Lab 3 Final Report | w203 Statistics for Data Science

Determinants of Crime in North Carolina

Prepared by Sophia Ayele & Navya Sandadi

1. Introduction

In this paper, we examine the factors associated with crime rates in North Carolina. The purpose of this analysis is to help a North Carolina political campaign to generate policy recommendations applicable to local government. For this analysis, we employ a cross-sectional dataset first compiled for the paper C. Cornwell and W. Trumball (1994), "Estimating the Economic Model of Crime with Panel Data," Review of Economics and Statistics 76, 360-366.

Our paper seeks to address the following research question:

What factors are associated with crime rates in the counties of North Carolina?

We begin by conducting an exploratory analysis to inform our model building process, following which, we evaluate four linear regression models. Finally, we discuss our results, their limitations, and policy implications.

2. Exploratory Analysis & Model Building Process

Our analysis seeks to measure the factors associated with crime rates in North Carolina counties. We operationalize crime rate as the ratio of crimes in North Carolina to its population. The focus of the exploratory data analysis is to select appropriate independent variables to include in our linear regression models. We also identify and address data quality issues, outliers, and multicollinearity in the available independent variables.

2.1 Data Quality Issues

We identified and addressed the following data quality issues in the dataset.

- **Null Values**: The dataset has 6 rows of NA observations. We removed these rows from the dataset.
- **Duplicate Rows**: There are duplicate observation for county ID 193. We removed one of the duplicate observations.
- **Invalid Values**: Probability of arrest and probability of conviction should only contain values between 0 and 1. However, 1 observation for probability of arrest and 10 observations for probability of conviction have values greater than 1. We do not use probability of conviction in our analysis and we dropped the one invalid observation in probability of arrest (see Independent Variable Selection section below for more information). One observation is tagged as being in both the west and central regions (county ID 71). We do not have a way to determine which region this county belongs in, therefore, we dropped this observation from the dataset.
- **Outliers**: We found several outlier values. These include a moderate outlier in percent young male (one value of .25 while the other values are .15 or below), a moderate outlier in tax per capita (one value of 119.8, while other values are below 80), and an outlier in the service industry wage variable (one value exceeds 2,000 while all other values are below 500). It is difficult to know whether these outliers are due to errors in the data or represent actual values. As explained in subsequent sections, we address each of these outliers differently.
- **Inconsistent Units**: The percentage young male variable was a percent rather than a decimal (e.g. 25 vs. 0.25). We rescaled this variable to a decimal for consistency with the other percentage and probability variables.
- **Omitted Counties**: The dataset contains observations for 90 out of 100 counties in North Carolina. If the omitted counties have different characteristics than the counties included in the dataset, our analysis may suffer from selection bias.
- **Data Recency**: This analysis is being conducted in 2019, but relies on data from 1987 (1980 in the case of census variables). It is likely that important population and demographic shifts have taken place over the last 30 years that may make this data unreliable for policy-making in 2019.

After addressing data quality issues, our final dataset includes observations for 88 out of 100 counties in North Carolina.

2.2 Categorization of variables

There are 25 variables in the dataset. Besides 'county' and 'year', the remaining variables can be categorized as:

- Demographic
- Geographic
- Economic
- Crime

The table below summarizes the variables included in the dataset. (see Table #1 below).

In [ ]:
```
### Install and load necessary packages ###

#install.packages("matrixStats")
#install.packages("plyr")
#install.packages("stargazer")
#install.packages("corrplot")
#install.packages("car")
#install.packages("lmtest")
#install.packages("sandwich")
#install.packages("ggpubr")

library(plyr)
library(ggplot2)
library(matrixStats)
library(stargazer)
library(dplyr)
library(corrplot)
library(car)
library(lmtest)
library(sandwich)
library(gridExtra)
library(ggpubr)
```

In [2]:
```
### Load data ###
data = read.csv("crime_v2.csv")
```

In [3]:
```
#### Summarize and Transform Data ####

# Remove rows that are all NA
#data[!complete.cases(data),]
data <- na.omit(data)
#dim(data)
#data

# Remove duplicate row
data <- distinct(data, .keep_all = FALSE)
#dim(data)
#data

# Change prbconv to numeric
data$prbconv <- as.numeric(as.character(data$prbconv))

# Rescale pctmin80 to decimal to match other variables
data$pctmin80_rescaled <- data$pctmin80/100

# Plot of service industry outlier
#plot(data$wser)

# Generate median wage variable
data$median_wage <- apply(data[,c("wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg", "wfed", "wsta", "wloc")],1, median, na.rm = TRUE)

# Drop incorrect value in arrest var (probability greater than 1)
data  <- subset(data, prbarr < 1)

# Remove county tagged in both regions
data$dup_region <- 0
data$dup_region <- with(data, ifelse(west == 1 & central == 1 , 1, dup_region))
#table(data$dup_region)
data <- subset(data, !dup_region == 1)
data <- subset(data, select=-c(dup_region))

# Print summary of dataframe
stargazer(data, type = "text", title="Descriptive statistics", digits=1)

# Print data types
#str(data)
```

```
Descriptive statistics
=================================================================
Statistic         N  Mean  St. Dev.  Min  Pctl(25) Pctl(75)   Max
-----------------------------------------------------------------
county            88 100.8   58.9     1     50.5    151.5     197
year              88 87.0    0.0     87      87       87      87
crmrte            88 0.03    0.02   0.01    0.02     0.04     0.1
prbarr            88  0.3    0.1     0.1    0.2      0.3      0.7
prbconv           88  0.5    0.3     0.1    0.3      0.6      2.1
prbpris           88  0.4    0.1     0.2    0.4      0.5      0.6
avgsen            88  9.5    2.6     5.4    7.4      11.4    17.4
polpc             88 0.002  0.001   0.001  0.001    0.002    0.004
density           88  1.4    1.5    0.000   0.5      1.6      8.8
taxpc             88 38.3    13.2   25.7    30.8     41.1    119.8
west              88  0.2    0.4      0       0        0        1
central           88  0.4    0.5      0       0        1        1
urban             88  0.1    0.3      0       0        0        1
pctmin80          88 26.1    16.9    1.5    10.1     38.3    64.3
wcon              88 286.2   47.5   193.6  250.8    315.3    436.8
wtuc              88 407.8   75.0   187.6  374.0    437.3    613.2
wtrd              88 210.5   34.1   154.2  190.4    223.4    354.7
wfir              88 321.1   54.5   170.9  284.0    340.6    509.5
wser              88 275.5  209.7   133.0  227.7    276.7   2,177.1
wmfg              88 334.5   88.4   157.4  286.7    356.9    646.8
wfed              88 441.9   60.2   326.1  395.9    477.8    598.0
wsta              88 357.9   43.7   258.3  329.0    384.1    499.6
wloc              88 311.1   27.1   239.2  297.2    327.4    388.1
mix               88  0.1    0.1    0.02    0.1      0.2      0.5
pctymle           88  0.1    0.02    0.1    0.1      0.1      0.2
pctmin80_rescaled 88  0.3    0.2    0.02    0.1      0.4      0.6
median_wage       88 314.8   37.9   231.7  293.7    329.5    436.8
-----------------------------------------------------------------
```

Table 2 (below), shows the variables that were selected for our final model.

Table 2: Variables Selected for Analysis

	Selected Variables	Variables Not Selected
Demographic	population density percent young male percent minority	urban
Economic	tax per capita	median weekly wage
Geographic	central region west region	
Criminal Justice System	***crime rate - dependent variable*** probability of arrest	police per capita probability of conviction probability of sentence average sentence days offense mix
Other		county year

2.3 Dependent Variable Selection

The dataset contains two variables that measure crime:

- ***Crime Rate*** - crimes committed per person
- ***Offense Mix*** - ratio of crimes involving face to face contact (robbery, assault, rape) vs. those that do not

We selected crime rate as the dependent variable for our analysis because it directly relates to our research question about the incidence of crimes. Offense mix addresses the severity of crimes committed vs. the incidence of crimes. Offense mix could be a useful dependent variable for a secondary analysis of factors that contribute to the severity of crimes, but that is outside the scope of this analysis.

2.4 Independent Variable Selection

We used the following criteria to select independent variables to include in our final models:

- Good data quality
- Highly correlation with crime rate
- Logical connection to crime rate
- Low multicollinearity
- Not a proxy for crime rate

We examined histograms of each variable to look for outliers and other anomalies. We also conducted correlation tests between our dependent variable (crime rate) and potential independent variables. To check for multicollinearity among potential independent variables, we also conducted correlation tests between continuous variables and t-tests for binary variables. Lastly, we performed a sanity check on our model selection by discussing logical links between independent variables and crime rate and whether any potential independent variables might also be proxies for crime rate.

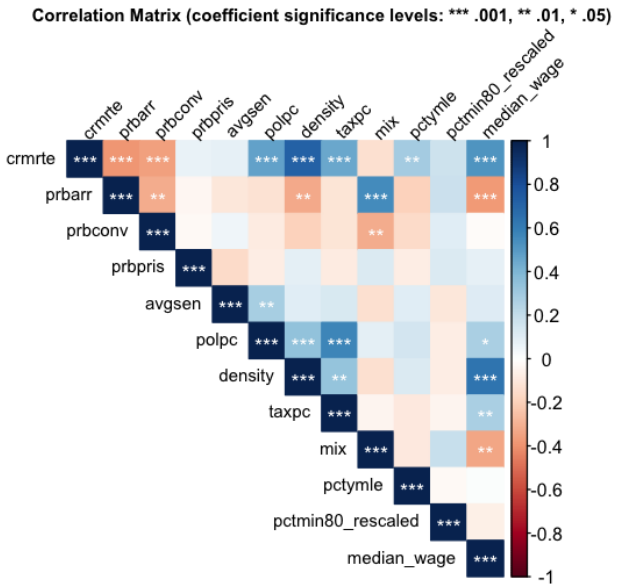
The Correlation Matrix below shows the degree of correlation between the continuous variables in the dataset. The color scale indicates the value of the correlation coefficient and the stars indicate whether a correlation is statistically significant. Correlations with no stars were not found to be statistically significant.

