The Effect of Public Transportation on Criminal Homicides In Five American Cities

Jason Baker, Joanna Chae, Paul Coletti, Alexandra DeKinder,
Peter Kolodziej, Rui Liu, Jess Matthew
Columbia University

Author Note

This paper is the culmination of a graduate-level Advanced Data Analysis course within the Columbia University department of Statistics.

GR5291 – Advanced Data Analysis, Professor Demissie Alemayehu

Correspondence concerning this paper should be addressed to the authors via their respective email addresses.

Contact: jjb2225@columbia.edu, jjc2274@columbia.edu, pec2123@columbia.edu, ad3540@columbia.edu, pgk2115@columbia.edu, rl3023@columbia.edu, jm4742@columbia.edu

Abstract

The rate of criminal homicide in the United States has declined by more than 50% since the 1990s. Scholars have attributed this decline to increased policing, mass incarceration, and greater neighborhood cohesion (Zimring 2006; Levitt 2004; Travis, Western, and Redburn 2014; Sharkey, Torrats-Espinosa, and Taykar 2017). This study considers how structural characteristics of neighborhoods, specifically public transportation access, affects criminal homicide in five cities across the United States – Baltimore, MD; Chapel Hill, NC; Chicago, IL; Cincinnati, OH; Los Angeles, CA; and New York City, NY. Using zero-inflated negative binomial models, we find that a greater density of public transportation stops within a census tract is associated with a smaller number of criminal homicide incidents. This relationship is more pronounced in Chicago, Il; Los Angeles, CA; and Cincinnati, OH. Future studies should explore how public transportation may prevent criminal homicide from occurring.

The Effect of Public Transportation on Criminal Homicides in Five American Cities

Public transportation has been growing around the United States in order to provide a cleaner, cheaper way to efficiently travel. Yet, a concern around extended public transportation might be the fear of an increased amount of crime. Because public transportation may allow criminals to escape the crime scene after breaking laws, there could be an increase in crime in their area if there is more public transportation, and, therefore, greater access to it.

This paper outlines an observational study regarding the relationship between the number of public transportation stops in a census tract and the amount of crime there. For five different United States cities, this study looks into the relationship between the number of homicides in a census tract from January 2010 through September 2019 and the number of public transportation stops in that census tract. The study then tests whether any such relationships are significant in any of the five cities included.

The study uses data from the Police Data Initiative. As part of the initiative, 53 unique police districts publish crime incident data online. It also takes into consideration the American Community Survey from 2008-2010, published by the United States Census Bureau, to obtain a number of additional sociodemographic characteristics for each census tract. This helps appropriately ascribe meaning to the impact of a census tract's public transportation accessibility on the number of homicides in that tract, in a larger and more complete context. These characteristics include proportions of residents who were Black, Hispanic, Asian, non-White, single parents, foreign-born, have a college degree or higher, and live in poverty. Finally, the data includes the proportion of people who are 16 and older and have a commute shorter than 15 minutes, as well as the population density of each census tract.

Objectives

A preliminary objective is to test the necessity of such a study and understand in what context its results may be useful. A literature review revealed that past studies have primarily focused on finding whether there is a correlation between theft and public transportation in various areas around the United States. However, whether structural characteristics of neighborhoods, such as public transportation, affect homicide rates is an open question that this study intends to explore, thereby adding to the conversation about crime and public transportation.

The primary objective, then, is to investigate the connection between homicides and public transportation, and to determine whether the connection is meaningful in any way. Although the data includes visualizations and analysis for five different cities, the study first focuses on Chicago. This was an intentional selection by the authors, who surmised that Chicago's transportation network would be the most indicative of the five cities because it is among the most comprehensive. Chicago has the second largest transportation network in the United States with eight subway lines and 129 busway routes. The city is also an appealing case because it has the highest number of homicides of the six different cities observed in the study. The goals for each city thereafter are the same: to determine whether there is a correlation between murder and public transportation.

To accomplish this larger objective, the study first aims to visualize the dataset in an appropriate way. The study then looks to perform exploratory data analysis on the dataset to understand which model may be most useful in describing the relationship between public transportation and criminal homicide, and then to choose an appropriate model to test the significance of the relationship. Finally, the study aims to test whether the assumptions of the model are violated by the data, and evaluate the performance of the model.

Methods and Theory

To assess the utility of and precedence for the study, the authors conducted a comprehensive literature review. A literature review is a general examination of previously published information regarding a particular topic. Often limited to a certain time period, literature reviews require human inferences and decision-making to derive the importance of the examined texts and place it in context. Although a literature review can include extended synthesis of the literature, the literature review in this study was a process by which the authors sought and summarized any previously published literature to determine what sort of prior investigation, if any, had been done on the relationship between criminal homicide and public transportation. During a literature review, it is not necessary to contribute anything new to the source material, but only to gain insight into its content.

Before conducting the data analysis, the authors had to perform four stages of data cleaning. First, each incident of crime was assigned a United States census tract code based on its longitude and latitude. Each tract is populated by an average of 4,000 residents and are smaller subdivisions of a county. Second, because districts provided varying levels of detail about crime type for each incident, the descriptions had to be streamlined and standardized according to the categories of violent crime as defined by the Federal Bureau of Investigation. The authors utilized string matching techniques to convert local police crime descriptions to these standard categories. These categories are: criminal homicide, forcible rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle-theft, and arson. Third, a new dataframe was created to summarize the incident data by census tract, such that the number of criminal homicide cases that occurred would be available for each census tract. Hence, while the raw data observations each represented a single incident, the data cleaning process yielded a dataset for each city in which an observation represented a unique census tract.

Fourth, the authors calculated a density of public transportation stops per square mile in every census tract. All the census tracts in each of the six cities were matched with the geographic coordinates corresponding to public transportation stop locations, which were extracted from each city's respective general transit feed. The total number of stops in each tract was divided by the area of the tract.

The six districts this study observes were chosen based on whether the data was geocoded, and the amount of time over which data was available. Ultimately, the six chosen districts captured what the authors perceived to be a good breadth of United States cities vis-à-vis culture, crowdedness, and geographic location: Baltimore, MD; Chapel Hill, NC; Chicago, IL; Cincinnati, OH; Los Angeles, CA; and New York City, NY. However, through the data cleaning process, it became clear that Chapel Hill, NC had very few cases of criminal homicide. Although it would have provided meaningful heterogeneity as a city that has a less dense public transportation system, the analysis did not include Chapel Hill, NC because of this lack of data.

Exploratory data analysis entailed visualizing the data using various techniques. A Tableau dashboard helped the authors to understand the data subject to certain temporal filters. Meanwhile, an R script was used to visualize a heatmap of all the crime specifically for Chicago.

This heatmap can be seen in Figure A##, and relates the crime density to the heavy rail system in Chicago through superimposition. Finally, a map of all of Chicago's public transportation (Figure A##) helped to understand what to expect as far as the number of stops in a census tract.

Further exploratory data analysis included using histograms to get a sense for the distribution of criminal homicide per census tract in each of the five cities. These histograms were generated with R code, and are a way of grouping observations into bins of roughly equal size to visualize how data is distributed. The height of each bar corresponds with how many observations fall into the given bin. It also gives a good sense of how the data is spread out.

The literature review revealed that most frequently, a fixed-effects ordinary least squares (OLS) regression model is the primary model considered among the examined studies. A fixed effects OLS model takes the form $y_{it} = X_{it}\beta + \alpha_i + e_{it}$, where $i = 1, ..., N$ observations and $t = 1, ..., T$ time periods. α_i is an individual effect that cannot be observed. These models included social determinants, along with binary variables indicating whether or not public transportation existed in a given neighborhood, as dependent variables. Further, they determined the type of neighborhood by the quantity of residential buildings versus commercial buildings in the neighborhood.

Based on the results of the exploratory data analysis and the literature review, the first model tested was an ordinary least squares linear regression model. This model, which takes the form $y_i = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon_i$ for $i = 1, ..., N$ intends to find a set of coefficients $\hat{\beta}_0, ..., \hat{\beta}_p$ such that the quantity $MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$ is minimized. It assumes that the errors are uncorrelated across observations, that the error is normally distributed with mean zero and constant variance, that predictors are uncorrelated with the error term, that no predictor is a linear combination of the other predictors, and that the relationship between the predictor variables and the dependent variable is linear in the first place. Another model the authors tested was a log-transformed least squares regression model, which uses the equation $\ln ln\left(y_i\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon_i$ for $i = 1, ..., N$ to minimize the same objective function.

After it became clear that assumptions of the original models were violated or the coefficients were not significant, the final model used is a zero-inflated negative binomial model. This model is a modification of a negative binomial model. It is appropriate when the data is a mixture of two processes – one that only produces zeros and one that follows a negative binomial distribution. The simple Poisson model, $\ln ln\ (\mu) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$, is not appropriate in this case for two reasons: first, the data has a large number of census tracts with zero homicides (which can be seen in the histogram of homicides in Chicago in Figure A#); second, the Poisson model has the assumption that mean and variance is equal which the data does not meet.

The first problem of zeros can be addressed by using a zero-inflated Poisson model which assumes that some of the zero values grew from the Poisson process, but that there are other tracts that we would not expect to have a homicide occur in the first place. The zero-inflated Poisson models combines these two processes. It assumes that some tracts would not expect to have a homicide occur, a reasonable assumption because homicide is a very rare crime and census tracts have a wide range of population densities.

The second problem with the regular Poisson model can be addressed by instead using the negative binomial model. Violating the assumption that the mean and variance are equal means the Poisson model is prone to overdispersion, the term for when the sample mean is substantially smaller than the sample variance. Overdispersion is problematic because it leads to underestimation of the standard error, and an increased Type I error rate. The negative binomial model is an alternative to the Poisson model that is more robust to overdispersion; its distribution has probability mass function $f(k) = \frac{\Gamma(r+k)}{\Gamma(k+1)\Gamma(r)} p^{r}(1 - p)^{k}$, where $r > 0$, and $0 < p < 1$ are parameters. The distribution has mean $\mu = r\frac{1-p}{p}$ and variance $\sigma^{2} = r\frac{1-p}{p^{2}}$.

To test the multicollinearity assumption among the variables included in the zero-inflated Poisson and negative binomial models, the authors used a correlation plot. A correlation plot is a table showing how strongly correlated each pair of variables is. Each cell in the plot indicates the strength of the correlation between a pair of variables. The reported figure may be any measure of correlation, but it is most common to use the Pearson correlation coefficient $r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^{2}}\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^{2}}}$. Typically, the plots are used as a diagnostic, as it is in this study.

The results of the plot led to a Principal Components Analysis on all the socio-demographic variables in the dataset. A principal components analysis is a dimensionality reduction technique which uses orthogonal transformation to convert data into a set of linearly uncorrelated variables arranged by the extent to which they explain the variability in the original data. The transformation is defined by mapping each observation to a new vector of principal component scores using a set of weights. Each subsequent principal component can be found by subtracting the already-found principal components from the original data. In this study, the principal components analysis led further to creation of a scale ranging from 1 to 5 to be included as an independent variable in the regression models.

Results

The first goal of determining the context of the study was achieved with a literature review. According to the Department of Justice, since 1990, there has been a decline in the rate of criminal homicide in the United States. The rate of homicides in 2010 was 4.8 homicides per 1000,000 residents, which was nearly half of the rate from 1992, at 9.3 homicides per 100,000 residents. Zimring (2006) first referred to this trend as the “Great crime decline.” Many explanations have been proposed, including an increase in policing, mass incarceration, the decline of the crack epidemic, and even the legalization of abortion (Levit 2004; Travis, Western, and Redburn 2014). Sharkey et al (2017) also propose that greater community engagement as facilitated by non-profit organizations has also contributed to the decline. Because most studies found that crimes occur more frequently in neighborhoods with easier access to public transportation (Neiss, 2015), the study in this paper was deemed worthwhile of pursuit and execution.

Exploratory data analysis revealed that the distribution of the number of public transportation stops per square mile by census tract was found to be right skewed for all of the cities. Histograms in Figures A##-A## verify this observation. To address the skew in the distribution, we standardized this measure, such that one-unit increase corresponds with an increase of one standard deviation above the mean.

The initial model of ordinary least squares was abandoned because certain assumptions were not met. The plots of residuals versus fitted values in Figures A## and A## show that multicollinearity and heteroskedasticity (nonconstance of error variance) could be present in the model. The Q-Q plots in Figures A## and A## show that the data is skewed for the linear and log-transformed models.

The correlation plot in Figure A## shows that many of the socio-demographic characteristics are correlated with one another. For example, the share of people who identify as Hispanic is highly correlated with the share of foreign-born residents in a census tract.

The Principal Components Analysis resulted in creation of five new variables on which to regress the independent variable, corresponding with the first five principal components of the demographic data.

Finally, the outputs of the zero-inflated negative binomial models after principal components selection, and the zero-inflated binomial models, shown in Figures A##-A##, reveal that there is significance in a few cities, but not all five. Specifically, Los Angeles and New York City have significant coefficient estimates for the number of transportation stops per square mile, with p-values of 0.0232 and $4.39E-5$, respectively. These p-values are significant at the $\alpha = 0.05$ level, suggesting that there are areas of the country where public transportation stops may be correlated with the number of criminal homicides in a certain census tract.

Conclusions

The study set out to explore various relationships between the density of public transportation stops in a census tract with the number of criminal homicides in that tract. While some of the results confirmed initial suspicions, like the model for Los Angeles, California and New York City, New York, others were not as significant, like for Cincinnati, Ohio. Even in a case where a significant result was achieved, there is no guarantee that the correlation implies a causal relationship. Although certain demographic variables were taken into account, a confounding factor like population size, political makeup of a census tract, or repeat offenders living nearby might have a greater effect on the number of homicides. Further research on criminal homicides might extend the studies to more cities in the united states, or a longer time period than 2010-2019, or it may explore whether access to lesser public transportation like rideshare cellphone applications or pay-to-ride bikes and scooters have any relationship with crime like homicide, all of which seem to reasonable along the same line of thinking as buses and trains. Tests on these variables and others may indicate if an increase in public transportation does in fact lead to an increase in criminal homicides.

References

https://www.bjs.gov/content/pub/pdf/htus8008.pdf
https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf
https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/offense-definitions
https://collected.jcu.edu/cgi/viewcontent.cgi?article=1003&context=jep
Crime Data:
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard5cd6-ry5g
Chicago Transportation Authority: https://www.transitchicago.com
US Census Data: https://opportunityinsights.org/data/

Di, Wang. "The Impact of Mass Transit of Public Security – A Study of Bay Area Rapid Transit in San Francisco." *Transportation Research Procedia*, Elsevier, 8 June 2017, https://www.sciencedirect.com/science/article/pii/S2352146517304362.

Levitt, Steven D. 2004. "Understanding Why Crime Fell in the 1990s: Four Factors That Explain the Decline and Six That Do Not." *Journal of Economic Perspectives* 18(1):163-90.

Neiss, Morga. 2015. "Does Public Transit Affect Crime? The Addition of a Bus Line in Cleveland." *The Journal of Economics and Politics* 22:17.

