Lab 3: Policy Recommendations for Reducing Crime

Ammara Essa, Curtis Lin, and David Wheeler

A-

Table of Contents:

+ generally good report.

- comment and interpret on summary statistics
- focus interpretation and discussion on main variables
- don't disregard the data when interpreting the estimates

Introduction, variable selection & research question

Back to Table of Contents

Introduction

Public safety is a major concern for constituents and elected officials, especially during elections. By identifying determinants of crime (represented by *crmrte*: crimes comitted per person) in North Carolina (NC), we can develop policy suggestions to help the campaign deliver on their promise of reducing crime rates.

Variable selection

Of the 22 candidate explanatory variables (25-*crmrte-year-county* = 22) only a subset can feasibly be altered with cheap and simple changes by politicians at the state level.

- By process of elimination the wage variables will likely not be simple or cheap to alter by politicians. Moreover, the potential links between crime and weekly wages are likely too nuanced and will be met with resistance if lowered.
- To change demographic variables like *pctmin80* and *pctymyle* without being accused of prejudice seems futile. To attempt such a task would not be cheap or simple.
- The *mix* variable is not responsive to change by politicians, unless the politicians are also criminals.

could inform about which demographic to focus on?

- Similarly, the geographic variables like *west*, *urban* and *central* are not amenable to change, unless politicians want to relocate people without infringing on their rights.
- Likewise, this applies to *density*- it seems unlikely for a politician to enforce laws that regulate the number of people that can live in a given space. However, long-term housing policies may alleviate population density but quickly or cheaply.
- Although tax revenues per capita (*taxpc*) can be changed by politicians, one needs only to recall the most recent federal attempts to change taxes to note that changes to our tax structure are not simple and cheap, but are often met with resistance.
- Finally, to ask politicians to change the average sentence length (*avgsen*) and the 'probability' of prison sentences (*prbpris*) appears to be asking too much. Again, although there is precedence for these types of changes, they are certainly not simple and cheap and will likely require an arduous process punctuated by protests and episodes of public resistance.

Although these variables may not have potential to be changed (simply and cheaply) by politicians, we will still include some in our model building process as covariates and to control for their influences on crime. Finally, this leaves several variables which could be simply and cheaply altered by politicians to reduce crime: *polpc*, *prbarr*, and *prbconv*

Research question

Can better policing, indicated by *polpc*, *prbarr*, *prbconv*, reduce crime rates (*crmrte* = crimes committed per person) in NC?

Importance: This question is important because, if we can improve policing, then it follows that crime rates in NC should decline. Moreover, changes in policing are relatively simple and cheap to implement. Finally, if crime rates can be reduced by better policing then we can provide relatively cheap, simple, and actionable policy recommendations to the campaign.

Hypothesized model that explains crimes rate (crimes committed per person):

$$crmrte = \beta_0 + \beta_1 polpc + \beta_2 prbarr + \beta_3 prbconv + \beta_4 prbarr \cdot prbconv + u$$

Null hypotheses:

1. Police per capita (*polpc*) has no effect on y, crimes committed per person.
2. 'Probability' of arrest (*prbarr*) has no effect on y, crimes committed per person.
3. 'Probability' of conviction (*prbconv*) has no effect on y, crimes committed per person.
4. The 'probability' of conviction (*prbconv*) does not depend on the 'probability' of arrest (*prbarr*).

- $H_0^{(1)}: \frac{\partial crmrte}{\partial polpc} = 0 \Rightarrow \beta_1 = 0$
- $H_0^{(2)}: \frac{\partial crmrte}{\partial prbarr} = 0 \Rightarrow \beta_2 = 0$
- $H_0^{(3)}: \frac{\partial crmrte}{\partial prbconv} = 0 \Rightarrow \beta_3 = 0$
- $H_0^{(4)}: \frac{\partial crmrte}{\partial prbconv}\Big|_{arrested} = 0 \Rightarrow \beta_4 = 0$

Alternative hypotheses:

1. Police per capita (*polpc*) has an effect on y, crimes committed per person.

2. 'Probability' of arrest (*prbarr*) has an effect on y, crimes committed per person.
3. 'Probability' of conviction (*prbconv*) has an effect on y, crimes committed per person.
4. The 'probability' of conviction (*prbconv*) depends on the 'probability' of arrest (*prbarr*).

- $H_0^{(1)}: \beta_1 \neq 0$
- $H_0^{(2)}: \beta_2 \neq 0$
- $H_0^{(3)}: \beta_3 \neq 0$
- $H_0^{(4)}: \beta_4 \neq 0$

good

Covariates of interest

Although, the variables above will be key in providing advice and guidance to the political campaign, we need to include other variables to control for other sources of variation.

- geography (e.g. *west*, *urban*, and *central*) since crime is often associated with place & resultant models could be used to inform policies for specific locales.
- demographics (e.g. *density*, *pctmin80* and *pctymle*) since crime is often associated with demographics & resultant models could be used to help demographics of interest reduce crime.
- economics (e.g. *taxpc*, *wfed*, *wtuc*, and *wtrd*) since wages could conceivably be associated with crime rates & resultant models could be used to justify wage change policies to reduce crime rates.
- imprisonment (e.g. *prbpris* and *avgsen*) since a high 'probability' of a prison sentence and a long average sentence length could reduce the likelihood that criminals reoffend.

Context good

- We recognize that the dataset used herein is a cross sectional sample and provides a snapshot of the population in time (1987). Thus, this limited dataset may not be best for causal inference. However, based on our understanding of the variables, we will attempt to draw conclusions and provide recommendations using this data. Moreover, we will discuss a few key sources of omitted variable bias that we believe have been potentially introduced since we can no longer control for them using multi-year panel data.
- NC has 100 counties, the latest incorporated in 1911 (https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina) (https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina)). However, the dataset available to us only captures 90 unique counties. With that said, we will use this data to draw conclusions regarding crime rates in NC.

cite the US census instead! Wiki can contain mistakes

- Based on the summary of average prison sentence (Table 1), we see that the sentence length is approximately 10 days on average and at most approximately 21 days. This may be an indicator that law enforcement is dealing with more misdemeanor crimes than felony crimes.
- Finally, while most of the variables were collected in 1987, two of the variables, *pctmin80* and *pctymle* were collected in 1980! ∴ any effect that these variables have on crime rates may be confounded by time- we may not be able tease apart differences due to *pctmin80* and *pctymle* from differences in time.

Descriptive anlayses | exploratory data analysis

Back to Table of Contents

Install and invoke packages

In [9]:
```
# install.packages("car")
# install.packages("GGally")
# install.packages("scatterplot3d")
# install.packages("lmtest")
# install.packages("sandwich")
# install.packages("corrplot", repos="http://cran.us.r-project.org")
library(Hmisc)
library(corrplot)
library(car)
library(stargazer)
library(GGally)
library(scatterplot3d)
library(lmtest)
library(sandwich)
```

Load data

In [10]:
```
df = read.csv("crime_v2.csv",header=T)
```

Check classes of each variable

Note that *prbconv* is treated as a factor, *county*, *west*, *central* and *urban* are treated as numeric variables, and *year* is a constant.

In [11]:
```
1 str(df)
```
just show summary table

```
'data.frame':   97 obs. of  25 variables:
 $ 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 : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52
3 59 78 42 86 ...
 $ 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 ...
```

Convert *prbconv* from factor to numeric since it expresses a *continuous* outcome

In [74]:
```
1 df$prbconv = as.numeric(as.character(df$prbconv))
```

Convert *county*, *west*, *central*, and *urban* from numerics to factors

In [13]:
```
1 df$county= factor(df$county)
2 df$west= factor(df$west)
3 df$central= factor(df$central)
4 df$urban= factor(df$urban)
```

Table 1: Descriptive statistics

not an issue.

Note:

- *prbarr* and *prbconv* variables have values that exceed 1.
- the variables are expressed at different scales, from 0.002 to 2177.1
- the weekly wage variables look like they have been censored/bottom and or top-coded. Perhaps extreme values were ommitted...

In [14]:
```
stargazer(df, type="text",
          title="Descriptive statistics",
          digits= 1)
```

comment!

```
Descriptive statistics
=============================================================
Statistic N  Mean  St. Dev.  Min  Pctl(25) Pctl(75)   Max
-------------------------------------------------------------
year      91 87.0    0.0    87.0    87.0     87.0    87.0
crmrte    91 0.03    0.02   0.01    0.02     0.04     0.1
prbarr    91  0.3     0.1    0.1     0.2      0.3      1.1
prbconv   91  0.6     0.4    0.1     0.3      0.6      2.1
prbpris   91  0.4     0.1    0.2     0.4      0.5      0.6
avgsen    91  9.6     2.8    5.4     7.3      11.4    20.7
polpc     91 0.002  0.001   0.001   0.001    0.002    0.01
density   91  1.4     1.5    0.000   0.5      1.6      8.8
taxpc     91 38.1    13.1   25.7    30.7     40.9    119.8
pctmin80  91 25.5    17.0    1.3     9.8      38.1    64.3
wcon      91 285.4   47.5   193.6   250.8    314.8    436.8
wtuc      91 411.7   77.3   187.6   374.6    443.4    613.2
wtrd      91 211.6   34.2   154.2   190.9    225.1    354.7
wfir      91 322.1   53.9   170.9   286.5    345.4    509.5
wser      91 275.6  206.3   133.0   229.7    280.5   2,177.1
wmfg      91 335.6   87.8   157.4   288.9    359.6    646.8
wfed      91 442.9   59.7   326.1   400.2    478.0    598.0
wsta      91 357.5   43.1   258.3   329.3    382.6    499.6
wloc      91 312.7   28.2   239.2   297.3    329.2    388.1
mix       91  0.1     0.1    0.02    0.1      0.2      0.5
pctymle   91  0.1     0.02    0.1    0.1      0.1      0.2
-------------------------------------------------------------
```

comment. Does this seem reasonable?

From str(df), a total of 97 observations are included in the dataset. However, in the stargazer table the number of observation is 91 (N = 91). The difference of number of observations between str(df) and stargazer table indicates that there might be 6 missing values (97-91 = 6) in the dataset. The missing values in the dataset will be identified in the following section.

Remove the *year* variable - it is a constant

you don't need to do that, just don't include in your regression model.

In [15]:
```
df$year = NULL
```

Locate missing values, NAs

Table 2: Missing values

Note that rows 92-97 are all missing data, NAs. There are 6 missing values in the dataset.

In [16]:
```
1 apply(is.na(df), 2, which)
```

county	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	⋯	wtuc	wtrd	wfi
92	92	92	92	92	92	92	92	92	92	⋯	92	92	92
93	93	93	93	93	93	93	93	93	93	⋯	93	93	93
94	94	94	94	94	94	94	94	94	94	⋯	94	94	94
95	95	95	95	95	95	95	95	95	95	⋯	95	95	95
96	96	96	96	96	96	96	96	96	96	⋯	96	96	96
97	97	97	97	97	97	97	97	97	97	⋯	97	97	97

Remove rows with missing values, NAs, for analyses

In [17]:
```
1 crime.narm <- na.omit(df)
```

Identify duplicate rows

Note two duplicate entries for county 193.

In [18]:
```
1 crime.narm[duplicated(crime.narm) == TRUE,]
```

	county	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	wes
89	193	0.0235277	0.266055	0.588859	0.423423	5.86	0.00117887	0.8138298	28.51783	

In [19]:
```
1 nrow(crime.narm)
```

91

Delete duplicate row

In [20]:
```
1 crime.narm = crime.narm[-89,]
```

In [21]:
```
1 nrow(crime.narm)
```

90

In the *prbarr* & *prbconv* variables, there are instances which are larger than 1.

In [22]:
```
paste("Summary of prbarr variable")
summary(df$prbarr)
paste("Summary of prbconv variable")
summary(df$prbconv)
```

'Summary of prbarr variable'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.09277 0.20568 0.27095 0.29492 0.34438 1.09091       6
```

'Summary of prbconv variable'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.06838 0.34541 0.45283 0.55128 0.58886 2.12121       6
```

The $prbconv$ has 10 cases with $p > 1$

In [23]:
```
lenprbconv <- length(df[which(df$prbconv > 1),]$prbconv)
paste("Number of prbconv > 1: ", lenprbconv)
```

'Number of prbconv > 1: 10'

The $prbarr$ has 1 case with $p > 1$

In [24]:
```
lenprbarr <- length(df[which(df$prbarr > 1),]$prbarr)
paste("Number of prbarr > 1: ", lenprbarr)
```

'Number of prbarr > 1: 1'

good

Justification for the inclusion of *prbarr* and *prconv* values > 1

As we can see, the two variables described by the study as *prbarr* and *prbconv* have values > 1. The first instinct may be to discard the values > 1, since the variable is supposedly a probability and probability *p* can only take valid values $0 \leq p \leq 1$. However, the research by C. Cornwell and W. Trumball (1994)study defines the variables as

- The probability of arrest (P_a or *prbarr* in the dataset) is the ratio of arrests to offenses
- The probability of conviction (P_c or *prbconv* in the dataset) is the ratio of convictions to arrests

In this case of *prbconv*, there are few reasons why the ratio of convictions to arrests may be greater than 1. The police could make a single arrest of multiple people, which could result in multiple convictions. Or a single person could be arrested but convicted for multiple crimes. Therefore, we do not consider *prbconv* to strictly indicate probability. Moreover, instead of truncating entries > 1 to 1 or discarding the values completely, we will leave the entries in the dataset used for the analysis. We have used the same approach for *prbarr* and again did not discard values > 1.

More formally, in probability theory, all variables should follow $f(x) \in [0, 1] \text{ for all } x \in \Omega$. However, the $prbarr$ and $prbconv$ indicated the ratio of arrests to offenses and the ratio to conviction, respectively. These two variables do not satisfy the axioms of probability theory. However, the $prbarr > 1$ and $prbconv > 1$ were kept for analysis because of the reasons described above.

The dimensions of the final dataframe

In [25]:
```
dim(crime.narm)
```

90 24

Transformations

As noted above, all continuous variables, except *avgsen* and the identifier *county*, were log-transformed because some of these variables are (i) expressed on different scales, from 0.002 to 2177.1 (**table 1**), (ii) not all linearly related (**figures A3-A6**), (iii) skewed (except: *prbpris, wtuc, wfir, and wloc*) (**figure A1 & A2**), and (iv) interpretation will be aided by log-log model specification. Categorical variables, such as *west*, *central*, etc...were not log-transformed. Moreover, *avgsen* was not log-transformed to aid interpretability. Here, we noted most variables are not normally distributed (**Figure A1**). Distributions of variables after log-transformation resolved most instances of non-symmetric distributions (**Figure A2**). Moreover, we verified that none of the transformed values were 0 or negative, therefore variables could transform without any undesirable modifications such as adding a constant value.

skewness and asymmetry are not necessarily problems.

Relationships among variables

Log-transform variables for correlation matrix (this will only be used for the matrix below)

In [26]:
```
log_df = log(crime.narm[,c(2:9,13:24)])
```

Bind log-transformed and untransformed data

In [27]:
```
cd = cbind(log_df,
           crime.narm$county,
           crime.narm$west,
           crime.narm$central,
           crime.narm$urban)
```

Correlation matrix

In [28]:
```
cm = rcorr(as.matrix(cd),type=c("spearman"))
```

Figure 1a: Correlation matrix. Blue and red circles represent positive and negative Spearman' correlations, respectively. Spearman was used instead of Pearson to accommodate categorical variables.

