Lab 4

Shan He, Joanna Huang, Tiffany Jaya

18 December 2017

```
As this is a policy exercise, you should do your best to address the campaign’s questions from a causal
```

Introduction

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")
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
```

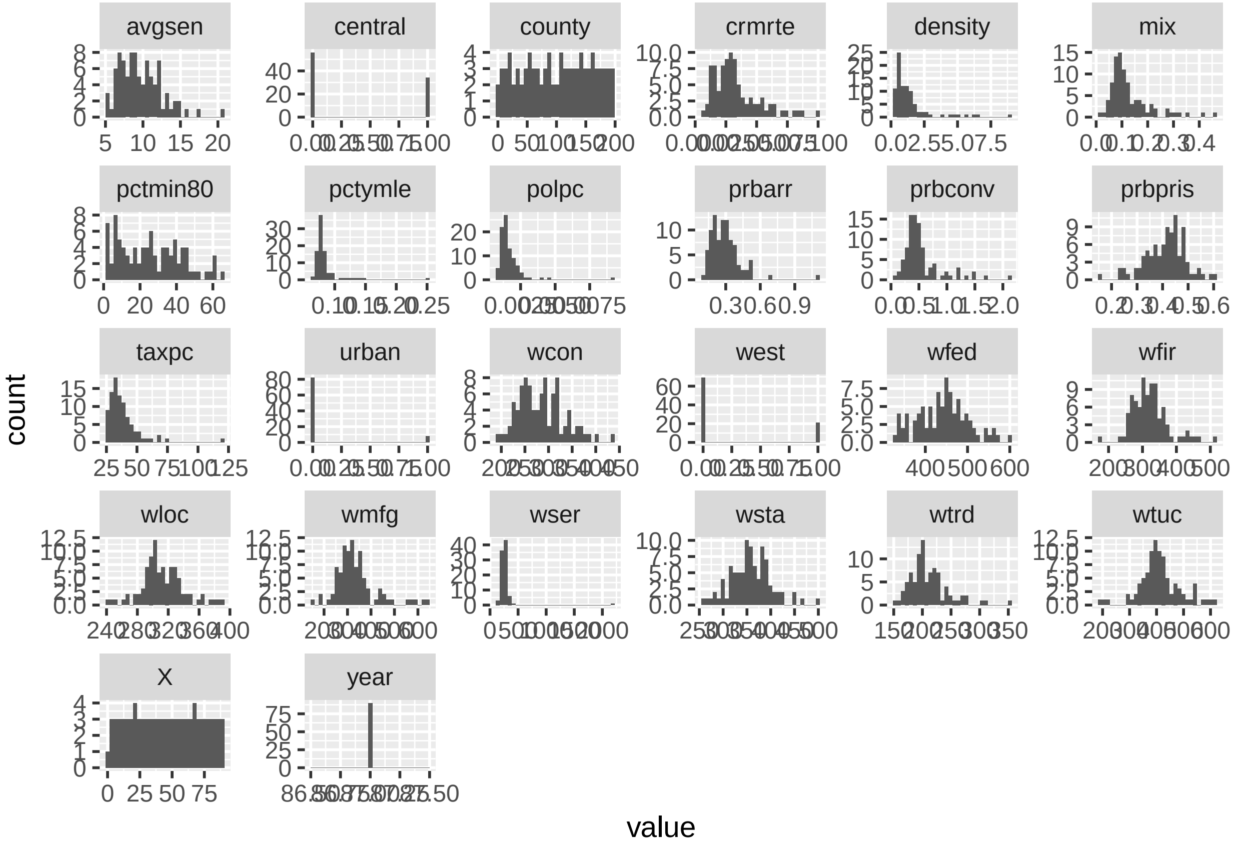
There are no missing values.

```r
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.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`.
```



Dependent variable: *crmrte* has a slight positive skew -> perform a log transform Categorical variables: *central*, *urban*, *west*

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

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.