Adjusting for Bounding and Time-in-Sample Effects in the National Crime Victimization Survey (NCVS) Property Crime Rate Estimation

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

by

Hui Yang

Graduate Program in Statistics

The Ohio State University

2016

Dissertation Committee:

Elizabeth A. Stasny, Co-Advisor
Asuman Turkmen, Co-Advisor
Catherine A. Calder
Eloise K. Kaizar
Abstract

In this study, we deal with two problems: rotation group bias and lack of bounding information for producing crime rate estimates using the National Crime Victimization Survey (NCVS) data. Rotation group bias (time-in-sample effect) is the bias that occurs in a longitudinal survey when respondents provide less information as the number of times they are interviewed increases. Another kind of bias comes from the fact that respondents tend to remember serious events, like crimes, as having happened more recently than they actually did. At the first interviews, there is no previous record in the survey of crimes, and, as a result, more crimes are reported in the first interviews. Before 2006, the information from the first interviews in the NCVS was only used for bounding purposes and not in published crime rate estimates. However, in 2006, the Bureau of Justice Statistics (BJS) and the Census Bureau started to use the unbounded survey data for computing yearly crime rate estimates. As a result of the change in 2006, the estimated yearly crime rates have to be adjusted to eliminate the bias caused by lack of bounding information in the first interviews.

We created a longitudinal property crime data file to track 26759 households through all seven of their NCVS interviews over a period of six years, recording the number of property crimes that were reported by each household in each interview. The main goal of this study is to develop a statistical model that will allow us to adjust the estimated crime rates for both time-in-sample and bounding biases. We fit
zero-inflated count models, which allow extra zeros in standard Poisson count models, to model this property crime count data using a Bayesian approach. We also fit multinomial models to small domains longitudinal count data to study the bounding and time-in-sample effects. In the multinomial count model we also provide a way to deal with the missingness in the data. Adjustments to the estimated yearly household property crime rates are made using the model-fitting results from the multinomial count models.

**Key words:** rotation group bias, zero-inflated count models, multinomial count models, Bayesian models
This is dedicated to my family.
Acknowledgements

I would like to gratefully and sincerely thank Dr. Stasny and Dr. Turkmen for their guidance, understanding and patience.

I would like to thank the Department of Statistics at the Ohio State University, especially those members of my doctoral committee, for their input, valuable discussions and accessibility.
Vita

September 14, 1984 ................................Born - LiaoNing, China

2007 .................................................B.S. Fluid Mechanics,
TsingHua University, China

2009 .................................................M.S. Fluid Mechanics,
TsingHua University, China

2013 .................................................M.S. Chemical Engineering,
The Ohio State University

2013-present .................................Graduate Teaching Associate,
The Ohio State University

Fields of Study

Major Field: Statistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 The National Crime Victimization Survey (NCVS)</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Bounding and time-in-sample effects in the NCVS</td>
<td>3</td>
</tr>
<tr>
<td>1.2.1 Panel and rotating panel design of the NCVS</td>
<td>3</td>
</tr>
<tr>
<td>1.2.2 Rotation group bias in the NCVS</td>
<td>7</td>
</tr>
<tr>
<td>1.2.3 Forward telescoping – bounding effect</td>
<td>9</td>
</tr>
<tr>
<td>1.3 Organization of this thesis</td>
<td>13</td>
</tr>
<tr>
<td>2. The NCVS data</td>
<td>14</td>
</tr>
<tr>
<td>2.1 NCVS data structure</td>
<td>14</td>
</tr>
<tr>
<td>2.2 Creation of longitudinal property crime data by panel and rotation group</td>
<td>16</td>
</tr>
<tr>
<td>2.2.1 Household-level identification information</td>
<td>16</td>
</tr>
<tr>
<td>2.2.2 Incident-level property crime counts</td>
<td>18</td>
</tr>
<tr>
<td>2.2.3 Combining household weights and property crime count data</td>
<td>19</td>
</tr>
<tr>
<td>2.3 Preliminary descriptive data analysis</td>
<td>22</td>
</tr>
</tbody>
</table>
List of Tables

