# Lab 3: Reducing Crime

*Guangyu (Gary) Pei, Kirby Bloom and Sophia Huang*

*April 15, 2018*

## Contents

## List of Figures

## List of Tables

# 1 Introduction

The purpose of this report was to examine the crime statistics in selected counties in North Carolina, particularly to understand the factors that cause criminal incidents (crime rate) and to investigate the relationship between crime rate and related factors, such as population density, demographic composition, and law enforcement. Based on the findings, recommendations can be made on policies to reduce the crime rate.

Using crime rate as the dependent variable, this report investigated the following research questions:

- To what extent did urbanization as reflected by population density affect the crime rate in some North Carolina counties?
- To what extent did the certainty of punishment (i.e., criminals are expected to be caught and face punishment) reduce crime rate in North Carolina counties?
- What were the effect of the young male minority population on the crime rate in the areas studied?
- Did the severity of punishment, such as long prison sentences, have effect on the crime rate in North Carolina counties?

The report covers (a) exploratory data analysis of the crime dataset in Section 2; (b) modeling procedure in Section 3 and data analysis in Section 4; (c) discussion on omitted variables in Section 5; and (d) conclusion in Section 6. [1]

# 2 Exploratory Data Analysis

Total observations contained 97 records (n=97). Three issues in the data file required data cleaning procedure.

1. Missing values: The last 6 rows of the data file contained missin values.
2. Variable recode: Recode categorical variable, `prbconv`, into numeric variable.
3. Duplicated records in the data file were removed.

```
crime = read.csv("crime_v2.csv")
# drop NA's in the last 6 rows
crime = crime %>% na.omit()
# prbconv was considered as factor due to NA, convert back to numeric
crime$prbconv = as.numeric(as.character(crime$prbconv))
```

```
duplicated_county = crime$county[which(duplicated(crime$county))]
# The duplicated entries with example fields
crime[crime$county==duplicated_county, 1:8]
```

```
   county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc
88    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
89    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
```

```
# remove the duplicated entry
crime = crime[!duplicated(crime$county),]
```

```
# The largest wser
crime[which.max(crime$wser), ]
```

```
   county year    crmrte   prbarr prbconv  prbpris avgsen     polpc
84    185   87 0.0108703 0.195266 2.12121 0.442857   5.38 0.0012221
     density    taxpc west central urban pctmin80     wcon    wtuc
84 0.3887588 40.82454    0       1     0  64.3482 226.8245 331.565
       wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
84 167.3726 264.4231 2177.068 247.72 381.33 367.25 300.13 0.04968944
      pctymle
84 0.07008217
```

Figure 1 indicated that county 185 had extreme high weekly wages for service industry (variable `wser`) with value of $2177.07, The average annual income of the county was $113208. As shown in the report by BLS, average weekly

[1] Appendix provides additional information such as `R` packages used in this report, full correlation matrix and additional plots etc for completeness. The reader does not need to read the appendix in order to understand this report.

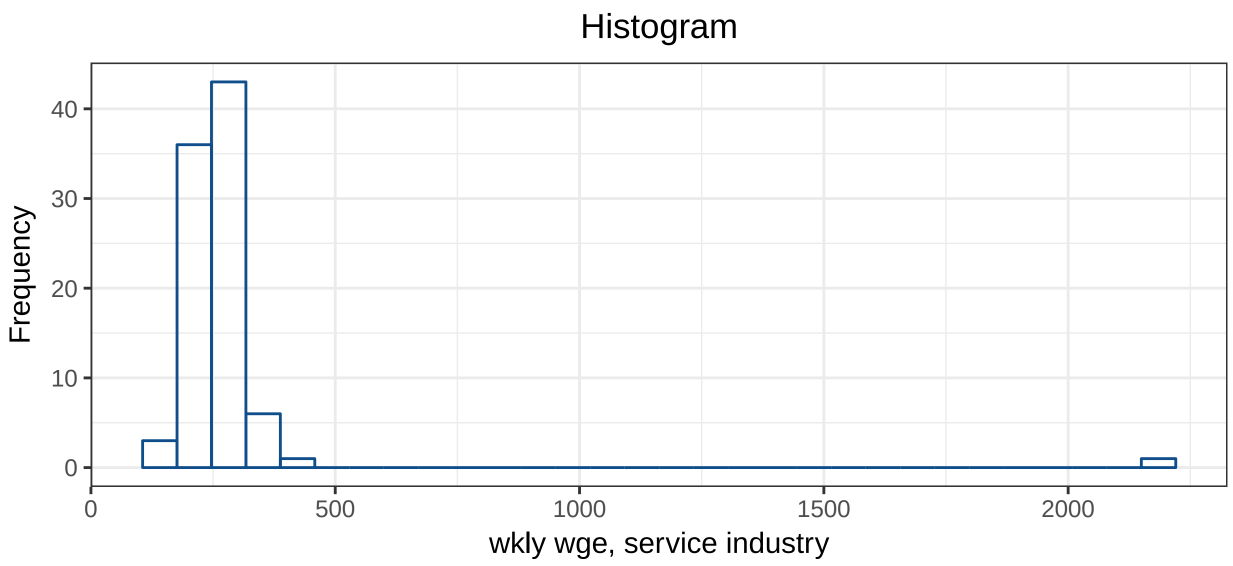

Figure 1: Extreme large weekly wages for `wser`

wage of North Carolina was $932 with the highest weekly wage of $1,254 in County Durham. Granted, the data in this report was in 2016 while we were looking at data in 1987. However, inflation was a possible reason for the wage increase from 1987 to 2016. Moreover, wage in non-urban county tended to be lower than urban city. Therefore, we believe that one county contained a erroneous data row. Second, the probability of conviction (variable `probconv`) of county 185 is 2.12121. The value was an outlier. Given these observations, we believe county 185 data was erroneous and we exclude it from our analysis as shown below.

```r
# The fowlloing function is from attached Lab3Unitls.R, which is shown in Appendix.
print(hist_plot(crime, xvar = "wser", xlab = "wkly wge, service industry"))
```

```r
# Drop the entry 84 with the unrealistic wage entry
crime = crime[-which.max(crime$wser),]
```

Finally, a new categorical variable `county_loc` was created as a location indicator. The variable combined indicator variables `west`, `central`, and `urban`. `county_loc` with values, "ER","W","C", and "U". They represented east rural counties, west counties, central counties and urban counties respectively.

```r
# create county_loc and combine west, central and urban
crime = crime %>% mutate(county_loc = case_when(west == 0 & central == 0 & urban == 0 ~ "ER",
    west == 1 & central == 0 & urban == 0 ~ "W", central == 1 & urban == 0 ~ "C",
    urban == 1 ~ "U")) %>% mutate(county_loc = factor(county_loc, levels=c("ER","W","C","U")))
```

Several records in variables, probability of arrest (`prbarr`) and probability of conviction (`prbconv`) contained values that were greater than 1. Because multiple offenses and multiple convictions are possible from an arrest, in addition to false arrests, we included these data points.

The final data file for analysis contained 89 records. Table 1 provides the summary of the data.

```r
stargazer(crime %>%
              select(-one_of(c("county_loc","west","central","urban","county","year"))),
          summary.stat = c("min","p25","median","mean","p75","max"),
          type = "latex", title = "Summary of crime data", label = "tab:summary", header=FALSE)
```

```r
brks = seq(0, 0.12, 0.005)
distribution_plts(crime, brks = brks)
```

The range and statistics of these variables in above table suggested no obvious errors.

In the following data visualization section, we choose variable `crmrte` (crimes committed per person) as our dependent variable and density, probability of conviction, probability of arrest, percent of young male, and percent of minority etc. as independent variables.

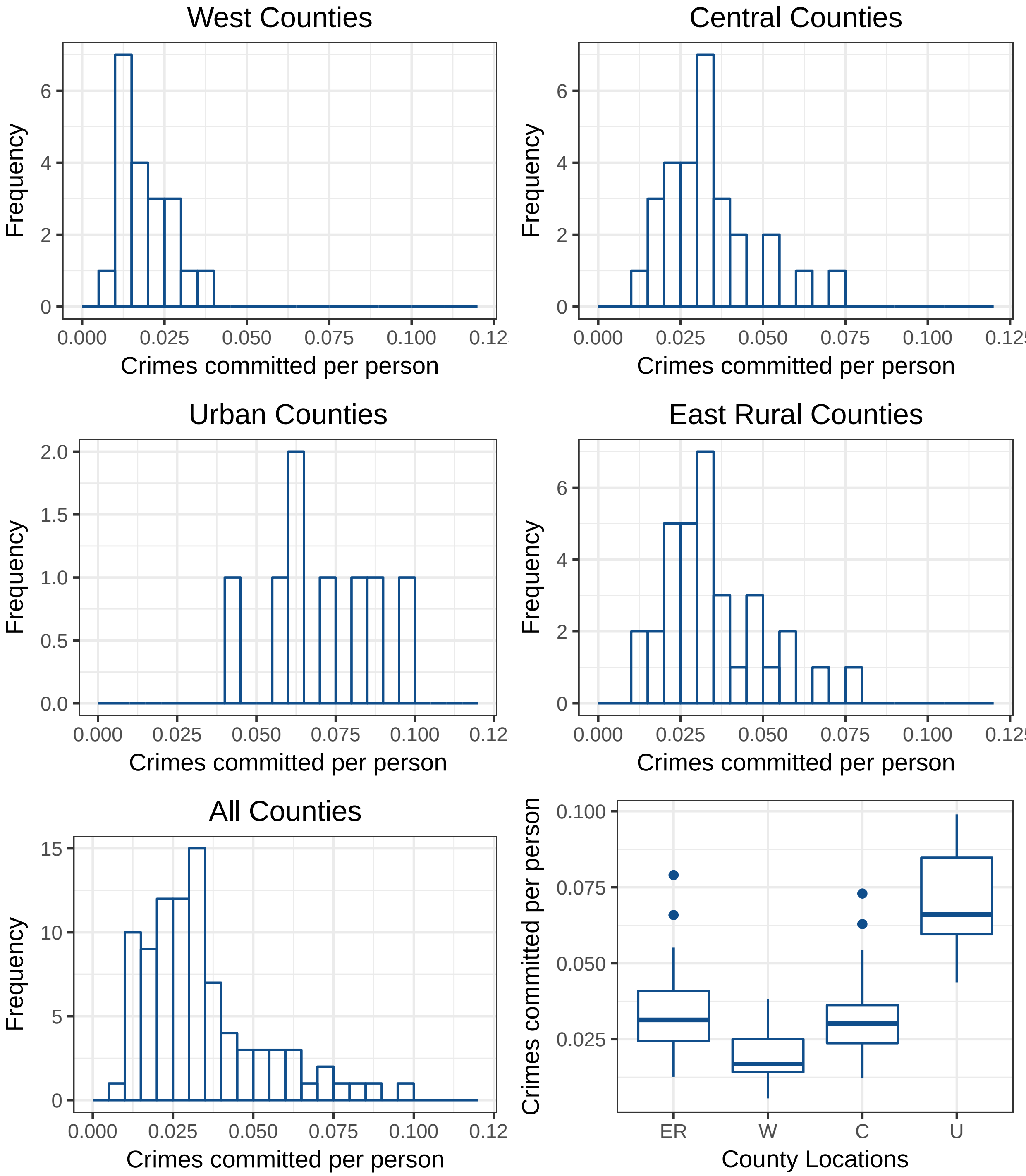

Figure 2: Distribution characteristics of variable *crmrte* at various geographic areas

Table 1: Summary of crime data

| Statistic | Min | Pctl(25) | Median | Mean | Pctl(75) | Max |
|---|---|---|---|---|---|---|
| crmrte | 0.006 | 0.022 | 0.030 | 0.034 | 0.041 | 0.099 |
| prbarr | 0.093 | 0.207 | 0.272 | 0.296 | 0.345 | 1.091 |
| prbconv | 0.068 | 0.343 | 0.451 | 0.533 | 0.574 | 1.671 |
| prbpris | 0.150 | 0.364 | 0.421 | 0.410 | 0.458 | 0.600 |
| avgsen | 5.450 | 7.420 | 9.120 | 9.737 | 11.510 | 20.700 |
| polpc | 0.001 | 0.001 | 0.001 | 0.002 | 0.002 | 0.009 |
| density | 0.00002 | 0.548 | 0.996 | 1.447 | 1.570 | 8.828 |
| taxpc | 25.693 | 30.696 | 34.870 | 38.131 | 41.072 | 119.761 |
| pctmin80 | 1.284 | 10.005 | 24.312 | 25.279 | 38.061 | 61.942 |
| wcon | 193.643 | 250.836 | 281.426 | 286.011 | 315.164 | 436.767 |
| wtuc | 187.617 | 375.235 | 406.504 | 411.798 | 441.595 | 613.226 |
| wtrd | 154.209 | 191.172 | 203.016 | 211.411 | 224.722 | 354.676 |
| wfir | 170.940 | 288.462 | 317.308 | 322.264 | 342.682 | 509.466 |
| wser | 133.043 | 229.015 | 253.010 | 253.970 | 276.263 | 391.308 |
| wmfg | 157.410 | 289.430 | 321.900 | 337.025 | 360.210 | 646.850 |
| wfed | 326.100 | 403.150 | 449.840 | 443.308 | 478.480 | 597.950 |
| wsta | 258.330 | 329.220 | 357.690 | 357.633 | 383.720 | 499.590 |
| wloc | 239.170 | 297.190 | 308.050 | 312.417 | 329.160 | 388.090 |
| mix | 0.020 | 0.081 | 0.102 | 0.130 | 0.152 | 0.465 |
| pctymle | 0.062 | 0.075 | 0.078 | 0.084 | 0.084 | 0.249 |

The histograms, Figure 2 showed the distribution of`crmrte` for the west, central, east rural, urban counties and combined from all counties. The boxplot suggested that the urban counties had higher crime rate than other counties. Distributions for west, central and east rural counties were similar. Figure 2 suggested that increaing of density (urbanization) was associated with crime rate (`crmrte`).

Crime rate increased with population density while decreased with probability of arrest and conviction. Probability of sentence and average sentence days had weak positive correlation. Crime rate increased with police per capita, which was unexpected. However, police per capita increased with people density and high population density corresponded to high crime rate.

```
correlation.matrix = cor(crime %>%
              select(-one_of(c("county_loc","west","central","urban","county","year"))))
ggcorrplot(correlation.matrix, colors = c("#6D9EC1", "white", "#E46726"),
           tl.cex = 8, outline.color = "white", legend.title = "Correlation") +
theme(legend.position="top", legend.direction = "horizontal", plot.margin = margin(0,0,0,0,"pt"),
      legend.text=element_text(size=8), legend.title=element_text(size=8))
```

To investigate the overall patterns of correlations among all variables, we plot Figure 3 to show the heat map of correlation matrix[2]. Several features of the data stand out. First, all wages are highly positively correlated as indicated by the square area from variable `wcon` to variable `wloc`. Second, the correlations are high among variables `density`, `taxpc` and `polpc` which can be easily identified in the square area formed by these three variables. Third, all variables related to wages are correlated with `density`, `taxpc` and `polpc`, which is clear from the two symmetric rectangle areas formed by the intersection of these variables. It is particularly true with `density`. Clearly, these visible features indicate that urbanization is associated with the strong correlations among these variables. Finally, if we focus on our outcome variable `crmrte`, it is negatively correlated with law enforcement variables, namely, `prbarr` and `prbconv`. Furthermore, `crmrte` is positively correlated with all aforementioned variables that are associated with urbanization especially with variable `density`.

```
scatter_plts(crime, xvars = c("density", "prbarr", "prbconv", "prbpris","avgsen","polpc"),
             xlabs = c("People per square mile", "Probability of arrest",
                       "Probability of conviction", "Probability of prison sentence",
                       "Average sentence (days)", "Police per capita"))
```

[2] The full correlation matrix is provided in Table 6.

