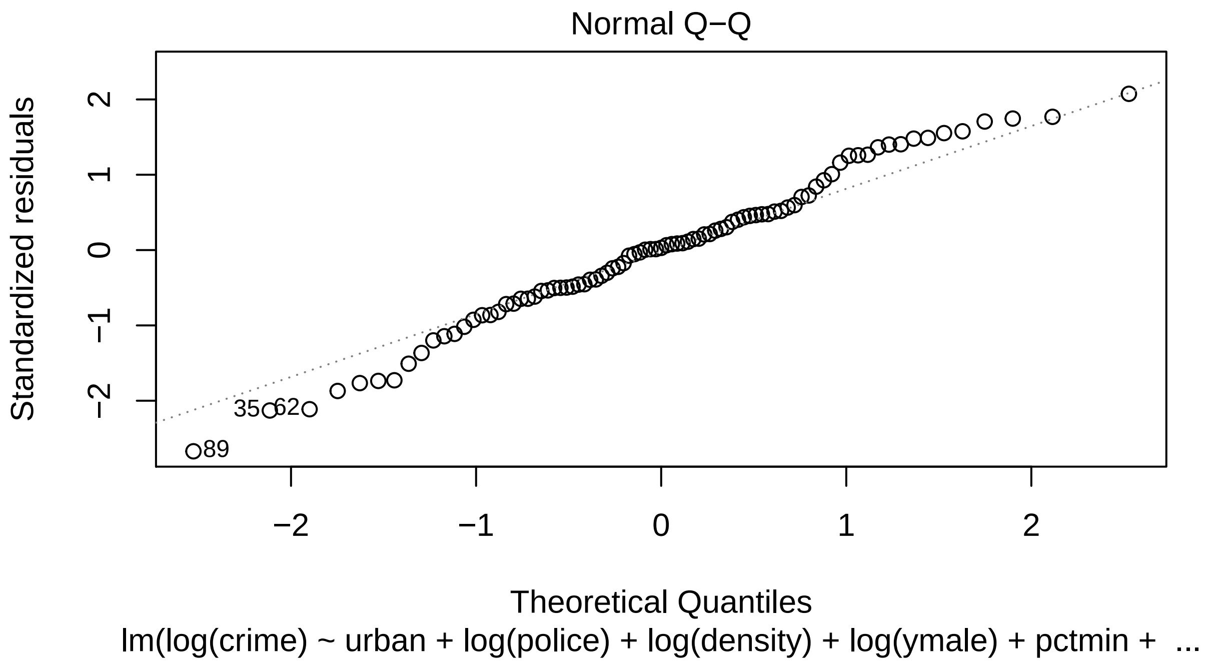
```
##        urban  log(police) log(density)   log(ymale)       pctmin
##     2.102954     1.534408     2.118472     1.160271     1.078172
##          tax
##     1.556989
```

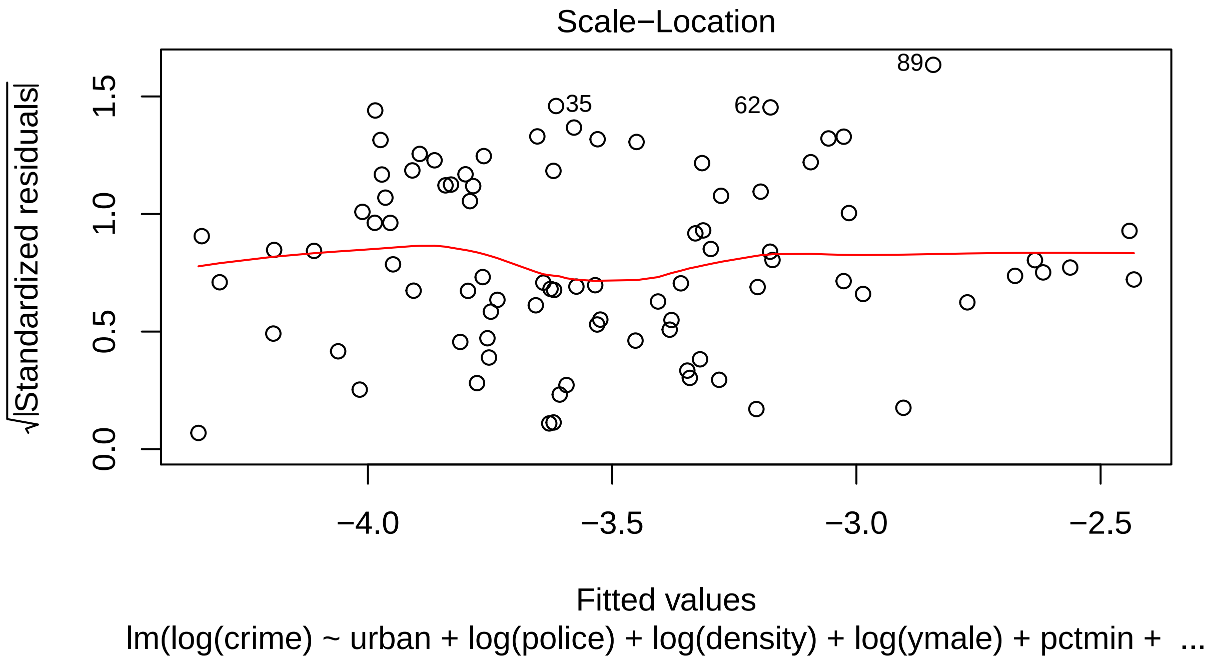
None of the VIFs values seems alarming.