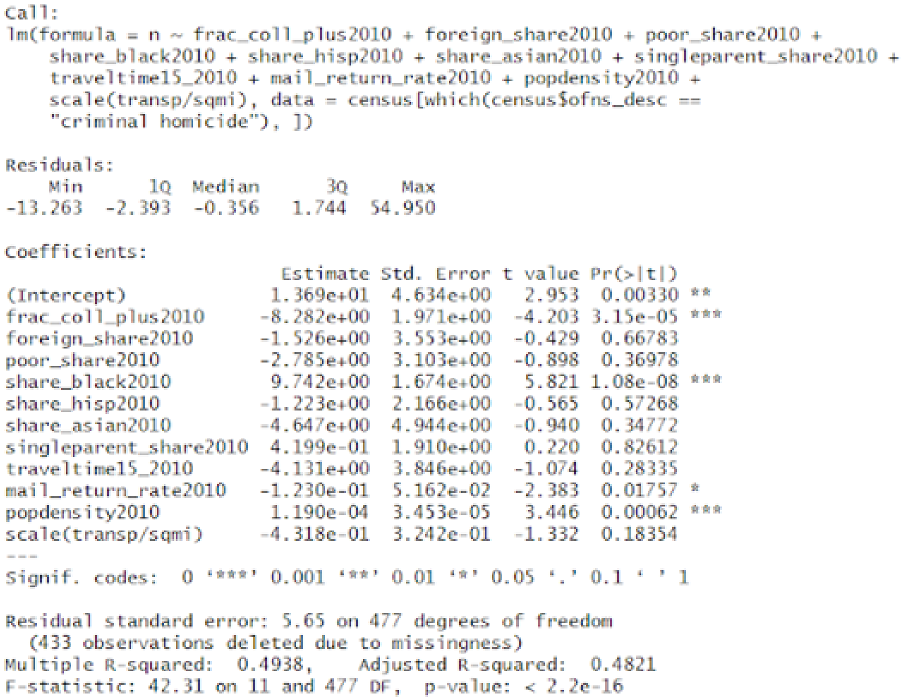
Sharkey Patric T, Gerard Torrats-Espinosa, and Delaram Takyar. 2017. Community and the crime decline: The causal effect of local nonprofits on violent crime. *American Sociological Review* 82(6): 1214-1240.

Travis, Jeremy, Bruce Western, and Steve Redburn, eds. 2014. *The Growth of Incarceration in the United States: Exploring Causes and Consequences*. Washington, DC: National Academies Press.

Willoughby, Jack. 2014. "The Effect of Public Transportation on Crime: An Analysis of Durhams' Bull City Connector." *Urban Economics*, https://sites.duke.edu/urbaneconomics/?p=1215.

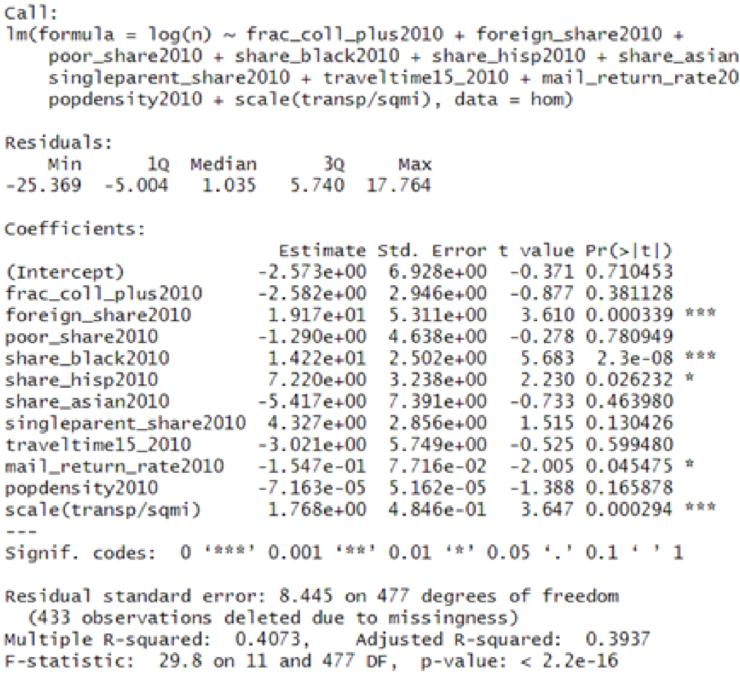
Zimring, Franklin E. 2006. *The Great American Crime Decline*. New York: Oxford University Press.

Appendix A



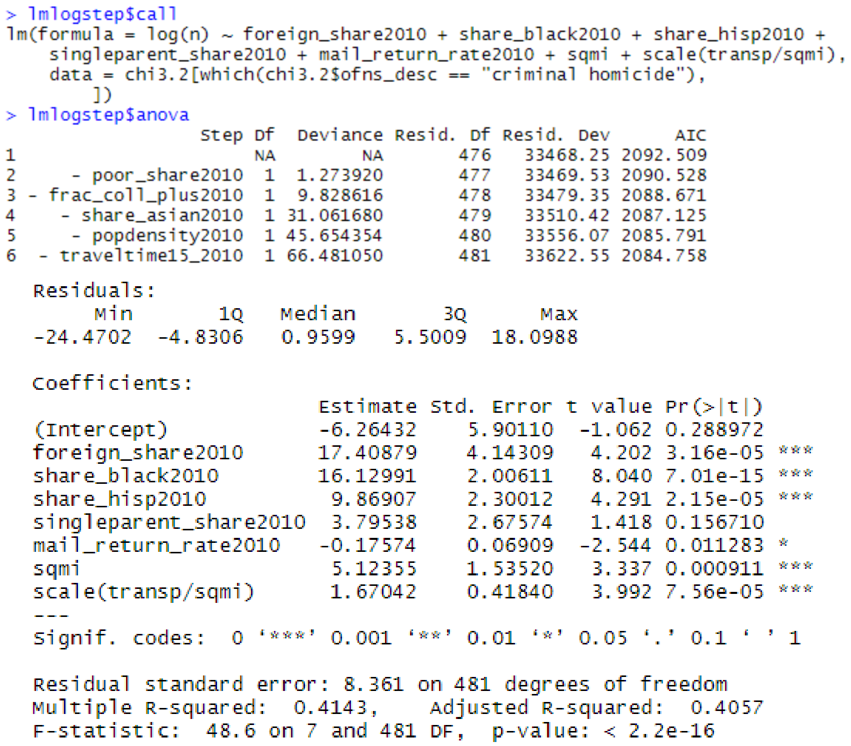
```
Call:
lm(formula = n ~ frac_coll_plus2010 + foreign_share2010 + poor_share2010 +
    share_black2010 + share_hisp2010 + share_asian2010 + singleparent_share2010 +
    traveltime15_2010 + mail_return_rate2010 + popdensity2010 +
    scale(transp/sqmi), data = census[which(census$ofns_desc ==
    "criminal homicide"), ])

Residuals:
    Min      1Q  Median      3Q     Max 
-13.263  -2.393  -0.356   1.744  54.950 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            1.369e+01  4.634e+00   2.953  0.00330 ** 
frac_coll_plus2010    -8.282e+00  1.971e+00  -4.203 3.15e-05 ***
foreign_share2010     -1.526e+00  3.553e+00  -0.429  0.66783    
poor_share2010        -2.785e+00  3.103e+00  -0.898  0.36978    
share_black2010        9.742e+00  1.674e+00   5.821 1.08e-08 ***
share_hisp2010        -1.223e+00  2.166e+00  -0.565  0.57268    
share_asian2010       -4.647e+00  4.944e+00  -0.940  0.34772    
singleparent_share2010  4.199e-01  1.910e+00   0.220  0.82612    
traveltime15_2010     -4.131e+00  3.846e+00  -1.074  0.28335    
mail_return_rate2010  -1.230e-01  5.162e-02  -2.383  0.01757 *  
popdensity2010         1.190e-04  3.453e-05   3.446  0.00062 ***
scale(transp/sqmi)    -4.318e-01  3.242e-01  -1.332  0.18354    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.65 on 477 degrees of freedom
  (433 observations deleted due to missingness)
Multiple R-squared:  0.4938,    Adjusted R-squared:  0.4821 
F-statistic: 42.31 on 11 and 477 DF,  p-value: < 2.2e-16
```

Table A1. Output of an ordinary least squares regression on the number of criminal homicides in a census tract on all other available variables.



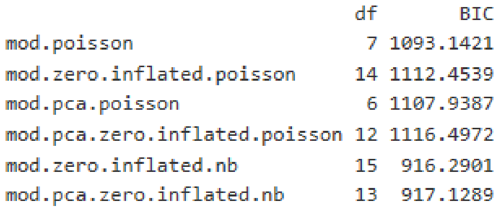
```
Call:
lm(formula = log(n) ~ frac_coll_plus2010 + foreign_share2010 +
    poor_share2010 + share_black2010 + share_hisp2010 + share_asian2010 +
    singleparent_share2010 + traveltime15_2010 + mail_return_rate2010 +
    popdensity2010 + scale(transp/sqmi), data = hom)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.369  -5.004   1.035   5.740  17.764 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -2.573e+00  6.928e+00  -0.371 0.710453    
frac_coll_plus2010    -2.582e+00  2.946e+00  -0.877 0.381128    
foreign_share2010      1.917e+01  5.311e+00   3.610 0.000339 ***
poor_share2010        -1.290e+00  4.638e+00  -0.278 0.780949    
share_black2010        1.422e+01  2.502e+00   5.683  2.3e-08 ***
share_hisp2010         7.220e+00  3.238e+00   2.230 0.026232 *  
share_asian2010       -5.417e+00  7.391e+00  -0.733 0.463980    
singleparent_share2010  4.327e+00  2.856e+00   1.515 0.130426    
traveltime15_2010     -3.021e+00  5.749e+00  -0.525 0.599480    
mail_return_rate2010  -1.547e-01  7.716e-02  -2.005 0.045475 *  
popdensity2010        -7.163e-05  5.162e-05  -1.388 0.165878    
scale(transp/sqmi)     1.768e+00  4.846e-01   3.647 0.000294 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.445 on 477 degrees of freedom
  (433 observations deleted due to missingness)
Multiple R-squared:  0.4073,    Adjusted R-squared:  0.3937 
F-statistic:  29.8 on 11 and 477 DF,  p-value: < 2.2e-16
```

Table A2. Output of a log-transformed least squares regression on the number of homicides in a census tract on all other available variables.



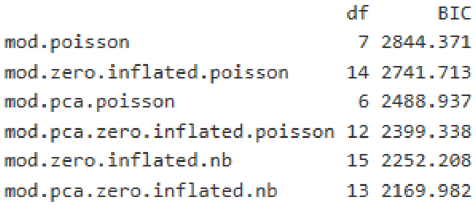
```
> lmlogstep$call
lm(formula = log(n) ~ foreign_share2010 + share_black2010 + share_hisp2010 +
    singleparent_share2010 + mail_return_rate2010 + sqmi + scale(transp/sqmi),
    data = chi3.2[which(chi3.2$ofns_desc == "criminal homicide"),
        ])
> lmlogstep$anova
                 Step Df   Deviance Resid. Df Resid. Dev      AIC
1                     NA         NA       476   33468.25 2092.509
2     - poor_share2010  1   1.273920       477   33469.53 2090.528
3 - frac_coll_plus2010  1   9.828616       478   33479.35 2088.671
4    - share_asian2010  1  31.061680       479   33510.42 2087.125
5     - popdensity2010  1  45.654354       480   33556.07 2085.791
6  - traveltime15_2010  1  66.481050       481   33622.55 2084.758

  Residuals:
       Min       1Q   Median       3Q      Max
  -24.4702  -4.8306   0.9599   5.5009  18.0988

  Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
  (Intercept)             -6.26432    5.90110  -1.062 0.288972
  foreign_share2010       17.40879    4.14309   4.202 3.16e-05 ***
  share_black2010         16.12991    2.00611   8.040 7.01e-15 ***
  share_hisp2010           9.86907    2.30012   4.291 2.15e-05 ***
  singleparent_share2010   3.79538    2.67574   1.418 0.156710
  mail_return_rate2010    -0.17574    0.06909  -2.544 0.011283 *
  sqmi                     5.12355    1.53520   3.337 0.000911 ***
  scale(transp/sqmi)       1.67042    0.41840   3.992 7.56e-05 ***
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

  Residual standard error: 8.361 on 481 degrees of freedom
  Multiple R-squared:  0.4143,    Adjusted R-squared:  0.4057
  F-statistic:  48.6 on 7 and 481 DF,  p-value: < 2.2e-16
```

Table A3. The output of a stepwise regression on the number of criminal homicides in a census tract on all available variables.



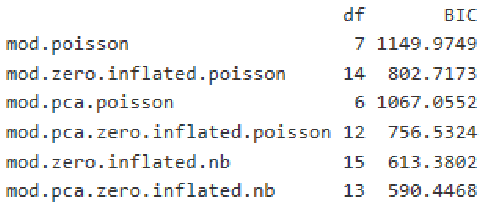
```
                             df       BIC
mod.poisson                   7 1093.1421
mod.zero.inflated.poisson    14 1112.4539
mod.pca.poisson               6 1107.9387
mod.pca.zero.inflated.poisson 12 1116.4972
mod.zero.inflated.nb         15  916.2901
mod.pca.zero.inflated.nb     13  917.1289
```

Table A4. The BIC of Poisson, zero-inflated Poisson, PCA-Poisson, zero-inflated PCA-Poisson, zero-inflated Negative Binomial, and zero-inflated PCA-Negative Binomial models for Baltimore, MD.



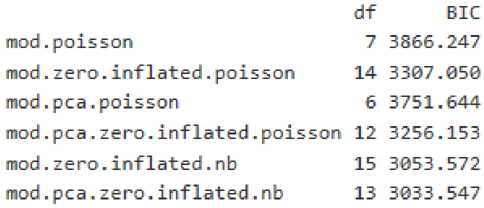
```
                             df      BIC
mod.poisson                   7 2844.371
mod.zero.inflated.poisson    14 2741.713
mod.pca.poisson               6 2488.937
mod.pca.zero.inflated.poisson 12 2399.338
mod.zero.inflated.nb         15 2252.208
mod.pca.zero.inflated.nb     13 2169.982
```

Table A5. The BIC of Poisson, zero-inflated Poisson, PCA-Poisson, zero-inflated PCA-Poisson, zero-inflated Negative Binomial, and zero-inflated PCA-Negative Binomial models for Chicago, IL.



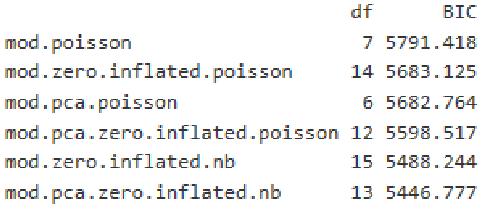
```
                               df       BIC
mod.poisson                     7 1149.9749
mod.zero.inflated.poisson      14  802.7173
mod.pca.poisson                 6 1067.0552
mod.pca.zero.inflated.poisson  12  756.5324
mod.zero.inflated.nb           15  613.3802
mod.pca.zero.inflated.nb       13  590.4468
```

Table A6. The BIC of Poisson, zero-inflated Poisson, PCA-Poisson, zero-inflated PCA-Poisson, zero-inflated Negative Binomial, and zero-inflated PCA-Negative Binomial models for Cincinnati, OH.



