W203: Lab 3

Deepak Nagaraj

3/19/2018

About the dataset:

- A selection of counties in North Carolina
- Original: Cornwell and Trumball (1994)

What to do:

- Understand determinants of crime
- Generate policy suggestions applicable to local government
- Provide research for a political campaign

What to do for Week 1:

- Identify variables of interest
- Any transformations for each variable?
- Support from EDA?
- What covariates can identify causal effect? Which ones are problematic (multicollinearity or dampening)
- Produce 3 models:
  - One model with only explanatory variables of key interest (and no covariates)
  - Above, plus covariates that increase accuracy without introducing bias
  - Above, plus most other covariates
- Regression table, via stargazer
- Discussion of 5-10 omitted variables, for each: how it affects

How to do:

- Use OLS regression
- Omitted variables will be a major obstacle
- Aim for causal estimates, clearly explaining how omitted variables may affect conclusions

Reading the data

Let us first read the data.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
crime0 = read.csv("crime_v2.csv")
colnames(crime0)
```

```
##  [1] "county"   "year"     "crmrte"   "prbarr"   "prbconv"  "prbpris"
##  [7] "avgsen"   "polpc"    "density"  "taxpc"    "west"     "central"
## [13] "urban"    "pctmin80" "wcon"     "wtuc"     "wtrd"     "wfir"
## [19] "wser"     "wmfg"     "wfed"     "wsta"     "wloc"     "mix"
## [25] "pctymle"
```

```r
# Commented for conciseness
# structure(crime0)

# prbconv is a factor: convert to float instead
# commented for conciseness
# levels(crime0$prbconv)
crime0$prbconv = as.numeric(levels(crime0$prbconv))[crime0$prbconv]
```

```
## Warning: NAs introduced by coercion
```

```r
crime = crime0 %>%
  filter(prbconv <= 1.0) %>%
  filter(prbarr <= 1.0) %>%
  filter(prbpris <= 1.0) %>%
  filter(!(west == 1 & central == 1)) %>%
  select(-year)
# Commented for conciseness
# structure(crime)
```

Our dependent variable is going to be *crmrte*. We want to come up with a model that can predict crime rate.

The following columns are interesting:

- county: county number
- prbarr: Probability of arrest (ratio: arrest/offense)
- prbconv: Probability of conviction (ratio: conviction/arrest)
- prbpris: Probability of prison sentence (ratio: prison/total convictions)
- avgsen: Average sentence in days
- polpc: Police per capita
- density: People per square mile
- taxpc: Tax revenue per capita
- west/central/urban: Geographical location
- pctmin80: % minority (1980)
- wcon/wtuc/wtrd/wfir/wser/wmfg/wfed/wsta/wloc: Weekly wages across sectors
- mix: Offense mix: face-to-face/other
- pctymle: % young male

Possible collinearity pairs we should check for:

- urban and density
- prbarr, prbconv, prbpris
- wages across sectors

Anything else?

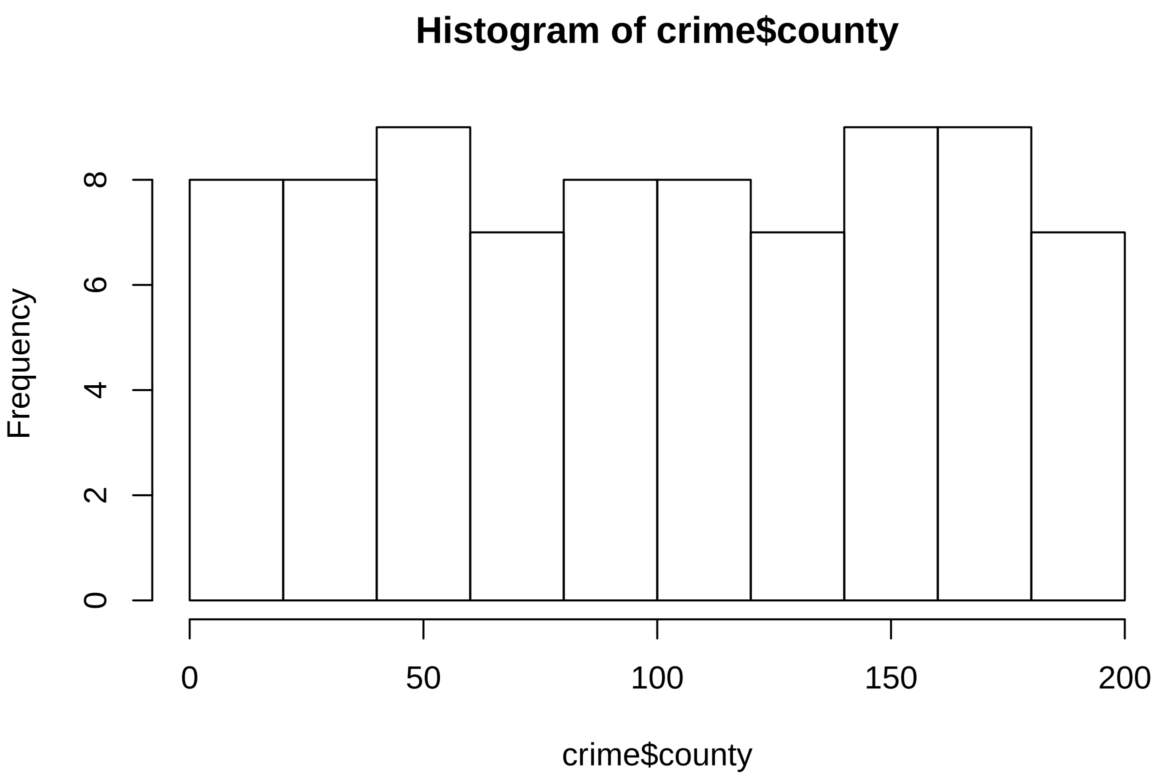
Interesting causal questions:

- Can low probabilities of arrest, conviction, or sentence drive high crime rate? Answer: Yes (for arrest)
- Can low sentencing period cause high crime rate? Answer: No
- Can fewer police per capita cause high crime rate?
- Can very high or very low density cause crime?
- Can lower tax revenue cause crime?
- Can geographical location cause crime?
- Is crime higher in urban areas, certain counties?
- Can high % of young males drive crime?
- Can low wages cause crime?
- Can high numbers of minorities cause crime, esp hate crime?

County and crime rate

Crime seems to be uniformly distributed across the counties.