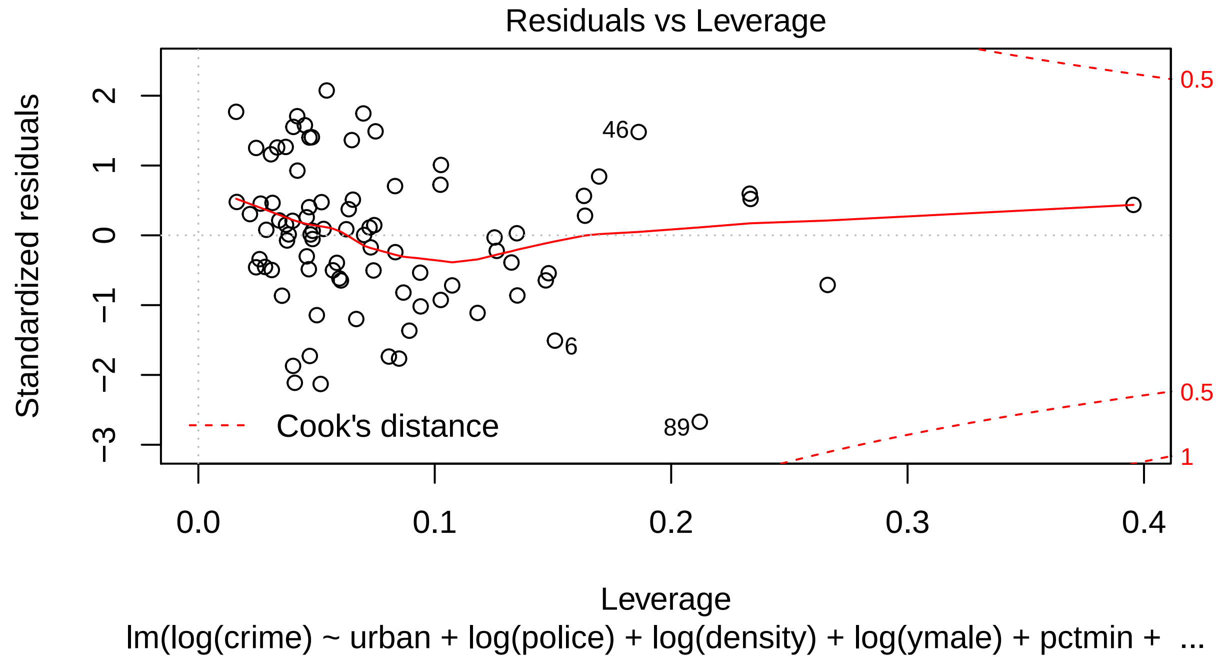
We will now take a look at the diagnostic plots for our second model.

```
plot(model2)
```


Residuals vs Fitted
Residuals
0.6
0.2
−0.2
−0.6
−4.0
−3.5
−3.0
−2.5
35
62
89
Fitted values
lm(log(crime) ~ urban + log(police) + log(density) + log(ymale) + pctmin + ...


