Lab 3: Reducing Crime - Final Report

W203 Instructional Team - Prof. Penner - S2 - Mon 6:30pm

Adam Sohn, Alvin Lim, Brian Neesby, Khyati Tripathi ¶

0.0 Abstract

This report is the documentation of the creation and interpretation of OLS regression models to characterize the determinants of crime in 1987 North Carolina for use in a 1988 North Carolina gubenatorial campaign. Covariates examined from a dataset from C. Cornwell and W. Trumball (1994) were in the categories of Law Enforcement, Wealth, Identity, Population Density, and Region.

Data were cleaned and examined through an EDA process and increasing covariate-count models were compared for model fit and parsimony. An initial model consisting of two seemingly sound contributors (wealth as proxied by tax per-capita, and residential density) demonstrated utility and was able to explain more than 57% of the variance in crime rate. A second model containing two additional variables of interest (probability of conviction and probability of arrest) as well as their interaction was both more parsimonious and explained more crime variance (68%).

An exploratory model with almost all data set variables included was then assessed. This had superior AIC and adjusted-R^2, and highlighted additional regressors of significance, demonstrating the insufficiency of both Model 1 and Model 2. Finally, Model 4 was generated based on the findings from Model 3, and this final model had the lowest AIC and highest adjusted-R^2 (80% variance explained) of all models and was retained as the best model for the purposes of this analysis.

The resultant model is: $crmrte = -0.020 + 0.005density + 0.0004taxpc - 0.029prbcov - 0.074prbarr + 0.051prconv * prbarr + 0.017log(pctymle) + 0.0003pctmin80 + 0.00004wfed$

These findings are used to recommend that the Jordan campaign focus on policies which would:

- Improve effectiveness of police investigations and prosecutions to increase arrest and conviction rates.
- Focus law enforcement effots in areas of high population density and tax revenue.
- Leverage zoning laws to prevent increase in population density.

Also, an examination of omitted variables leads to a recommendation that further study be done into:

- Community relations with police
- Reduction in eviction rate

1.0 Introduction

From 1960, when the North Carolina state government began tracking state crime statistics, through 1987, the overall crime rate has been on a generally increasing trend. Overall crime per capita has grown over 400% (http://www.disastercenter.com/crime/nccrimn.htm) in this time. The electorate is up in arms and the 1988 Robert B Jordan gubernatorial campaign could win voter trust by exploiting this issue.

The team analyzed recent (1987) North Carolina crime data to understand the relationship between the crime rate and variables that are responsive to political influence: Wealth and Population Density. It is intended to identify these relationships and expound on their political value.

Research Question:

When considering population density and wealth as variables related to crime rate, what is the combination of density & wealth that is most likely to define a zone of maximal return on political investment?

2.0 The Initial Data Loading and Cleaning

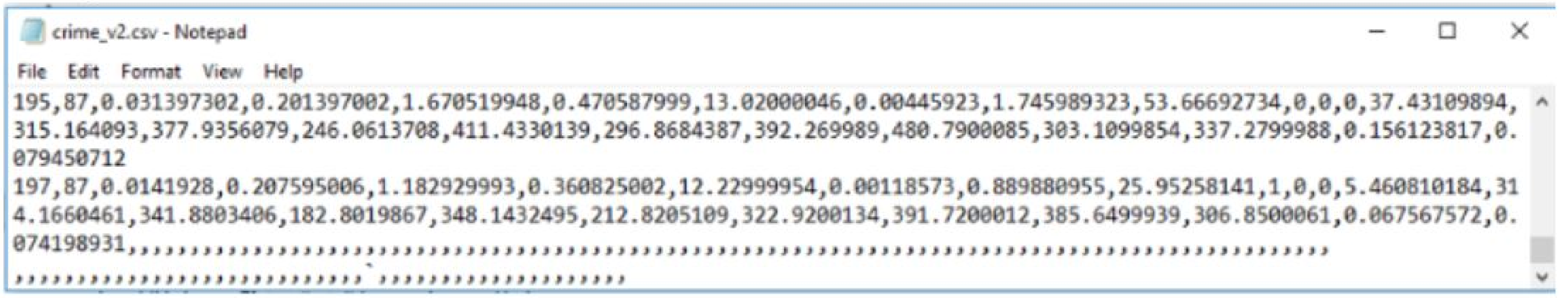
A first step towards data cleaning was to identify the outcome (response) variable. As the customer is interested in examining determinants of crime, two variable options are Crime Rate (crmrte) and Offense Mix (mix), with the latter being a ratio of face-to-face crime vs. other crime. As 'crmrte' illustrates overall instances (quantity) of crime as opposed to 'mix' illustrating the character (quality) of the crime, it is the more directly relevant of the two metrics to the research question. 'crmrte' was chosen as the outcome variable and 'mix' was not considered in this analysis.

Exploratory Data Analysis (EDA) is detailed below:

- 2.1. Data Load and Clean: Ensure dataset is complete and workable for analysis.
- 2.2. Scatterplot & Leverage Analysis
- 2.3. Variable Analysis: Correlation w/ crime rate
- 2.4. Variable Analysis: Determine over-leverage points
- 2.5 Variable Analysis: Determining any top-coding/bottom-coding
- 2.6 Variable Analysis: Empirical data exploration

2.1 Data Load and Clean:

Data file has an odd ending of ",,,`,,\" which results in NAs in created dataframe.


crime_v2.csv - Notepad

File Edit Format View Help

195,87,0.031397302,0.201397002,1.670519948,0.470587999,13.02000046,0.00445923,1.745989323,53.66692734,0,0,0,37.43109894,
315.164093,377.9356079,246.0613708,411.4330139,296.8684387,392.269989,480.7900085,303.1099854,337.2799988,0.156123817,0.
079450712
197,87,0.0141928,0.207595006,1.182929993,0.360825002,12.22999954,0.00118573,0.889880955,25.95258141,1,0,0,5.460810184,31
4.1660461,341.8803406,182.8019867,348.1432495,212.8205109,322.9200134,391.7200012,385.6499939,306.8500061,0.067567572,0.
074198931,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,`,,,,,,,,,,,,,,,,,,,,,

As part of data cleaning, NAs are removed.
As part of transformation, pctymle variable is multiplied with 100.

In [37]:

```
#Libary Import
library(stargazer) # used for formating
library(car) # used to plot scatterplotmatrix
library (ggplot2) #used for graphic plot
library(tidyr) # used for cleaning dataset of NA
library(repr) # used to show all columns in head/tail.

options(warn=-1) # To remove warnings
options(repr.matrix.max.cols = 30) #Show all columns in head/tail

#Importing Data
crime_df = read.csv("crime_v2.csv")

#Initial Data View
head(crime_df,1)
cat('Note that the final 6 rows of imported data are NA.')
tail(crime_df,7)
```

county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	central	urban	pctmin80	wcon	w
1	87	0.0356036	0.29827	0.527595997	0.43617	6.71	0.00182786	2.422633	30.99368	0	1	0	20.2187	281.4259	4

Note that the final 6 rows of imported data are NA.

	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	central	urban	pctmin80	wcon
91	197	87	0.0141928	0.207595	1.182929993	0.360825	12.23	0.00118573	0.889881	25.95258	1	0	0	5.46081	314.1
92	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
93	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
94	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
95	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
96	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
97	NA	NA	NA	NA	`	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

In [2]:
```
#Cleaning data to remove final",,,`,,"
crime_df <- read.csv("crime_v2.csv", na.strings ="``") #convert tilde to NA for solve in next step.
crime_df <- crime_df %>% drop_na() #Drop NAs

#Demonstrating fix
head(crime_df,2)
tail(crime_df,2)
```

county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	central	urban	pctmin80	wcon	wtu
1	87	0.0356036	0.298270	0.527596	0.43617	6.71	0.00182786	2.422633	30.99368	0	1	0	20.21870	281.4259	408
3	87	0.0152532	0.132029	1.481480	0.45000	6.35	0.00074588	1.046332	26.89208	0	1	0	7.91632	255.1020	376

	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	central	urban	pctmin80	wcon
90	195	87	0.0313973	0.201397	1.67052	0.470588	13.02	0.00445923	1.745989	53.66693	0	0	0	37.43110	315.1641
91	197	87	0.0141928	0.207595	1.18293	0.360825	12.23	0.00118573	0.889881	25.95258	1	0	0	5.46081	314.1660

2.2 Scatterplot & Leverage Analysis

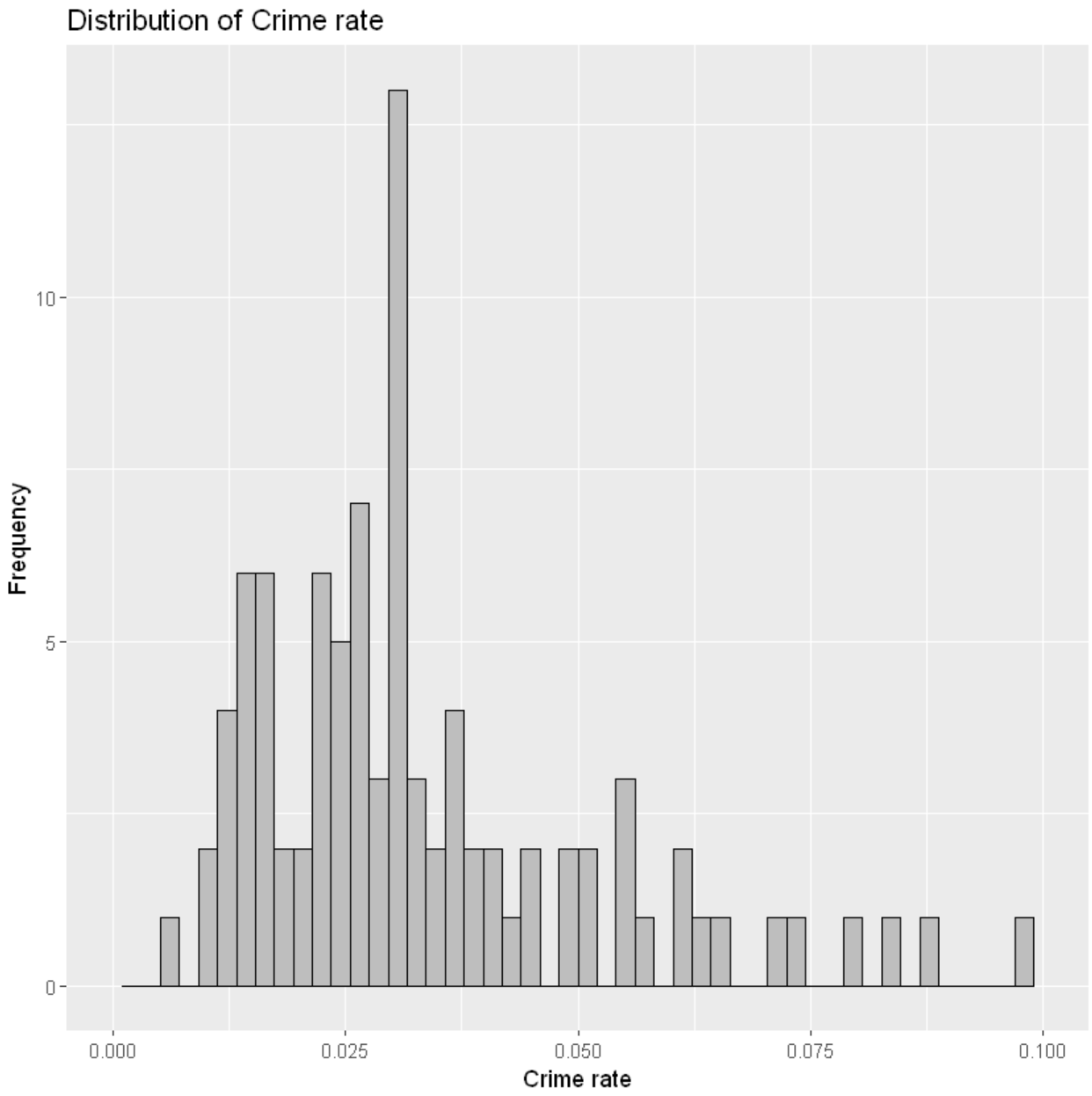
There are six counties with missing attribute values. Note that the range of 'pctymle' should be between 0 and 100. Hence, 'pctymle' was transformed to percentage by multiplying by 100 to provide parity with treatment of pctmin80.

After filtering records with missing values and pre-EDA data transformation, leverage analysis is is performed on variables which will ultimately become models 1, 2, and 4 (best model). Other variables in data set are omitted from EDA for brevity.