Table  Page

1.1 NCVS panel and rotation group chart (1999-2004) ...................... 6
1.2 2005 bounding adjustment factors, by time-in-sample ..................... 12
2.1 Panel 1 rotation group 1 household weights over the seven interviews 19
2.2 A portion of the property crime count data file .......................... 20
2.3 List of household-level covariates .......................................... 22
2.4 Indicators for categorical covariates ....................................... 26
2.5 Unweighted household property crime rates by each covariate .......... 28
3.1 Posterior percentiles of coefficients: Poisson, NB, and ZIP count models 50
3.2 Comparison of models by DIC .............................................. 50
3.3 Posterior percentiles for coefficients in the ZIP(14) model ............... 51
3.4 Comparison of ZIP models by DIC ........................................ 55
3.5 Posterior percentiles for coefficients in the ZIP(9) model ................ 55
3.6 Posterior percentiles for coefficients in the mean of the ZIP2(8) model 57
3.7 Posterior percentiles for the $\pi$ in the ZIP(9) and the $\pi_j$s in the ZIP2(8) model .............................................................. 59
3.8 Comparison of the ZIP model and ZIP2 model by DIC ................... 59
6.1 Number of households by number of reported crimes and interview sequence. 

127
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Rate of violent victimization, by time-in-sample, 2003 (Technical Notes, Criminal Victimization, 2006)</td>
<td>8</td>
</tr>
<tr>
<td>2.1</td>
<td>Example for NCVS annual hierarchical data structure</td>
<td>15</td>
</tr>
<tr>
<td>2.2</td>
<td>Property crime rates for the 36 panel and rotation groups that completed all seven interviews between 1999 and 2004</td>
<td>23</td>
</tr>
<tr>
<td>2.3</td>
<td>Bar graph of NCVS property crime counts for the 26759 households that completed all seven interviews between 1999 and 2004</td>
<td>24</td>
</tr>
<tr>
<td>2.4</td>
<td>Bar graph and table of household income categories</td>
<td>27</td>
</tr>
<tr>
<td>3.1</td>
<td>Histograms of the prior distributions for the $\pi$ in the ZIP or the $\pi_j$s in the ZIP2 model</td>
<td>46</td>
</tr>
<tr>
<td>3.2</td>
<td>Posterior distributions of coefficients (with intercept): Poisson, NB, and ZIP count models</td>
<td>47</td>
</tr>
<tr>
<td>3.3</td>
<td>Posterior distributions of coefficients (without intercept): Poisson, NB, and ZIP count models</td>
<td>49</td>
</tr>
<tr>
<td>3.4</td>
<td>Coefficients in the ZIP(14) model (including intercept)</td>
<td>52</td>
</tr>
<tr>
<td>3.5</td>
<td>Coefficients in the ZIP(14) model (excluding intercept)</td>
<td>52</td>
</tr>
<tr>
<td>3.6</td>
<td>Coefficients in the ZIP(9) model (including intercept)</td>
<td>54</td>
</tr>
<tr>
<td>3.7</td>
<td>Coefficients in the ZIP(9) model (excluding intercept)</td>
<td>54</td>
</tr>
</tbody>
</table>
3.8 Coefficients in the ZIP2(8) model (including intercept)..........................57
3.9 Coefficients in the ZIP2(8) model (excluding intercept).........................58
3.10 Boxplots of the posterior distributions for the $\pi$ in the ZIP(9) and the
    $\pi_j$s in the ZIP2(8) model..................................................60
3.11 Comparison of the actual and predicted crime rates.............................62
3.12 Histograms of the actual and predicted crime rates under five models
    (binwidth=0.5, left closed and right open intervals)..........................63
4.1 Time series plots for 18 domains data.............................................78
4.2 Sequences of observed relative frequencies of reporting zero crimes and
    proportions of missing by interview sequence................................100
4.3 Boxplots of posterior distributions of random effects ($\alpha$) in the small
domain Bayesian multinomial Model 4.2.3.1..................................103
4.4 Boxplots of posterior distributions of probabilities of missing an inter-
    view ($\pi$) in the small domain Bayesian multinomial Model 4.2.3.1......103
4.5 Boxplots of posterior distributions of probabilities of reporting zero
    crimes ($w$) in the small domain Bayesian multinomial Model 4.2.3.1...104
4.6 Boxplots of posterior distributions of the covariate coefficients ($\beta$) in
    the small domain Bayesian multinomial Model 4.2.3.1.....................104
4.7 Posterior means of probabilities from Model 4.2.3.1 (green curve) and
    Model 4.2.3.2 (red curve) and the observed relative frequencies (black
curves) for reporting zero crimes and missing a response by interview
    sequence.................................................................108
4.8 Posterior means of probabilities from Model 4.2.3.3 (red curves) and
    the observed relative frequencies (black curves) for reporting zero crimes
    and missing a response by interview sequence...............................111
4.9 Differences between the observed relative frequencies and the posterior
    means of probabilities from Model 4.2.3.3 for reporting zero crimes and
    missing a response by interview sequence....................................111
5.1 Adjusted base-weighted property crime rates and credible intervals for years 1999-2004
Chapter 1: Introduction