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−2
−1
0
1
2
35
62
89
Theoretical Quantiles
lm(log(crime) ~ urban + log(police) + log(density) + log(ymale) + pctmin + ...





In comparison to our first model, we have very similar results regarding our classical linear model assumptions. Everything that was said about MLR1 through MLR4 in our first model remains true for our second model.

The homoskedasticity assumption (MLR5) looks less violated in our second model, as the band of data points in the 'Residuals vs Fitted' plot looks more even and the spline curve in the Scale-Location plot looks more horizontal. It is however still violated and we will continue to use heteroskedasticity-robust errors.

The normality assumption (MLR6) looks more violated in our second model than it did in our first model. It however still stands true, that we can rely on asymptotics under the same circumstances as in our first model.

The 'Residuals vs Leverage' plot does not show any data points with very high influence.

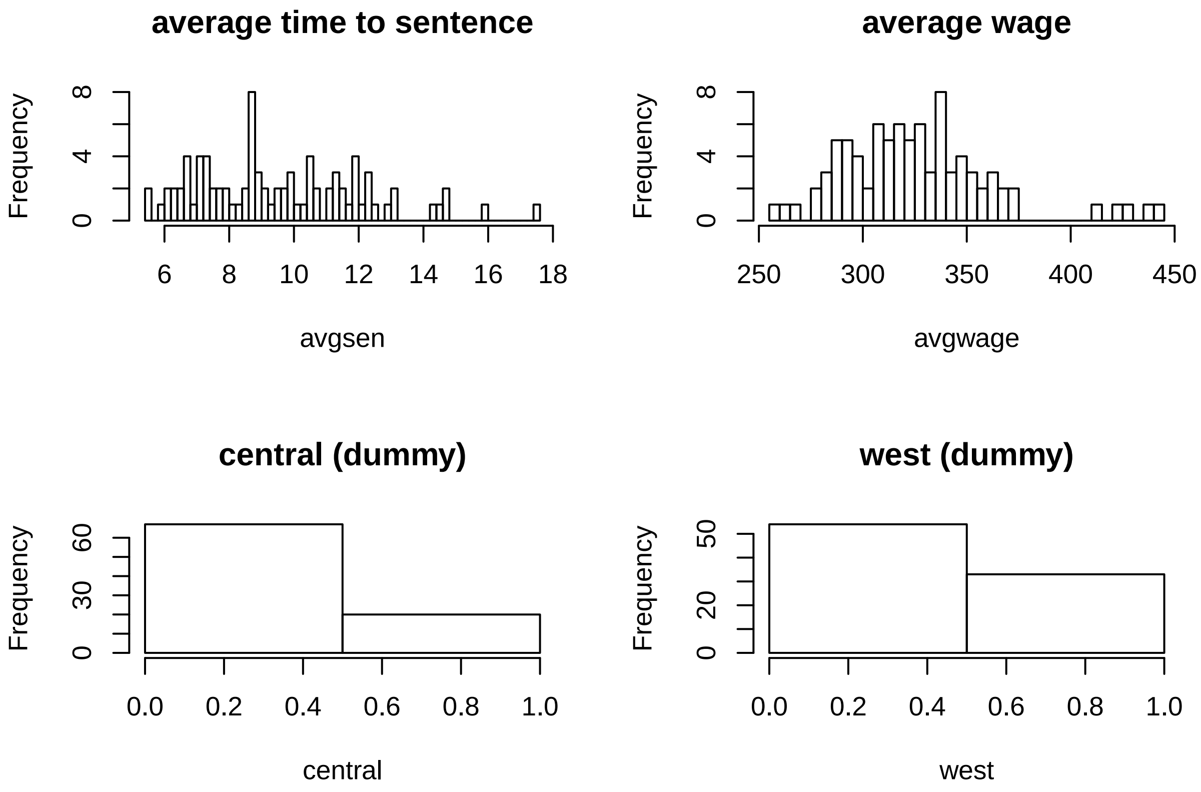
Model with the most covariates

We will add the two dummy variables west and central to the model. We will furthermore add the average time to sentence (avgsen) and the average wage (avgwage) to the model.