In [4]:

```
### Histograms, correlations, and t-tests for sanity checks ###

#options(repr.plot.width=6, repr.plot.height=4)
#hist(data$crmrte)
#hist(data$prbarr)
#hist(data$density)
#hist(data$taxpc)
#hist(data$pctymle)
#hist(data$pctmin80)
#hist(data$prbconv)
#hist(data$prbpris)
#hist(data$avgsen)
#hist(data$polpc)
#hist(data$median_wage)
#hist(data$mix)
#table(data$urban)
#table(data$central)
#table(data$west)

# Correlation btwn crime rate and potential independent variables
# plot(data$crmrte, data$prbarr)
# cor.test(data$crmrte, data$prbarr)
# plot(data$crmrte, data$prbconv)
# cor.test(data$crmrte, data$prbconv)
# plot(data$crmrte, data$prbpris)
# cor.test(data$crmrte, data$prbpris)
# plot(data$crmrte, data$avgsen)
# cor.test(data$crmrte, data$avgsen)
# plot(data$crmrte, data$polpc)
# cor.test(data$crmrte, data$polpc)
# plot(data$crmrte, data$density)
# cor.test(data$crmrte, data$density)
# plot(data$crmrte, data$taxpc)
# cor.test(data$crmrte, data$taxpc)
# plot(data$crmrte, data$pctmin80_rescaled)
# cor.test(data$crmrte, data$pctmin80_rescaled)
# plot(data$crmrte, data$pctymle)
# cor.test(data$crmrte, data$pctymle)
# plot(data$crmrte, data$mix)
# cor.test(data$crmrte, data$mix)
# plot(data$crmrte, as.numeric(data$median_wage))
#cor.test(data$crmrte, data$median_wage)

# Correlation btwn independent variables
# plot(data$density, data$taxpc)
#cor.test(data$density, data$taxpc)
#plot(data$density, data$pctmin80_rescaled)
#cor.test(data$density, data$pctmin80_rescaled)
# plot(data$density, data$pctymle)
#cor.test(data$density, data$pctymle)
# plot(data$density, data$prbarr)
#cor.test(data$density, data$prbarr)
# plot(data$taxpc, data$pctmin80_rescaled)
# cor.test(data$taxpc, data$pctmin80_rescaled)
# plot(data$taxpc, data$pctymle)
# cor.test(data$taxpc, data$pctymle)
# plot(data$taxpc, data$prbarr)
# cor.test(data$taxpc, data$prbarr)
# plot(data$pctymle, data$pctmin80_rescaled)
# cor.test(data$pctymle, data$pctmin80_rescaled)
# plot(data$pctymle, data$prbarr)
# cor.test(data$pctymle, data$prbarr)
# plot(data$pctymle, data$prbarr)
# cor.test(data$pctymle, data$prbarr)

# T-tests for urban
#t.test(data$crmrte ~ data$urban)
#t.test(data$prbarr ~ data$urban)
#t.test(data$prbconv ~ data$urban)
#t.test(data$avgsen ~ data$urban)
#t.test(data$polpc ~ data$urban)
#t.test(data$density ~ data$urban)
# t.test(data$taxpc ~ data$urban)
# t.test(data$pctmin80 ~ data$urban)
# t.test(data$pctymle ~ data$urban)
# t.test(data$median_wage ~ data$urban)

# T-tests for regions
#t.test(data$crmrte ~ data$central)
# t.test(data$prbarr ~ data$central)
# t.test(data$prbconv ~ data$central)
# t.test(data$avgsen ~ data$central)
# t.test(data$polpc ~ data$central)
# t.test(data$density ~ data$central)
# t.test(data$taxpc ~ data$central)
# t.test(data$pctmin80 ~ data$central)
# t.test(data$pctymle ~ data$central)
# t.test(data$median_wage ~ data$central)

#t.test(data$crmrte ~ data$west)
# t.test(data$prbarr ~ data$west)
# t.test(data$prbconv ~ data$west)
# t.test(data$avgsen ~ data$west)
# t.test(data$polpc ~ data$west)
# t.test(data$density ~ data$west)
# t.test(data$taxpc ~ data$west)
# t.test(data$pctmin80 ~ data$west)
# t.test(data$pctymle ~ data$west)
# t.test(data$median_wage ~ data$west)

# t.test(data$crmrte ~ data$other_region)
# t.test(data$prbarr ~ data$other_region)
# t.test(data$prbconv ~ data$other_region)
# t.test(data$avgsen ~ data$other_region)
# t.test(data$polpc ~ data$other_region)
```

```
# t.test(data$density ~ data$other_region)
# t.test(data$taxpc ~ data$other_region)
# t.test(data$pctmin80 ~ data$other_region)
# t.test(data$pctymle ~ data$other_region)
# t.test(data$median_wage ~ data$other_region)
```

In [5]:
```
# Correlation plot of continous variables
continous <- data[,c(3:10, 24:27)]
correlations <- cor(continous)
#correlations
options(repr.plot.width=5, repr.plot.height=5)
res1 <- cor.mtest(continous, conf.level = .95)
corrplot(correlations, p.mat = res1$p, method = "color", type = "upper",
         sig.level = c(.001, .01, .05), pch.cex = .9,
         insig = "label_sig", pch.col = "white",
         tl.col = "black", tl.srt = 45, tl.cex = .75)
title("Correlation Matrix (coefficient significance levels: *** .001, ** .01, * .05)", line = 2.5, cex.main=.75)
```



Demographic Variables - We hypothesize that urban, population-dense areas are likely to have a higher incidence of crime. We also hypothesize that the urban indicator variable is a proxy for population density, therefore, we only included one of these variables in our models. We selected population density (the number of people per square mile) because it is highly correlated with crime rate (corr coef. 0.7261648, p-value 8.119e-16), because it is a continuous variable that allows for a more nuanced estimate, and because only a few counties are urban (8 or 9%).

We included percent young male because it is moderately correlated with crime rate and we hypothesize that young males are more likely to commit crimes than other groups. There is one moderate outlier in percent young male (one value of .25 while the other values are .15 or below), but it is not enough of an outlier to raise suspicion that the value may be an error. We also included percent minority in our model. While this variable is not significantly correlated with crime rate (corr coef. 0.1781311, p-value 0.0949), we hypothesize that it could be an important predictor of crime rate because it could account for disparities in opportunity.

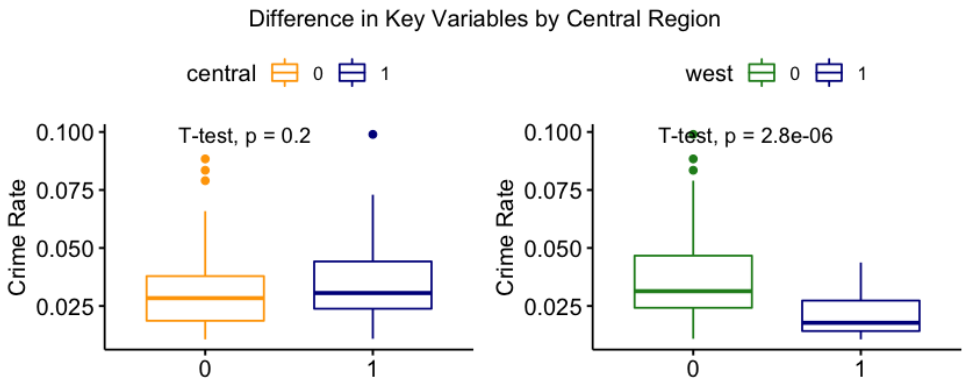
Economic Variables - We hypothesize that poverty and lack of opportunity are likely drivers of crime. Therefore, we also hypothesize that more affluent areas are likely to have lower crime rates. The dataset includes two variables that can be used as a proxy for affluence: nine average weekly wage variables across different sectors and tax revenue per capita. There is likely a high correlation among the average weekly wage variables within each county, therefore, we combined these variables into a median wage variable. We used the median rather than the average wage across sectors because there is an outlier in the service industry wage variable for one county (one value exceeds $2,000 while all other values are below 500$).

There is a low correlation between tax per capita and median weekly wage (corr coef. 0.2658057, p-value 0.01134). Since both of these variables likely act as a proxy for the affluence of a county, we decided to only include one in our model. Although both variables are moderately correlated with crime rate, we selected tax per capita for our final model because it is likely a better predictor of overall affluence of a county than median wage. This is because tax revenue per capita measures income as well as wealth and business revenue. There is one moderate outlier in tax per capita (one value of 119.8, while other values are below 80), but it is not enough of an outlier to raise suspicions that the value may be incorrect.

Geographic Variables - We hypothesize that region specific effects could impact crime rate. These effects could include level of urbanization, demographic factors, and economic factors. The dataset includes proxies for some of these factors, including population density, percent minority, and percent young male, but there are likely other region specific factors that are not captured. For example, high unemployment rates could indicate lack of legitimate opportunities to earn a living. T-tests show significant differences in mean crime rates for counties in the West Region vs. other regions, but we do not see a significant difference for counties in the Central Region.

The dataset only includes indicator variables for the west and central regions. There are 33 counties tagged in the central region, 32 counties tagged in the west region, and 35 observations that are not tagged as being in either region. One county is tagged as being in both the west and central regions. We dropped this county from our final analysis dataset because we do not have a way to correct the issue (the dataset only has county IDs and not county names). We included both the central and west region indicators in our models. We do not need to include an indicator for the counties not in central or west because that effect is already captured by the other two categories.