In [3]:
```
## Simplifying all variable names
crmrte <- crime_df$crmrte
prbarr <- crime_df$prbarr
prbconv <- crime_df$prbconv
prbpris <- crime_df$prbpris
avgsen <- crime_df$avgsen
polpc <- crime_df$polpc
density <- crime_df$density
taxpc <- crime_df$taxpc
west <- crime_df$west
central <- crime_df$central
urban <- crime_df$urban
pctmin80 <- crime_df$pctmin80
wcon <- crime_df$wcon
wtuc <- crime_df$wtuc
wtrd <- crime_df$wtrd
wfir <- crime_df$wfir
wser <- crime_df$wser
wmfg <- crime_df$wmfg
wfed <- crime_df$wfed
wsta <- crime_df$wsta
wloc <- crime_df$wloc
mix <- crime_df$mix
pctymle <- crime_df$pctymle * 100
```

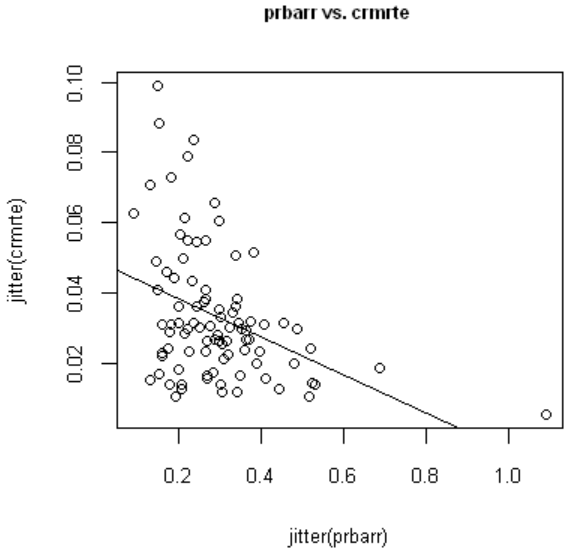
In [4]:
```
#Plot showing Distribution of Crime Rate
(ggplot(crime_df, aes(x=crmrte)) + geom_histogram(bins=50, color="black", fill="gray") + xlab("Crime rate") +
 scale_x_continuous(limits=c(0, 0.1)) + ylab("Frequency") +
ggtitle("Distribution of Crime rate"))
cat("Note that crime rate appears to have a normal curve with a tail towards higher crime.")
```

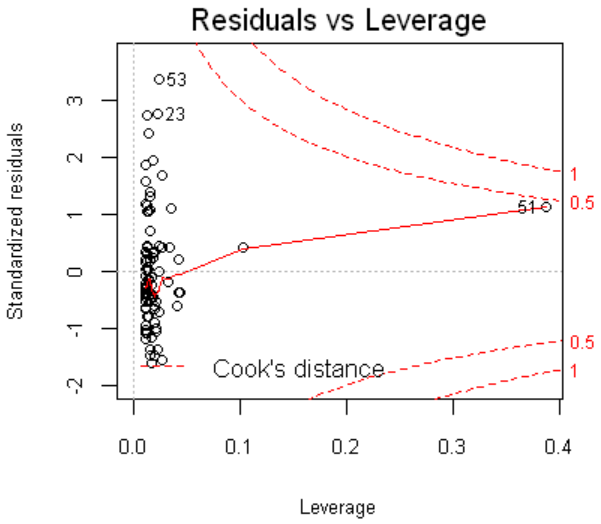
```
Note that crime rate appears to have a normal curve with a tail towards higher crime.
```



In [6]:
```
par(mfrow=c(2,2))
plot(jitter(prbarr), jitter(crmrte), main="prbarr vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8,
     cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8)
model1 = lm(crmrte ~ prbarr, data = crime_df)
abline(model1)
plot(model1, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.
8)
cat("Note that crmte appears to increase as prbarr declines. This is expected, as a lack of apprehension begets repeat offe
nders. There is one data point that is high leverage, but not influential (Cook's distance < 1). Scatter does not clearly d
emonstrate homoskedasticity (residuals spread) or heteroskedasticity (few data points prior to spread).")
```

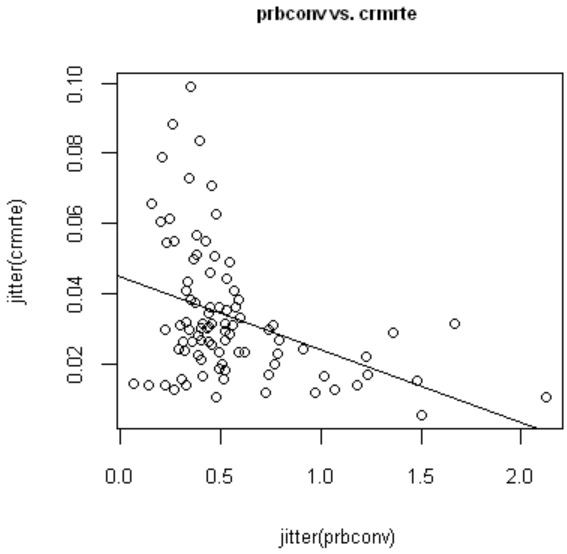
```
Note that crmte appears to increase as prbarr declines. This is expected, as a lack of apprehension begets repeat offender
s. There is one data point that is high leverage, but not influential (Cook's distance < 1). Scatter does not clearly demon
strate homoskedasticity (residuals spread) or heteroskedasticity (few data points prior to spread).
```

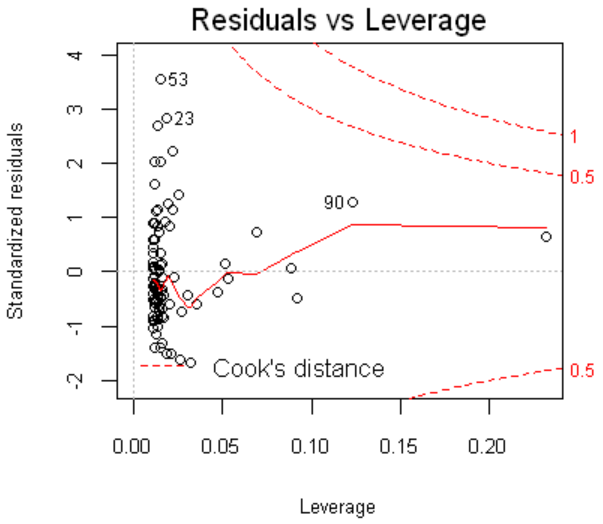




In [7]:
```
par(mfrow=c(2,2))
plot(jitter(prbconv), jitter(crmrte), main="prbconv vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8)
model2 = lm(crmrte ~ prbconv, data = crime_df)
abline(model2)
plot(model2, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.
8)
cat("Note that crmte appears to increase as prbconv declines. This is expected, as a lack of conviction begets repeat offen
ders. There is one data point that is high leverage, but not influential (Cook's distance < 1). Scatter does not clearly de
monstrate homoskedasticity (residuals spread) or heteroskedasticity (few data points prior to spread).")
```

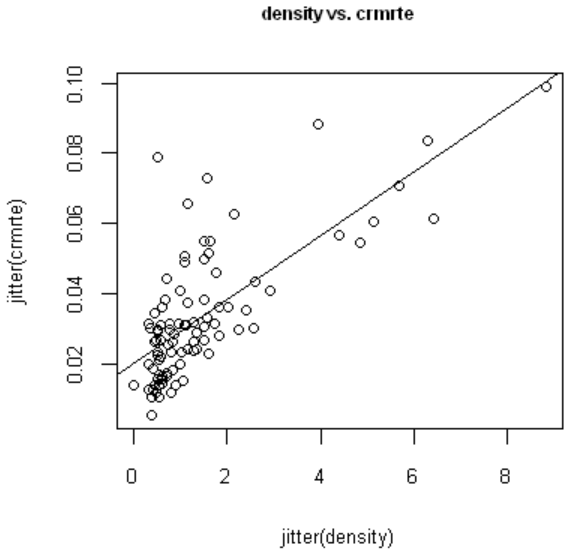
```
Note that crmte appears to increase as prbconv declines. This is expected, as a lack of conviction begets repeat offenders.
There is one data point that is high leverage, but not influential (Cook's distance < 1). Scatter does not clearly demonstr
ate homoskedasticity (residuals spread) or heteroskedasticity (few data points prior to spread).
```

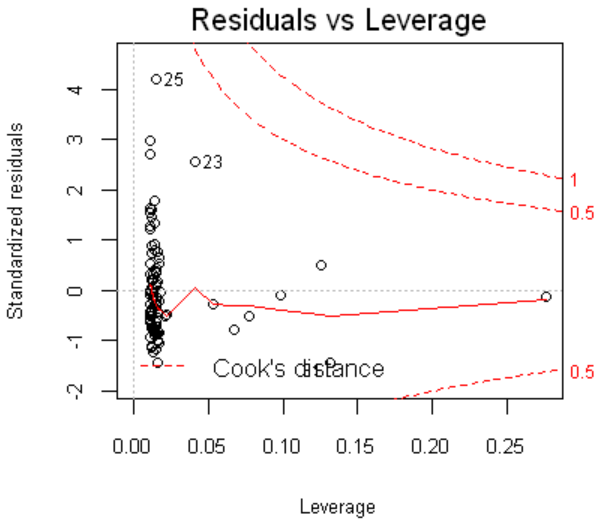




In [8]:
```
par(mfrow=c(2,2))
plot(jitter(density), jitter(crmrte), main="density vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8)
model6 = lm(crmrte ~ density, data = crime_df)
abline(model6)
plot(model6, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.
8)
cat("Note that crmte appears to increase as density increases. This is expected, as proximity exacerbates antisocial tenden
cies. There is one data point that is high leverage, but not influential (Cook's distance < 1). Scatter does not clearly de
monstrate homoskedasticity (residuals spread) or heteroskedasticity (few data points prior to spread).")
```

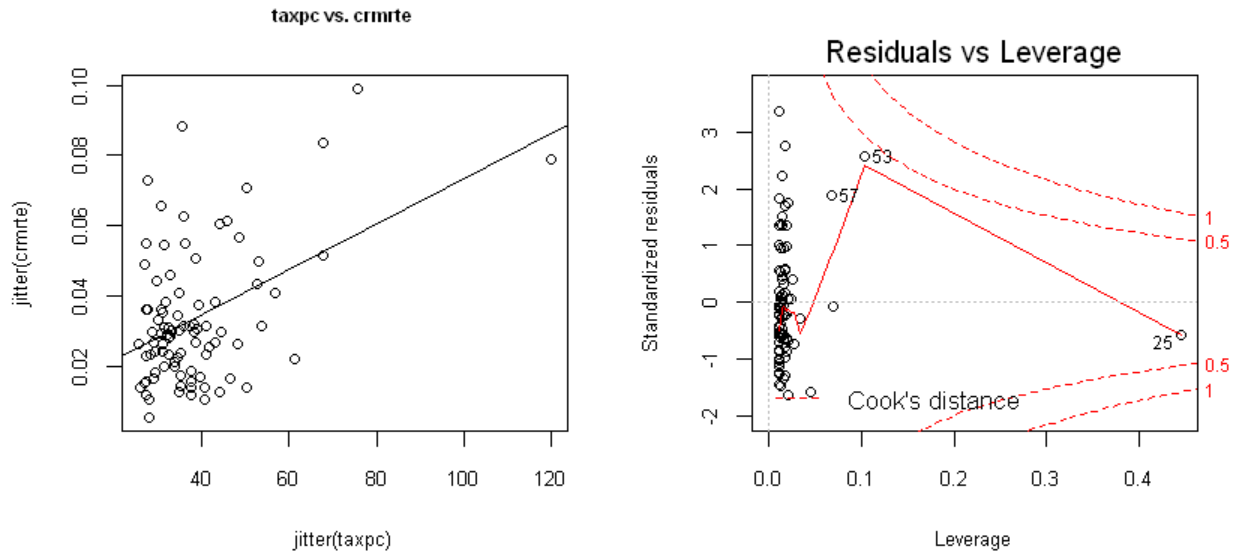
```
Note that crmte appears to increase as density increases. This is expected, as proximity exacerbates antisocial tendencies.
There is one data point that is high leverage, but not influential (Cook's distance < 1). Scatter does not clearly demonstr
ate homoskedasticity (residuals spread) or heteroskedasticity (few data points prior to spread).
```





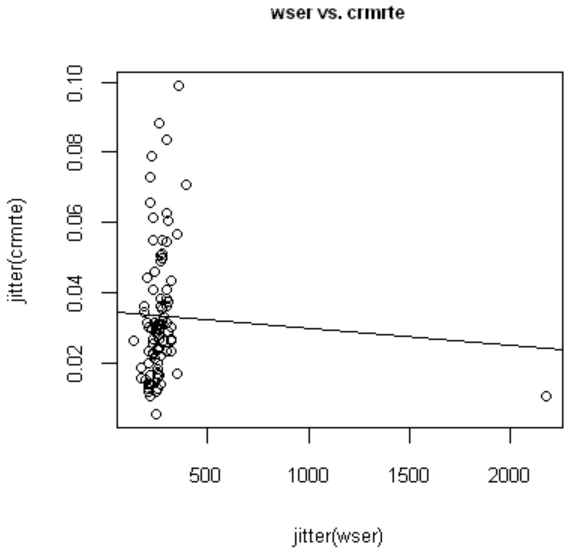
In [9]:
```
par(mfrow=c(2,2))
plot(jitter(taxpc), jitter(crmrte), main="taxpc vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8)
model7 = lm(crmrte ~ taxpc, data = crime_df)
abline(model7)
plot(model7, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.
8)
cat("Note that crmte appears to increase as taxpc increases. This is unexpected, as wealthier areas are not typically assoc
iated with higher crime. There is one data point that is high leverage, but not influential (Cook's distance < 1). Scatter
 does not clearly demonstrate homoskedasticity (residuals spread) or heteroskedasticity (few data points prior to spread)."
)
```

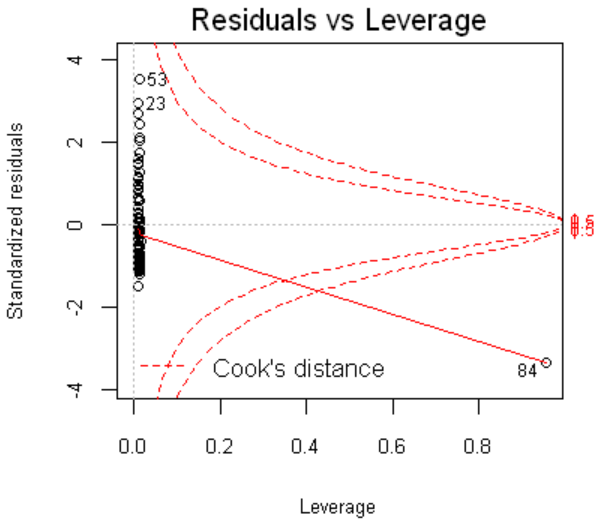
```
Note that crmte appears to increase as taxpc increases. This is unexpected, as wealthier areas are not typically associated
with higher crime. There is one data point that is high leverage, but not influential (Cook's distance < 1). Scatter does n
ot clearly demonstrate homoskedasticity (residuals spread) or heteroskedasticity (few data points prior to spread).
```



In [10]:
```
par(mfrow=c(2,2))
plot(jitter(wser), jitter(crmrte), main="wser vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8)
model16 = lm(crmrte ~ wser, data = crime_df)
abline(model16)
plot(model16, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=
0.8)
cat("Note that crmte relationship to wser is not clear due to a highly influentual (Cook's distance > 1) point. wser is onl
y included in EDA as it will recieve tranform treatment for outlier prior to use in model 3.")
```