Note:

- There is no perfect multicollinearity (**figures A3-A6**) among the explanatory variables - although *pctmin80* vs *west* is close. discuss why
- *ln(crmrte)* is correlated, when $\alpha = 0.05$, with all but 7 candidate explanatory variables. Note the *strong* correlations with several variables that politicians can actually influence, for example

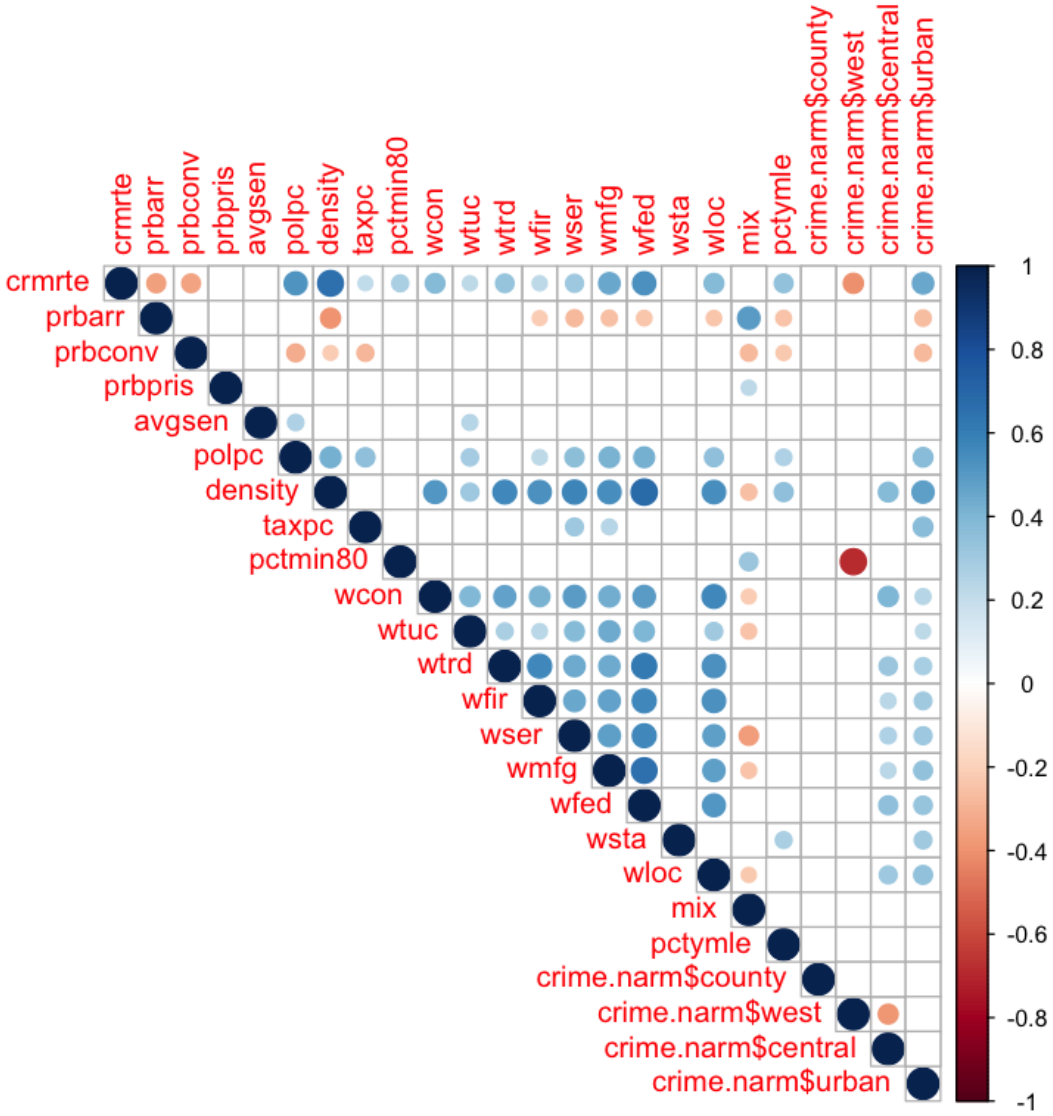
Unclear

now? Add 2 sentences with more details

polpc. Although *wfed* is also *strongly* correlated with crime rates, it might be challenging for politicians to changes federal employee salaries without substantial resistance.

- See the plot below for the relationship between *crmrte* and the categorical variables.
- Finally, more detailed scatterplot matrices are presented in the appendix (**figures A3-A6**).

In [29]:
```
# Plot options
backup_options <- options()
options(backup_options)
# Continuous data
corrplot(cm$r,
         type = "upper",
        #order='hclust',
         p.mat = cm$P,
         sig.level = 0.05,
         insig = "blank")
```



good **Figure 1b. Relationships between crime and continuous variables that are highly correlated with crime rates: police per capita and density.**

Note that crime rates are comparatively high in urban (red) areas and low in western (black). Therefore, *urban* might be an important covariate in subsequent models. There are clear and steep relationships between *ln(crmrte)* & *ln(density)* and *ln(crmrte)* & *ln(polpc)*. There are two outliers, *sensu lato*, from rows 51 and 79. The former is an outlier if *ln(polpc)* is regessed on *ln(crmrte)*- it may *pull* the slope towards 0. The latter observation is likely influential (see Cook's distance discussion below) if *ln(density)* is regressed on *ln(crmrte)*. Finally, when the outliers are removed the slopes appear more linear than when the outliers were included. These outliers are inspected in detail in the appendix. – state your conclusion

In [30]:

```
# Plot conditions
par(mfrow = c(2,3))
options(repr.plot.height = 15,
        repr.plot.width = 20,
        repr.plot.pointsize = 20)

# Crime rate vs polpc | urban
plot(log(crime.narm$crmrte)~log(crime.narm$polpc),
    cex=2,pch=19,col=crime.narm$urban,
    cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
    main= "Crime rate vs police per capita in \
urban(red) and non-urban(black) areas\n",
    ylab="log(crmrte)",xlab="log(polpc)")
text(log(crime.narm$crmrte)~log(crime.narm$polpc),
     labels = ifelse(log(crime.narm$polpc) > -5,
     row.names(crime.narm), NA),
     pos=3,cex=2)

# Crime rate vs polpc | west
plot(log(crime.narm$crmrte)~log(crime.narm$polpc),
    cex=2,pch=19,col=crime.narm$west,
    cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
    main= "Crime rate vs police per capita in \
western(red) and non-western(black) areas\n",
    ylab="log(crmrte)",xlab="log(polpc)")
text(log(crime.narm$crmrte)~log(crime.narm$polpc),
     labels = ifelse(log(crime.narm$polpc) > -5,
     row.names(crime.narm), NA),
     pos=3,cex=2)

# Crime rate vs polpc | no outlier
plot(log(crime.narm[-51,2])~log(crime.narm[-51,7]),
    cex=2,pch=19,col="black",
    cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
    main= "Crime rate vs police per capita\n with outlier(51) removed
    ylab="log(crmrte)",xlab="log(polpc)")

# Crime rate vs density |urban
plot(log(crime.narm$crmrte)~log(crime.narm$density),
    cex=2,pch=19,col=crime.narm$urban,
    cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
    main= "Crime rate vs population density in\
urban (red) & non-urban (black) areas\n",
    ylab="log(crmrte)",xlab="log(density)")
text(log(crime.narm$crmrte)~log(crime.narm$density),
    labels = ifelse(log(crime.narm$density) < -8 |
                    log(crime.narm$crmrte) < -5,
    row.names(crime.narm), NA),
    pos=3,cex=2)

# Crime rate vs density | west
plot(log(crime.narm$crmrte)~log(crime.narm$density),
    cex=2,pch=19,col=crime.narm$west,
    cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
    main= "Crime rate vs population density in\
western (red) & non-western (black) areas\n",
```

```
    ylab="log(crmrte)",xlab="log(density)")
text(log(crime.narm$crmrte)~log(crime.narm$density),
    labels = ifelse(log(crime.narm$density) < -8 |
                    log(crime.narm$crmrte) < -5,
    row.names(crime.narm), NA),
    pos=3,cex=2)

# Crime rate vs density | no outlier
plot(log(crime.narm[-79,2])~log(crime.narm[-79,8]),
    cex=2,pch=19,col="black",
    cex.main=2.5,cex.axis=2.5,cex.lab=2.5,
    main= "Crime rate vs population density\n with outlier(79) remove
    ylab="log(crmrte)",xlab="log(density)")


```

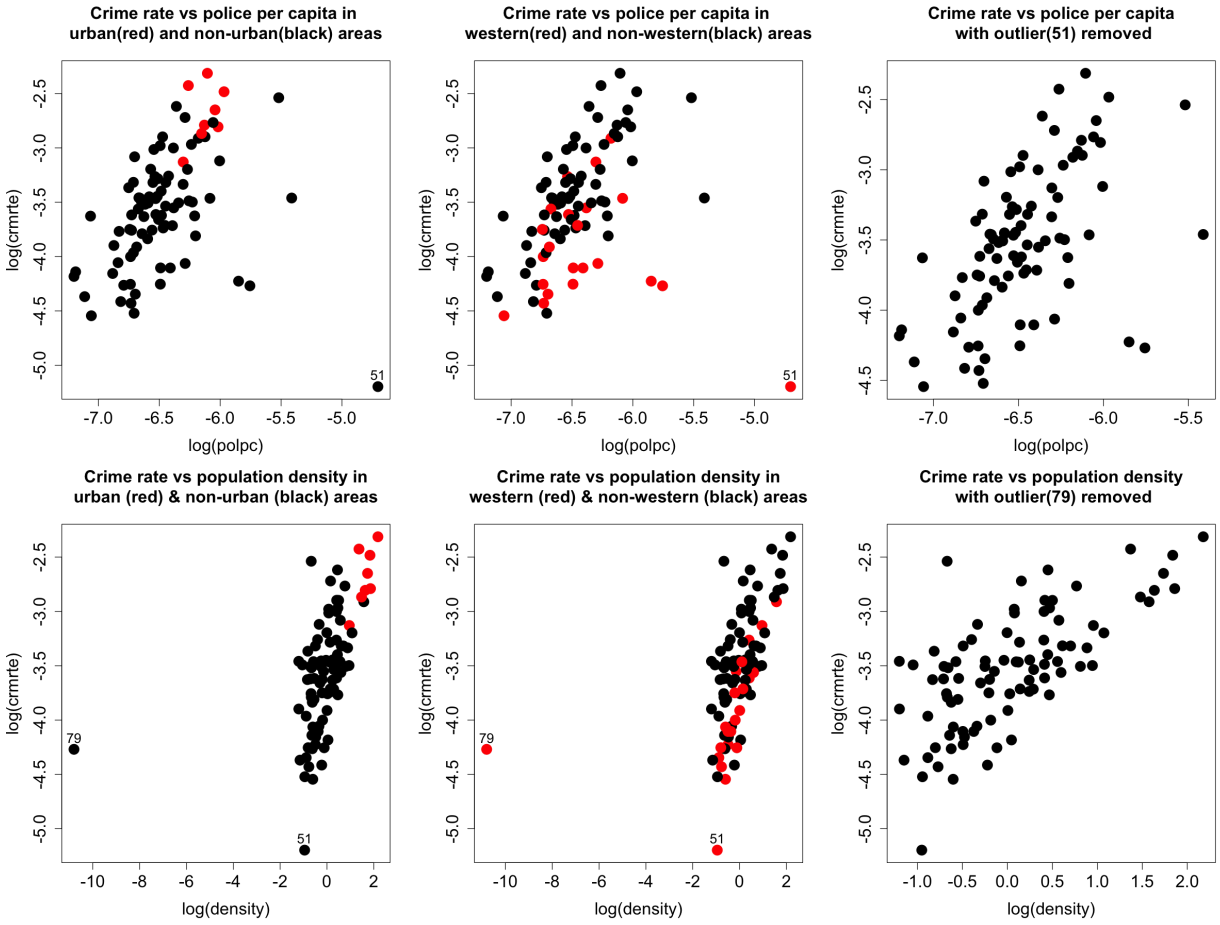
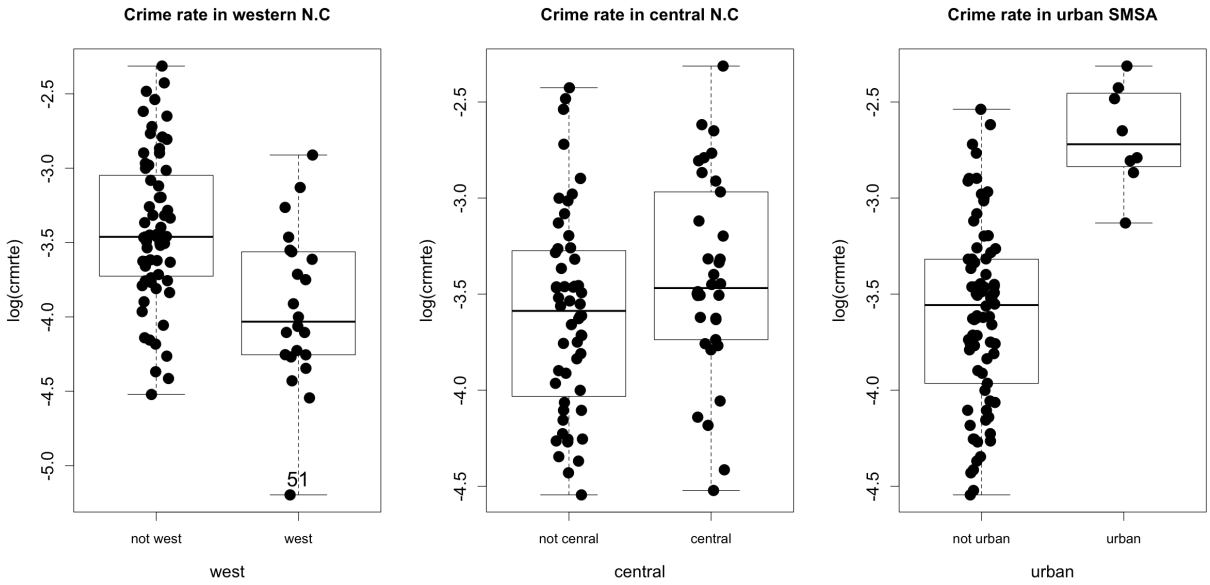


Figure 1c: Relationship between crime and categorical variables.

Note that crime rates differ between western & not-western and urban & not-urban regions in NC. Also note the unbalanced sampling sizes for urban vs not-urban regions. For central NC, crime rates look comparable within regions. Finally, note that the crime rate from county 51 is unexpectedly low.

In [31]:

```
# Plot conditions
par(mfrow = c(1,3))
options(repr.plot.height = 10,
        repr.plot.width = 20,
        repr.plot.pointsize = 25)

# Crime in western NC
boxplot(log(crime.narm$crmrte) ~ as.numeric(crime.narm$west),
        cex=1.5, pch=19, xaxt='n', cex.main=2.5, cex.axis=2.5, cex.la
        main= "Crime rate in western N.C",
        ylab="log(crmrte)",xlab="west")
stripchart(log(crime.narm$crmrte) ~ as.numeric(crime.narm$west),
           vertical = TRUE, method="jitter", add=TRUE,
           cex=2.5, pch=20, col='black')
text(log(crime.narm$crmrte)~as.numeric(crime.narm$west),
     labels = ifelse(log(crime.narm$crmrte) < -5,
                     row.names(crime.narm), NA),
     pos=3, cex=3)
axis(1, at=c("1", "2"), labels=c("not west","west"),cex.axis=2)

# Crime in central NC
boxplot(log(crime.narm$crmrte) ~ as.numeric(crime.narm$central),
        cex= 1.5, pch=19,xaxt = 'n',cex.main=2.5,cex.axis=2.5,cex.lab
        main= "Crime rate in central N.C",
        ylab="log(crmrte)",xlab="central")
stripchart(log(crime.narm$crmrte) ~ as.numeric(crime.narm$central),
           vertical = TRUE, method="jitter", add=TRUE,
           cex=2.5, pch=20, col='black')
text(log(crime.narm$crmrte)~as.numeric(crime.narm$central),
    labels = ifelse(log(crime.narm$crmrte) < -5,
    row.names(crime.narm), NA),
    pos=3, cex=3)
axis(1, at=c("1", "2"), labels=c("not cenral","central"),cex.axis=2)

# Crime in urban NC
boxplot(log(crime.narm$crmrte) ~ as.numeric(crime.narm$urban),
        cex= 1.5, pch=19, xaxt='n', cex.main=2.5, cex.axis=2.5, cex.l
        main= "Crime rate in urban SMSA",
        ylab="log(crmrte)",xlab="urban")
stripchart(log(crime.narm$crmrte) ~ as.numeric(crime.narm$urban),
           vertical = TRUE, method="jitter", add=TRUE,
           cex=2.5, pch=20, col='black')
text(log(crime.narm$crmrte)~as.numeric(crime.narm$urban),
    labels = ifelse(log(crime.narm$crmrte) < -5,
    row.names(crime.narm), NA),
    pos=3, cex=3)
axis(1, at=c("1", "2"), labels=c("not urban","urban"),cex.axis=2)
```



ok **Summary**

From the exploratory data analysis above we learned that, despite several anomalous observations, all but 7 candidate explanatory variables are correlated with crime rate. Of those variables, we will use (i) the 3 variables that we think politicians can influence as key variables in our models, (ii) covariates of interest that might improve model fit and enable use to control for other sources of variability, and (ii) all remaining variables to assess the robustness of our models.

what about the outliers? Dropped or not?

