Lab 4

Shan He, Joanna Huang, Tiffany Jaya

18 December 2017

TODO: As this is a policy exercise, you should do your best to address the campaign's questions from a causal perspective. At the same time, you should clearly explain the limitations of your analysis, and provide discussion around whether your estimates suffer from endogeneity bias.

Introduction

TODO: A brief introduction

Exploratory Data Analysis

TODO: An initial exploratory analysis. Detect any anomalies, including missing values, top-coded or bottom-coded variables, etc.

```
library(car) # lm
library(ggplot2) # ggplot
library(lmtest) # bptest
```

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

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

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

```
library(plm) # vcovHC
```

```
## Loading required package: Formula
```

```
library(sandwich)
library(stargazer) # stargazer
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
library(tidyr) # gather
```

```
data <- read.csv("crime.csv")
str(data)
```

```
## 'data.frame':    90 obs. of  26 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
```

```
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : num  0.528 1.481 0.268 0.525 0.477 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

```
summary(data)
```

```
##        X             county           year        crmrte
##  Min.   : 1.00   Min.   :  1.0   Min.   :87   Min.   :0.005533
##  1st Qu.:23.25   1st Qu.: 51.5   1st Qu.:87   1st Qu.:0.020604
##  Median :45.50   Median :103.0   Median :87   Median :0.030002
##  Mean   :45.50   Mean   :100.6   Mean   :87   Mean   :0.033510
##  3rd Qu.:67.75   3rd Qu.:150.5   3rd Qu.:87   3rd Qu.:0.040249
##  Max.   :90.00   Max.   :197.0   Max.   :87   Max.   :0.098966
##      prbarr           prbconv           prbpris           avgsen
##  Min.   :0.09277   Min.   :0.06838   Min.   :0.1500   Min.   : 5.380
##  1st Qu.:0.20495   1st Qu.:0.34422   1st Qu.:0.3642   1st Qu.: 7.375
##  Median :0.27146   Median :0.45170   Median :0.4222   Median : 9.110
##  Mean   :0.29524   Mean   :0.55086   Mean   :0.4106   Mean   : 9.689
##  3rd Qu.:0.34487   3rd Qu.:0.58513   3rd Qu.:0.4576   3rd Qu.:11.465
##  Max.   :1.09091   Max.   :2.12121   Max.   :0.6000   Max.   :20.700
##      polpc              density           taxpc             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   :0.2333
##  3rd Qu.:0.0018856   3rd Qu.:1.5693   3rd Qu.: 41.01   3rd Qu.:0.0000
##  Max.   :0.0090543   Max.   :8.8277   Max.   :119.76   Max.   :1.0000
##     central           urban            pctmin80          wcon
##  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   :0.3778   Mean   :0.08889   Mean   :25.713   Mean   :285.4
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:38.183   3rd Qu.:315.0
##  Max.   :1.0000   Max.   :1.00000   Max.   :64.348   Max.   :436.8
##       wtuc             wtrd             wfir             wser
```

```
##  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
##       wmfg            wfed            wsta            wloc
##  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             pctymle
##  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
```

```
# sample size n = 90
nrow(data)
```

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

```
# number of variables = 26
ncol(data)
```

```
## [1] 26
```

```
# number of prbarr where probability is greater than 1
sum(data$prbarr > 1)
```

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