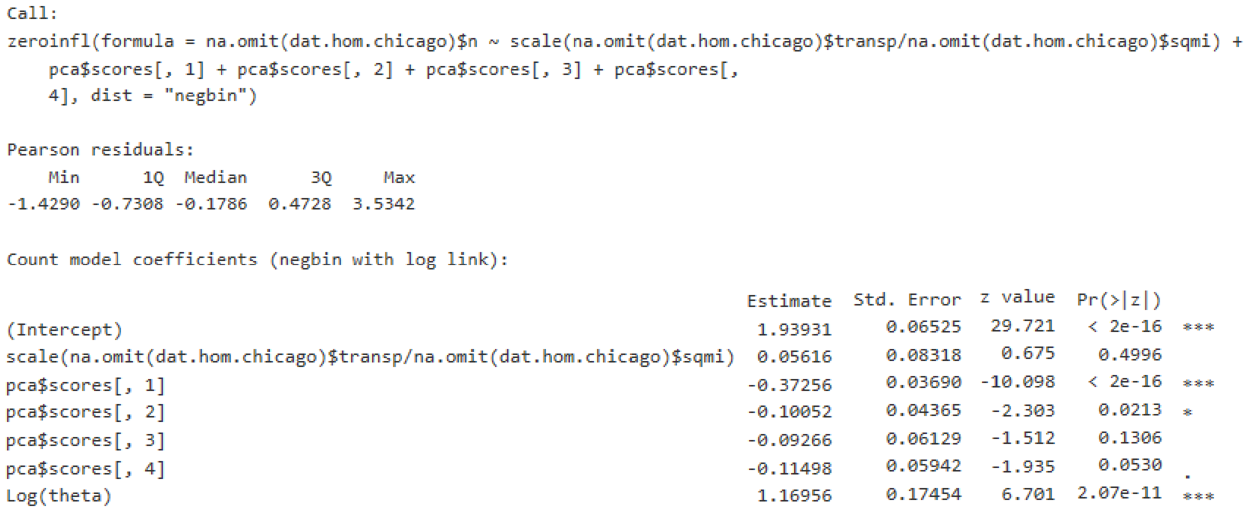
```
                              df      BIC
mod.poisson                    7 3866.247
mod.zero.inflated.poisson     14 3307.050
mod.pca.poisson                6 3751.644
mod.pca.zero.inflated.poisson 12 3256.153
mod.zero.inflated.nb          15 3053.572
mod.pca.zero.inflated.nb      13 3033.547
```

Table A7. The BIC of Poisson, zero-inflated Poisson, PCA-Poisson, zero-inflated PCA-Poisson, zero-inflated Negative Binomial, and zero-inflated PCA-Negative Binomial models for Los Angeles, CA.



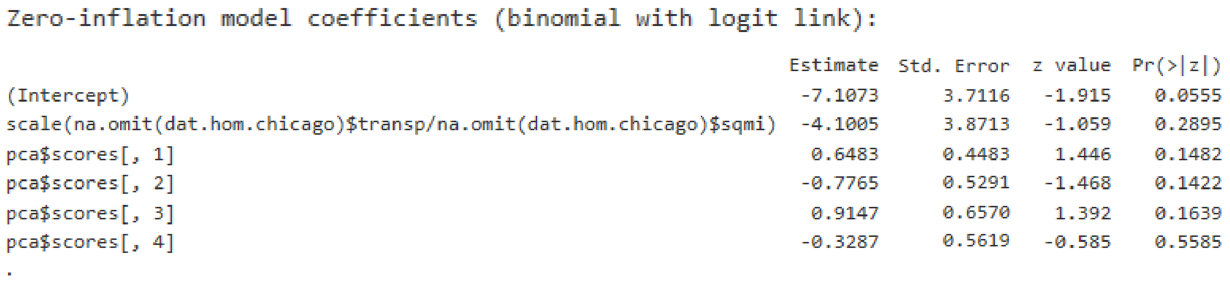
```
                              df      BIC
mod.poisson                    7 5791.418
mod.zero.inflated.poisson     14 5683.125
mod.pca.poisson                6 5682.764
mod.pca.zero.inflated.poisson 12 5598.517
mod.zero.inflated.nb          15 5488.244
mod.pca.zero.inflated.nb      13 5446.777
```

Table A8. The BIC of Poisson, zero-inflated Poisson, PCA-Poisson, zero-inflated PCA-Poisson, zero-inflated Negative Binomial, and zero-inflated PCA-Negative Binomial models for New York City, NY.

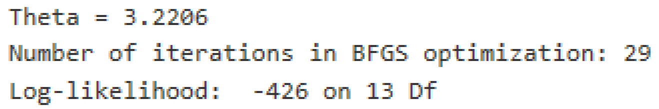


```
Call:
zeroinfl(formula = na.omit(dat.hom.chicago)$n ~ scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) + 
    pca$scores[, 1] + pca$scores[, 2] + pca$scores[, 3] + pca$scores[, 
    4], dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.4290 -0.7308 -0.1786  0.4728  3.5342 

Count model coefficients (negbin with log link):
                                                                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                                            1.93931    0.06525  29.721  < 2e-16 ***
scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi)  0.05616    0.08318   0.675   0.4996    
pca$scores[, 1]                                                       -0.37256    0.03690 -10.098  < 2e-16 ***
pca$scores[, 2]                                                       -0.10052    0.04365  -2.303   0.0213 *  
pca$scores[, 3]                                                       -0.09266    0.06129  -1.512   0.1306    
pca$scores[, 4]                                                       -0.11498    0.05942  -1.935   0.0530 .  
Log(theta)                                                             1.16956    0.17454   6.701 2.07e-11 ***
```

Table A9. The estimated coefficients for a zero-inflated negative binomial model, and their levels of significance, in Baltimore, MD.

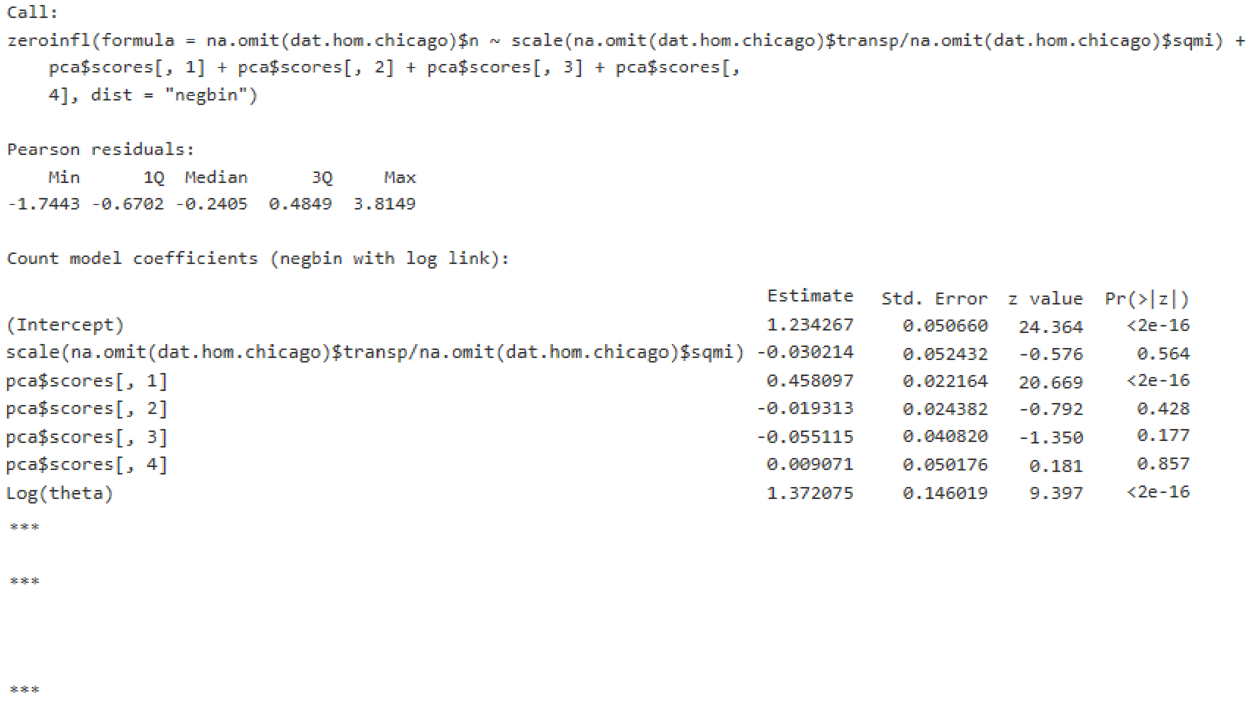


```
Zero-inflation model coefficients (binomial with logit link):
                                                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                             -7.1073     3.7116  -1.915   0.0555
scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi)  -4.1005     3.8713  -1.059   0.2895
pca$scores[, 1]                                                          0.6483     0.4483   1.446   0.1482
pca$scores[, 2]                                                         -0.7765     0.5291  -1.468   0.1422
pca$scores[, 3]                                                          0.9147     0.6570   1.392   0.1639
pca$scores[, 4]                                                         -0.3287     0.5619  -0.585   0.5585
.
```



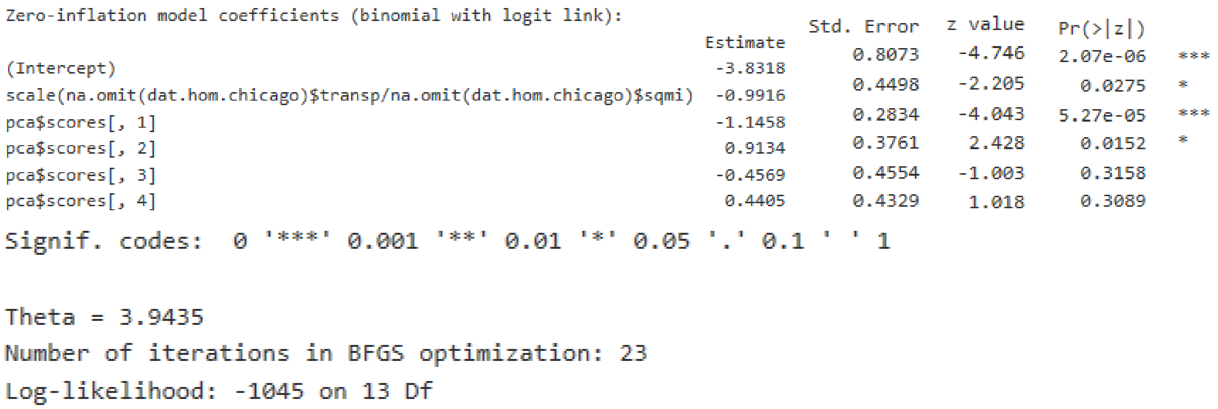
```
Theta = 3.2206
Number of iterations in BFGS optimization: 29
Log-likelihood:  -426 on 13 Df
```

Table A10. The estimated coefficients for a zero-inflated binomial model, and their levels of significance, in Baltimore, MD.



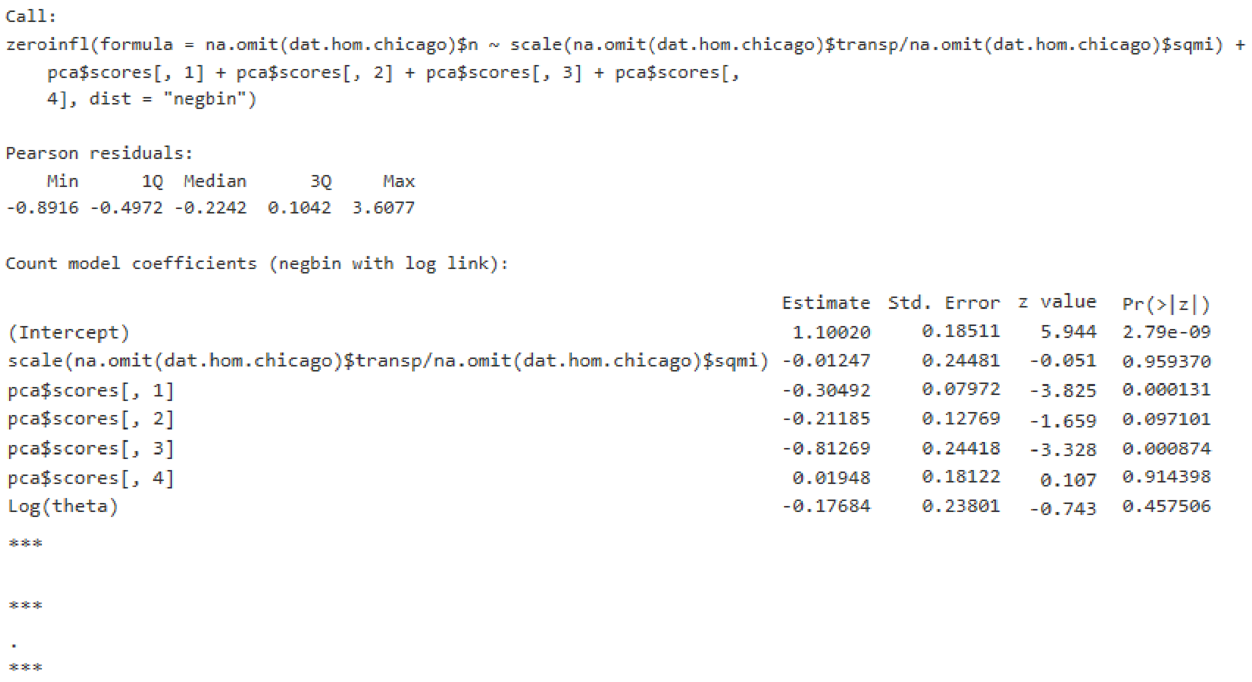
```
Call:
zeroinfl(formula = na.omit(dat.hom.chicago)$n ~ scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) +
    pca$scores[, 1] + pca$scores[, 2] + pca$scores[, 3] + pca$scores[,
    4], dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.7443 -0.6702 -0.2405  0.4849  3.8149

Count model coefficients (negbin with log link):
                                                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                           1.234267   0.050660  24.364   <2e-16
scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) -0.030214   0.052432  -0.576    0.564
pca$scores[, 1]                                                       0.458097   0.022164  20.669   <2e-16
pca$scores[, 2]                                                      -0.019313   0.024382  -0.792    0.428
pca$scores[, 3]                                                      -0.055115   0.040820  -1.350    0.177
pca$scores[, 4]                                                       0.009071   0.050176   0.181    0.857
Log(theta)                                                            1.372075   0.146019   9.397   <2e-16
***

***



***
```

Table A11. The estimated coefficients for a zero-inflated negative binomial model, and their levels of significance, in Chicago, IL.



