Predictive Policing for Crime in Chicago

Izabela Litwin

Contents

Introduction

The goal of this study is to understand how demographic and geographic factors relate to narcotic-related crime in Chicago. In particular, we are interested in knowing whether narcotic-related crimes depend on demographic and geographic factors and whether cannabis and noncannabis-related crimes are correlated. In this study, I present stable, interpretable, theoretically valid regression models to predict narcotic crime in Chicago that address these research questions.

Exploratory Data Analysis

Part 1: Univariate EDA

The data analyzed in this report comes from narcotic-related crime reports across 2012 per Census block group. There are 15 variables and 2102 observations. Each observation represents a block group. Within each block group, we have variables from the 2010 US Census and the 2011 American Community Survey that provide additional information. The counts of narcotic-related crimes are split in two: cannabis and non-cannabis related cases.

All variables besides zone and ward are continuous. In my analysis, I decided to treat zone as a categorical variable and ward as continuous because of the large number of categories within it that are based on geographical locations in Chicago (they are neighboring locations/similar to each other). The predictor variables used in this report are TotalCrime, CrimeNC and CrimeC. Other variables are used as response variables.

Table 1: Summary of Continuous Variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
PopulationTotal	2,102	1,255.076	546.371	56	892.2	1,494.8	11,309
income.male	2,102	36,425.680	20,838.340	2,499	22,375.8	45,403.8	238,512
income.female	2,102	28,666.500	14,888.440	2,499	18,481	35,943	235,167
age.male	2,102	33.413	9.088	4.100	27.700	38.900	74.300
age.female	2,102	36.166	9.303	13.800	29.600	41.500	82.600
Ward	2,102	25.088	14.564	1	13	38	50
latitude	2,102	41.858	0.095	41.648	41.776	41.940	42.022
longitude	2,102	−87.682	0.062	−87.939	−87.721	−87.642	−87.528
CrimeC	2,102	19.593	7.145	4	15	23	61
CrimeNC	2,102	16.553	9.077	1	10	21	104
pctWhite	2,102	0.422	0.326	0.000	0.026	0.711	0.985
pctBlack	2,102	0.378	0.417	0.000	0.024	0.937	0.998
pctAsian	2,102	0.048	0.087	0.000	0.001	0.057	0.939
CrimeTotal	2,102	36.146	13.356	13	27	43	144

Based on the histograms, we can say that the distribution of the population is right skewed with a mean of 1255.1 and standard deviation of 546.4. The distributions of income (male), income (female), age (male) and age (female) are also right skewed with respective means, medians and ranges specified in the table below. The distribution of 50 wards is relatively uniform. The distributions of cannabis-related crimes, non-cannabis related and total crimes are right-skewed with respective means of 19.6, 16.6, and 36.15. The distribution of pctBlack is bimodal close to 0 and 1. The distributions of pctWhite and pctAsian are right-skewed with modes close to 0.

Additionally, population, income.male, income.female, pctAsian have very skewed distributions because of the outliers. There are a few outliers in the data set. For example, the group 682 in ward 35 with population of 1514 and total crime of 144. Observation 2091 has the largest population of 11309. Observation 2034 in ward 25 has the highest pctAsian of 0.9386.

Part 2: Multivariate EDA

Given the correlation matrix (in the Appendix), we can see that CrimeTotal, CrimeC and CrimeNC are strongly correlated which is understandable. When it comes to other variables, some of them are highly correlated. In particular, pctBlack and pctWhite have a correlation of -0.91, pctBlack and latitude have a correlation of -0.58 and income (female) and income (male) have a correlation of 0.52.

Based on the plots graphing the three responses (TotalCrime, CrimeNC and CrimeC) against all other predictors with trendlines, we can say that there are few interesting trends in the data set. There seem to be positive relationships between income (male) and CrimeC, income (female) and CrimeC. Also, there seems to be a slight negative relationship between CrimeNC and age for females and males. There appears to be no relationship between the geographical variables and total crime. There is a slight positive relationship between CrimeC and longitude. There seems to be a negative relationship between total crime and pctAsian. There seems to be no relationship between zones and crime.

Plots analyzing the relationships between all available variables and the three responses (TotalCrime, CrimeNC and CrimeC) are included in the Appendix.

Part 3: Geographical Patterns

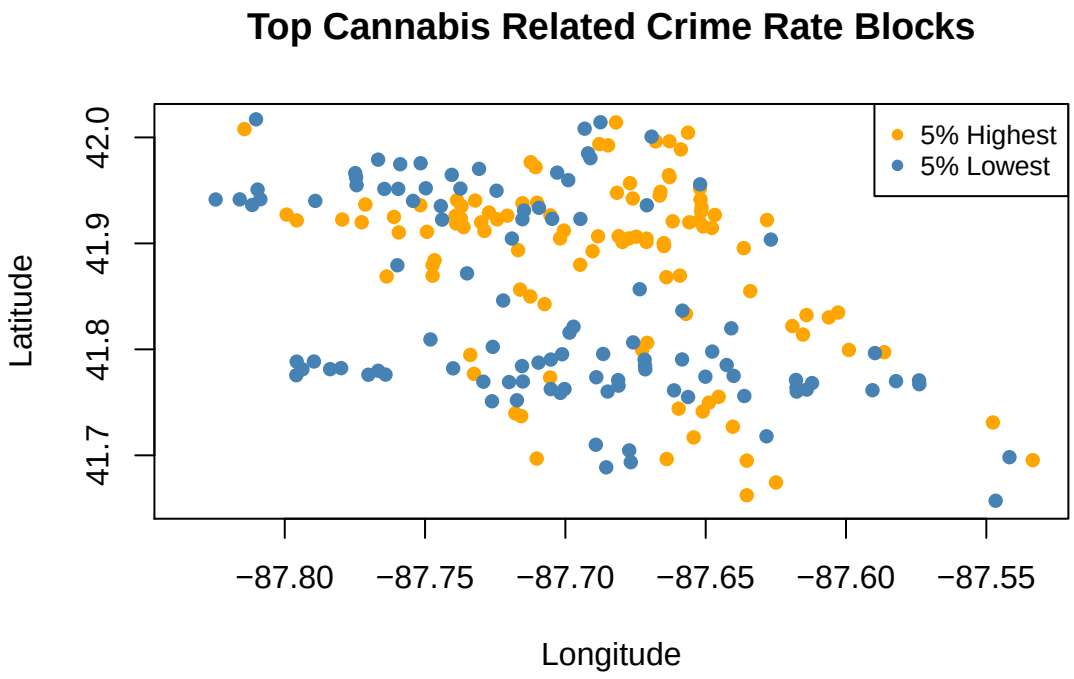

Figure 1: Geographical Patterns for Blocks with High and Low Cannabis Related Crime Rates

To see if there is a geograhical pattern in the data for high and low crime rates, upon identifying the top 5% highest, and the bottom 5% lowest crime rate blocks, I plotted them using longitude and latitude variables.

For cannabis related crime, I see a similar pattern as for non cannabis related crime rate blocks. Most of high crime blocks are concentrated in the higher latitude (between 41.85 and 42). Most of low crime blocks are concentrated in [41.7,41.8] and [41.9,42].

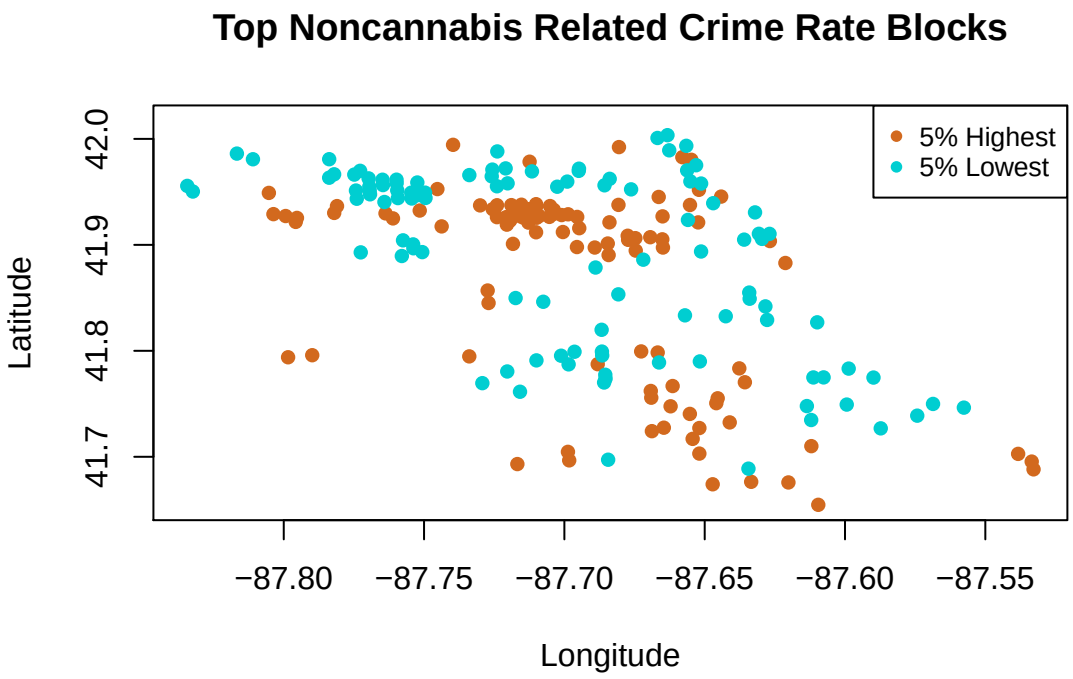

Figure 2: Geographical Patterns for Blocks with High and Low Noncannabis Related Crime Rates

For non cannabis related crime rate blocks, it appears that most of high crime blocks are concentrated in the higher latitude (above 41.85). Most of low crime blocks are concentrated above 41.9.