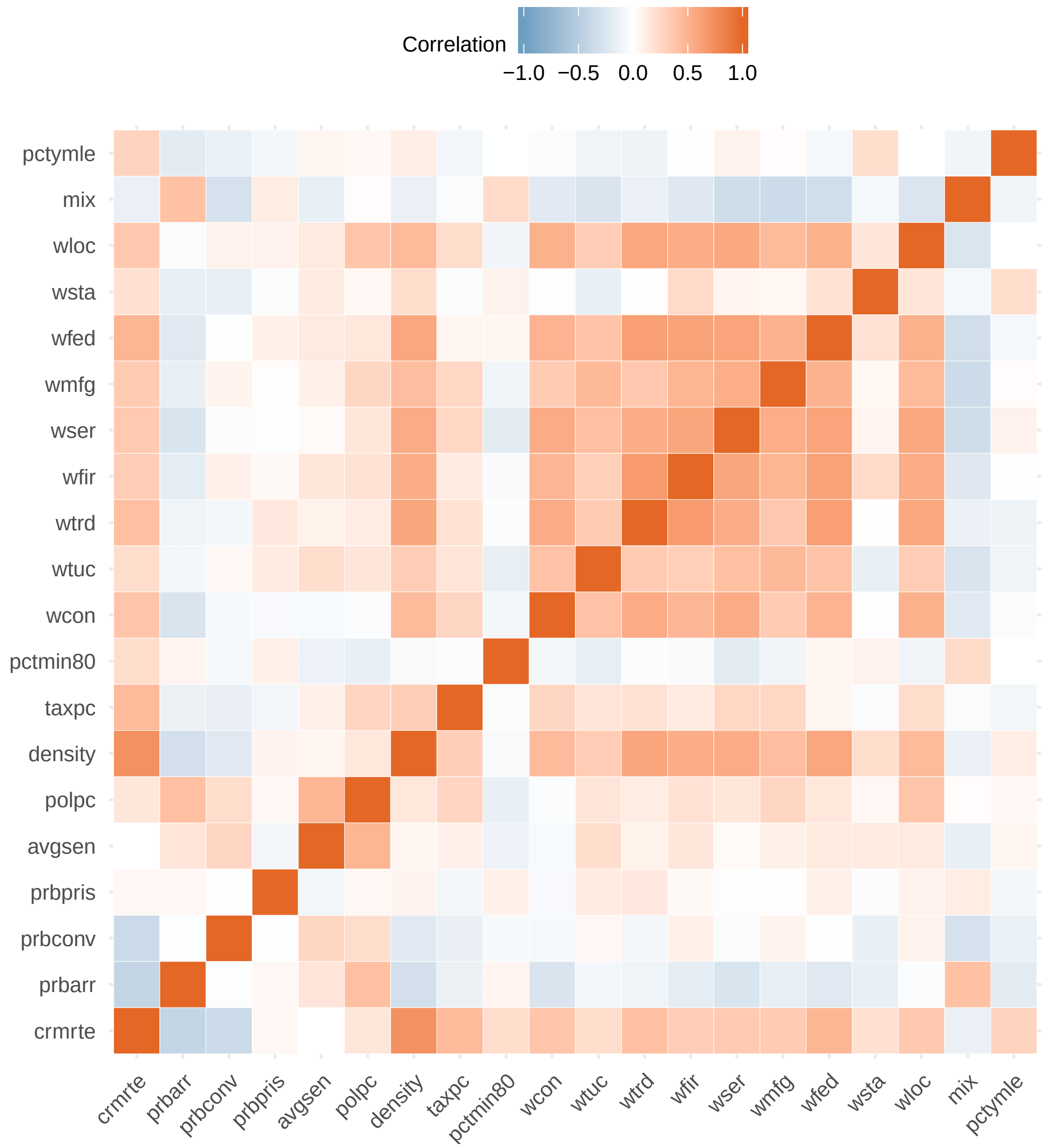

Figure 3: Correlation Heatmap

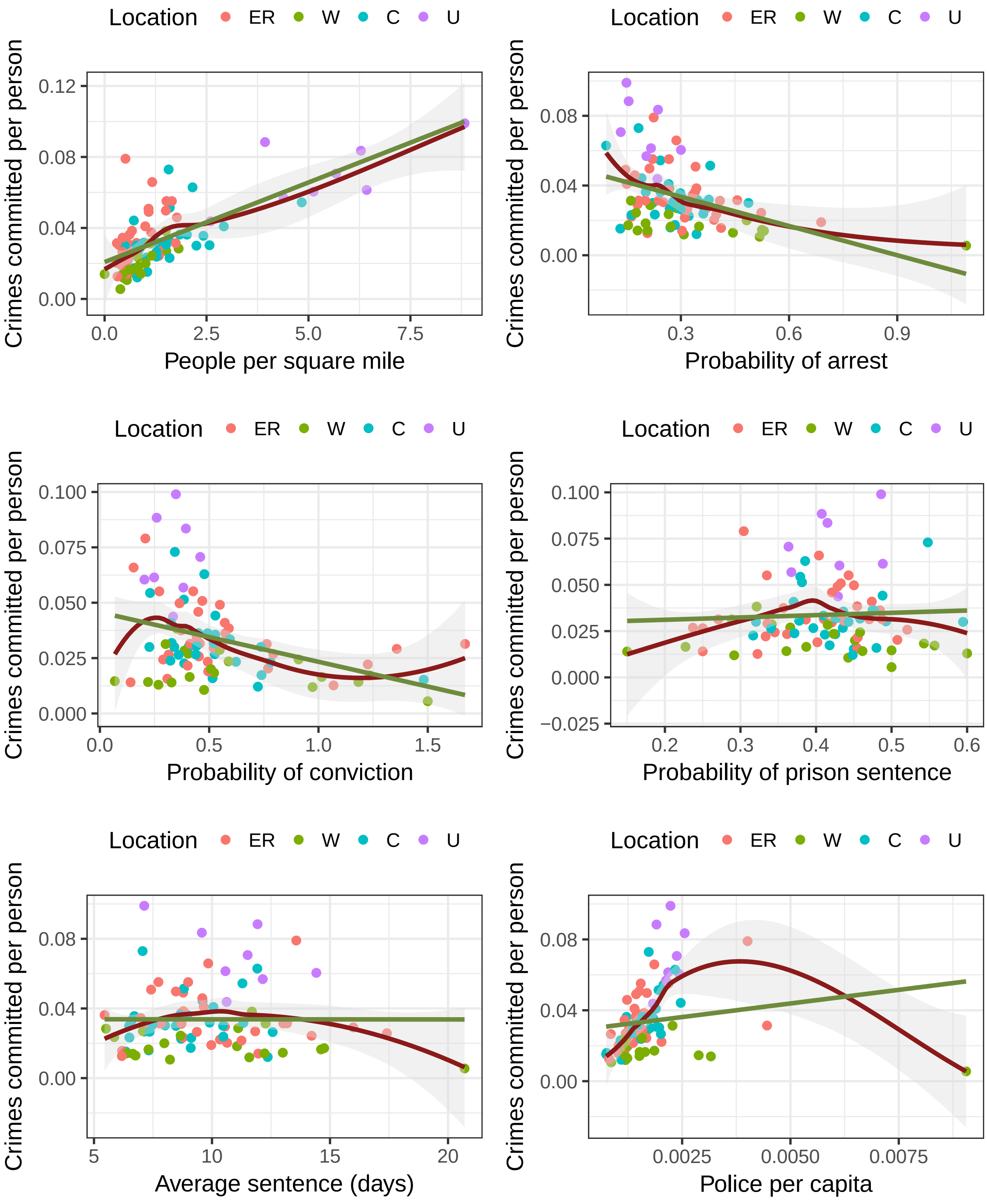

Figure 4: Scatter plots for variables *density*, *prbarr*, *prbconv*, *prbpris*, *avgsen* and *polpc*

Figure 4 showed bivariate scatter plots between several independent variables and our outcome variable. These scatter plots suggested that crime rate increased as density increased and decreased as probability of arrest and conviction increased. Probability of sentence and average sentence days had slightly positive correlation. Crime rate increased with police per capita, which may seem unexpected. Police per capita and crime rate increased as density increased.

```
scatter_plts(crime, xvars = c("taxpc", "pctmin80", "mix", "pctymle", "wcon", "wtuc"),
             xlabs = c("Tax revenue per capita", "Percent of minority, 1980",
                       "Offense mix: face-to-face/other", "Percent young male",
                       "Weekly wage, construction", "Weekly wage, trns, util, commun"))
```

Figure 5 showed the scatter plots between crime rate and variables `taxpc`, `pctmin80`, `mix`, `pctymle`, `wcon` and `wtuc`. `taxpc`. These plots suggested medium correlation between crime rate, pecent young male, and percent miniority.

Figure 15 and Figure 16 plot the scatter plots for the rest of variables in the data set for completeness. Allthese weekly wage variables exhibit very similar characteristics as `wcon` and `wtuc` in Figure 5.

```
brks = seq(0, 9, 0.5)
distribution_plts(crime, xvar = "density", brks = brks, xlab = "People per square mile")
```

Scatter plots indicated that `density` (people per square mile) was a strong proxy for urbanization and crime rate had strong correlation with density as shown in the scatter plot. The plot csuggested high population density was directly associated with high degree of urbanization.

Figure 6 plotted the histograms for counties located in the west, central, east rural areas and urban areas. It indicated that urban counties had high population density while west, central and east rural counties have similar low density distributions.

```
brks = seq(0, 1.2, 0.05)
distribution_plts(crime, xvar = "prbarr", brks = brks, xlab = "Probability of arrest")
```

```
brks = seq(0, 2.2, 0.05)
distribution_plts(crime, xvar = "prbconv", brks = brks, xlab = "Probability of conviction")
```

Figure 7 showed the characteristics of variable `prbarr`. Urban counties had lower probability of arrest while we show previously that urban counties had higher crime rate. This indicated the public policy was not developed along with the process of urban development. Similarly, Figure 8 showed that urban counties had distribution toward lower probability of convictions. Again, this could be due to law enforcement and justice systems in urban counties were not able to handle all criminal cases from arrests.

```
multiple_hist_plts(crime,
                   histvars = c("prbpris", "avgsen", "pctmin80", "pctymle", "mix", "polpc"),
                   xlabs = c("Probability of prison sentence","avg. sentence, days",
                             "perc. minority, 1980", "percent young male",
                             "offense mix: face-to-face/other", "police per capita"))
```

Figure 9 showed the histograms for the variables `prbpris`, `avgsen`, `pctmin80`, `pctymle`, `mix` and `polpc`. No specific error data points stood out. The percentage of minority spread wide range from couple percentages to 60% and yet the correlation with crime rate was very low.

The histograms for the rest of variables indicated less relevancy with variables mainly responsed to `density` variable. For completeness, curious reader can find these histograms in Figure 17 and Figure 18.

# 3 Models

## 3.1 Indepdent variable selection and transformation

EDA results indicated that all metric variables are reasonably distributed. No variable has highly skewed distributions and/or with range of values that span several order of magnitudes. Thus, transformation such as logarithmic transformation on the variables was not necessary. In the following model build process, we examined the residuals to identify any additional needs for transformation based on the characteristics of the residuals from our models.

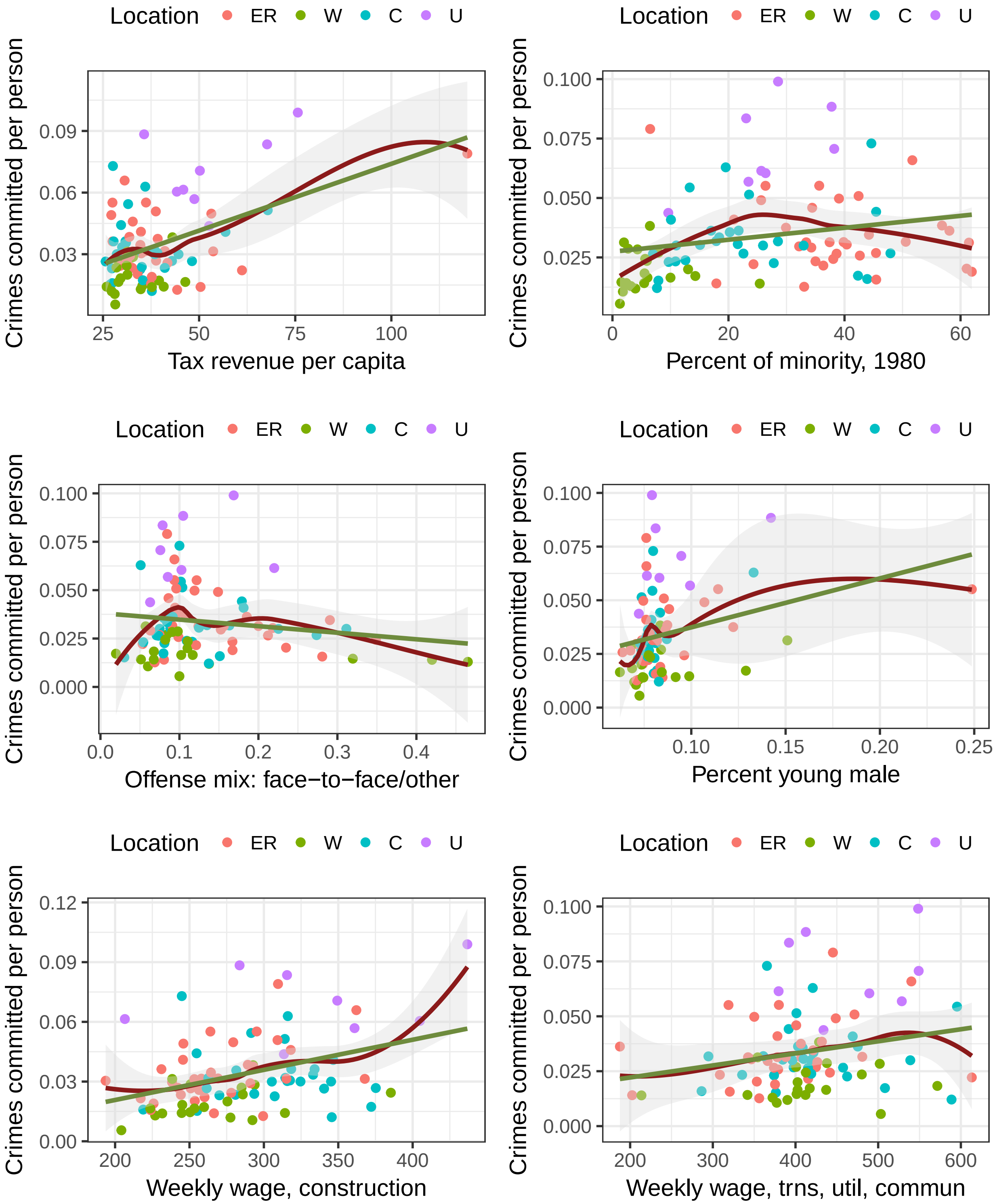

Figure 5: Scatter plots for variables *taxpc*, *pctmin*80, *mix*, *pctymle*, *wcon* and *wtuc*

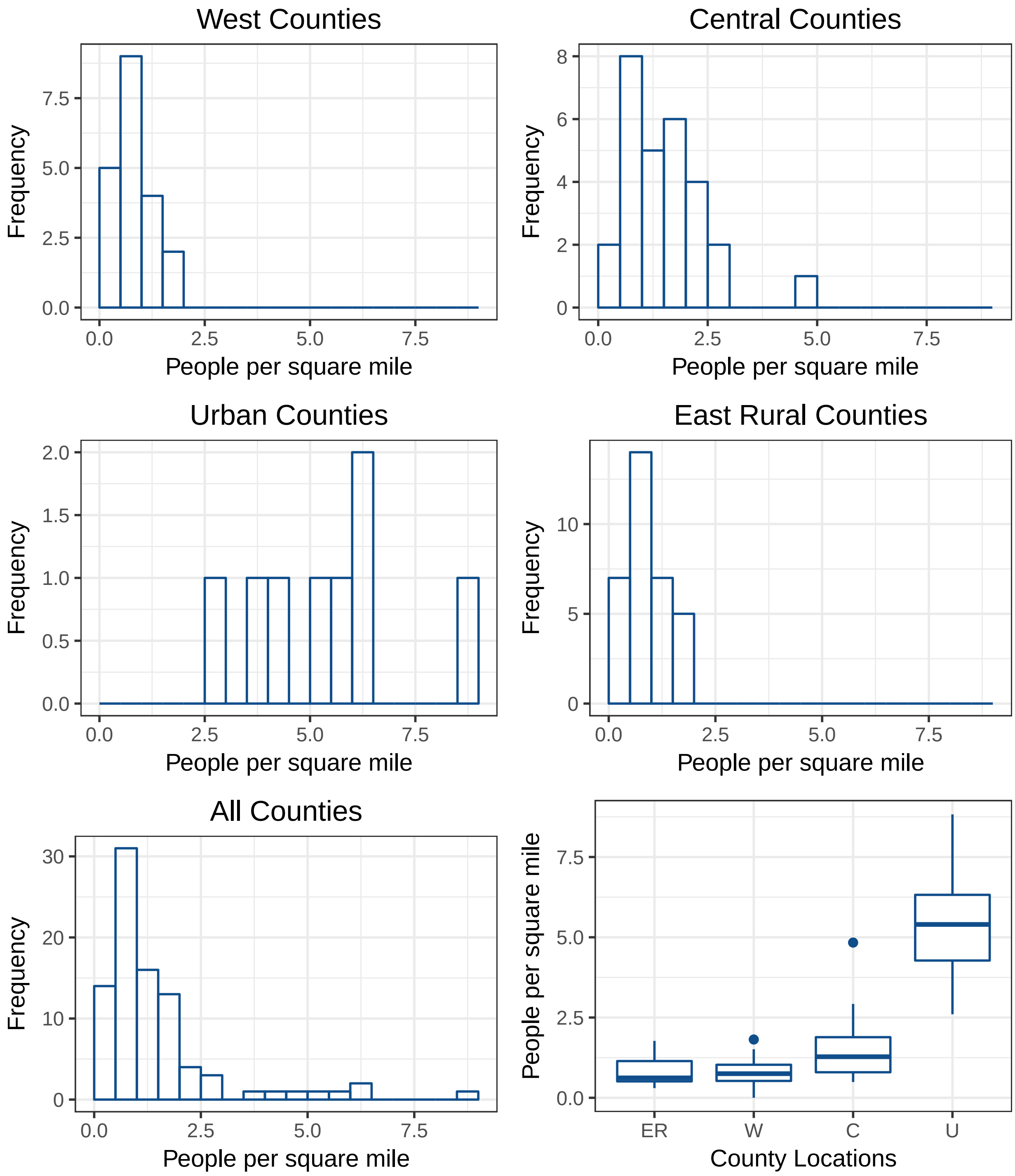

Figure 6: Distribution characteristics of variable *density* at various geographic areas