In [6]:
```
options(repr.plot.width=7, repr.plot.height=3)

# Box plots with t-tests for crime rate by region
p_central_crmrte <- ggboxplot(data, x = "central", y = "crmrte", color = "central", palette = c("orange", "blue4")) + xlab("") + ylab("C
p_west_crmrte <- ggboxplot(data, x = "west", y = "crmrte", color = "west", palette = c("forestgreen", "blue4")) + xlab("") + ylab("Crime
grid.arrange(p_central_crmrte, p_west_crmrte, nrow=1, top="Difference in Key Variables by Central Region")

# Box plots with t-test for key variables by central
#p_central_crmrte <- ggboxplot(data, x = "central", y = "crmrte", color = "central", palette = c("orange", "blue4")) + xlab("") + ylab(
#p_central_density <- ggboxplot(data, x = "central", y = "density", color = "central", palette = c("orange", "blue4")) + xlab("") + ylai
#p_central_prbarr <- ggboxplot(data, x = "central", y = "prbarr", color = "central", palette = c("orange", "blue4")) + xlab("") + ylab(
#p_central_taxpc <- ggboxplot(data, x = "central", y = "taxpc", color = "central", palette = c("orange", "blue4")) + xlab("") + ylab("Ta
#p_central_pctymle <- ggboxplot(data, x = "central", y = "pctymle", color = "central", palette = c("orange", "blue4")) + xlab("") + ylai
#p_central_pctmin80 <- ggboxplot(data, x = "central", y = "pctmin80", color = "central", palette = c("orange", "blue4")) + xlab("") + y
#grid.arrange(p_central_crmrte, p_central_density, p_central_taxpc, p_central_prbarr, p_central_pctymle, p_central_pctmin80, ncol=2, to

# # Box plots with t-test for key variables by west
#p_west_crmrte <- ggboxplot(data, x = "west", y = "crmrte", color = "west", palette = c("forestgreen", "blue4")) + xlab("") + ylab("Crir
#p_west_density <- ggboxplot(data, x = "west", y = "density", color = "west", palette = c("forestgreen", "blue4")) + xlab("") + ylab("Pc
#p_west_prbarr <- ggboxplot(data, x = "west", y = "prbarr", color = "west", palette = c("forestgreen", "blue4")) + xlab("") + ylab("Prol
#p_west_taxpc <- ggboxplot(data, x = "west", y = "taxpc", color = "west", palette = c("forestgreen", "blue4")) + xlab("") + ylab("Tax Va
#p_west_pctymle <- ggboxplot(data, x = "west", y = "pctymle", color = "west", palette = c("forestgreen", "blue4")) + xlab("") + ylab("Pc
#p_west_pctmin80 <- ggboxplot(data, x = "west", y = "pctmin80", color = "west", palette = c("forestgreen", "blue4")) + xlab("") + ylab(
#grid.arrange(p_west_crmrte, p_west_density, p_west_taxpc, p_west_prbarr, p_west_pctymle, p_west_pctmin80, ncol=2, top="Differences in
```



Criminal Justice System Variables - We hypothesize that criminal justice system policies and practices could impact crime rate. For example, counties that focus more on justice system diversion and rehabilitation may have lower arrest and conviction rates than those that follow a more traditional criminal justice system model. Besides the crime rate variable that we selected as our dependent variable, the dataset includes several variables related to the criminal justice system. Of these variables, we found three variables to be significantly correlated with crime rate.

Probability of arrest and probability of conviction are negatively correlated with crime rate (corr coef. -0.379276, p-value 0.0002469 and corr coef. -0.3530406, p-value 0.0006902, respectively) and police per capita is positively correlated with crime rate (corr coef. 0.4804399, p-value 1.89e-06). The probability of arrest variable included one value greater than one (1.09). Since we were unable to determine whether this should be 0.109, 0.0109 or some other value, we dropped this observation to avoid introducing error into our model (corr coef. and p-value above calculated after dropping this observation).

The probability of conviction variable also included ten values greater than one. We decided not to use this variable because it would have involved dropping 10 observations (our dataset was already relatively small, n=88) or making assumptions about these values that might have introduced error into our model. Additionally, because both probability of arrest and probability of conviction are negatively correlated with crime rate, we hypothesize that both of these variables capture the deterrent effect of the criminal justice system.

Police per capita is positively correlated with crime rate indicating that counties with higher crime rates have more police per capita. This variable is likely measuring the crime rate, to some extent, and should not be included in our models. Additionally, police per capita is moderately correlated with tax revenue per capita indicating that there is likely some degree of collinearity between these two variables.

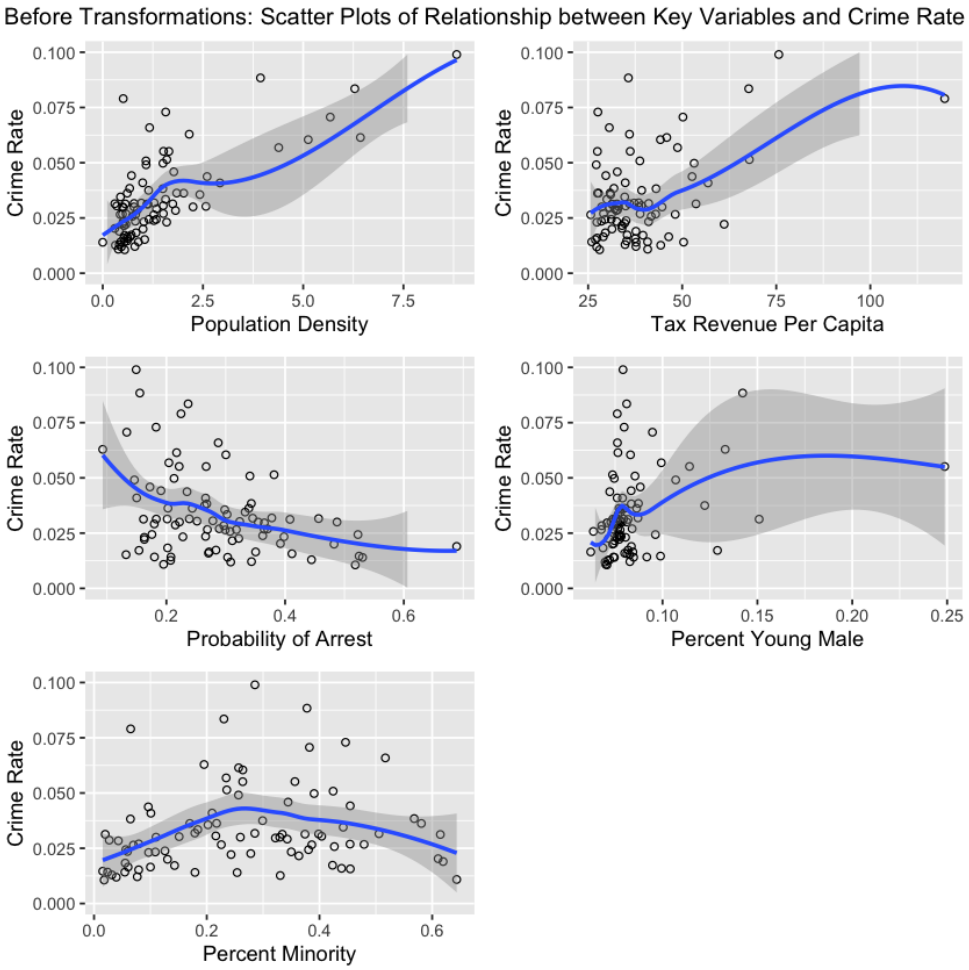
Ultimately, we only included probability of arrest in our final model because this variable is significantly correlated with crime rate and because we hypothesize that probability of conviction, probability of sentence, and average sentence days are also likely proxies for justice system deterrents. We do not hypothesize that offense mix (the severity of crimes) impacts crime rates. In fact, this variable could also measure crime rate, to some extent, because areas with higher crime per capita may also be more likely to have a higher incidence of severe crimes. Therefore, this variable should not be included in our model.

Other Variables - We did not include county ID or year in our analysis. Since the data is all from the same year, 1987, the year variable does not provide useful information. While there could be unique county specific effects, including indicators for all 90 counties was not viable.

2.5 Data Transformations

Once we selected our key variables, we used scatter plots to evaluate the relationship between the continuous independent variables and crime rate (see scatter plots below). The lowess line (blue line) finds the curve of best fit without assuming that the relationship between the variables is linear.

In [7]:
```
# Correlation plots with lowess line to reveal relationship between key variables and crime rate
options(repr.plot.width=7, repr.plot.height=7)
p_density <- ggplot(data = data, aes(x = density , y = crmrte)) + geom_point(pch = 1) + xlab("Population Density") + ylab("Crime Rate")
p_prbarr <- ggplot(data = data, aes(x = prbarr, y = crmrte)) + geom_point(pch = 1) + xlab("Probability of Arrest") + ylab("Crime Rate")
p_taxpc <- ggplot(data = data, aes(x = taxpc, y = crmrte)) + geom_point(pch = 1) + xlab("Tax Revenue Per Capita") + ylab("Crime Rate")
p_pctymle <- ggplot(data = data, aes(x = pctymle, y = crmrte)) + geom_point(pch = 1) + xlab("Percent Young Male") + ylab("Crime Rate")
p_pctmin80_rescaled <- ggplot(data = data, aes(x = pctmin80_rescaled, y = crmrte)) + geom_point(pch = 1) + xlab("Percent Minority") + y

grid.arrange(p_density, p_taxpc, p_prbarr, p_pctymle, p_pctmin80_rescaled, ncol=2, top="Before Transformations: Scatter Plots of Relati
```



The relationship between population density and crime rate is mostly linear. Although there is some variation between smaller and larger values, this is mostly driven by a small group of urban counties with higher crime rates. While these values are outliers, they likely have unique characteristics that are important to include in our models. The relationship between tax revenue per capita, probability of arrest, and percent young male all exhibit exponential characteristics. In both tax revenue per capita and percent young male this relationship is largely driven by outlier points. Finally, the relationship between percent minority and crime rate appears to be parabolic. In the following sections, we discuss transformations that we attempt in order to reveal linearity in the relationships between these variables. We do not apply any transformation to population density since the relationship with crime rate is already mostly linear.