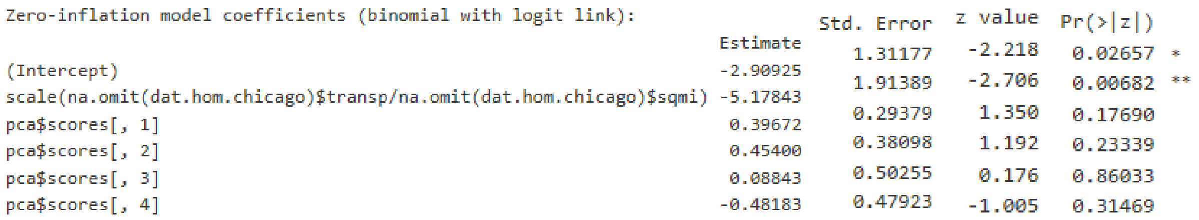
```
Zero-inflation model coefficients (binomial with logit link):
                                                                        Std. Error  z value   Pr(>|z|)
                                                              Estimate
(Intercept)                                                   -3.8318   0.8073    -4.746   2.07e-06  ***
scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi)  -0.9916   0.4498    -2.205    0.0275   *
pca$scores[, 1]                                               -1.1458   0.2834    -4.043   5.27e-05  ***
pca$scores[, 2]                                                0.9134   0.3761     2.428    0.0152   *
pca$scores[, 3]                                               -0.4569   0.4554    -1.003    0.3158
pca$scores[, 4]                                                0.4405   0.4329     1.018    0.3089
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 3.9435
Number of iterations in BFGS optimization: 23
Log-likelihood: -1045 on 13 Df
```

Table A12. The estimated coefficients for a zero-inflated binomial model, and their levels of significance, in Chicago, IL.

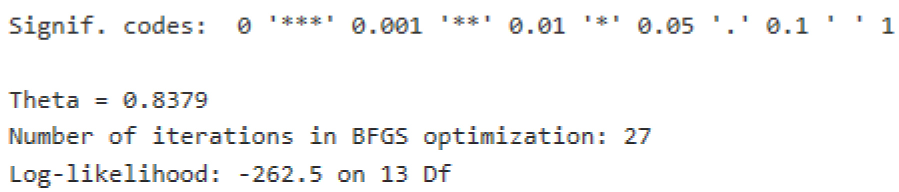


```
Call:
zeroinfl(formula = na.omit(dat.hom.chicago)$n ~ scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) +
    pca$scores[, 1] + pca$scores[, 2] + pca$scores[, 3] + pca$scores[,
    4], dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.8916 -0.4972 -0.2242  0.1042  3.6077

Count model coefficients (negbin with log link):

                                                                      Estimate Std. Error z value  Pr(>|z|)
(Intercept)                                                           1.10020    0.18511   5.944  2.79e-09
scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) -0.01247    0.24481  -0.051  0.959370
pca$scores[, 1]                                                       -0.30492    0.07972  -3.825  0.000131
pca$scores[, 2]                                                       -0.21185    0.12769  -1.659  0.097101
pca$scores[, 3]                                                       -0.81269    0.24418  -3.328  0.000874
pca$scores[, 4]                                                        0.01948    0.18122   0.107  0.914398
Log(theta)                                                            -0.17684    0.23801  -0.743  0.457506
***


***
.
***
```

Table A13. The estimated coefficients for a zero-inflated negative binomial model, and their levels of significance, in Cincinnati, OH.

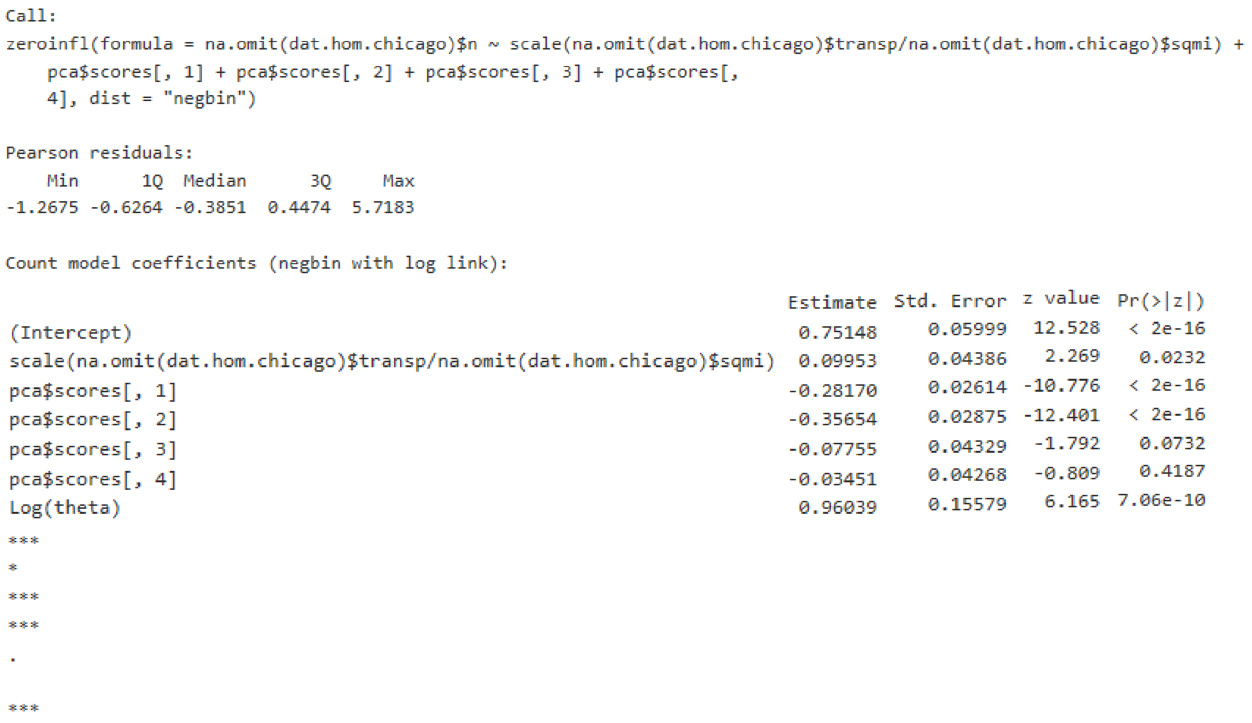


```
Zero-inflation model coefficients (binomial with logit link):
                                                                          Std. Error  z value  Pr(>|z|)
                                                              Estimate
(Intercept)                                                   -2.90925    1.31177   -2.218   0.02657 *
scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) -5.17843    1.91389   -2.706   0.00682 **
pca$scores[, 1]                                                0.39672    0.29379    1.350   0.17690
pca$scores[, 2]                                                0.45400    0.38098    1.192   0.23339
pca$scores[, 3]                                                0.08843    0.50255    0.176   0.86033
pca$scores[, 4]                                               -0.48183    0.47923   -1.005   0.31469
```



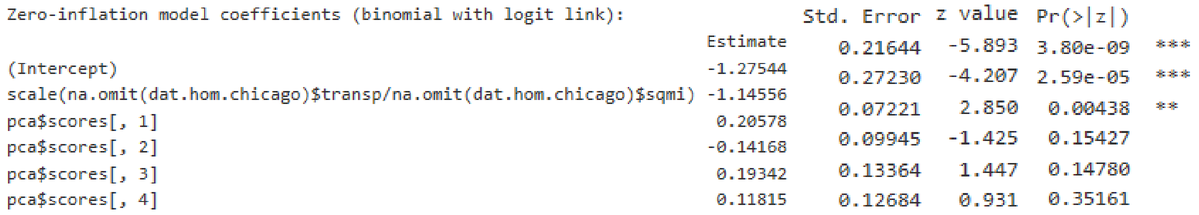
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 0.8379
Number of iterations in BFGS optimization: 27
Log-likelihood: -262.5 on 13 Df
```

Table A14. The estimated coefficients for a zero-inflated binomial model, and their levels of significance, in Cincinnati, OH.

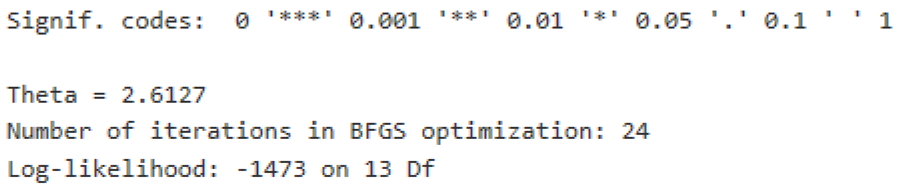


```
Call:
zeroinfl(formula = na.omit(dat.hom.chicago)$n ~ scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) +
    pca$scores[, 1] + pca$scores[, 2] + pca$scores[, 3] + pca$scores[,
    4], dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.2675 -0.6264 -0.3851  0.4474  5.7183

Count model coefficients (negbin with log link):
                                                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                              0.75148    0.05999  12.528  < 2e-16
scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi)  0.09953    0.04386   2.269   0.0232
pca$scores[, 1]                                                         -0.28170    0.02614 -10.776  < 2e-16
pca$scores[, 2]                                                         -0.35654    0.02875 -12.401  < 2e-16
pca$scores[, 3]                                                         -0.07755    0.04329  -1.792   0.0732
pca$scores[, 4]                                                         -0.03451    0.04268  -0.809   0.4187
Log(theta)                                                               0.96039    0.15579   6.165 7.06e-10
***
*
***
***
.

***
```

Table A15. The estimated coefficients for a zero-inflated negative binomial model, and their levels of significance, in Los Angeles, CA.

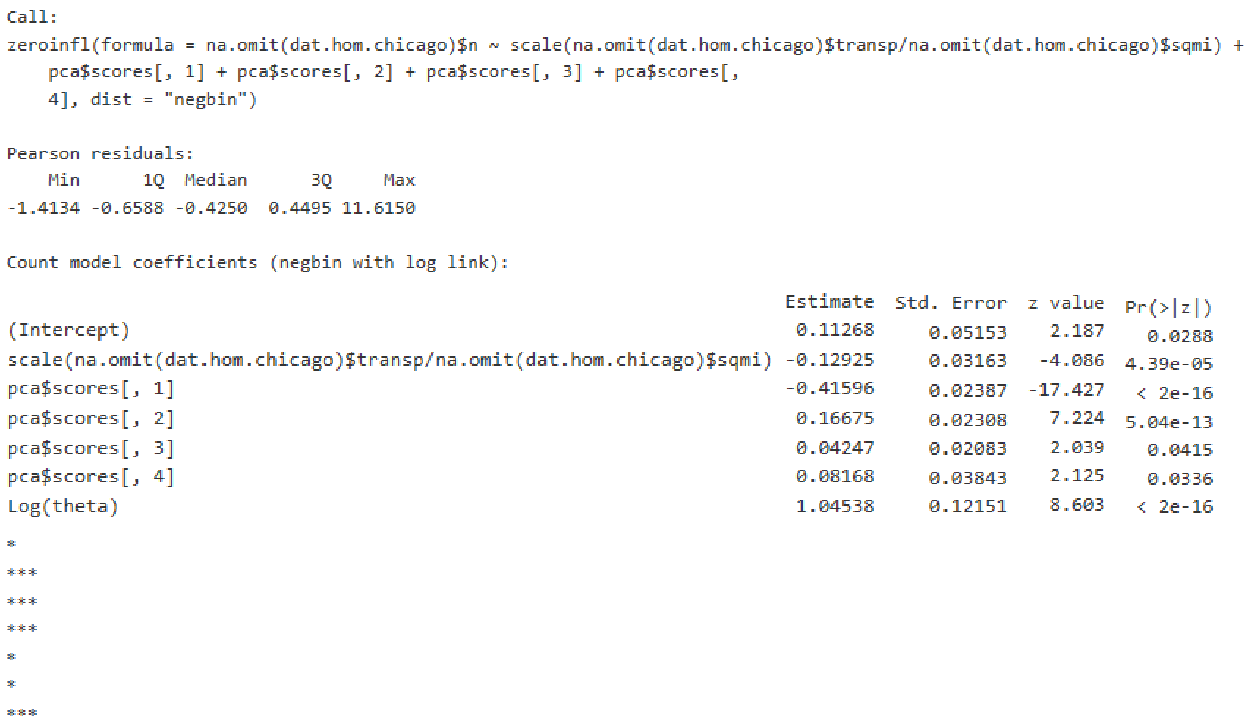


```
Zero-inflation model coefficients (binomial with logit link):
                                                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                           -1.27544    0.21644  -5.893 3.80e-09 ***
scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) -1.14556    0.27230  -4.207 2.59e-05 ***
pca$scores[, 1]                                                        0.20578    0.07221   2.850  0.00438 **
pca$scores[, 2]                                                       -0.14168    0.09945  -1.425  0.15427
pca$scores[, 3]                                                        0.19342    0.13364   1.447  0.14780
pca$scores[, 4]                                                        0.11815    0.12684   0.931  0.35161
```



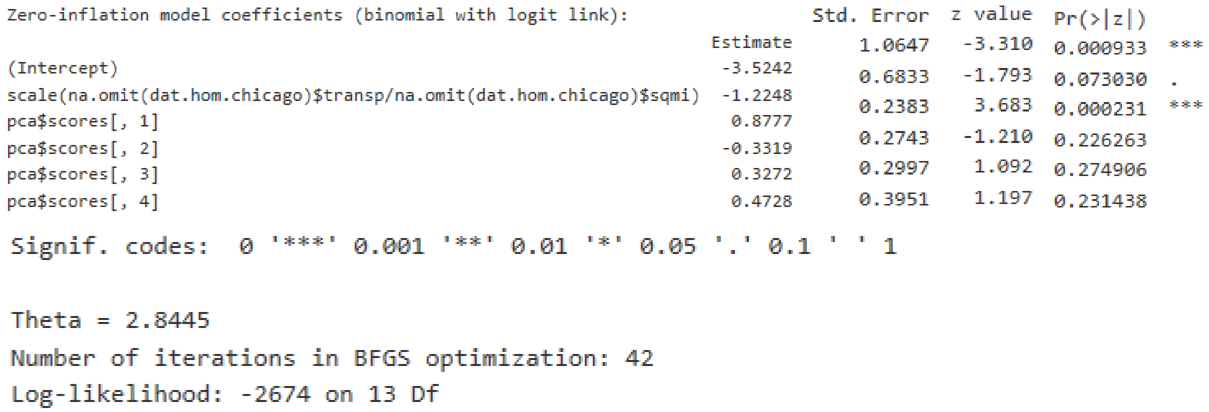
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 2.6127
Number of iterations in BFGS optimization: 24
Log-likelihood: -1473 on 13 Df
```

Table A16. The estimated coefficients for a zero-inflated binomial model, and their levels of significance, in Los Angeles, CA.