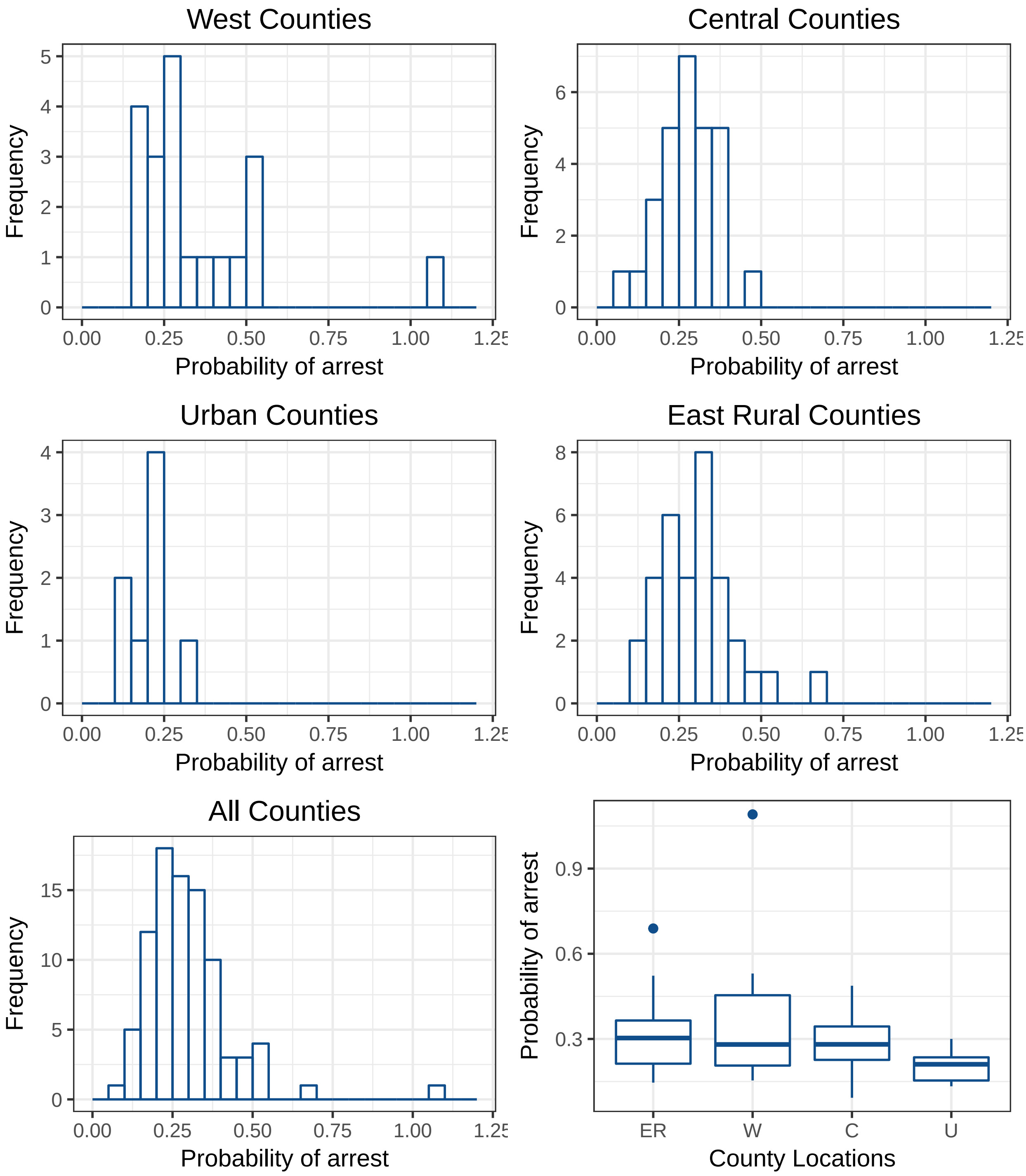

Figure 7: Distribution characteristics of variable *prbarr* at various geographic areas

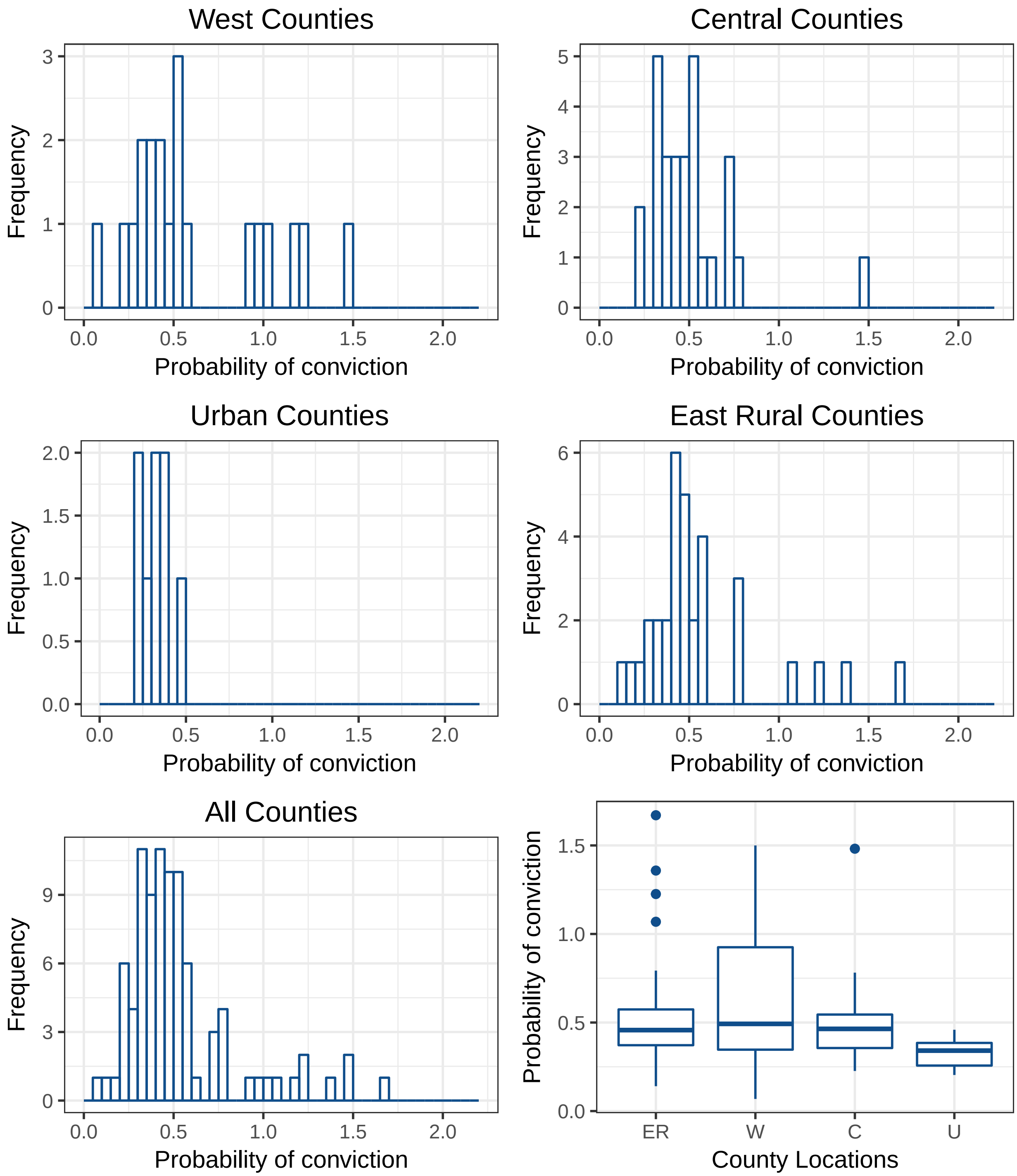

Figure 8: Distribution characteristics of variable *prbconv* at various geographic areas

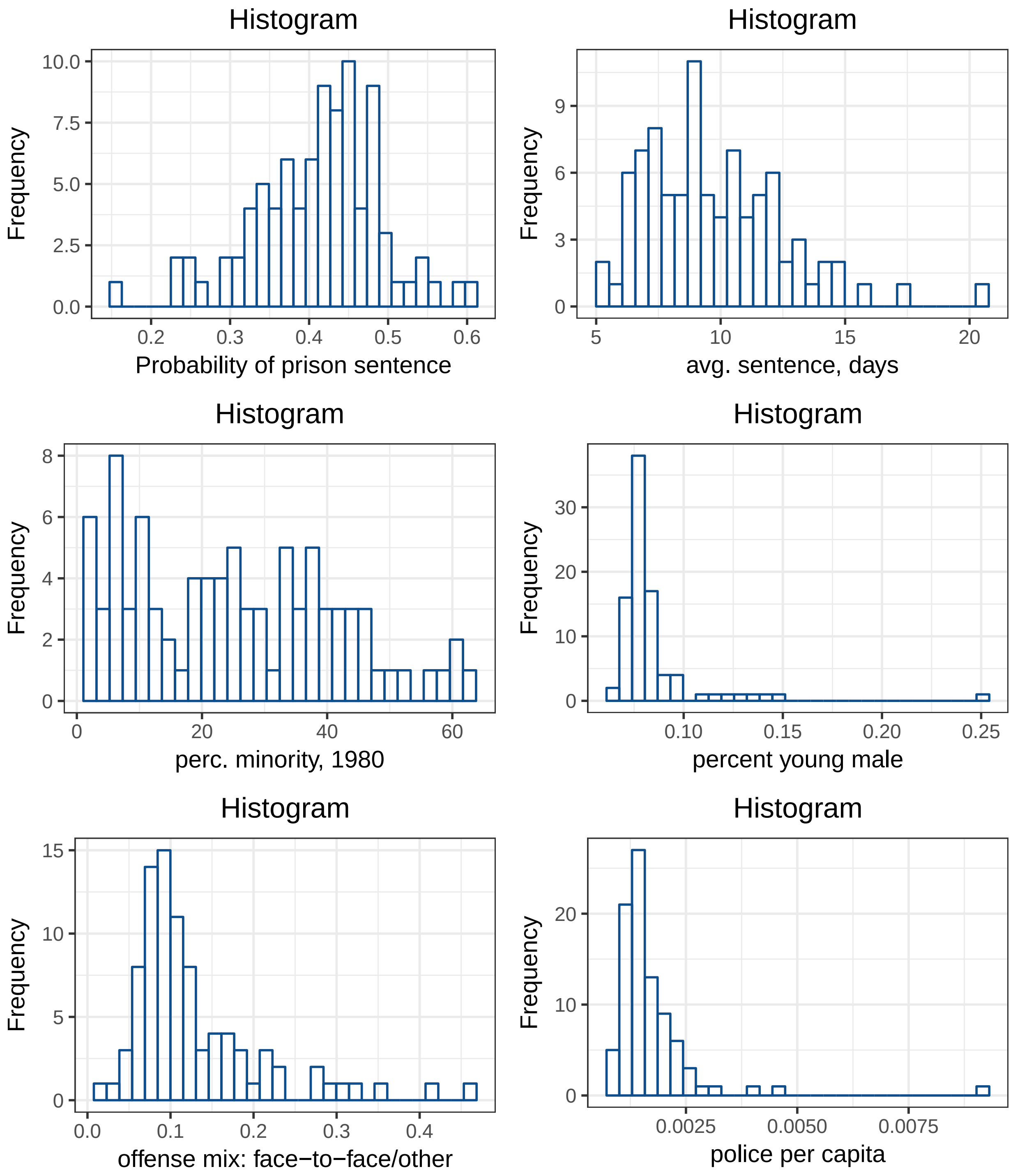

Figure 9: Histograms of variables $prbpris$, $avgsen$, $pctmin80$, $pctymle$, $mix$ and $polpc$

## 3.2 Models with only the explanatory variables of key interest

Our first basic Model 1a is expressed as following.

$$\text{crmrte} = \beta_0 + \beta_1 \times \text{density} + u$$

```
(model1a = lm (crmrte ~ density, data = crime))
```

```

Call:
lm(formula = crmrte ~ density, data = crime)

Coefficients:
(Intercept)      density
   0.020776     0.008973
```

```
ols_diag_plot(model1a, ncol = 2)
```

Figure 10 showed the diagnosis plot of our OLS model. The fitted value verse residuals showed the residuals were close to zero but overall below the perfect zero horizontal line. The zero conditional mean assumption is clearly violated and it indicates that there are non-trial omitted variables for Model 1a. Scale-location plot indicates certain heteroskedasticity of this model. The normal Q-Q plot showed residuals deviates from normal distribution. There was no data points whose Cook's distance is too high to be concerned. We tested the statistical significance using heteroskedasticity-robust residual estimates as shown below.

```
coeftest(model1a, vcov = vcovHC)
```

```

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.02077623 0.00180328  11.521 < 2.2e-16 ***
density     0.00897318 0.00079398  11.302 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result showed that `density` coefficient was statistically significant.

Our second basic Model 1b was expressed as following.

$$\text{crmrte} = \beta_0 + \beta_0 \times \text{prbarr} + \beta_1 \times \text{prbconv} + u$$

```
(model1b = lm (crmrte ~ prbarr + prbconv, data = crime))
```

```

Call:
lm(formula = crmrte ~ prbarr + prbconv, data = crime)

Coefficients:
(Intercept)       prbarr      prbconv
    0.06289     -0.05706     -0.02290
```