```
# number of prbconv where probability is greater than 1
sum(data$prbconv > 1)
```

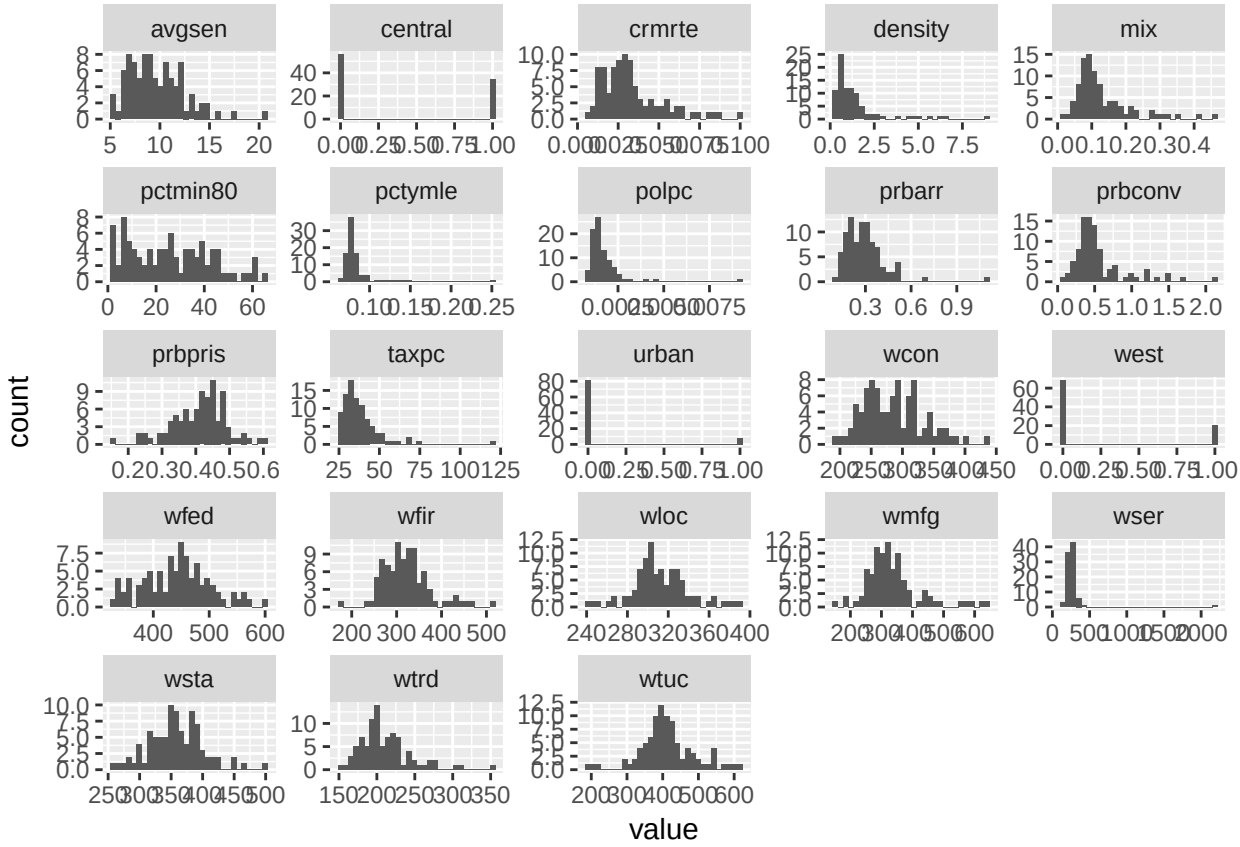
```
## [1] 10
```

```
# verify number of missing values = 0
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
```

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

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




```
# just list outliers for variables that are non-categorical
subset.data <- data[!(names(data) %in% c("X", "county", "year", "central", "urban", "west"))]

# list all outliers
list_all_outliers <- function (var) {
  outliers <- sort(boxplot.stats(var)$out)
  return(paste(length(outliers), ": ", paste(outliers, collapse=", ")))
}
lapply(subset.data, list_all_outliers)
```

```
## $crmrte
## [1] "5 :  0.0729478970170021, 0.0790162980556488, 0.0834982022643089, 0.0883848965167999, 0.09896589
##
## $prbarr
## [1] "2 :  0.689023971557617, 1.09090995788574"
##
## $prbconv
## [1] "11 :  0.972972989082336, 1.01538002490997, 1.06896996498108, 1.18292999267578, 1.22561001777649
##
## $prbpris
## [1] "1 :  0.150000005960464"
##
## $avgsen
## [1] "1 :  20.7000007629395"
##
```