We will still not include variables such as the 'probability' of arrest, the 'probability' of conviction, the 'probability' of prison sentence and the proportion of face to face crimes. Including these variables would somehow imply that we believe that people who commit crimes are aware of these 'probabilities'/ratios. We do not believe that this is the case. We highly doubt that people who commit crimes are aware of such differences between different counties that are all under the jurisdiction of the same local government. Even if there are correlations between those variables and crime, we would expect those to be spurious.

We will now take a look at the avgsen, avgwage, central and west variables.

```r
par(mfrow=c(2,2))
hist(Crime_data$avgsen, breaks = 50, main = "average time to sentence", xlab = "avgsen")
hist(Crime_data$avgwage, breaks = 50, main = "average wage", xlab = "avgwage")
hist(Crime_data$central, breaks = 2, main = "central (dummy)", xlab = "central")
hist(Crime_data$west, breaks = 2, main = "west (dummy)", xlab = "west")
```



Both metric variables (average time to sentence and the average wage) have some skew, but a simple transformation (such as log transformation) would not help in these cases, which is why we will use the variables as they are.

$$\log(crime) = \beta_0+\beta_1\cdot\log(police)+\beta_2\cdot\log(density)+\beta_3\cdot\log(ymale)+\beta_4\cdot pctmin+\beta_5\cdot tax+\beta_6\cdot urban+\beta_7\cdot west+\beta_8\cdot central+\beta_9\cdot avg$$

```
model3 = lm(log(crime) ~ west + central + urban + log(police) + log(density) +
              log(ymale) +  pctmin + tax + avgsen + avgwage, data = Crime_data)
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                  Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -0.33793001  1.31466850 -0.2570   0.79784
## west         -0.14634328  0.09964353 -1.4687   0.14605
## central      -0.22065935  0.15414773 -1.4315   0.15639
## urban         0.07209853  0.17643573  0.4086   0.68395
## log(police)   0.38654326  0.19259815  2.0070   0.04831 *
## log(density)  0.43679166  0.09056040  4.8232 7.104e-06 ***
## log(ymale)    0.21157977  0.13755212  1.5382   0.12816
## pctmin        0.00884669  0.00373311  2.3698   0.02034 *
## tax          -0.00280272  0.00477788 -0.5866   0.55921
## avgsen       -0.02661713  0.01189421 -2.2378   0.02816 *
## avgwage       0.00018702  0.00157203  0.1190   0.90562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will once again take a look at the Akaike Information Criterion (AIC).

```
AIC(model3)
```

```
## [1] 21.87915
```

The AIC of our third model (21.8791495) is pretty much the same as the AIC of our second model (21.7315838). This means that adding four more independent variables to our model did not really improve the 'ratio' between fit and parsimony. In other words: The relative quality of our second model is approximately equal to the relative quality of our third model. Since the second model is also much simpler to understand than the third model, we would prefer it over the third model. Altogether, the second model seems to offer the best relative quality of all our models.

Next we will check if the four new parameters have joint statistical significance. We will once again use the wald test, which generalizes the F-test of overall significance, but allows for a heteroskedasticity-robust covariance matrix.