Main Analysis

Back to Table of Contents

For inferential analysis, we would like to achieve 3 goals

1. Model 1 will only include the explanatory variables of key interest.
2. Model 2 will include the covariates to find to balance between parsimony and accuracy and control for other sources of variation.
3. Model 3 will include all explanatory variables and demonstrate the robustness of our model specification.

Model 1: Model with only key variables of interest (m_key)

$$\ln(crmrte) = \beta_0 + \beta_1 \ln(polpc) + \beta_2 \ln(prbarr) + \beta_3 \ln(prbconv) + \beta_4 ln(prbarr) \cdot ln(prbconv) + u$$

In [32]:
```
m_key <- lm(log(crmrte) ~ log(polpc) + log(prbarr) + log(prbconv)
                        + log(prbconv)*log(prbarr), data = crime.narm
```

Test the hypotheses that the slopes, β_j, are equal to zero for *polpc*, *prbarr*, and *prbconv*

1.

$$H_0^{(1)} : \beta_1 ln(polpc) = 0 \text{ vs. } H_A^{(1)} : \beta_1 ln(polpc) \neq 0$$

In [33]:
```
# Hypothesis - H_0: Test if slopes are equal to zero
H0_p = c("log(polpc)=0")
# Wald/F test
linearHypothesis(
    m_key, H0_p,
    vcov=vcovHC(m_key),
    singular.ok=TRUE
)
```

Res.Df	Df	F	Pr(>F)
86	NA	NA	NA
85	1	0.541528	0.4638261

> Fail to reject the $H_0^{(1)} : \beta_1 ln(polpc) = 0$. *polpc* should stay in the model. The implications of this result will be discussed below.

2.

$$H_0^{(2)} : \beta_2 ln(prbarr) = 0 \text{ vs. } H_A^{(2)} : \beta_2 ln(prbarr) \neq 0$$

In [34]:
```
# Hypothesis - H_0: Test if slopes are equal to zero
H0_prba = c("log(prbarr) = 0")
# Wald/F test
linearHypothesis(
    m_key, H0_prba,
    vcov=vcovHC(m_key),
    singular.ok=TRUE
)
```

Res.Df	Df	F	Pr(>F)
86	NA	NA	NA
85	1	0.6383648	0.4265305

> Fail to reject the $H_0^{(1)} : \beta_2 ln(prbarr) = 0$. *prbarr* should stay in the model. The implications of this result will be discussed below.

3.

$$H_0^{(3)} : \beta_3 ln(prbconv) = 0 \text{ vs. } H_A^{(3)} : \beta_3 ln(prconv) \neq 0$$

In [35]:
```
# Hypothesis - H_0: Test if slopes are equal to zero
H0_prbc = c("log(prbconv) = 0")
# Wald/F test
linearHypothesis(
    m_key, H0_prbc,
    vcov=vcovHC(m_key),
    singular.ok=TRUE
)
```

Res.Df	Df	F	Pr(>F)
86	NA	NA	NA
85	1	0.001011247	0.974706

> Fail to reject the $H_0^{(1)}: \beta_3 ln(prbconv) = 0$. *prbconv* should stay in the model. The implications of this result will be discussed below.

4.

$$H_0^{(4)}: \beta_4 ln(prbarr) \cdot ln(prbconv) = 0 \text{ vs. } H_A^{(4)}: \beta_4 ln(prbarr) \cdot ln(prbconv) \neq 0$$

In [36]:
```
# Hypothesis - H_0:
# Test whether interactions between prbarr and prbconv could be dropp
H0 = c("log(prbarr):log(prbconv) = 0")
# Wald/F test
linearHypothesis(
    m_key, H0,
    vcov=vcovHC(m_key),
    singular.ok=TRUE
)
```

Res.Df	Df	F	Pr(>F)
86	NA	NA	NA
85	1	0.2653271	0.6078207

> Fail to reject the null hypothesis H_0: $\beta_4 ln(prbarr) \cdot ln(prbconv) = 0$. Therefore, the interaction between *prbarr* and *prbconv* could not be dropped from the model.

Model 2: A Model with key variables plus covariates (m_asscociated)

$$\ln(crmrte) = \beta_0 + \beta_1 \ln(polpc) + \beta_2 \ln(prbarr) + \beta_3 \ln(prbconv) + \beta_4 ln(prbarr) \cdot ln(prbconv) + \beta_5 \ln(prbpris) + \beta_6 urban + \beta_7 \ln(taxpc) + \beta_8 \ln(wfed) + \beta_9 \ln(density) + \beta_{10} \ln(pctmin80) + u$$

In [37]:
```
m_associated <- lm(log(crmrte) ~ log(polpc) + log(prbarr) + log(prbco
                                 + log(prbarr)*log(prbconv)
                                 + log(prbpris)
                                 + log(taxpc) + log(wfed) + factor(urba
                                 + log(density) + log(pctmin80), data =
```

Model 3: A Model all variables (m_all)

$$\ln(crmrte) = \beta_0 + \beta_1 \ln(polpc) + \beta_2 \ln(prbarr) + \beta_3 \ln(prbconv) + \beta_4 \ln(prbpris) + \beta_5 urban + \beta_6 \ln(taxpc) + \beta_7 \ln(wfed) + \beta_8 \ln(density) + \beta_9 \ln(pctmin80) + \beta_{10} \ln(wcon) + \beta_{11} \ln(wtuc) + \beta_{12} \ln(wtrd) + \beta_{13} \ln(wfir) + \beta_{14} \ln(wser) + \beta_{15} \ln(wmfg) + \beta_{16} \ln(wsta) + \beta_{17} \ln(wloc) + \beta_{18} avgsen + \beta_{19} west + \beta_{20} \ln(pctymle) + \beta_{21} \ln(mix) + u$$

In [38]:
```
m_all <- lm(log(crmrte) ~ log(polpc) + log(prbarr) + log(prbconv)
                          + log(prbpris) + avgsen + log(taxpc
                          + factor(west) + factor(urban)
                          + log(wfed) + log(wcon) + log(wtuc)
                          + log(wtrd) + log(wfir) + log(wser)
                          + log(wmfg) + log(wsta) + log(wloc)
                          + log(density) + log(pctmin80) + lo
                          + log(mix), data = crime.narm)
```

Robust standard errors

Covariance matrix was applied for estimating a robust covariance matrix of variables. These are used because, as documented below, we violate the assumption of homoscedasticity.

In [39]:
```
# Compute robust standard errors
se.m_key = sqrt(diag(vcovHC(m_key)))
se.m_associated = sqrt(diag(vcovHC(m_associated)))
se.m_all = sqrt(diag(vcovHC(m_all)))
```

Akaike Information Criteria (AIC)

This is used, in addition to the adjusted-R^2, to assess model fit.

In [40]:
```
# Compute AIC values
m_key$AIC <- AIC(m_key)
m_associated$AIC <- AIC(m_associated)
m_all$AIC <- AIC(m_all)
```

how do you calculate this? Standard OLS does not have a likelihood function.

Table 3: Regression table. The coefficients, robust standard errors, p-values, adjusted-R^2 and AIC values are presented and discussed below.

Coefficients and standard errors are interpreted below.

In [41]:
```
# We pass the standard errors into stargazer through the se argument.
stargazer(m_key, m_associated, m_all,
          type = "text",
          keep.stat=c("n", "adj.rsq", "aic"),
          se = list(se.m_key, se.m_associated, se.m_all),
          column.labels = c("key", "associated", "all"),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
=================================================================
                                Dependent variable:
                            ----------------------------
                                       log(crmrte)
                              key      associated     all
                              (1)         (2)         (3)
-----------------------------------------------------------------
log(polpc)                   0.296       0.317      0.507*
                            (0.403)     (0.307)    (0.235)

log(prbarr)                  -0.542      -0.399    -0.512***
                            (0.678)     (0.391)    (0.136)

log(prbconv)                 -0.032      -0.153     -0.293*
                            (1.021)     (0.576)    (0.137)

log(prbpris)                             -0.192     -0.315
                                        (0.231)    (0.217)

avgsen                                              -0.035
                                                   (0.019)

log(taxpc)                               0.048      0.049
                                        (0.286)    (0.287)

factor(west)1                                       0.125
                                                   (0.121)

log(wfed)                                0.679      0.555
                                        (0.566)    (0.549)

log(wcon)                                           0.199
                                                   (0.224)

log(wtuc)                                           0.102
                                                   (0.359)

log(wtrd)                                           0.362
                                                   (0.348)

log(wfir)                                          -0.365
                                                   (0.417)

log(wser)                                          -0.337*
                                                   (0.167)

log(wmfg)                                           0.031
```

			(0.225)
log(wsta)			0.007
			(0.371)
log(wloc)			0.047
			(0.694)
factor(urban)1		0.011	0.058
		(0.250)	(0.240)
log(density)		0.098	0.138
		(0.179)	(0.151)
log(pctmin80)		0.247***	0.273***
		(0.052)	(0.060)
log(prbarr):log(prbconv)	0.326	0.215	
	(0.633)	(0.355)	
log(pctymle)			0.156
			(0.261)
log(mix)			0.051
			(0.106)
Constant	-2.665	-7.525	-5.155
	(3.535)	(5.425)	(4.746)
Observations	90	90	90
Adjusted R2	0.474	0.772	0.790
Akaike Inf. Crit.	96.385	26.409	27.767
Note:	*p<0.05; **p<0.01; ***p<0.001		

so you didn't drop the outliers?

don't discuss practical significance of insignificant variables as if they were statistically signif.

Notes about model fit are in the appendix

← just focus on the key variables and 1-2 sentences on anything surprising.

Interpretation of coefficients of Model 1 (m_key model):

use proper word instead

- $\beta_1 ln(polpc)$: For each 1% increase in *polpc*, crime rate (*crmrte*) increases by 0.296%. This coefficient is statistically insignificant and the standard error is large relative to the estimate. The p-value associated with H_0: $\beta_1 = 0$ is $p > 0.05$- the slope is not different from 0. From a practical perspective, this may be significant - any increase in crime rates is significant from the victims perspective. Moreover, since the direction of causality is unknown, it seems likely that we could recommend more police per captia in areas with relatively high crime rates and reduce, not increase, crime rates. In other words, areas with more crime might need more police to maintain the safety.

unclear. You suggest disregarding the data?

- $\beta_2 ln(prbarr)$ and $\beta_3 ln(prbconv)$: For each 1% increase in *prbarr*, *crmrte* decreases by 0.542%. For each 1% increase in *prbconv*, *crmrte* decreases by 0.032%. *prbarr* and *prbconv* failed to satisfy the axioms of probability theory as described above ("Justification for the inclusion of prbarr and prconv values > 1"). These coefficients are not statistically significant. The p-values associated with H_0: $\beta_2 = 0$ and β_3 is $p > 0.05$- the slopes are not different from 0. From a practical perspective, these coefficient represent a small but perhaps important

decrease in crime rates/person. Perhaps people are less likely to commit a crime if they are more likely to be arrested or more afraid of being arrested. In addition, people may be scared to be convicted after they get arrested. Perhaps stauncher policing, arrest and conviction criteria may reduce crime slowly.

- $\beta_4 ln(prbarr) \cdot ln(prbconv)$: There is not an interaction between *prbarr* and *prbconv*. The p-value associated with H_0: $\beta_4 = 0$ is $p > 0.05$. In other words, the effect of *ln(prbarr)* on *ln(crmrte)* does not depend on the values of *ln(prbarr)*. From a practical perspective this means that politicians need not consider the interaction of *prbarr* and *prbconv* when implementing policies to reduce crime- if it sufficient to *prbarr* and *prbconv* separately.
- Finally, the intercept, β_0, is -2.665 and not significant. The logarithm of *crmrte* is -2.665 when *polpc*, *prbarr*, *prbconv*, and the interaction between *prbarr* and *prbconv* are 1 and u is 0. In other words, when transformed to the original units, $e^{-2.665} = 0.069$, crime rate per person is low, 0.069, even when there is one police officer per person, and *prbarr* and *prbconv* are high, at 1.

Interpretation of coefficients of Model 2 (m_associated model):

- $\beta_1 ln(polpc)$, $\beta_2 ln(prbarr)$, $\beta_3 ln(prbconv)$ The effects and directions of these key explanatory varibles and the interaction between *prbarr* and *prbconv* are similar to m_key model as described above. For each 1% increase in *polpc*, *crmrte* increases by 0.317%. For each 1% increase in *prbarr*, *crmrte* decreases by 0.399%. For each 1% increase in *prbconv*, the *crmrte* decreases by 0.153%. These coefficients of these key explanatory varibles are not statistically and practically significant. The p-value associated with H_0: $\beta_1 = 0$, β_2 or β_3 is $p > 0.05$- the slopes are not different from 0.
- $\beta_4 ln(prbarr) \cdot ln(prbconv)$: As above, there is not an interaction between *prbarr* and *prbconv*. The p-value associated with H_0: $\beta_4 = 0$ is $p > 0.05$. In other words, the effect of *prbarr* on *crmrte* does not depend on the values of *prbarr*. Finally, the standard error for this interaction is less than for m_key. From a practical perspective this means that politicians need not consider the interaction of *prbarr* and *prbconv* when implementing policies to reduce crime- if it sufficient to *prbarr* and *prbconv* separately.
- $\beta_5 ln(prbpris)$: For every 1% increase in *prbpris*, there is 0.192% decrease in *crmrte*. While this coefficient may of limited practical significance in that prospect of prison time may lead to few crimes being the committed, in this model, this coefficient is statistically insignificant ($p > 0.05$).
- $\beta_6 urban$: Controlling for all other variables, on average, the crime rate in urban areas is 1.1% more compared to non urban area. This coefficient of this covariate is neither statistically nor practically significant. The p-value associated with H_0: $\beta_6 = 0$ is $p > 0.05$. From a practical perspective, this coefficient may imply that the urban areas may have higher crime rates, although n is small and the effect is weak.
- $\beta_7 (taxpc)$: For every 1% increase in *taxpc*, *crmrte* increases by 0.048%. This coefficient of this covariate is neither statistically nor practically significant. The p-value associated with H_0: $\beta_7 = 0$ is $p > 0.05$- the slope is not different from 0. From a practical perspective, the county with high tax revenue may imply more rich people and/or bussiness activities which may, in turn, be associated with increased crime rates. However, the effect is week because the coefficient is small and statistically insignificant.
- $\beta_8 ln(wfed)$: For every 1% increase in *wfed*, *crmrte* increases 0.679%. This coefficient of this covariate is neither statistically nor practically significant. The p-value associated with H_0: $\beta_8 = 0$ is $p > 0.05$- the slope is not different from 0. When converted back to the original

units, this coefficient represents a 1.97 unit increase in crime rates (crimes committed/person) for every unit increase in weekly wages for federal employees.

- $\beta_9 ln(density)$: For every 1% increase *density*, *crmrte* increases by 0.098%. This coefficient of this covariate is neither statistically nor practically significant. The p-value associated with H_0: $\beta_9 = 0$ is $p > 0.05$- the slope is not different from 0. From a practical perspective, this coefficient suggests that counties with more people have slightly hire crime rates. If it was ethical, we could relocate people and or encourage people to move to more rural, sparsely populated areas.
- $\beta_{10} ln(pctmin80)$: For every 1% increase in *pctmin80*, *crmrte* increases by 0.247%. This coefficient of this covariate is statistically and practically significant. The p-value associated with H_0: $\beta_{10} = 0$ is $p < 0.001$- the slope is different from 0. From a practical perspective, this coefficient seems to represent an important driver of crime rates/person.
- Finally, the intercept, β_0, is -7.527 and not significant. The logarithm of *crmrte* is -7.527 when the covariates are 1 and u is 0. In other words, when transformed to the original units, $e^{-7.527} = 0.0005$, crime rate per person is low, 0.0005, even when covariates are fixed at 1.

or location of which to investigate further?
"Driver" is a causal word. Be careful with using that.