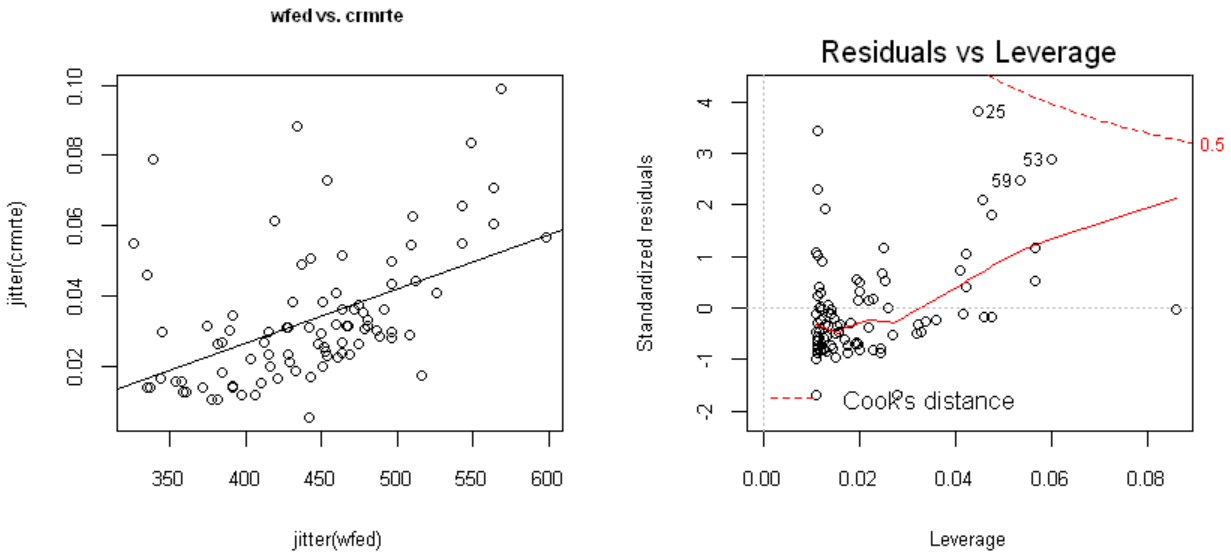
```
Note that crmte relationship to wser is not clear due to a highly influentual (Cook's distance > 1) point. wser is only inc
luded in EDA as it will recieve tranform treatment for outlier prior to use in model 3.
```





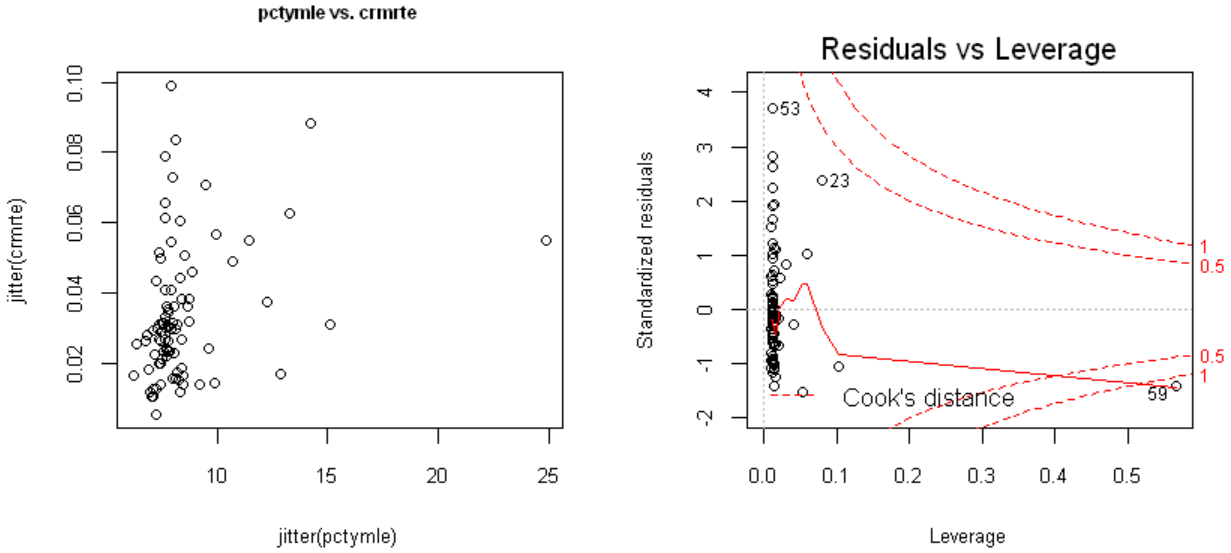
In [74]:
```
par(mfrow=c(2,2))
plot(jitter(wfed), jitter(crmrte), main="wfed vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8)
model18 = lm(crmrte ~ wfed, data = crime_df)
abline(model18)
plot(model18, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=
0.8)
cat("Note that crmte appears to increase as wfed increases. This is unexpected, as one does not associate federal employees
with local crime performance. There are several data that are high leverage, but not influential (Cook's distance < 1). Sca
tter does not clearly demonstrate homoskedasticity (residuals spread) or heteroskedasticity (residuals are consistently spr
ead).")
```

```
Note that crmte appears to increase as wfed increases. This is unexpected, as one does not associate federal employees with
local crime performance. There are several data that are high leverage, but not influential (Cook's distance < 1). Scatter
does not clearly demonstrate homoskedasticity (residuals spread) or heteroskedasticity (residuals are consistently spread).
```



In [11]:
```
par(mfrow=c(2,2))
plot(jitter(pctymle), jitter(crmrte), main="pctymle vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8)
model22 = lm(crmrte ~ pctymle, data = crime_df)
abline(model22)
plot(model22, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=
0.8)
cat("Note that crmte relationship to pctymle is unclear given the spread of results at any given value for pctymle. This is
unsurprising, yet not expected. There is one data point that is high leverage and influential (Cook's distance > 1). Scatte
r does not clearly demonstrate homoskedasticity (residuals spread) or heteroskedasticity (no trend to evaluate spread ove
r).")
```

```
Note that crmte relationship to pctymle is unclear given the spread of results at any given value for pctymle. This is unsu
rprising, yet not expected. There is one data point that is high leverage and influential (Cook's distance > 1). Scatter do
es not clearly demonstrate homoskedasticity (residuals spread) or heteroskedasticity (no trend to evaluate spread over).
```



2.3. Variable Analysis: Correlation w/ crime rate

Correlation analysis was used to study bivariate interaction between covariates, and between covariates and the response variable. This helps to highlight variables that will perform well in model and areas of potential multicollinearity.

As expected, covariates 'urban' and 'density' are highly correlated (0.82) and therefore would be expected to be a detriment to the parsimony of models 1 & 2 if both included. Similarly, many wage covariates are highly correlated.

Interestingly, covariates related to the justice system are generally not highly correlated, with the exception of 'polpc' and 'prbarr' at 0.43.

In [16]:
```
cat("Bivariate interaction between covariates")
round(cor(crime_df[, c('crmrte', 'prbarr', 'prbconv', 'prbpris', 'avgsen', 'polpc', 'density', 'taxpc', 'west', 'central',
'urban', 'pctmin80', 'wcon',
'wtuc', 'wtrd', 'wfir', 'wser', 'wmfg', 'wfed', 'wsta', 'wloc', 'mix', 'pctymle')], method="pearson"),2)
```

```
Bivariate interaction between covariates
```

	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	central	urban	pctmin80	wcon	wtuc	wtrd	wfir	wser	wr
crmrte	1.00	-0.39	-0.39	0.05	0.03	0.17	0.73	0.45	-0.35	0.17	0.62	0.19	0.39	0.23	0.41	0.33	-0.05	0.3
prbarr	-0.39	1.00	-0.06	0.05	0.18	0.43	-0.30	-0.14	0.17	-0.17	-0.21	0.05	-0.25	-0.07	-0.10	-0.17	-0.13	-0.
prbconv	-0.39	-0.06	1.00	0.01	0.15	0.17	-0.23	-0.13	0.05	-0.05	-0.20	0.06	-0.12	-0.01	-0.13	0.03	0.46	0.0
prbpris	0.05	0.05	0.01	1.00	-0.10	0.05	0.08	-0.09	-0.04	0.16	0.05	0.10	-0.06	0.13	0.14	0.03	0.04	0.0
avgsen	0.03	0.18	0.15	-0.10	1.00	0.49	0.08	0.10	0.08	-0.14	0.15	-0.15	-0.03	0.21	0.08	0.16	-0.15	0.1
polpc	0.17	0.43	0.17	0.05	0.49	1.00	0.16	0.28	0.14	-0.04	0.16	-0.16	-0.02	0.17	0.11	0.19	-0.02	0.2
density	0.73	-0.30	-0.23	0.08	0.08	0.16	1.00	0.32	-0.14	0.36	0.82	-0.07	0.45	0.33	0.58	0.54	0.04	0.4
taxpc	0.45	-0.14	-0.13	-0.09	0.10	0.28	0.32	1.00	-0.19	0.04	0.35	-0.02	0.26	0.16	0.17	0.12	0.07	0.2
west	-0.35	0.17	0.05	-0.04	0.08	0.14	-0.14	-0.19	1.00	-0.40	-0.09	-0.64	-0.18	0.10	-0.13	-0.02	-0.06	-0.
central	0.17	-0.17	-0.05	0.16	-0.14	-0.04	0.36	0.04	-0.40	1.00	0.16	-0.04	0.40	0.18	0.37	0.28	0.19	0.1
urban	0.62	-0.21	-0.20	0.05	0.15	0.16	0.82	0.35	-0.09	0.16	1.00	0.02	0.32	0.22	0.42	0.40	0.06	0.4
pctmin80	0.19	0.05	0.06	0.10	-0.15	-0.16	-0.07	-0.02	-0.64	-0.04	0.02	1.00	-0.11	-0.20	-0.08	-0.09	0.19	-0.
wcon	0.39	-0.25	-0.12	-0.06	-0.03	-0.02	0.45	0.26	-0.18	0.40	0.32	-0.11	1.00	0.41	0.56	0.49	-0.01	0.3
wtuc	0.23	-0.07	-0.01	0.13	0.21	0.17	0.33	0.16	0.10	0.18	0.22	-0.20	0.41	1.00	0.36	0.33	-0.02	0.4
wtrd	0.41	-0.10	-0.13	0.14	0.08	0.11	0.58	0.17	-0.13	0.37	0.42	-0.08	0.56	0.36	1.00	0.67	-0.02	0.3
wfir	0.33	-0.17	0.03	0.03	0.16	0.19	0.54	0.12	-0.02	0.28	0.40	-0.09	0.49	0.33	0.67	1.00	0.01	0.4
wser	-0.05	-0.13	0.46	0.04	-0.15	-0.02	0.04	0.07	-0.06	0.19	0.06	0.19	-0.01	-0.02	-0.02	0.01	1.00	0.0
wmfg	0.35	-0.15	0.02	0.01	0.12	0.27	0.44	0.26	-0.01	0.18	0.40	-0.11	0.35	0.46	0.36	0.49	0.01	1.0
wfed	0.49	-0.21	-0.06	0.09	0.14	0.16	0.59	0.06	-0.17	0.34	0.42	0.03	0.51	0.40	0.64	0.62	0.02	0.5
wsta	0.20	-0.16	-0.13	-0.03	0.13	0.05	0.22	-0.03	-0.08	0.09	0.30	0.10	-0.02	-0.16	0.00	0.24	0.04	0.0
wloc	0.35	-0.03	0.05	0.08	0.12	0.38	0.45	0.21	-0.09	0.32	0.33	-0.12	0.51	0.34	0.59	0.56	0.08	0.4
mix	-0.13	0.41	-0.30	0.12	-0.14	0.03	-0.14	-0.04	-0.01	-0.09	-0.06	0.20	-0.20	-0.25	-0.13	-0.21	-0.17	-0.
pctymle	0.29	-0.18	-0.16	-0.08	0.07	0.05	0.12	-0.09	-0.04	-0.10	0.09	-0.02	-0.02	-0.10	-0.11	0.01	-0.04	0.0

2.4 Variable Analysis: Determine over-leverage points

Over-leverage points degrade the ability of a covariate to fit a model.

Three variables found have influential data points (Cook's distance > 1)

- 'polpc' (Police per capita) - Data Point 51 (Row = 52)
- 'wser' (Service Industry Wage) - Data Point 84 (Row = 85)
- 'pctymle' (Percent Male) - Data Point 59 (Row = 60)

The decision criteria used for applying transformations to influential data points is that all 3 statements below are true for the tranformation:

- Effective to reduce Cook's distance to < 1
- Result is interpretable. This does not apply for model 3 which is not intended to be an interpretted specification.
- Will be used as a model covariate.

'polpc'

Tranform: No
Police levels are difficult to interpret as high police presence could logically be both a cause of a low crime rate as well as an effect of high crime rate. Therefore, as techniques are not yet learned to separate causal from acillary effects, 'polpc' was omitted from modelling.

'wser'

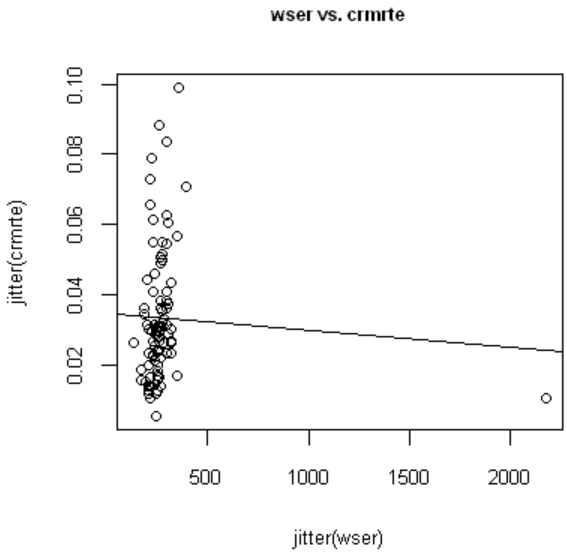
Tranform: Yes
Transforming 'wser' by taking the inverse brings the influential point < 1 Cook's distance. Note that while applying an inverse transform lends to difficult interpretation, this transformation is acceptable as this covariate will only be used for model 3, which is an exploratory model containing almost all variables in the data set.