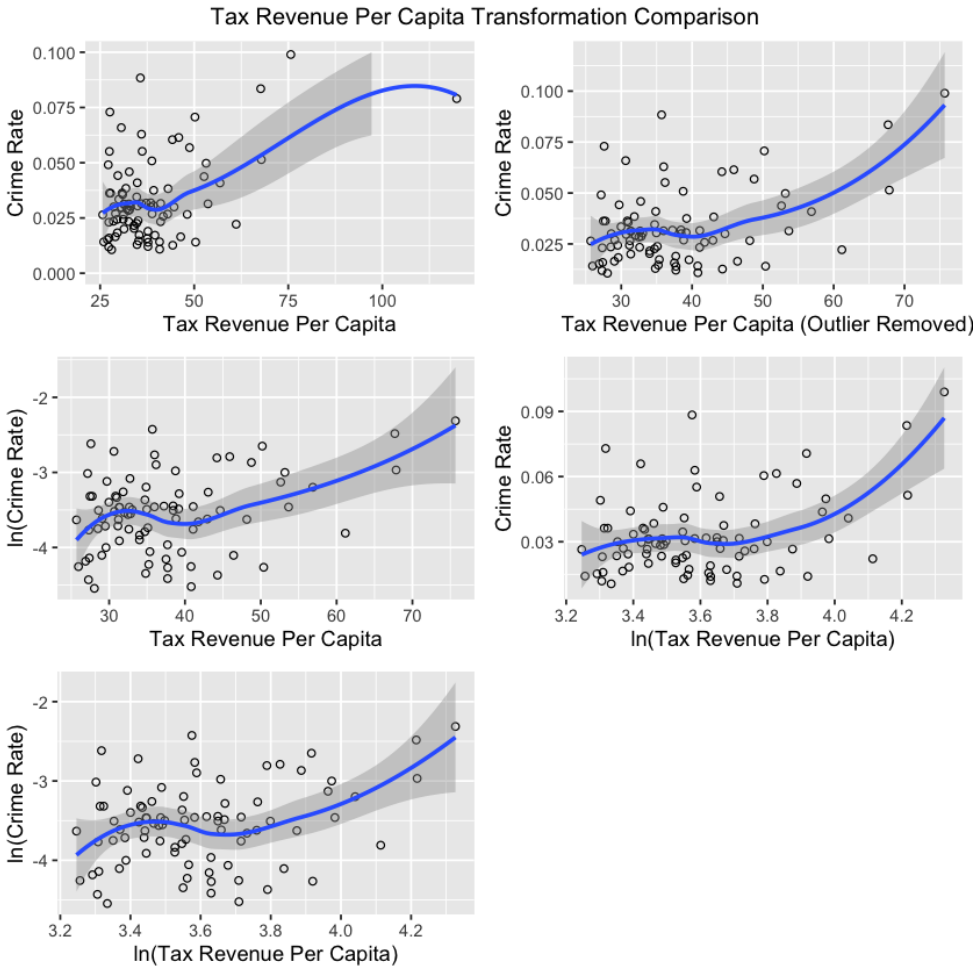
Since there are no negative or zero values in our selected variables, logging is technically an appropriate transformation. However, taking the log of values close to zero can amplify measurement error in data because the log of values close to zero will have a steeper slope than values not close to zero. Several of our selected variables have values between 0 and 1 (crime rate, probability of arrest, percent minority, percent young male), therefore, it is important to be mindful of this downside of logs. We still decided to include logs in our models because they: (1) are easier to interpret, (2) reveal linearity in the data better than other transformations (see specific examples in the next section), and (3) make theoretical sense in the context of this data (e.g. it makes sense to talk about percentage changes for these variables).

2.5.1 Tax Revenue Per Capita

We began by removing the outlier in tax revenue per capita. We then tried logging crime rate, tax revenue per capita, and both variables. These additional transformations do not provide much improvement. Therefore, we decided to remove the outlier, but not apply additional transformations to this variable.

In [8]:
```
# Create indicator variable for percent minority
data$high_pctmin <- as.numeric(data$pctmin80_rescaled >= .26)

# Check mean crmrte for % minority < 26% and >=26%
#aggregate(data$crmrte,list(data$high_pctmin),mean)

# Create dataset with outliers in taxpc and pctmyle removed
no_outliers <- data[data$taxpc < 100, ]
no_outliers <- no_outliers[no_outliers$pctymle < .2, ]
#no_outliers
```

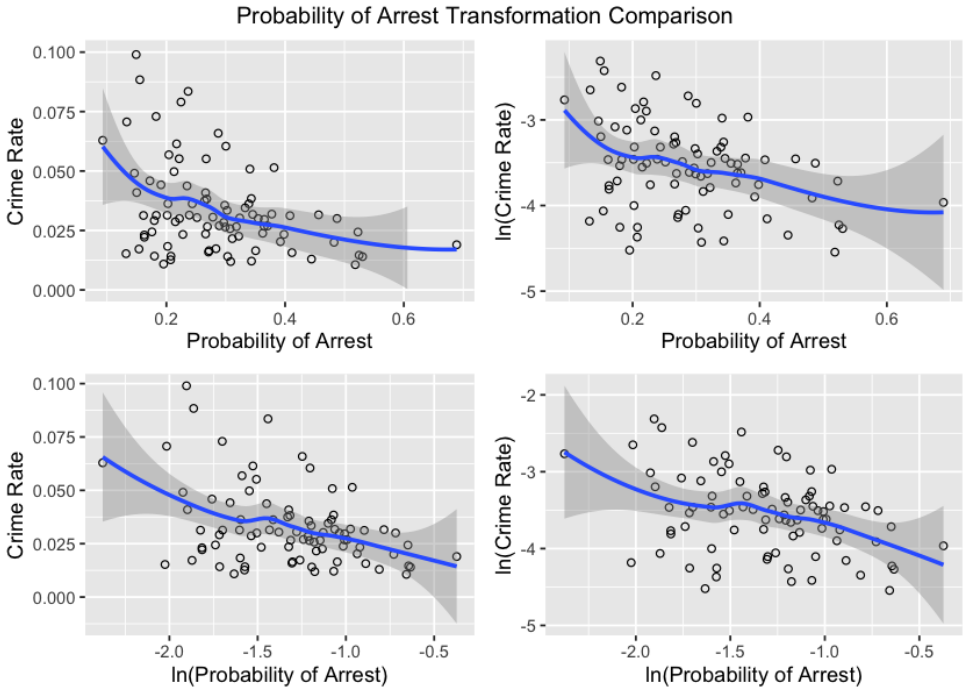
In [9]:
```
# Correlation plot comparison for taxpc
options(repr.plot.width=7, repr.plot.height=7)
p_taxpc_no_outlier <- ggplot(data = no_outliers, aes(x = taxpc, y = crmrte)) + geom_point(pch = 1) + xlab("Tax Revenue Per Capita (Outl
p_taxpc_plog <- ggplot(data = no_outliers, aes(x = log(taxpc), y = crmrte)) + geom_point(pch = 1) + xlab("ln(Tax Revenue Per Capita)") +
p_taxpc_olog <- ggplot(data = no_outliers, aes(x = taxpc, y = log(crmrte))) + geom_point(pch = 1) + xlab("Tax Revenue Per Capita") + yla
p_taxpc_dlog <- ggplot(data = no_outliers, aes(x = log(taxpc), y = log(crmrte))) + geom_point(pch = 1) + xlab("ln(Tax Revenue Per Capita
#p_taxpc_sqrt <- ggplot(data = no_outliers, aes(x = sqrt(taxpc), y = crmrte)) + geom_point(pch = 1) + xlab("Square Root of Tax Value Pe
#p_taxpc_dsqrt <- ggplot(data = no_outliers, aes(x = sqrt(taxpc), y = sqrt(crmrte))) + geom_point(pch = 1) + xlab("Square Root of Tax V

grid.arrange(p_taxpc, p_taxpc_no_outlier, p_taxpc_olog, p_taxpc_plog, p_taxpc_dlog, ncol=2, top="Tax Revenue Per Capita Transformation
```



2.5.2 Probability of Arrest

Next, we tried logging crime rate, probability arrest, and both variables. The plot with the log of probability of arrest appears to reveal the most linear relationship between the variables. We decided to apply this transformation to this variable. An advantage of only logging probability of arrest, rather than also logging crime rate, is that introducing the log of crime rate might not be appropriate for the other independent variables.

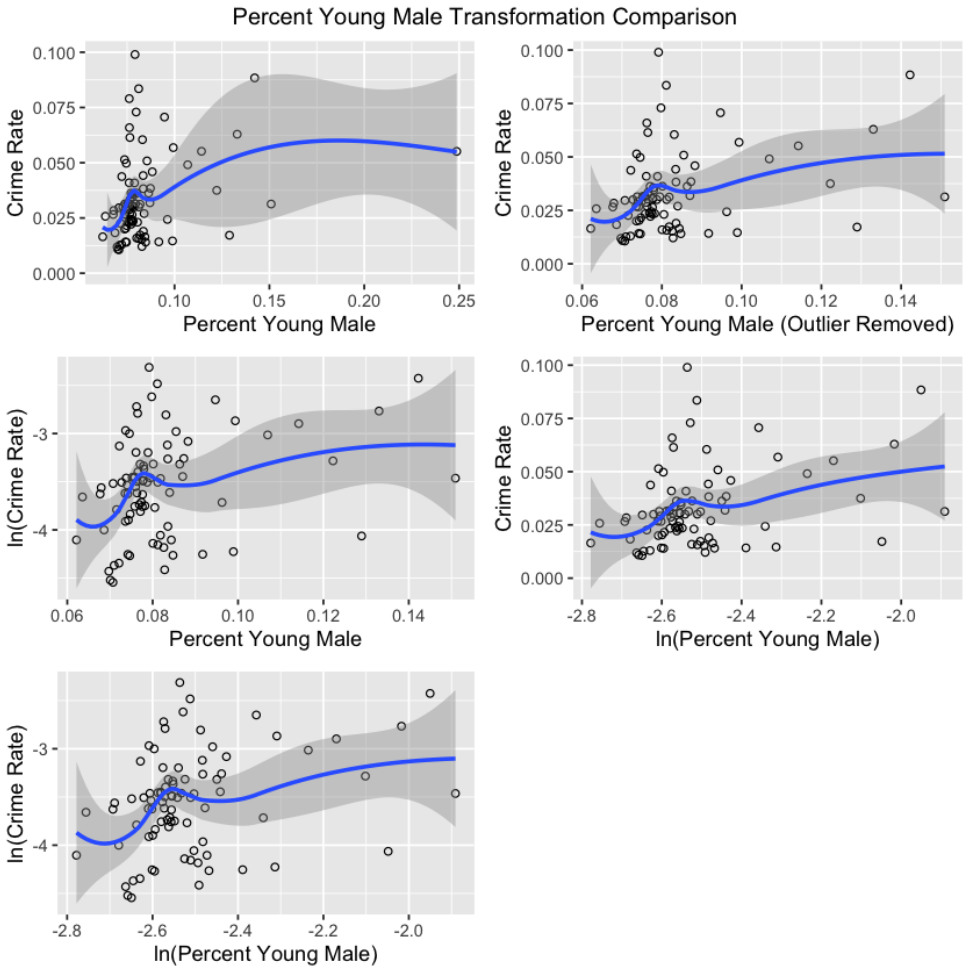
In [10]:
```
# Correlation plot comparison for prbarr
options(repr.plot.width=7, repr.plot.height=5)
p_prbarr_olog <- ggplot(data = no_outliers, aes(x = prbarr, y = log(crmrte))) + geom_point(pch = 1) + xlab("Probability of Arrest") + y
p_prbarr_plog <- ggplot(data = no_outliers, aes(x = log(prbarr), y = crmrte)) + geom_point(pch = 1) + xlab("ln(Probability of Arrest)"
p_prbarr_dlog <- ggplot(data = no_outliers, aes(x = log(prbarr), y = log(crmrte))) + geom_point(pch = 1) + xlab("ln(Probability of Arre
grid.arrange(p_prbarr, p_prbarr_olog, p_prbarr_plog, p_prbarr_dlog, ncol=2, top="Probability of Arrest Transformation Comparison")
```



2.5.3 Percent Young Male

We began by removing the outlier in percent young male. We then tried logging crime rate, percent young male, and both variables. These additional transformations do not provide much improvement. Therefore, we decided to remove the outlier, but not apply additional transformations to this variable.