```
ols_diag_plot(model1b, ncol = 2)
```

```
coeftest(model1b, vcov = vcovHC)
```

```

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.062886   0.016246  3.8708 0.0002107 ***
prbarr      -0.057059   0.039218 -1.4549 0.1493385
```

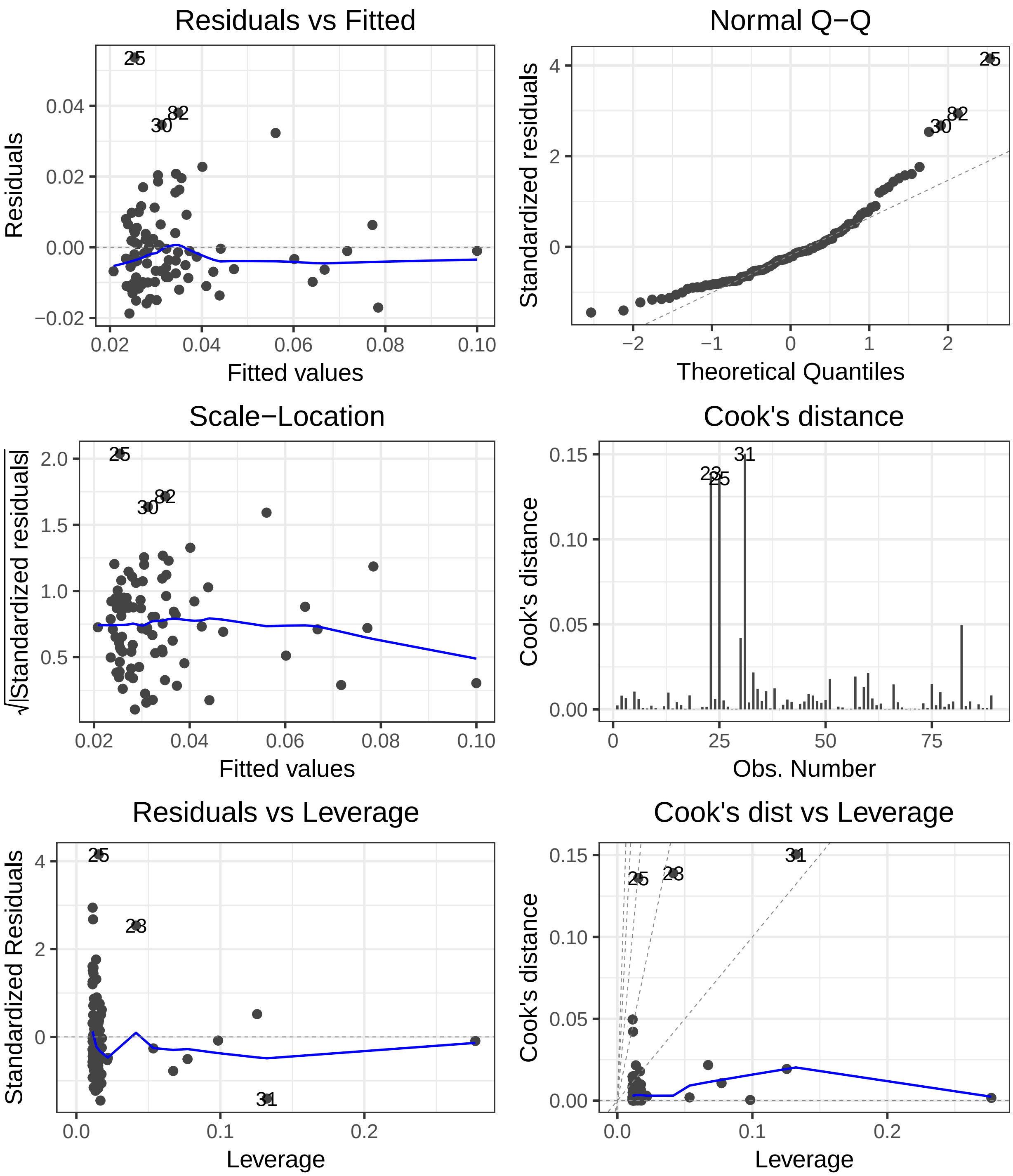

Figure 10: Model 1a OLS diagnostics

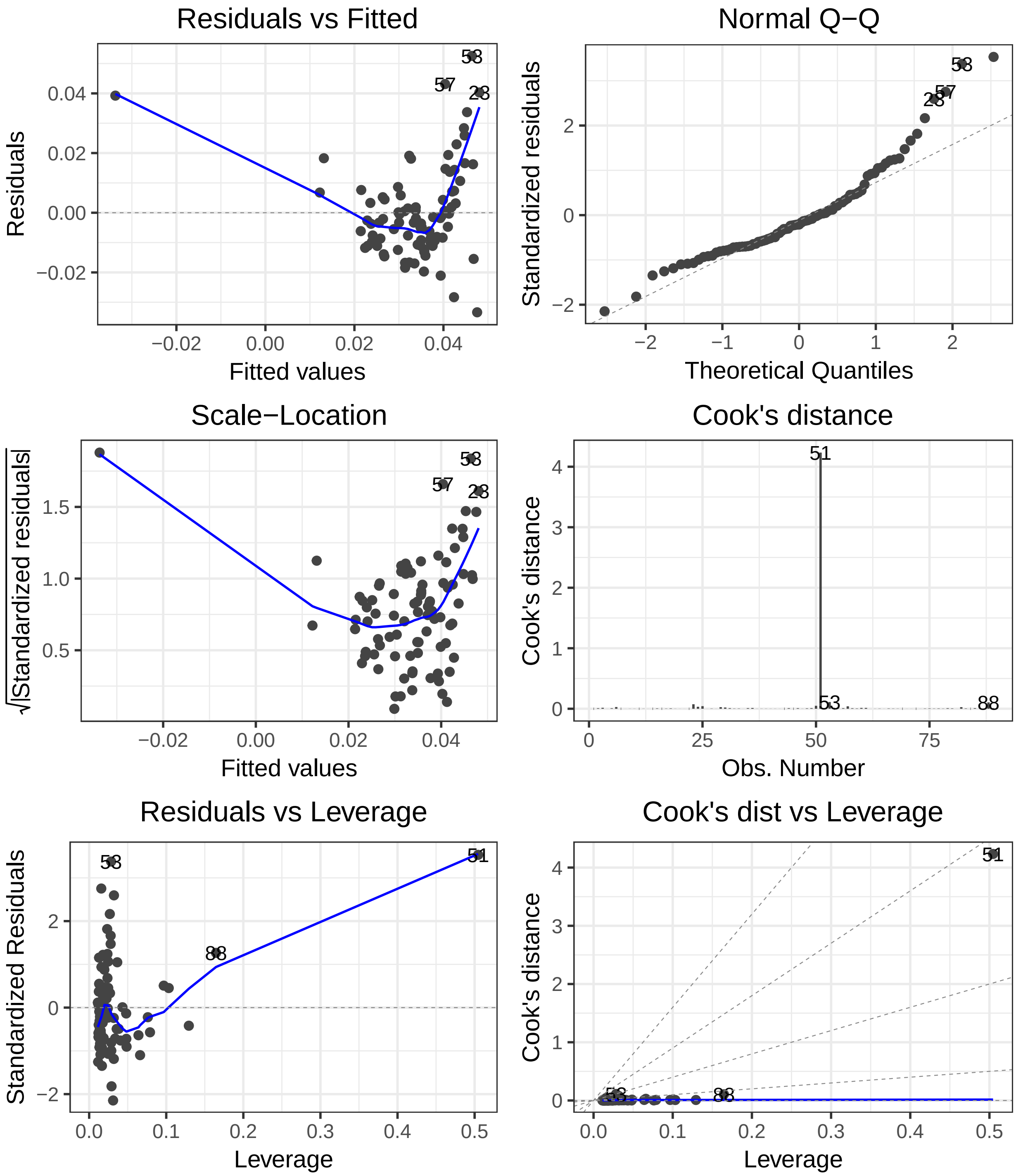

Figure 11: Model 1b OLS diagnostics

```
prbconv     -0.022903   0.010541 -2.1728 0.0325483 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 11 showed the diagnosis plot of Model 1b and the above coefficient test indicated that coefficient of `prbarr` is statistically significant. Both residual verse fitted value plot and scale-location plot showed that one data point significantly changed the shape of the curve. The leverage plot and Cook's distance plot identify that observation number 51 had extreme high influence with Cook's distance value of 4.2.

```
crime[51, c("county","crmrte","prbarr","prbconv","west","central","urban")]
```

```
   county    crmrte  prbarr prbconv west central urban
51    115 0.0055332 1.09091     1.5    1       0     0
```

```
# remove the high influence point observation 51
crime = crime[-51,]
# update both model 1a and 1b
model1a = lm (crmrte ~ density, data = crime)
model1b = lm (crmrte ~ prbarr + prbconv, data = crime)
ols_diag_plot(model1b, ncol = 2)
```

As shown above, observation number 51 is a county in west and had extreme high value of `prbconv` at 1.5. Based on residual plots in Figure 11, the value of Cook's distance and examination of data values in observation number 51, the removal of observation number 51 is justified. Thus, we recompute the Model 1b and plot the OLS diagnosis in Figure 12. Clearly, there is a lot improvement. We did not find any unusual about the data points identified in the diagnosis plot. We recompute the coefficients test for Model 1a and Model 1b as shown below. After removing observation 51, coefficients of both `prbarr` and `prbconv` in Model 1b are statistically significant and levels are dramatically different from early results that use data point 51 in computation.

```
# Retest model1a and model1b
coeftest(model1a, vcov = vcovHC)
```

```

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 0.02113261 0.00179814  11.752 < 2.2e-16 ***
density     0.00887463 0.00078768  11.267 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(model1b, vcov = vcovHC)
```

```

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  0.0781603  0.0091456  8.5462 4.476e-13 ***
prbarr      -0.0950532  0.0195226 -4.8689 5.126e-06 ***
prbconv     -0.0320996  0.0074034 -4.3358 3.970e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(model1b)
```

```
  prbarr  prbconv 
1.101629 1.101629 
```

We also calculated the VIF for Model 1b. There is no perfect multicollinearity between `prbarr` and `prbconv`.

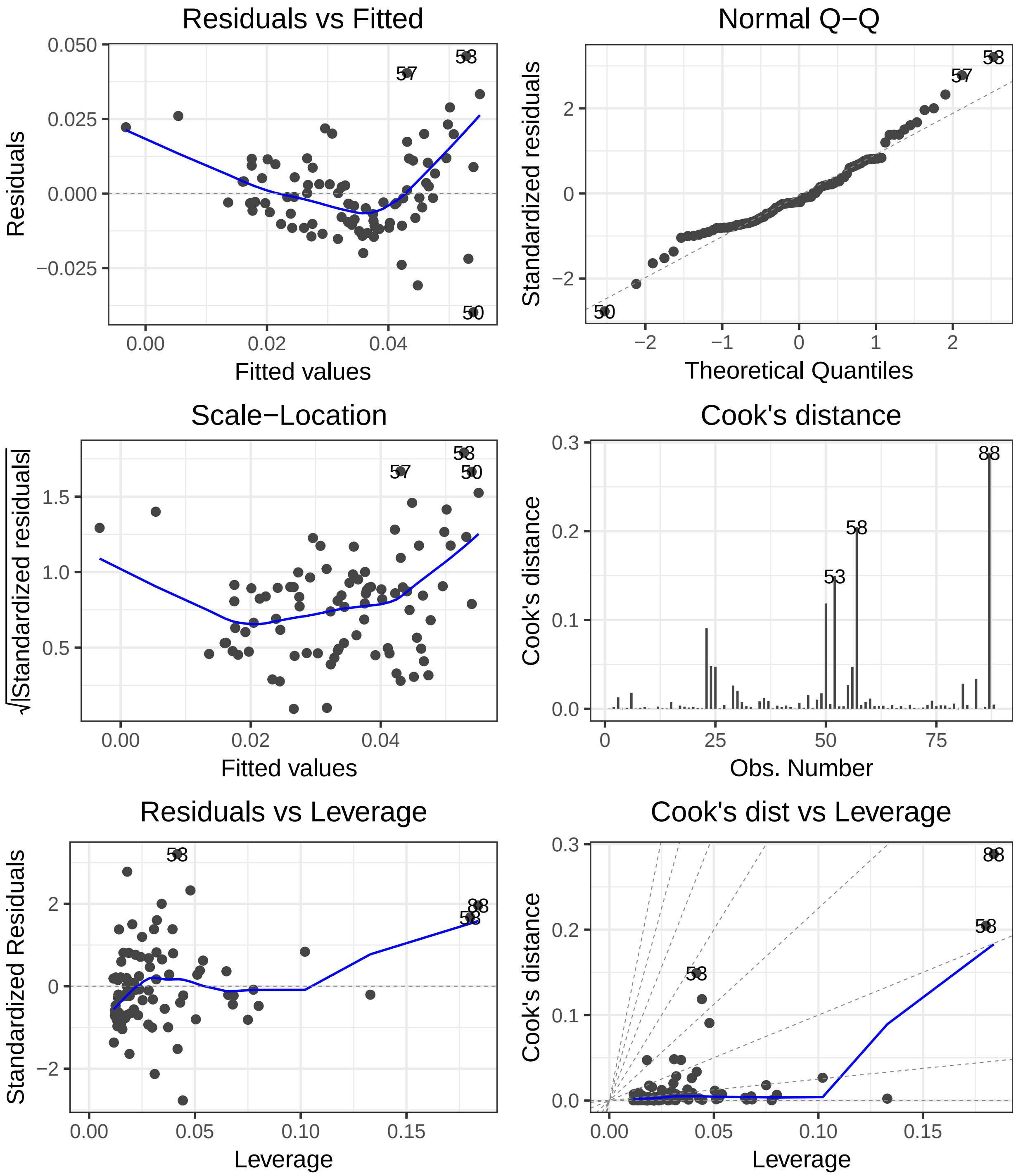

Figure 12: Model 1b with influential observation 51 removed

## 3.3 Model 2, 3, & 4

### 3.3.1 Model 2

Model 2 examined the effects of variables `density`, `prbarr` and `prbconv`. This model reflects our EDA. It focuses on our researc questions. Our Model 2 is expressed as following.

$$\text{crmrte} = \beta_0 + \beta_1 \times \text{density} + \beta_2 \times \text{prbarr} + \beta_3 \times \text{prbconv} + u$$

```
(model2 = lm (crmrte ~  density + prbarr + prbconv, data = crime))
```

```

Call:
lm(formula = crmrte ~ density + prbarr + prbconv, data = crime)

Coefficients:
(Intercept)      density       prbarr      prbconv  
   0.050084     0.006782    -0.053277    -0.020276  
```

```
ols_diag_plot(model2, ncol = 2)
```

Figure 13 showed the diagnosis plots for the second model. The residuals vs. fitted value plot suggested that residuals were reasonably close to zero horizontal line. There was a little up tick on the left side of graph, indicating a mild violation of zero conditional mean. Clearly, there are many omitted variables that can contribute to this violation.

Both Scale-location plot and residual-fitted value plot suggested that heteroskedasticity exists and heteroskedasticity robust estimators should be used. The following used heteroskedasticity robust error estimates to perform test on coefficients of our Model 2. All coefficients in our Model 2 were statistically significant.

```
coeftest(model2, vcov = vcovHC)
```

```

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  0.0500842  0.0092209  5.4316 5.337e-07 ***
density      0.0067817  0.0010632  6.3783 9.281e-09 ***
prbarr      -0.0532773  0.0165424 -3.2206  0.001820 ** 
prbconv     -0.0202758  0.0062443 -3.2471  0.001676 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The normal Q-Q plot in Figure 13 indicated that the error terms of our second model was very close to normal distribution. Nevertheless, because our sample size was 88 and more than 35, we rely solely on OLS asymptotics. No other responses were needed to address the slight deviation from normal distributions.

Cook's distance plots indicated that there was no influential points in our model.