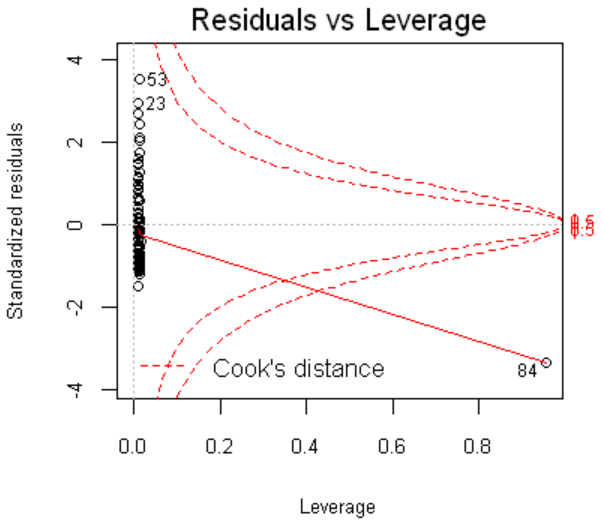
'pctymle'

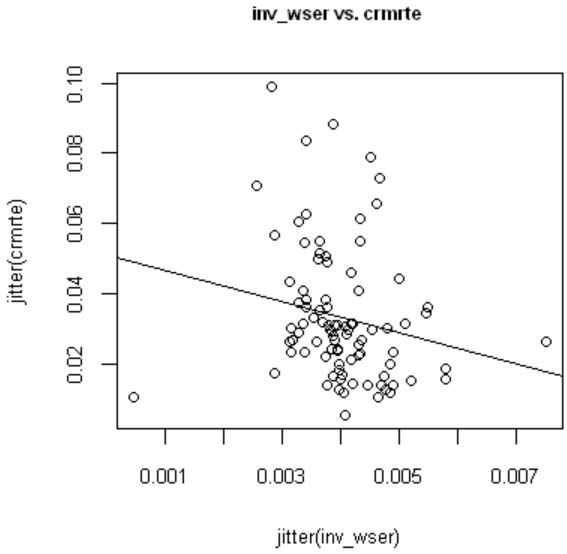
Tranform: Yes
Transforming 'pctymle' by taking the log brought the influential point < 1 Cook's distance. This transformed variable is interpretable, as log translates to the outcome variable changing with a percentage change of the covariate rather than unit change of the covariate.

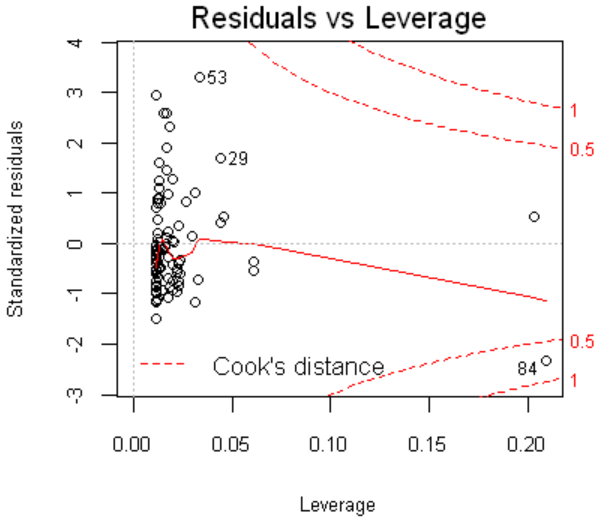
In [12]:

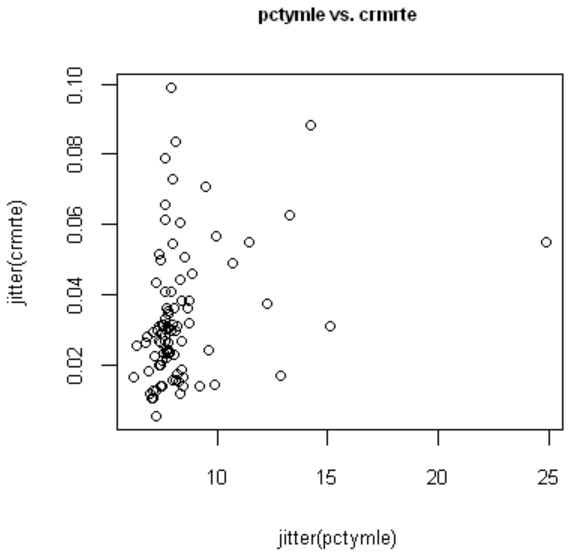
```
par(mfrow=c(2,2))

#Non-transformed: wser
plot(jitter(wser), jitter(crmrte), main="wser vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8)
model16 = lm(crmrte ~ wser, data = crime_df)
abline(model16)
plot(model16, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=
0.8)

#Transformed: inv_wser
inv_wser = 1/(wser)
plot(jitter(inv_wser), jitter(crmrte), main="inv_wser vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8)
model16 = lm(crmrte ~ inv_wser, data = crime_df)
abline(model16)
plot(model16, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=
0.8)

#Non-transformed: pct_ymle
plot(jitter(pctymle), jitter(crmrte), main="pctymle vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8)
model22 = lm(crmrte ~ pctymle, data = crime_df)
abline(model22)
plot(model22, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=
0.8)

#Transformed: log(pct_ymle)
log_pctymle = log(pctymle)
plot(jitter(log_pctymle), jitter(crmrte), main="log(pctymle) vs. crmrte", cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8
)
model22 = lm(crmrte ~ log_pctymle, data = crime_df)
abline(model22)
plot(model22, which = 5, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=0.8, cex.lab=.8, cex.axis=.8, cex.main=.7, cex.sub=
0.8)
```

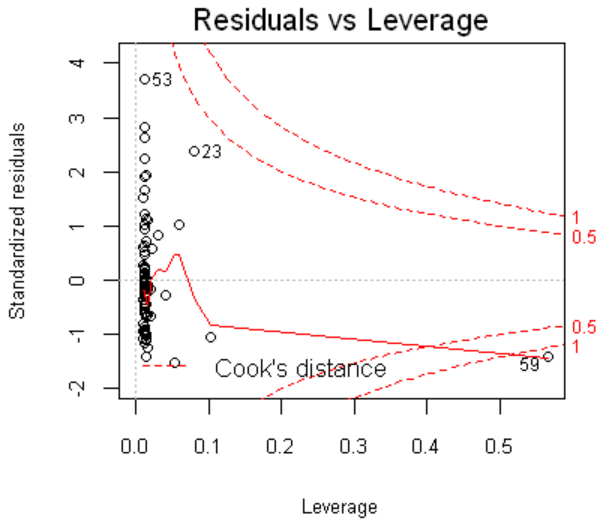

wser vs. crmrte
jitter(crmrte)
0.02
0.04
0.06
0.08
0.10
500
1000
1500
2000
jitter(wser)

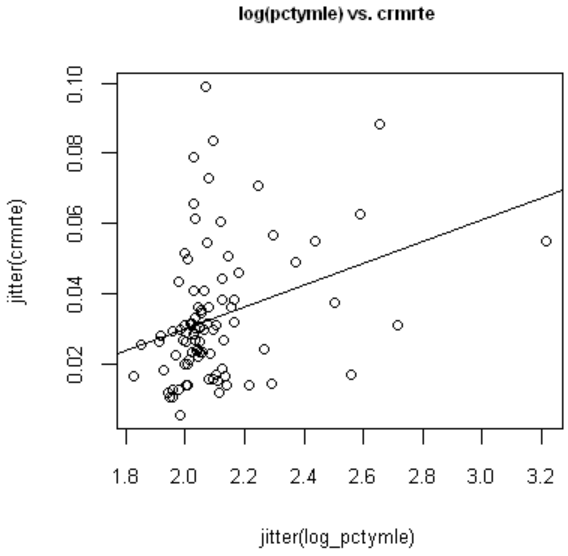

Residuals vs Leverage
Standardized residuals
4
2
0
-2
-4
53
23
84
Cook's distance
0.5
0.5
0.0
0.2
0.4
0.6
0.8
Leverage

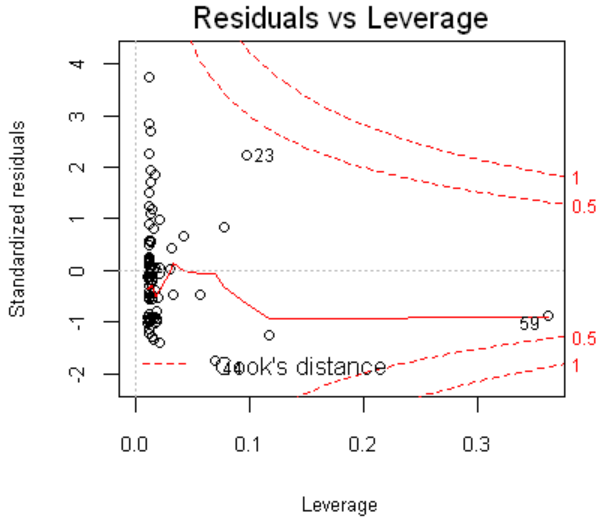

inv_wser vs. crmrte
jitter(crmrte)
0.02
0.04
0.06
0.08
0.10
0.001
0.003
0.005
0.007
jitter(inv_wser)


Residuals vs Leverage
Standardized residuals
4
3
2
1
0
-1
-2
-3
53
29
84
Cook's distance
1
0.5
0.5
1
0.00
0.05
0.10
0.15
0.20
Leverage


pctymle vs. crmrte
jitter(crmrte)
jitter(pctymle)
0.02
0.04
0.06
0.08
0.10
10
15
20
25


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
53
23
59
1
0.5
0.5
1
-2
-1
0
1
2
3
4
0.0
0.1
0.2
0.3
0.4
0.5


log(pctymle) vs. crmrte
jitter(crmrte)
jitter(log_pctymle)
0.02
0.04
0.06
0.08
0.10
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.2


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
23
59
1
0.5
0.5
1
-2
-1
0
1
2
3
4
0.0
0.1
0.2
0.3

2.5 Variable Analysis: Empirical data exploration

Variables found to be questionable are discussed below.

polpc (Police per capita)

Notice on the lower right-hand side of the scatterplot, Madison county (county 115) has an extremely high crime rate and yet a very high police presence. Police presence is positively corrrelated with crime. This is likely not because high police presence causes more crime. Instead, crime likely creates demand for police.

The anomalous character of Madison county is likely due to having three police departments.

- Hot Springs Police Dept
- Madison County Sheriff
- Marshall Police Dept

This is unusual for a county of its size, but likely the result of interplay of county and municipal jurisdictions. This county exceeds the next closest county by 40 percentage-points.

prbconv (Probability of conviction)

Notice on the lower right-hand side of the scatterplot, 10 data points are > 1 (100%), indicating higher convicted person count than arrested person count. This suggests that counties could be charging individuals with multiple crimes for a single arrest. Though these practices could vary between counties, no data elimination treatment is recommended.

density (People per square mile)

Density appears to be univerally too low by a factor of ~500. For example, Wake County (County 83) had density of ~450 (http://www.wakegov.com/data/bythenumbers/PublishingImages/WC%20Population%20Density.jpg) in 1987, whereas dataset shows density of ~0.8. However, as scale of numbers is consistent inter-county, variable will be used as-is. Regardless, recommendation is that data be examined post-analysis for verity.

wser (Weekly wage, service industry)

Only one data point, Warren County (county 185), is an outlier at 2177. This county exceeds the next closest county by a factor of 5. Salary is ~ $ 100k, which is \$ 242k in 2019 dollars. This is not an impossibility, however, therefore no grounds for removing the datapoint. This datapoint did exert influence prompting log transformation.

3.0 The Model Building Process

3.1 - Attempted Transformations

Most attempted transformations resulted in variable interpretation problems. Additionally, often model R^2 decreased or AIC increased.

The following three transformations were implemented:

1. The probability of punishment -- Multiplying 'prbarr' by 'prbconv'. If used to replace 'prbarr' or 'prbconv', R^2 decreased (as expected) and AIC score increased. Low AIC ~ better parsimony. Instead, both variables were retained and interaction term was added to model 2. This allowed the model to benefit from the contribution of either or both variables in the event of incomplete moderation, giving it more explanatory power.
2. wser -- As noted in section 2.4, this variable was put through an inverse tranformation in order to mitigate an inlfuential (Cook's distance > 1) point.
3. pctymle -- As noted in section 2.4, this variable was tranformed with a log function in order to mitigate an influential (Cook's distance > 1) point.

3.2 - Assumptions in Building the Model

The following variables were not considered for model use:

1. polpc - Discussion in section 2.4 notes difficulty separating causal and ancillary effects.
2. mix - Discussion in section 2.0 notes 'mix' as an alternative attribute of crime which is not pursuant to Research Question.
3. county - Omitted as it is just the State-county-level FIPS number. It unhelpful as a categorical variable. Because each record represents a single county, usage would require 91 individual covariants (not recommended). Higher level groupings exist for geography such as central and west.

3.3 - Determining the models

The following models were chosen:

1. $model\_1 : \beta_0 + \beta_1 density + \beta_2 taxpc + u$
2. $model\_2 : \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + u$
3. $model\_3 : \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + \beta_6 prbpris + \beta_7 avgsen + \beta_8 west + \beta_9 central + \beta_{10} pctmin80 + \beta_{11} wcon + \beta_{12} wtuc + \beta_{13} wtrd + \beta_{14} wfir + \beta_{15} wser + \beta_{16} wmfg + \beta_{17} wfed + \beta_{18} wsta + \beta_{19} wloc + \beta_{20} log(pctymle) + u$

Model 1 (Base Model): density, taxpc

Based on the research question, density and taxpc are the best proxies for population density and regional wealth profile.

$$model\_1 : \beta_0 + \beta_1 density + \beta_2 taxpc + u$$

Regarding the choice to of 'urban' vs. 'density' - Multicollinearity is high between 'urban' and 'density' at `cor(urban, density) = 0.82` . This prompts a choice to be made to use one, not both, of these variables as covariates. 'urban' is a categorical simplification of 'density'. Between the choices 'urban'/'density', 'density' was chosen for the modeling benefit provided by its continuous, cardinal values vs. binary values of 'urban'.

*Model 2: Model 1 + prbconv, prbarr, prbconv*prbarr*

$$model\_2 : \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + u$$

Additional covariates for model 2 (density, taxpc, prbconv, prbarr, prbconv*prbarr) were chosen for characterization power which will aid the policy component of the research question. Furthermore, EDA showed those variables that were understood to not correlate with the explanatory variables of the base model while maintaining an attendant associative "effect" on the crime rate. The two variables that had a notable correlation with the crime rate while also having some correlative impact on the base variables were the two deterrence variables in the dataset -- namely, the probability of arrest (prbarr) and probability of conviction (prbcon). This also matches intuitive understandings that not only does deterrence reduce the crime rate, but the propensity, and perhaps the ability, to hire police, investigators, and judicial officials -- who in turn pursue arrests and convictions -- is highly dependent on taxpayer funding.

Initially attempted as a transformation, the interaction term prbconv*prbarr works better as an interaction term (moderation analysis) so as to allow retaining the main effects of the two constituent variables. This interaction term works well test the influence each component variable has on one another other when predicting crime.

Regarding 'avgsen', this variable has a positive, albeit marginal, correlation with 'crmrte'. This suggests exclusion of this variable from Models 1 and 2. 'avgsen' will have a negative coefficient when other variables are included in Model 3, but the effect is insignificant.