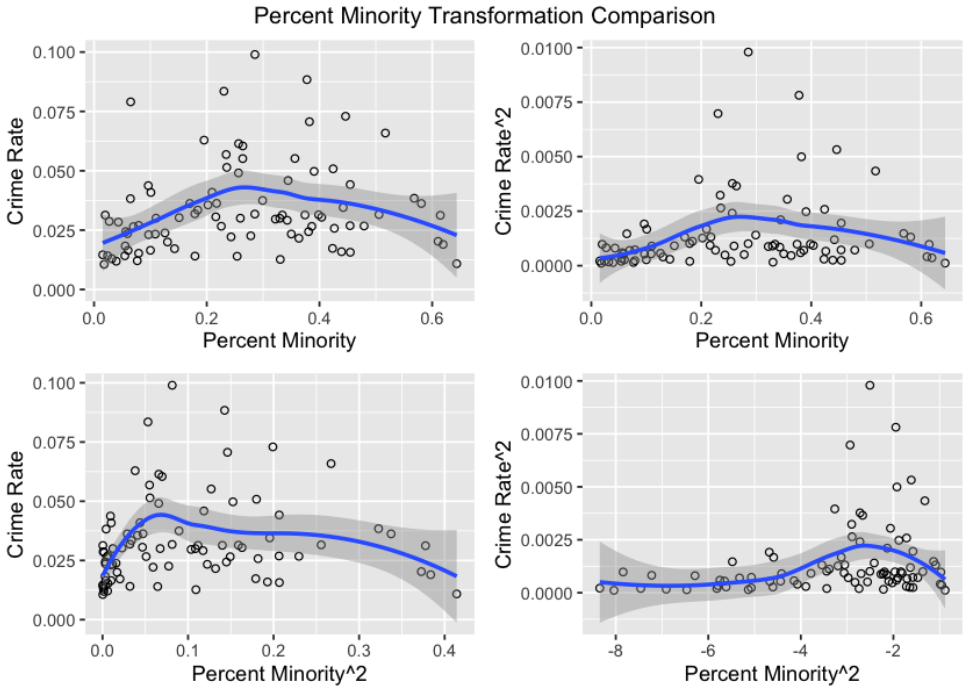
In [11]:
```
# Correlation plot comparison for pctymle
options(repr.plot.width=7, repr.plot.height=7)
p_pctymle_no_outlier <- ggplot(data = no_outliers, aes(x = pctymle, y = crmrte)) + geom_point(pch = 1) + xlab("Percent Young Male (Out
p_pctymle_olog <- ggplot(data = no_outliers, aes(x = pctymle, y = log(crmrte))) + geom_point(pch = 1) + xlab("Percent Young Male") + y
p_pctymle_plog <- ggplot(data = no_outliers, aes(x = log(pctymle), y = crmrte)) + geom_point(pch = 1) + xlab("ln(Percent Young Male)")
p_pctymle_dlog <- ggplot(data = no_outliers, aes(x = log(pctymle), y = log(crmrte))) + geom_point(pch = 1) + xlab("ln(Percent Young Ma
#p_pctymle_sqrt <- ggplot(data = no_outliers, aes(x = sqrt(pctymle), y = crmrte)) + geom_point(pch = 1) + xlab("Square Root of Percent
#p_pctymle_dsqrt <- ggplot(data = no_outliers, aes(x = sqrt(pctymle), y = sqrt(crmrte))) + geom_point(pch = 1) + xlab("Square Root of

grid.arrange(p_pctymle, p_pctymle_no_outlier, p_pctymle_olog, p_pctymle_plog, p_pctymle_dlog, ncol=2, top="Percent Young Male Transforr
```



2.5.4 Percent Minority

The relationship between percent minority and crime rate appears to be parabolic, therefore, we tried squaring crime rate, percent minority, and both variables. These transformations do not provide much improvement. We see a change in the direction of the line at 26% minority, so we decided to create an indicator variable (1 for values at or above 26% and 0 for values less than 26%). Although the line appears to be decreasing above 26% minority in the scatter plot, the mean crime rate is actually higher for counties with more than 26% minorities due the clustering of low data points in the bottom right of the plot (3.73 vs. 2.97).

In [12]:
```
# Correlation plot comparison for pctmin80_rescaled
options(repr.plot.width=7, repr.plot.height=5)
p_pctmin80_rescaled_psq <- ggplot(data = no_outliers, aes(x = pctmin80_rescaled, y = crmrte^2)) + geom_point(pch = 1) + xlab("Percent M
p_pctmin80_rescaled_osq <- ggplot(data = no_outliers, aes(x = pctmin80_rescaled^2, y = crmrte)) + geom_point(pch = 1) + xlab("Percent M
p_pctmin80_rescaled_dsq <- ggplot(data = no_outliers, aes(x = log(pctmin80_rescaled^2), y = crmrte^2)) + geom_point(pch = 1) + xlab("Pe
grid.arrange(p_pctmin80_rescaled, p_pctmin80_rescaled_psq, p_pctmin80_rescaled_osq, p_pctmin80_rescaled_dsq, ncol=2, top="Percent Minor
```



3. Regression Models

In this section, we fit four linear regression models to test our assumptions about the determinants of crime. In these models we include demographic, geographic, economic, and criminal justice system factors that we selected from the dataset (outlined in Table 2 above). Models 3 and 4 also include the data transformations discussed in the previous section. The p-values for the coefficients in each model are calculated using heteroskedasticity robust standard errors.

In the following sections, we discuss the justification for what we include in each model and evaluate model robustness based on adherence to the six Classical Linear Model (CLM) assumptions for the multiple linear regression model. For each model, we examine plots to check for zero conditional mean, heteroskedasticity, and normality of the error term. We also evaluate the Breusch-Pagan test for heteroskedasticity, Shapiro-Wilk normality test, and the variance inflation factor (VIF) for models with more than one independent variable.

In [13]:
```
# Define models
m1 = lm(data$crmrte ~ density, data=data)
m2 = lm(data$crmrte ~ density + taxpc + pctymle + pctmin80_rescaled + prbarr, data=data)
m3 = lm(no_outliers$crmrte ~ density + taxpc + pctymle + high_pctmin + log(prbarr), data=no_outliers)
m4 = lm(no_outliers$crmrte ~ density + taxpc + pctymle + high_pctmin + log(prbarr) + central + west, data=no_outliers)

# Get robust standard errors from robust covariance matrix
invisible((se.m1 = sqrt(diag(vcovHC(m1)))))
invisible((se.m2 = sqrt(diag(vcovHC(m2)))))
invisible((se.m3 = sqrt(diag(vcovHC(m3)))))
invisible((se.m4 = sqrt(diag(vcovHC(m4)))))


# Pass robust standard errors into stargazer
stargazer(m1, m2, m3, m4, type="text",
          omit.stat = c("f"),
          se = list(se.m1, se.m2, se.m3, se.m4),
          star.cutoffs = c(0.05, 0.01, 0.001),
          dep.var.labels=c("Crime Rate", ""),
          covariate.labels=c("Population Density", "Tax Value Per Capita", "Percent Young Male", "Percent Minority", "Probability of A
```

```
================================================================================================
                                          Dependent variable:
                                ---------------------------------------------------------------
                                                    Crime Rate
                                      (1)             (2)             (3)             (4)
------------------------------------------------------------------------------------------------
Population Density                 0.009***        0.007***        0.009***        0.009***
                                    (0.001)         (0.002)         (0.001)         (0.001)

Tax Value Per Capita                                0.0004          0.0001          0.00001
                                                   (0.0004)        (0.0002)        (0.0002)

Percent Young Male                                  0.178**         0.204*          0.194*
                                                    (0.066)         (0.099)         (0.087)

Percent Minority                                    0.029**
                                                    (0.009)

Probability of Arrest                               -0.029*
                                                    (0.012)

High Percent Minority                                               0.010***         0.005
                                                                    (0.002)         (0.004)

ln(Probability of Arrest)                                           -0.006          -0.005
                                                                    (0.004)         (0.004)

Central Region                                                                      -0.005
                                                                                    (0.004)

West Region                                                                         -0.008*
                                                                                    (0.004)

Constant                           0.021***         -0.006          -0.010          -0.001
                                    (0.002)         (0.013)         (0.009)         (0.010)

------------------------------------------------------------------------------------------------
Observations                           88              88              86              86
R2                                    0.526           0.712           0.718           0.735
Adjusted R2                           0.521           0.694           0.700           0.711
Residual Std. Error       0.013 (df = 86) 0.010 (df = 82) 0.010 (df = 80) 0.010 (df = 78)
================================================================================================
Note:                                                          *p<0.05; **p<0.01; ***p<0.001
```