Interpretation of coefficients of Model 3 (m_all model):

- $\beta_1 ln(polpc)$: For every 1% in increase in *polpc*, *crmrte* increases by 0.507%. This is counter-intuitive but can be indicative of the fact that when crime increases, more police resources per capita may be deployed and not that more police per capita causes a higher crime rate. The null hypothesis that H_0: $\beta_1 = 0$ can be rejected, though it should be noted that the associated p value only satisfies ($p < 0.05$).

good

- $\beta_2 ln(prbarr)$: For every 1% increase in *prbarr*, we see that *crmrte* reduces by 0.512%. This coefficient is statically significant and thus reject the null hypothesis H_0: $\beta_2 = 0$ since the associated p value $p < 0.001$. This decrease is also practically significant since there is the potential to impact the crime in an area in an actionable and positive manner. One reason for this negative relation between the independent and explanatory variable may be that once a person is arrested, their opportunity to commit crime is reduced.

it's the other way around

ok

- $\beta_3 ln(prbconv)$: For every 1% increase in *prbarr*, there is 0.293% decrease in *ln(crmrte)*. This coefficient isstatistically significant and thus reject the null hypothesis H_0: $\beta_3 = 0$ since the associated p value $p < 0.05$. This impact of this coefficient may beof small but important practically significant in that perhaps people are less likely tocommit a crime if they face the prospect of conviction.

ok

- $\beta_4 ln(prbpris)$: For every 1% increase in *ln(prbpris)*, there is 0.315% decrease in *ln(crmrte)*. While this coefficient may of limited practical significance in that prospect of prison time may lead to few crimes being the committed, in this model, this coefficient is statistically insignificant ($p > 0.05$).
- $\beta_5 urban$: Controlling for all other variables, on average, *crmrte* in urban areas is 5.8% more compared to non urban area. This coefficient of this covariate is statistically insignificant. The p-value associated with H_0: $\beta_5 = 0$ is $p > 0.05$. From a practical perspective, this coefficient may imply that the urban areas may have higher crime rates.
- $\beta_6 ln(taxpc)$: For every 1% increase in *ln(taxpc)*, the *ln(crmrte)* increases by 0.049%. This coefficient of this covariate is neither statistically nor practically significant. The p-value associated with H_0: $\beta_6 = 0$ is $p > 0.05$- the slope is not different from 0. From a practical perspective, the county with high tax revenue may imply more rich people and/or business activities; furthermore, potentially leading the increase of crime rate. However, the effect is week because the coefficient is small and statistically insignificant.

- Wage coefficients: $\beta_7 ln(wfed)$, $\beta_10 ln(wcon)$, $\beta_{11} ln(wtuc)$, $\beta_{12} ln(wtrd)$, $\beta_{15} ln(wmfg)$, $\beta_{17} ln(wloc)$, $\beta_{13} ln(wfir)$, $\beta_{14} ln(wser)$,*and* $\beta_{16} ln(wsta)$: We see here that when all types of wages are included, any 10% increase in $wfed$, $wcon$, $wtuc$,$wtrd$,$wmfg$, $wloc$, or $wsta$ is associated with no more than a 6% increase in *ln(crmrte)*, with the impact of federal wage being the highest at 5.6% and manufacturing wages having the lowestimpact at 0.07% increase in crime rate. In many cases, the impact to crime rate is practically neglible, e.g in the case of state employees' wages. Any 10% increase in $wfir$ is associated with a 3.65% decrease in crime rate but again the impact is practically insignificant. On the other hand, Any 10% increase in $wser$ is associated with a 3.37% decrease in crime rate and the impact is statistically ($p < 0.05$) and practically significant. Moreover, all the coefficients, except *wser*, related to wage are statistically insignificant (each associated p , $p > 0.05$) and we fail to reject the null hypothesis for each of these variables H_0: $\beta_j = 0$for j = $wfed$, $wcon$, $wtuc$, $wtrd$, $wmfg$, $wloc$, $wfir$, $wser$ or $wsta$.

or reflects value of crime/reporting bias?

- We recognize that some of the interpretations for wage variables are counter intuitive e.g. higher wages in some categories are associated with higher crime ratewhereas higher wages in other categories are associated with lower crime rates. As mentioned previously, the wage variables are most likely top and bottom coded andthe extreme values have been censored. Combined with the various sources of omitted variable bias discussed in later sections, we believe that conclusions based on wages may be inaccurate.
- $\beta_8 ln(density)$: For every 1% in increase *ln(density)*, *ln(crmrte)* increases by 0.138% which is practically insignificant. This is also statistically insignificant ($p > 0.05$) and we fail to reject the null hypothesis that the slope is 0 i.e. H_0: $\beta_{density} = 0$.
- $\beta_9 ln(pctmin80)$: For every 1% increase in ln(pctmin80)*, the *ln(crmrte)* increases by 0.273%. This coefficient of this covariate is statistically and practically significant. The p-value associated with H_0: $\beta_9 = 0$ is $p < 0.001$- the slope is different from 0. From a practical perspective, this coefficient seems to represent an important driver of crime rates/person.
- $\beta_{18} avgsen$: For 1 day increase in *avgsen*, there appears to be a 3.5% decrease in *ln(crmrte)*. While this coefficient may be of practical significance in that perhaps longer the a prison sentences provide less opportunities to commit crime or even that the experience of prison deters recidivism, in this model $\beta_{18} avgsen$ coefficient is statistically insignificant ($p > 0.05$) and we fail to reject the null hypothesis that the slope is 0 i.e. H_0: $\beta_{avgsen} = 0$.
- $\beta_{19} ln(west)$: Controlling for all other variables, on average, the *crmrte* in the Western NC areas is 12.5% more compared to areas not categorized as west. This coefficient of this covariate is statistically insignificant. The p-value associated with H_0: $\beta_{19} = 0$ is $p > 0.05$. From a practical perspective, this coefficient may imply that that western NC has a much higher crime rate than those areas not categorized as western NC, although none of the EDA supports this conclusion.
- $\beta_{20} ln(pctmle)$: For every 1% increase in ln(pctmle)*, *ln(crmrte)*increases by 0.156%. This coefficient of this covariate is statistically and practically insignificant. The p-value associated with H_0: $\beta_{20} = 0$ is $p > 0.05$- the slope is different from 0.
- $\beta_{21} ln(mix)$: For every 1% increase in *ln(mix)*, *ln(mix)* increases by 0.051%. This coefficient of this covariate is both statistically and practically insignificant. The p-value associated with H_0: $\beta_{21} = 0$ is $p > 0.05$- the slope is different from 0.
- Finally, the intercept, β_0, is -5.5155 and not significant. The logarithm of *crmrte* is -5.5155 when the covariates are 1 and u is 0. In other words, when transformed to the original units, $e^{-5.5155} = 0.004$, crime rate per person is low, 0.004, even when covariates are fixed at 1.

Diagnostics

Back to Table of Contents

Figure 2a: Diagnostic plots for model with only key explanatory variables (m_key).

Note that these plots will be discussed below.

In [42]:
```
options(repr.plot.height = 3, repr.plot.width = 8, repr.plot.pointsiz
par(mfrow = c(1,4))
plot(m_key, which=1, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=2)
plot(m_key, which=3, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=2)
plot(m_key, which=4, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=1)
```

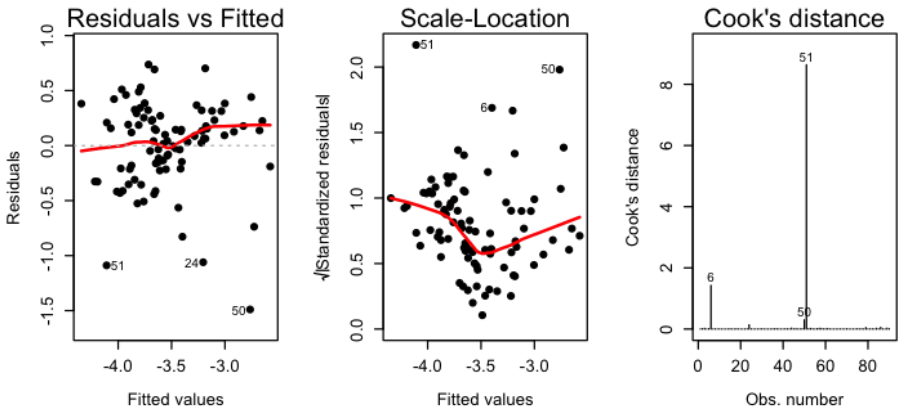


Figure 2b: Diagnostic plots for model with key variables + covariates of interest (m_associated).

Note that these plots will be discussed below.

In [43]:
```
options(repr.plot.height = 3, repr.plot.width = 8, repr.plot.pointsiz
par(mfrow = c(1,4))
plot(m_associated, which=1, pch=19, cex=1, cex.lab=1, cex.main=4, lwd
plot(m_associated, which=3, pch=19, cex=1, cex.lab=1, cex.main=4, lwd
plot(m_associated, which=4, pch=19, cex=1, cex.lab=1, cex.main=4, lwd
```

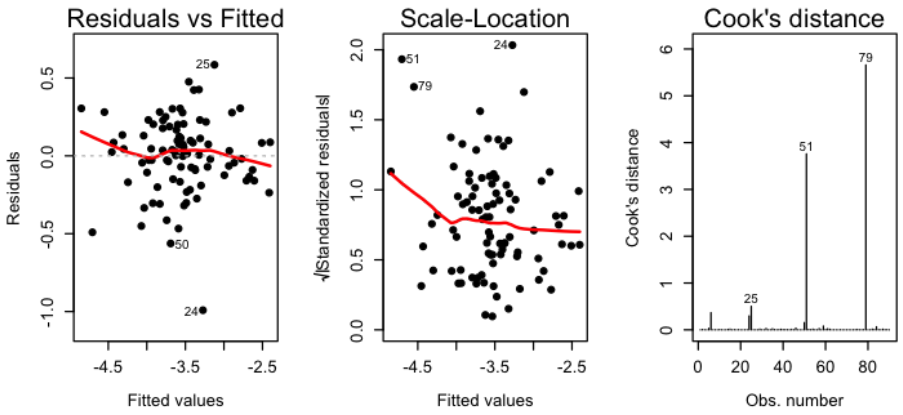
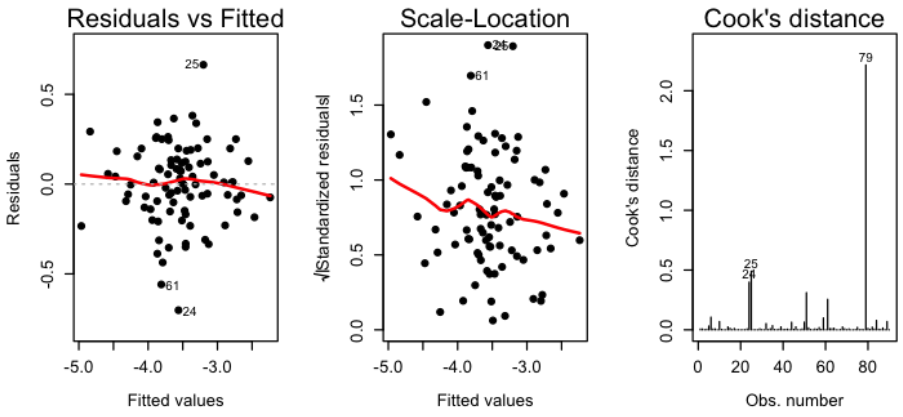


Figure 2c: Diagnostic plots for model with all explanatory variables (m_all).

Note that these plots will be discussed below.

In [44]:
```
options(repr.plot.height = 3, repr.plot.width = 8, repr.plot.pointsiz
par(mfrow = c(1,4))
plot(m_all, which=1, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=2)
plot(m_all, which=3, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=2)
plot(m_all, which=4, pch=19, cex=1, cex.lab=1, cex.main=4, lwd=1)
```



good

Summary of diagnostic plots and assumptions for all three models: m_all, m_associated, and m_key

1. **Linear in parameters**
   - As seen in our models above, all of our models are linear and additive in the parameters, the $\beta_0, \beta_1, \ldots \beta_k$
2. **Random sampling**
   - We assume that are n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$ are independent and identically distributed, iid. It is possible that this assumption is violated but it is difficult to know with certainly from the information provided.

ok. which means what for your analysis?

3. **No perfect collinearity**
   - As documented in the scatterplot matrices below (**Figs A3-A6**) and those in the appendix, there is no *perfect* collinearity. Although, there are certainly explanatory variables that are correlated, these variables are not *perfectly* correlated. ∴ the explanatory variables are not redundant.
4. **Zero-conditional mean**
   - The zero-conditional mean assumption, $E(u|x_1, x_2, \ldots, x_k) = 0$ is not satisfied, *sensu stricto*, for any of the models presented above. The residual versus fitted values plots (**Figs 2a-c**) document the extent to which these models violate this assumption. If the zero mean assumption were satisfied then we should expect to see a horizontal red line, centered at zero (ASYNC lecture 12.5). In contrast we see that our estimates of u, the residuals, deviate from this expectation, especially for m_key and m_associated.
5. **Homoskedasticity**

- Homoskedasticity does not appear to be satisfied for all models presented above. Evidence for heteroskedasticity is provided by plots and Breusch-pagan tests, below. When **figures 2a-c** are inspected, we see minor to acute violations of this assumption in the residual versus fitted value plots (**Figs 2a-c**) and the scale-location plots. If variances were homoskedastic we should see uniform horizontal bands/scatters of points across these plots. However, note that we see non-constant variability in the residuals and standardized residuals across the range of fitted values. Likewise, Breusch-pagan tests for each model rejected the H_0: residuals are homoscedastic. Since we did not satisfy this assumption, robust standard errors were used.

Note about outliers and their influence

- From **figures 2a-c** above we can see values with large residuals, large standardized residuals, and two observations, 51 and 79, with large Cook's distance (as noted in ASYNC lecture 11.10) that is greater than 1. Given that we have no evidence that these observations were recorded incorrectly, we will keep these observations in the dataframe, despite their influence on the regression surface.

ok

Breusch Pagan test for homoscedasticity

> H_0: residuals are homoscedastic

Squared residuals are regressed on explanatory variables

In [45]:
```
# m_key
bp_key = lm(m_key$residuals**2 ~ log(polpc) + log(prbarr) + log(prbco
                        + log(prbconv)*log(prbarr), data = crime.narm
# m_associated
bp_assoc = lm(m_key$residuals**2 ~ log(polpc) + log(prbarr) + log(prb
                                + log(prbarr)*log(prbconv) +
                                + log(prbpris) + avgsen +
                                + log(taxpc) + log(wfed) +
                                + factor(west) + factor(urban) + log(d
# m_all
bp_all = lm(m_key$residuals**2 ~ log(polpc) + log(prbarr) + log(prbco
                                  + log(prbpris) + avgsen + log(taxpc
                                  + factor(west) + factor(urban) + fa
                                  + log(wfed) + log(wcon) + log(wtuc)
                                  + log(wtrd) + log(wfir) + log(wser)
                                  + log(wmfg) + log(wsta) + log(wloc)
                                  + log(density) + log(pctmin80) + lo
                                  + log(mix), data = crime.narm)
```

Test statistics

In [46]:
```
# m_key
bpts_k = nobs(bp_key) * summary(bp_key)$r.squared
paste("Breusch-Pagan test-statistic:", bpts_k)
# m_associated
bpts_as = nobs(bp_assoc) * summary(bp_assoc)$r.squared
paste("Breusch-Pagan test-statistic:", bpts_as)
# m_all
bpts_al = nobs(bp_all) * summary(bp_all)$r.squared
paste("Breusch-Pagan test-statistic:", bpts_al)
```

'Breusch-Pagan test-statistic: 11.0058263780772'

'Breusch-Pagan test-statistic: 24.2752619331665'

'Breusch-Pagan test-statistic: 37.4205416315331'

df

In [47]:
```
# m_key
bpdf_k = length(bp_key$coefficients) - 1
# m_associated
bpdf_as = length(bp_assoc$coefficients) - 1
# m_all
bpdf_al = length(bp_all$coefficients) - 1
```

p-value

In [48]:
```
# m_key
paste("p-value:",
      1-pchisq(bpts_k, bpdf_k))
# m_associated
paste("p-value:",
      1-pchisq(bpts_as, bpdf_as))
# m_all
paste("p-value:",
      1-pchisq(bpts_al, bpdf_al))

```

'p-value: 0.0264986118703746'

'p-value: 0.0116213319034999'

'p-value: 0.0150557449473256'

good

As noted above, since the assumption of homoscedasticity was not satisfied, robust standard errors were used (as recommended in ASYNC, unit 12).