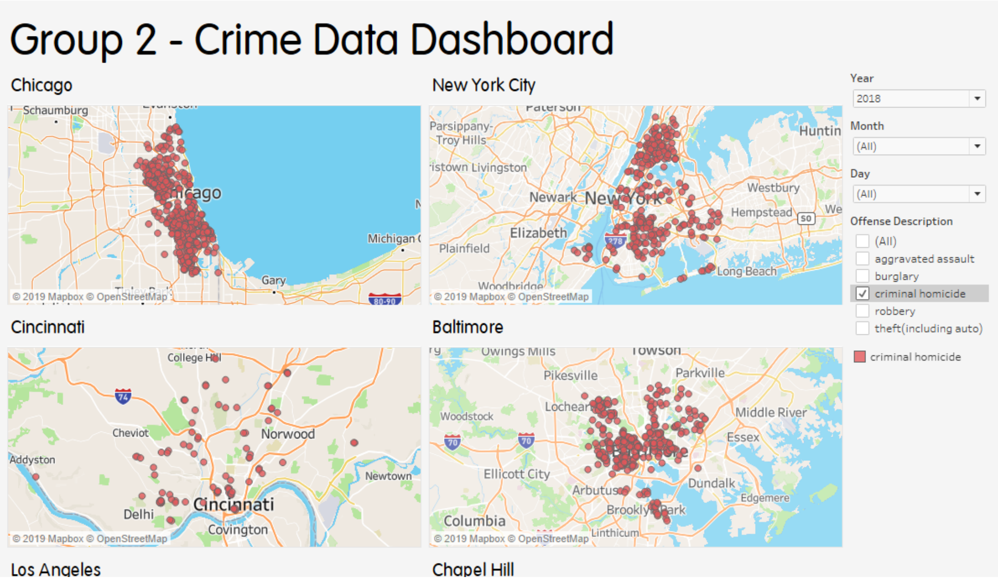
```
Call:
zeroinfl(formula = na.omit(dat.hom.chicago)$n ~ scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) +
    pca$scores[, 1] + pca$scores[, 2] + pca$scores[, 3] + pca$scores[,
    4], dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.4134 -0.6588 -0.4250  0.4495 11.6150

Count model coefficients (negbin with log link):

                                                                         Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)                                                               0.11268     0.05153    2.187    0.0288
scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) -0.12925     0.03163   -4.086  4.39e-05
pca$scores[, 1]                                                          -0.41596     0.02387  -17.427   < 2e-16
pca$scores[, 2]                                                           0.16675     0.02308    7.224  5.04e-13
pca$scores[, 3]                                                           0.04247     0.02083    2.039    0.0415
pca$scores[, 4]                                                           0.08168     0.03843    2.125    0.0336
Log(theta)                                                                1.04538     0.12151    8.603   < 2e-16

*
***
***
***
*
*
***
```

Table A17. The estimated coefficients for a zero-inflated negative binomial model, and their levels of significance, in New York City, NY.



```
Zero-inflation model coefficients (binomial with logit link):                     Std. Error  z value  Pr(>|z|)
                                                                   Estimate      1.0647   -3.310  0.000933 ***
(Intercept)                                                         -3.5242      0.6833   -1.793  0.073030 .
scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi) -1.2248     0.2383    3.683  0.000231 ***
pca$scores[, 1]                                                      0.8777      0.2743   -1.210  0.226263
pca$scores[, 2]                                                     -0.3319      0.2997    1.092  0.274906
pca$scores[, 3]                                                      0.3272      0.3951    1.197  0.231438
pca$scores[, 4]                                                      0.4728
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 2.8445
Number of iterations in BFGS optimization: 42
Log-likelihood: -2674 on 13 Df
```

Table A18. The estimated coefficients for a zero-inflated binomial model, and their levels of significance, in New York City, NY.



Group 2 - Crime Data Dashboard

Chicago

New York City

Cincinnati

Baltimore

Los Angeles

Chapel Hill

Year: 2018

Month: (All)

Day: (All)

Offense Description

- ☐ (All)
- ☐ aggravated assault
- ☐ burglary
- ☑ criminal homicide
- ☐ robbery
- ☐ theft(including auto)

🟥 criminal homicide

Figure A1. Interactive Tableau dashboard visualizing the raw crime data through filters.

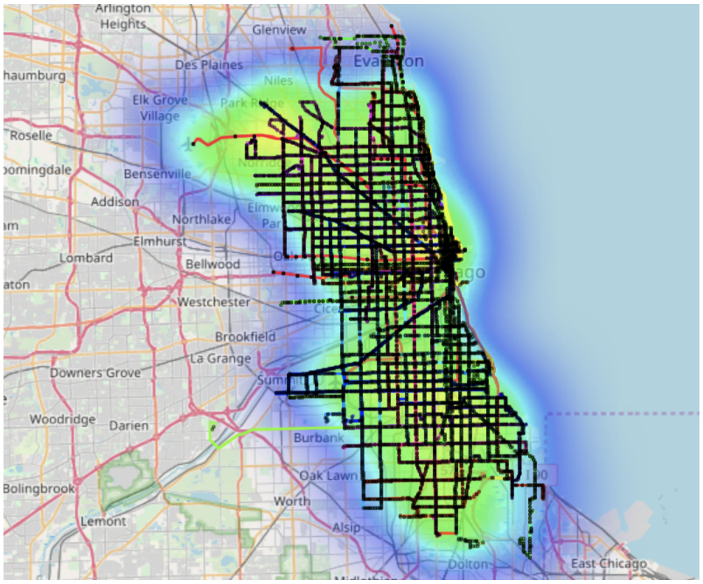


Figure A2. Heatmap of crime overlaid on the main public transportation system in Chicago, IL.

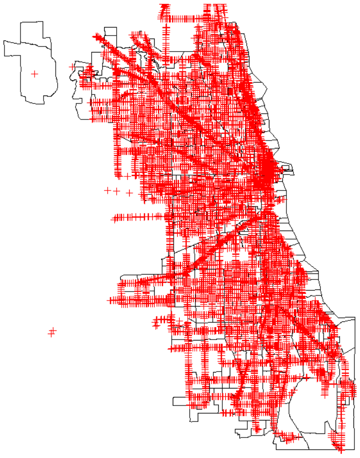


Figure A3. Map of Chicago, Illinois's bus and heavy rail public transportation routes.

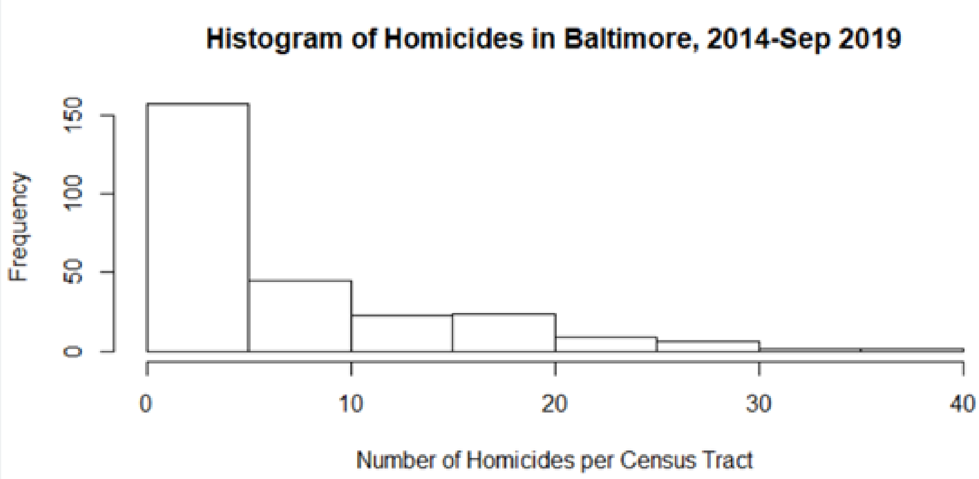


Figure A4. Histogram of criminal homicides by census tract in Baltimore, MD.

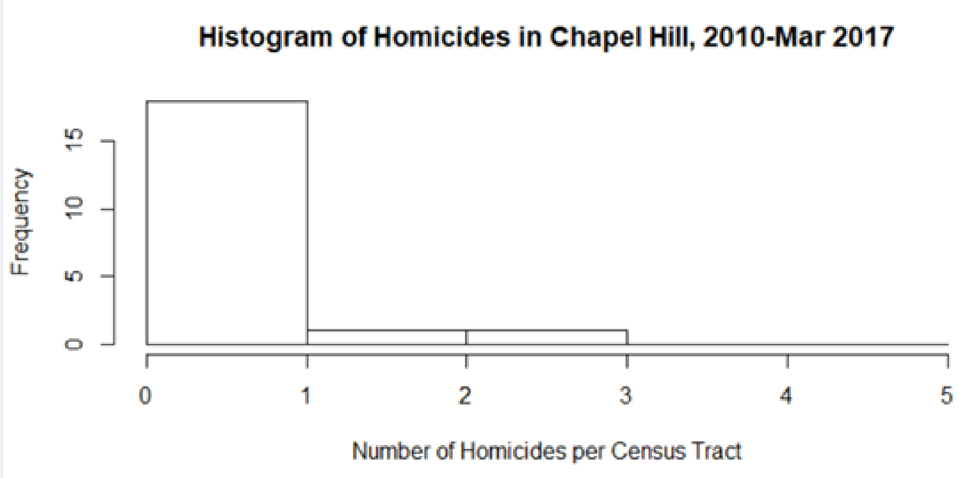


Figure A5. Histogram of criminal homicides by census tract in Chapel Hill, NC.

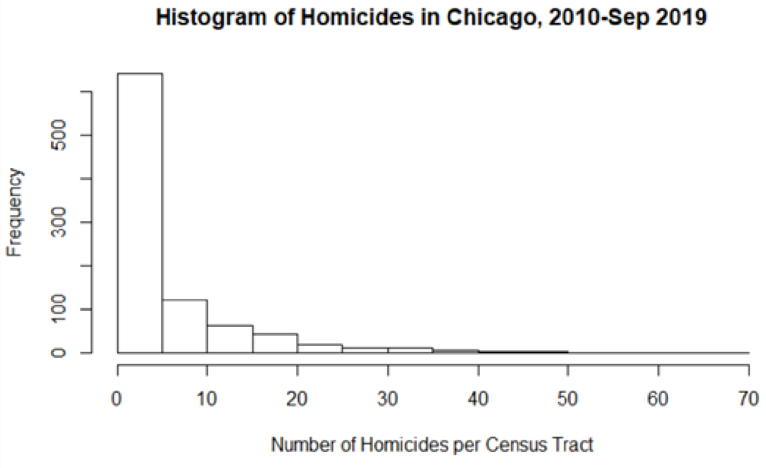


Figure A6. Histogram of criminal homicides by census tract in Chicago, IL.

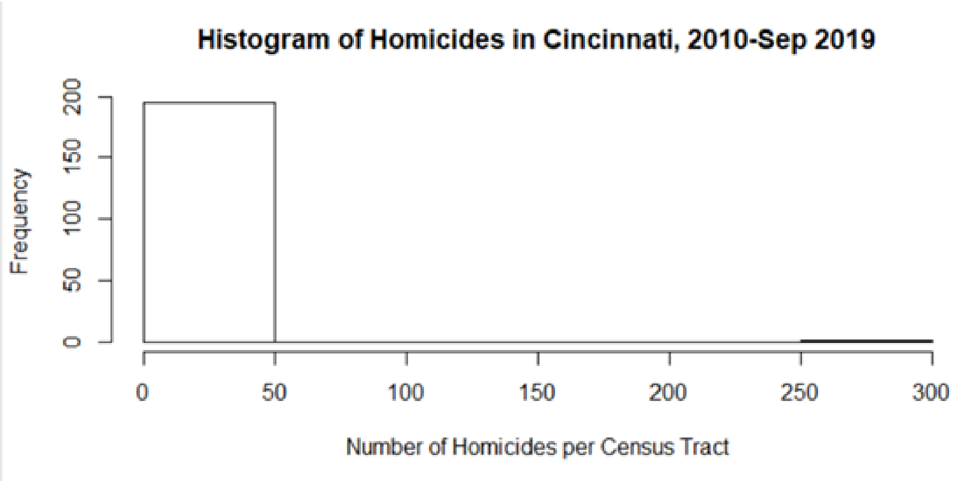


Figure A7. Histogram of criminal homicides by census tract in Cincinnati, OH.

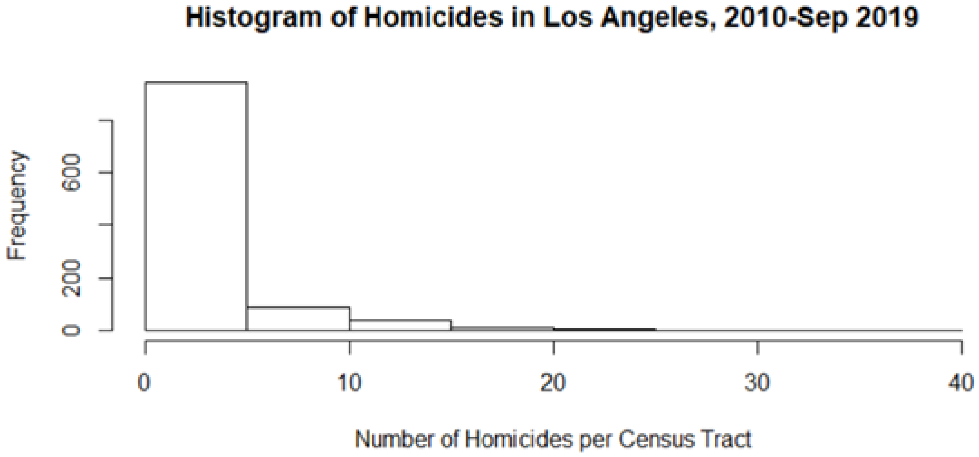


Figure A8. Histogram of criminal homicides by census tract in Los Angeles, CA.

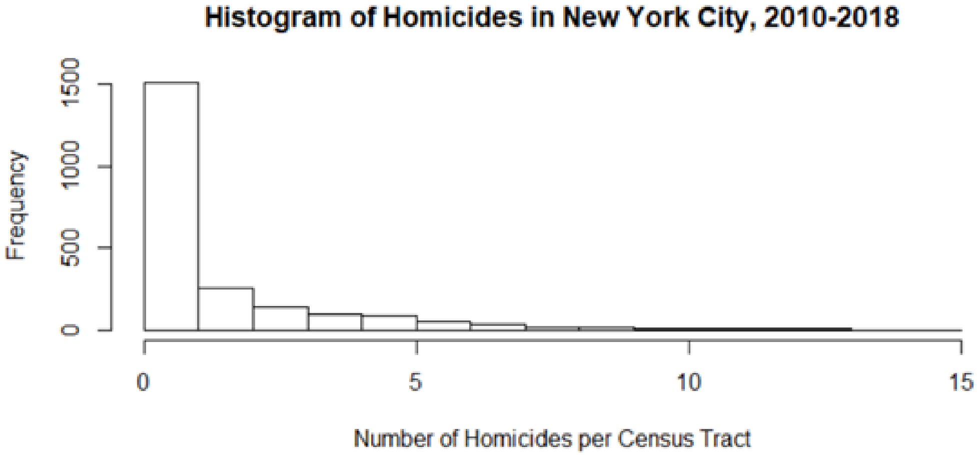


Figure A9. Histogram of criminal homicides by census tract in New York City, NY.