For both cannabis related and non cannabis related crimes, the distribution of longitude is relatively bell shaped, meaning that most of the crimes are in [-87.75,-87.65] range. The difference between cannabis related and non cannabis related crimes in terms of lattitude is that for low crime blocks for noncanabis related crimes, there is a concentration of blocks above 41.95 (mode) and for canabis related crimes there is a concentration below 41.8 (mode). There seem to be a division between those two around the latitude of 41.85.

Initial Modeling & Diagnostics

Part 4: Multiple Linear Regression Models

I constructed two candidate multiple linear regression models for the total crime count (cannabis and non-cannabis crimes added together) per block.

$$
\begin{aligned}
CrimeTotal_i = \\
&\beta_0 + \beta_{log(PopulationTotal)} log(PopulationTotal)_i + \\
&\beta_{log(income.male)} log(income.male)_i + \\
&\beta_{log(income.female)} log(income.female)_i + \\
&\beta_{age.male} age.male_i + \\
&\beta_{age.female} age.female_i + \\
&\beta_{Ward} Ward_i + \\
&\beta_{latitude} latitude_i + \\
&\beta_{longitude} longitude_i + \\
&\beta_{pctWhite} pctWhite_i + \\
&\beta_{pctBlack} pctBlack_i + \\
&\beta_{pctAsian} pctAsian_i + \\
&\beta_{zone} zone_i + \epsilon_i
\end{aligned}
$$

In the first model, I decided to include all variables except cannabis and non-cannabis crimes because those would simply produce a perfect, meaningless fit. I decided to log transform the following variables: Population, income.male, income.female because the distributions for those variables were very skewed. pctAsian could be log transformed too but I decided not to do it because some data points have pctAsian=0, which cannot be log transformed. All of variables are treated as continuous variables besides zone which is treated as a factor. I decided to treat Ward as a factor because there are 50 different divisions of the city of Chicago and if I were to treat is as a factor, I would produce too many coefficients in the regression. I decided to include all variables (excluding CrimeNC and CrimeC) because I wanted to see their significance.

$$
\begin{aligned}
CrimeTotal_i = \\
&\beta_0 + \beta_{latitude} latitude_i + \\
&\beta_{pctAsian} pctAsian_i + \\
&\beta_{zone} zone_i + \epsilon_i
\end{aligned}
$$

For the second model, I decided to remove some of the explanatory variables in order to see if others that are correlated with those become significant. In particular, to select the variables that should be used in the second model, I used a stepwise model selection using Schwartz' Bayesian Information Criterion. I chose the model with the lowest BIC. The graph is included in the Appendix.

Part 5: Model Selection

In order to identify a model with the best prediction performance, I performed 5-fold cross-validation to choose between Models 1 and 2 as predictors. For each fold of cross-validation, I created the test set by randomly selecting n=421 (420 for some folds) for each k = 1,..,5.

Model 2 is better than Model 1 because its average cross-validation error and std. deviation are smaller than those for Model 1. For Model 1, the estimated cross-validated prediction error is 172.28 with the standard deviation of 9.13 (measure of uncertainty).

By looking at the diagnostic plots and the distributions of residuals (Figures XXXXX), we can say that those models perform very similarly as the distributions of residuals look nearly identical. Since Model 2 has lower average cross-validation error and std. deviation, Model 2 appears to be a better choice than Model 1.

Table 2: Regression Models

	Dependent variable:	
	CrimeTotal	
	(1)	(2)
log(PopulationTotal)	−1.067	
	(0.793)	
log(income.male)	0.232	
	(0.686)	
log(income.female)	1.334**	
	(0.663)	
age.male	−0.028	
	(0.041)	
age.female	−0.030	
	(0.040)	
Ward	0.040	
	(0.030)	
latitude	−40.035***	−36.573***
	(7.869)	(7.083)
longitude	7.583	
	(5.998)	
pctWhite	−5.661*	
	(2.923)	
pctBlack	−5.963***	
	(2.247)	
pctAsian	−15.286***	−10.673***
	(4.246)	(3.678)
as.factor(zone)1	8.792***	9.081***
	(1.317)	(1.284)
Constant	2,369.907***	1,562.419***
	(579.890)	(295.785)
Observations	2,102	2,102
R^2	0.042	0.035
Adjusted R^2	0.037	0.034
Residual Std. Error	13.108 (df = 2089)	13.130 (df = 2098)
F Statistic	7.702*** (df = 12; 2089)	25.372*** (df = 3; 2098)
Note:		*p<0.1; **p<0.05; ***p<0.01

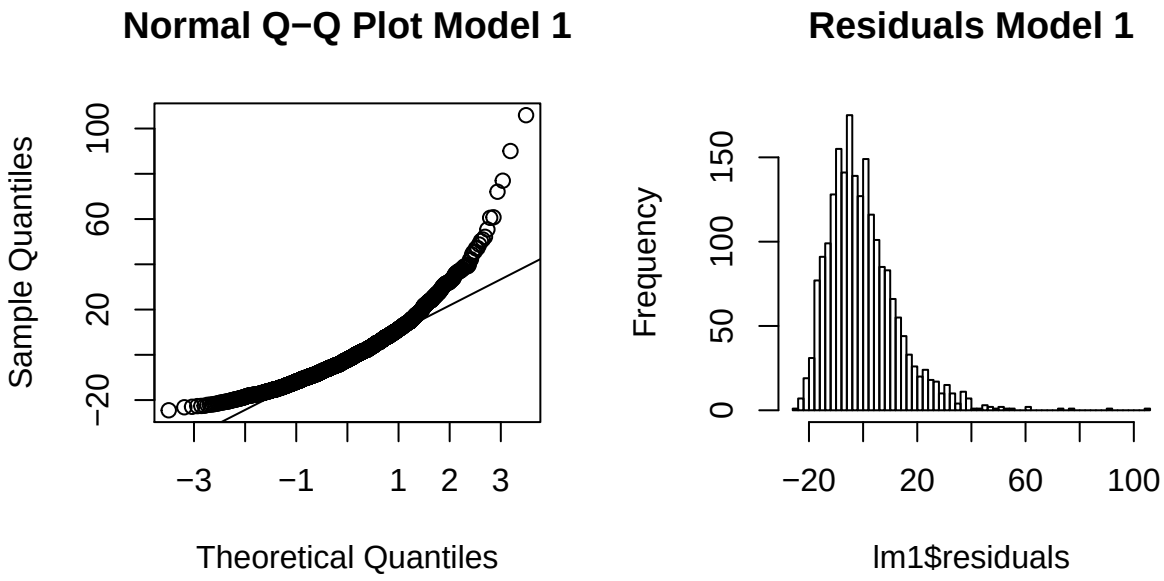

Figure 3: Model 1 Residual Diagnostics

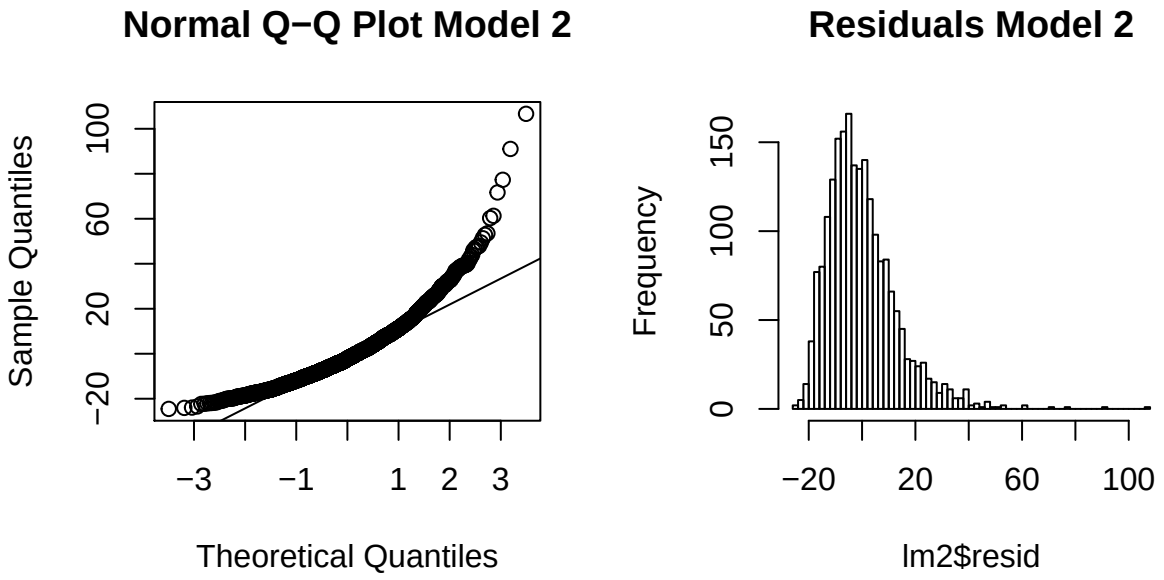

Figure 4: Model 2 Residual Diagnostics

Table 3: Average squared prediction errors

k	Model1	Model2
1	156.2985	156.4761
2	160.1634	159.3747
3	207.7491	205.2741
4	162.1530	161.4029
5	178.6121	178.8860

Table 4: Average and SD of cross-validation error values

Model	Mean	SD
1	172.9952	9.487815
2	172.2828	9.129262

Part 6: Model Diagnostics

In order to assess the model fit, I produced the diagnostic plots (Figure XXXXX). From the diagnostic plots, we can say that there are some extreme residuals that may point to a non-linear relationship. There is no relationship between residuals and all predictors and fitted values. From the plot showing the squared residuals, we can infer that the variance is non-constant. The normal Q-Q plot allows us to verify the Gaussian error assumption. It is clear that there are deviations from the Q-Q line implying that the error is not approximately normal.