```r
hist(crime$county)
```

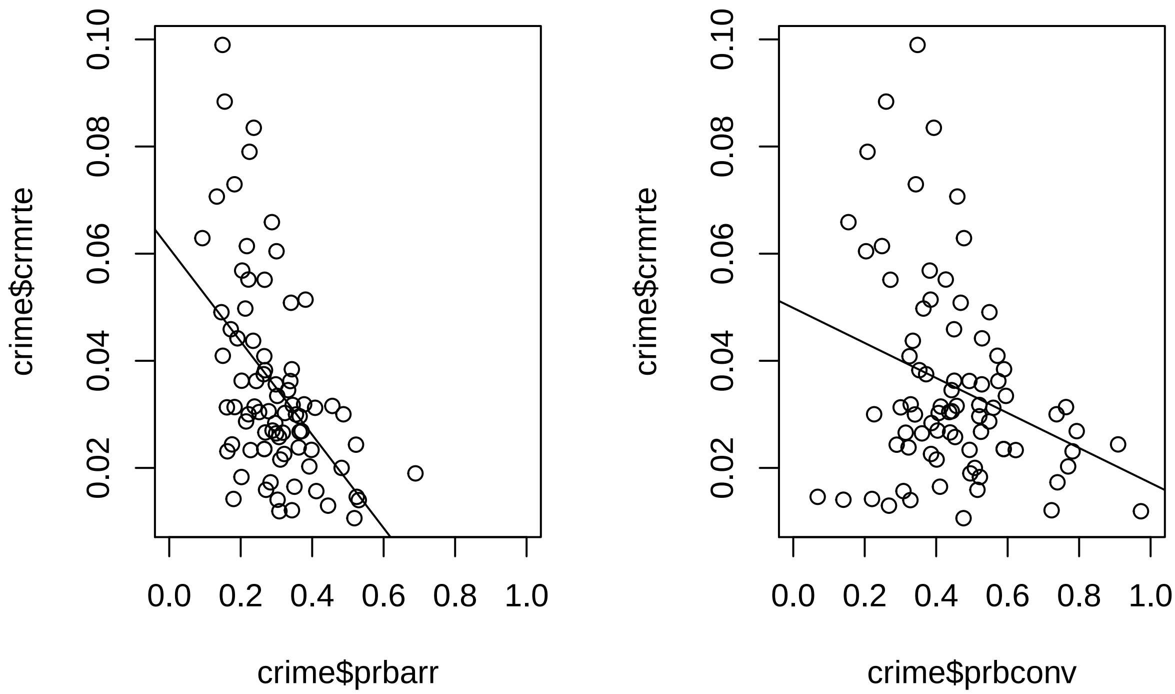


Arrests and crime rate

We see arrests and conviction driving down crime rate:

```r
par(mfrow = c(1, 2))

plot(crime$prbarr, crime$crmrte, xlim=c(0,1.0))
m = lm(crime$crmrte ~ crime$prbarr)
abline(m)

plot(crime$prbconv, crime$crmrte, xlim=c(0,1.0))
```

```r
m = lm(crime$crmrte ~ crime$prbconv)
abline(m)
```



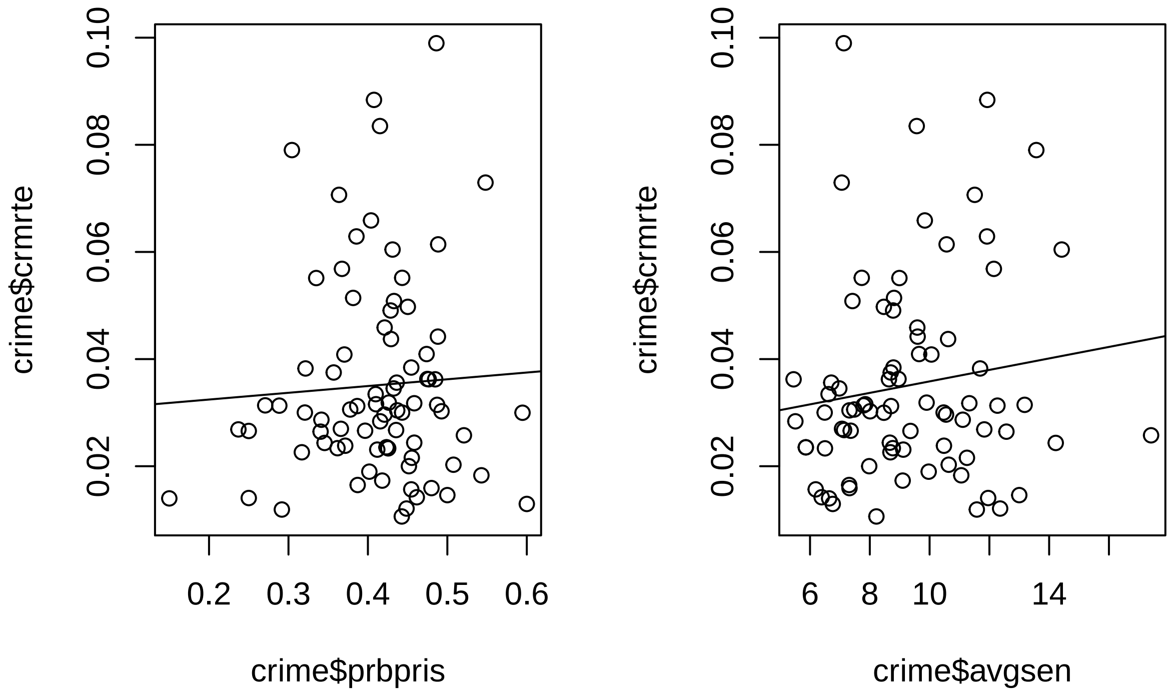
Policy recommendation: enable infrastructure to allow for catching of criminals.

TODO: Check for multicollinearity in the arrest, conviction and prison rates.

Prison sentence and crime rate

However, not much effect based on whether prison sentence happened and how long the sentence was.

```r
par(mfrow = c(1, 2))

plot(crime$prbpris, crime$crmrte)
m = lm(crime$crmrte ~ crime$prbpris)
abline(m)