```
## $polpc
## [1] "4 :  0.00316379009746015, 0.00400961982086301, 0.00445923022925854, 0.00905433017760515"
##
## $density
## [1] "8 :  3.93455100059509, 4.38875865936279, 4.8347339630127, 5.12442398071289, 5.6744966506958, 6.1
##
## $taxpc
## [1] "6 :  56.8621063232422, 61.1525115966797, 67.6796340942383, 67.8479766845703, 75.6724319458008,
##
## $pctmin80
## [1] "0 :  "
##
## $wcon
## [1] "1 :  436.766632080078"
##
## $wtuc
## [1] "9 :  187.617263793945, 202.429153442383, 213.675216674805, 548.323852539062, 548.986511230469, 5
##
## $wtrd
## [1] "5 :  277.292510986328, 279.227264404297, 306.083526611328, 308.576232910156, 354.676116943359"
##
## $wfir
## [1] "7 :  170.940170288086, 430.069702148438, 435.110717773438, 441.141296386719, 453.172210693359, 4
##
## $wser
## [1] "4 :  133.043060302734, 354.300720214844, 391.308074951172, 2177.06811523438"
##
## $wmfg
## [1] "7 :  157.410003662109, 494.299987792969, 560.780029296875, 567.059997558594, 588.989990234375, 6
##
## $wfed
## [1] "0 :  "
##
## $wsta
## [1] "1 :  499.589996337891"
##
## $wloc
## [1] "5 :  239.169998168945, 246.649993896484, 379.769989013672, 386.119995117188, 388.089996337891"
##
## $mix
## [1] "8 :  0.273622035980225, 0.28078818321228, 0.290488421916962, 0.311355322599411, 0.319526642560
##
## $pctymle
## [1] "10 :  0.0989191979169846, 0.0993558466434479, 0.106941692531109, 0.114216551184654, 0.122244790
```

Glaring outliers: 1. Probability > 1 for prbarr and prbconv 2. Outlier 2177 in wser

```
# number of urban
sum(data$urban == 1)
```

```
## [1] 8
```

```
# number of central
sum(data$central == 1)
```

```
## [1] 34
```

```r
# number of west
sum(data$west == 1)
```

```
## [1] 21
```

```r
# determine if a county can have multiple categorical label: urban, central, west
urban.county <- data[data$urban == 1, ]$county
central.county <- data[data$central == 1, ]$county
west.county <- data[data$west == 1, ]$county
# can a county be labeled as both urban and central?
uc.county <- c(urban.county, central.county)
uc.county[duplicated(uc.county)]
```

```
## [1]  63  67  81 119 183
```

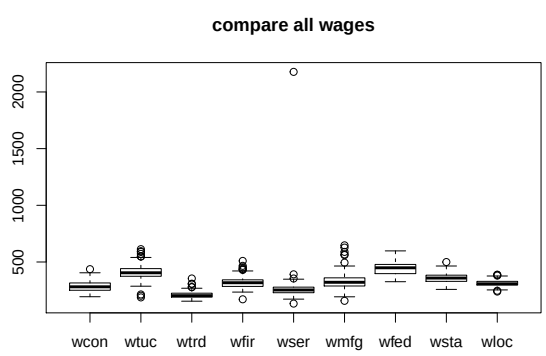
```r
# can a county be labeled as both urban and west?
uw.county <- c(urban.county, west.county)
uw.county[duplicated(uw.county)]
```

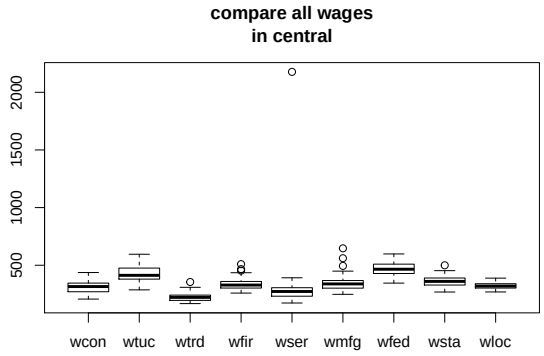
```
## [1] 21
```