Robustness

Back to Table of Contents

ok.

Omitted Variable Bias:

focus on key variables only.

We recognize that crimes is a complex issue, depending on a wide variety of factors. As we can see, in even the most aggresive model (m_all), the Adjusted $R^2 = 0.790$ indicates that nearly 80% of the variation in crime per person can be explained by the explanatory variables we have chosen for the model. Even in this all-inclusive model, 20% of the variation in crime rate remains unexplained due to variables we either cannot measure or have access to. Here we will discuss a few such possible sources of omitted variable bias,

Poverty

In and of itself, higher poverty may not necessarily cause an increase in crime rates. However, poverty and the lack of steady income, especially at levels needed to cover basic living needs may drive individuals towards crime to make ends meet or acquire material items they may not be able to access otherwise. While the dataset provides information on wage, it does not provide a benchmark of the poverty line, which would allow us to compare how the poverty varies from across counties. Moreover, impoverished communities tend to have less access to quality education and thus better career opportunities which can otherwise lead to more financial stability. Thus, we ~~estimate~~ conjecture/guess/hypothesize that

- Poverty has a positive effect on crime and the coefficient $\beta_{poverty} > 0$
- Poverty is negatively correlated to wages i.e. higher poverty leads to lower wages. Given that $\beta_{poverty} > 0$, the wage coefficients are negatively biased.
- Poverty is negatively correlated to taxes i.e. people in poverty will pay less in taxes. Given that $\beta_{poverty} > 0$, β_{taxpc} is negatively biased.
- Since the dataset shows that most NC counties are categorized as non-urban, poverty may be considered to be positively correlated to population density. The reason why context is important is because we may have dense urban areas that are not generally fiscally constrained e.g. San Francisco or Manhattan. Given that $\beta_{poverty} > 0$, $\beta_{density}$ is positively biased. which means..?

Education

Education can have impact on crime rates. Better education may lead to better and steadier employment which can in turn impact income. By measuring the impact of education on crime rates, policies can be implemented to ensure proper funding and resources are given to improve education in disadvantaged communities. We should point out that in certain cases, highly educated people will still commit common crimes. Thus, we estimate that

- Education has a negative effect on crime and the coefficient $\beta_{education} < 0$
- Education is postively correlated to wages i.e. wages may increase with better education. Given that $\beta_{education} < 0$, the wage coefficients are negatively biased due to lack of information regarding education.
- Education is postively correlated to taxes since we assume that better education leads to better wages and in turn results in more taxes being paid. Given that $\beta_{education} < 0$, β_{taxpc} is negatively biased.

Employment

Another socioeconomic factor (along with of poverty and education), employment status can impact crime rates. We believe that steady employment and subsequent income generated can help reduce the tendency for a person to commit crime. Thus, we estimate that

- Employment has a negative effect on crime and the coefficient $\beta_{employment} < 0$
- Not surprisingly, employment is postively correlated to wages. Given that $\beta_{employment} < 0$, the wage coefficients are negatively biased due to lack of information regarding employment.
- Employment is postively correlated to taxes. Given that $\beta_{employment} < 0$, β_{taxpc} is negatively biased.

Family cohesiveness
We believe there is merit to the idea that individuals from stable families, especially where both parents are active in the child's upbringing and character building may be less inclined to break the law. While ambiguous to quantify, sometimes young individuals may feel that they don't receive enough support at home and use deliquency as a way to get their family's attention. Thus, we estimate that

- Greater family cohesiveness has a negative effect on crime and the coefficient $\beta_{fam\_cohesiveness} < 0$
- Family cohesiveness is negatively correlated to *pctmle* in that a closer, possibly even a loving family dynamic, may encourage young males to stay away from crime and focus their energy on positive activities instead. Given that $\beta_{fam\_cohesiveness} < 0$, β_{pctmle} is positively biased.

Gang influence Another important factor to consider is presence and degree of influence of gangs, especially dense environments. Youth, especially young males, may join gangs for a variety of reasons (e.g. fulfilling basic essentials, the promise of protection, and simply peer pressure to fit in). Thus we estimate,

- Higher gang influence has a positive effect on crime and the coefficient $\beta_{gang} > 0$
- Gang influence is postively correlated to population *density* in that a densely populated are more susceptible to criminal activity by gang members. Given that $\beta_{gang} > 0$, $\beta_{density}$ is positively biased.
- Gangs are postively correlated to police acivity i.e. more gang presence may lead to more police and more arrests, convictions and prison time. Given that $\beta_{gang} > 0$, β_{polpc}, β_{prbarr}, $\beta_{prbconv}$, $\beta_{prbpris}$ are positively biased.
- Lastly, gangs tend to have more males than females. As mentioned above, young males are often recruited and as such, we estimate that gang influence is positively correlated to percent of younger males. Given that $\beta_{gang} > 0$, β_{pctmle} is positively biased.

these are the ones you care about

discuss how this might affect your conclusion.

Type of crime
The type of crime is an important factor that has not been captured. The *mix* variable does not provide enough granular information. If we have captured the type of crime in a few indicators variables (e.g. Use of gun, sexual assualt, trespassing etc), public officials could focus attention to specific solutions. e.g. if assaults involving a deadly weapon are most common, this could lead to policy changes involving better gun control or if sexual assault is the most pressing issue, then specific community outreach and educational programs may be introduced. Given the wide range of possibilities, estimating the direction of the bias is non-trivial.

Day of week / Time of Day
Knowing when and where crimes occur most frequently can help public safety officials plan effectively and better allocate resources. This missing variable can potentially bias the *polpc* explanatory variable though the direction of the bias is a bit unclear.

Location

Better knowledge of where the felonies are occurring (indoors / outdoors / type of location / intersection) can also help public safety officials curb crime by focusing on problem areas(e.g. if the data shows that crime is occurring in abandoned buildings more so than street intersections, policies can be implemented to better deal vacant properties, perhaps even convert them into spaces that positively impact the community. Again, this variable is potentially correlated to the *polpc* variable though the direction of the bias is a bit unclear.

Conclusions

Back to Table of Contents

Summary

The relationships between law enforcement (*polpc*), certainty of punishment (*prbarr* and *prbconv*), and crime rates (*crmtre*) are not unprecendented. Our models weakly corroborate this precedence and intuition. Although we can not claim that *polpc*, *prbarr*, and *prbconv* are the main determinants of crime, since the slopes were not different than 0, we can not negate their importance, since we failed to reject the hypothesis that they can be removed from the model individually. Interestly, these explanatory variables do not jointly contribute to our m_key and could have been removed from the model (see appendix). Sources of variation that might have contributed to these results are proposed below.

- These variables do not exist in a vacuum- we need to control for other covariates, for example those in m_associated. When covariates are added in m_associated and m_all the coefficient estimates for *polpc*, *prbarr*, *prbconv* change and the slopes become significantly different from zero (p-value < 0.05) (**table 3**). In all cases the standard errors decrease in size when covariates are added, from m_key to m_associated to m_all. Thus, these explanatory variables are not robust to alternative model specifications. As noted in the omitted variable bias discussion, granular knowledge of time and location of crime may enable politicians to deploy more police in areas with high crime.
- Outlying observations with influence. Observation 51, from county 115, likely pulled the regression surface away from the cloud of observations that defined the relationship between *crmrte* and *polpc* (**figure 1b**).
- Omitted variables may have deminished or negated the effects of *polpc*, *prbarr*, and *prbconv* on *crmrte*.
- The effects of *polpc* on crime are likely not independent of *prbarr* and *prbconv*, as suggested above and illustrated weakly in **figure A6**.
- The small sample size may have decreased our power.

good

Actionable Recommendations

While it might be tempting for politicians to increase police presence and mandate more arrests as a quick and easy solution, it is misguided to think that simply increasing police activity will decrease crime. We advise the politicians to consider crime in the context of location, demographics, and socio-economic factors (see the omitted variable bias section). With that said, in the context of police activity, we advise the following to the political campaign of NC:

Very good!

- increase police presence in areas of high crime, like densely populated urban areas (**figure 1b**). Omitted variables like time and specific locations could also be used to deploy resources optimally.
- provide training to officers to enable **informed** arrests, since 'probability' of arrests is negatively related to crime rates (**figure A6**). Arrests should be fair and not influenced by prejudice or an artificial quota that needs to be met.

Finally we think it might be valuable to increase community engagement between law enforcement and citizens. This will help erode mistrust and establish partnerships betweens citizens and the police that serve their neighborhoods. This would be cheap and simple and would not exhaust county budgets. In conclusion, while overhauling the law enforcement presence and practices may not be simple and cheap, certain steps can be taken to use resources more efficiently and reduce crime.

Appendix

Back to Table of Contents

Supplementary figures.

Figure A1: Distributions of variables

- Note that most variables are not normally distributed

In [49]:
```
options(repr.plot.height = 15, repr.plot.width = 20, repr.plot.points
par(mfrow = c(5,4))
par(mar=c(6,5,4,2))
hist(crime.narm$crmrte,breaks=20, main = "crmrte",cex.lab=1.5,cex.mai
hist(crime.narm$prbarr,breaks=20, main = "prbarr",cex.lab=1.5,cex.mai
hist(crime.narm$prbconv,breaks=20, main = "prbconv",cex.lab=1.5,cex.m
hist(crime.narm$prbpris,breaks=20, main = "prbpris",cex.lab=1.5,cex.m
hist(crime.narm$avgsen,breaks=20, main = "avgsen",cex.lab=1.5,cex.mai
hist(crime.narm$polpc,breaks=20, main = "polpc",cex.lab=1.5,cex.main=
hist(crime.narm$density,breaks=20, main = "density",cex.lab=1.5,cex.m
hist(crime.narm$taxpc,breaks=20, main = "taxpc",cex.lab=1.5,cex.main=
hist(crime.narm$pctmin80,breaks=20, main = "pctmin80",cex.lab=1.5,cex
hist(crime.narm$wcon,breaks=20, main = "wcon",cex.lab=1.5,cex.main=2.
hist(crime.narm$wtuc,breaks=20, main = "wtuc",cex.lab=1.5,cex.main=2.
hist(crime.narm$wtrd,breaks=20, main = "wtrd",cex.lab=1.5,cex.main=2.
hist(crime.narm$wfir,breaks=20, main = "wfir",cex.lab=1.5,cex.main=2.
hist(crime.narm$wser,breaks=20, main = "wser",cex.lab=1.5,cex.main=2.
hist(crime.narm$wmfg,breaks=20, main = "wmfg",cex.lab=1.5,cex.main=2.
hist(crime.narm$wfed,breaks=20, main = "wfed",cex.lab=1.5,cex.main=2.
hist(crime.narm$wsta,breaks=20, main = "wsta",cex.lab=1.5,cex.main=2.
hist(crime.narm$wloc,breaks=20, main = "wloc",cex.lab=1.5,cex.main=2.
hist(crime.narm$mix,breaks=20, main = "mix",cex.lab=1.5,cex.main=2.5,
hist(crime.narm$pctymle,breaks=20, main = "pctymle",cex.lab=1.5,cex.m
```

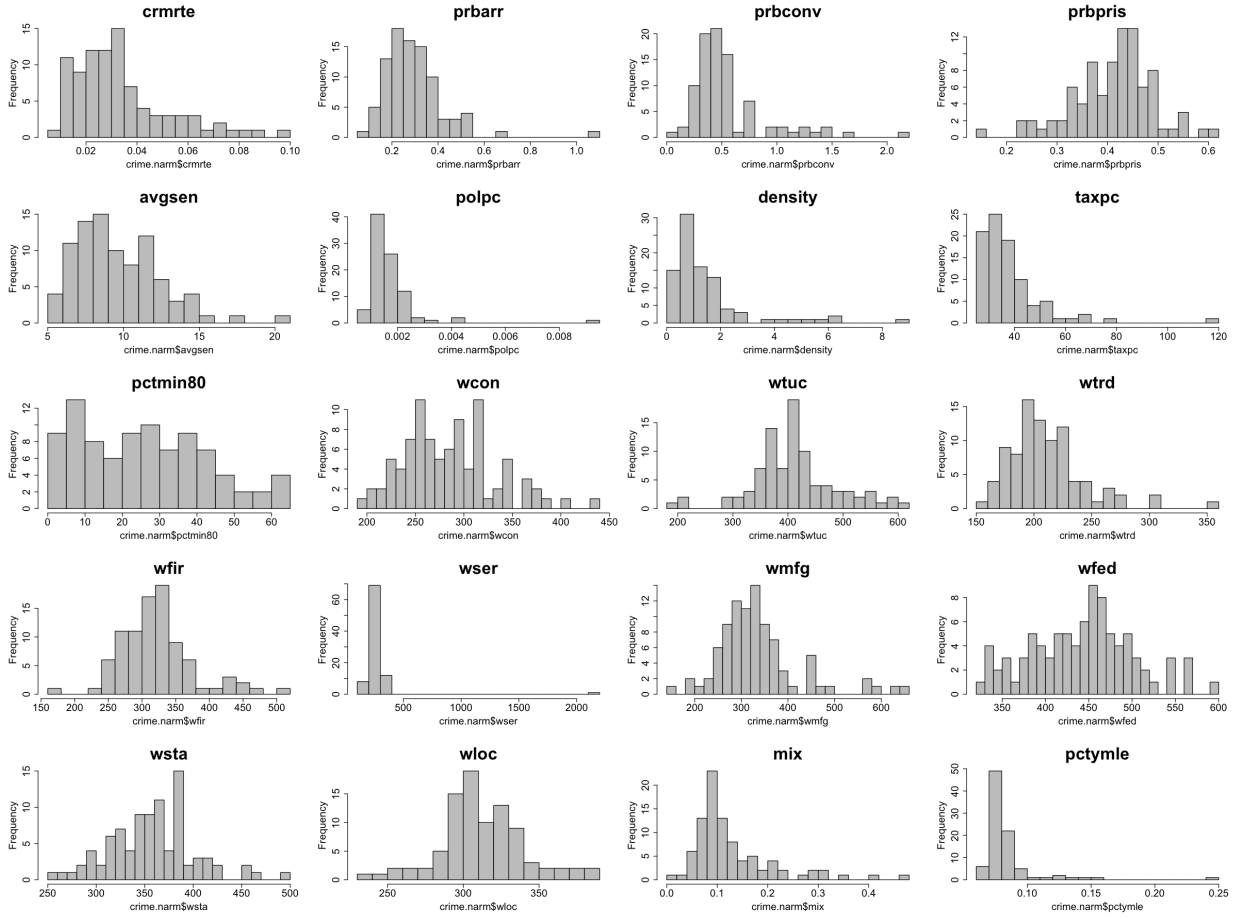
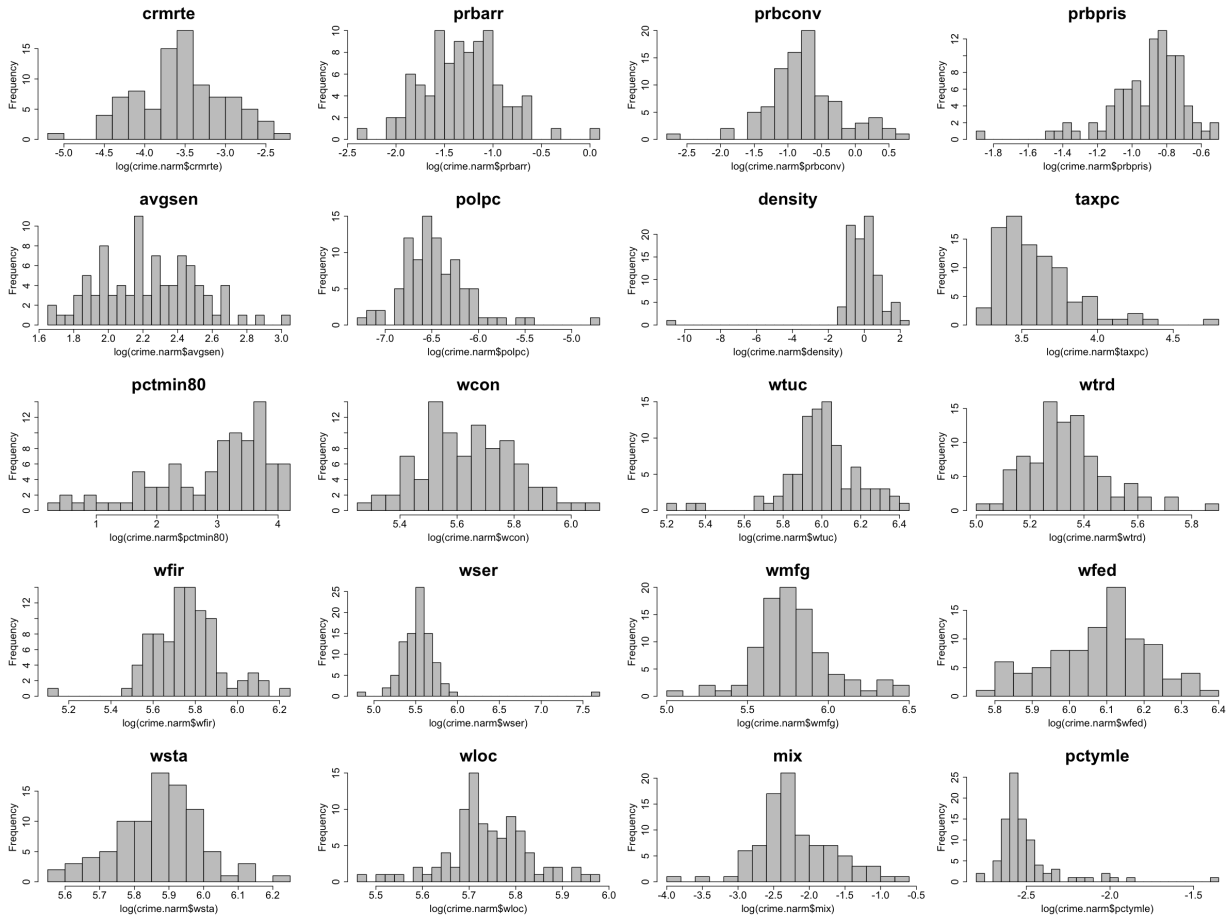