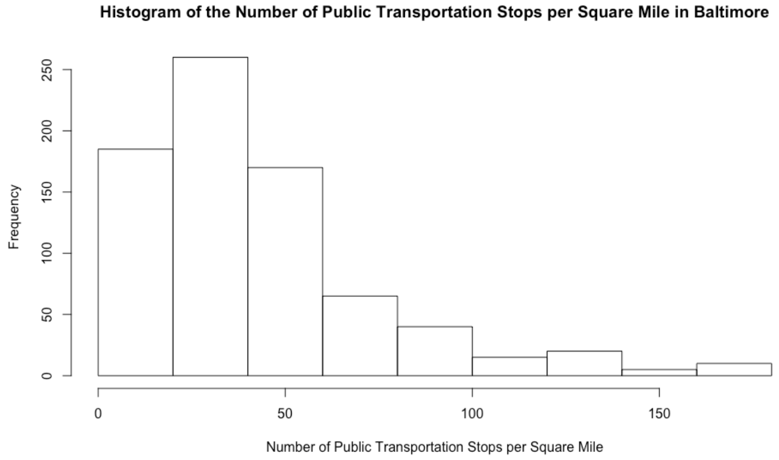


Figure A10. Histogram of public transportation stops by census tract density in Baltimore, MD.

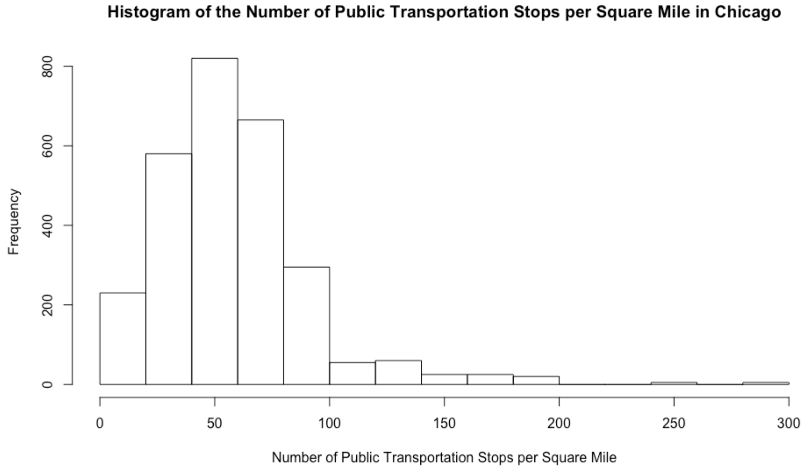


Figure A11. Histogram of public transportation stops by census tract density in Chicago, IL.

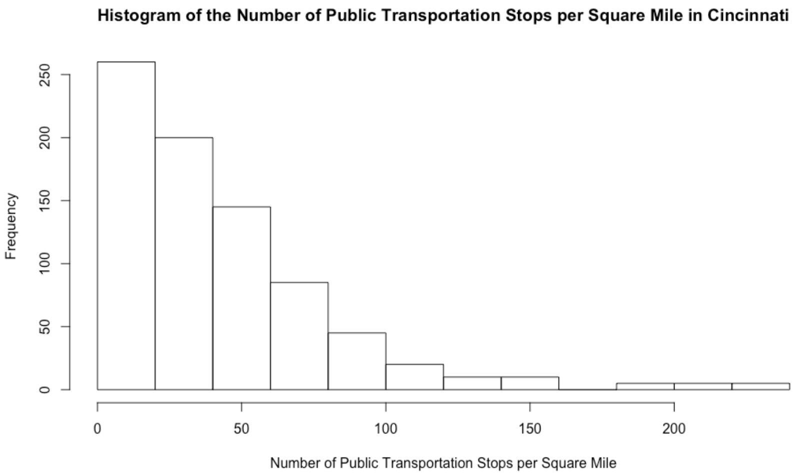


Figure A12. Histogram of public transportation stops by census tract density in Cincinnati, OH.

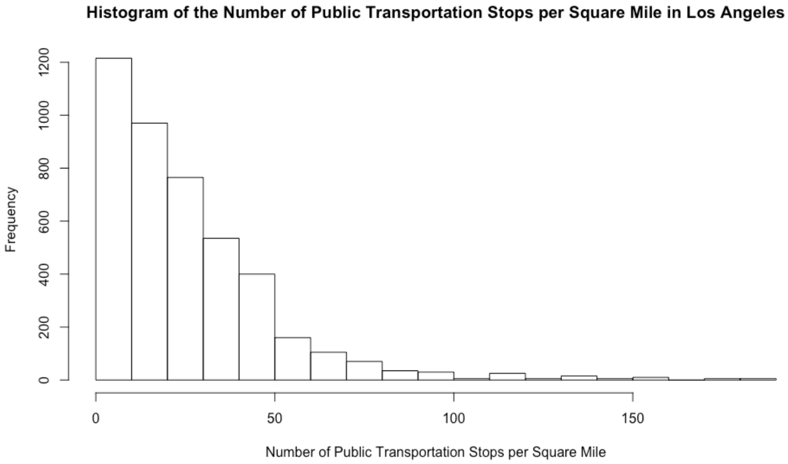


Figure A13. Histogram of public transportation stops by census tract density in Los Angeles, CA.

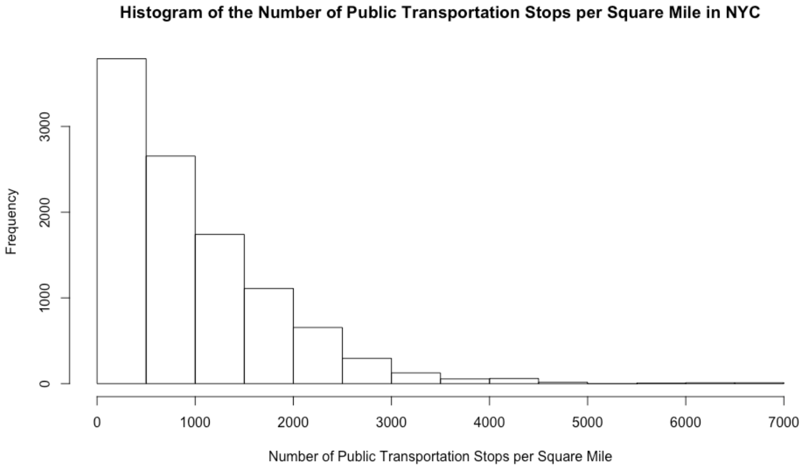


Figure A14. Histogram of public transportation stops by census tract density in New York City, NY.

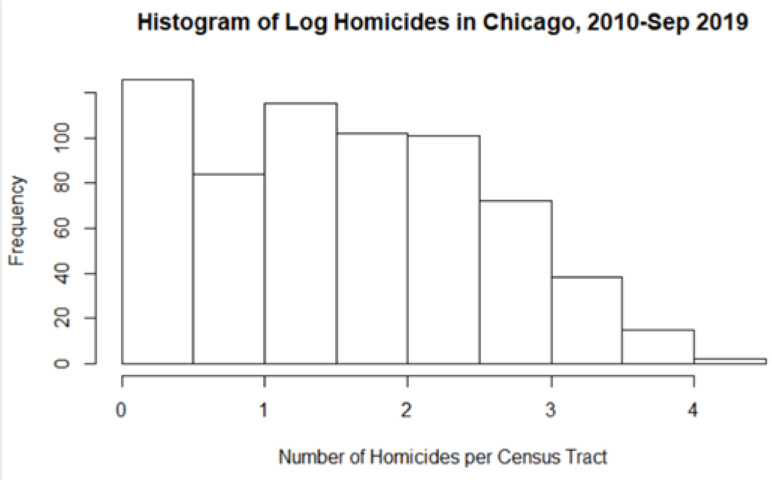


Figure A15. Histogram of the log of criminal homicides per census tract in Chicago, IL.

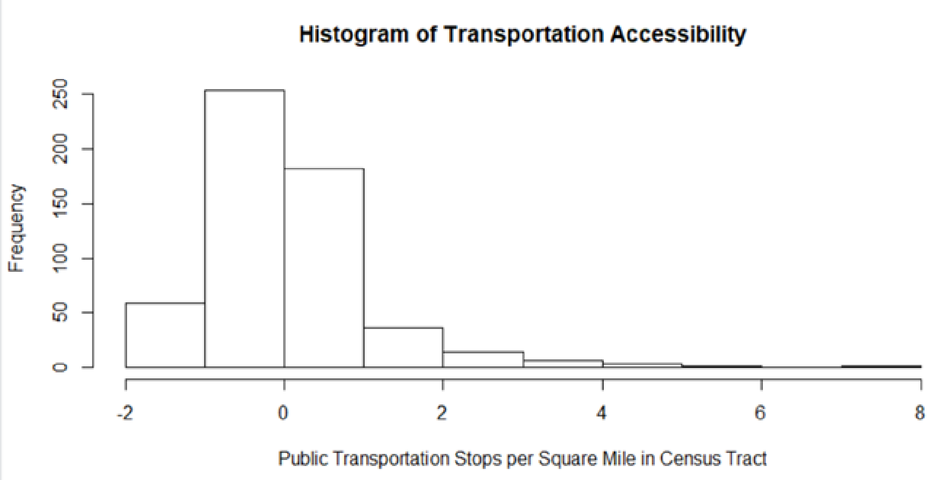


Figure A16. Histogram of standardized public transportation stops by census tract in Chicago, IL.

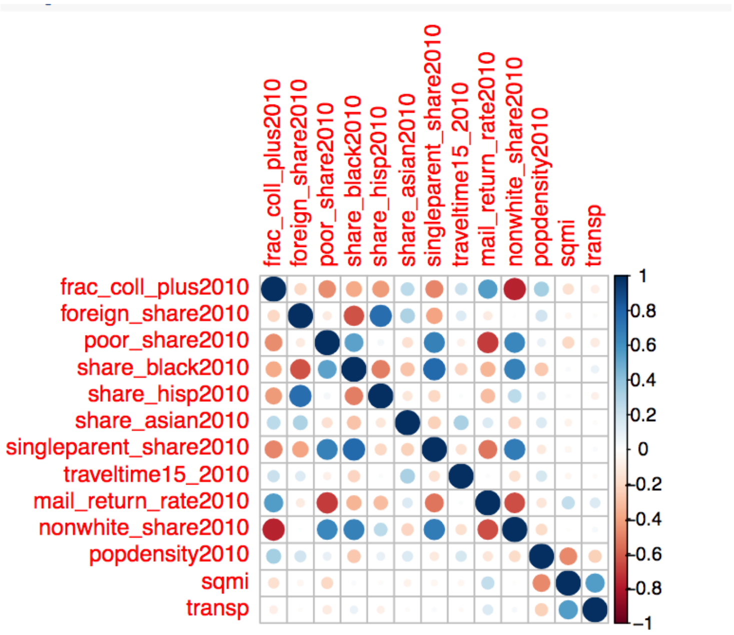


Figure A17. Correlation plot. Stronger colors and larger circles indicate stronger correlation. Reds represent a negative correlation, while blues represent a positive correlation.

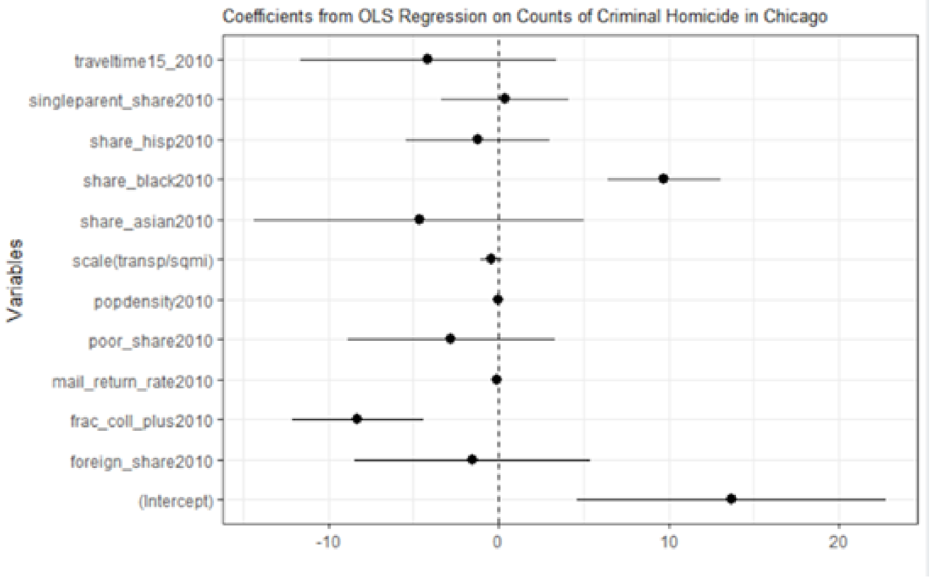


Figure A8. Violin plot of the distribution of SLG in each league

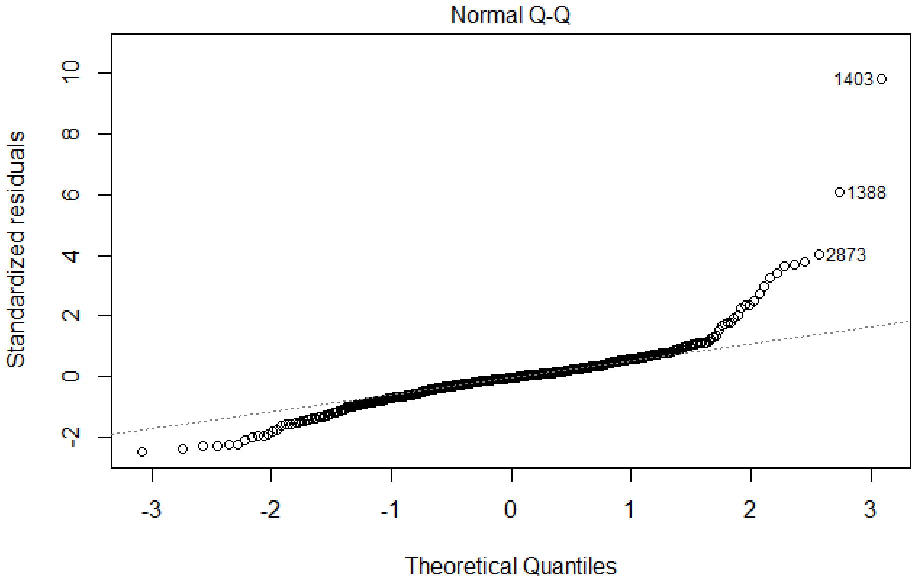


Figure A18. Q-Q plot for ordinary least squares regression.