```
wald2 = waldtest(model2, model3, vcov = vcovHC)
wald2
```

```
## Wald test
##
## Model 1: log(crime) ~ urban + log(police) + log(density) + log(ymale) +
##     pctmin + tax
## Model 2: log(crime) ~ west + central + urban + log(police) + log(density) +
##     log(ymale) + pctmin + tax + avgsen + avgwage
##   Res.Df Df      F Pr(>F)
## 1     80
## 2     76  4 1.7057 0.1575
```

The test results indicate no joint statistical significance of the four new coefficients.

We will also take a look at the variance inflation factors (VIFs).

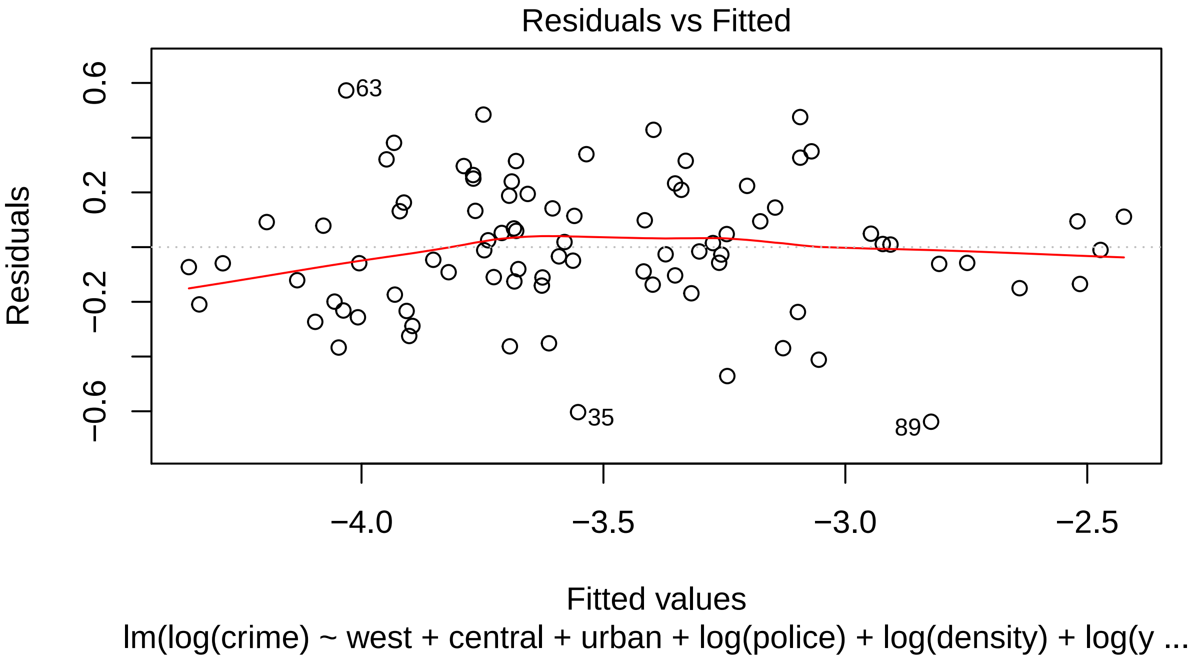
```
vif(model3)
```