plot(crime$avgsen, crime$crmrte)
m = lm(crime$crmrte ~ crime$avgsen)
abline(m)
```

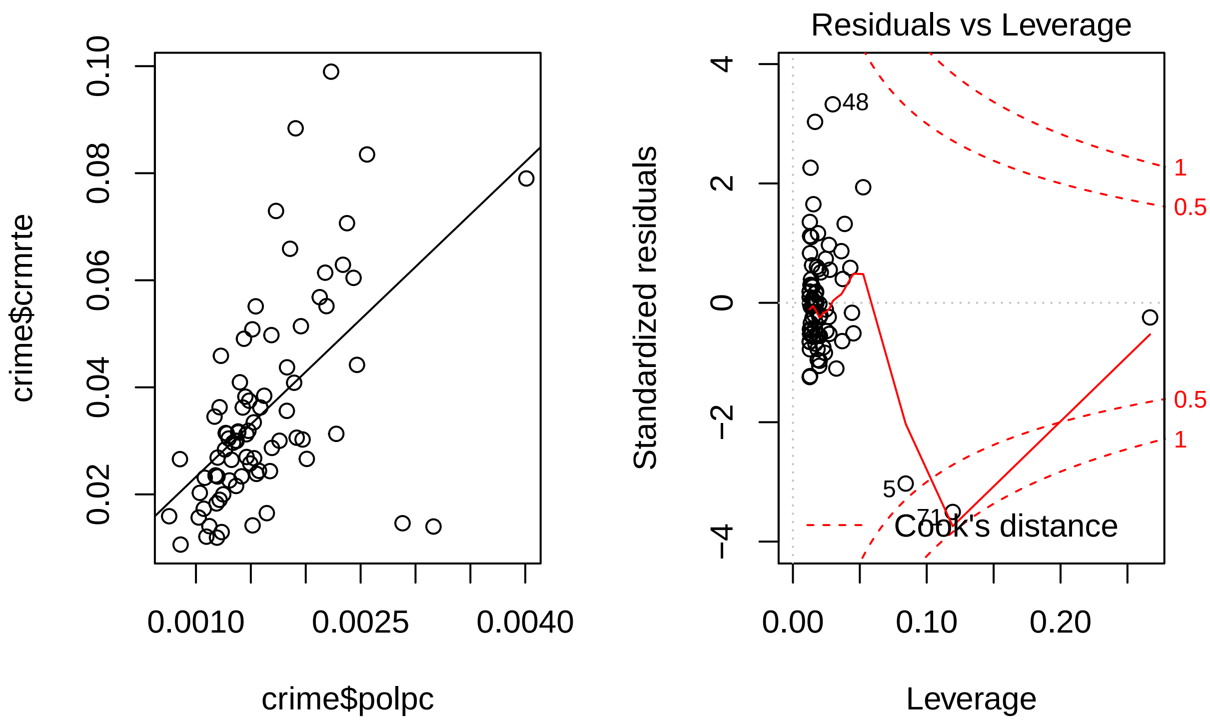


Policy recommendation: Probability of prison sentence does not affect crime rate much; it is higher where average sentences are high.

Police and crime rate

We see higher police per capita associated with higher crime rate. This could be because we are deploying more police in higher crime areas (effect, not cause) or perhaps the additional police are not being effective enough in deterring crime.

```
par(mfrow = c(1, 2))

plot(crime$polpc, crime$crmrte)
m = lm(crime$crmrte ~ crime$polpc)
abline(m)

plot(m, which=5)
```

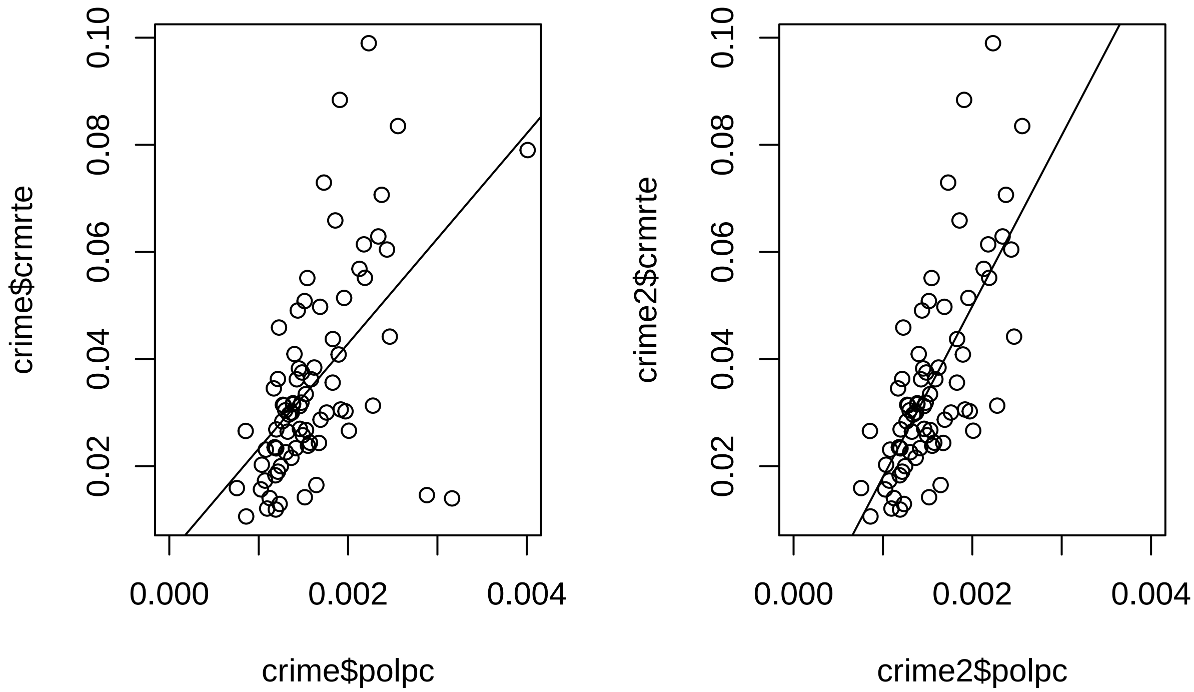


The residuals graph shows that row 72 has a lot of leverage, but it still falls short of Cook's distance of 1. So we will keep it as-is. Let us look at the row though. We can also drop the row and see how the graph changes.

```
crime %>% slice(71) %>% select(everything())
```

```
## # A tibble: 1 x 24
##   county crmrte prbarr prbconv prbpris avgsen   polpc  density taxpc  west
##    <int>  <dbl>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl>    <dbl> <dbl> <int>
## 1    173 0.0140  0.530   0.328   0.150   6.64 0.00316  2.03e-5  37.7     1
## # ... with 14 more variables: central <int>, urban <int>, pctmin80 <dbl>,
## #   wcon <dbl>, wtuc <dbl>, wtrd <dbl>, wfir <dbl>, wser <dbl>,
## #   wmfg <dbl>, wfed <dbl>, wsta <dbl>, wloc <dbl>, mix <dbl>,
## #   pctymle <dbl>
```

```
par(mfrow = c(1, 2))

plot(crime$polpc, crime$crmrte, xlim=c(0,0.004))
m = lm(crime$crmrte ~ crime$polpc)
abline(m)