Model 3: Model 2 + prbpris, avgsen, west, central, pctmin80, wcon, wtuc, wtrd, wfir, inv_wser, wmfg, wfed, wsta, wloc, log(pctymle)

All other variables except polpc, urban, mix, and county are included. The reasons for the excluded variables are explained in above sections.

3.4 - Evaluating Potential Models

First we examined the six Classical Linear Model (CLM) assumptions.

In [13]:
```
model_1 <- lm(crmrte ~ density + taxpc)
model_2 <- lm(crmrte ~ density + taxpc + prbarr*prbconv)
model_3 <- lm(crmrte ~ density + taxpc + prbarr*prbconv + log(pctymle) + prbpris + avgsen + west + central + pctmin80 +
              wcon + wtuc + wtrd + wfir + inv_wser + wmfg + wfed + wsta + wloc)
```

CLM1 (Gauss-Markov 1): Linear in Parameters

All the models - Model 1, Model 2, Model 3 are linear models in the form of: $model\_1 : \beta_0 + \beta_1 x1 + \beta_2 x2 + \ldots + \beta_k xk + u$

CLM2 (Gauss-Markov 2): Random Sampling

For all the models, data used is assumed to be a random sample of data, independent and identically distributed (iid).

CLM3 (Gauss-Markov 3): Sample Variation in the Explanatory Variable

The variance inflation factor (VIF) test is employed to determine severity of multicollinearity between covariates within a model. VIF test measures the ratio of the variance in a model by the variance of a model with one term alone. VIF > 4 may indicate high multicollinearity. VIF > 10 indicate large standard errors that could affect the model's prediction ability.

For Model 1, the VIF test shows 1.11 (under 10). Hence, we do not have evidence of significant multicolinearity. The VOF values are as follows:

In [14]:
```
# VIF test for CLM 3 - multicolinearity
round(vif(model_1),1)
```

density	1.1
taxpc	1.1

For Model 2, the VIF test shows the following:

In [15]:
```
# VIF test for CLM 3 - multicolinearity
round(vif(model_2),1)
```

density	1.4
taxpc	1.1
prbarr	4.4
prbconv	3.5
prbarr:prbconv	5.9

All VIF < 10, although two > 4. Since none of the data is > 10, there should be no affect on ability to model the data.

For Model 3, the VIF test shows the following:

In [16]:
```
# VIF test for CLM 3 - multicolinearity
round(vif(model_3),1)
```

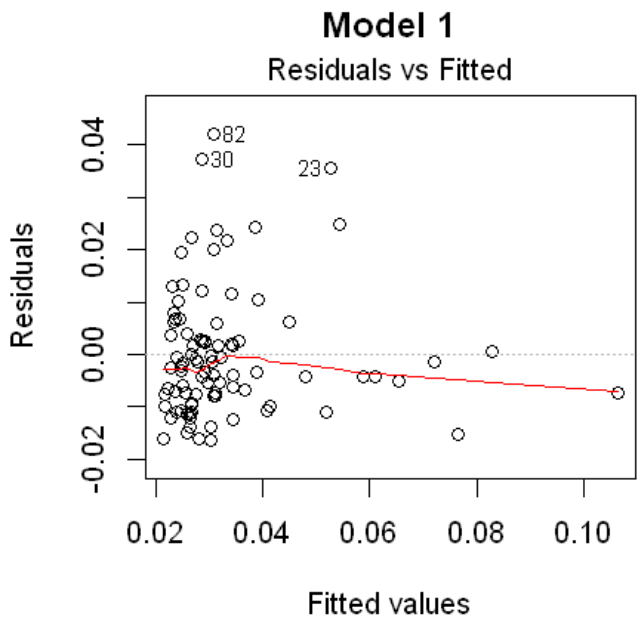
density	2.4
taxpc	1.6
prbarr	5.5
prbconv	4.5
log(pctymle)	1.4
prbpris	1.2
avgsen	1.6
west	3.4
central	2.1
pctmin80	2.8
wcon	2.2
wtuc	1.7
wtrd	3.1
wfir	2.7
inv_wser	1.9
wmfg	1.9
wfed	2.8
wsta	1.6
wloc	2.5
prbarr:prbconv	8.2

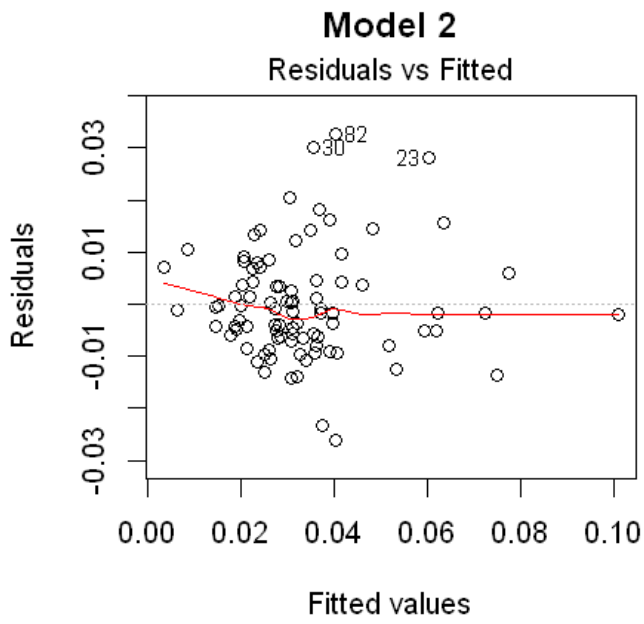
Note that for model 3, prbarr, prbconv, and prbarr*prbconv (the indicator variable) have a 4 < VIF < 10. This is not unexpected because there will likely be multicollinearity between an interaction term and its constituent terms; removing the interaction term from the VIF analysis reduces all values to < 5. In any case, all VIF values are < 10 even with the interaction term included, so there are no multicollinearity issues.

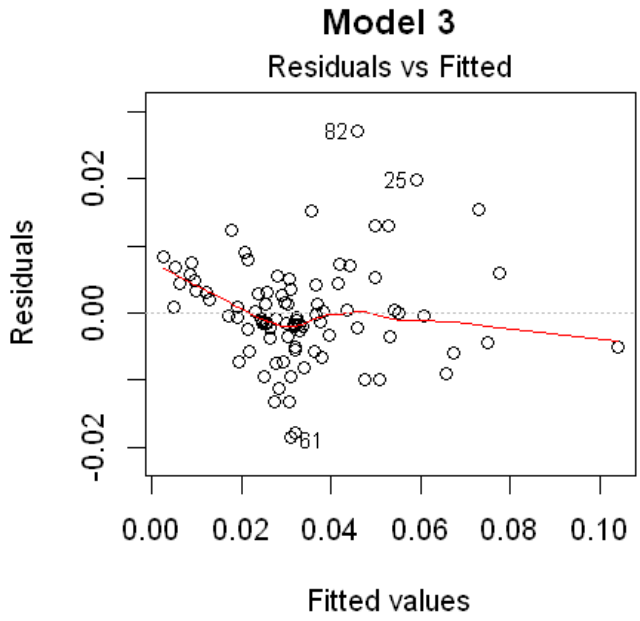
CLM4 (Gauss-Markov 4): Zero Conditional Mean

For all the models - Model 1, Model 2, Model 3 - the Plot of Residuals vs Fitted shows a reasonably horizontal line for conditional mean for errors centered on zero.

In [17]:
```
#Checking for CLM 4 - zero conditional error mean
par(mfrow=c(1,2))
options(repr.plot.width=8, repr.plot.height=4)
plot(model_1, which=1, main = "Model 1")
plot(model_2, which=1, main = "Model 2")
plot(model_3, which=1, main = "Model 3")
```







CLM5 (Gauss-Markov 5): Homoskedasticity

The Breusch-Pagan test for Model 1 shows a p value of 0.9 and Model 3 shows p value of 0.1, which are greater than 0.05 and hence not significant. The null hypothesis for a Breusch-Pagan test is that the model has Homoskedasticity. Therefore, p-values should be bigger than .05 or else we will reject the null hypothesis (ie. indicate heteroskedasticity). Also, the residuals vs fitted plot (above) does not show a pattern of residual spread change for different values of the input variable. Inspecting the plots visually, there may be evidence of heteroskedasticity toward the highest values of X (or the fitted values which is a linear combination of the x's) for Model 1 and the lowest and highest values of the x's for Model 3. This may also be due to the fact that there appears to be less data points at these extremes.

In [38]:
```
#checking for CLM 5 - nonconstant variance test
#install.packages("lmtest",repos = "http://cran.us.r-project.org")
library(lmtest)
cat("Model 1: ")
bptest(model_1)
cat("Model 3: ")
bptest(model_3)
```

```
Model 1:

        studentized Breusch-Pagan test

data:  model_1
BP = 0.1195, df = 2, p-value = 0.942

Model 3:

        studentized Breusch-Pagan test

data:  model_3
BP = 25.937, df = 20, p-value = 0.1679
```

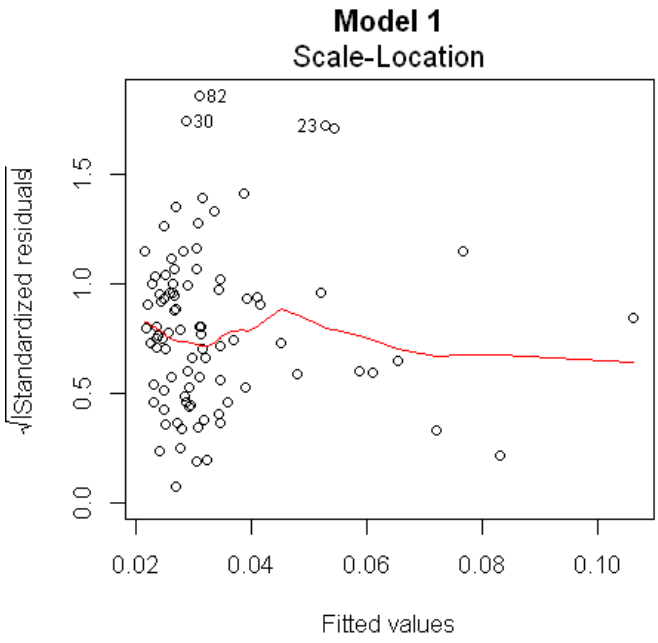
The Breusch-Pagan (BP) test for Model 2 shows a p value of 0.02. Though a significant result, examination of the residual vs fitted value plot does not reveal a clear pattern of variance scatter. While, once again, the plot shows that it is possible that heteroskedasticity exists at extreme values of the x's (low and high), the fewer datapoints provide a lack of clarity on this point. Although we could perhaps conclude that the homoskedasticity assumption has not been violated, we will choose to use heteroskedasticity-robust standard errors.

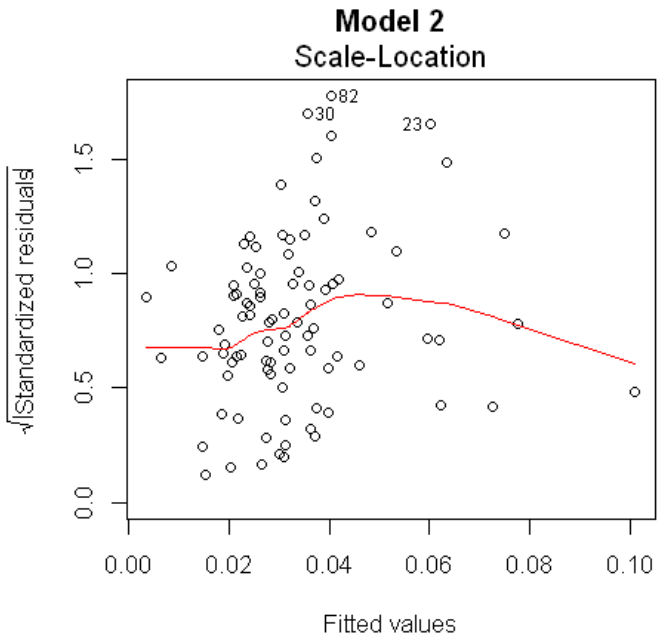
In [19]:
```
cat("Model 2: ")
bptest(model_2)
```

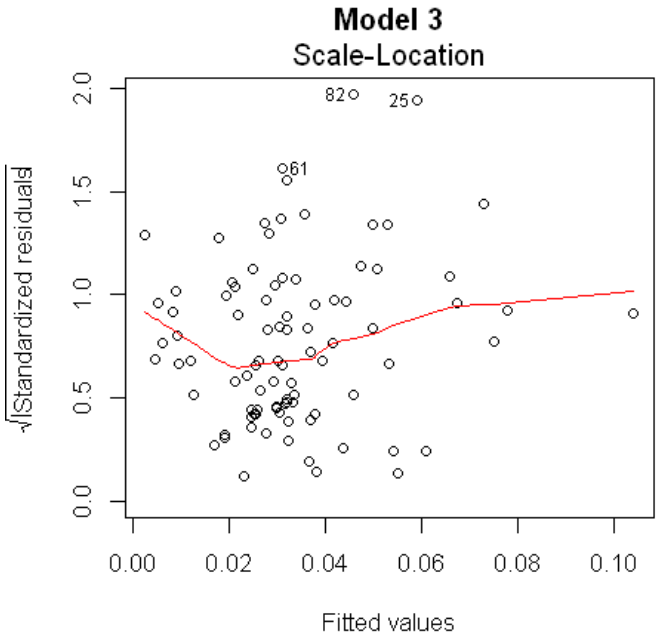
```
Model 2:

        studentized Breusch-Pagan test

data:  model_2
BP = 13.625, df = 5, p-value = 0.01818
```

Scale-location plots are also a good way of detecting problems with homoskedasticity. Here, unlike the above Residuals vs Fitted plot, the y-axis is replaced with the square-root of the absolute value of the residual. Tasking the absolute value means that X's with greater variance will show up as having a more positive value. Additionally, the square-root spreads the points out so that they do not all cluster toward 0. Ideally, we should see a horizontal smoothing curve with equally spread out points. Overall, these plots appear to demonstrate homoskedasticity, but we will use heteroskedasticity-robust tools in an abudance of caution.