This dissertation studies the National Crime Victimization Survey (NCVS) data to adjust the annual estimate of the national property crime rate for rotation group bias and bounding effects. This first chapter introduces the history and design of the NCVS and also the underlying biases that are inheritant in the design of the survey.

1.1 The National Crime Victimization Survey (NCVS)

The National Crime Victimization Survey (NCVS) is a national survey of the U.S. population which focuses on collecting victimization data from respondents in sampled housing units. It is carried out by the US Census Bureau for the US Bureau of Justice Statistics and was started in 1972 as the National Crime Survey (NCS). A redesign of the survey was implemented in 1992, at which time the survey was renamed the NCVS. The NCVS collects data on both personal-level and household-level crimes by interviewing a household reference person and other household members age 12 or older living in sampled housing units (for additional details on the survey, see NACJD, “About NCVS, National Crime Victimization Survey Resource Guide”, ICPSR website). The interviewed households in the NCVS are selected by sampling addresses of housing units from the Master Address File used for the Decennial Census. The survey covers crimes that are both reported to the police and not reported
to the police by directly interviewing the sampled households either in person or by phone/computer interviews.

The NCVS mainly contains two types of information about the interviewed individuals and households. The first type of information is the basic demographic information for the interviewed individuals or households, such as the gender of the household reference person, household income, whether the household is in a city or rural area, and how many people age 12 or older are in the interviewed household. The second type of information is about any crimes experienced by the interviewed household and individuals. During each interview, respondents are asked whether there was any crime committed against them in the previous six-month period. The survey categorizes crimes as “personal” or “household” level and “property” or “violent” crimes. Personal crimes include rape and sexual attack, robbery, aggravated and simple assault, which are violent crimes, and purse-snatching/pocket-picking, which is a property crime. Household-level property crimes include household burglary, theft, and motor vehicle theft (NACJD, “About NCVS, National Crime Victimization Survey Resource Guide”, ICPSR website; “National Crime Victimization Survey: Unbounded Data Codebook”, Bureau of Justice Statistics, 2004). The household reference person reports for the whole household on household-level crimes that are committed against the whole household (Menard, 2007).

The NCVS provides a good source of crime victimization data with continuously collected data records for over 40 years. In addition, the detailed information provided in NCVS data makes it possible to study many aspects of the underlying mechanics of criminal victimization. By analyzing the NCVS crime victimization data, we can gain a great amount of knowledge about crime, such as which region of the nation
suffers most from crime, what characteristics of a household made it more vulnerable to household crimes and what kind of cities are safer with fewer crimes (Lynch and Addington, 2006). The Criminal Victimization Report published by the Bureau of Justice Statistics each year summarizes national violent and property crime rates based on analysis of NCVS data.

1.2 Bounding and time-in-sample effects in the NCVS

The bounding effect and rotation-group bias (time-in-sample effect) are two major sources of bias in the NCVS. Since they are caused by the survey design, it is not possible to avoid them. Therefore, we start by describing the design of the NCVS survey to gain a better understanding of these two types of biases.