# TODO: Check what's special in rows 71, 5, 23
crime2 = crime %>% slice(-71) %>% slice(-5) %>% slice(-22)
plot(crime2$polpc, crime2$crmrte, xlim=c(0,0.004))
m = lm(crime2$crmrte ~ crime2$polpc)
abline(m)
```



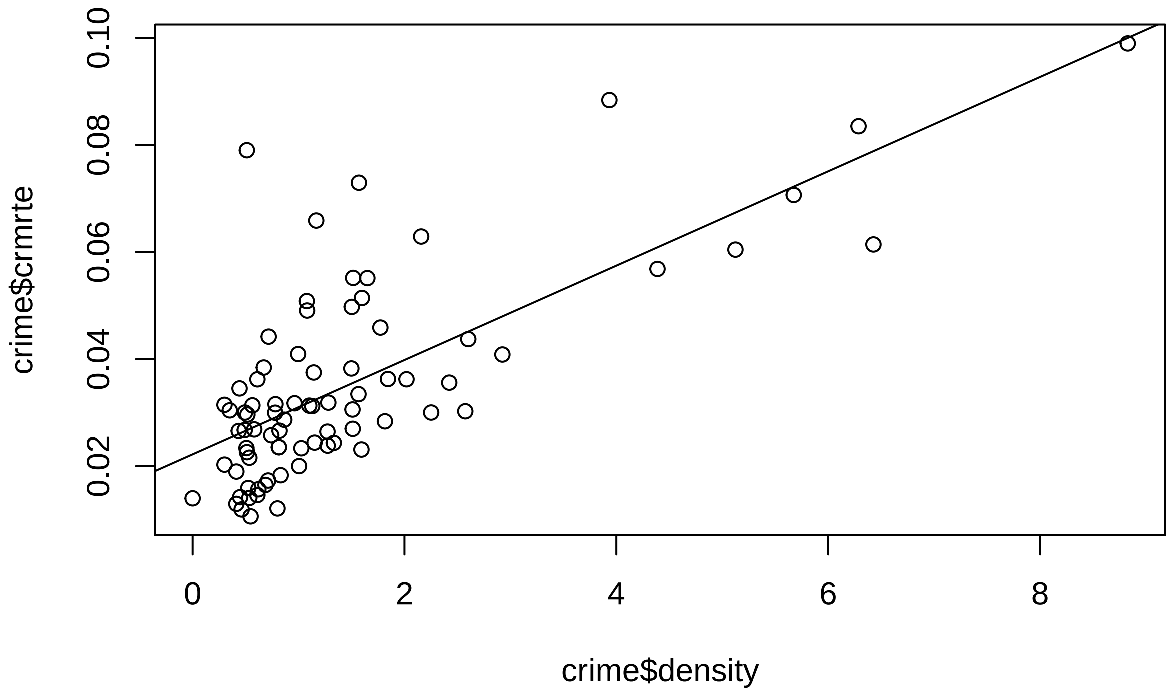
The graph has a higher slope if we remove the outliers.

Policy recommendation: increase effectiveness of police.

Density

Let us check if population density affects crime.

```
plot(crime$density, crime$crmrte)
m = lm(crime$crmrte ~ crime$density)
abline(m)
```

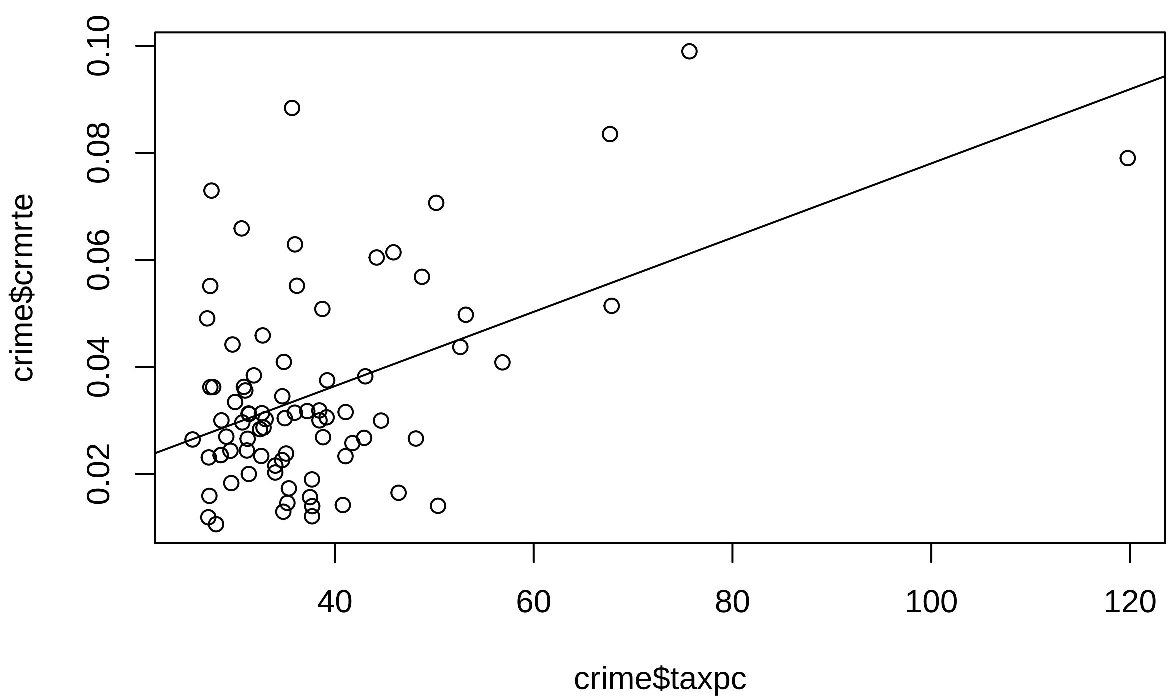


It looks like crime rate goes up with density, although data is sparse at higher densities.

Tax revenue

How does tax revenue per capita affect crime rate?

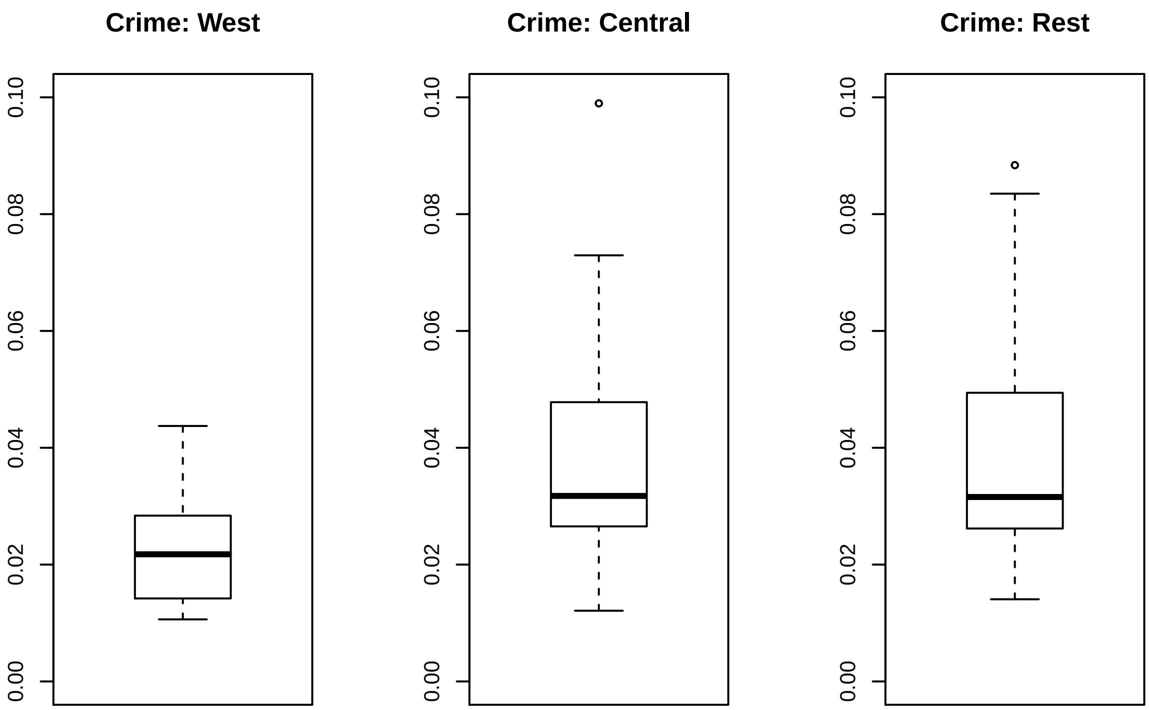
```
plot(crime$taxpc, crime$crmrte)
m = lm(crime$crmrte ~ crime$taxpc)
abline(m)
```



We see that crime rate goes up as tax revenue per capita goes up. This could be because higher tax revenue implies higher income and therefore higher chance for theft or burglary.

Geographic location