```
stargazer(vif(model2), type = "latex", header=FALSE, label = "tab:model2xs",
          title="Variance Inflation Factors (VIF) of Model 2 Predictors")
```

Table 2: Variance Inflation Factors (VIF) of Model 2 Predictors

| density | prbarr | prbconv |
|---|---|---|
| 1.279 | 1.349 | 1.249 |

No additional error message was called out from `R` `lm` command, there was no perfect multicollinearity in the independent variables. Variance Inflation Factors (VIF) among the predictors were computed to to verify with VIF values. Table 2 showed that the value of VIFs are low and thus confirms there was no violation of CLM assumption on perfect multicollinearity.

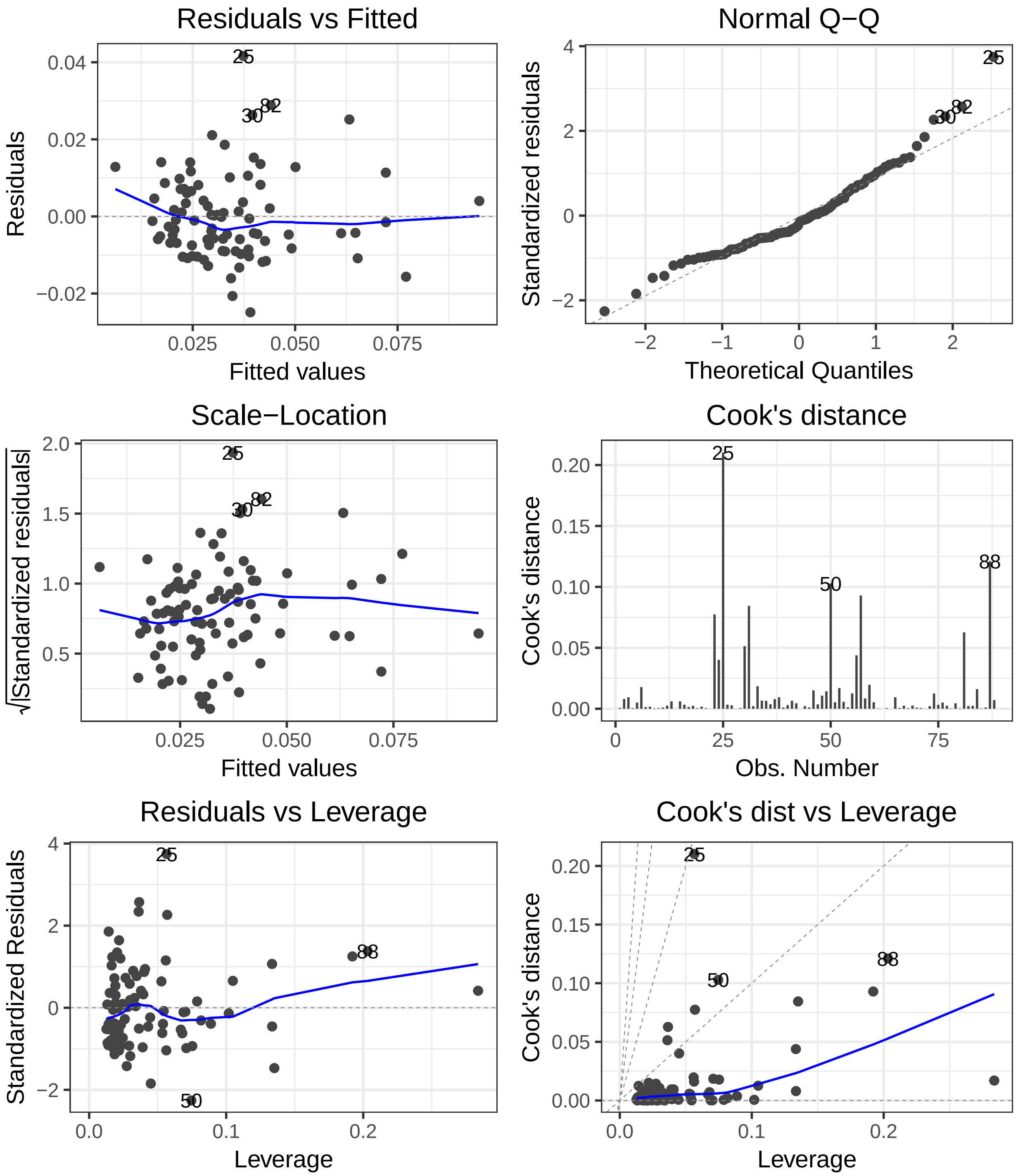

Figure 13: Model 2 OLS diagnostics

Finally, there are 100 counties in North Carolina. Our county sample size was 88, which was near population total. The samples are representative in term of geographic locations. However, since the we only have data for year 1987, the data are clustered in time. Therefore, it violates the true random sampling assumption of CLM. However, we can not address this due to limitation of dataset.

| CLM assumptions | Assessments | Responses/Comments |
|---|---|---|
| CLM.1 Linearity in parameters | Satisfied | Model is clearly linear in parameters. |
| CLM.2 Random Sampling | Violated | Data is clustered in time. Need addtional aggregation data or with a multi-year panel analysis for complete response. |
| CLM.3 Multicollinearity | Satisfied | No perfect multicollinearity among our predictors per Table 2. |
| CLM.4 Zero conditional Mean | Violated | Our estimation is likely biased. We will perform omitted variable analysis in the next section. |
| CLM.5 Homoskedasticity | Violated | Use heteroskedasticity-robust tools. |
| CLM.6 Normality of error terms | Slightly violated | Because we have a large sample, we can rely on OLS asymptotics. |

Table 3: Summary of CLM assumptions

In summary, Table 3 provides the complete assessment of CLM assumptions for Model 2.

```
(model3 = lm (crmrte ~ density + prbarr + prbconv + prbpris + avgsen + pctymle + pctmin80 + mix,
              data = crime))
```

```
Call:
lm(formula = crmrte ~ density + prbarr + prbconv + prbpris +
    avgsen + pctymle + pctmin80 + mix, data = crime)

Coefficients:
(Intercept)      density       prbarr      prbconv      prbpris
  0.0370846    0.0066989   -0.0581382   -0.0209287   -0.0056335
     avgsen      pctymle     pctmin80          mix
  0.0002175    0.0816301    0.0003697   -0.0096419
```

```
ols_diag_plot(model3, ncol = 2)
```

### 3.3.2 Model 3

Model 3 included most of the rest of variables except variables can be the outcome variable of `density`. These include the wage related variables, `polpc` and `taxpc`. Model 3 is expressed as following.

$$\text{crmrte} = \beta_0+\beta_1\times\text{density}+\beta_2\times\text{prbarr}+\beta_3\times\text{prbconv}+\beta_4\times\text{prbpris}+\beta_5\times\text{avgsen}+\beta_6\times\text{pctymle}+\beta_7\times\text{pctmin80}+\beta_8\times\text{mix}+u$$

Figure 14 showed the diagnosis plots for the third model. The OLS diagnosis plots of the Model 3 exhibited very similar characteristics as Model 2. The assessment of CLM assumptions of Model 3 is the same as Model 2.

We used heteroskedasticity robust error estimation to test Model 3 coefficients as shown below.

```
coeftest(model3, vcov = vcovHC)
```

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  3.7085e-02  1.2802e-02  2.8968  0.004875 **
density      6.6989e-03  1.1142e-03  6.0124 5.357e-08 ***
prbarr      -5.8138e-02  1.3950e-02 -4.1675 7.817e-05 ***
prbconv     -2.0929e-02  6.7720e-03 -3.0905  0.002760 **
prbpris     -5.6335e-03  1.3843e-02 -0.4070  0.685143
```

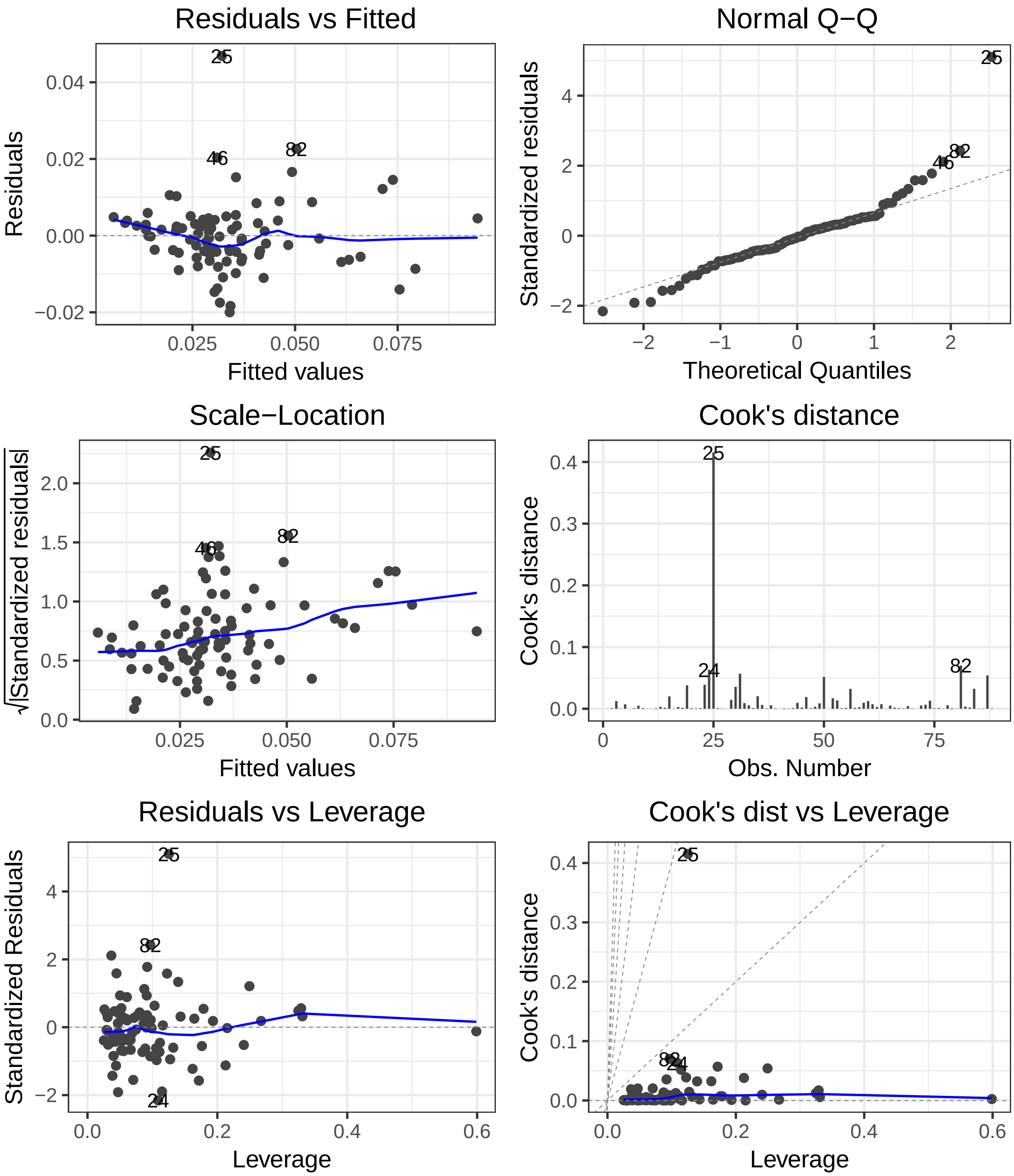

Figure 14: Model 3 OLS diagnostics

```
avgsen       2.1755e-04  5.4878e-04  0.3964  0.692860
pctymle      8.1630e-02  4.3356e-02  1.8828  0.063412 .
pctmin80     3.6967e-04  6.2699e-05  5.8960 8.749e-08 ***
mix         -9.6419e-03  1.2701e-02 -0.7591  0.450040
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
linearHypothesis(model3, c("prbpris = 0", "avgsen = 0"), vcov = vcovHC)
```

```
Linear hypothesis test

Hypothesis:
prbpris = 0
avgsen = 0

Model 1: restricted model
Model 2: crmrte ~ density + prbarr + prbconv + prbpris + avgsen + pctymle +
    pctmin80 + mix

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1     81
2     79  2 0.1103 0.8957
```

```r
##### Joint Hypothesis Testing
# We want to test whether the four additional performance indicators are jointly
# significant.  That is, whether the coefficients for prbpris, avgsen, pctymle
# and mix are all zero.
waldtest(model2, model3, vcov = vcovHC)
```

```
Wald test

Model 1: crmrte ~ density + prbarr + prbconv
Model 2: crmrte ~ density + prbarr + prbconv + prbpris + avgsen + pctymle +
    pctmin80 + mix
  Res.Df Df      F    Pr(>F)
1     84
2     79  5 10.121 1.627e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(model3)
```

```
 density   prbarr  prbconv  prbpris   avgsen  pctymle pctmin80      mix
1.307755 1.854988 1.389217 1.081495 1.086871 1.115013 1.095398 1.547346
```

The low values in VIFs of Model 3 confirm that there is no need to be concerned about violation of no perfect multicollinearity CLM assumption.

Only `density`, `prbarr`, `prbconv` and `pctmin80` are statistically significant. We fail to reject the null hypothesis that coefficients of `prbpris`, `avgsen`, `pctymle` and `mix` are zero.

To test whether the difference in fit is significant between Model 2 and Model 3, we perform the Wald test as shown above. We reject the null hypothesis in the joint hypothesis testing. Therefore, there is support that variables `prbpris`, `avgsen`, `pctymle`, `pctmin80` and `mix` are jointly not zero. In particular, `pctymle` and `pctmin80` show the low $p$-value. This leads to our model 4 as described below.

### 3.3.3 Model 4 - Examine the effect of young male percentage and minority variables

Our background knowledge suggest that young male minorities are correlated with the crime rate (Viels, Shaw, & Whiteman, 2010). This phenomenon escalates in the presence of urbanization. When residents with different social economic statuses and backgrounds gather in the same areas without cultural affinity or a common ground, criminal incidents increase, especially when policy makers fail to provide relevant policies that efficiently facilitate development. In addition, an increasing crime rate does not only disturb the peace in an area. Young male minority-related crimes often lead to more serious social problems. Public policies that address these issues are therefore necessary in any urbanization planning efforts.

To this end, Model 4 focused on the effect of demographic variables, young male percentage and minority percentage on the crime rate.

$$\text{crmrte} = \beta_0 + \beta_1 \times \text{density} + \beta_2 \times \text{prbarr} + \beta_3 \times \text{prbconv} + \beta_4 \times \text{pctymle} + \beta_5 \times \text{pctmin80} + u$$

```
(model4 = lm (crmrte ~ density + prbarr + prbconv + pctymle + pctmin80, data = crime))
```

```

Call:
lm(formula = crmrte ~ density + prbarr + prbconv + pctymle +
    pctmin80, data = crime)

Coefficients:
(Intercept)      density       prbarr      prbconv      pctymle
  0.0357666    0.0067340   -0.0610760   -0.0200137    0.0873037
   pctmin80
  0.0003573
```

```
coeftest(model4, vcov = vcovHC)
```

```

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  3.5767e-02  1.0942e-02  3.2686 0.0015803 **
density      6.7340e-03  1.0308e-03  6.5328 5.084e-09 ***
prbarr      -6.1076e-02  1.3805e-02 -4.4243 2.951e-05 ***
prbconv     -2.0014e-02  5.6502e-03 -3.5422 0.0006579 ***
pctymle      8.7304e-02  3.8131e-02  2.2896 0.0246152 *
pctmin80     3.5732e-04  6.6513e-05  5.3722 7.124e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The diagnosis plots for Model 4 indicated similar characteristics as previous models. HHeteroskedasticity robust error estimation for Model 4 coefficients suggested no assumptions were violated. Similar to test from previous models (Models 1, 2, and 3), the low values in VIFs of Model 4 confirm that there is no need to be concerned about violation of no perfect multicollinearity CLM assumption.

### 3.3.4 Wald Statistics

Wald statistics suggested adding new variables had impact to the model

```
waldtest(model1b, model2, vcov = vcovHC)
```

```
Wald test

Model 1: crmrte ~ prbarr + prbconv
Model 2: crmrte ~ density + prbarr + prbconv
  Res.Df Df      F    Pr(>F)
1     85
2     84  1 40.683 9.281e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(model1b, model4, vcov = vcovHC)
```

```
Wald test

Model 1: crmrte ~ prbarr + prbconv
Model 2: crmrte ~ density + prbarr + prbconv + pctymle + pctmin80
  Res.Df Df      F    Pr(>F)
1     85
2     82  3 18.575 2.758e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 4 Regression Table

Table 4 showed the comparison of model specifications. We tested the statistical significance to compare the model specification accordingly. The coefficients of main variables `density`, `prbarr` and `prbconv` are close across model 2, 3 and 4. Model 3 and model 4 also have similar coefficients estimates for variables `pctymle` and `pctmin80`. It is important to notice that `prbpris` and `avgsen` are not statistically significant in all models.

```
se.model1a = sqrt(diag(vcovHC(model1a)))
se.model1b = sqrt(diag(vcovHC(model1b)))
se.model2 = sqrt(diag(vcovHC(model2)))
se.model3 = sqrt(diag(vcovHC(model3)))
se.model4 = sqrt(diag(vcovHC(model4)))
stargazer(model1a, model1b, model2, model3, model4, type = "latex", label = "tab:OLSTab",
          se = list(se.model1a, se.model1b, se.model2, se.model3, se.model4),
          column.labels = c("Model 1a", "Model 1b", "Model 2", "Model 3", "Model 4"),
          model.numbers = FALSE,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Linear Models Specification Table", header=FALSE,
          add.lines=list(c("AIC", round(AIC(model1a),1), round(AIC(model1b),1),
                           round(AIC(model2),1), round(AIC(model3),1),
                           round(AIC(model4),1))),
          omit.stat = "f")
```

Clearly, the more predictor variables, the higher $R^2$ values. The $R^2$ improvements from the model 2, 3 and 4 to the model 1b and 1a are noticable. We also report Akaike Information Criterion (AIC) in Table 4. AIC metric penalizes models with more parameters. Using Burnham & Anderson recommended Rule of Thumb for comparing AIC's, both model 3 and 4 has more than 10 lower AIC than model 2. Thus, AIC model fit metric suggests that there is strong evidence that both model 3 and 4 fit better than model 2 with our dataset. There is some evidence that model 3 is better than model 4 since AIC difference is about 5.