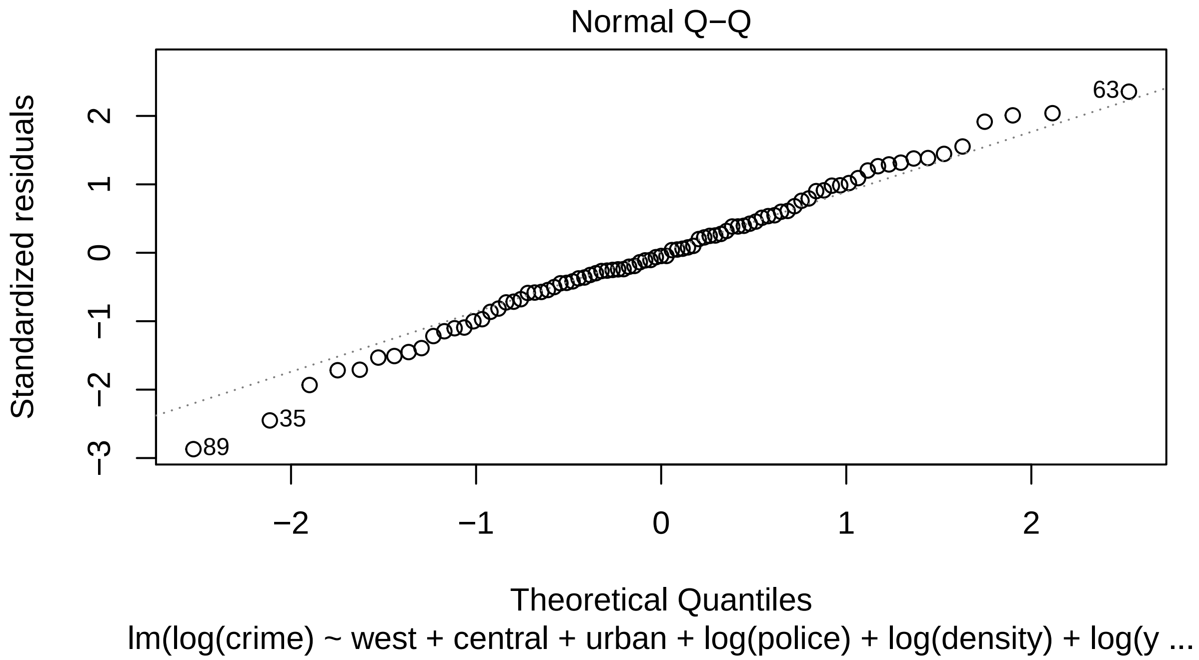
```
##         west      central        urban  log(police) log(density)
##     2.245781     3.955742     2.447472     1.821986     4.045743
##   log(ymale)       pctmin          tax       avgsen      avgwage
##     1.261710     3.258771     1.818724     1.190341     2.469172
```

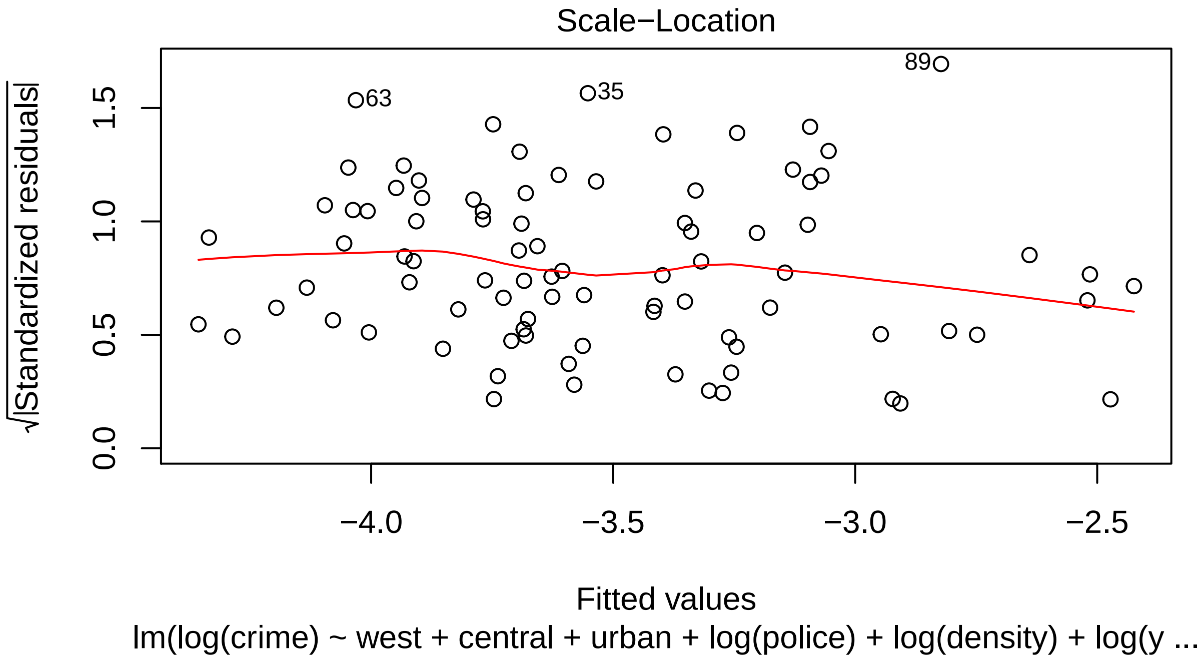
While the variance in our third model did increase quite a bit in comparison to the variance of our second model, the VIF values are not yet alarming.

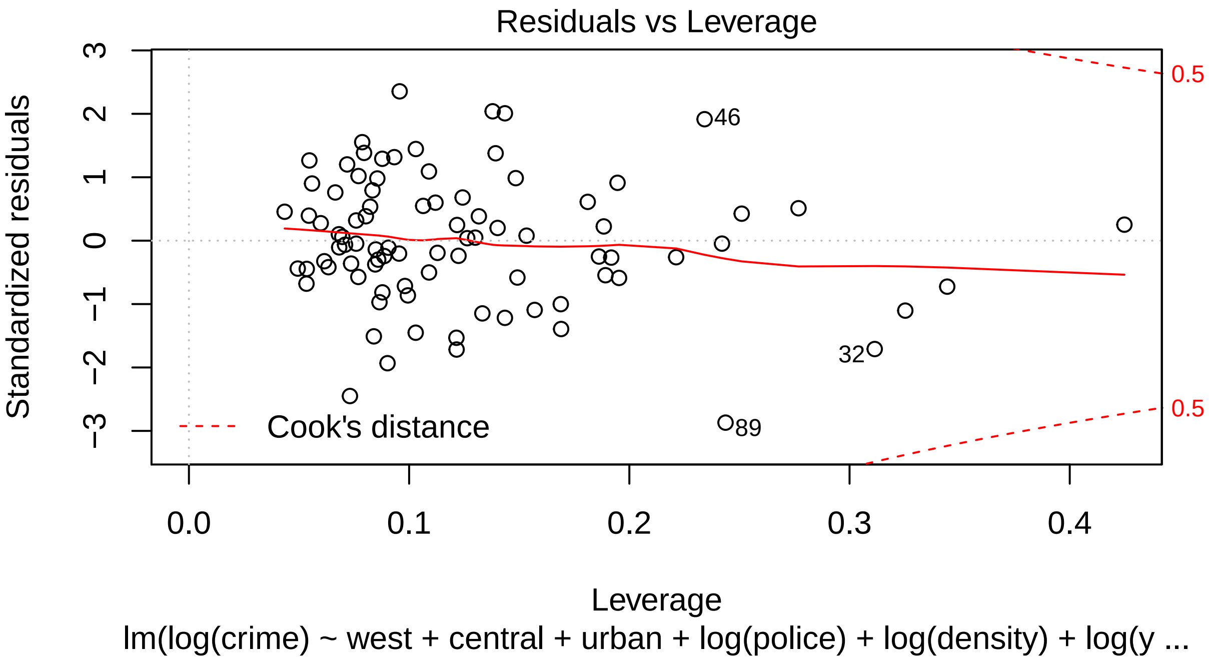
Next we will take a look at the diagnostic plots.