3.1 Model 1 Evaluation

$$crmrte = \beta_0 + \beta_1 density + u$$

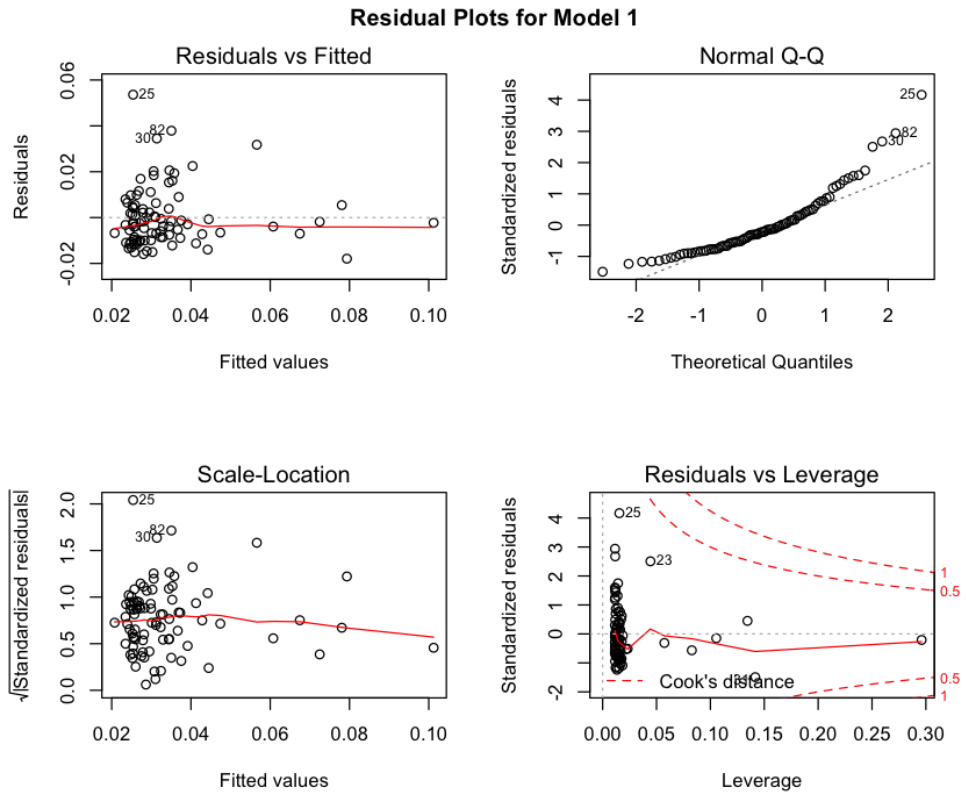
Overview: Our base model only includes population density which, in a bivariate model, explains 52.2% of the variation in the crime rate in North Carolina (adjusted R-squared). We include this bivariate model because it helps to demonstrate the importance of the population density variable in our analysis. The coefficient of population density (0.009) is highly statistically significant.

Robustness:

1. **MLR.1 Linearity in Parameters** - By definition this assumption always holds unless restrictions are put on the error term. The models that we assessed are all linear in their parameters.
2. **MLR.2 Random Sampling** - While our data is not from a random sample, it does comprise observations from the majority of counties in North Carolina. Although we have a large number of counties, sampling bias may affect our models if counties not included in the data are missing for a systematic reason.
3. **MLR.3 No Perfect Collinearity** - Since this model only includes one variable, this is not a concern.
4. **MLR.4 Zero Conditional Mean** - The residuals vs. fitted values plot for Model 1 (below) shows that the red spline curve is flat for the most part except for a tiny bump to the left. Our data does not majorly violate this assumption.
5. **MLR.5 Homoskedasticity** - The residuals vs. fitted values and scale-location plots for Model 1 (below) show that our error term is not exactly homoskedastic. Rather than a consistent band of data points, our data is fanned out on the left side and slightly caves in as we move to the right. However, this could be due to the few outliers on the top left. The output for Breusch-Pagan test has a p-value of 0.98, thus, we fail to reject the null hypothesis that the data is homoskedastic. The residual vs. leverage plot for Model 1 (below) shows that none of the data points are outside of Cook's Distance.
6. **MLR.6 Normality** - The Normal Q-Q plot (below) shows that our model violates this assumption. We see the data points diverge from the diagonal line in the bottom left and top right corner of the plot. Running a Shapiro Wilk test confirms the violation, with a p-value less than 0.05. However, our sample size is large enough that it is unlikely that we need to worry about violating this assumption.

Conclusion: Except for lack of homoskedasticity, the model does not violate any CLM assumptions. While this model helps to highlight the importance of population density as a predictor of crime rate, it is too sparse and does not include other independent variables that may impact crime crime rate. To fully evaluate the research question, we will need to include other potential determinants of crime rate.

In [14]:
```
options(repr.plot.width=7, repr.plot.height=6)

# Residual plots for model 1
par(mfrow=c(2,2))
plot(m1)
title("Residual Plots for Model 1", outer=TRUE, line = -2)
```



In [15]:
```
# Breusch-Pagan test for heteroskedasticity
bptest(m1)

# Shapiro Wilk test for normality
shapiro.test(m1$residuals)
```

```
        studentized Breusch-Pagan test

data:  m1
BP = 0.00059426, df = 1, p-value = 0.9806


        Shapiro-Wilk normality test

data:  m1$residuals
W = 0.87318, p-value = 3.915e-07
```

3.2 Model 2 Evaluation

$$crmrte = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 pctymle + \beta_4 pctmin80\_rescaled + \beta_5 prbarr + u$$

Overview: This model tests the majority of determinants of crime rate that we selected from the dataset, including demographic, criminal justice system, and economic factors. We introduce four additional independent variables: tax per capita, percent young male, percent minority, and probability of arrest. All the variables, except tax per capita, are statistically significant. With the addition of these four variables we are now able to explain 69.4% of the variation in crime rate, 17.3% more than the base model.

The adjusted R-squared improves significantly for this model, indicating that the model now explains 71.2% of the variation in crime rate vs, 52.6% in Model 1. The coefficient of population density has the same level of statistical significance but decreases slightly from 0.009 to 0.007. Percent young male and percent minority have moderate significance with coefficients of 0.178 and 0.029 respectively. Probability of arrest has lower statistical significance with a coefficient of -0.029. Tax per capita is the only variable that is statistically insignificant with a coefficient of 0.0004.

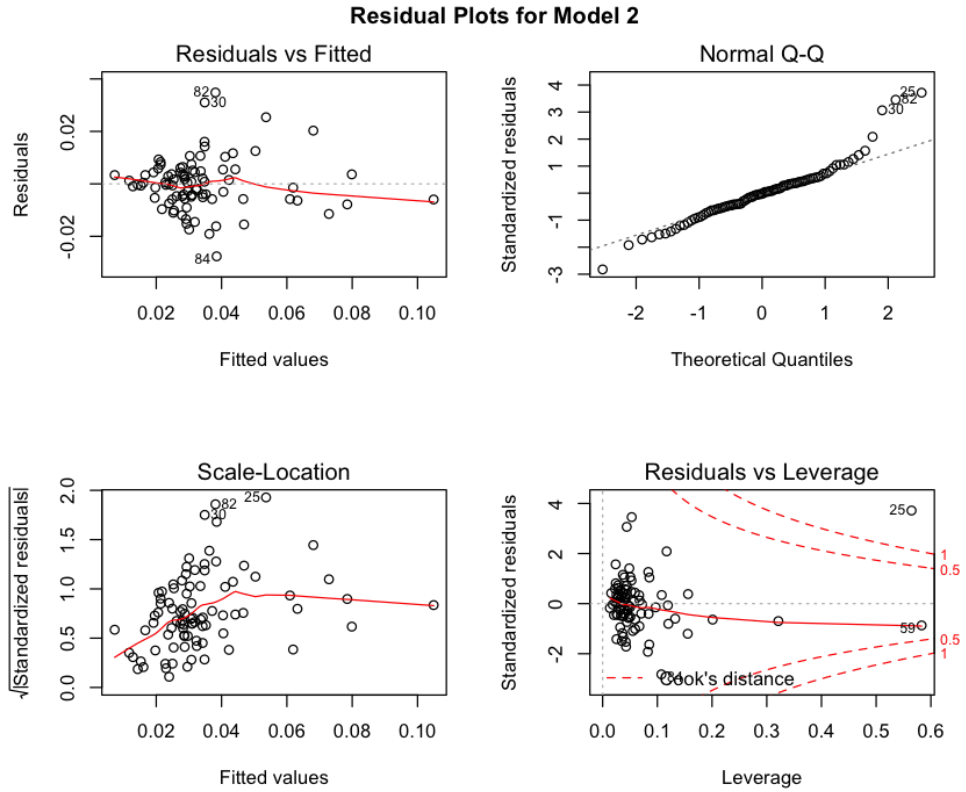
Robustness:

1. **MLR.1 Linearity in Parameters** - By definition this assumption always holds unless restrictions are put on the error term. The models that we assessed are all linear in their parameters.
2. **MLR.2 Random Sampling** - While our data is not from a random sample, it does comprise observations from the majority of counties in North Carolina. Although we have a large number of counties, sampling bias may affect our models if counties not included in the data are missing for a systematic reason.
3. **MLR.3 No Perfect Collinearity** - We ran correlations between our indicator variables to check for perfect collinearity. The variables included in our models have either no statistically significant correlation or a low correlation. The output of VIF test for our model is slightly above 1 at a maximum of 1.2. All of the coefficients in this model have values less than 2. Since only values greater than 10 indicate a collinearity problem, we do not have perfect multicollinearity.
4. **MLR.4 Zero Conditional Mean** - The residuals vs. fitted values plot for Model 2 (below) shows that our data violates this assumption. If this assumption was true, we would expect to see the data centered around zero as we move from the left side to the right side of the plot. However, this could be due to the one data point on the extreme right that the red spline curve is slightly dipping on the right.
5. **MLR.5 Homoskedasticity** - The residuals vs. fitted values and scale-location plots for Model 2 (below) show that our error term is not homoskedastic. Rather than a consistent band of data points, our data clusters on the left side of the plots and fans out as you move to the right. The output for Breusch-Pagan test has a p-value of 0.01, allowing us to reject the null hypothesis that the data is homoskedastic. The residual vs. leverage plot for Model 2 (below) also shows that we have one data point outside of Cook's Distance and data point very close to the Cook's Distance line. This indicates that we outliers in the data that needs further investigation.