In [20]:
```
par(mfrow=c(2,2))
options(repr.plot.width=8, repr.plot.height=8)
plot(model_1, which=3, main = "Model 1")
plot(model_2, which=3, main = "Model 2")
plot(model_3, which=3, main = "Model 3")
```







CLM 6:

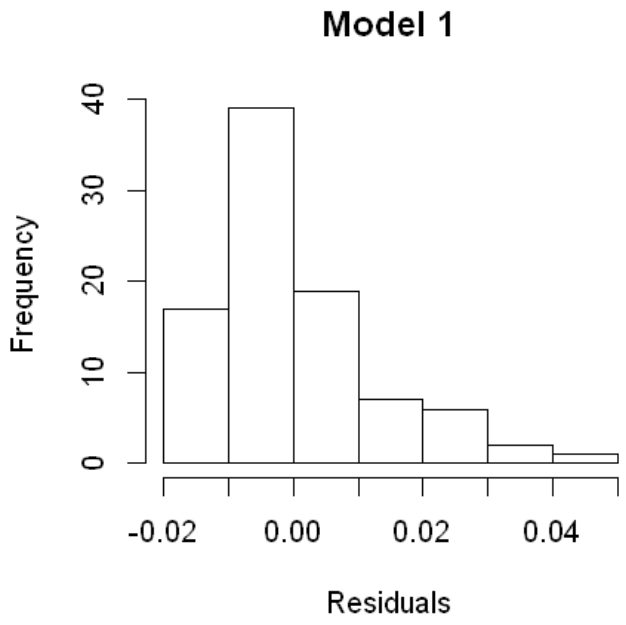
For Model 1, a Shapiro-Wilk analysis of model 1 demonstrates marked departure from normality of residuals. It has been suggested that with a large enough sample size -- thanks to a version of the central limit thereom -- non-normality of residuals can be overcome with a large enough sample size. With 91 samples, which is greater than the rule of thumb of 30, asymptotics should prevail and our estimators should have a normal distribution.

In [21]:
```
shapiro.test(model_1$residuals)
```

```
        Shapiro-Wilk normality test

data:  model_1$residuals
W = 0.88991, p-value = 1.315e-06
```

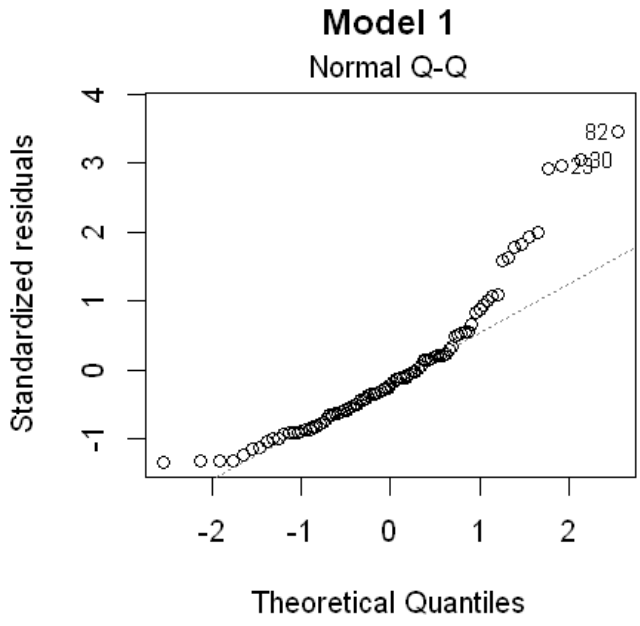
A histogram of the residuals show a positive skew.

In [22]:
```
par(mfrow=c(1,2))
options(repr.plot.width=8, repr.plot.height=4)
hist(model_1$residuals, main = "Model 1", xlab="Residuals")
```



Considering that the Q-Q plot and histogram indicate a skewed distribution with a longer right tail, however, it is likely that this model tends to under-predict crime (positive Y - $\hat{Y}$) more than it over-predicts it.

In [23]:
```
par(mfrow=c(1,2))
options(repr.plot.width=8, repr.plot.height=4)
plot(model_1, which=2, main = "Model 1")
```



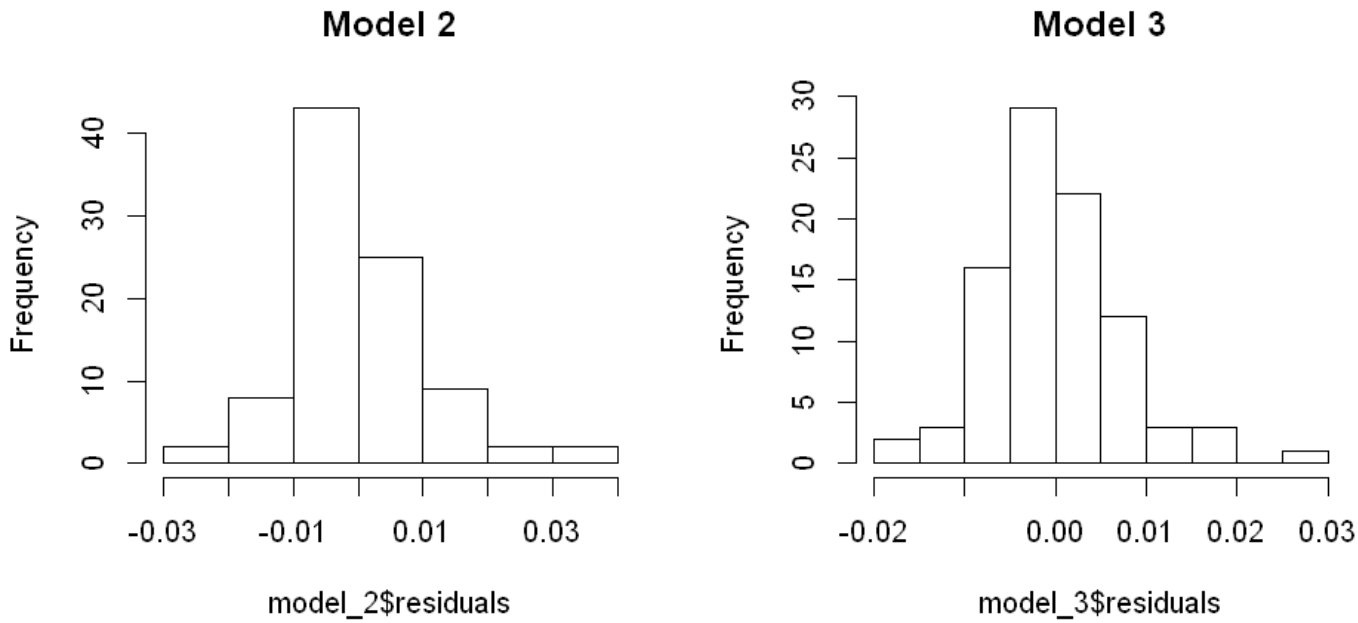
For Model 2 and Model 3, though the Shapiro-Wilk result suggests departure from normality, the Q-Q plot and histogram of residuals show little deviation. The Shapiro-Wilk test is known to be oversensitive, especially for data sets around 100 or larger. Considering our sample size of 91 and how the residual plots look, it seems reasonable to assume sufficient normality for regression.

In [24]:
```
cat("Model 2")
shapiro.test(model_2$residuals)
cat("Model 3")
shapiro.test(model_3$residuals)
```

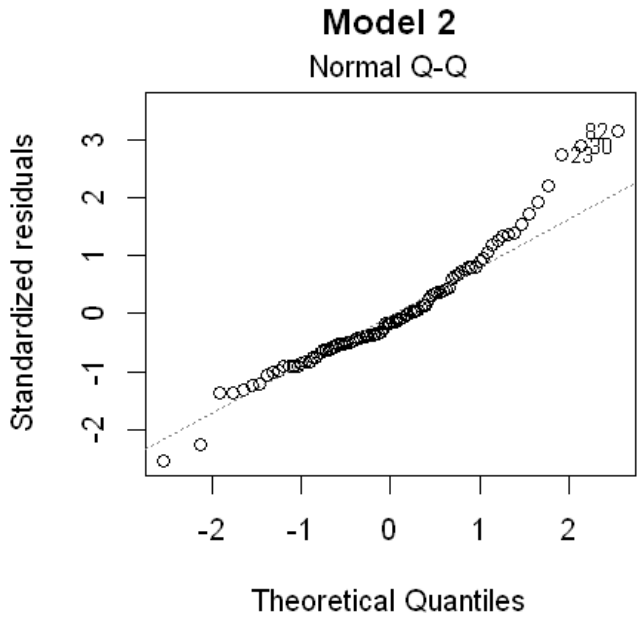
```
Model 2

        Shapiro-Wilk normality test

data:  model_2$residuals
W = 0.95628, p-value = 0.003911

Model 3

        Shapiro-Wilk normality test

data:  model_3$residuals
W = 0.96554, p-value = 0.01641
```

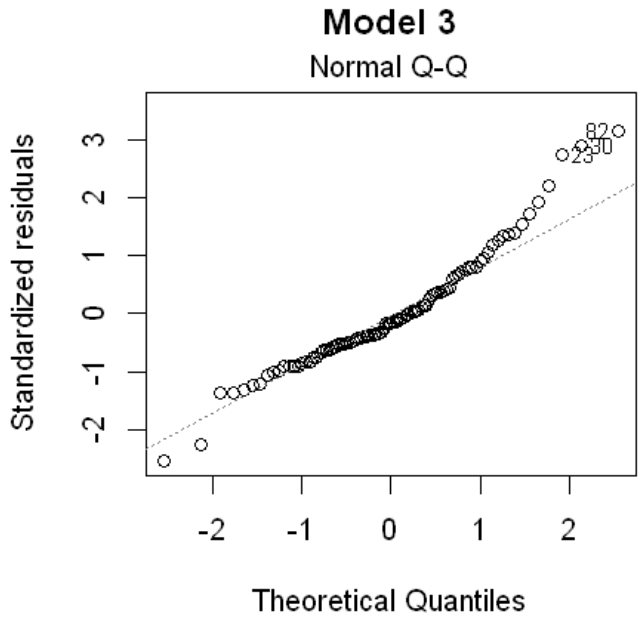
A histogram of the residuals shows a fairly normal distribution.

In [25]:
```
par(mfrow=c(1,2))
options(repr.plot.width=8, repr.plot.height=4)
hist(model_2$residuals, main = "Model 2")
hist(model_3$residuals, main = "Model 3")
```



In [26]:
```
par(mfrow=c(1,2))
options(repr.plot.width=8, repr.plot.height=4)
plot(model_2, which=2, main = "Model 2")
plot(model_2, which=2, main = "Model 3")
```





4.0 Regression Models: Base Model

We start with the following model.

$$model\_0 : \beta_0 + u$$

Model 0 above uses mean crime as a point estimate of crime in all cases.

For Model 1, Model 2 and Model 3:

- Null Hypothesis $H0$: The input variables of the model are not associated with crime rate; the model does not predict crime better than model 0 (mean crime rate as an estimator).
- Alternative Hypothesis HA: The input variables of the model are associated with crime rate; the model has some non-zero coefficients and predicts crime better than simply using the mean crime value.

$$model\_1 : \beta_0 + \beta_1 density + \beta_2 taxpc + u$$

We can evalute the effects of explanatory variables on the crime rate.

This model attempts to explain crime rate by looking at two variables that are theoretically expected to increase crime. These are the following:

- taxpc - high tax revenue per capita should result in higher crime
- density - high density should be associated with higher crime

This seems to match intuition:

- The high tax revenue per capita -- The high revenue per capita may increase opportunities for crime, since the location the crime was committed in was a location of greater wealth. Note that the crime rate is not necessarily based on the residency of the perpetrator, but on the county in which the criminal activity occurred. This is the county that would have jurisdiction to conduct the arrest.
- Density -- High density areas provide opportunities for more crime.

4.1 Regression Model: Second Model

$$model\_2 : \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + +\beta_5 prbconv * prbarr + u$$

For the second model, we will add variables that deal with deterrence:

- prbarr - a high probability of arrest may reduce crime
- prbconv - a high probability of conviction may reduce crime
- prbconv*prbarr - the interaction between probability of arrest and probability of conviction

These are negatively correlated with the crime rate. The higher the probability of arrest and conviction, the more that offenders, who might otherwise commit crimes, choose not to, for fear of being caught and convicted.

Additionally, arrests without conviction are expected to be less effective than arrests with convictions. The interaction between the two main variables is of interest here.

4.2 Regression Model: Third Model

$$model\_3 : \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + \beta_6 prbpris + \beta_7 avgsen + \beta_8 west + \beta_9 central + \beta_{10} pctmin80 + \beta_{11} wcon + \beta_{12} wtuc + \beta_{13} wtrd + \beta_{14} wfir + \beta_{15} inv\_wser + \beta_{16} wmfg + \beta_{17} wfed + \beta_{18} wsta + \beta_{19} wloc + \beta_{20} log(pctymle) u$$

For the final third model, all variables are added (save the few that have been determined to be inadmissible) into the model to see how robust key variables are to the introduction of other variables. Transformations (log of pctymle and inverse of wser) will only come into play here.

4.3 The Regression Table

Note the Regression Table. Coefficients for all key variables remain fairly constant when all variables are added into the analysis.

AIC does not change with addition of a variable but rather changes due to composition of predictors; this statistic is a measure of quality of model fit and is therefore commonly used for model selection. R^2 is percentage of variance that can be explained by a model. Models with higher R^2 value are better as they have stronger power to explain variance in the outcome variable.

In the table below, Model 2 has a lower AIC (-561.1) than Model 1 (-537.6). Model 2 is a better fit than Model 1 and also has noticeably better adjusted-R^2 (0.68 vs. 0.57). This means 68% of the variance in crime can be explained by Model 2.

Similarly, Model 3 performs better than Model 2 with an adjusted-R^2 of 0.80 (highest) while also demonstrating a more negative AIC.

All the variables chosen for Model 1 and Model 2 significantly contribute to the models (p-value < 0.05).

Model 3 contains additional significant contributors - pctmin80, pctymle, and wfed. The superior AIC and adjusted-R^2 of this model suggest that these three variables should also be considered when attempting to predict crime.