```
plot(model3)
```




Normal Q−Q
Standardized residuals
−3
−2
−1
0
1
2
−2
−1
0
1
2
63
35
89
Theoretical Quantiles
lm(log(crime) ~ west + central + urban + log(police) + log(density) + log(y ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−4.0
−3.5
−3.0
−2.5
63
35
89
Fitted values
lm(log(crime) ~ west + central + urban + log(police) + log(density) + log(y ...



In comparison to our first model, we have very similar results regarding all of our classical linear model assumptions. Everything that was said about MLR1 through MLR6 in our first model remains true for our third model.

The 'Residuals vs Leverage' plot does not show any data points with very high influence.

Regression table

Table 1 shows a regression table for all three models. Please note that stargazer would not print the AIC at the time of the creation of this report. Please refer to the above anlysis for the AIC.

```r
se.model1 = sqrt(diag(vcovHC(model1)))
se.model2 = sqrt(diag(vcovHC(model2)))
se.model3 = sqrt(diag(vcovHC(model3)))

stargazer(model1, model2, model3, type = "latex",
          header = FALSE,
          title = "Linear Models Predicting Crime",
          keep.stat = c("rsq", "n", "aic"),
          se = list(se.model1, se.model2, se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

Statistical and practical significance

Please note that all of our models are calculated with heterokedasticity robust standard errors. Please note also, that statistical significance is largely influenced by the sample size. The number of observations in each model is about 87, so some of the statistical significance might just be due to the (not large, but still notable) sample size.

Table 1: Linear Models Predicting Crime

	Dependent variable:		
	log(crime)		
	(1)	(2)	(3)
west			−0.146
			(0.100)
central			−0.221
			(0.154)
urban		−0.030	0.072
		(0.140)	(0.176)
log(police)	0.323	0.276	0.387^{*}
	(0.180)	(0.165)	(0.193)
log(density)	0.393^{***}	0.463^{***}	0.437^{***}
	(0.057)	(0.050)	(0.091)
log(ymale)	0.311	0.282^{*}	0.212
	(0.162)	(0.123)	(0.138)
pctmin		0.013^{***}	0.009^{*}
		(0.002)	(0.004)
tax		0.00003	−0.003
		(0.004)	(0.005)
avgsen			-0.027^{*}
			(0.012)
avgwage			0.0002
			(0.002)
Constant	−0.669	−1.391	−0.338
	(1.218)	(1.013)	(1.315)
Observations	87	87	87
R^2	0.577	0.753	0.775
Note:	*p<0.05; **p<0.01; ***p<0.001		

We will analyse the statistical and practical significance of our models.

Model 1:

Statistical significance: Only the coefficient for log(density) has (high) statistical signficance. The other two coefficients, log(police) and log(ymale) have no statistical significance.