6. **MLR.6 Normality** - The Normal Q-Q plot (below) shows that our model violates this assumption. We see the data points diverge from the diagonal line in the bottom left and top right corner of the plot. Running a Shapiro Wilk test confirms the violation, with a p-value of 0.006. However, our sample size is large enough that it is unlikely that we need to worry about violating this assumption.

Conclusion: Model 2 strikes a good balance between accuracy and parsimony and the reduction in degrees of freedom is compensated by an increase in predictive power. However, violations of the CLM assumptions and outliers indicate that the model could potentially benefit from data transformations in order to produce more accurate estimates of coefficients.

In [16]:
```
# Residual plots for model 2
par(mfrow=c(2,2))
plot(m2)
title("Residual Plots for Model 2", outer=TRUE, line = -2)
```



In [17]:
```
# Breusch-Pagan test for heteroskedasticity
bptest(m2)

# VIF test for multicollinearity
vif(m2)

# Shapiro Wilk test for normality
shapiro.test(m2$residuals)
```

```
        studentized Breusch-Pagan test

data:  m2
BP = 19.889, df = 5, p-value = 0.001311
```

density	1.26426119671963
taxpc	1.15702194334385
pctymle	1.06589791591537
pctmin80_rescaled	1.03484462263727
prbarr	1.18337014310068

```
        Shapiro-Wilk normality test

data:  m2$residuals
W = 0.95807, p-value = 0.006185
```

3.3 Model 3 Evaluation

$$crmrte = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 pctymle + \beta_4 high\_pctmin + \beta_5 log(prbarr) + u$$

Overview: In this model we introduce several data transformations. As explained in the data transformations section, these include removing outlier values in tax revenue per capita and percent young male, taking the log of the probability of arrest, and introducing an indicator variable for high percentage of minorities (at or above 26%). The removal of two outliers reduces the sample size of the underlying data to 86 observations, from 88 in the previous models. Other than the variable transformations, Model 3 is identical to Model 2. This allows us to compare the effect of adding the variable transformations.

The adjusted R-squared improves slightly for this model, indicating that the model now explains 71.8% of the variation in crime rate vs. 71.2% in Model 2. Although this is a small increase in adjusted R-squared, the inclusion of transformations also results in other improvements (described below). In Model 3, the coefficient on population density retains the same level of statistical significance, but increases from 0.007 to 0.009. Percent young male decreases in significance from the 0.01 level to the 0.05 level, but the coefficient

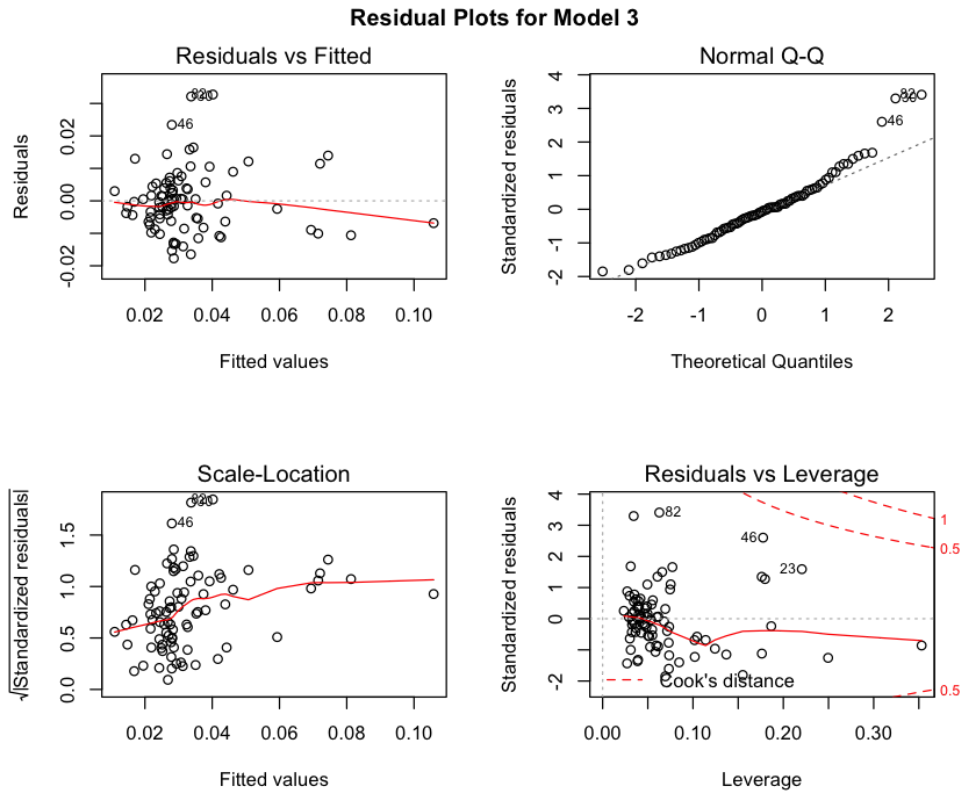
increases from 0.178 to 0.204. As with percent minority, the new high percent minority indicator has a positive relationship with crime rate and is statistically significant at the 0.001 level with a coefficient of 0.010. The log of the probability of arrest continues to have a negative relationship with with crime rate, but is no longer significant. As with Model 2, tax revenue per capita remains insignificant.

Robustness:

1. **MLR.1 Linearity in Parameters** - By definition this assumption always holds unless restrictions are put on the error term. The models that we assessed are all linear in their parameters.
2. **MLR.2 Random Sampling** - While our data is not from a random sample, it does comprise observations from the majority of counties in North Carolina. Although we have a large number of counties, sampling bias may affect our models if counties not included in the data are missing for a systematic reason.
3. **MLR.3 No Perfect Collinearity** - We ran correlations between our indicator variables to check for perfect collinearity. The variables included in our models have either no statistically significant correlation or a low correlation. This is confirmed through the low VIFs for each coefficient in this model (see below). All of the coefficients in this model have values less than 2. Values greater than 10 indicate a collinearity problem.
4. **MLR.4 Zero Conditional Mean** - The residuals vs. fitted values plot for Model 3 (below) shows that our data violates this assumption. If this assumption was true, we would expect to see the data centered around zero as we move from the left side to the right side of the plot. On the right hand side of the plot some of the deviation from zero conditional mean is likely due to the sparsity of data for larger fitted values.
5. **MLR.5 Homoskedasticity** - The pattern of the data points in the residuals vs. fitted values and scale-location plots for Model 3 (below) does appear to be more random than in Model 2, but the data still clusters on the left side of the plots. This is likely due, in part, to the sparsity of data points with higher fitted values. The Breusch-Pagan test for heteroskedasticity (below) is not significant (p-value 0.2852), therefore, we fail to reject the null hypothesis of homoskedasticity for Model 3. We also see the spread of the residuals shrink from Model 2 to Model 3, indicating that our predictions are likely more accurate. The residual vs. leverage plot for Model 3 (below) shows that no data points are outside of Cook's Distance, an improvement over Model 2. This indicates that it is unlikely that outliers are unduly influencing the model.
6. **MLR.6 Normality** - The Normal Q-Q plot (below) shows that the variable transformations appear to have improved normality in the left tail, but the model still violates the assumption of normality of the error term. This is confirmed by the Shapiro Wilk test of normality (below) where we reject the null hypothesis of normality (p-value = 0.00223). However, our sample size is large enough that it is unlikely that we need to worry about violating this assumption.

Conclusion: Improvements in zero conditional mean, heteroskedasticity, and normality indicate that this model likely produces less biased coefficient estimates than Model 2. Both Models 2 and 3 test the majority of determinants of crime rate that we selected from the dataset (demographic, criminal justice system, and economic factors). However, they do not include geographic determinants.

In [18]:
```
# Residual plots for model 3
par(mfrow=c(2,2))
plot(m3)
title("Residual Plots for Model 3", outer=TRUE, line = -2)
```



In [19]:
```
# Breusch-Pagan test for heteroskedasticity
bptest(m3)

# VIF test for multicollinearity
vif(m3)

# Shapiro Wilk test for normality
shapiro.test(m3$residuals)
```

```
        studentized Breusch-Pagan test

data:  m3
BP = 6.222, df = 5, p-value = 0.2852
```

density	1.56859574009779
taxpc	1.3976835338896
pctymle	1.1656115668176
high_pctmin	1.01506852155128
log(prbarr)	1.29297858463822

```
        Shapiro-Wilk normality test

data:  m3$residuals
W = 0.95, p-value = 0.00223
```

3.4 Model 4 Evaluation

$$crmrte = \beta_0 + \beta_1 density + \beta_2 taxpc + \beta_3 pctymle + \beta_4 high\_pctmin + \beta_5 log(prbarr) + \beta_6 central + \beta_7 west + u$$

Overview: To fully test our research question, Model 4 includes all of the potential determinants of crime rate that we selected from the dataset (demographic, criminal justice system, economic, and geographic factors). As well as the data transformations introduced in Model 3, we introduce two additional independent variables: the west and central region indicators. The addition of these two variables reduces the degrees of freedom, but does not provide a large increase in adjusted R-squared. The model now explains 71.1% of the variation in crime rate vs. 70% of the variation in Model 3.

Both the central region and west region indicators have a negative relationship with crime rate, but the central region is insignificant, while the west region is significant at the 0.05 level with a coefficient of -0.008. The addition of the region indicators does not change the significance level or coefficient for population density. However, it does cause changes in both percent young male and the high percent minority indicator.

Percent young male remains significant at the 0.05 level, but the coefficient drops slightly from 0.204 to 0.194. The high percent minority indicator goes from being significant at the 0.001 level to being insignificant. These changes in the demographic variables indicate that the region variables likely also explain demographic differences and are picking up some of the signal of the demographic variables. Both tax value per capita and log of probability of arrest remain insignificant in Model 4.