Possible imporvements to the model include adding varaibles such as log(income.female), pctBlack and pctWhite which make the distribution of residuals look more normal (in the Appendix).

Part 7: Transformations

Additionally, in this model income (female) is log transformed because its distribution is skewed. pctAsian distribution is skewed and could be log transformed. However, there are 0 values, so it cannot be log transformed.

Part 8: Addressed all Model assumptions?

I was not able to address all concerns about the model assumptions. It would be better if residuals were more normally distributed (followed the qq line) but what I have right now is good enough. Also, the variance should be constant but is not (some outliers).

Results

Part 9: A Relationship Between Crime Rate and Geographic and Demographic Variables

To see if there seem to be a relationship between total crime rate and geographic and demographic variables, I fit the following 3 models.

- Model 1: Crime vs. all (geographic and demographic) variables

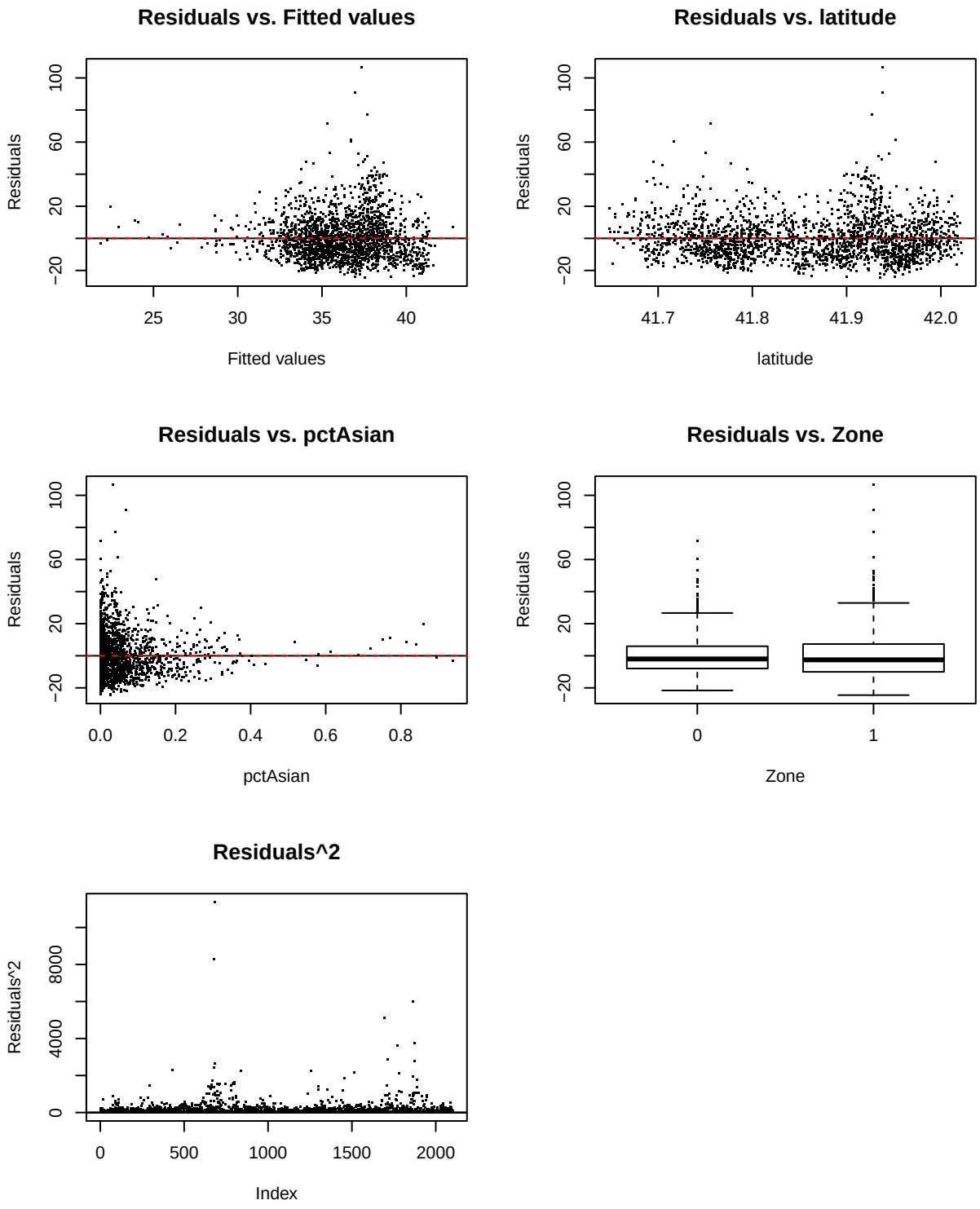

Figure 5: Model 2 Diagnostic Plots

$$
\begin{aligned}
CrimeTotal_i = \\
&\beta_0 + \beta_{log(PopulationTotal)} log(PopulationTotal)_i + \\
&\beta_{log(income.male)} log(income.male)_i + \\
&\beta_{log(income.female)} log(income.female)_i + \\
&\beta_{age.male} age.male_i + \\
&\beta_{age.female} age.female_i + \\
&\beta_{Ward} Ward_i + \\
&\beta_{latitude} latitude_i + \\
&\beta_{longitude} longitude_i + \\
&\beta_{pctWhite} pctWhite_i + \\
&\beta_{pctBlack} pctBlack_i + \\
&\beta_{pctAsian} pctAsian_i + \\
&\beta_{zone} zone_i + \epsilon_i
\end{aligned}
$$

- Model 2: Crime vs. demographic variables

$$
\begin{aligned}
CrimeTotal_i = \\
&\beta_0 + \beta_{log(PopulationTotal)} log(PopulationTotal)_i + \\
&\beta_{log(income.male)} log(income.male)_i + \\
&\beta_{log(income.female)} log(income.female)_i + \\
&\beta_{age.male} age.male_i + \\
&\beta_{age.female} age.female_i + \\
&\beta_{pctWhite} pctWhite_i + \\
&\beta_{pctBlack} pctBlack_i + \\
&\beta_{pctAsian} pctAsian_i + \epsilon_i
\end{aligned}
$$

- Model 3: Crime vs. geographic variables

$$
\begin{aligned}
CrimeTotal_i = \\
&\beta_0 + \beta_{Ward} Ward_i + \\
&\beta_{latitude} latitude_i + \\
&\beta_{longitude} longitude_i + \\
&\beta_{zone} zone_i + \epsilon_i
\end{aligned}
$$

I bootstraped the 5-fold cross-validation analysis for the models that were called Model 1 and Model 2, and Model 1 and Model 3.

I created 200 bootstrap samples each consisting of n = 2102 rows from the data set selected at random with replacement. I used the "resampling cases" form of the bootstrap.

For each bootstrap sample, I randomly divided the observations into 5 disjoint sets of equal size. Treating each of the 5 folds as test data and the other 4 as training data, I calculated prediction error for each model.

Then I computed the difference between the averages of those prediction errors for both models and called it $E(T_j*)$.

Then I drew a normal q-q plot of the $T*$ values and added the qqline. After confirming that they look like a sample of normal random variables, I run a t test to test the null hypothesis that $E(T_j*) = 0$.

Testing Model 1 vs. Model 2

Test error ($E(T_j*)$) is defined as the mean of the difference between the prediction error for Model 1 and prediction error for Model 2.