1.2.1 Panel and rotating panel design of the NCVS

The NCVS uses a rotating panel design to guard against respondent fatigue and to avoid replacing the entire sample at once. The NCVS samples are labeled with the letter “J” followed by a number, e.g. J1, J2 and so on. The letter “J” tells us that this is an NCVS sample instead of a sample for another survey, and the numbers after the “J” tell us the sequence of a sample (“National Crime Victimization Survey: Unbounded Data Codebook”, Bureau of Justice Statistics, 2004). For example, sample J20 is the 20th sample in NCVS. In each sample, there are about 50000 households that are randomly selected from across the nation by sampling addresses of housing units from the Master Address File used for the Decennial Census. Each sample of approximately 50000 households is divided into six sub-groups (rotation groups), each with about 8300 different households. The six rotation groups from one sample do
not enter the NCVS at the same time, instead, they enter the survey one by one at 6-month intervals. Each rotation group of sampled households remains in the survey for 3\(\frac{1}{2}\) years, and all residents age 12 or older in each sampled household are interviewed at 6-month intervals for seven waves, each time collecting data for the previous six months. After the seventh interview, the whole rotation group leaves the survey and is replaced by a new rotation group from the next sample. Thus, every six months, a rotation group that has been contacted for seven interviews is dropped from the survey, and at the same time a rotation group from the next sample is added into the survey for its first interview. Thus, during any given 6-month period, seven rotation groups are interviewed: six of them are previously-interviewed rotation groups from the old sample, and one rotation group from the new sample is having its bounding (first) interview (“National Crime Victimization Survey, Technical Documentation”, Bureau of Justice Statistics, 2014). At each interview, the rotation groups are all experiencing a different sequence of interviews and there will always be a rotation group that is completing the first (or second, or third, or fourth, or fifth, or sixth, or seventh) interview. In other words, there are never two rotation groups taking the same wave of interviews at the same time. This is why they are called rotation groups and such a survey design is called a rotating panel design.

The rotating panel design of the NCVS enables preservation of the overall sample size. In this way, the NCVS design ensures that the total reported crime counts (or the weighted crime counts) by all the respondents in each interview do not differ too much from each other and the crime rate estimates derived from the survey data stay stable across each interview period.
Each rotation group is further divided into six panels. Panel 1 is interviewed in January and July of each year, panel 2 is interviewed in February and August, panel 3 is interviewed in March and September, panel 4 in April and October, panel 5 in May and November, and panel 6 in June and December. Thus, panels 1, 2 and 3 are interviewed in the first and third quarters each year, while panels 4, 5, and 6 are interviewed in the second and fourth quarters.

To better illustrate the panel and rotating panel design in the NCVS, we use the data in our time period of interest (1999-2004) as an example. For rotation group 1, sample J21, the first (bounding) interviews were in 2001’s first and second quarters (months January to June for panels 1 to 6), and the last (7th) interviews were in 2004’s first and second quarters. For rotation group 2, sample J21, the first (bounding) interviews were in 2001’s third and fourth quarters (months July to December for panels 1 to 6), and the last (7th) interviews were in 2004’s third and fourth quarters. For rotation group 3, sample J20, the first (bounding) interviews were in 1999’s first and second quarters, and the last (7th) interviews were in 2002’s first and second quarters. For rotation group 4, sample J20, the first (bounding) interviews were in 1999’s third and fourth quarters, and the last (7th) interviews were in 2002’s third and fourth quarters. For rotation group 5, sample J20, the first (bounding) interviews were in 2000’s first and second quarters, and the last (7th) interviews were in 2003’s first and second quarters. For rotation group 6, sample J20, the first (bounding) interviews were in 2000’s third and fourth quarters, and the last (7th) interviews were in 2003’s third and fourth quarters.