Robustness:

1. **MLR.1 Linearity in Parameters** - By definition this assumption always holds unless restrictions are put on the error term. The models that we assessed are all linear in their parameters.
2. **MLR.2 Random Sampling** - While our data is not from a random sample, it does comprise observations from the majority of counties in North Carolina. Although we have a large number of counties, sampling bias may affect our models if counties not included in the data are missing for a systematic reason.
3. **MLR.3 No Perfect Collinearity** - We ran correlations between our indicator variables to check for perfect collinearity. The variables included in our models have either no statistically significant correlation or a low correlation. This is confirmed through the low VIFs for each coefficient in this model (see below). All of the coefficients in this model have values less than 2.5. Values greater than 10 indicate a collinearity problem.
4. **MLR.4 Zero Conditional Mean** - The data appears to be more centered around zero in the residuals vs. fitted values plot for Model 4 (below) vs. the Model 3 plot. However, while there is less deviation from zero on the right hand side of the plot, there is now more deviation on the left hand side of the plot. As with previous models, some of the deviation from zero conditional mean is likely due to the sparsity of data for larger fitted values.
5. **MLR.5 Homoskedasticity** - The pattern of the data points in the residuals vs. fitted values and scale-location plots for Model 4 (below) is similar to that of the Model 3 plots, however, the spread of the residuals has increased. This indicates that we have less precision in predictions. The Breusch-Pagan test for heteroskedasticity (below) is not significant (p-value 0.512), therefore, we fail to reject the null hypothesis of homoskedasticity for Model 4. As with Model 3, the residual vs. leverage plot for Model 4 (below) shows that no data points are outside of Cook's Distance.
6. **MLR.6 Normality** - The Normal Q-Q plot (below) is similar to that of Model 3 and still shows a violation of the normality assumption at both tails. This is confirmed by the Shapiro Wilk test of normality (below) where we reject the null hypothesis of normality (p-value = 0.000424). However, our sample size is large enough that it is unlikely that we need to worry about violating this assumption.

Conclusion: The addition of the region indicators does not cause a meaningful increase in adjusted R-squared. It also appears that they measure demographic factors that we have already controlled for in the model. Moreover, Model 4 shows an increase in the spread of the residuals, indicating a decrease in accuracy. Overall, the addition of the region variables does not appear to be worth the loss in precision.

In [ ]:
```
# Residual plots for model 4
par(mfrow=c(2,2))
plot(m4)
title("Residual Plots for Model 4", outer=TRUE, line = -2)
```

In [ ]:
```
# Breusch-Pagan test for heteroskedasticity
bptest(m4)

# VIF test for multicollinearity
vif(m4)

# Shapiro Wilk test for normality
shapiro.test(m4$residuals)
```

3.5 Model Comparison

In our analysis of each model (previous section), Model 3 appears to best reflect our understanding of the determinants of crime and strike the best balance between accuracy and parsimony. The transformations introduced in Model 3 appear to reduce the spread of the residuals and therefore, the accuracy of our predictions. They also remove outliers that may unduly influence the predictions in Model 2. In addition, Model 3 aligns well with our research question because it tests the majority of the potential determinants of crime rate that we selected from the data (demographics, criminal justice system, and economic factors). Although we had originally hypothesized that region could be an important determinant, it appears that region proxies demographic differences that we already control for with percent minority and percent young male.

To further compare the models we ran analysis of variance (ANOVA) tests to see which models provide the best, most parsimonious fit of the data (below). Since we introduced transformations and dropped outliers in Models 3 and 4, we only compared the first two models without the transformations and the last two models with the transformations. In the test of Model 1 and Model 2, we reject the null hypothesis that the simpler model has a better fit (p-value 2.40431e-08), indicating that there is a benefit to adding the additional independent variables. In the test of Model 3 and 4, we fail to reject the null hypothesis that the simpler model, Model 3, has a better fit than the more complex model, indicating that there is limited benefit to adding the additional geographic indicator variables.

As we continue to see moderate violations of the CLM assumptions with Model 3, we ran one additional test on Model 3. The Ramsey Regression Equation Specification Error Test (RESET) evaluates the null hypothesis that no higher-order polynomial terms are necessary. This test can indicate whether more data transformations or a non-linear model are necessary to accurately model the data. The RESET test (below) fails to reject the null hypothesis (p-value 0.0632) that no higher-order polynomial terms are necessary. This indicates that the linear functional form is appropriate for this data and that it is unlikely that we need to add additional higher-order data transformations.

We conclude that Model 3 provides the most accurate estimates of the factors associated with crime rate in North Carolina counties.

In [22]:
```
# ANOVA test to compare m1 and m2
anova(m1, m2)

# ANOVA test to compare m2 and m3
anova(m3, m4)
```

A df[,6]: 2 × 6

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
86	0.014456231	NA	NA	NA	NA
82	0.008793822	4	0.005662409	13.20011	2.40431e-08

A df[,6]: 2 × 6

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
80	0.007878525	NA	NA	NA	NA
78	0.007403049	2	0.0004754762	2.504856	0.08823932

In [23]:
```
# RESET test for non-linear functional form for m3
resettest(m3, type="regressor")
```

```
        RESET test

data:  m3
RESET = 1.8763, df1 = 10, df2 = 70, p-value = 0.0632
```

3.6 Interpretation of Results

Model 3 indicates that in North Carolina,

- Holding other variables constant, population density, percent young male, and high percent minority are significantly associated with increased crime rates. Percentage of young males shows the largest association with increased crime rates. - Holding other variables constant, an increase of 10 percentage points in the percentage of young males is associated with a 2.04 percentage point increase in crime rate.
- Holding other variables constant, having 26 percent or more minorities is associated with a 1 percentage point increase in crime rate.
- Holding other variables constant, an additional 10 people per square mile is associated with a 0.09 percentage point increase in crime rate.

4. Omitted Variables

Omitted variable bias occurs when a variable that is correlated with the dependent variable and one or more independent variables are left out of a model. We hypothesize that the following omitted variables may be impacting the accuracy of our results.

Bias in Policing - Although we have a variable for probability of arrest, it does not tell us how many of the arrests are false arrests or due to bias in policing. For example, there could be racial bias at play where police target certain groups for stop and check/arrest which could lead to a higher probability of arrests. Omitted variable bias would likely shift the slope coefficient for probability of arrest towards zero. This is because we hypothesize that bias in policing is positively correlated with crime rate and positively correlated with probability of arrest, which has a negative slope coefficient. We hypothesize that not having this variable introduces moderate bias into our model.

Recidivism Rate - Recidivism is a tendency to relapse into criminal behavior. We expect that this variable would be positively correlated with crime rate and the probability of arrest. Similar to bias in policing (above), we would expect the omitted variable bias will drive the slope coefficient for probability of arrest towards zero. We also hypothesize that this omitted variable introduces moderate bias into our model.

Disparities in Wages - The dataset provides average weekly wage for different industries, but does not provide information about disparities in wages. Income inequality is an important factor in understanding opportunities available to residents of a county, poverty, and affluence. Omitted variable bias will move the slope coefficient of tax revenue per capita toward zero. This is because we hypothesize that wage disparity is positively correlated with crime rate and negatively correlated with tax revenue per capita, which has a positive slope coefficient. However, it is also possible that the distribution of wages does not result in a change in overall tax revenue.

If disparities in wages are greater among minorities vs. non-minorities, we might also see an effect on percent minority. Omitting wage disparities from the model would likely bias the slope coefficient of percent minority away from zero. This is because percent minority has a positive slope coefficient and we hypothesize that wage disparity is positively correlated with crime rate and percent minority. Overall, we hypothesize that this omitted variable introduces moderate bias into our model.

Unemployment Rate - One of the biggest factors that encourages people to engage in crime is unemployment. Omitted variable bias for this variable will shift the slope coefficient for probability of arrest toward zero. This is because we hypothesize that unemployment rate is positively correlated with crime rate and probability of arrest, which has a negative slope coefficient. We hypothesize that this omitted variable has a moderate effect on our model. This variable could potentially be a proxy for 'taxpc' variable, as people do not pay income tax when unemployed.

High school dropout rate - We assume that there is a positive correlation between high school dropouts and crime rate, as people who dropout of school have a hard time finding a well paying job and a job that is stable. There is also a positive correlation between high school dropout rate and probability of arrest, which makes the omitted variable bias positive. Since probability of arrest has a negative correlation with crime rate, omitted variable bias will drive the slope coefficient for probability of arrest toward zero. This variable might be measuring a similar effect as unemployment rate, so it would not be a good idea to include both in a model. High school dropout rate could be a proxy for percent minority as minority youth are more likely to dropout of high school than non-minority.

Table 3: Potentially omitted variables and direction of bias

Omitted Variable	Impacted Variable	Direction of Bias

Bias in Policing	Probability of arrest	Towards zero (less negative)
Recidivism Rate	Probability of arrest	Towards zero (less negative)
Disparities in Wages	Tax per capita	Towards zero (less positive)
Disparities in Wages	Percent Minority	Away from zero (more positive)
Unemployment Rate	Probability of arrest	Towards zero (less negative)
High School Dropout Rate	Probability of arrest	Towards zero (less negative)

Overall, we hypothesize that the omitted variables discussed above introduce moderate bias into our model. Based on our analysis, we could be overestimating the coefficients of probability of arrest, tax per capita, and percent minority. While the other omitted variables can be measured, bias in policing and recidivism rates are theoretically important, but hard to measure in practice.

5. Conclusion

Our analysis concludes that in North Carolina:

- Holding other variables constant, population density, percent young male, and high percent minority are significantly associated with increased crime rates.
- Holding other variables constant, the percentage of young males in a county shows the largest association with increased crime rates.

These results are moderately robust. Even after introducing data transformations, our selected model shows moderate violations of the CLM zero conditional mean, homoskedasticity, and normality assumptions. In interpreting these results, it is also important to remember that the original dataset only contained data for 90 out of 100 counties in North Carolina (four additional observations were dropped due to data quality issues and outlier values). If the omitted counties have different characteristics than the counties included in the dataset, our analysis may suffer from selection bias. Omitted variables may also be biasing our estimates. We hypothesize that we may be overestimating the coefficients of probability of arrest, tax per capita, and percent minority. Lastly, the data used in this analysis is more than 30 years old, which likely impacts its reliability for policy-making in 2019.

Based on our findings and model robustness, we recommend that local governments in North Carolina adopt the following policies:

- Provide more support and opportunities to unemployed young men and minorities.
- Focus crime prevention interventions in areas with higher population densities.

We also recommend the following next steps for this analysis:

- Obtain more recent data to ensure accuracy of results.
- Obtain data for all counties to ensure that the analysis is truly representative of all counties in North Carolina.
- Include data for potentially omitted variables, if available.