Figure A2: Distributions of variables after log-transformation

- Note that log-transformation resolved most instances of non-symetric distributions

In [50]:

```
options(repr.plot.height = 15, repr.plot.width = 20, repr.plot.points
par(mfrow = c(5,4))
par(mar=c(6,5,4,2))
hist(log(crime.narm$crmrte),breaks=20, main = "crmrte",cex.lab=1.5,ce
hist(log(crime.narm$prbarr),breaks=20, main = "prbarr",cex.lab=1.5,ce
hist(log(crime.narm$prbconv),breaks=20, main = "prbconv",cex.lab=1.5,
hist(log(crime.narm$prbpris),breaks=20, main = "prbpris",cex.lab=1.5,
hist(log(crime.narm$avgsen),breaks=20, main = "avgsen",cex.lab=1.5,ce
hist(log(crime.narm$polpc),breaks=20, main = "polpc",cex.lab=1.5,cex.
hist(log(crime.narm$density),breaks=20, main = "density",cex.lab=1.5,
hist(log(crime.narm$taxpc),breaks=20, main = "taxpc",cex.lab=1.5,cex.
hist(log(crime.narm$pctmin80),breaks=20, main = "pctmin80",cex.lab=1.
hist(log(crime.narm$wcon),breaks=20, main = "wcon",cex.lab=1.5,cex.ma
hist(log(crime.narm$wtuc),breaks=20, main = "wtuc",cex.lab=1.5,cex.ma
hist(log(crime.narm$wtrd),breaks=20, main = "wtrd",cex.lab=1.5,cex.ma
hist(log(crime.narm$wfir),breaks=20, main = "wfir",cex.lab=1.5,cex.ma
hist(log(crime.narm$wser),breaks=20, main = "wser",cex.lab=1.5,cex.ma
hist(log(crime.narm$wmfg),breaks=20, main = "wmfg",cex.lab=1.5,cex.ma
hist(log(crime.narm$wfed),breaks=20, main = "wfed",cex.lab=1.5,cex.ma
hist(log(crime.narm$wsta),breaks=20, main = "wsta",cex.lab=1.5,cex.ma
hist(log(crime.narm$wloc),breaks=20, main = "wloc",cex.lab=1.5,cex.ma
hist(log(crime.narm$mix),breaks=20, main = "mix",cex.lab=1.5,cex.main
hist(log(crime.narm$pctymle),breaks=20, main = "pctymle",cex.lab=1.5,
```



Correlations between dependent variable (*crmrte*) and independent variables

Figure A3: Crime and geography

- Note that (i) there is no perfect collinearity between the explanatory variables (ii) *ln(density)* is clearly related to *ln(crmrte)*.

In [51]:
```
options(repr.plot.height = 15, repr.plot.width = 20, repr.plot.points
pairs(~log(crmrte)
      + log(density)
      + jitter(as.numeric(west),0.5)
      + jitter(as.numeric(central),0.5)
      + jitter(as.numeric(urban),0.5),
      data = crime.narm,
      cex.labels=3, lower.panel = NULL,
      upper.panel=panel.smooth,
      pch=1,cex=3,lwd=3,col="grey55")
```

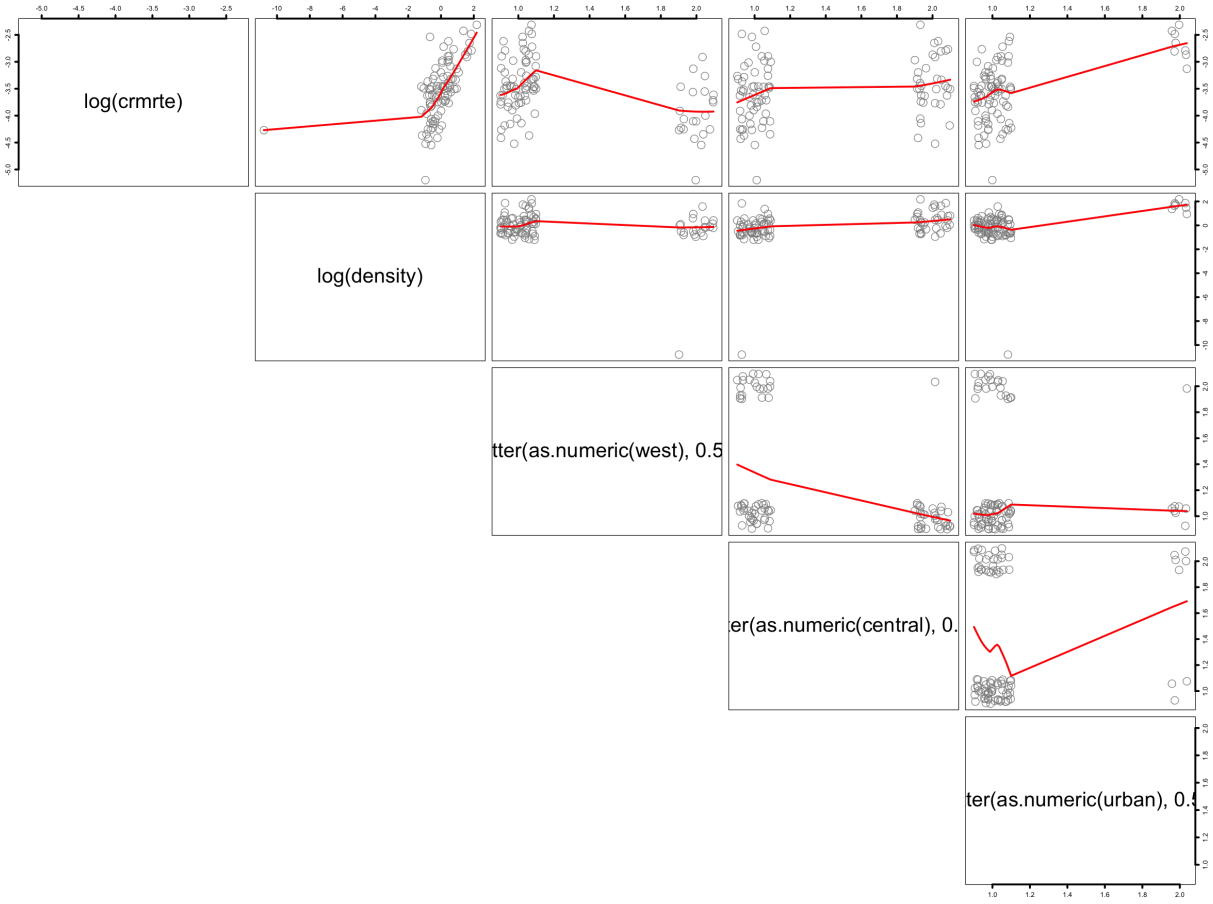


Figure A4: Crime and demographics

- Note that (i) there is no perfect collinearity between the explanatory variables (ii) *ln(polpc)* is clearly related to *ln(crmrte)*.

In [52]:
```
options(repr.plot.height = 15, repr.plot.width = 20, repr.plot.points
pairs(~log(crmrte)
      + log(polpc)
      + log(density)
      + log(pctmin80)
      + log(pctymle),
      data = crime.narm,
      cex.labels=3, lower.panel = NULL,
      upper.panel=panel.smooth,
      pch=1,cex=3,lwd=3,col="grey55")
```

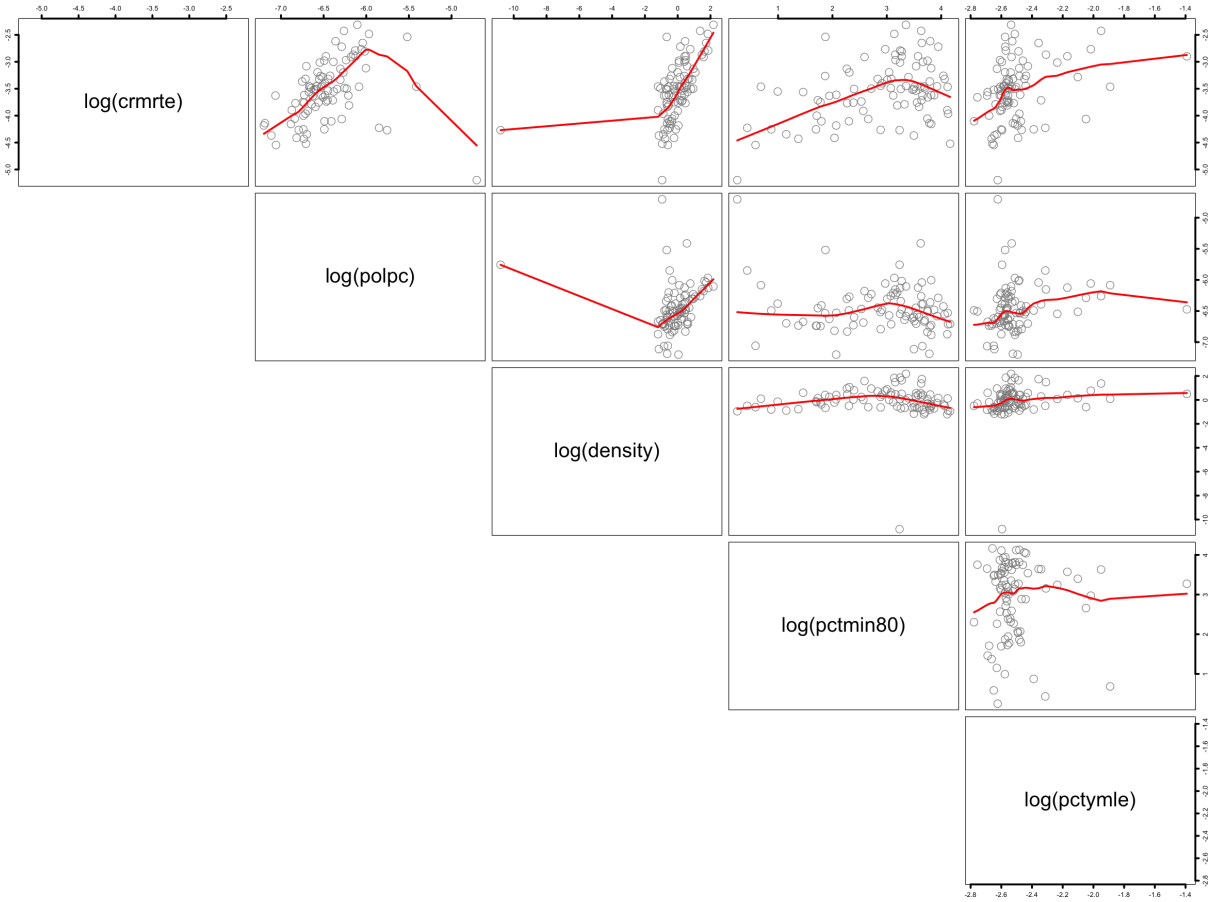


Figure A5: Crime and economics

- Note that (i) there is no perfect collinearity between the explanatory variables (ii) *ln(wfed)*, *ln(wcon)*, and *ln(wmfg)* are clearly related to *ln(crmrte)*.