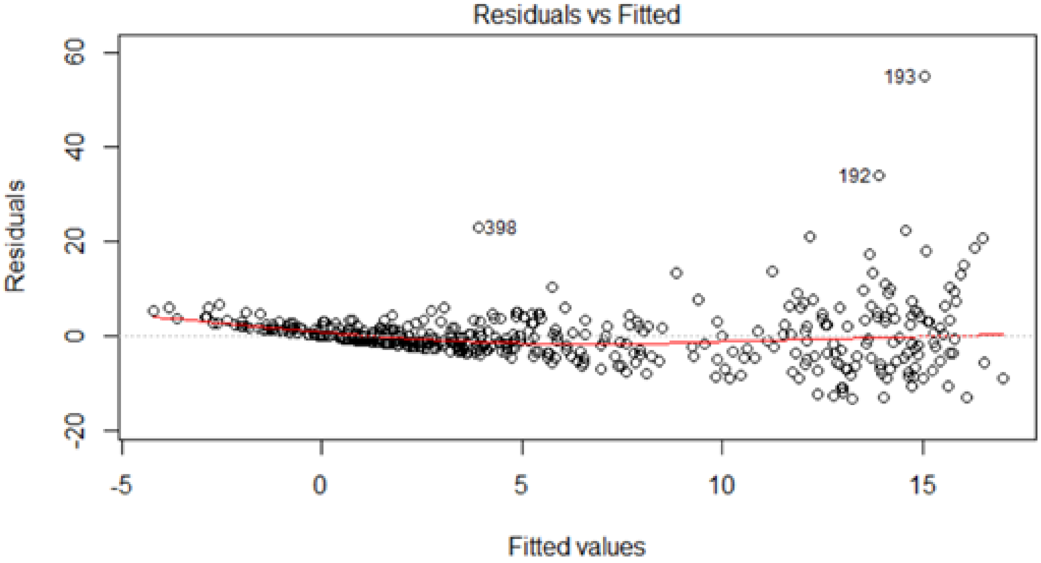


Figure A19. Residuals versus fitted values for ordinary least squares regression.

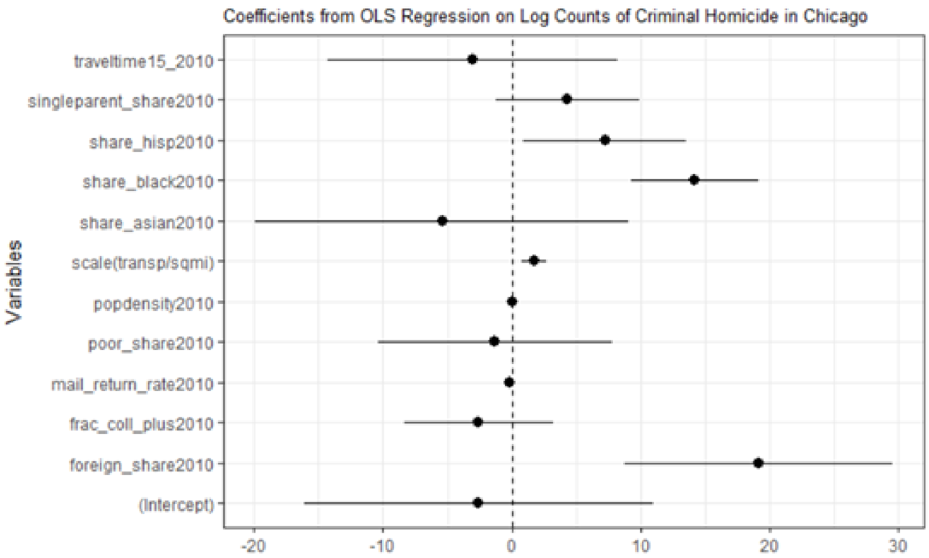


Figure A20. Significance plot for all variable coefficient estimates in ordinary least squares regression.

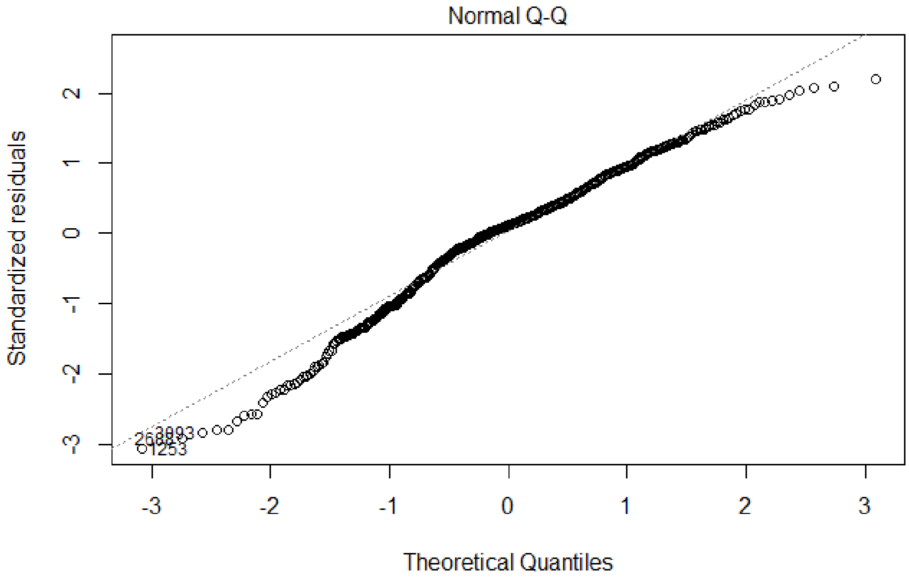


Figure A21. Q-Q plot for log-transformed least squares regression.

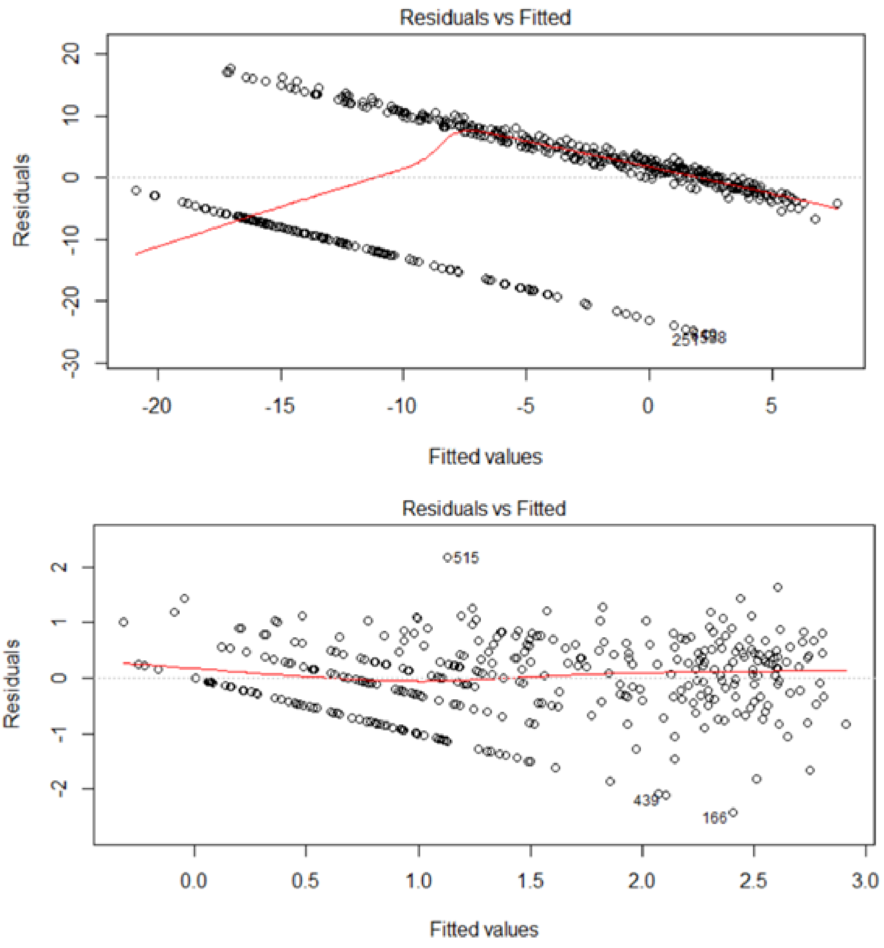


Figure A22. Residuals versus fitted values for log-transformed least squares regression; the bottom plot includes only those census tracts for which there was at least one criminal homicide.

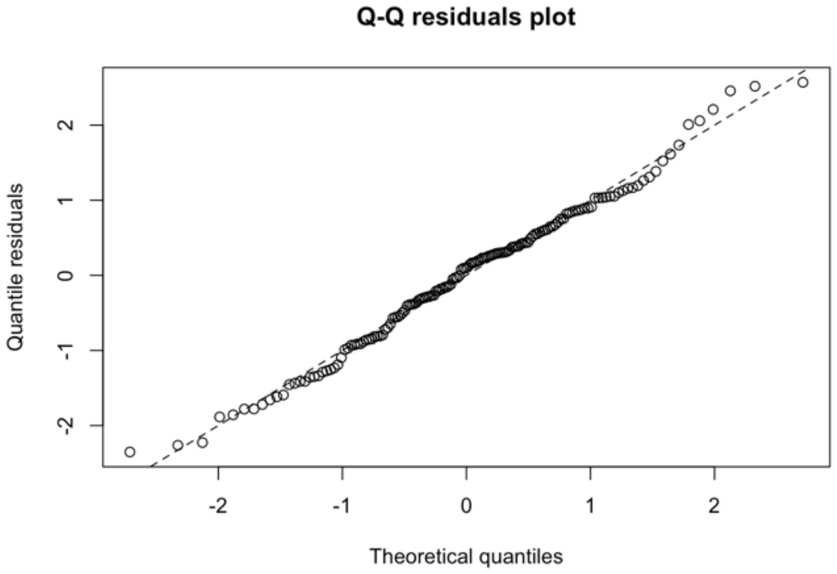


Figure A23. Q-Q plot for zero-inflated negative binomial model for Baltimore, MD.

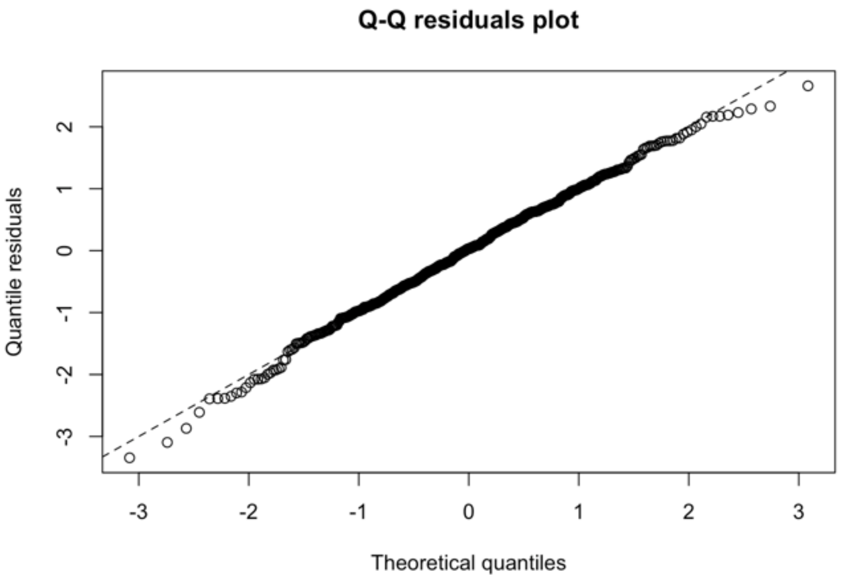


Figure A24. Q-Q plot for zero-inflated negative binomial model for Chicago, IL.

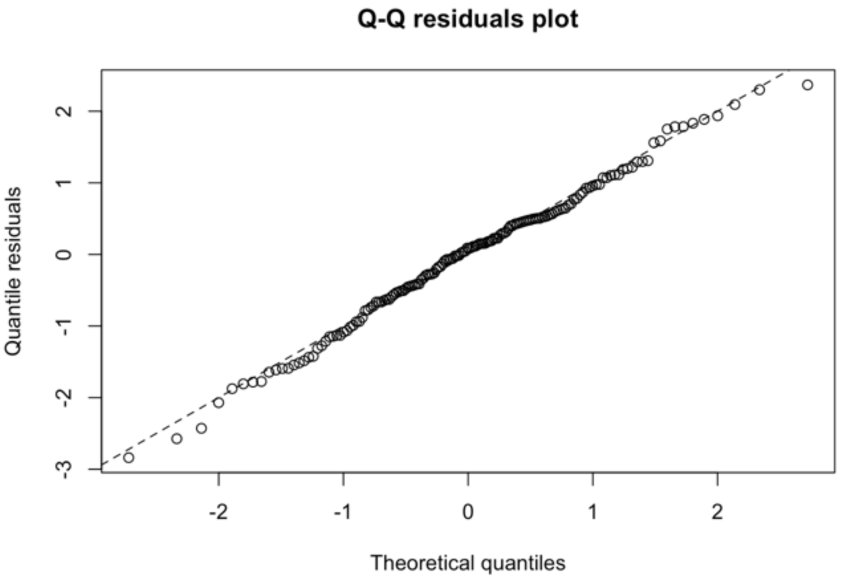


Figure A25. Q-Q plot for zero-inflated negative binomial model for Cincinnati, OH.

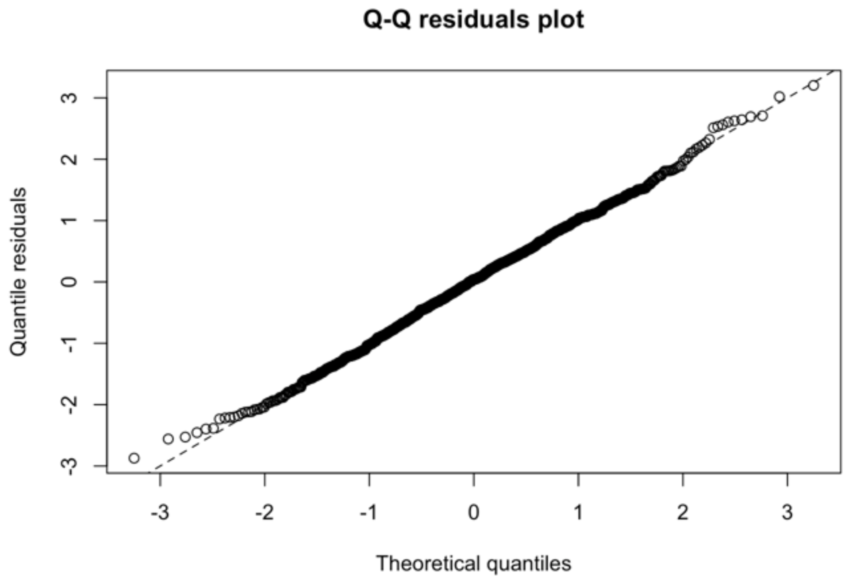


Figure A26. Q-Q plot for zero-inflated negative binomial model for Los Angeles, CA.

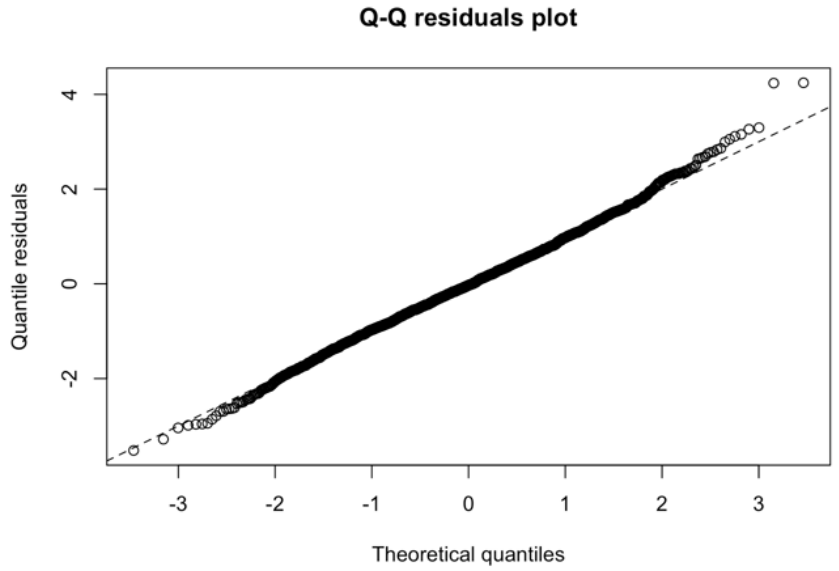


Figure A27. Q-Q plot for zero-inflated negative binomial model for New York City, NY.