```r
# I have verified that there is no county labeled as both central and west
# counties that have more than one categorical label
all.county <- c(urban.county, central.county, west.county)
all.county[duplicated(all.county)]
```

```
## [1]  63  67  81 119 183  21
```

```r
# compare all the wages
boxplot(as.vector(data[c(16:24)]), main="compare all wages")
# compare all wages in central
central.data <- data[data$central == 1, ]
boxplot(as.vector(central.data[c(16:24)]), main="compare all wages\nin central")
```

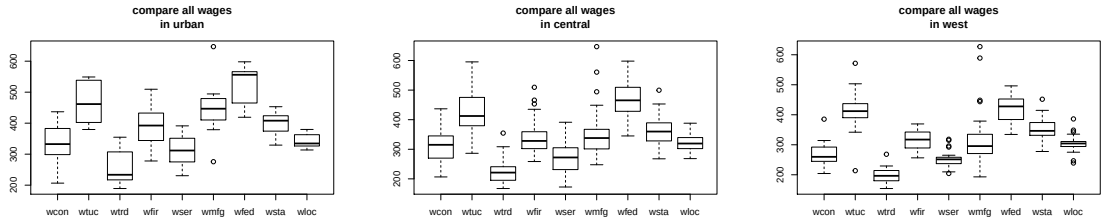





```r
# compare all wages in urban
urban.data <- data[data$urban == 1, ]
boxplot(as.vector(urban.data[c(16:24)]), main="compare all wages\nin urban")
# compare all wages in central with outlier removed
central.data[central.data$wser == 2177.06811523438, ]$wser <- mean(central.data$wser)
boxplot(as.vector(central.data[c(16:24)]), main="compare all wages\nin central")
# compare all wages in west
west.data <- data[data$west == 1, ]
boxplot(as.vector(west.data[c(16:24)]), main="compare all wages\nin west")
```