## Practical significance and policy studies

Model 2 indicated that the average crime rate was 0.05. For every unit increase in the probability of arrest, the crime rate decreased by -0.058. For every unit increase in the probability of conviction, the crime rate decreased by 0.02.

Moreover, the data suggested that for urban counties, the median crime rate was 0.066 and median probability of arrests was 0.211. In contrast, for rural counties, the median crime rate was 0.0289 and median probability of arrests was 0.285.

Using our model 2, we studied the following cases:

(1) What will be median crime rate of urban counties if its median `prbarr` is the same as rural counties while keeping all others the same?
(2) What will be median crime rate of urban counties if its median `prbconv` is the same as rural counties while keeping all others the same?
(3) What will be median crime rate of urban counties if both of its median `prbarr` and `prbconv` are the same as rural counties while keeping all others the same?

Table 4: Linear Models Specification Table

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | | | crmrte | | |
| | Model 1a | Model 1b | Model 2 | Model 3 | Model 4 |
| density | 0.009$^{***}$ | | 0.007$^{***}$ | 0.007$^{***}$ | 0.007$^{***}$ |
| | (0.001) | | (0.001) | (0.001) | (0.001) |
| prbarr | | $-0.095^{***}$ | $-0.053^{**}$ | $-0.058^{***}$ | $-0.061^{***}$ |
| | | (0.020) | (0.017) | (0.014) | (0.014) |
| prbconv | | $-0.032^{***}$ | $-0.020^{**}$ | $-0.021^{**}$ | $-0.020^{***}$ |
| | | (0.007) | (0.006) | (0.007) | (0.006) |
| prbpris | | | | $-0.006$ | |
| | | | | (0.014) | |
| avgsen | | | | 0.0002 | |
| | | | | (0.001) | |
| pctymle | | | | 0.082 | 0.087$^{*}$ |
| | | | | (0.043) | (0.038) |
| pctmin80 | | | | 0.0004$^{***}$ | 0.0004$^{***}$ |
| | | | | (0.0001) | (0.0001) |
| mix | | | | $-0.010$ | |
| | | | | (0.013) | |
| Constant | 0.021$^{***}$ | 0.078$^{***}$ | 0.050$^{***}$ | 0.037$^{**}$ | 0.036$^{**}$ |
| | (0.002) | (0.009) | (0.009) | (0.013) | (0.011) |
| AIC | -511.6 | -488.2 | -531.1 | -553.5 | -558.4 |
| Observations | 88 | 88 | 88 | 88 | 88 |
| R$^2$ | 0.528 | 0.397 | 0.638 | 0.750 | 0.747 |
| Adjusted R$^2$ | 0.522 | 0.383 | 0.626 | 0.724 | 0.731 |
| Residual Std. Error | 0.013 (df = 86) | 0.015 (df = 85) | 0.011 (df = 84) | 0.010 (df = 79) | 0.010 (df = 82) |
| *Note:* | | | | | $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001 |

(4) What will be median crime rate of urban counties if its median `prbarr` is 75% and its median `prbconv` is 50% while keeping all others the same?

The following `R` code used model 2 to compute the expected crime rate for the above proposed studies.

```
# beta0=intercept, beta1 = density slope, beta2 = prbarr slope, beta3 = prbconv slope
beta0 = model2$coefficients[1]
beta1 = model2$coefficients[2]
beta2 = model2$coefficients[3]
beta3 = model2$coefficients[4]
median_df = crime %>% group_by(county_loc) %>%
    summarise(mdn_crmrte = median(crmrte), mdn_density = median(density),
              mdn_prbarr = median(prbarr), mdn_prbconv = median(prbconv)) %>% data.frame()
# Probability of arrests if U has same median crime rate as ER
urban_mdn_crmrte = median_df[median_df$county_loc=="U", "mdn_crmrte"]
urban_mdn_density = median_df[median_df$county_loc=="U", "mdn_density"]
urban_mdn_prbarr = median_df[median_df$county_loc=="U", "mdn_prbarr"]
urban_mdn_prbconv = median_df[median_df$county_loc=="U", "mdn_prbconv"]
er_mdn_crmrte = median(median_df[median_df$county_loc!="U", "mdn_crmrte"])
er_mdn_density = median(median_df[median_df$county_loc!="U", "mdn_density"])
er_mdn_prbarr = median(median_df[median_df$county_loc!="U", "mdn_prbarr"])
er_mdn_prbconv = median(median_df[median_df$county_loc!="U", "mdn_prbconv"])
(urban_case1 = beta0 + beta1*urban_mdn_density + beta2*er_mdn_prbarr +
        beta3*urban_mdn_prbconv)[["(Intercept)"]]
```

```
[1] 0.06481716
```

```
(urban_case2 = beta0 + beta1*urban_mdn_density + beta2*urban_mdn_prbarr +
        beta3*er_mdn_prbconv)[["(Intercept)"]]
```

```
[1] 0.06606429
```

```
(urban_case3 = beta0 + beta1*urban_mdn_density + beta2*er_mdn_prbarr +
        beta3*er_mdn_prbconv)[["(Intercept)"]]
```

```
[1] 0.06232527
```

```
(urban_case4 = beta0 + beta1*urban_mdn_density + beta2*0.75 +
        beta3*0.5)[["(Intercept)"]]
```

```
[1] 0.03660579
```

Comparing with current median crime rate of urban counties, the median crime rate was reduced by 1.8%, -0.039%, 5.6%, and 45%, for cases 1, 2, 3 and 4 respectively. Result in case (1) suggested that matching `prbarr` alone will reduce the crime rate while case (2) suggested that increasing `prbconv` alone cannot. However, in case (3), we notice that if urban counties match both `prbarr` and `prbconv` with rural counties, there was significant reduction in crime rate. Finally, as we increase both `prbarr` and `prbconv` to ideal levels in case (4), the crime rate was reduced by nearly half. Clearly, urban counties need to have policies for more aggressive law enforcement and efficient justice system to reduce crime rates.

Similarly, our Model 4 suggested that the average crime rate was 0.036. Model 4 results suggested that for every unit increase in the probability of arrest, the crime rate decreased by -0.061. For every unit increase in the probability of conviction, the crime rate decreased by 0.02. On the contrary, For every percent increase in the young male in the community, the crime rate increased by 0.09. For every percent increased in the minority in the community, the crime rate increased by 0.04.

# 5 Omitted Variables

True model with omitted variable $x_k$.

$$\text{crmrte} = \beta_0 + \beta_1 \times \text{density} + \beta_2 \times \text{prbarr} + \beta_3 \times \text{prbconv} + \beta_k \times x_k + u$$

$$x_k = \delta_0 + \delta_1 \times \text{density} + \delta_2 \times \text{prbarr} + \delta_3 \times \text{prbconv} + v$$

$$\text{crmrte} = (\beta_0 + \beta_k\delta_0) + (\beta_1 + \beta_k\delta_1) \times \text{density} + (\beta_2 + \beta_k\delta_2) \times \text{prbarr} + (\beta_3 + \beta_k\delta_3) \times \text{prbconv} + (u + \beta_k v)$$

We have $\beta_0 > 0$, $\beta_1 > 0$, $\beta_2 < 0$ and $\beta_3 < 0$.

Examining Omitted Variables.

(1) Poverty level `povtlv`: the variable measured the percentage of population who lived at poverty level. We expect high poverty level corresponds to high crime rate. Thus, $\beta_k$ was greater than 0 for variable `povtlv`. We assumed `povtlv` was negatively correlated with people density so $\delta_1$ was less than 0. This leads negative omitted variable bias (OMVB) for $\beta_1$. Since $\beta_1 > 0$, this indicated the bias was towards 0. High value `povtlv` leads more offenses. If law enforcement capacity remained same, the number of arrests remain the same. More offenses would reduce the probability of arrests. Thus, $\delta_2$ was less than 0. OMVB was negative for for $\beta_2$. Because $\beta_2 < 0$, this bias was away from zero. Since probability of conviction was the ratio between convicted and arrested, $\delta_3$ was 0. Thus, there is no impact on $\beta_3$.

(2) Unemployment rate `unemprt`: this variable measures unemployment rate of the population. Usually high unemployment rate leads high crime rate. Thus, $\beta_k$ is greater than 0 for variable `unemprt`. We assume `unemprt` is negatively correlated with people density so $\delta_1$ is less than 0. This leads negative OMVB for $\beta_1$. Since $\beta_1 > 0$, this means the bias is towards 0. High value `unemprt` leads more offenses. If law enforcement capacity remains the same, more offenses will reduce the probability of arrests. Thus, $\delta_2$ is less than 0. OMVB is negative for for $\beta_2$. Because $\beta_2 < 0$, this bias is away from zero. Assume the ratio between convicted and arrested remains the same, $\delta_3$ is 0. Thus, there is no impact on $\beta_3$.

(3) Job availability `jobv`: this variable measures the job availability in a region. A high value represents more jobs are available. The job availability has opposite effect from unemployment rate. High job availability means more opportunities to find a job and that leads to low crime rate. Thus, $\beta_k$ is less than 0 for variable `jobv`. People density is usually positively associated with job availability. Therefore, $\delta_1$ is greater than 0. This leads negative OMVB for $\beta_1$. Since $\beta_1 > 0$, this means the bias is towards 0. High value `jobv` leads less offenses. If law enforcement capacity remains the same, less number of offenses increases the ratio between number of arrests and number of offenses. Therefore, it increases the probability of arrests. We have $\delta_2$ is greater than 0. This makes positive OMVB for $\beta_2$. Because $\beta_2 < 0$, this bias is towards to 0. Assume the ratio between convicted and arrested remains the same, $\delta_3$ is 0. Thus, there is no impact on $\beta_3$.

(4) Citizen's attitude towards crime `crimeatti`: this variable measures the citizen's attitude towards crime. The higher value of `crimeatti` represent higher attitude against crime. Therefore, $\beta_k$ is less than 0 by the definition of the variable `crimeatti`. People density is independent from citizen's attitude towards crime. Thus, $\delta_1$ is 0 and this omitted variable has no impact on $\beta_1$. High value of `crimeatti` leads low number of offenses. Assume that law enforcement capacity remains the same. The probability of arrests will increase with increases of `crimeatti`. So, $\delta_2$ is greater than 0. This makes OMVB is negative for $\beta_2$. Because $\beta_2 < 0$, this bias is away from 0. High value of `crimeatti` can make it easier to deliver conviction by juries. Thus, $\delta_3$ is greater than 0. So, OMVB is negative for $\beta_3$. Since $\beta_2 < 0$, this bias is also away from 0.

(5) Crime report practice of citizenry `crimerpt`: this variable measures the ratio between reported crimes to all witnessed crimes by citizens. The higher value of `crimerpt` will deter crime and reduce crime rate. Therefore, $\beta_k$ is less than 0. People density is independent from `crimerpt`. Thus, $\delta_1$ is 0 and this omitted variable has no impact on $\beta_1$. Due to limited law enforcement capacity, high value of `crimerpt` increases number of offenses and reduces the probability of arrests. $\delta_2$ is less than 0. The OMVB is positive for $\beta_2$. Since $\beta_2 < 0$, this bias is towards 0. Also, `crimerpt` will not impact the ratio between convicted and arrested. This omitted variable has no impact on $\beta_3$.

(6) Family conditions with respect to divorce and family cohesiveness `famcond`: The high value of this variable represents more cohesive family condition. We know from our background knowledge that the more cohesive family the less likely generates criminals. Thus, $\beta_k$ is less than 0. Higher people density is expected to lower value in `famcond` and $\delta_1$ is less than 0. This leads OMVB is positive for $\beta_1$. Because $\beta_1 > 0$, this bias is away from 0. High value of `famcond` reduces number of offenses. That increases probability of arrests since law enforcement capacity remains the same. Therefore, $\delta_2$ is greater than 0. This leads the OMVB is negative for $\beta_2$. Since $\beta_2 < 0$, this bias is away from 0. `famcond` will not impact the the ratio between convicted and arrested. This omitted variable has no impact on $\beta_3$.

(7) Jail Overcrowding `jailcap`; The recent impacts of overcrowded jails has lead to shorter sentences severed. It has also lead to a lower level of sentences given, replaced by probation and community service. A higher value

of this variable represents the county's jail capacity being close to full. This can lead to an increase in crime rate for certain crimes as offenders know they will get off easily. Therefore, $\beta_k$ is greater than 0. People density can cause an increase to `jailcap`. However, `jailcap` does not have an impact of the density so OMVB is 0. An increase in this value has an increase on the amount of crimes committed leading to more arrests, yielding $\delta_2$ is greater than 0. Community service is considered a measure of probation and not conviction, yielding $\delta_3$ being less than 0.

(8) Percentage of time not served, `pctnotsrvd`; This variable represents the amount of time not served in jail for a given sentence. For various reasons, someone may spend a reduced amount of time in jail. A higher value represents less time served than sentenced. This can lead to an increase in crime rate for certain crimes as offenders know they will get off easier than advertised. Therefore, $\beta_k$ is greater than 0. People density can cause an increase to `pctnotsrvd`. However, `pctnotsrvd` does not have an impact of the density so OMVB is 0. An increase in this value has an increase on the amount of crimes committed leading to more arrests, yielding $\delta_2$ greater than 0. More crimes committed will lead to more convictions yielding $\delta_3$ being greater than 0.

Table 5 summarizes the impact of aforementioned omitted variables.

| $x_k$ | $\beta_k$ | $\delta_1$ | $\delta_2$ | $\delta_3$ | Bias directions |
|---|---|---|---|---|---|
| `povtlv` | $> 0$ | $< 0$ | $< 0$ | $= 0$ | Negative OMVB for $\beta_1$ and towards 0. Negative OMVB for $\beta_2$ and away from 0. No impact on $\beta_3$. |
| `unemprt` | $> 0$ | $< 0$ | $< 0$ | $= 0$ | Negative OMVB for $\beta_1$ and towards 0. Negative OMVB for $\beta_2$ and away from 0. No impact on $\beta_3$. |
| `jobv` | $< 0$ | $> 0$ | $> 0$ | $= 0$ | Negative OMVB for $\beta_1$ and towards 0. Positive OMVB for $\beta_2$ and towards 0. No impact on $\beta_3$. |
| `crimeatti` | $< 0$ | $= 0$ | $> 0$ | $> 0$ | No impact on $\beta_1$ . Negative OMVB for $\beta_2$ and away from 0. Negative OMVB for $\beta_3$ and away from 0. |
| `crimerpt` | $< 0$ | $= 0$ | $< 0$ | $= 0$ | No impact on $\beta_1$ . Positive OMVB for $\beta_2$ and towards 0. No impact on $\beta_3$. |
| `famcond` | $< 0$ | $< 0$ | $> 0$ | $= 0$ | Positive OMVB for $\beta_1$ and away from 0. Negative OMVB for $\beta_2$ and away from 0. No impact on $\beta_3$. |
| `pctnotsrvd` | $> 0$ | $= 0$ | $> 0$ | $> 0$ | No impact on $\beta_1$ . Positive OMVB for $\beta_2$ and $\beta_3$. |
| `jailcap` | $> 0$ | $= 0$ | $> 0$ | $< 0$ | No impact on $\beta_1$ . Positive OMVB for $\beta_2$ and Negative OMVB for $\beta_3$ |

Table 5: Summary of Omitted Variable Bias

## 6 Conclusions

In this report, we examined the dataset of crime statistics for counties in North Carolina. The data confirmed our hypothesis of that population density was a proxy of urbanization and was a key factor that affected crime rate. Even though enhancing the law enforcement in the developing areas such as certainty of punishment and arrest rates (ie., criminals expect to get caught and convicted) reduced the crime rate, it is only a bandage to current problem. Policies that address the source of crime such as educating and providing resources to foster younger generation especially minority and male are the long term solution reducing the crime.

Contradicting to our initial intuitions, our results suggest that there is no support for the arguement that with more severe punishment such longer prison sentences alone could reduce crime rate. Moreover, our results are likely biased due to the cross-section nature of the data set.