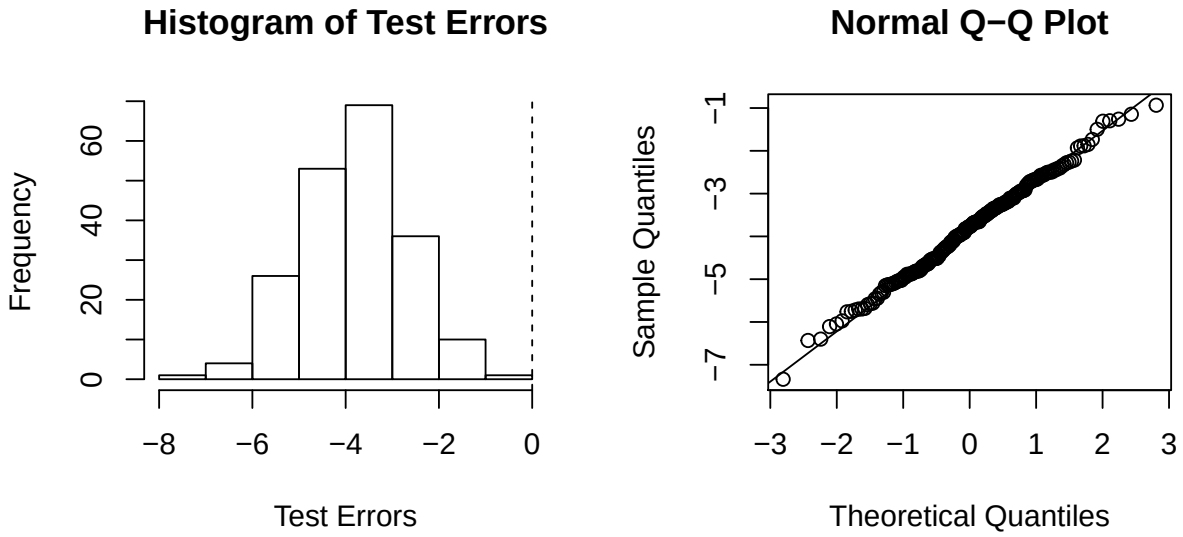

Figure 6: T* Distribution (Model 1 vs. Model 2)

It looks like Model 1 is uniformly better than Model 2. Every difference is negative. T_j* values look like a sample of normal random variables.

$$H_0 : E(T_j*) = 0$$
$$H_A : E(T_j*) \neq 0$$

Table 5: T Test

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-3.826631	-47.50599	1.518705e-110	199	-3.985473	-3.667789	One Sample t-test	two.sided

The q-q plot is remarkably straight, and the t test rejects the null hypothesis that $T_j* = 0$ at virtually all levels. It looks like Model 1 is actually better at predicting.

Testing Model 1 vs. Model 3

Test error ($E(T_j*)$) is defined as the mean of the difference between the prediction error for Model 1 and prediction error for Model 3.

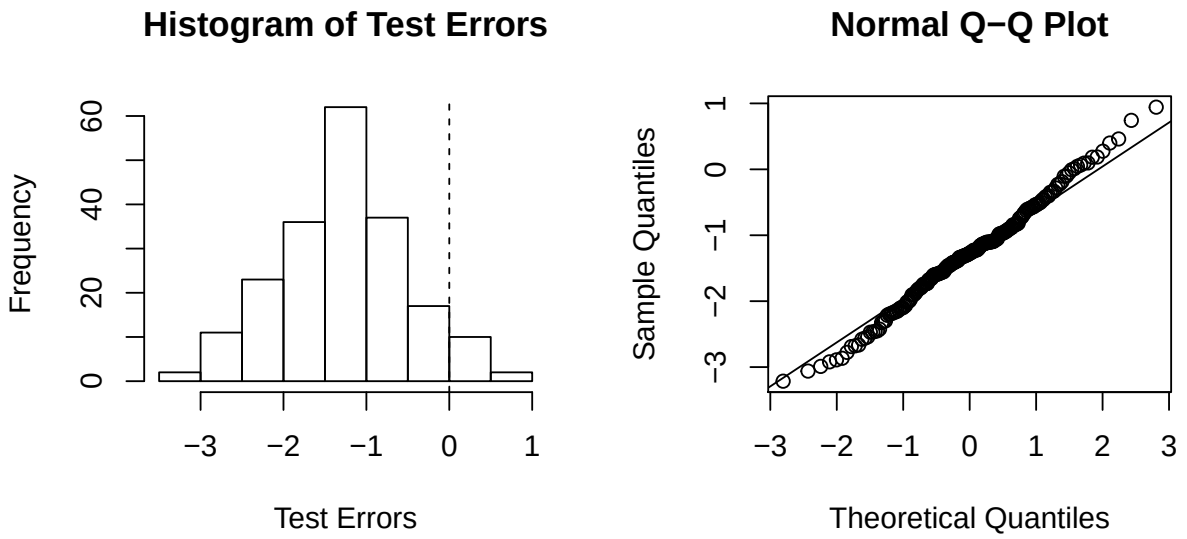

Figure 7: T* Distribution (Model 1 vs. Model 3)

10 test errors (6% of 200 T values) were positive, meaning that only for those 10 instances Model 3 performed better than Model 1. Therefore, for most instances, Model 1 performed better than Model 3. T_j* values look like a sample of normal random variables.

$$H_0 : E(T_j*) = 0$$
$$H_A : E(T_j*) \neq 0$$

Table 6: T Test

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-1.28171	-23.54324	1.97829e-59	199	-1.389065	-1.174356	One Sample t-test	two.sided

The q-q plot is remarkably straight, and the t test rejects the null hypothesis that $T_j* = 0$ at virtually all levels. It looks like Model 1 is better at predicting.

Conclusions: * Testing Model 1 vs. Model 2 - the null hypothesis is rejected. Therefore, there seem to be a relationship between total crime rate and demographic variables. * Testing Model 1 vs. Model 3 - the null hypothesis is rejected. Therefore, there seem to be a relationship between total crime rate and geographic variables.

As a result, there seem to be a relationship between total crime rate and geographic and demographic variables.

Part 10: Relationship between being above or below the river affect the cannabis-related versus non-cannabis-related crime counts

In order to test whether being above or below the river affect the cannabis-related versus non-cannabis-related crime counts in a block tract differently, I decided to set up the data and regression model so that one model is nested in a more general model. I started with the following two regression models:

$$CrimeC_i = \beta_0 + \beta_{ZoneC} Zone_i + \epsilon_i$$
$$CrimeNC_i = \beta_0 + \beta_{ZoneNC} Zone_i + \epsilon_i$$

I appended the second dataset onto the first dataset. I generated a dummy variable, Dummy, that equals 1 if the data came from CrimeNC and 0 if the data came from CrimeC. Then I generated the interaction between Zone and Dummy. I used the following formula.

$$Crime_i = \beta_0 + \beta_{Dummy} Dummy_i + \beta_{Zone} Zone_i + \beta_{Dummy*Zone} Dummy * Zone_i + \epsilon_i$$

The coefficient for the interaction between the Dummy variable and Zone shows the difference between the two initial slopes: β_{ZoneNC} and β_{ZoneC}

We would like to test:

$$H_0 : \beta_{ZoneNC} = \beta_{ZoneC}$$
$$H_A : \beta_{ZoneNC} \neq \beta_{ZoneC}$$

This is equivalent to testing:

$$H_0 : \beta_{Dummy*Zone} = 0$$
$$H_A : \beta_{Dummy*Zone} \neq 0$$

Table 7: Regression Results

	Dependent variable:		
	CrimeC	CrimeNC	Crime
	(1)	(2)	(3)
as.factor(zone)1	0.880^{***}	1.663^{***}	
	(0.313)	(0.397)	
Dummy			-3.478^{***}
			(0.379)
as.factor(Zone)1			0.880^{**}
			(0.358)
Dummy:as.factor(Zone)1			0.783
			(0.506)
Constant	19.100^{***}	15.621^{***}	19.100^{***}
	(0.235)	(0.297)	(0.268)
Observations	2,102	2,102	4,204
R^2	0.004	0.008	0.040
Adjusted R^2	0.003	0.008	0.039
Residual Std. Error	7.133 (df = 2100)	9.042 (df = 2100)	8.144 (df = 4200)
F Statistic	7.882^{***} (df = 1; 2100)	17.521^{***} (df = 1; 2100)	58.019^{***} (df = 3; 4200)
Note:			*p<0.1; **p<0.05; ***p<0.01

The t value is 1.547 and p value is 0.122. By assuming our usual treshold of $\alpha = 0.05$, we fail to reject H0. This indicates that the regression coefficient B_NC is not significantly different from B_C. Therefore, not enough evidence is available to suggest that being above or below the river affect the cannabis-related versus non-cannabis-related crime counts in a block tract differently.

Part 11: Ranking of Wards with Highest Crime Rates

Table 8: Highest Crime

	Ward	CrimeTotal	PopulationTotal
35	35	2361	53049
1	1	2203	59392
21	21	2173	49637
34	34	2060	53472
19	19	2031	50471

Wards 35, 1, 21, 34 and 19 have the highest count of crime. The crime count is included in the table below.

Table 9: Highest Positive Residuals

Ward	Residuals	Population	Crime	CrimePerPopulationPerc	RankingPopulation	RankingResiduals
35	839.7261	53049	2361	4.450602	23	1
21	671.9627	49637	2173	4.377783	33	2
1	644.1058	59392	2203	3.709254	7	3
34	536.2173	53472	2060	3.852483	21	4
19	525.0162	50471	2031	4.024093	31	5

The wards with highest crimes (35, 1, 21, 34 and 19) have the highest positive residuals meaning that the actual value of CrimeTotal was higher than the predicted value of CrimeTotal. For those wards, the model underestimates CrimeTotal. Those Walds have populations in the middle of the range and high number of crimes. Those wards also have the highest crime/person rate. Therefore, correcting by population size is not reasonable. The ranking is included in the Appendix.