Table 1.1 shows the structure of the sample group rotation in the time period of interest to us. The two-digit numbers in the chart are the panel and rotation group
Table 1.1: NCVS panel and rotation group chart (1999-2004).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>31(1st)</td>
<td>32(1st)</td>
<td>33(1st)</td>
<td>34(1st)</td>
<td>35(1st)</td>
<td>36(1st)</td>
<td>31(2nd)</td>
<td>32(2nd)</td>
<td>33(2nd)</td>
<td>34(2nd)</td>
<td>35(2nd)</td>
</tr>
<tr>
<td></td>
<td>41(1st)</td>
<td>42(1st)</td>
<td>43(1st)</td>
<td>44(1st)</td>
<td>45(1st)</td>
<td>46(1st)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>41(2nd)</td>
<td>42(2nd)</td>
<td>43(2nd)</td>
<td>44(2nd)</td>
<td>45(2nd)</td>
<td>46(2nd)</td>
<td>41(3rd)</td>
<td>42(3rd)</td>
<td>43(3rd)</td>
<td>44(3rd)</td>
<td>45(3rd)</td>
</tr>
<tr>
<td></td>
<td>51(1st)</td>
<td>52(1st)</td>
<td>53(1st)</td>
<td>54(1st)</td>
<td>55(1st)</td>
<td>56(1st)</td>
<td>51(2nd)</td>
<td>52(2nd)</td>
<td>53(2nd)</td>
<td>54(2nd)</td>
<td>55(2nd)</td>
</tr>
<tr>
<td>2001</td>
<td>31(5th)</td>
<td>32(5th)</td>
<td>33(5th)</td>
<td>34(5th)</td>
<td>35(5th)</td>
<td>36(5th)</td>
<td>31(6th)</td>
<td>32(6th)</td>
<td>33(6th)</td>
<td>34(6th)</td>
<td>35(6th)</td>
</tr>
<tr>
<td></td>
<td>41(4th)</td>
<td>42(4th)</td>
<td>43(4th)</td>
<td>44(4th)</td>
<td>45(4th)</td>
<td>46(4th)</td>
<td>41(5th)</td>
<td>42(5th)</td>
<td>43(5th)</td>
<td>44(5th)</td>
<td>45(5th)</td>
</tr>
<tr>
<td></td>
<td>61(2nd)</td>
<td>62(2nd)</td>
<td>63(2nd)</td>
<td>64(2nd)</td>
<td>65(2nd)</td>
<td>66(2nd)</td>
<td>61(3rd)</td>
<td>62(3rd)</td>
<td>63(3rd)</td>
<td>64(3rd)</td>
<td>65(3rd)</td>
</tr>
<tr>
<td>11(1st)</td>
<td>12(1st)</td>
<td>13(1st)</td>
<td>14(1st)</td>
<td>15(1st)</td>
<td>16(1st)</td>
<td>11(2nd)</td>
<td>12(2nd)</td>
<td>13(2nd)</td>
<td>14(2nd)</td>
<td>15(2nd)</td>
<td>16(2nd)</td>
</tr>
<tr>
<td></td>
<td>21(1st)</td>
<td>22(1st)</td>
<td>23(1st)</td>
<td>24(1st)</td>
<td>25(1st)</td>
<td>26(1st)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>31(7th)</td>
<td>32(7th)</td>
<td>33(7th)</td>
<td>34(7th)</td>
<td>35(7th)</td>
<td>36(7th)</td>
<td>31(8th)</td>
<td>32(8th)</td>
<td>33(8th)</td>
<td>34(8th)</td>
<td>35(8th)</td>
</tr>
<tr>
<td></td>
<td>41(6th)</td>
<td>42(6th)</td>
<td>43(6th)</td>
<td>44(6th)</td>
<td>45(6th)</td>
<td>46(6th)</td>
<td>41(7th)</td>
<td>42(7th)</td>
<td>43(7th)</td>
<td>44(7th)</td>
<td>45(7th)</td>
</tr>
<tr>
<td></td>
<td>51(5th)</td>
<td>52(5th)</td>
<td>53(5th)</td>
<td>54(5th)</td>
<td>55(5th)</td>
<td>56(5th)</td>
<td>51(6th)</td>
<td>52(6th)</td>
<td>53(6th)</td>
<td>54(6th)</td>
<td>55(6th)</td>
</tr>
<tr>
<td></td>
<td>61(4th)</td>
<td>62(4th)</td>
<td>63(4th)</td>
<td>64(4th)</td>
<td>65(4th)</td>
<td>66(4th)</td>
<td>61(5th)</td>
<td>62(5th)</td>
<td>63(5th)</td>
<td>64(5th)</td>
<td>65(5th)</td>
</tr>
<tr>
<td></td>
<td>21(2nd)</td>
<td>22(2nd)</td>
<td>23(2nd)</td>
<td>24(2nd)</td>
<td>25(2nd)</td>
<td>26(2nd)</td>
<td>21(3rd)</td>
<td>22(3rd)</td>
<td>23(3rd)</td>
<td>24(3rd)</td>
<td>25(3rd)</td>
</tr>
<tr>
<td>2003</td>
<td>51(7th)</td>
<td>52(7th)</td>
<td>53(7th)</td>
<td>54(7th)</td>
<td>55(7th)</td>
<td>56(7th)</td>
<td>51(8th)</td>
<td>52(8th)</td>
<td>53(8th)</td>
<td>54(8th)</td>
<td>55(8th)</td>
</tr>
<tr>
<td></td>
<td>61(6th)</td>
<td>62(6th)</td>
<td>63(6th)</td>
<td>64(6th)</td>
<td>65(6th)</td>
<td>66(6th)</td>
<td>61(7th)</td>
<td>62(7th)</td>
<td>63(7th)</td>
<td>64(7th)</td>
<td>65(7th)</td>
</tr>
<tr>
<td></td>
<td>11(5th)</td>
<td>12(5th)</td>
<td>13(5th)</td>
<td>14(5th)</td>
<td>15(5th)</td>
<td>16(5th)</td>
<td>11(6th)</td>
<td>12(6th)</td>
<td>13(6th)</td>
<td>14(6th)</td>
<td>15(6th)</td>
</tr>
<tr>
<td></td>
<td>21(4th)</td>
<td>22(4th)</td>
<td>23(4th)</td>
<td>24(4th)</td>
<td>25(4th)</td>
<td>26(4th)</td>
<td>21(5th)</td>
<td>22(5th)</td>
<td>23(5th)</td>
<td>24(5th)</td>
<td>25(5th)</td>
</tr>
<tr>
<td>2004</td>
<td>11(7th)</td>
<td>12(7th)</td>
<td>13(7th)</td>
<td>14(7th)</td>
<td>15(7th)</td>
<td>16(7th)</td>
<td>11(8th)</td>
<td>12(8th)</td>
<td>13(8th)</td>
<td>14(8th)</td>
<td>15(8th)</td>
</tr>
<tr>
<td></td>
<td>21(6th)</td>
<td>22(6th)</td>
<td>23(6th)</td>
<td>24(6th)</td>
<td>25(6th)</td>
<td>26(6th)</td>
<td>21(7th)</td>
<td>22(7th)</td>
<td>23(7th)</td>
<td>24(7th)</td>
<td>25(7th)</td>
</tr>
</tbody>
</table>