# Appendix

## All additional R code in the file "Lab3Unitls.R"

```
load_packages = function (pkgs = c("ggplot2", "dplyr", "gridExtra", "ggthemes",
                                   "car", "lmtest", "sandwich", "stargazer",
                                   "ggfortify","ggcorrplot")) {
    out <- lapply(pkgs, library, character.only = T)
    return (out)
}

init_doc = function() {
    load_packages()
}

hist_plot = function(df, xvar = "crmrte", brks = NULL,
                     filcol = "transparent", lcol = "dodgerblue4",
                     xlab = "Crimes committed per person", ylab = "Frequency",
                     tle = "Histogram") {
    ret = ggplot(df, aes_string(x=xvar)) +
        geom_histogram(breaks = brks, fill = filcol, col = lcol) +
        labs(x = xlab, y = ylab, title = tle) +
        theme_bw() + theme(plot.title = element_text(hjust = 0.5))
}

scatter_hist_plot = function (df, xvar = "prbarr", yvar = "crmrte",
                              ptcolvar = "county_loc", ptsize = 1.5,
                              xlab = "Probability of arrest",
                              ylab = "Crimes committed per person",
                              htle = "Histogram of prbarr",
                              histylab = "Frequency",
                              nbins = 20,
                              lmcol = "darkolivegreen4",
                              smcol = "firebrick4", fillcol = "transparent",
                              histlcol = "dodgerblue4") {
  ret = list()
  ret[[1]] = ggplot(df, aes_string(x = xvar, y = yvar, colour = ptcolvar)) +
      geom_point(size = ptsize) +
      scale_color_discrete(name="Location") +
      geom_smooth(method="loess", col=smcol, fill="gainsboro") +
      geom_smooth(method="lm", se=F, col = lmcol) +
      labs(x = xlab, y = ylab) + theme_bw() +
      theme(legend.position="top", legend.direction = "horizontal")
  ret[[2]] = ggplot(df, aes_string(x = xvar)) +
      geom_histogram(bins = nbins, fill = fillcol, col = histlcol) +
      labs(x = xlab, y = histylab, title = htle) + theme_bw() +
      theme(plot.title = element_text(hjust = 0.5))
  args.list = c(ret, list(nrow = 1))
  do.call(grid.arrange, args.list)
}

ols_diag_plot = function(model, plts=1:6, ncol = 3, labsz = 3) {
    ret = autoplot(model, which = plts, ncol = ncol, label.size = labsz) +
        theme_bw() + theme(plot.title = element_text(hjust = 0.5))
    return(ret)
}

distribution_plts = function(df, xvar = "crmrte", filvar = "county_loc",
                             brks = NULL,
                             xlab = "Crimes committed per person",
                             ylab = "Frequency",
                             types = c("'W'", "'C'", "'U'", "'ER'"),
                             tles = c("West Counties", "Central Counties",
                                      "Urban Counties","East Rural Counties",
                                      "All Counties"),
                             lcol = "dodgerblue4") {

  plts = list()
  for (i in 1:length(types)) {
    plts[[i]] = hist_plot(df %>% filter_(paste(filvar, "==", types[i])),
                          xvar = xvar, brks = brks, lcol = lcol, tle = tles[i],
                          xlab = xlab, ylab = ylab)
  }
  all_index = length(types) + 1
  plts[[all_index]] = hist_plot(df, xvar = xvar, brks = brks,  lcol = lcol,
                                xlab = xlab, ylab = ylab, tle = tles[all_index])
  boxplt_idx = all_index + 1
  plts[[boxplt_idx]] = ggplot(df, aes_string(x = filvar, y = xvar)) +
    geom_boxplot(color = lcol) + theme_bw() + labs(x = "County Locations",
                                                   y = xlab)
  args.list = c(plts, list(ncol = 2))
  do.call(grid.arrange, args.list)
}

scatter_plts = function(df, xvars = c("prbarr"), yvar = "crmrte",
                        xlabs = c("Probability of arrest"),
                        ylab = "Crimes committed per person",
                        lmcol = "darkolivegreen4", smcol = "firebrick4",
                        ptcolvar = "county_loc", ptsize = 1.5, ncls = 2){
  plts = list()
  pltidx = 1
  for (i in xvars) {
    plts[[pltidx]] = ggplot(df, aes_string(x = i, y = yvar, colour = ptcolvar)) +
        geom_point(size = ptsize) +
        scale_color_discrete(name="Location") +
        geom_smooth(method="loess", col=smcol, fill="gainsboro") +
        geom_smooth(method="lm", se=F, col = lmcol) +
        labs(x = xlabs[pltidx], y = ylab) + theme_bw() +
        theme(legend.position="top", legend.direction = "horizontal")
    pltidx = pltidx + 1
  }
  args.list = c(plts, list(ncol = ncls))
  do.call(grid.arrange, args.list)
}

multiple_hist_plts = function(df, histvars = c("prbpris"),
                              xlabs = c("Probability of prison sentence"),
                              ncls = 2) {
  plts = list()
  pltidx = 1
  for (i in histvars) {
    plts[[pltidx]] = hist_plot(df, xvar = i, xlab = xlabs[pltidx])
    pltidx = pltidx + 1
  }
  args.list = c(plts, list(ncol = ncls))
  do.call(grid.arrange, args.list)
```

```
}
```

## R software environment

The R software version and support packages information used for generating this document is given below.

```r
sessionInfo()
```

```
R version 3.4.3 (2017-11-30)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: OS X El Capitan 10.11.6

Matrix products: default
BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] bindrcpp_0.2     ggcorrplot_0.1.1 ggfortify_0.4.3  stargazer_5.2.1
 [5] sandwich_2.4-0   lmtest_0.9-35    zoo_1.8-1        car_2.1-6
 [9] ggthemes_3.4.0   gridExtra_2.3    dplyr_0.7.4      ggplot2_2.2.1

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.14       nloptr_1.0.4       pillar_1.1.0
 [4] compiler_3.4.3     plyr_1.8.4         bindr_0.1
 [7] tools_3.4.3        digest_0.6.13      lme4_1.1-15
[10] evaluate_0.10.1    tibble_1.4.2       gtable_0.2.0
[13] nlme_3.1-131       lattice_0.20-35    mgcv_1.8-22
[16] pkgconfig_2.0.1    rlang_0.1.6        Matrix_1.2-12
[19] parallel_3.4.3     yaml_2.1.16        SparseM_1.77
[22] stringr_1.2.0      knitr_1.17         MatrixModels_0.4-1
[25] rprojroot_1.2      grid_3.4.3         nnet_7.3-12
[28] glue_1.2.0         R6_2.2.2           rmarkdown_1.8
[31] minqa_1.2.4        purrr_0.2.4        tidyr_0.8.0
[34] magrittr_1.5       codetools_0.2-15   backports_1.1.2
[37] scales_0.5.0       htmltools_0.3.6    MASS_7.3-47
[40] splines_3.4.3      assertthat_0.2.0   pbkrtest_0.4-7
[43] colorspace_1.3-2   quantreg_5.34      stringi_1.1.6
[46] lazyeval_0.2.1     munsell_0.4.3
```

## Additional Exploratory Data Analysis results

```r
stargazer(correlation.matrix, type = "latex", header=FALSE, label = "tab:cor",
          title="Correlation Matrix", font.size = "scriptsize", digits = 2,
          column.sep.width = "0pt", float.env = "sidewaystable")
```

```r
scatter_plts(crime, xvars = c("wtrd", "wfir", "wser", "wmfg"),
             xlabs = c("Weekly wage, whlesle, retail trade", "Weekly wage, fin, ins, real est",
                       "Weekly wage, service industry", "Weekly wage, manufacturing"))
```

```r
scatter_plts(crime, xvars = c("wfed", "wsta","wloc"),
             xlabs = c("Weekly wage, fed employees", "Weekly wage, state employees",
```

Table 6: Correlation Matrix

| | crmrte | prbarr | prbconv | prbpris | avgsen | polpc | density | taxpc | pctmin80 | wcon | wtuc | wtrd | wfir | wser | wmfg | wfed | wsta | wloc | mix | pctymle |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| crmrte | 1 | -0.41 | -0.37 | 0.05 | -0.001 | 0.16 | 0.73 | 0.46 | 0.22 | 0.38 | 0.23 | 0.42 | 0.33 | 0.35 | 0.34 | 0.48 | 0.20 | 0.36 | -0.15 | 0.29 |
| prbarr | -0.41 | 1 | -0.02 | 0.05 | 0.17 | 0.42 | -0.31 | -0.14 | 0.07 | -0.27 | -0.08 | -0.11 | -0.18 | -0.26 | -0.16 | -0.22 | -0.16 | -0.03 | 0.41 | -0.19 |
| prbconv | -0.37 | -0.02 | 1 | -0.01 | 0.27 | 0.22 | -0.22 | -0.16 | -0.06 | -0.06 | 0.05 | -0.08 | 0.10 | -0.03 | 0.08 | -0.01 | -0.16 | 0.08 | -0.29 | -0.15 |
| prbpris | 0.05 | 0.05 | -0.01 | 1 | -0.09 | 0.05 | 0.08 | -0.09 | 0.10 | -0.05 | 0.13 | 0.15 | 0.04 | -0.02 | 0.01 | 0.09 | -0.03 | 0.08 | 0.12 | -0.08 |
| avgsen | -0.001 | 0.17 | 0.27 | -0.09 | 1 | 0.49 | 0.06 | 0.09 | -0.13 | -0.05 | 0.22 | 0.09 | 0.16 | 0.04 | 0.09 | 0.14 | 0.13 | 0.14 | -0.16 | 0.06 |
| polpc | 0.16 | 0.42 | 0.22 | 0.05 | 0.49 | 1 | 0.16 | 0.28 | -0.16 | -0.03 | 0.17 | 0.12 | 0.19 | 0.16 | 0.27 | 0.16 | 0.05 | 0.39 | 0.02 | 0.05 |
| density | 0.73 | -0.31 | -0.22 | 0.08 | 0.06 | 0.16 | 1 | 0.32 | -0.06 | 0.45 | 0.33 | 0.59 | 0.54 | 0.55 | 0.44 | 0.59 | 0.22 | 0.46 | -0.15 | 0.11 |
| taxpc | 0.46 | -0.14 | -0.16 | -0.09 | 0.09 | 0.28 | 0.32 | 1 | -0.03 | 0.27 | 0.17 | 0.19 | 0.13 | 0.26 | 0.26 | 0.06 | -0.04 | 0.22 | -0.04 | -0.09 |
| pctmin80 | 0.22 | 0.07 | -0.06 | 0.10 | -0.13 | -0.16 | -0.06 | -0.03 | 1 | -0.08 | -0.17 | -0.03 | -0.05 | -0.20 | -0.09 | 0.06 | 0.09 | -0.10 | 0.23 | -0.004 |
| wcon | 0.38 | -0.27 | -0.06 | -0.05 | -0.05 | -0.03 | 0.45 | 0.27 | -0.08 | 1 | 0.40 | 0.56 | 0.48 | 0.55 | 0.34 | 0.50 | -0.02 | 0.52 | -0.21 | -0.03 |
| wtuc | 0.23 | -0.08 | 0.05 | 0.13 | 0.22 | 0.17 | 0.33 | 0.17 | -0.17 | 0.40 | 1 | 0.34 | 0.32 | 0.42 | 0.46 | 0.39 | -0.15 | 0.33 | -0.27 | -0.11 |
| wtrd | 0.42 | -0.11 | -0.08 | 0.15 | 0.09 | 0.12 | 0.59 | 0.19 | -0.03 | 0.56 | 0.34 | 1 | 0.66 | 0.54 | 0.36 | 0.64 | 0.01 | 0.58 | -0.14 | -0.12 |
| wfir | 0.33 | -0.18 | 0.10 | 0.04 | 0.16 | 0.19 | 0.54 | 0.13 | -0.05 | 0.48 | 0.32 | 0.66 | 1 | 0.59 | 0.49 | 0.62 | 0.24 | 0.55 | -0.23 | 0.004 |
| wser | 0.35 | -0.26 | -0.03 | -0.02 | 0.04 | 0.16 | 0.55 | 0.26 | -0.20 | 0.55 | 0.42 | 0.54 | 0.59 | 1 | 0.54 | 0.61 | 0.07 | 0.58 | -0.34 | 0.09 |
| wmfg | 0.34 | -0.16 | 0.08 | 0.01 | 0.09 | 0.27 | 0.44 | 0.26 | -0.09 | 0.34 | 0.46 | 0.36 | 0.49 | 0.54 | 1 | 0.51 | 0.06 | 0.45 | -0.36 | 0.02 |
| wfed | 0.48 | -0.22 | -0.01 | 0.09 | 0.14 | 0.16 | 0.59 | 0.06 | 0.06 | 0.50 | 0.39 | 0.64 | 0.62 | 0.61 | 0.51 | 1 | 0.19 | 0.52 | -0.33 | -0.07 |
| wsta | 0.20 | -0.16 | -0.16 | -0.03 | 0.13 | 0.05 | 0.22 | -0.04 | 0.09 | -0.02 | -0.15 | 0.01 | 0.24 | 0.07 | 0.06 | 0.19 | 1 | 0.17 | -0.07 | 0.22 |
| wloc | 0.36 | -0.03 | 0.08 | 0.08 | 0.14 | 0.39 | 0.46 | 0.22 | -0.10 | 0.52 | 0.33 | 0.58 | 0.55 | 0.58 | 0.45 | 0.52 | 0.17 | 1 | -0.26 | -0.005 |
| mix | -0.15 | 0.41 | -0.29 | 0.12 | -0.16 | 0.02 | -0.15 | -0.04 | 0.23 | -0.21 | -0.27 | -0.14 | -0.23 | -0.34 | -0.36 | -0.33 | -0.07 | -0.26 | 1 | -0.10 |
| pctymle | 0.29 | -0.19 | -0.15 | -0.08 | 0.06 | 0.05 | 0.11 | -0.09 | -0.004 | -0.03 | -0.11 | -0.12 | 0.004 | 0.09 | 0.02 | -0.07 | 0.22 | -0.005 | -0.10 | 1 |