```
# compare all wages in urban and central

# compare all wages in urban and west
```



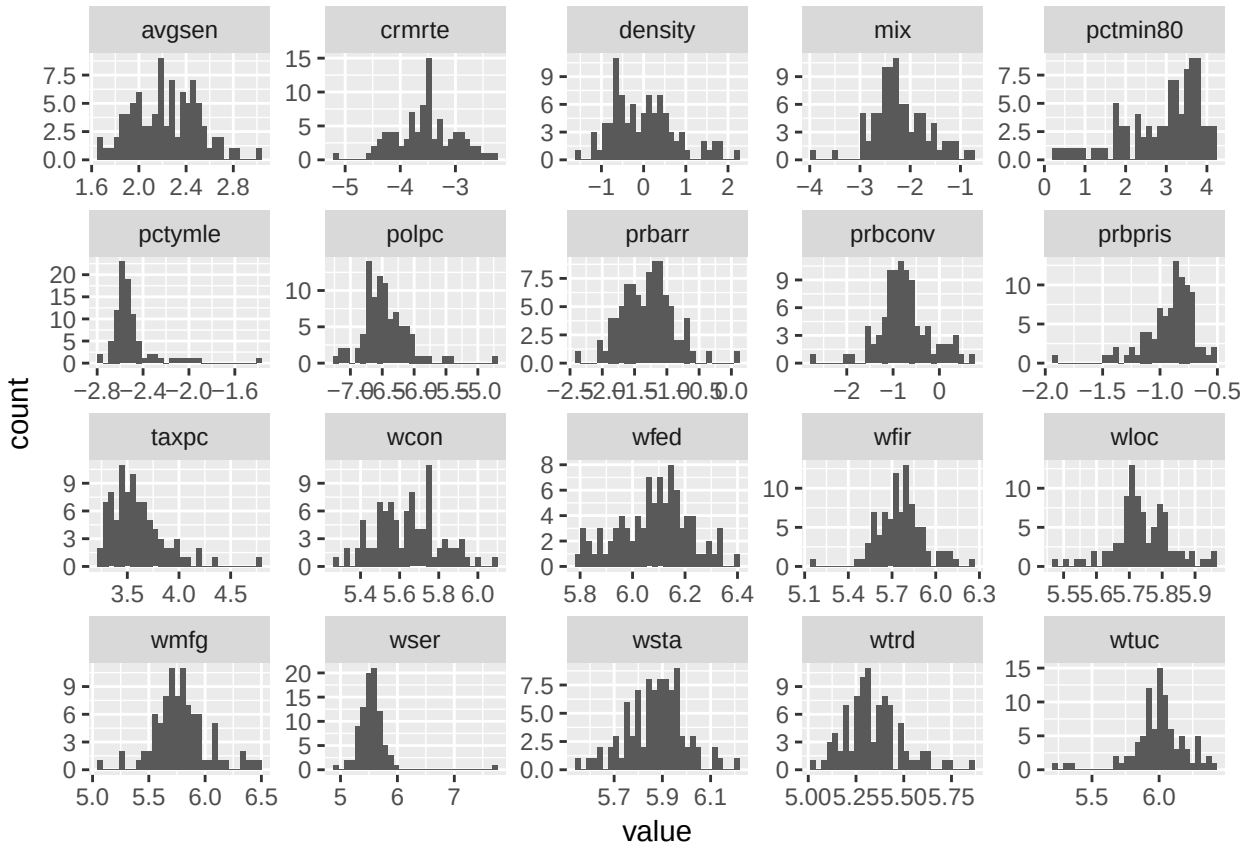
It is really weird that the outlier happens to be in central and not in urban.

Observations

- All observations were recorded for the year of 1987.
- *urban*, *central*, *west* are categorical variables.
- Most wages variables (wcon, wfed, wfir, wmfg, wser, wsta, wtrd) have a positively skewed distribution which may be due to few number people getting paid above the average.
- wtuc, wloc appear to be normally distributed.
- prbarr, prbconv appear to be positively skewed while prbpris is more negatively skewed.
- avgsen appears to be positively skewed.
- crmrte, density, polpc, prbconv: The histogram indicates that these variables have positive skews. Given the variables have a meaningful zero-point, we can take the log for a more normal distribution.