```
par(mfrow = c(1, 3))

crime_west = crime %>% filter(west == 1)
crime_central = crime %>% filter(central == 1)
crime_rest = crime %>% filter(west == 0 & central == 0)
boxplot(crime_west$crmrte, main="Crime: West", ylim=c(0,0.10))
boxplot(crime_central$crmrte, main="Crime: Central", ylim=c(0,0.10))
boxplot(crime_rest$crmrte, main="Crime: Rest", ylim=c(0,0.10))
```

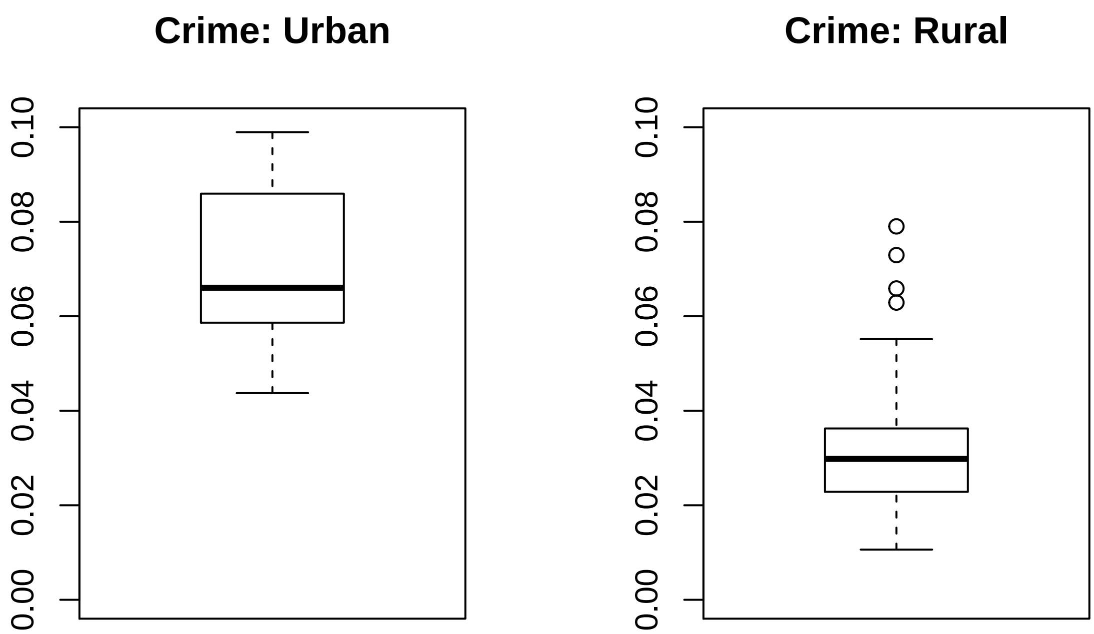


Crime rate is higher in the central region, with one clear outlier (1 in 10). But crime range is higher in the remaining regions.

Urban vs rural

Is crime rate higher in urban or rural areas?

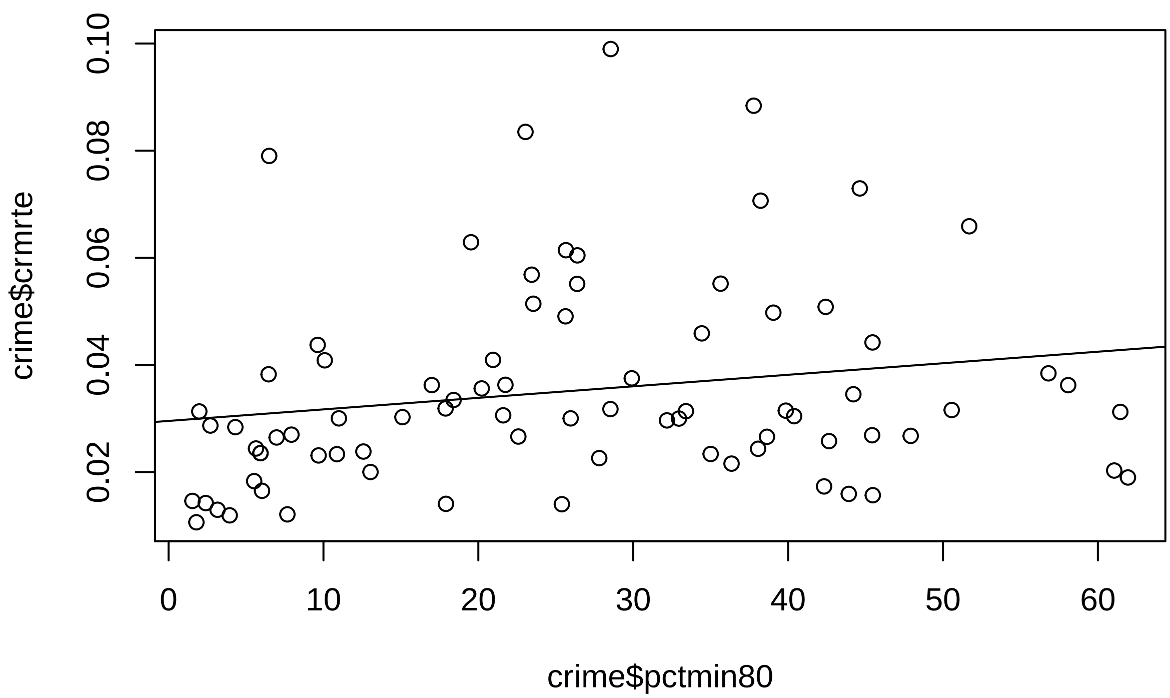
```
par(mfrow = c(1, 2))

crime_urban = crime %>% filter(urban == 1)
crime_rural = crime %>% filter(urban == 0)
boxplot(crime_urban$crmrte, main="Crime: Urban", ylim=c(0,0.10))
boxplot(crime_rural$crmrte, main="Crime: Rural", ylim=c(0,0.10))
```



Clearly, crime rate is higher in urban areas. We need to focus on urban areas.

Minorities