In [27]:
```
library(sandwich)
cov_1 <- vcovHC(model_1, type = "HC")
robust.se_1 <- sqrt(diag(cov_1))
cov_2 <- vcovHC(model_2, type = "HC")
robust.se_2 <- sqrt(diag(cov_2))
cov_3 <- vcovHC(model_3, type = "HC")
robust.se_3 <- sqrt(diag(cov_3))

library(stargazer)
stargazer(model_1, model_2, model_3, type = "text",
          #report = "vc", # Don't report errors, since we haven't covered them
          title = "Linear Models",
          se=list(NULL, robust.se_1, robust.se_2, robust.se_3),
          #keep.stat = c("rsq", "n"),
          #omit.table.layout = "n",
          add.lines=list(c("AIC", round(AIC(model_1),1), round(AIC(model_2),1), round(AIC(model_3),1)))
         ) # Omit more output related to errors
```

Linear Models

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
density	0.008***	0.006***	0.005***
	(0.001)	(0.001)	(0.001)
taxpc	0.0003***	0.0003*	0.0004***
	(0.0001)	(0.0002)	(0.0001)
prbarr		-0.071	-0.076***
			(0.018)
prbconv		-0.027	-0.026***
			(0.006)
log(pctymle)			0.020
prbpris			0.002
avgsen			-0.0001
west			0.001
central			-0.003
pctmin80			0.0003
wcon			0.00002
wtuc			0.00000
wtrd			0.00001
wfir			-0.00004
inv_wser			2.643
wmfg			-0.00001
wfed			0.0001
wsta			-0.00003
wloc			0.0001
prbarr:prbconv		0.044	0.050***
			(0.012)
Constant	0.009**	0.042***	-0.048***
	(0.004)	(0.006)	(0.009)
AIC	-537.6	-561.1	-591
Observations	91	91	91
R2	0.584	0.699	0.844
Adjusted R2	0.574	0.681	0.799
Residual Std. Error	0.012 (df = 88)	0.011 (df = 85)	0.008 (df = 70)
F Statistic	61.644*** (df = 2; 88)	39.454*** (df = 5; 85)	18.943*** (df = 20; 70)

Note: *p<0.1; **p<0.05; ***p<0.01

4.4 Potential Model 4

Based on Model 3's highest performance, we suggest Model 4, which eliminates model 3's insignificant terms.

$$model\_4 : \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + \beta_6 pctmin80 + \beta_7 wfed + \beta_8 log(pctymle) + u$$

In [28]:
```
model_4 <- lm(crmrte ~ density + taxpc + prbarr*prbconv + log(pctymle) + pctmin80 + wfed)
```

CLM1 (Gauss-Markov 1):

Model 4 is a linear models in the form of: $model\_4 : \beta_0 + \beta_1 x1 + \beta_2 x2 + \ldots + \beta_k xk + u$

CLM2 (Gauss-Markov 2):

Data used is assumed to be random sample of data, independent and identically distributed (iid).

CLM3 (Gauss-Markov 3):

There is evidence of multicollinearity between an interaction term and its constituent terms. As mentioned above, this is expected; removing the interaction term from the VIF analysis reduces all values to less than 5. In any case, all VIF values are smaller than 10 even with the interaction term included, so we can conclude that there are no multicollinearity issues.

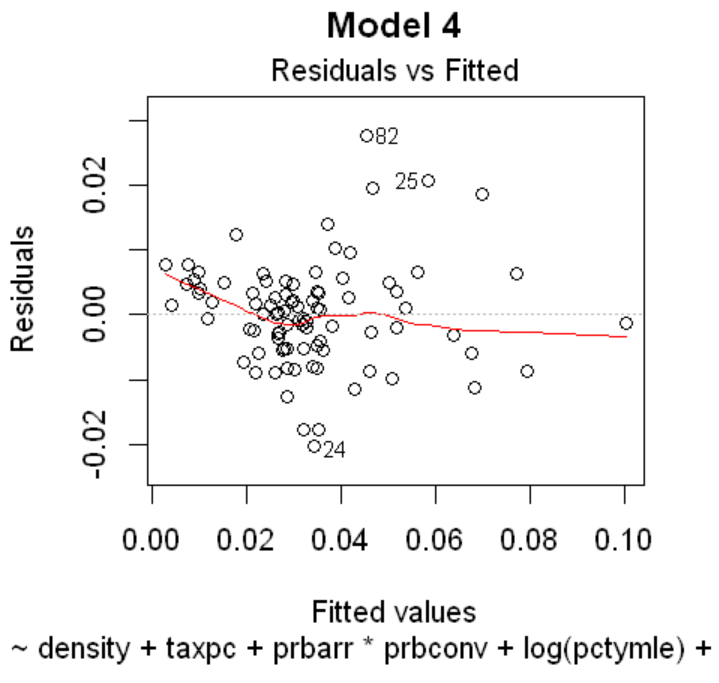
In [29]:
```
round(vif(model_4),1)
```

density	2
taxpc	1.2
prbarr	4.8
prbconv	3.8
log(pctymle)	1.2
pctmin80	1.1
wfed	1.6
prbarr:prbconv	6.3

CLM4 (Gauss-Markov 4):

Plot of Residuals vs Fitted shows a reasonably horizontal line for conditional mean for errors centered on zero.

In [30]:
```
#Checking for CLM 4 - zero conditional error mean
par(mfrow=c(1,1))
options(repr.plot.width=4, repr.plot.height=4)
plot(model_4, which=1, main = "Model 4")
```



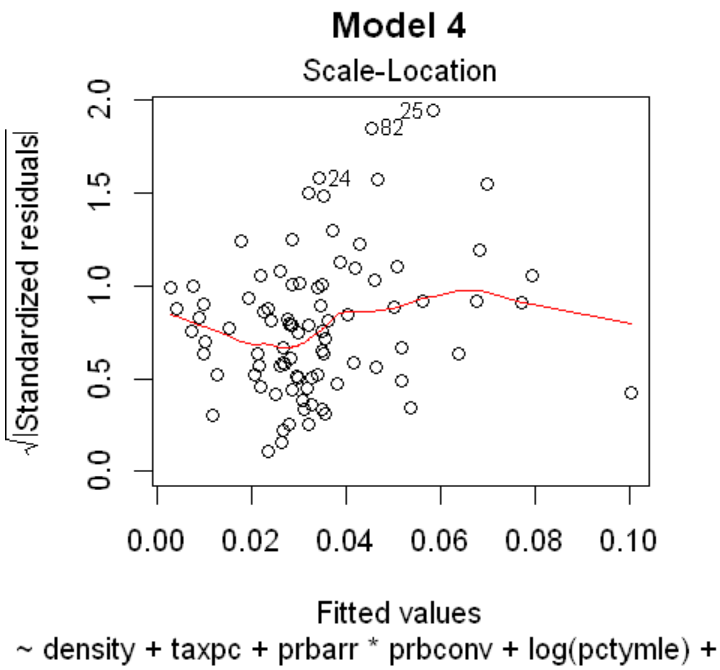
CLM5 (Gauss-Markov 5):

The Breusch-Pagan (BP) test for Model 4 shows a p value of 0.009. Though the BP test result is significant, an examination of the residual vs fitted value plot and the scale-location plot does not reveal a clear pattern of variance scatter. Although we could perhaps conclude that the homoskedasticity assumption has not been violated, we will choose to use heteroskedasticity-robust standard errors.

In [31]:
```
#checking for CLM 5 - nonconstant variance test
#ncvTest(model_4)
#install.packages("lmtest",repos = "http://cran.us.r-project.org")
library(lmtest)
bptest(model_4)
```

```
        studentized Breusch-Pagan test

data:  model_4
BP = 20.334, df = 8, p-value = 0.009144
```

In [32]:
```
par(mfrow=c(1,1))
options(repr.plot.width=4, repr.plot.height=4)
plot(model_4, which=3, main = "Model 4")
```



CLM 6:

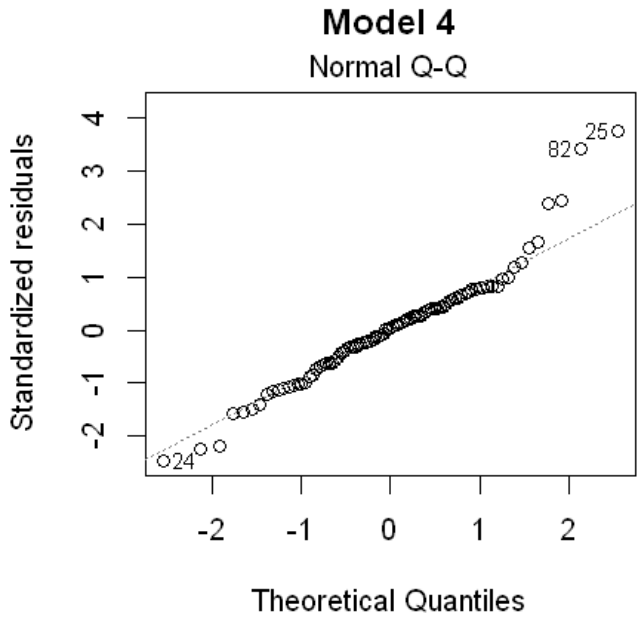
Shapiro-Wilk analysis shows some evidence of non-normality (p-value = 0.025).

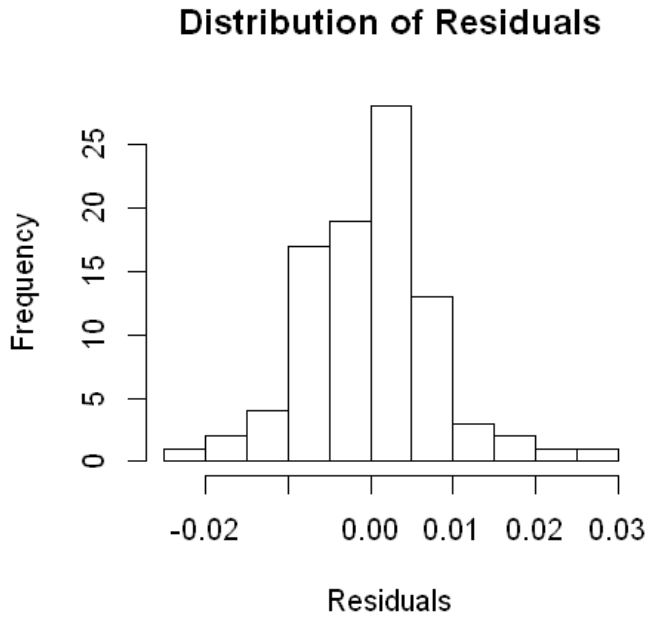
In [33]:
```
shapiro.test(model_4$residuals)
```

```
        Shapiro-Wilk normality test

data:  model_4$residuals
W = 0.96817, p-value = 0.02504
```

Additionally a Q-Q Plot of model 4 demonstrates a notable positive skew. Asymptotics based on a version of the central limit theorem should be helpful here (N > 30) in assuming a normal sampling distribution of our coefficients. Moreover, a histogram of the residuals appears normal.

In [34]:
```
par(mfrow=c(1,2))
options(repr.plot.width=8, repr.plot.height=4)
plot(model_4, which=2, main="Model 4")
hist(model_4$residuals, main="Distribution of Residuals", xlab="Residuals")
```





The new Regression Table is displayed below:

In [35]:
```
library(sandwich)
cov_4 <- vcovHC(model_4, type = "HC")
robust.se_4 <- sqrt(diag(cov_4))

library(stargazer)
stargazer(model_1, model_2, model_3, model_4, type = "text",
          #report = "vc", # Don't report errors, since we haven't covered them
          title = "Linear Models",
          se=list(NULL, robust.se_1, robust.se_2, robust.se_3, robust.se_4),
          #keep.stat = c("rsq", "n"),
          #omit.table.layout = "n",
         add.lines=list(c("AIC", round(AIC(model_1),1), round(AIC(model_2),1), round(AIC(model_3),1), round(AIC(model_4),1
)))
         ) # Omit more output related to errors
```

Linear Models

	Dependent variable:			
	crmrte			
	(1)	(2)	(3)	(4)
density	0.008***	0.006***	0.005***	0.005***
	(0.001)	(0.001)	(0.001)	(0.001)
taxpc	0.0003***	0.0003*	0.0004***	0.0004***
	(0.0001)	(0.0002)	(0.0001)	(0.0001)
prbarr		-0.071	-0.076***	-0.074***
			(0.018)	(0.017)
prbconv		-0.027	-0.026***	-0.029***
			(0.006)	(0.005)
log(pctymle)			0.020	0.017***
				(0.004)
prbpris			0.002	
avgsen			-0.0001	
west			0.001	
central			-0.003	
pctmin80			0.0003	0.0003***
				(0.0001)
wcon			0.00002	
wtuc			0.00000	
wtrd			0.00001	
wfir			-0.00004	
inv_wser			2.643	
wmfg			-0.00001	
wfed			0.0001	0.00004**
				(0.00002)
wsta			-0.00003	
wloc			0.0001	
prbarr:prbconv		0.044	0.050***	0.051***
			(0.012)	(0.013)
Constant	0.009**	0.042***	-0.048***	-0.020
	(0.004)	(0.006)	(0.009)	(0.025)
AIC	-537.6	-561.1	-591	-600.7
Observations	91	91	91	91
R2	0.584	0.699	0.844	0.818
Adjusted R2	0.574	0.681	0.799	0.800
Residual Std. Error	0.012 (df = 88)	0.011 (df = 85)	0.008 (df = 70)	0.008 (df = 82)
F Statistic	61.644*** (df = 2; 88)	39.454*** (df = 5; 85)	18.943*** (df = 20; 70)	45.918*** (df = 8; 82)

Note: *p<0.1; **p<0.05; ***p<0.01

As can be seen in the summary table above, Model 4 has the lowest (best) AIC and highest (best) adjusted-R^2 of all four models.

Furthermore, comparing all models to model_0 (no covariates -- mean crime as a point estimate of crime in all cases), all models support rejection of the null hypothesis shown below in favor of the alternative hypothesis.

- Null Hypothesis $H0$: The input variables of the model are not associated with crime rate; the model does not predict crime better than model 0 (mean crime rate as an estimator).
- Alternative Hypothesis HA: The input variables of the model are associated with crime rate; the model has some non-zero coefficients and predicts crime better than simply using the mean crime value.

Reference: $model\_0 : \beta_0 + u$

In [36]:
```
model_0 <- lm(crmrte ~ 1)
cat("Model_0 vs. Model_1")
waldtest(model_1, model_0, vcov = vcovHC)
cat('---\n\n')
cat("Model_0 vs. Model_2")
waldtest(model_2, model_0, vcov = vcovHC)
cat('---\n\n')
cat("Model_0 vs. Model_3")
waldtest(model_3, model_0, vcov = vcovHC)
cat('---\n\n')
cat("Model_0 vs. Model_4")
waldtest(model_4, model_0, vcov = vcovHC)
```

```
Model_0 vs. Model_1
```

Res.Df	Df	F	Pr(>F)
88	NA	NA	NA
90	-2	69.13072	9.002712e-19

```
---

Model_0 vs. Model_2
```

Res.Df	Df	F	Pr(>F)
85	NA	NA	NA
90	-5	44.05201	3.410418e-22

```
---

Model_0 vs. Model_3
```

Res.Df	Df	F	Pr(>F)
70	NA	NA	NA
90	-20	13.79941	5.087887e-17

```
---

Model_0 vs. Model_4
```

Res.Df	Df	F	Pr(>F)
82	NA	NA	NA
90	-8	42.06622	6.623337e-26

5.0 Omitted Variables Discussion

For analysis over variables that comprise the error term, Model 2 will be used as the baseline to compare against.
$model\_2 : crmrte = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + u$

Although model 4 is the most descriptive model, the additional variables in model 4 as compared to model 2 ('pctymle', 'pctmin80', 'wfed') have no percieved relation to the omitted variables chosen and would not be proxies for the omitted variables.