Part 12: Relationship Between Cannabis and Non-cannabis Related Police Reports

In order to see whether there is a relationship between cannabis and non-cannabis related police reports in each block group, I run the following model.

$$CrimeC_i = \beta_0 + \beta_{log(CrimeNC)} log(CrimeNC) + \epsilon_i$$

I decided to log transform CrimeNC because there are outliers that make the distribution very skewed (plot of CrimeC vc. CrimeNC).

I was interested in testing the following hypotheses:

$$H_0 : \beta_{log(CrimeNC)} = 0$$
$$H_A : \beta_{log(CrimeNC)} \neq 0$$

The coefficient for log(CrimeNC) is significant at <2e-16, therefore we reject H_0, so there seem to be a relationship between cannabis and non-cannabis related police reports in each block group.

To see what happens to this relationship when you control for the other variables, I fit the following model.

Table 10: Regression Results

	Dependent variable:	
	CrimeC	
	(1)	(2)
log(CrimeNC)	4.333***	4.265***
	(0.278)	(0.282)
log(PopulationTotal)		0.129
		(0.406)
log(income.male)		0.298
		(0.351)
log(income.female)		0.533
		(0.339)
age.male		−0.021
		(0.021)
age.female		−0.014
		(0.021)
Ward		−0.0004
		(0.016)
latitude		2.009
		(4.067)
longitude		13.543***
		(3.070)
pctWhite		−2.409
		(1.497)
pctBlack		−0.581
		(1.152)
pctAsian		−3.372
		(2.177)
as.factor(zone)1		1.536**
		(0.682)
Constant	8.024***	1,103.943***
	(0.757)	(297.314)
Observations	2,102	2,102
R^2	0.104	0.124
Adjusted R^2	0.103	0.119
Residual Std. Error	6.766 (df = 2100)	6.707 (df = 2088)
F Statistic	242.829*** (df = 1; 2100)	22.775*** (df = 13; 2088)
Note:		*p<0.1; **p<0.05; ***p<0.01

$$
\begin{aligned}
CrimeC_i = & \\
& \beta_0 + \beta_{log(CrimeNC)} log(CrimeNC)_i + \\
& \beta_{log(PopulationTotal)} log(PopulationTotal)_i + \\
& \beta_{log(income.male)} log(income.male)_i + \\
& \beta_{log(income.female)} log(income.female)_i + \\
& \beta_{age.male} age.male_i + \\
& \beta_{age.female} age.female_i + \\
& \beta_{Ward} Ward_i + \\
& \beta_{latitude} latitude_i + \\
& \beta_{longitude} longitude_i + \\
& \beta_{pctWhite} pctWhite_i + \\
& \beta_{pctBlack} pctBlack_i + \\
& \beta_{pctAsian} pctAsian_i + \\
& \beta_{zone} zone_i + \epsilon_i
\end{aligned}
$$

I am interested in testing the following hypotheses:

$$
\begin{aligned}
H_0 &: \beta_{log(CrimeNC)} = 0 \\
H_A &: \beta_{log(CrimeNC)} \neq 0
\end{aligned}
$$

When I control for the other variables (transformed), the coefficient for log(CrimeNC) is still significant at < 2e-16. This means that the additional predictors are not strongly related to log(CrimeNC). In other words, predictor variables are not strongly related, so there is no multicollinearity. CrimeNC is not correlated with all other predictors (besides CrimeTotal). This agrres with the correlation matrix graph in the Appendix.

Conclusions/Discussion

Part 13

There is enough evidence to support our hypothesis that demographic and geographic factors relate to narcotic-related crime in Chicago. Narcotic-related crimes depend on demographic and geographic factors. Cannabis and noncannabis-related crimes are correlated. Nevertheless, the correlation does not imply causation.

The analysis suggests that there is statistical evidence that narcotic-related crimes depend on demographic and geographic factors. Additionally, there is not enough evidence available to suggest that being above or below the river affect the cannabis-related versus non-cannabis-related crime counts in a block tract differently.

One of the possible reasons for this finding may be the fact that

Further analysis needs to be done on the relationship between the narcotic-related crime in Chicago and many other predictor variables that could potentially present a relationship with our response variable.

Appendix

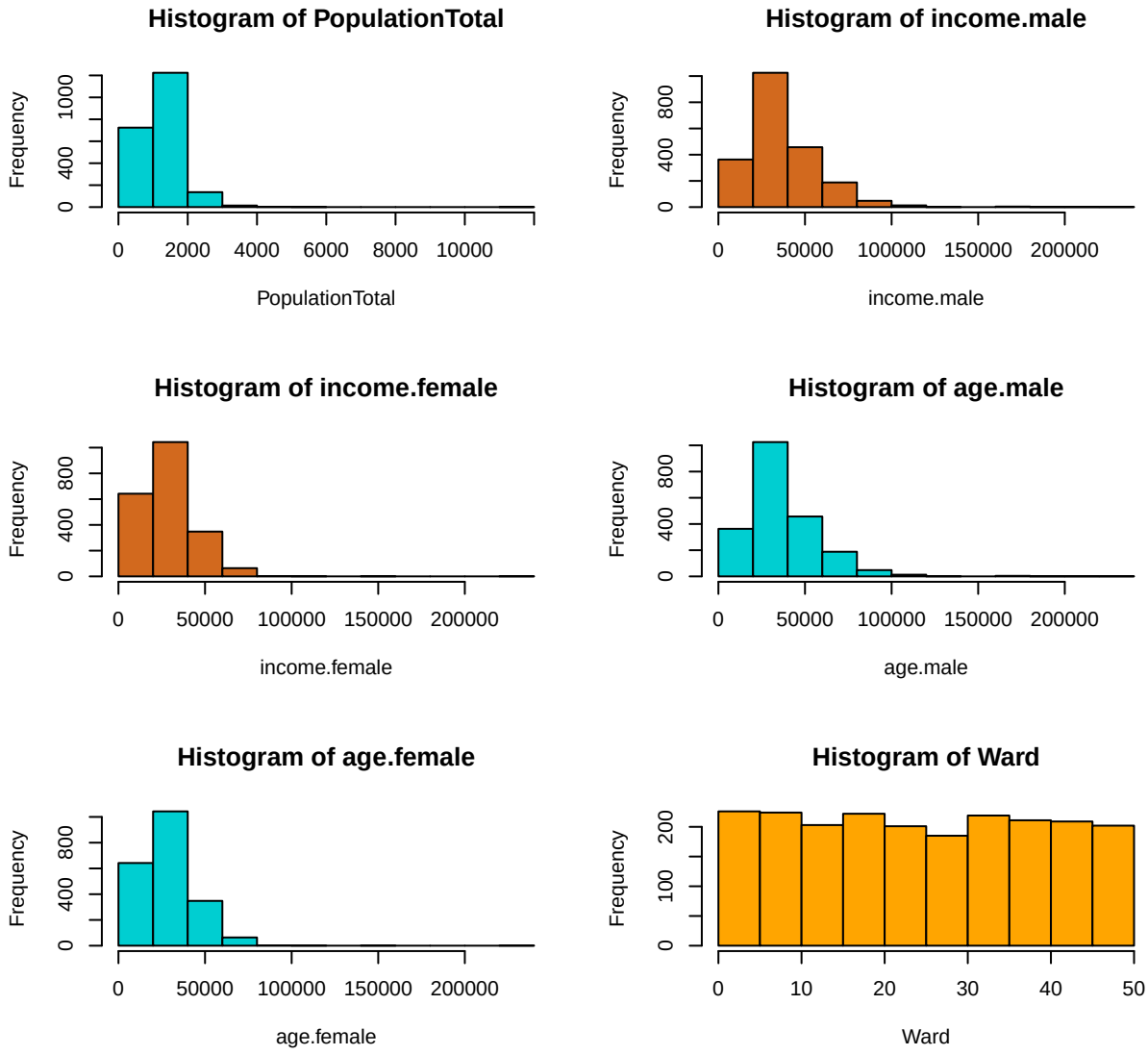

Figure 8: Univariate EDA (I)