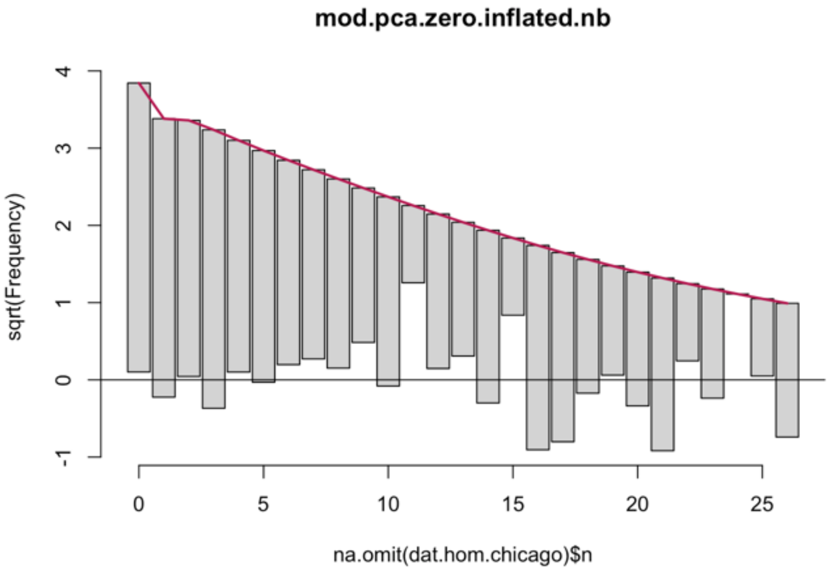


Figure A28. Output from Principal Components selection for Baltimore, MD.

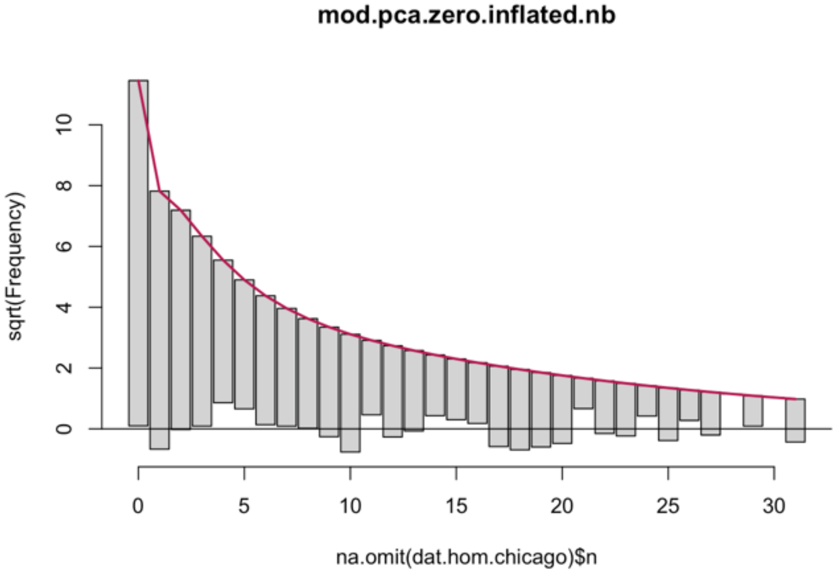


Figure A29. Output from Principal Components selection for Chicago, IL.

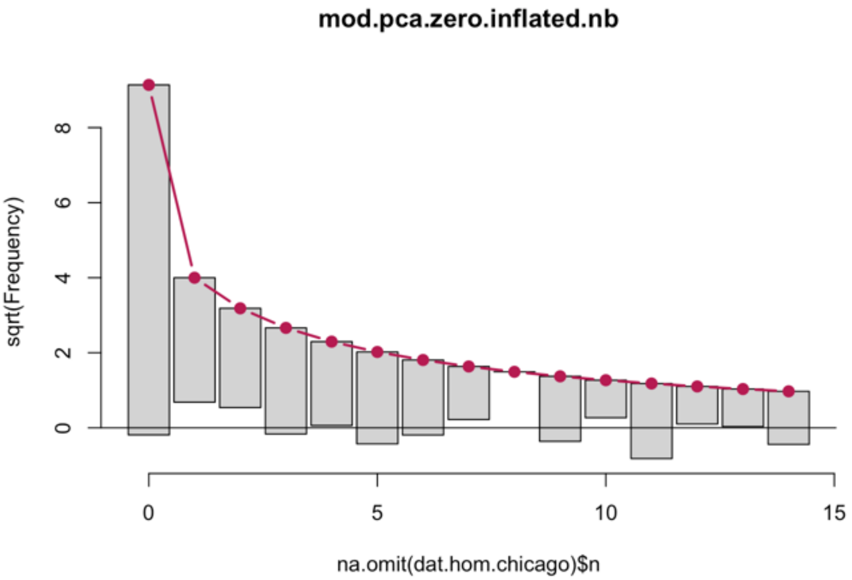


Figure A30. Output from Principal Components selection for Cincinnati, OH.

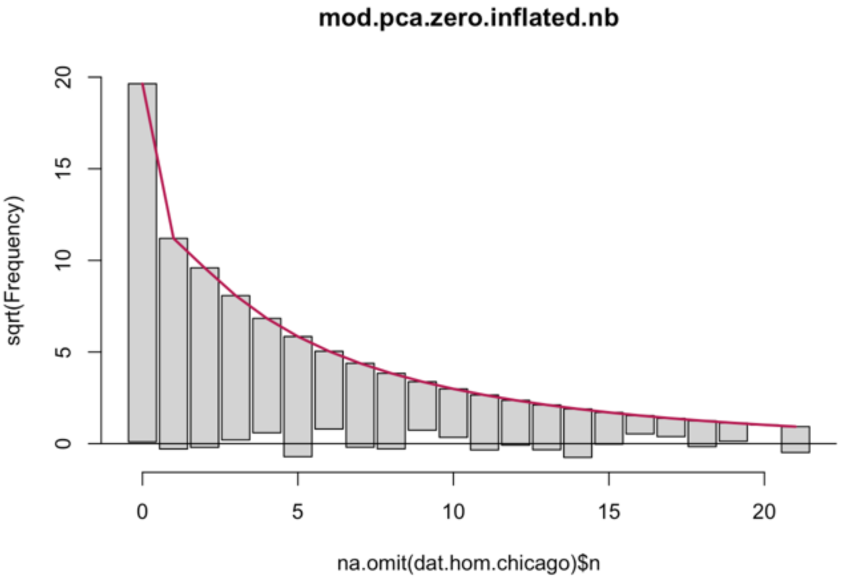


Figure A31. Output from Principal Components selection for Los Angeles, CA.

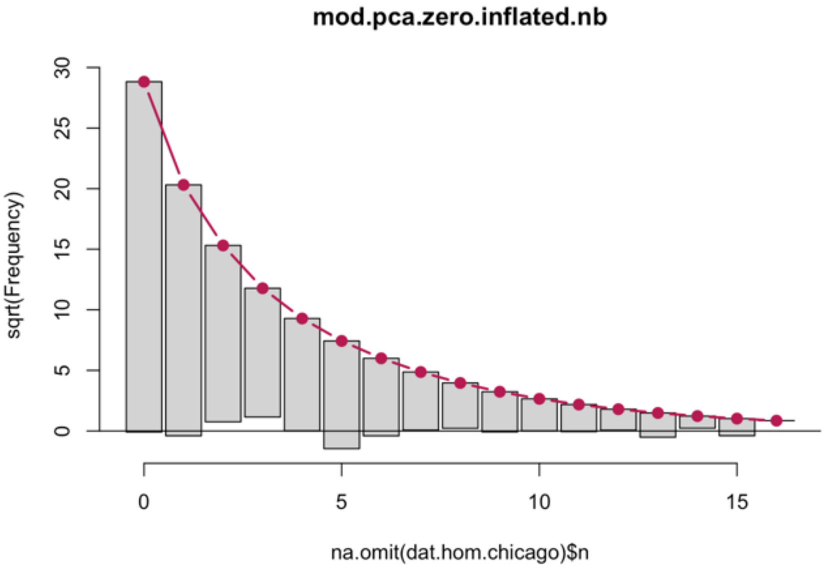
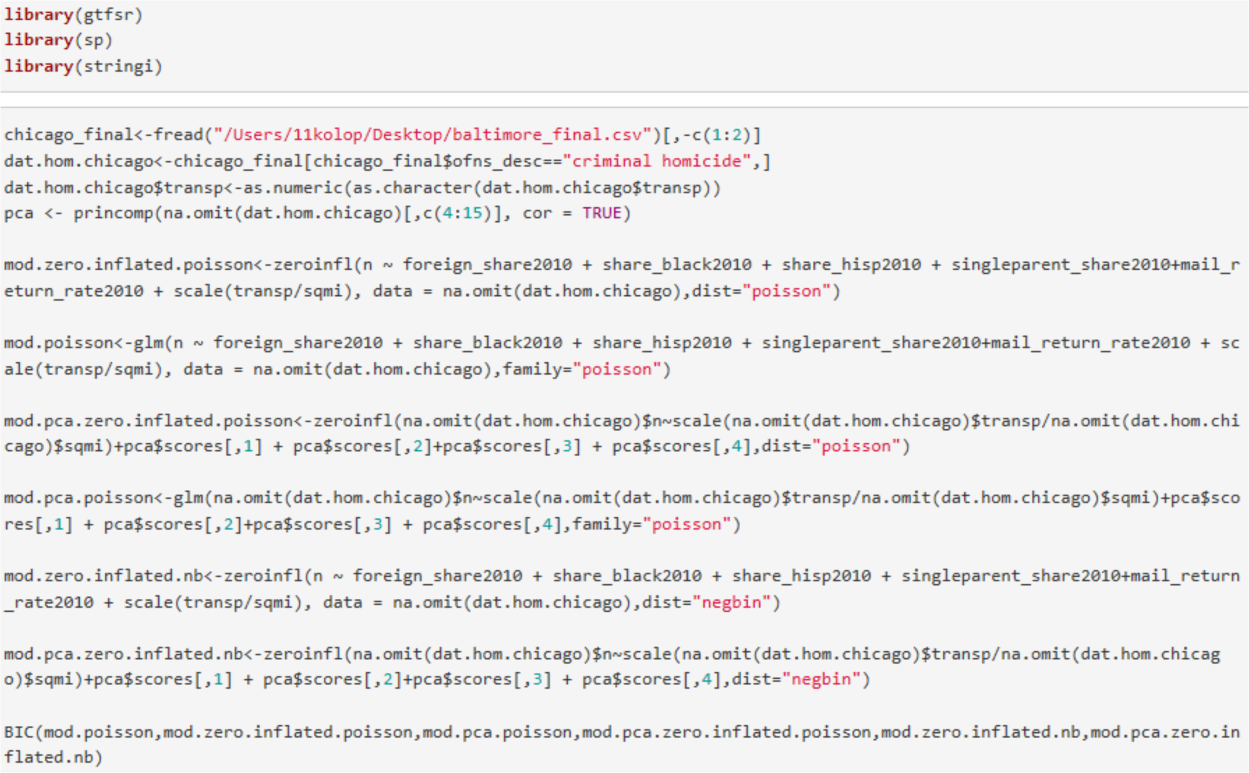


Figure A28. Output from Principal Components selection for New York City, NY.

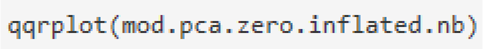
Appendix B



```
library(gtfsr)
library(sp)
library(stringi)
```

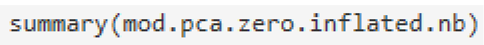
```
chicago_final<-fread("/Users/11kolop/Desktop/baltimore_final.csv")[,-c(1:2)]
dat.hom.chicago<-chicago_final[chicago_final$ofns_desc=="criminal homicide",]
dat.hom.chicago$transp<-as.numeric(as.character(dat.hom.chicago$transp))
pca <- princomp(na.omit(dat.hom.chicago)[,c(4:15)], cor = TRUE)

mod.zero.inflated.poisson<-zeroinfl(n ~ foreign_share2010 + share_black2010 + share_hisp2010 + singleparent_share2010+mail_r
eturn_rate2010 + scale(transp/sqmi), data = na.omit(dat.hom.chicago),dist="poisson")

mod.poisson<-glm(n ~ foreign_share2010 + share_black2010 + share_hisp2010 + singleparent_share2010+mail_return_rate2010 + sc
ale(transp/sqmi), data = na.omit(dat.hom.chicago),family="poisson")

mod.pca.zero.inflated.poisson<-zeroinfl(na.omit(dat.hom.chicago)$n~scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chi
cago)$sqmi)+pca$scores[,1] + pca$scores[,2]+pca$scores[,3] + pca$scores[,4],dist="poisson")

mod.pca.poisson<-glm(na.omit(dat.hom.chicago)$n~scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicago)$sqmi)+pca$sco
res[,1] + pca$scores[,2]+pca$scores[,3] + pca$scores[,4],family="poisson")

mod.zero.inflated.nb<-zeroinfl(n ~ foreign_share2010 + share_black2010 + share_hisp2010 + singleparent_share2010+mail_return
_rate2010 + scale(transp/sqmi), data = na.omit(dat.hom.chicago),dist="negbin")

mod.pca.zero.inflated.nb<-zeroinfl(na.omit(dat.hom.chicago)$n~scale(na.omit(dat.hom.chicago)$transp/na.omit(dat.hom.chicag
o)$sqmi)+pca$scores[,1] + pca$scores[,2]+pca$scores[,3] + pca$scores[,4],dist="negbin")

BIC(mod.poisson,mod.zero.inflated.poisson,mod.pca.poisson,mod.pca.zero.inflated.poisson,mod.zero.inflated.nb,mod.pca.zero.in
flated.nb)
```

Figure B1. R Code used to generate the final models for each city and calculate the BIC for different Poisson-family models. The displayed code is for Chicago, IL, but the code was analogous for the other four cities.



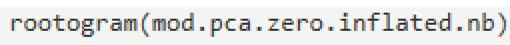
```
qqrplot(mod.pca.zero.inflated.nb)
```

Figure B2. R code used to generate the Q-Q plots for the zero-inflated PCA-Negative binomial models.



```
summary(mod.pca.zero.inflated.nb)
```

Figure B3. R code used to display the output, including coefficient estimates and levels of significant, for the zero-inflated PCA-negative binomial model for each city.



```
rootogram(mod.pca.zero.inflated.nb)
```

Figure B4. R code used to generate the rootogram for the zero-inflated PCA-negative binomial model for each city.