The table shows the rotation groups and panels for the NCVS survey from 1999 to 2004. The numbers are divided into rotation groups and panels, where the tens digit gives the rotation group number and the units digit gives the panel number. The parentheses indicate the sequence numbers for each interview. The red entries are from sample J21, while the others are from sample J20.

In Table 1.1, one may read across the rows for a change in panel number, and down the columns for a change in rotation group number. For example, “31(1st)” means the households in rotation group 3 panel 1 are having their 1\textsuperscript{st} interviews.

A rotation group is introduced semi-annually from the new sample into the survey. After 3\frac{1}{2} years, when all seven interviews for a housing unit are completed, it takes the
panels from a rotation group six months to be completely dropped out of the survey. For example, in sample J20 rotation group 3 was introduced into the survey in the period of January - June 1999 and dropped out of the survey in the period of January - June 2002. Rotation group 4 was introduced into the survey in the period of July - December 1999 and dropped out of the survey in the period of July - December 2002.

1.2.2 Rotation group bias in the NCVS

In the NCVS, some respondents tend to provide less and less information for the survey as the number of interviews increases; this is called rotation group bias (sometimes also called respondent fatigue) (Fienberg, 2006). The longer a household remains in the sample, the less likely it is to report crimes committed against it to the NCVS. This decline in the number of crimes reported cannot be attributed to actual changes in the level of crime or to sampling variation since the rotation group bias happens in repeated interviews with the same individuals and has a common pattern of crime reports declining over time. Rotation group bias is generally found in all sample surveys involving periodically repeated interviews with the same individuals (Bailar, 1975), and NCVS procedures are unable to prevent rotation group bias.