```
plot(crime$pctmin80, crime$crmrte)
m = lm(crime$crmrte ~ crime$pctmin80)
abline(m)
```

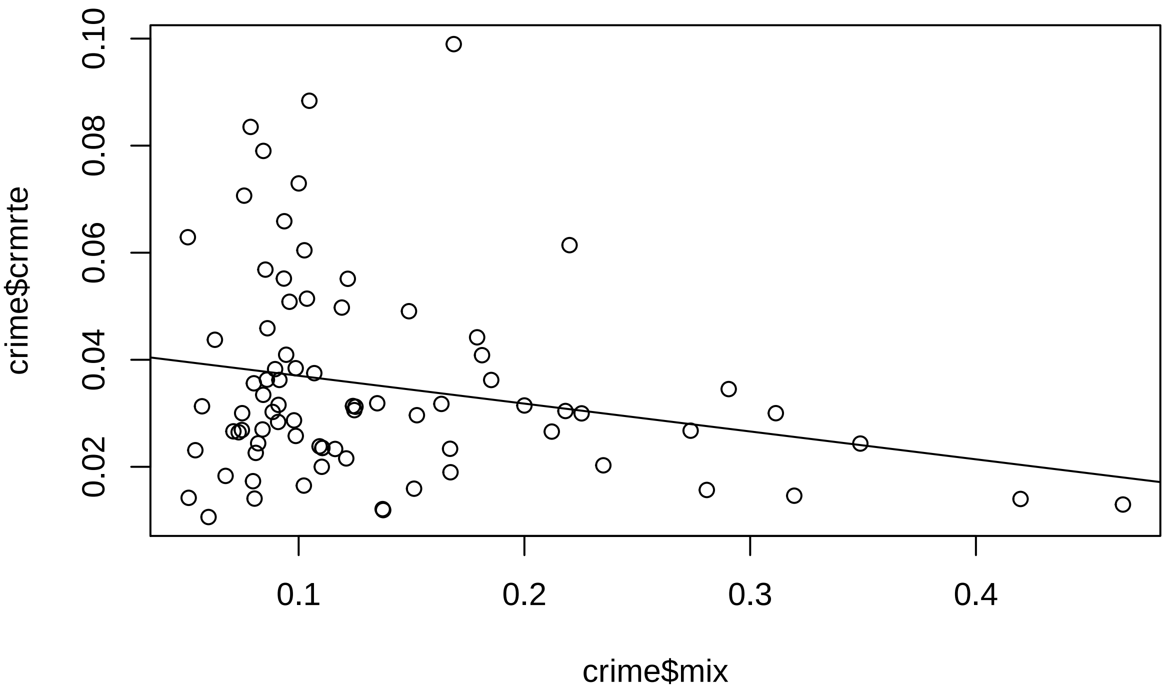


As minorities go up, we see a slight increase in crime rate.

Offense mix

Let's see how mix affects crime rate.

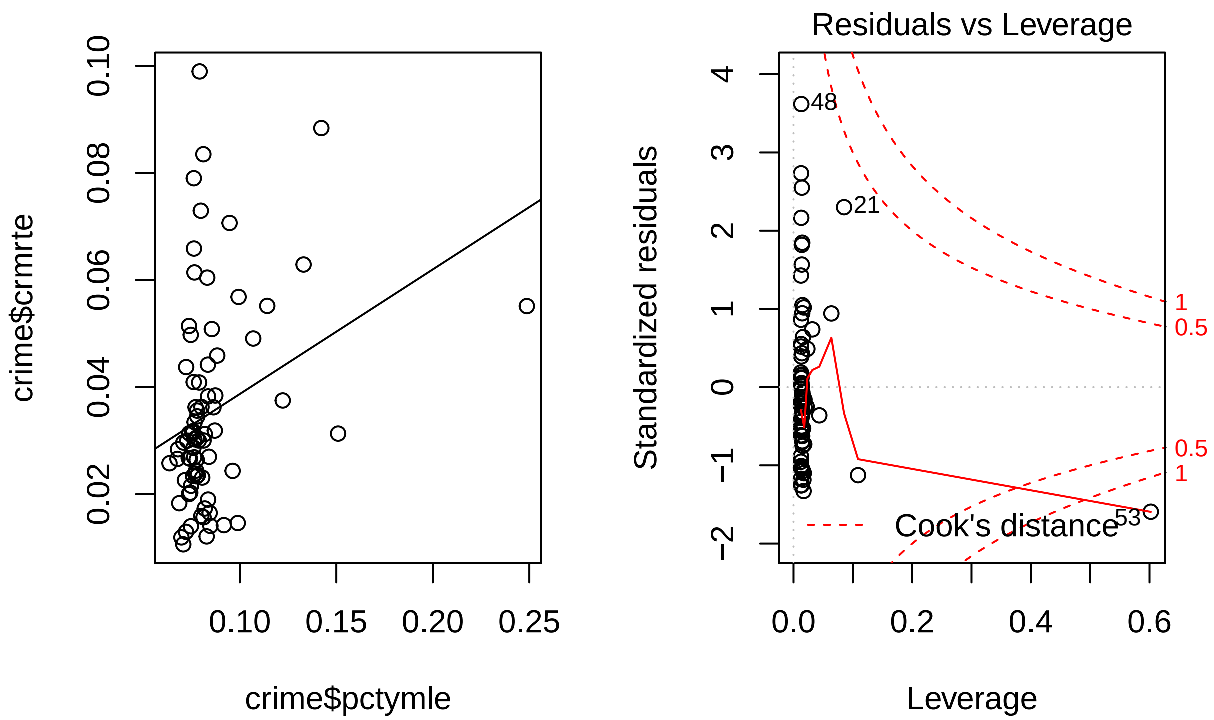
```r
plot(crime$mix, crime$crmrte)
m = lm(crime$crmrte ~ crime$mix)
abline(m)
```



We see that high crime rates are correlated with low crime mix: i.e. these crimes do not involve face-to-face interaction.

Young males