In [53]:

```
options(repr.plot.height = 20, repr.plot.width = 20, repr.plot.points
pairs(~log(crmrte)
      + log(taxpc)
      + log(wcon)
      + log(wtuc)
      + log(wtrd)
      + log(wfir)
      + log(wser)
      + log(wmfg)
      + log(wfed)
      + log(wsta)
      + log(wloc),
      data = crime.narm,
      cex.labels=2.5, lower.panel = NULL,
      upper.panel=panel.smooth,
      pch=1,cex=0.8,lwd=2,col="grey55")
```

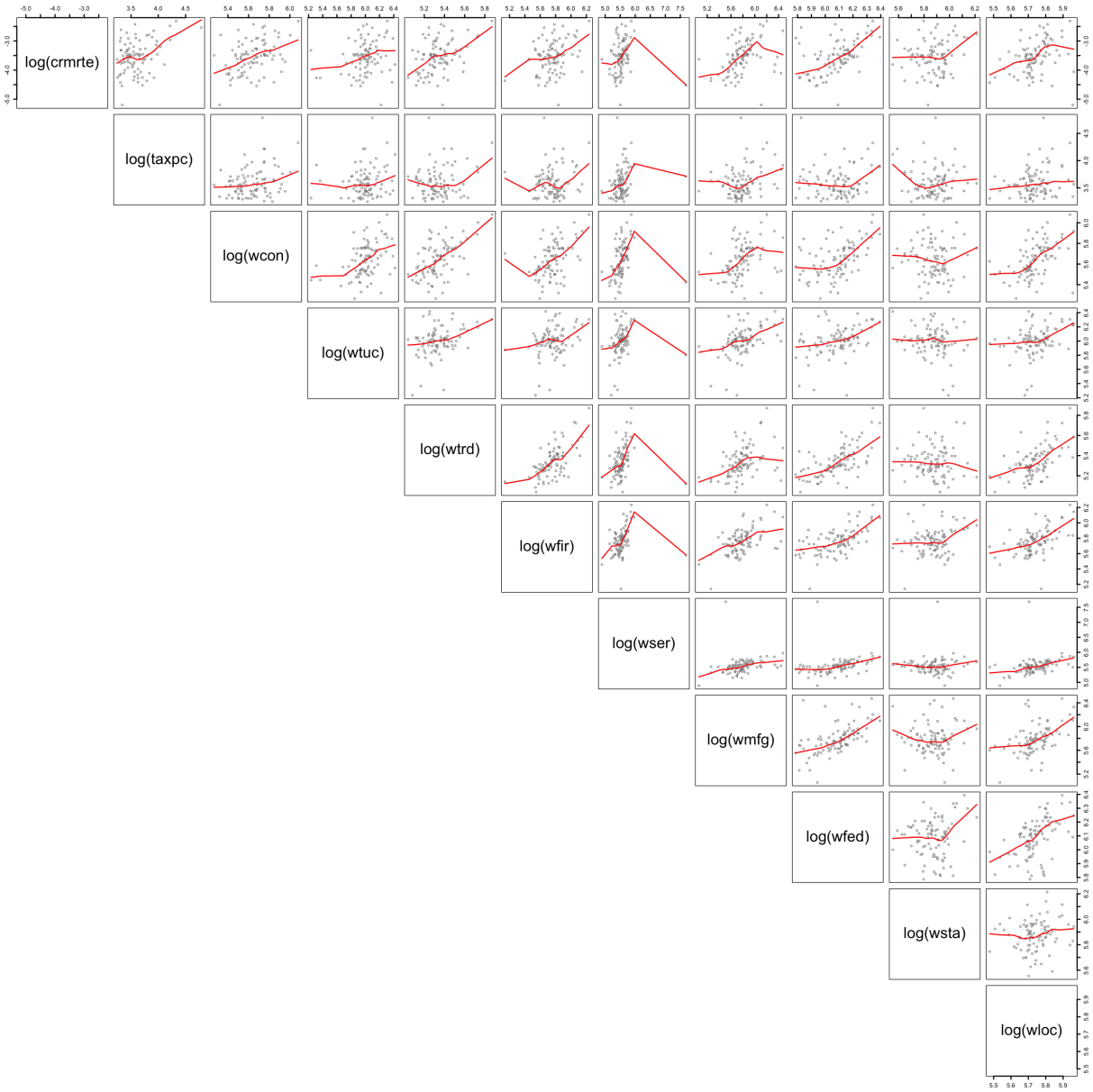
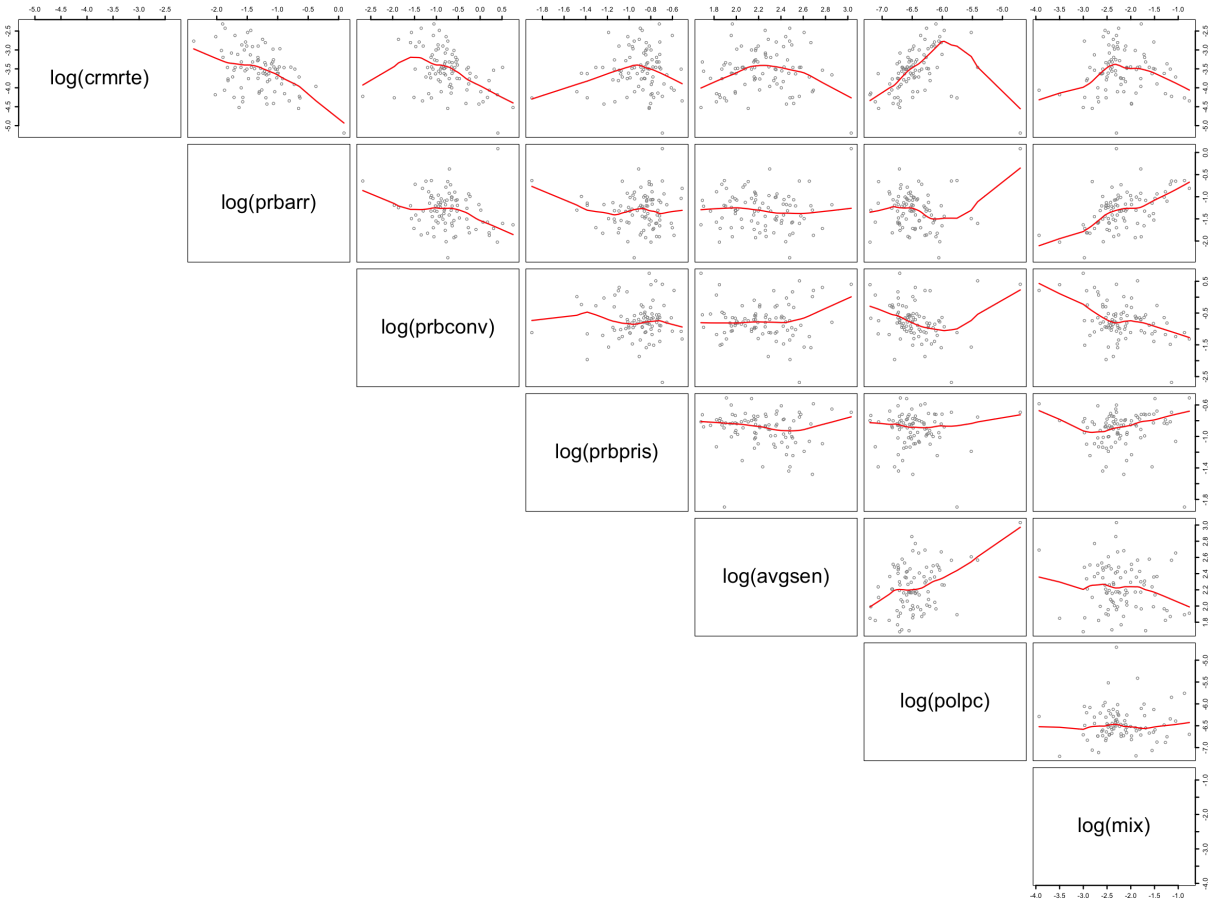


Figure A6: Crime and propensity to commit crime, be arrested, and sentence lengths

- Note that (i) there is no perfect collinearity between the explanatory variables (ii) *ln(prbarr)* and *ln(prconv)* are clearly related to *ln(crmrte)*.

In [54]:
```
options(repr.plot.height = 15, repr.plot.width = 20, repr.plot.points
pairs(~log(crmrte)
      + log(prbarr)
      + log(prbconv)
      + log(prbpris)
      + log(avgsen)
      + log(polpc)
      + log(mix),
      data = crime.narm,
      cex.labels=3, lower.panel = NULL,
      upper.panel=panel.smooth,
      pch=1,cex=1,lwd=2,col="grey55")
```



Outliers

Given that these observations may be influential in future models, future investigation is in order.

> Observation 79: note that the *density* is extremely low. There are $\frac{0.00002 \ people}{sq \ mile}$ in county 173!

In [55]:
```
1 subset(crime.narm, log(crime.narm$density) < -8)
```

	county	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west
79	173	0.0139937	0.530435	0.327869	0.15	6.64	0.00316379	2.03422e-05	37.72702	1

Observation 51: note that *polpc* is very high. There are 0.009 police per capita in county 115.

In [56]:
```
1 subset(crime.narm, log(crime.narm$polpc) > -5)
```

	county	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	··
51	115	0.0055332	1.09091	1.5	0.5	20.7	0.00905433	0.3858093	28.1931	1	··

Notes about model fit

The model fit, indicated by R^2 and AIC values, increased when all covariates were incorporated. The OLS model with covariates shows adjusted ($R^2 = 0.772$, AIC = 26.409) in the comparison to the model with only key exploratory variables ($R^2 = 0.474$, AIC = 96.385). In addition, even though the model with all variables provided the best prediction ($R^2 = 0.790$, AIC = 27.767), the risk of model overfitting and the cost (e.g. computing power, data collection, etc.) are increased at the same time.

The discrepancy between the coefficient of determination, R^2, for the nearly saturated model with all explanatory variables included ($R^2 = 0.790$) and the model with only the variables that politicians can influence simply and cheaply ($R^2 = 0.474$) is likely due to variety of sources and demands reconciliation. The full model, m_all, includes all explanatory variables while the simple model, m_key, includes 3 variables. It follows then that we accrue a $0.790 - 0.474 = 0.316 * 100 = 31.6\%$ increase in our ability to explain crime rate with the addition on 19 explanatory variables. Given that we can explain approximately 47% of the variation in crime rate with only 3 explanatory variables, the addition of 19 more explanatory variables to absorb 31.6% more variation in crime rate seems potentially wasteful.

From a broader perspective, it's more reasonable to communicate policy recommendations to the campaign using a model comprised of 3 versus 22 explanatory variables. The media and public have limited cognitive bandwidth. To ask the politicians to reduce crime rate by juggling 3 variables will be difficult enough, to include 19 more is ludicrous and beyond the scope of our goal.

Finally, from an economical perspective, why use the resources to explain crime rate with all variables when a model, m_associated, with comparable performance can be built with 10 variables? More importantly, by using the 3 variables from m_all, we could not only reduce the crime rate but also save the government resources to improve welfare, public services, senior care etc. If this study is to be repeated or reproduced elsewhere, we can save those investigators time, money, and ease of interpretation by only including the most salient explanatory variables.

Joint hypothesis test

$$H_0 : \beta_1 ln(polpc) = \beta_2 ln(prbarr) = \beta_3 ln(prbconv) = 0 \Rightarrow E(y|x_1, x_2, \ldots, x_n) = E(y)$$
$$H_A : \beta_1 ln(polpc) = \beta_2 ln(prbarr) = \beta_3 ln(prbconv) \neq 0 \Rightarrow \exists \beta_j \neq 0$$

In [57]:
```
# Hypothesis - H_0: Test if slopes are equal to zero
H0s = c("log(polpc) = 0","log(prbarr) = 0","log(prbconv) = 0")
# Wald test
linearHypothesis(
    m_key, H0s,
    vcov=vcovHC(m_key),
    singular.ok=TRUE
)
```

Res.Df	Df	F	Pr(>F)
88	NA	NA	NA
85	3	11.35223	2.462692e-06

Interestingly, we reject the null of the joint hypothesis described above. While we can not reject the hypotheses that these slopes are 0 alone, together they do not contribute towards describing y, *crmrte*.

Identify the independent variables which show significant contribution to crime rate

Does population density, tax revenue, and police force affect crime rate?

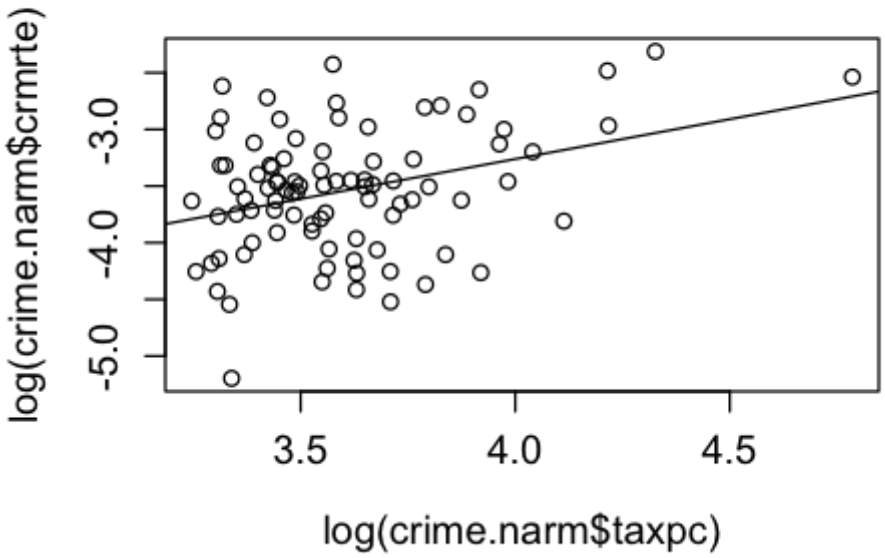
- Population, goverment income

In [58]:
```
# the explanatory variables of key interest
mp1 <- lm(log(crmrte) ~ log(density), data = crime.narm)
mp2 <- lm(log(crmrte) ~ log(polpc), data = crime.narm)
mp3 <- lm(log(crmrte) ~ log(taxpc), data = crime.narm)
```

In [59]:
```
# Compute robust standard errors
se.mp1 = sqrt(diag(vcovHC(mp1)))
se.mp2 = sqrt(diag(vcovHC(mp2)))
se.mp3 = sqrt(diag(vcovHC(mp3)))
# We pass the standard errors into stargazer through the se argument.
stargazer(mp1, mp2, mp3,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.mp1, se.mp2, se.mp3),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
==============================================
                   Dependent variable:
               -----------------------------
                         log(crmrte)
                  (1)        (2)       (3)
----------------------------------------------
log(density)     0.197
                (0.291)

log(polpc)                  0.417
                           (0.475)

log(taxpc)                           0.704***
                                      (0.207)

Constant       -3.525***   -0.847   -6.076***
                (0.055)    (3.104)   (0.750)

----------------------------------------------
Observations      90         90        90
Adjusted R2      0.235      0.071     0.105
==============================================
Note:           *p<0.05; **p<0.01; ***p<0.001
```

In [60]:
```
# check correlation between dependent variable and independent variab
options(repr.plot.height = 3, repr.plot.width = 4, repr.plot.pointsiz
plot(log(crime.narm$taxpc), log(crime.narm$crmrte))
abline(mp3)
```



> In the category of population and goverment income, *taxpc* significantly affects crime rate with positive correlation.

Does income affect crime rate?

- Employee incomes

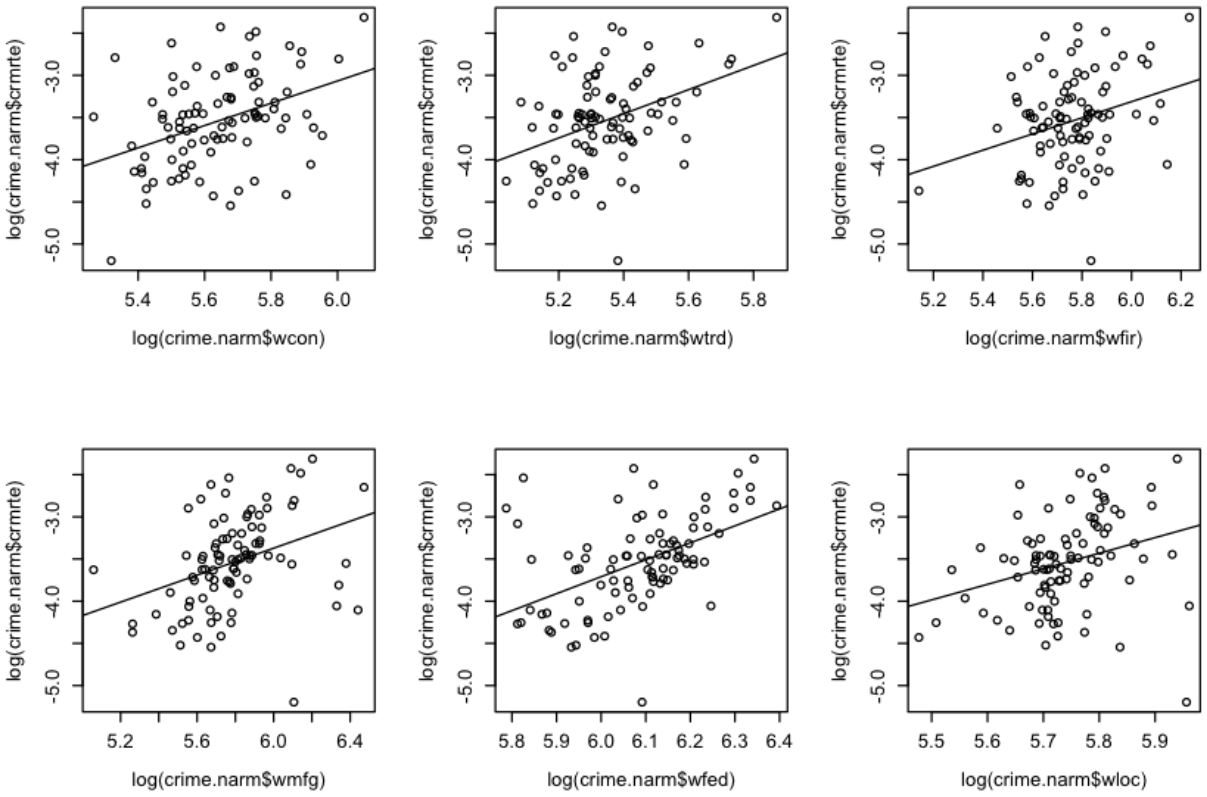
In [61]:
```
me1  <- lm(log(crmrte) ~ log(wcon), data = crime.narm)
me2  <- lm(log(crmrte) ~ log(wtuc), data = crime.narm)
me3  <- lm(log(crmrte) ~ log(wtrd), data = crime.narm)
me4  <- lm(log(crmrte) ~ log(wfir), data = crime.narm)
me5  <- lm(log(crmrte) ~ log(wser), data = crime.narm)
me6  <- lm(log(crmrte) ~ log(wmfg), data = crime.narm)
me7  <- lm(log(crmrte) ~ log(wfed), data = crime.narm)
me8  <- lm(log(crmrte) ~ log(wsta), data = crime.narm)
me9  <- lm(log(crmrte) ~ log(wloc), data = crime.narm)
```