Part 6

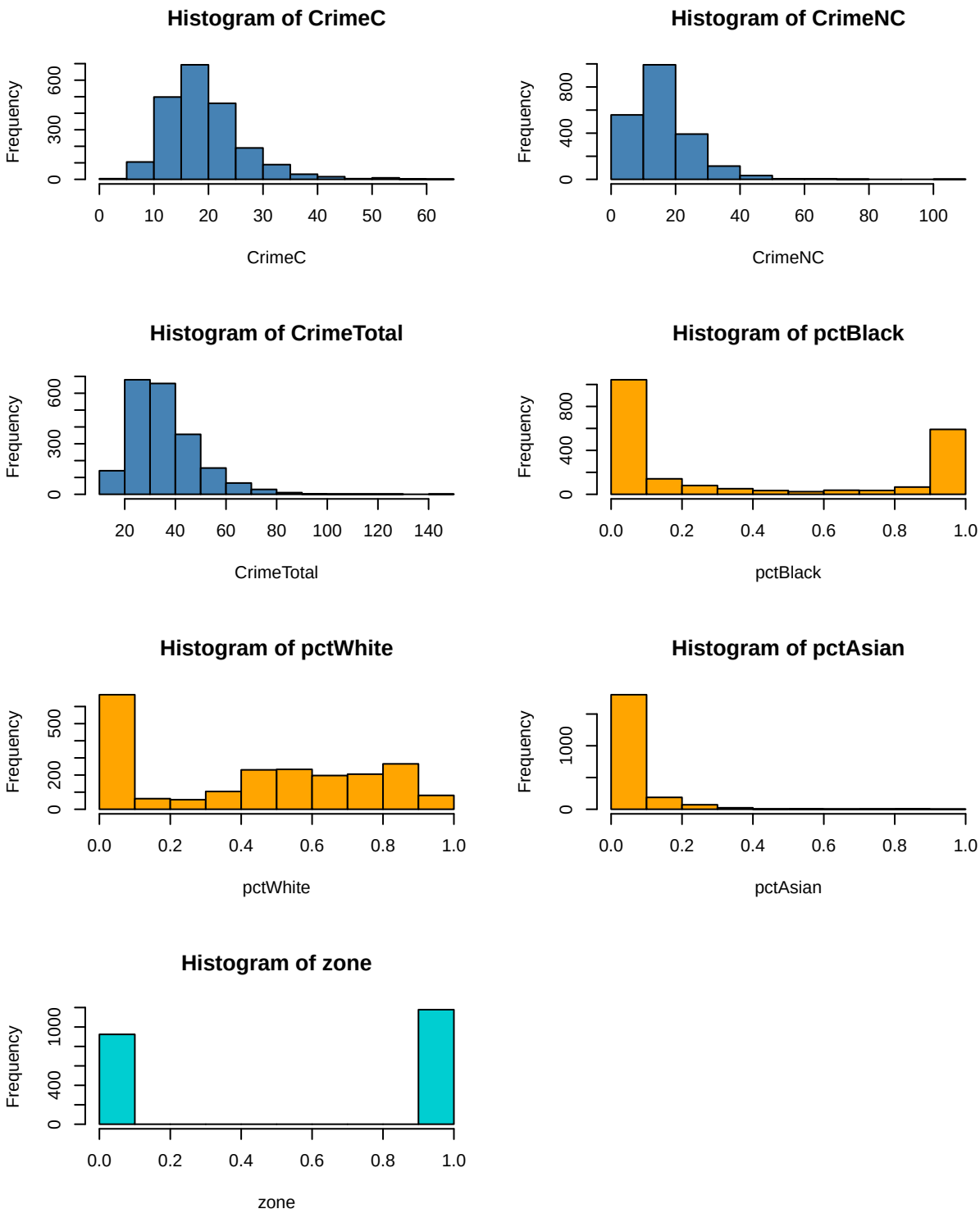

Figure 9: Univariate EDA (II)

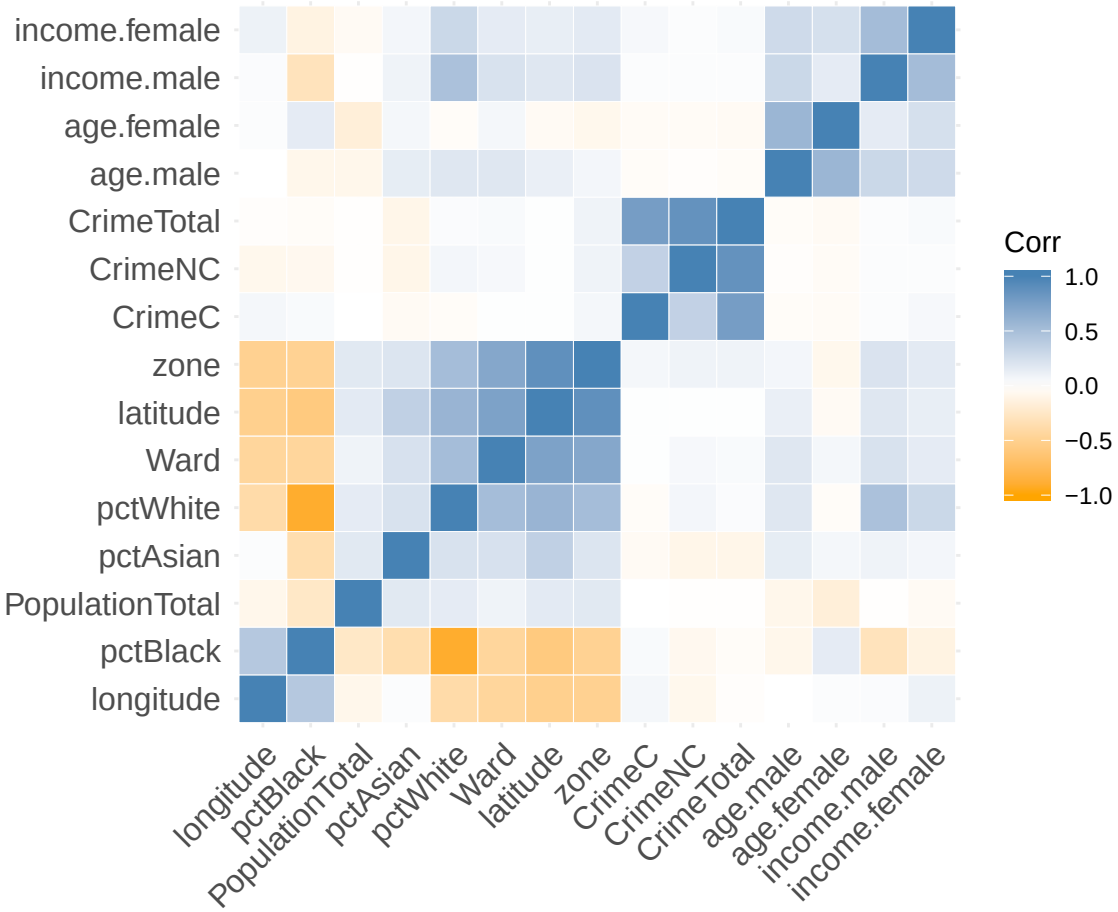

Figure 10: Pairwise correlations

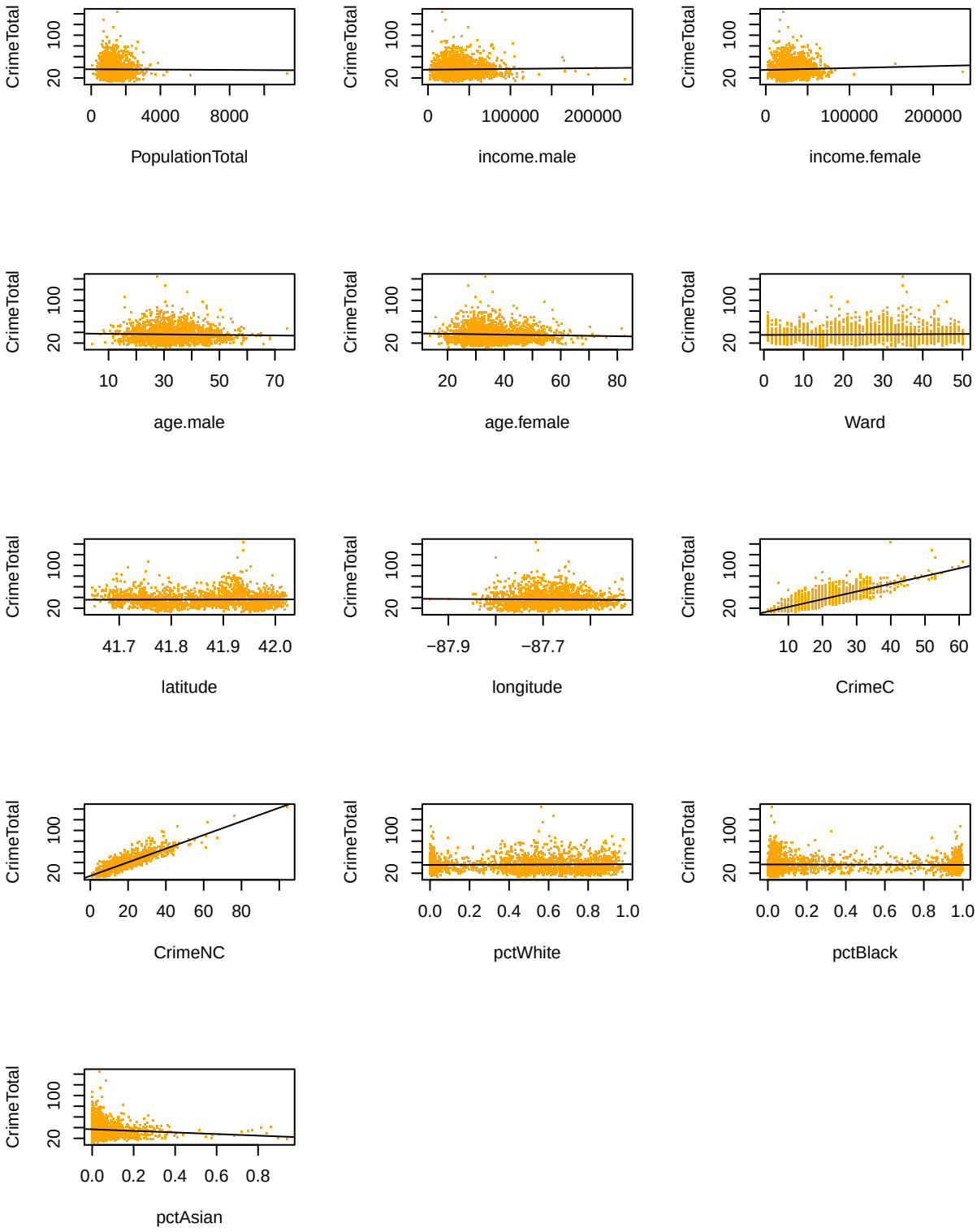

Figure 11: Multivariate EDA (CrimeTotal, Continuous)