```r
par(mfrow = c(1, 2))

plot(crime$pctymle, crime$crmrte)
m = lm(crime$crmrte ~ crime$pctymle)
abline(m)

plot(m, which=5)
```



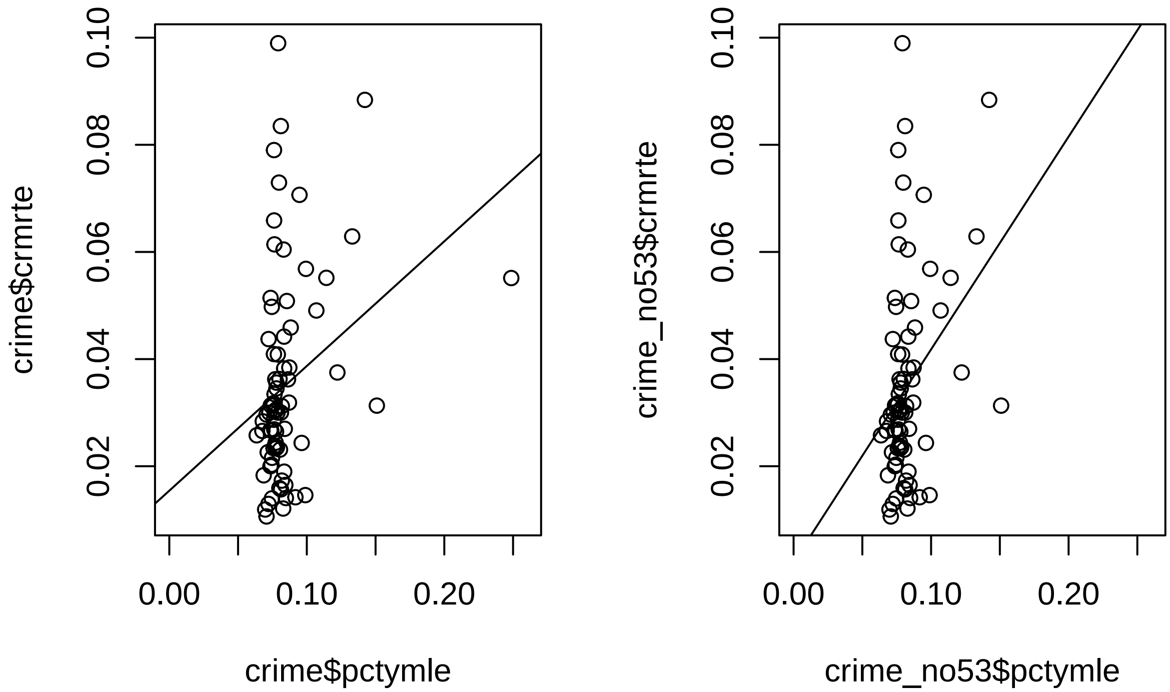
We see that crime rate goes up as we have higher percentage of young males. Row 53 is an outlier with large Cook's distance. Let us have a look at it. This is county 133, which has 5.5% crime rate and 25% young male population.

```
crime %>% slice(53) %>% select(everything())
```

```
## # A tibble: 1 x 24
##   county crmrte prbarr prbconv prbpris avgsen   polpc density taxpc  west
##    <int>  <dbl>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl> <dbl> <int>
## 1    133 0.0551  0.267   0.272   0.335   8.99 0.00154    1.65  27.5     0
## # ... with 14 more variables: central <int>, urban <int>, pctmin80 <dbl>,
## #   wcon <dbl>, wtuc <dbl>, wtrd <dbl>, wfir <dbl>, wser <dbl>,
## #   wmfg <dbl>, wfed <dbl>, wsta <dbl>, wloc <dbl>, mix <dbl>,
## #   pctymle <dbl>
```

Let us remove this data and try again:

```
par(mfrow = c(1, 2))

plot(crime$pctymle, crime$crmrte, xlim=c(0,0.26))
m = lm(crime$crmrte ~ crime$pctymle)
abline(m)

crime_no53 = crime %>% slice(-53)
plot(crime_no53$pctymle, crime_no53$crmrte, xlim=c(0,0.26))
m = lm(crime_no53$crmrte ~ crime_no53$pctymle)
abline(m)
```



There is a marked increase in slope. However, we have a wide band of crime rate in the 6-12% young male range. Our R^2 is only 9%.

```r
summary(m)
```

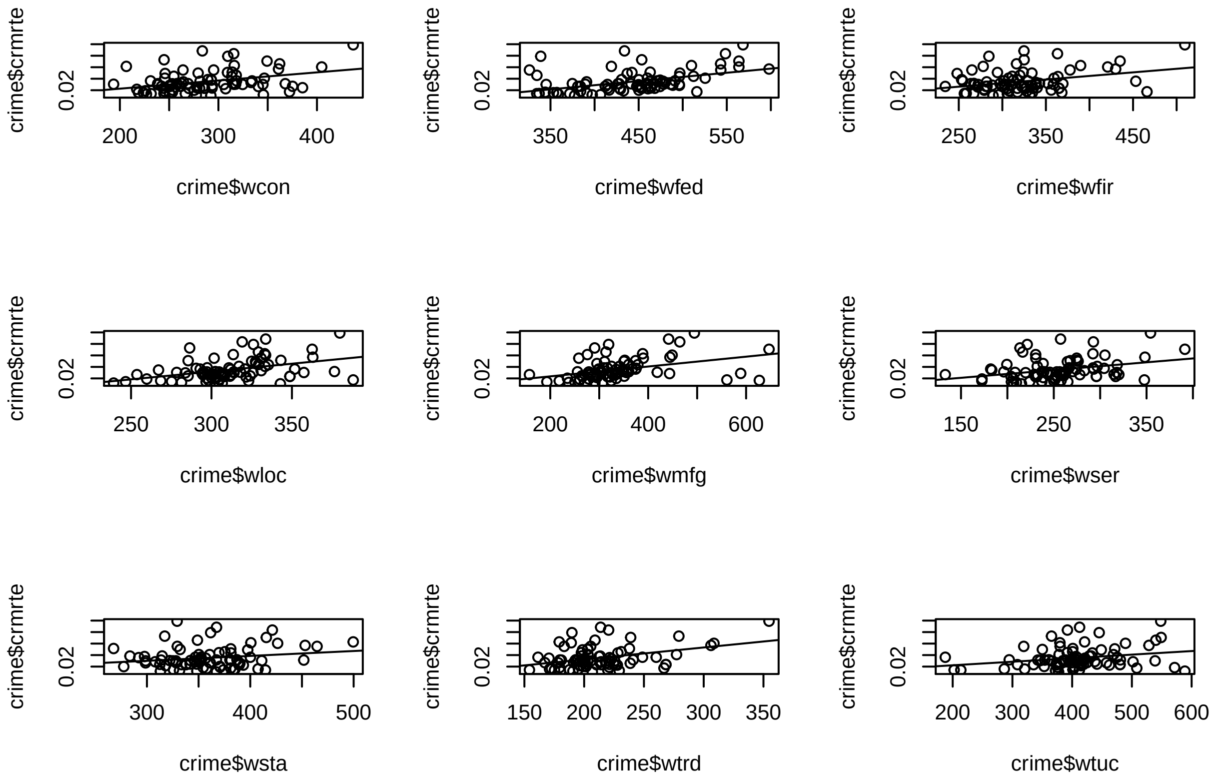
```
##
## Call:
## lm(formula = crime_no53$crmrte ~ crime_no53$pctymle)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.030719 -0.010584 -0.002399  0.005965  0.065431
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.002097   0.011096   0.189  0.85057
## crime_no53$pctymle 0.397109   0.132196   3.004  0.00359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01793 on 77 degrees of freedom
## Multiple R-squared:  0.1049, Adjusted R-squared:  0.09327
## F-statistic: 9.024 on 1 and 77 DF,  p-value: 0.003594
```

Wages and crime rate

As wage goes up, crime seems to go up.