In [62]:
```
# Compute robust standard errors
se.me1 = sqrt(diag(vcovHC(me1)))
se.me2 = sqrt(diag(vcovHC(me2)))
se.me3 = sqrt(diag(vcovHC(me3)))
se.me4 = sqrt(diag(vcovHC(me4)))
se.me5 = sqrt(diag(vcovHC(me5)))
se.me6 = sqrt(diag(vcovHC(me6)))
se.me7 = sqrt(diag(vcovHC(me7)))
se.me8 = sqrt(diag(vcovHC(me8)))
se.me9 = sqrt(diag(vcovHC(me9)))
# We pass the standard errors into stargazer through the se argument.
stargazer(me1, me2, me3, me4, me5, me6, me7, me8, me9,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.me1, se.me2, se.me3, se.me4, se.me5, se.me6, s
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
================================================================================
==============================
                                                 Dependent variable:
             ---------------------------------------------------------------
------------------------------
                                                     log(crmrte)
                (1)        (2)       (3)        (4)       (5)       (6)
(7)       (8)       (9)
--------------------------------------------------------------------------------
------------------------------
log(wcon)     1.322***
              (0.396)

log(wtuc)                0.574
                        (0.334)

log(wtrd)                          1.432***
                                   (0.331)

log(wfir)                                    0.959**
                                             (0.352)

log(wser)                                                0.093
                                                        (0.933)

log(wmfg)                                                          0.800**
                                                                   (0.305)


log(wfed)
2.010***

(0.473)

log(wsta)
0.686

(0.459)

log(wloc)
```

```
1.846*

(0.907)

Constant     -10.999*** -6.986*** -11.191*** -9.068*** -4.060  -8.170***
-15.768*** -7.573** -14.136**
              (2.237)    (1.995)   (1.770)    (2.020)  (5.142)  (1.747)
(2.898)   (2.695)   (5.181)

------------------------------------------------------------------------
----------------------------
Observations     90         90          90         90        90        90
90         90        90
Adjusted R2    0.147      0.034     0.142      0.073   -0.009    0.116
0.243     0.012     0.081
========================================================================
============================
Note:
*p<0.05; **p<0.01; ***p<0.001
```

In [63]:
```
options(repr.plot.height = 5, repr.plot.width = 7, repr.plot.pointsiz
par(mfrow = c(2,3))
plot(log(crime.narm$wcon), log(crime.narm$crmrte))
abline(me1)
plot(log(crime.narm$wtrd), log(crime.narm$crmrte))
abline(me3)
plot(log(crime.narm$wfir), log(crime.narm$crmrte))
abline(me4)
plot(log(crime.narm$wmfg), log(crime.narm$crmrte))
abline(me6)
plot(log(crime.narm$wfed), log(crime.narm$crmrte))
abline(me7)
plot(log(crime.narm$wloc), log(crime.narm$crmrte))
abline(me9)
```



In the category of employee income, *wcon*, *wtrd*, *wfir*, *wmfg*, *wfed*, and *wloc* significantly affect crime rate with positive correlation.

Do young males and the percent of population that is minority affect crime rate?

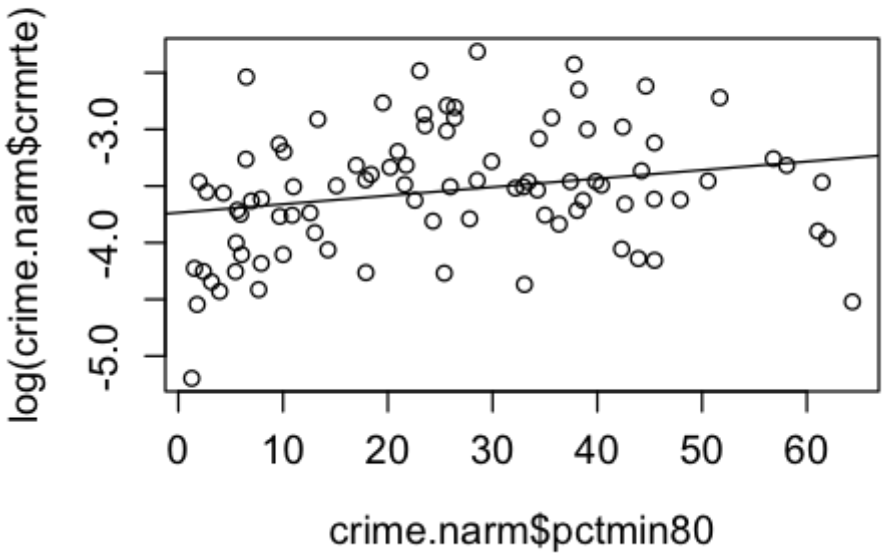
- minority and young male

In [64]:
```
mm1 <- lm(log(crmrte) ~ pctmin80, data = crime.narm)
mm2 <- lm(log(crmrte) ~ pctymle, data = crime.narm)
```

In [65]:
```
# Compute robust standard errors
se.mm1 = sqrt(diag(vcovHC(mm1)))
se.mm2 = sqrt(diag(vcovHC(mm2)))
# We pass the standard errors into stargazer through the se argument.
stargazer(mm1, mm2,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.mm1, se.mm2),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
==============================================
                  Dependent variable:
              ----------------------------
                      log(crmrte)
                    (1)             (2)
----------------------------------------------
pctmin80           0.008
                  (0.004)

pctymle                            6.509
                                  (3.739)

Constant         -3.735***        -4.089***
                  (0.115)          (0.311)

----------------------------------------------
Observations        90               90
Adjusted R2        0.044            0.067
==============================================
Note:          *p<0.05; **p<0.01; ***p<0.001
```

In [66]:
```
options(repr.plot.height = 3, repr.plot.width = 4, repr.plot.pointsiz
plot(crime.narm$pctmin80, log(crime.narm$crmrte))
abline(mm1)
```



In the category of minority and young male, *pctmin80* significantly affects crime rate with positive correlation.

Does geographic location affect crime rate?

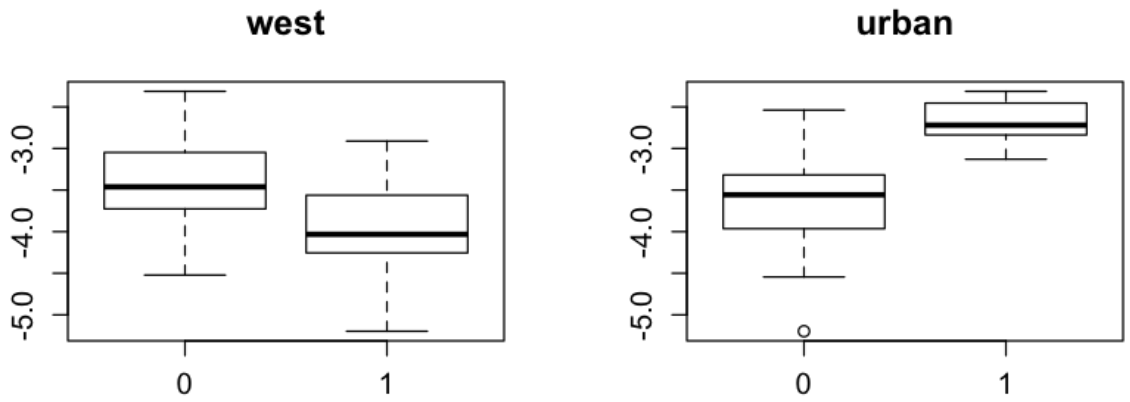
- geographic factors

In [67]:
```
mg1 <- lm(log(crmrte) ~ factor(west), data = crime.narm)
mg2 <- lm(log(crmrte) ~ factor(central), data = crime.narm)
mg3 <- lm(log(crmrte) ~ factor(urban), data = crime.narm)
```

In [68]:
```
# Compute robust standard errors
se.mg1 = sqrt(diag(vcovHC(mg1)))
se.mg2 = sqrt(diag(vcovHC(mg2)))
se.mg3 = sqrt(diag(vcovHC(mg3)))
# We pass the standard errors into stargazer through the se argument.
stargazer(mg1, mg2, mg3,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.mg1, se.mg2, se.mg3),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
================================================
                        Dependent variable:
                    -----------------------------
                              log(crmrte)
                       (1)        (2)        (3)
------------------------------------------------
factor(west)1      -0.526***
                    (0.129)

factor(central)1                0.208
                               (0.118)

factor(urban)1                            0.942***
                                           (0.115)

Constant           -3.413***  -3.620***  -3.625***
                    (0.061)    (0.075)    (0.055)

------------------------------------------------
Observations          90         90         90
Adjusted R2         0.162      0.023      0.233
================================================
Note:                *p<0.05; **p<0.01; ***p<0.001
```

In [69]:
```
options(repr.plot.height = 3, repr.plot.width = 7, repr.plot.pointsiz
par(mfrow = c(1,2))
plot(factor(crime.narm$west), log(crime.narm$crmrte), main = "west")
plot(factor(crime.narm$urban), log(crime.narm$crmrte), main = "urban"
```



In the category of geographic factors, *west* significantly affects crime rate with negative correlation. On the other hand, *urban* significantly affects crime rate with positive correlation

Do outcomes of crime affect crime rate?

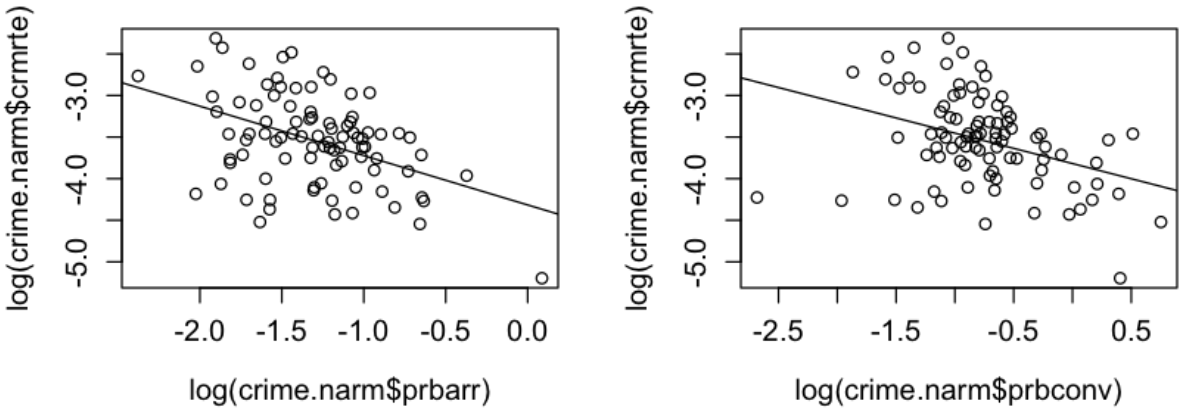
- crime outcomes

In [70]:
```
mc1 <- lm(log(crmrte) ~ log(prbarr), data = crime.narm)
mc2 <- lm(log(crmrte) ~ log(prbconv), data = crime.narm)
mc3 <- lm(log(crmrte) ~ log(prbpris), data = crime.narm)
mc4 <- lm(log(crmrte) ~ log(avgsen), data = crime.narm)
mc5 <- lm(log(crmrte) ~ log(mix), data = crime.narm)
```

In [71]:
```
# Compute robust standard errors
se.mc1 = sqrt(diag(vcovHC(mc1)))
se.mc2 = sqrt(diag(vcovHC(mc2)))
se.mc3 = sqrt(diag(vcovHC(mc3)))
se.mc4 = sqrt(diag(vcovHC(mc4)))
se.mc5 = sqrt(diag(vcovHC(mc5)))
# We pass the standard errors into stargazer through the se argument.
stargazer(mc1, mc2, mc3, mc4, mc5,
          type="text", keep.stat=c("n", "adj.rsq"),
          se = list(se.mc1, se.mc2, se.mc3, se.mc4, se.mc5),
          star.cutoffs=c(0.05, 0.01, 0.001)
         )
```

```
===============================================================
                         Dependent variable:
             ----------------------------------------------
                               log(crmrte)
                (1)       (2)       (3)       (4)       (5)
---------------------------------------------------------------
log(prbarr)  -0.594***
              (0.156)

log(prbconv)            -0.366*
                        (0.158)

log(prbpris)                       0.172
                                  (0.287)

log(avgsen)                                   0.046
                                             (0.294)

log(mix)                                                0.0004
                                                        (0.114)

Constant     -4.316*** -3.818*** -3.385*** -3.644*** -3.541***
              (0.199)   (0.113)   (0.276)   (0.644)   (0.252)

---------------------------------------------------------------
Observations    90        90        90        90        90
Adjusted R2    0.181     0.129    -0.006    -0.011    -0.011
===============================================================
Note:                           *p<0.05; **p<0.01; ***p<0.001
```

In [72]:
```
options(repr.plot.height = 3, repr.plot.width = 7, repr.plot.pointsiz
par(mfrow = c(1,2))
plot(log(crime.narm$prbarr), log(crime.narm$crmrte))
abline(mc1)
plot(log(crime.narm$prbconv), log(crime.narm$crmrte))
abline(mc2)
```



In the category of geographic factors, *prbarr* and *prbconv* significantly affect crime rate with negative correlation.

Test the regression model based on the variables with related characteristics

In [73]:
```
mp4 <- lm(log(crmrte) ~ log(density) + log(polpc)
          + log(taxpc), data = crime.narm)
me10 <- lm(log(crmrte) ~ log(wcon) + log(wtuc) + log(wtrd) +
         log(wfir) + log(wser) + log(wmfg) + log(wfed) +
         log(wsta) + log(wloc), data = crime.narm)
mm3 <- lm(log(crmrte) ~ pctmin80 + pctymle, data = crime.narm)
mg4 <- lm(log(crmrte) ~ factor(west) + factor(central) + factor(urban
mc6 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) +
         log(prbpris) + log(avgsen) + log(mix), data = crime.narm)

paste("adj.r.squared of population and goverment income: ", summary(m
paste("adj.r.squared of employee income: ", summary(me10)$adj.r.squar
paste("adj.r.squared of minority and young male: ", summary(mm3)$adj.
paste("adj.r.squared of geographic factors: ", summary(mg4)$adj.r.squ
paste("adj.r.squared of crime outcomes: ", summary(mc6)$adj.r.squared
```

'adj.r.squared of population and goverment income: 0.335445808769382'

'adj.r.squared of employee income: 0.255168337418916'

'adj.r.squared of minority and young male: 0.114261587697344'

'adj.r.squared of geographic factors: 0.360644623908643'

'adj.r.squared of crime outcomes: 0.396489655078683'

The OLS model of population and goverment income varibales shows adjusted $R^2 = 0.335$. The OLS model of employee income varibales shows adjusted $R^2 = 0.255$. The OLS model of minority and young male varibales shows adjusted $R^2 = 0.114$. The OLS model of geographic factors varibales shows adjusted $R^2 = 0.361$. The OLS model of crime outcomes varibales shows adjusted $R^2 = 0.396$. Therefore, the OLS models based on the categories do not provide competitive explanations of the variation in *crmrte*.

In [ ]:
```
```