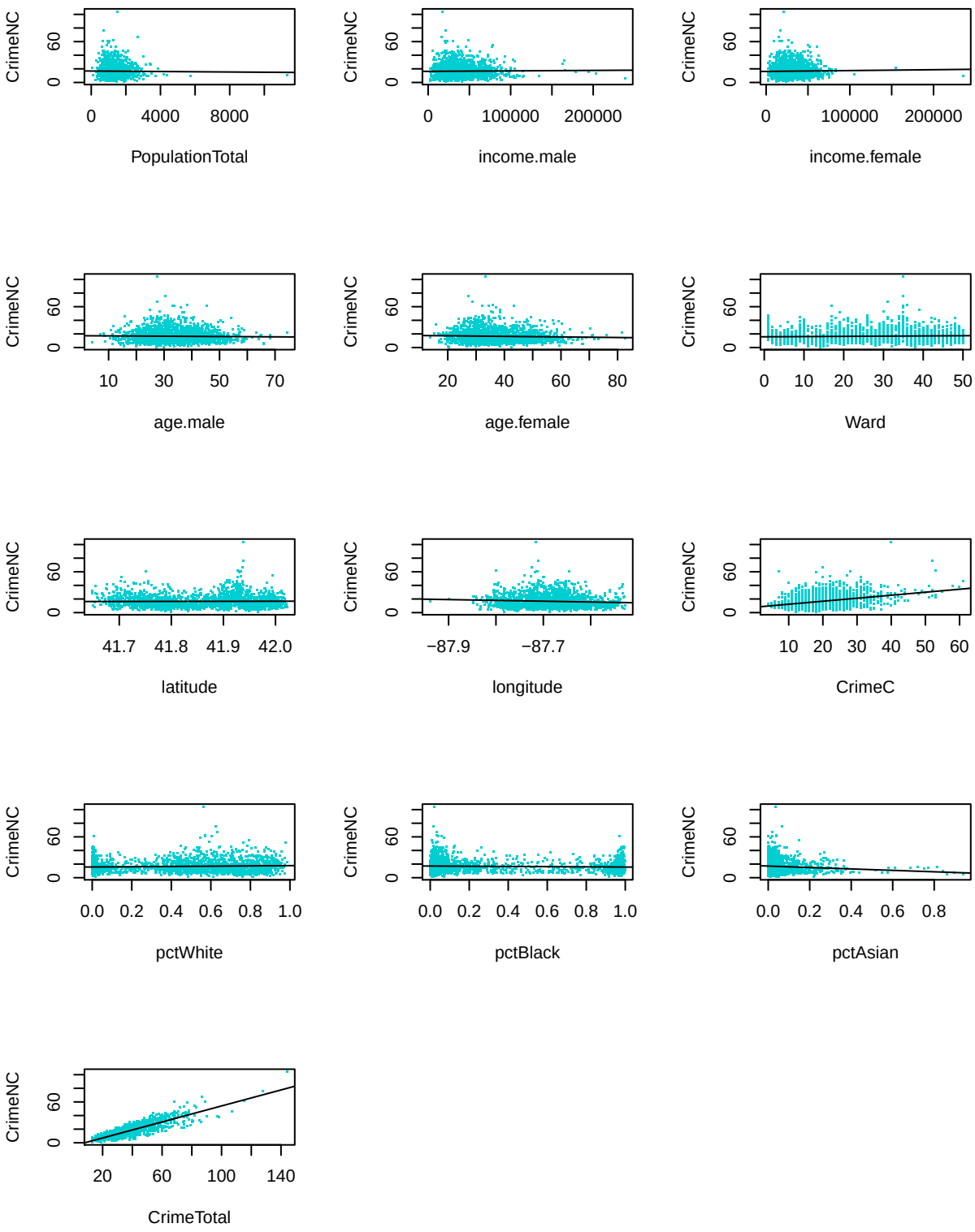

Figure 12: Multivariate EDA (CrimeNC, Continuous)

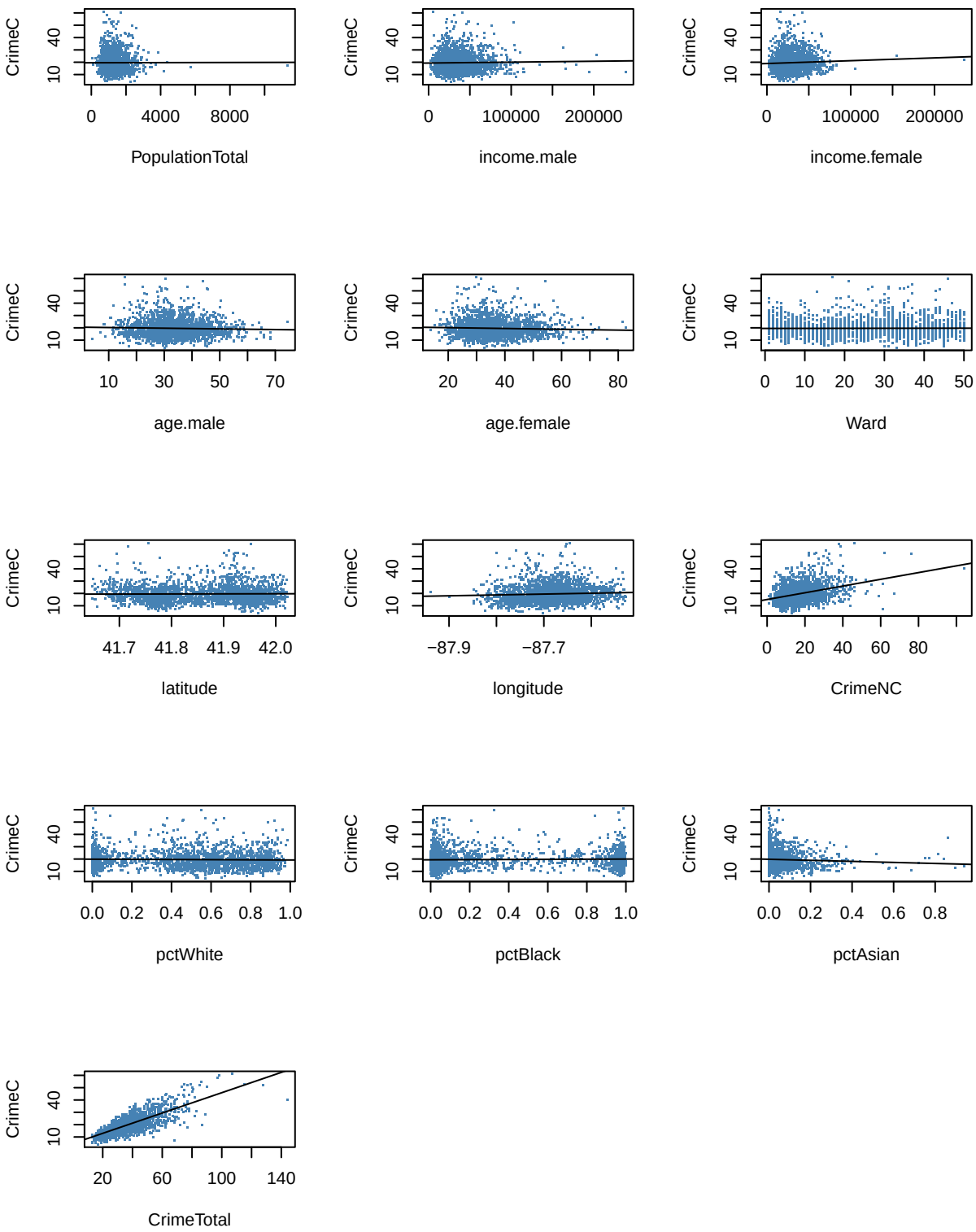

Figure 13: Multivariate EDA (CrimeC, Continuous)