```r
par(mfrow = c(3, 3))
m = lm(crime$crmrte ~ crime$wcon, data=crime)
plot(crime$wcon, crime$crmrte)
abline(m)
```

```r
m = lm(crime$crmrte ~ crime$wfed, data=crime)
plot(crime$wfed, crime$crmrte)
abline(m)

m = lm(crime$crmrte ~ crime$wfir, data=crime)
plot(crime$wfir, crime$crmrte)
abline(m)

m = lm(crime$crmrte ~ crime$wloc, data=crime)
plot(crime$wloc, crime$crmrte)
abline(m)

m = lm(crime$crmrte ~ crime$wmfg, data=crime)
plot(crime$wmfg, crime$crmrte)
abline(m)

m = lm(crime$crmrte ~ crime$wser, data=crime)
plot(crime$wser, crime$crmrte)
abline(m)

m = lm(crime$crmrte ~ crime$wsta, data=crime)
plot(crime$wsta, crime$crmrte)
abline(m)

m = lm(crime$crmrte ~ crime$wtrd, data=crime)
plot(crime$wtrd, crime$crmrte)
abline(m)

m = lm(crime$crmrte ~ crime$wtuc, data=crime)
plot(crime$wtuc, crime$crmrte)
abline(m)
```



Checking for collinearity

Let us look for pairs of variables with high correlation.

```r
# Build the matrix
crime_cor_matrix <- round(cor(crime), 2)
# It is symmetric
crime_cor_matrix[upper.tri(crime_cor_matrix, diag=TRUE)] <- NA
crime_cor_df <- as.data.frame(as.table(crime_cor_matrix))
# Select pairs with high correlation
crime_cor_df %>%
  filter(abs(Freq) >= 0.66) %>%
  arrange(desc(Freq)) %>%
  select(everything())
```

```
##      Var1    Var2 Freq
## 1   urban density 0.86
## 2 density  crmrte 0.72
## 3    wfir    wtrd 0.66
```

We see that *urban* and *density* have high correlation, so we can use one instead of both. There is also high correlation between wages in finance and investments $wfir$ vs. trade and retail *wtrd*. We can keep one instead of the other.

Selecting variables based on statistical significance

First, we will simply try to fit all the independent variables and then remove those that do not add a lot of statistical significance.

```r
# Source: https://www.youtube.com/watch?v=I4z3yjoEADY
m <- lm(crime$crmrte ~ ., crime)
summary(m)
```

```
##
## Call:
## lm(formula = crime$crmrte ~ ., data = crime)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0115618 -0.0036906 -0.0009302  0.0036460  0.0192144
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.196e-02  1.813e-02   0.660 0.512049
## county       2.584e-06  1.535e-05   0.168 0.866938
## prbarr      -5.223e-02  1.026e-02  -5.093 4.29e-06 ***
## prbconv     -7.328e-03  6.214e-03  -1.179 0.243272
## prbpris      1.100e-02  1.236e-02   0.890 0.377294
## avgsen      -8.389e-04  4.184e-04  -2.005 0.049832 *
## polpc        1.079e+01  2.617e+00   4.123 0.000125 ***
## density      4.869e-03  1.382e-03   3.523 0.000857 ***
## taxpc        2.010e-04  1.048e-04   1.918 0.060161 .
## west        -5.141e-03  4.266e-03  -1.205 0.233229
## central     -6.268e-03  2.702e-03  -2.320 0.024013 *
## urban        3.516e-03  6.297e-03   0.558 0.578843
## pctmin80     2.695e-04  9.368e-05   2.877 0.005667 **
## wcon         3.108e-05  2.670e-05   1.164 0.249400
## wtuc         1.281e-05  1.521e-05   0.842 0.403129
## wtrd         5.237e-05  4.196e-05   1.248 0.217235
## wfir        -4.966e-05  2.799e-05  -1.774 0.081446 .
## wser        -8.336e-05  3.054e-05  -2.729 0.008468 **
## wmfg        -2.522e-06  1.352e-05  -0.186 0.852772
## wfed         3.817e-05  2.500e-05   1.527 0.132472
## wsta        -5.022e-05  2.422e-05  -2.074 0.042689 *
## wloc         4.453e-05  4.512e-05   0.987 0.327864
## mix         -2.282e-02  1.398e-02  -1.632 0.108233
## pctymle      1.447e-01  4.436e-02   3.261 0.001893 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007224 on 56 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8531
## F-statistic: 20.94 on 23 and 56 DF,  p-value: < 2.2e-16
```

The default model has an R^2 value of 0.85.

We see that the following variables are significant (p $>$ 0.1):

- county
- prbconv
- prbpris
- west
- urban
- wcon

- wtuc
- wtrd
- wmfg
- wfed
- wloc
- mix

Let us try to build a newer model with the above (fewer) variables.

```
m <- lm(crime$crmrte ~ county + prbconv + prbpris +
          west + urban + wcon + wtuc + wtrd + wmfg +
          wfed + wloc + mix, crime)
summary(m)
```

```
##
## Call:
## lm(formula = crime$crmrte ~ county + prbconv + prbpris + west +
##     urban + wcon + wtuc + wtrd + wmfg + wfed + wloc + mix, data = crime)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.024233 -0.008562 -0.000780  0.006145  0.039782
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.368e-02  2.164e-02   0.632 0.529338
## county       2.552e-05  2.585e-05   0.987 0.327182
## prbconv     -2.639e-02  9.241e-03  -2.856 0.005704 **
## prbpris      9.783e-03  2.037e-02   0.480 0.632534
## west        -1.274e-02  3.660e-03  -3.481 0.000884 ***
## urban        2.835e-02  6.070e-03   4.671 1.49e-05 ***
## wcon         3.118e-05  4.538e-05   0.687 0.494427
## wtuc        -3.523e-06  2.563e-05  -0.137 0.891092
## wtrd        -3.887e-05  6.356e-05  -0.611 0.542956
## wmfg         8.367e-06  2.300e-05   0.364 0.717205
## wfed         3.666e-05  3.642e-05   1.007 0.317753
## wloc         4.229e-05  7.352e-05   0.575 0.567089
## mix         -3.475e-02  2.089e-02  -1.663 0.100914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01296 on 67 degrees of freedom
## Multiple R-squared:  0.5989, Adjusted R-squared:  0.527
## F-statistic: 8.336 on 12 and 67 DF,  p-value: 2.347e-09
```

Our R^2 dropped to 0.5.

Possible policy suggestions:

- Are we arresting criminals?
- Are we convicting too much or too little?
- Should we increase police presence?
- Is higher tax revenue going to cut down on crime?
- How is crime spread across West, Central and urban NC?
- How is crime correlated with % minority?
- Are low wages driving crime?
- Is there hate crime in the area ("mix"), and is it correlated with % minority?

- How is crime connected to % young male?