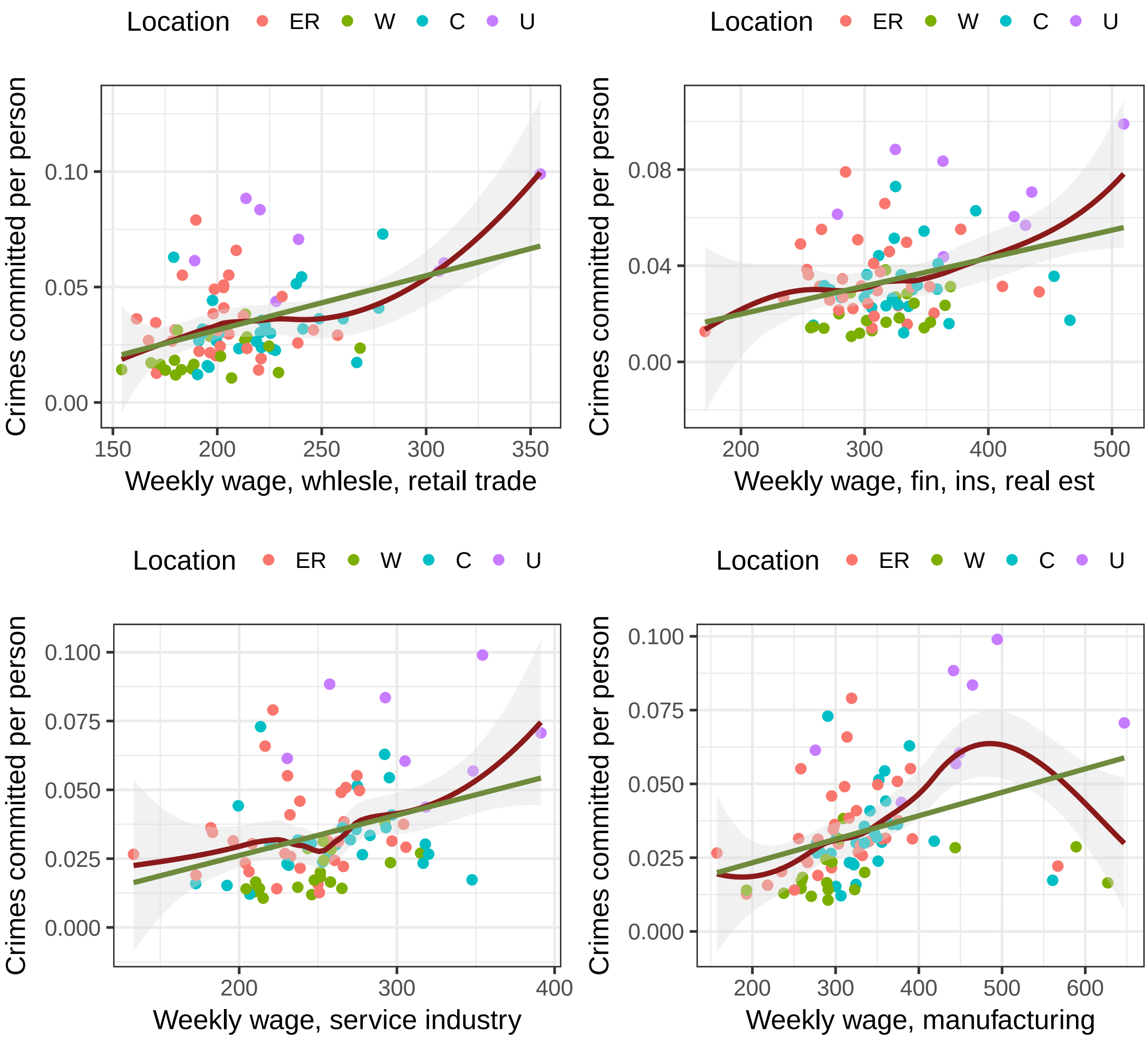

Figure 15: Scatter plots for variables $wtrd$, $wfir$, $wser$ and $wmfg$

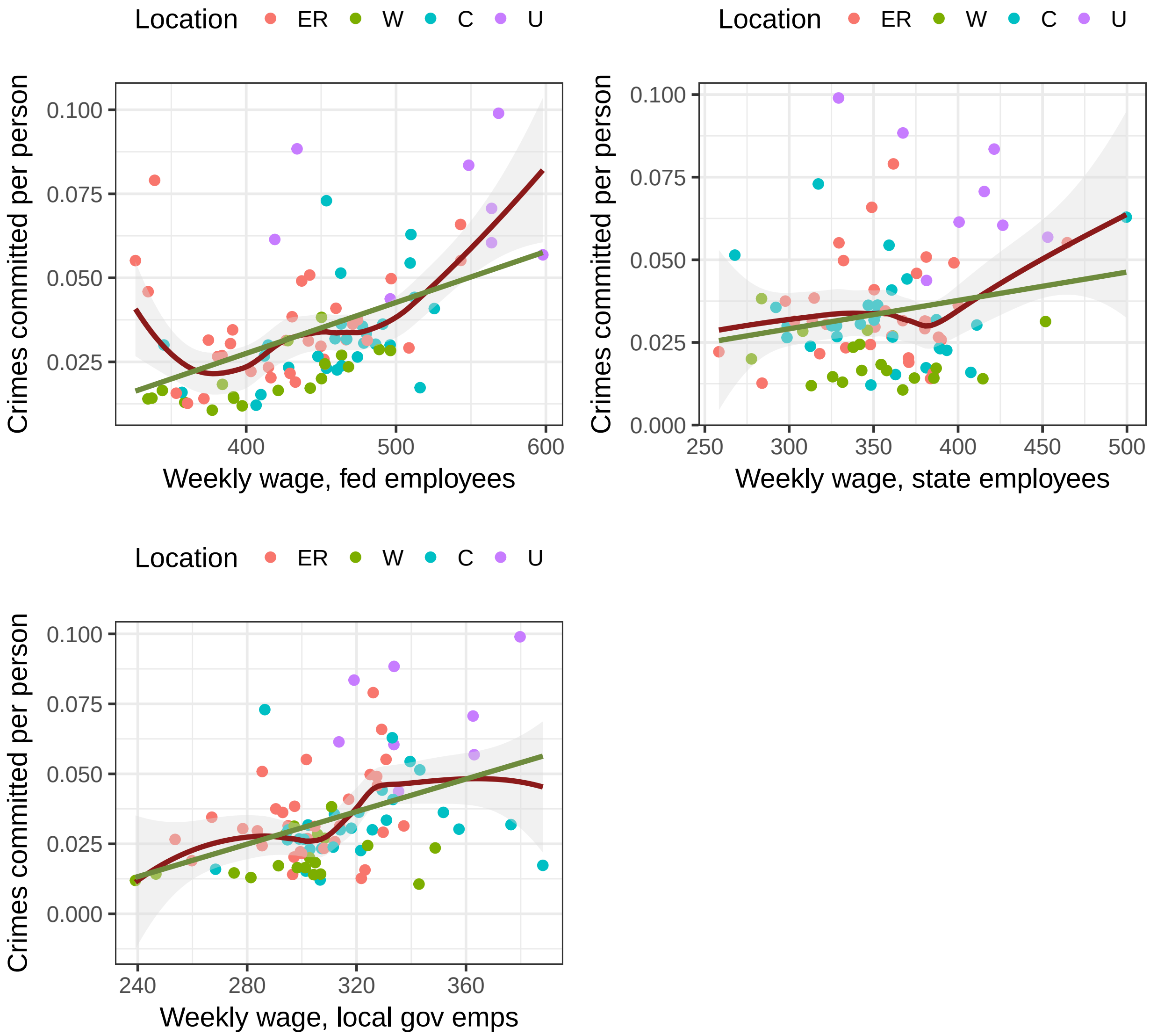

Figure 16: Scatter plots for variables $wfed$, $wsta$ and $wloc$

```
                       "Weekly wage, local gov emps"))
```

```
multiple_hist_plts(crime,
                   histvars = c("taxpc", "wcon", "wtuc", "wtrd", "wfir", "wser"),
                   xlabs = c("tax revenue per capita",
                             "Weekly wage, construction",
                             "Weekly wage, trns, util, commun",
                             "Weekly wage, whlesle, retail trade",
                             "Weekly wage, fin, ins, real est",
                             "Weekly wage, service industry"))
```

```
multiple_hist_plts(crime,
                   histvars = c("wmfg", "wfed", "wsta", "wloc"),
                   xlabs = c("Weekly wage, manufacturing",
                             "Weekly wage, fed employees",
                             "Weekly wage, state employees",
                             "Weekly wage, local gov emps"))
```

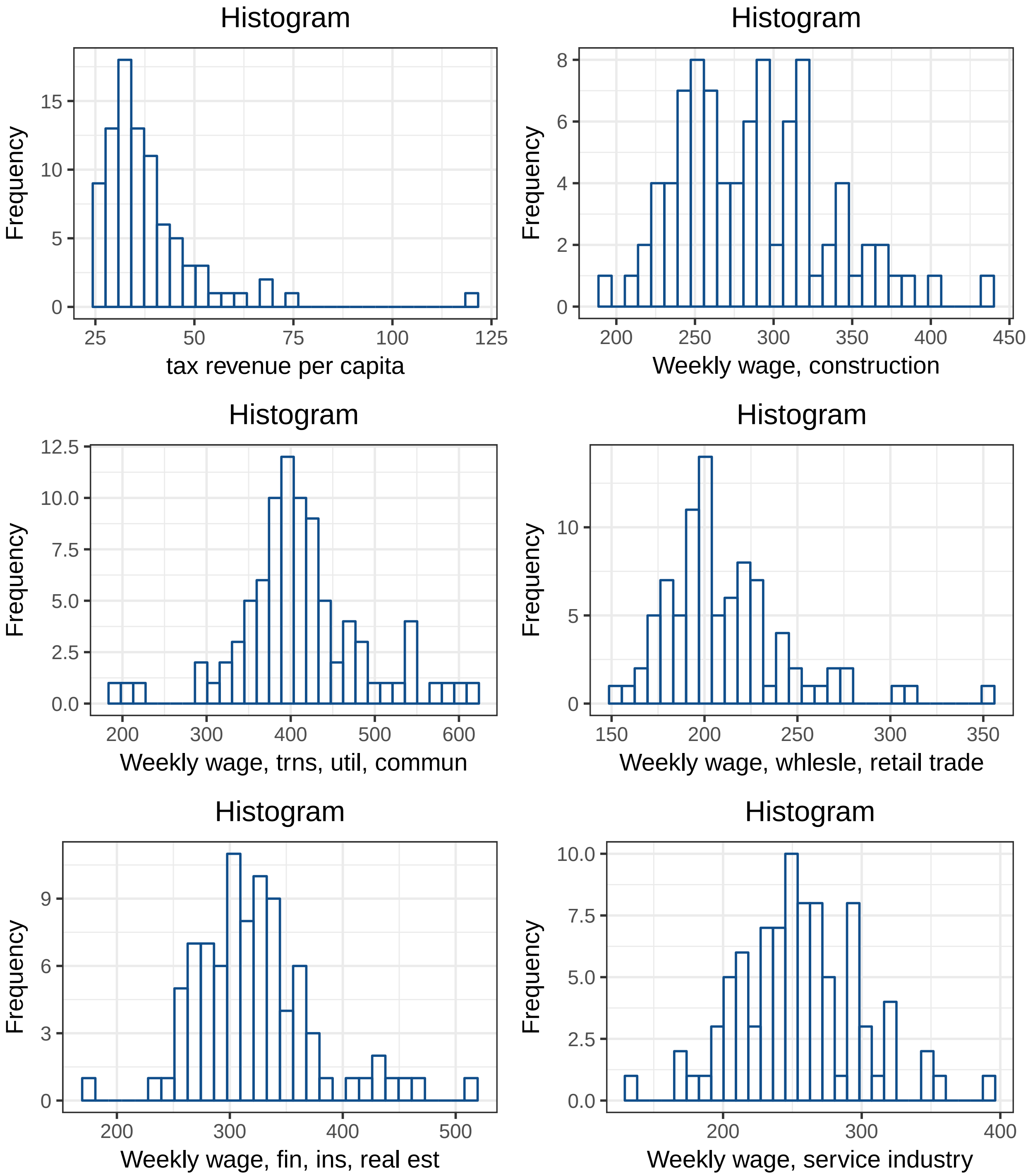

Figure 17: Histograms of variables $taxpc$, $wcon$, $wtuc$, $wtrd$, $wfir$ and $wser$

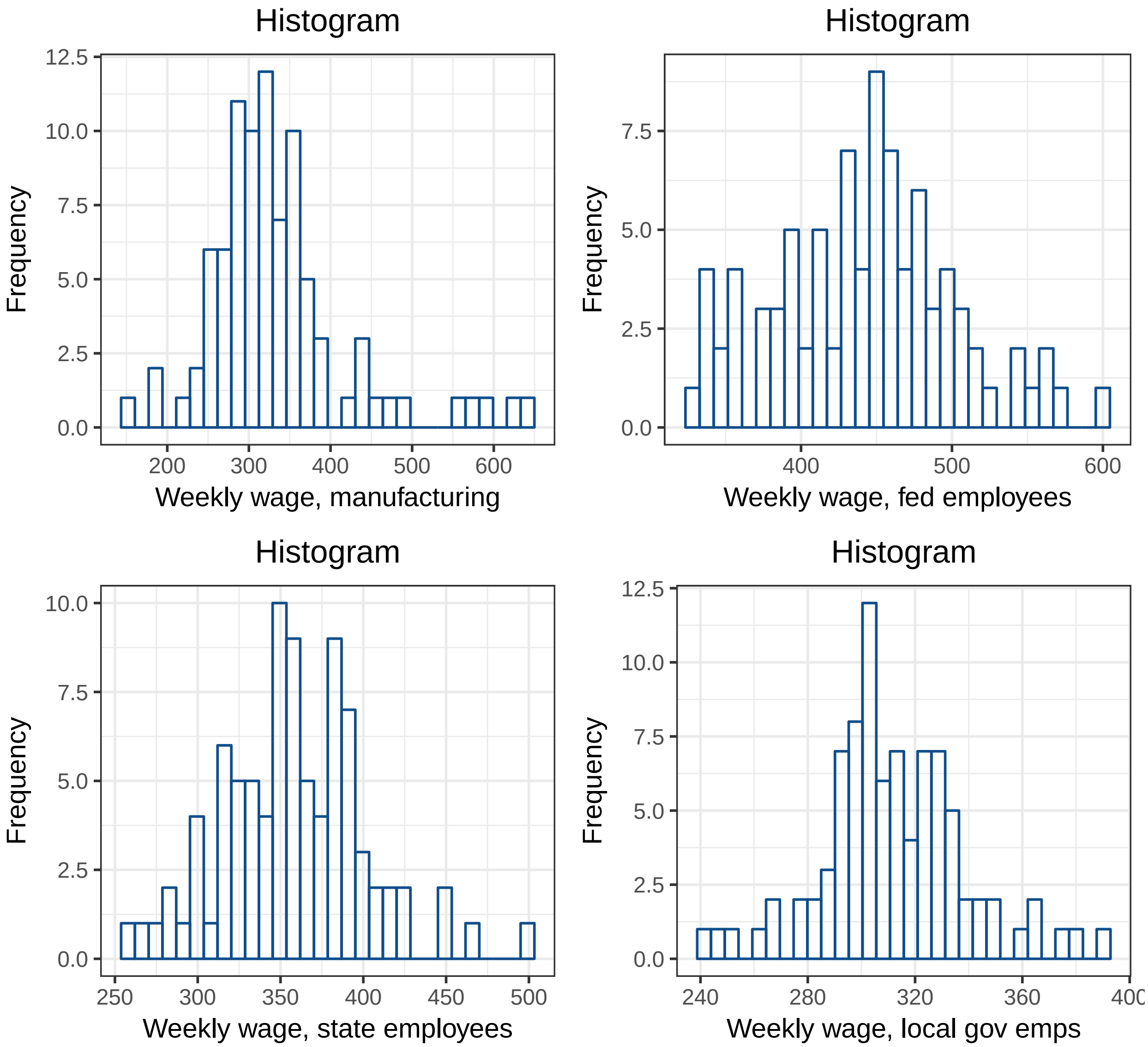

Figure 18: Histograms of variables $wmfg$, $wfed$, $wsta$, and $wloc$

### Additional Models Developed

The following included the model building process that led to Model 4

To examine the effects of demographic and law enforcement variables on the crime rates we built 3 models in addition to the base model. Four (4) models are presented as followings.

- Base Model estimated the effect of density on crime rate.

$$\text{crmrte} = \beta_0 + \beta_1 \times \text{density} + u$$

- Model 4-1 estimated the effect of law enforcement on crime rate using probabiliy of arrest rate and conviction rate variables.

$$\text{crmrte} = \beta_0 + \beta_1 \times \text{density} + \beta_2 \times \text{prbarr} + \beta_3 \times \text{prbconv} + u$$

- Model 4-2 estimated the effect of demographic profile on crime rate using the young male percentage and minority percentage in the area.

$$\text{crmrte} = \beta_0 + \beta_1 \times \text{density} + \beta_2 \times \text{pctmin80} + \beta_3 \times \text{pctymle} + u$$

- Model 4-3 estimated the effcet of density, law enforcement, and demographic variables on the crime rate.

$$\text{crmrte} = \beta_0 + \beta_1 \times \text{density} + \beta_2 \times \text{prbarr} + \beta_3 \times \text{prbconv} + \beta_4 \times \text{pctmin80} + \beta_5 \times \text{pctymle} + u$$

Model 4-2

```
model2 = lm (crmrte ~  density + pctymle + pctmin80, data = crime)
summary(model2)
```

```

Call:
lm(formula = crmrte ~ density + pctymle + pctmin80, data = crime)

Residuals:
      Min        1Q    Median        3Q       Max
-0.016285 -0.005944 -0.001916  0.003227  0.060154

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.012e-05  5.051e-03   0.010 0.992106
density     8.826e-03  8.134e-04  10.851  < 2e-16 ***
pctymle     1.631e-01  5.254e-02   3.103 0.002606 **
pctmin80    2.898e-04  7.518e-05   3.855 0.000226 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01152 on 84 degrees of freedom
Multiple R-squared:  0.6338,    Adjusted R-squared:  0.6208
F-statistic: 48.47 on 3 and 84 DF,  p-value: < 2.2e-16
```

Model 4-3

```
model3 = lm (crmrte ~ density + prbarr + prbconv + pctymle + pctmin80, data = crime)
summary(model3)
```

```

Call:
lm(formula = crmrte ~ density + prbarr + prbconv + pctymle +
    pctmin80, data = crime)

Residuals:
      Min        1Q    Median        3Q       Max
```

```
-0.019204 -0.004716 -0.000901  0.003730  0.048716

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.577e-02  7.295e-03   4.903 4.71e-06 ***
density      6.734e-03  7.686e-04   8.761 2.15e-13 ***
prbarr      -6.108e-02  1.166e-02  -5.238 1.23e-06 ***
prbconv     -2.001e-02  3.999e-03  -5.004 3.15e-06 ***
pctymle      8.730e-02  4.599e-02   1.898   0.0612 .
pctmin80     3.573e-04  6.482e-05   5.512 4.00e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009696 on 82 degrees of freedom
Multiple R-squared:  0.7466,    Adjusted R-squared:  0.7312
F-statistic: 48.32 on 5 and 82 DF,  p-value: < 2.2e-16
```