Practical significance: All three slope coefficients have practical significance. 1% increase in each (original) variable results in approximately between 0.311% and 0.393% crime rate.

Model 2:

Statistical significance: The coefficients for log(density) and pctmin, and log(ymale) are all statistically significant. The coefficients for log(density) and pctmin are even highly statistically significant.

Practical significance: Four of our coefficients turn out to be practically significant. For police, density and ymale, 1% increase in each variable results in 0.276%, 0.463% and 0.282% increase in crime rate respectively. A unit increase in the proportion of minorities and non-whites (measured in percent) is associated with a 1.3% increase in crime rate. We would consider the coefficients on urban and tax as not practially significant. The crime rate in urban areas is only 3% lower than it is in non-urban areas.

Model 3:

Statistical significance: This model shows statistical significance for the coefficients of the log(police), log(density) (high!), pctmin and avgsen variable.

Practical significance: The practical significance is similar to model 2. The added variables, avgsen and avgwage both don't have practical significance. The two new dummy variables west and central indicate that around 15% respectively 22% fewer crimes are committed in the western respectively central counties.

Causality

Omitting variables results in biased coefficients only if the omitted variables are correlated with independent variables that are included in the model.

When we think about crime, it is likely that quite a few determinants of crime have not been accounted for in our models.

One of those determinants would be education. Since we do not have any data on education, education would be part of our error term, u. Education however is likely to be positively correlated with wealth (endogeneity!). Since we do not have information on the overall average wage per county, we will use tax revenue per capita as a measure of wealth. We believe that tax (as measure of wealth) and education are positively correlated. We were not able to get a practically or statistically significant estimate for the coefficient on tax, but if we assume that the coefficient is actually positive (as indicated by our second regression model), we would have a positive omitted variable bias. Due to the missing significance of our estimate, this argument is not very strong.

In the crime literature, we furthermore found strong evidence that children with certain characteristics commit more crime when they become adults. Among those are is the proportion of children born to single mothers and raised without father. Again, these proportions are likely correlated with wealth (measured through our tax variable), because single mother families likely pay lower tax than families with both (working) parents. The correlation of these proportions with tax is therefore expected to be negative. In our second model, the parameter of tax is positive and hence the omitted variable bias would be negative. Again, the parameter of tax is neither practically nor statistically significant, which is why this argument is not very strong.

Another omitted variable that we believe has an influence on crime, is drug abuse. We believe that drug abuse is correlated with our density variable (higher drug abuse in denser areas). The correlation between density and drug abuse is likely positive and since the coefficient of our log(density) variable is also positive (throughout all models), we expect a positive omitted variable bias. The coefficient on the log(density) variable is both, practically and statistically significant, which is why we conclude that our model is likely biased.

We furthermore also believe that drug abuse is positively correlated with our young male variable. Since the log(ymale) variable has a positive coefficient throughout all our models, we expect a positive omitted variable bias. In model 2, the coefficient on the log(ymale) variable is both, practically and statistically significant, which is why we conclude that this model is likely biased.

Due to these missing independent variables, we do not believe that our models have a causal interpretation.

It is also important to note the coefficients that appear to have the wrong sign from a causal perspective. The coefficient for police (respectively log(police)) is positive in all the models we built. However, intuitively this does not make sense, because it indicates that more police leads to more crime. If police works as a deterrent for crime, then the coefficient should be negative. Maybe a high amount of police is to be interpreted as a response to a high amount of crime, rather than a predictor of crime.

Conclusion

Young Males: We observe a correlation between young males and the amount of crime. We believe some research and policy changes could be made in a variety of areas of concern such as youth education, drug awareness, healthcare and unemployment.

Police Force: We have also observed that counties with a high number of police per capita have higher crimes per person, indicating counties might have added more police force to control crime in those regions for the year of 1988. We believe this one time set of data might be insufficient and it might be more useful to look at a time-series of crime data, before and after adding a higher amount of police force to make more meaningful policy suggestions.

Density: The data reveals a high amount of crime in dense areas. Local government could focus more on these dense areas by making policy changes and improving law enforcement or by increasing public awareness (e.g. posters in subway stations).

Minority Proportion: We do not have practical significance on our statistical analysis regarding the proportion of minorites and hence cannot give a well grounded recommendation on this factor. It would however be interesting to study this topic in more depth.