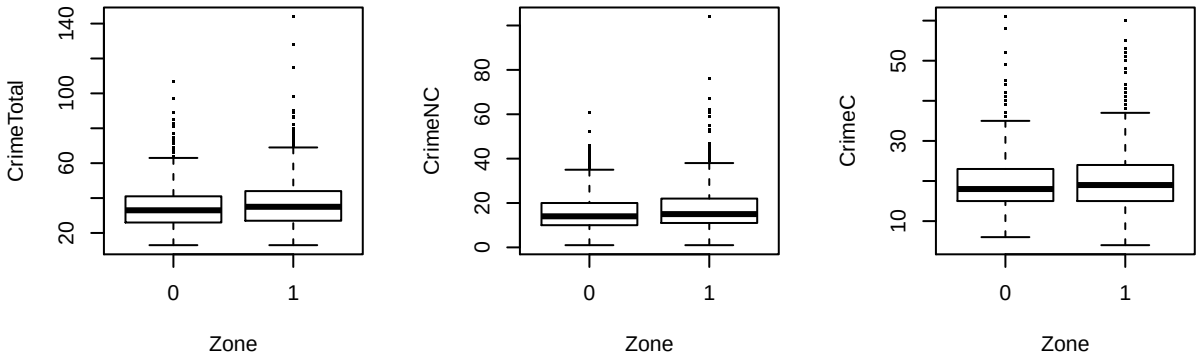

Figure 14: Multivariate EDA (categorical)

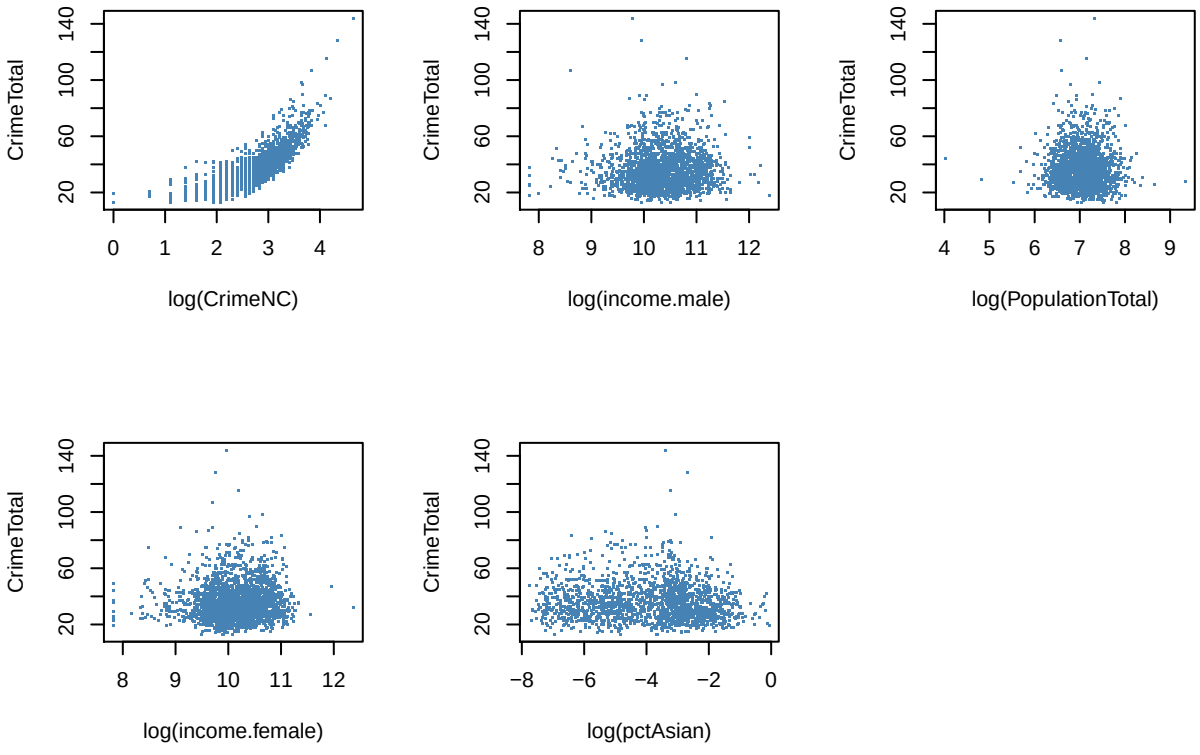

Figure 15: Log Transformed Variable Fit

Table 11: Alternative Model

	Dependent variable:
	CrimeTotal
latitude	-39.383^{***}
	(7.401)
pctAsian	-14.427^{***}
	(3.958)
as.factor(zone)1	8.945^{***}
	(1.285)
log(income.female)	1.235^{**}
	(0.607)
pctBlack	-4.795^{**}
	(1.970)
pctWhite	-4.336^{*}
	(2.539)
Constant	$1{,}671.435^{***}$
	(309.421)
Observations	2,102
R^2	0.039
Adjusted R^2	0.037
Residual Std. Error	13.110 (df = 2095)
F Statistic	14.310^{***} (df = 6; 2095)
Note:	*p<0.1; **p<0.05; ***p<0.01

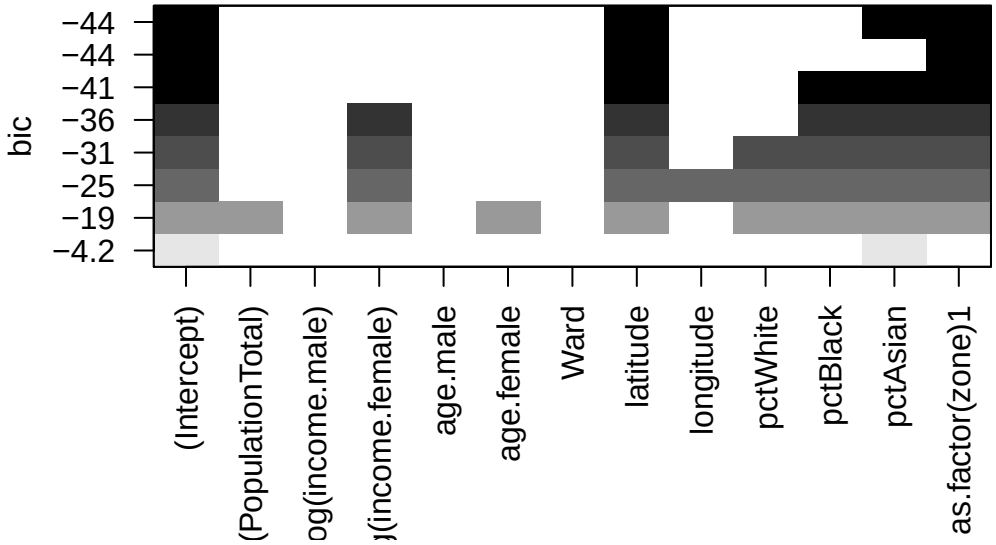

Figure 16: Regression Subset Selection (Schwartz' BIC)

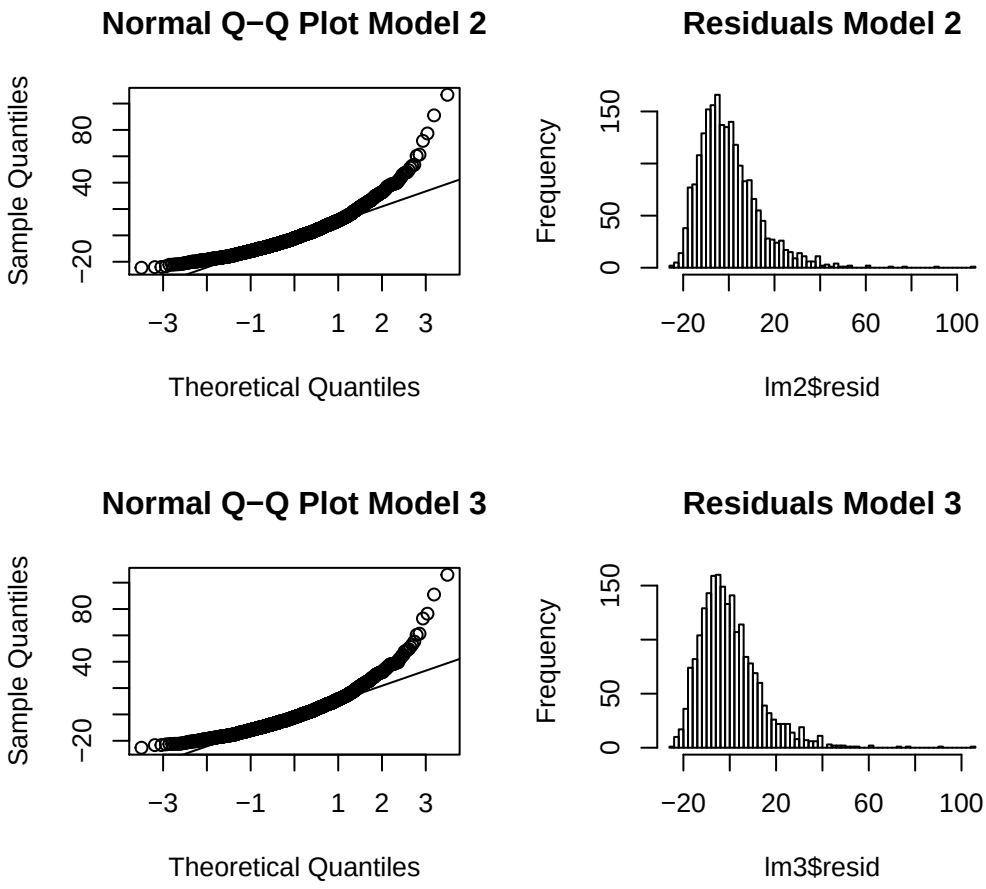

Figure 17: Model 2 and 3 Residual Diagnostics