Wealth of Adjacent Areas:

We see in our data that wealth (taxpc) is positively correlated with crime. While this tells us about where crime is committed, it does not necessarily tell us where the criminals live. It may be that adjacent county income differentials could be a driver of crime for the wealthier neighbor.

Expected direction: Higher adjacent wealth should bias crime rate towards zero (crime is lower - away from zero means crime is higher).

Bias size:

$model\_2: crmrte = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + \beta_6 adj\_wealth + u$
$adj\_wealth = \alpha_0 + \alpha_1 * density + u$
$adj\_wealth = \alpha_0 + \alpha_1 * taxpc + u$
$adj\_wealth = \alpha_0 + \alpha_1 * prbconv + u$
$adj\_wealth = \alpha_0 + \alpha_1 * prbarr + u$
$adj\_wealth = \alpha_0 + \alpha_1 * prbconv * prbarr + u$

adj_wealth (B6 > 0)	Term				
	B1 density	B2 taxpc	B3 prbconv	B4 prbarr	B5 probconv*prbarr)
Beta Coefficient Sign	**+**	**+**	**-**	**-**	**+**
alpha_1	+	~	~	~	~
Net effect of adj_wealth on covariate	If B1>0 and B6>0 then OMV B=B6*alpha_1 >0 and if B1>0 then OLS coefficient on density will be scaled away from zero (more positive) gaining statistical significance	As alpha_1 is not expected to be impacted as there is no relation between taxpc and adj_wealth	As alpha_1 is not expected to be impacted as there is no relation between prbconv and adj_wealth	As alpha_1 is not expected to be impacted as there is no relation between prbarr and adj_wealth	As alpha_1 is not expected to be impacted as there is no relation between the interaction term and adj_wealth

Proxy: 'taxpc' may proxy the omitted variable 'adj_wealth'.

Poverty Rate:

Another explanation for higher crime in high taxpc areas may be a large proportion of desperately poor people, who may be motivated by their circumstances to perform illegal activity.

Expected direction: Higher poor percentage should bias crime rate away from zero.

Bias size:

$model\_2: crmrte = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + \beta_6 pvrty + u$
$pvrty = \alpha_0 + \alpha_1 * density + u$
$pvrty = \alpha_0 + \alpha_1 * taxpc + u$
$pvrty = \alpha_0 + \alpha_1 * prbconv + u$
$pvrty = \alpha_0 + \alpha_1 * prbarr + u$
$pvrty = \alpha_0 + \alpha_1 * prbconv * prbarr + u$

pvrty (B6 > 0)	Term				
	B1 density	B2 taxpc	B3 prbconv	B4 prbarr	B5 probconv*prbarr)
Beta Coefficient Sign	**+**	**+**	**-**	**-**	**+**
alpha_1	+	-	~	~	~
Net effect of pvrty on covariate	If B1>0 and B6>0 then OMV B=B6*alpha_1>0 and if B1 >0 then OLS coefficient on density will be scaled away from zero (more positive) gaining statistical significance	If B2>0 and B6>0 then OMV B=B6*alpha_1 <0 and if B2>0 then OLS coefficient on density will be scaled away from zero (less positive) losing statistical significance	As alpha_1 is not expected to be impacted as there is no relation between prbconv and pvrty	As alpha_1 is not expected to be impacted as there is no relation between prbarr and pvrty	As alpha_1 is not expected to be impacted as there is no relation between the interaction term and pvrty

Proxy: 'taxpc' may proxy the omitted variable 'pvrty'.

Public Perception Score (police as viewed by public)

An omitted variable that could reduce crime is the public's perception of police. With a positive public perception of police, citizens are more likely to cooperate on criminal investigations leading to arrests. It is expected that density would be detrimental to the ability for the public and police to network as there are fewer community events on a per capita basis.

Expected direction: Higher perception scores should bias crime rate towards zero.

Bias size:

$model\_2: crmrte = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + \beta_6 percept + u$

$percept = \alpha_0 + \alpha_1 * density + u$

$percept = \alpha_0 + \alpha_1 * taxpc + u$

$percept = \alpha_0 + \alpha_1 * prbconv + u$

$percept = \alpha_0 + \alpha_1 * prbarr + u$

$percept = \alpha_0 + \alpha_1 * prbconv * prbarr + u$

percept (B6<0)	Term				
	B1 density	B2 taxpc	B3 prbconv	B4 prbarr	B5 probconv*prbarr)
Beta Coefficient Sign	**+**	**+**	**-**	**-**	**+**
alpha_1	-	~	~	+	~
Net effect of percept on covariate	If B1>0 and B6<0 then OMV B=B6*alpha_1>0 and if B1 >0 then OLS coefficient on density will be scaled away from zero (more positive) gaining statistical significance	As alpha_1 is not expected to be impacted as there is no relation between taxpc and percept	As alpha_1 is not expected to be impacted as there is no relation between prbconv and percept	If B4<0 and B6<0 then OMV B=B6*alpha_1<0 and if B4 < 0 then OLS coefficient on prbarr will be scaled away from zero (more negative) gaining statistical significance	As alpha_1 is not expected to be impacted as there is no relation between the interaction term and percept

Proxy: 'density' may proxy the omitted variable 'percept'.

Prosecutors per Capita

An omitted variable that could reduce crime is prosecutors per capita. The relationship expected is that wealthier areas can afford more prosectors and that prosecutors deliver convictions more reliably if they have sufficient head-count resources to better prepare cases.

Expected direction: Higher prosecutors per capita should bias crime rate towards zero.

Bias size:

$model\_2: crmrte = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + \beta_6 prsctrs + u$

$prsctrs = \alpha_0 + \alpha_1 * density + u$

$prsctrs = \alpha_0 + \alpha_1 * taxpc + u$

$prsctrs = \alpha_0 + \alpha_1 * prbconv + u$

$prsctrs = \alpha_0 + \alpha_1 * prbarr + u$

$prsctrs = \alpha_0 + \alpha_1 * prbconv * prbarr + u$

prsctrs (B6<0)	Term				
	B1 density	B2 taxpc	B3 prbconv	B4 prbarr	B5 probconv*prbarr)
Beta Coefficient Sign	**+**	**+**	**-**	**-**	**+**
alpha_1	~	+	+	~	~
Net effect of prsctrs on covariate	As alpha_1 is not expected to be impacted as there is no relation between density and prsctrs	If B2>0 and B6<0 then OMV B=B6*alpha_1<0 and if B2 >0 then OLS coefficient on taxpc will be scaled toward zero (less positive) losing statistical significance	If B3<0 and B6<0 then OMV B=B6*alpha_1>0 and if B3 < 0 then OLS coefficient on prbconv will be scaled toward zero (less negative) losing statistical significance	As alpha_1 is not expected to be impacted as there is no relation between prbarr and prsctrs	As alpha_1 is not expected to be impacted as there is no relation between the interaction term and prsctrs

Proxy: 'prbconv' and/or 'prbconv*prbarr' may proxy the omitted variable 'prsctrs'.

Recidivism

An omitted variable that could increase crime is recidivism. Recidivism rate directly translates to crimes committed. The relationship expected is that recidivism is a function of failures in the justice system.

Expected direction: Recidivism should bias crime rate away from zero.

Bias size:

$$model\_2: crmrte = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + \beta_6 rcdvsm + u$$
$$rcdvsm = \alpha_0 + \alpha_1 * density + u$$
$$rcdvsm = \alpha_0 + \alpha_1 * taxpc + u$$
$$rcdvsm = \alpha_0 + \alpha_1 * prbconv + u$$
$$rcdvsm = \alpha_0 + \alpha_1 * prbarr + u$$
$$rcdvsm = \alpha_0 + \alpha_1 * prbconv * prbarr + u$$

rcdvsm (B6>0)	Term				
	B1 density	B2 taxpc	B3 prbconv	B4 prbarr	B5 probconv*prbarr)
Beta Coefficient Sign	**+**	**+**	**-**	**-**	**+**
alpha_1	~	~	+	+	~
Net effect of rcdvsm on covariate	As alpha_1 is not expected to be impacted as there is no relation between density and rcdvsm	As alpha_1 is not expected to be impacted as there is no relation between taxpc and rcdvsm	If B3<0 and B6>0 then OMV B=B6*alpha_1>0 and if B3 < 0 then OLS coefficient on prbconv will be scaled toward zero (less negative) losing statistical significance	If B4<0 and B6>0 then OMV B=B6*alpha_1>0 and if B4 < 0 then OLS coefficient on prbarr will be scaled toward zero (less negative) losing statistical significance	As alpha_1 is not expected to be impacted as there is no relation between the interaction term and rcdvsm

Proxy: 'prbconv', 'prbarr', and/or 'prbconv*prbarr' may proxy the omitted variable 'rcdvsm'.

Eviction Rate

An omitted variable that could increase crime is eviction rate. Eviction rate could be expected to vary with taxpc (proxy for poverty rate), however study (https://github.com/UCB-INFO-PYTHON/Project2PetitSohnREPO/blob/master/W200%20Fall18%20_%20Thursday%2C%204_00%20_%20Project%202%20_%20Petit%20Sohn.pdf) indicates that this is not always true, and there are local reasons which may drive eviction. Eviction is a dimension of wealth and therefore could affect crime.

Expected direction: Eviction rate should bias crime rate away from zero.

Bias size:

$$model\_2: crmrte = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 prbconv + \beta_4 prbarr + \beta_5 prbconv * prbarr + \beta_6 eviction + u$$
$$eviction = \alpha_0 + \alpha_1 * density + u$$
$$eviction = \alpha_0 + \alpha_1 * taxpc + u$$
$$eviction = \alpha_0 + \alpha_1 * prbconv + u$$
$$eviction = \alpha_0 + \alpha_1 * prbarr + u$$
$$eviction = \alpha_0 + \alpha_1 * prbconv * prbarr + u$$

	Term				
eviction (B6>0)	B1 density	B2 taxpc	B3 prbconv	B4 prbarr	B5 probconv*prbarr)
Beta Coefficient Sign	**+**	**+**	**-**	**-**	**+**
alpha_1	~	-	~	~	~
Net effect of eviction on covariate	As alpha_1 is not expected to be impacted as there is no relation between density and eviction	If B2>0 and B6<0 then OMV B=B6*alpha_1<0 and if B2 >0 then OLS coefficient on taxpc will be scaled toward zero (less positive) losing statistical significance	As alpha_1 is not expected to be impacted as there is no relation between prbconv and eviction	As alpha_1 is not expected to be impacted as there is no relation between prbarr and eviction	As alpha_1 is not expected to be impacted as there is no relation between the interaction term and eviction

Proxy: 'taxpc' may proxy the omitted variable 'eviction'.

6.0 Conclusion

The following variables as part of model 4 were found to have significant explanatory power regarding the crime rate. Below is delineated each variable and their respective coefficient when model 4 variables are included in the model.

- Density, or people per square mile (density): $\beta = 0.005$
  - Every added person per square mile is correlated with 5 more crimes per thousand people.
- Tax revenue per-capita (taxpc): $\beta = 0.0004$
  - Here, because tax revenue and crime rate are both per capita, we can articulate the effect as follows: Each extra dollar in tax revenue is correlated with 4 more crimes per ten thousand people.
- Probability of arrest (prbarr): $\beta = -0.074$
  - This can be interpreted to mean that every new conviction for a given arrest is correlated with 74 less crimes per one thousand people.
- Probability of conviction (prbconv): $\beta = -0.0029$
  - This can be interpreted to mean that every new conviction for a given arrest is correlated with 29 less crimes per ten thousand people.
- Percentage change in population of young male (pctymle). $\beta = 0.017$
  - This can be interpreted to mean that every one-percent increase in young male population is correlated with 17 more crimes per thousand people.
- Change in population of minorities (pctmin80) $\beta = 0.0003$
  - This can be interpreted to mean that every percentage-point increase in minority population is correlated with 3 more crimes per ten thousand people.
- Wage of federal workers: $\beta = 0.00004$
  - This can be interpreted to mean that every additional dollar of wage for federal workers is correlated with 4 more crimes per hundred-thousand people.
- Product of prbarr, prbconv: $\beta = 0.051$
  - This can be interpreted to mean that a one-unit increase in the product of the probability of arrest and the probability of conviction is correlated with 51 more crimes per thousand people.

Reference: Research Question:

> When considering population density and wealth as variables related to crime rate, what is the combination of density & wealth that is most likely to define a zone of maximal return on political investment?

Response to Research Question: Statistical Response:

- Areas of higher density are likely to have a crime rate higher than areas of lower density.
- Areas of higher wealth are more likely to have a crime rate higher than areas of lower wealth.
- Based on above relationship, focus on areas that are dense and higher wealth.
- It is shown that prosecuting crime ('prbarr', 'prbconv') reduces crime.

Response to Research Question: Practical Resposne (Policy Recommendation):

- Improve effectiveness of police investigations and prosecutions -- Implement policies that facilitate arrests and convictions. One policy example might be expediting judicial review of warrants. Increasing the arrest and conviction rate will have a notable effect in reducing crime.
- Justice system funding should be directed toward dense areas with higher tax revenue per capita. In practice, this is likely the urban and suburban centers of North Carolina. The fact that criminal activity is correlated with dense areas is probably unsurprising. Putting greater police and first-responder resources in counties that already have higher tax revenue may be an unpopular policy. Although the data would indicate that crimes occur in these counties, indication exists that data supporting wealth profile of adjacent areas may offer deeper expla
- Hold density steady or reduce density where possible -- Population density has a ceiling as defined by zoning laws which is a lever firmly in the political realm.
- Study further the concept of improving relations between police and their communities to improve law enforcement outcomes.
- Study further reducing the eviction rate to improve wealth and crime rate outcomes.