```
# taking the log of the data excluding x, county, year and categorical variables
plot.log.data <- log(data[!(names(data) %in% c("X", "county", "year", "urban", "central", "west"))])
ggplot(gather(plot.log.data), aes(value)) +
       facet_wrap(~key, scales="free") +
       geom_histogram()
```

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



Observations after taking logs: It looks like the log transformation of variables cmrte, mix, prbconv, wfed, wfir, wmfg, and wsta have made the distributions quite normal. This will help ensure the errors of the model are normal.

Proposed Model

Based on the exploratory data analysis and general intution about the potential determinants of crime, an initial proposed model specification and coefficient expectation are:

- Sum of all wages: Summing the wage variables across sectors may identify whether income inequality between counties may explain the difference in crime rates. Initial thoughts are that sum of wages (higher incomes) lead to lower crime rates, so a linear relationship between sum of wages and crimes committed per person is expected with a negative coefficient.
- log(prbarr) Probability of arrest: There is the possibility that the probability of arrest is positively correlated with crime rate, as higher numbers of arrest can increase the number of crimes recorded.
- tax revenue per capita (taxpc): Lack of government-funded resources can take the form of a lack of educational opportunities or employment options, thus leading to rise in crime rates.

Model 1: $log(crmrte) = B_0 + B_1(allwages) + B_2 log(prbarr) + B_3(taxrev) + u$

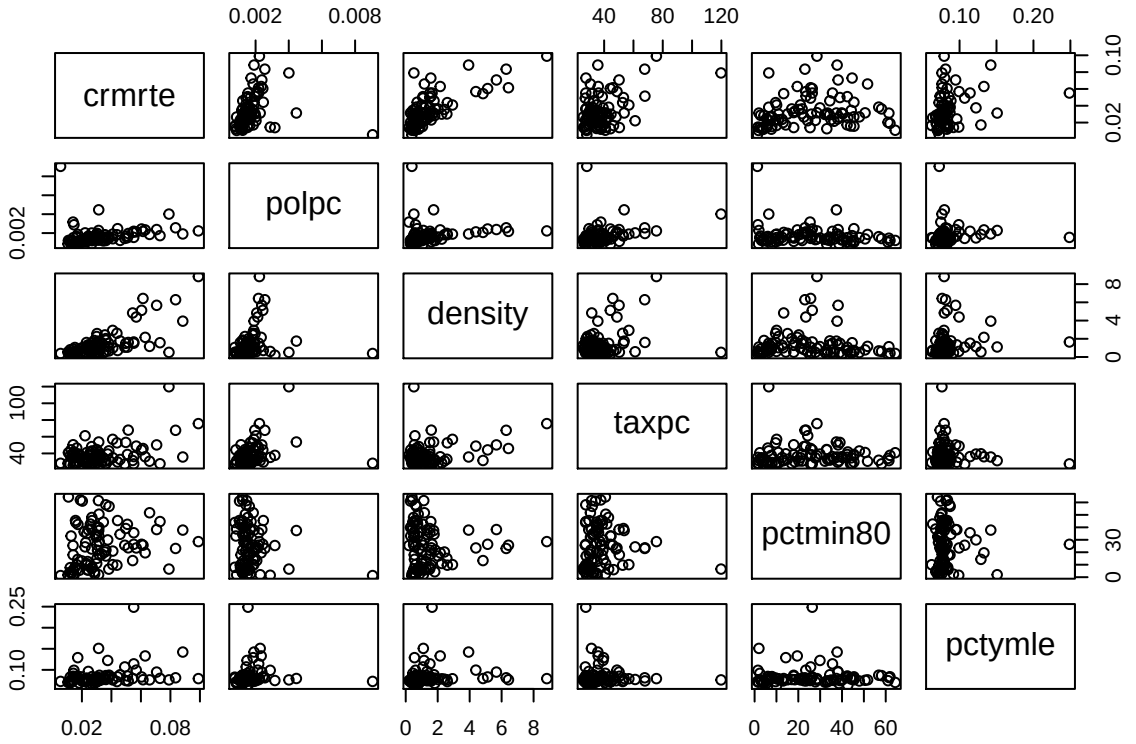
Model Building Process

TODO: A model building process, supported by exploratory analysis. Your EDA should be interspersed with, and support, your modeling decisions. In particular, you should use exploratory techniques to address *

What transformations to apply to variables and what new variables should be created. * What variables should be included in each model * Whether model assumptions are met

Potential independent variables: polpc (police per capita), density (people per sq mile), taxpc (tax revenue per capita), pctmin80 (percentage minority in 1980), pctymle (percentage young male)

```
pairs(crmrte ~ polpc + density + taxpc + pctmin80 + pctymle, data=data)
```



Model Specifications

TODO: * One model with only the explanatory variables of key interest (possibly transformed, as determined by your EDA), and no other covariates. * One model that includes key explanatory variables and only covariates that you believe increase the accuracy of your results without introducing bias (for example, you should not include outcome variables that will absorb some of the causal effect you are interested in). This model should strike a balance between accuracy and parsimony and reflect your best understanding of the determinants of crime. * One model that includes the previous covariates, and most, if not all, other covariates. A key purpose of this model is to demonstrate the robustness of your results to model specification.

For your first model, a detailed assessment of the 6 CLM assumptions. For additional models, you should check all assumptions, but only highlight major differences from your first model in your report.

A well-formatted regression table summarizing your model results. Make sure that standard errors presented in this table are valid. Also, be sure to comment on both statistical and practical significance.

Causality

TODO: A detailed discussion of causality. In particular, include a discussion of what variables are not included in your analysis and the likely direction of omitted variable bias. Highlight any coefficients you find that appear to have the wrong sign from a causal perspective, and explain why this is the case.

Conclusion

TODO: A brief conclusion with a few high-level takeaways.