The NCVS “Technical Notes” (Criminal Victimization, 2006) states that, “provided the sample remains balanced (i.e., the households interviewed during each wave of the survey are approximately evenly distributed into different interview sequences), the effects of rotation group bias should change little over time, having no effect on the survey’s year-to-year rate estimates”. In other words, the rotation group bias should affect each year’s crime estimates in the same way. However, it still biases the crime rate estimates at a single point in time.
Figure 1.1 shows the different levels of reporting violent victimizations by the sequences of all interviews that were conducted in 2003. The downward trend in the plot of the violent crime rate estimates in Figure 1.1 clearly illustrates the rotation group bias.

Fay and Li (2010) find evidence of time-in-sample effects by constructing a longitudinal data set for the years 1998 to 2004 containing waves 2 to 7. However, there is no generally accepted adjustment for the rotation group bias in NCVS estimates to date. The Census Bureau currently ignores the rotation group bias in estimating yearly crime rates. The rotating panel design of the NCVS makes adjusting for rotation group bias both difficult and complicated. In this work, we attempt to quantify the rotation group bias by fitting models to the longitudinal data on property crime.
counts for 36 panel and rotation groups using NCVS data sets for the years from 1999 to 2004.

1.2.3 Forward telescoping – bounding effect

In the NCVS, respondents in the sampled households are interviewed for a total of seven times during a $3\frac{1}{2}$ years period. In each interview the respondents are asked whether there were any crimes committed against them or their household in the previous six months. In their first interviews, because having a crime committed against you is often a highly stressful event, respondents might recall and report a crime as occurring within the 6-month period when, in fact, it occurred earlier. This kind of bias is called forward telescoping in the survey research literature.

Prior to 2006, the NCVS used a bounding interview procedure to guard against this kind of respondent recall bias. Before 2006, the data collected in the first interviews were not used in producing crime rate estimates in published reports. The purpose of the first interview was to provide bounding information for the second interview. Then, when in the second interview a respondent reported a crime as having happened in the six month period previous to the second interview, the interviewer would compare to the information from the first interview of the same household to identify whether the reported crime was a previously-reported crime (Fay and Li, 2010). The second-wave interview then serves to bound the third-wave interview and so on. Before 2006 the production of estimates from the sample of NCVS households included six rotation groups (i.e., seven interview rotation groups minus the bounding group). There were, however, still situations in which data collected from an unbounded interview would be included in deriving national crime victimization
estimates (Lynch and Addington, 2006). One of these situations was when an entire household moved from the sampled housing unit sometime during the 3\(\frac{1}{2}\) year period and was replaced by another household. Information from the new household’s first interview would be unbounced. These replacement respondents are both unbounced and mobile. Addington (2005) studied the effects of both bounding and mobility in the NCVS and provided a method to disentangle the two effects through logistic regression. Addington concluded that bounding only had significant effects on reports of property victimization, while moving only significantly affected reports of violent victimization.

Since the second interview is bounded by the first interview and the time-in-sample bias is not yet a big problem, the second wave interviews are thought to give crime-rate estimates “closest to the truth”. The 3\(^{rd}\) wave interviews tend to have a smaller number of crimes reported than the actual number. Then in the 4\(^{th}\) and later interviews, the respondents report even fewer and fewer crimes than what actually happened. Thus, when using the 2\(^{nd}\) through 7\(^{th}\) waves of interviews to estimate the yearly crime rates, there will be underestimation of the crime rate as the 3\(^{rd}\) through 7\(^{th}\) waves of interviews will suffer from rotation group bias and tend to have “fewer than the truth” crimes reported. Even the 2\(^{nd}\) interview may suffer from rotation group bias, of course, since the respondents gain experience from the first interviews.

To remain within the Bureau of Justice Statistics (BJS) budget, BJS and the Census Bureau decided to make use of the data from the bounding interviews for the sampled households entering the survey in and after 2006. In other words, they no longer exclude the first wave interviews when producing yearly crime rate estimates using NCVS data. Before 2006, the 1\(^{st}\) wave bounding interviews are not included.
in the public-use NCVS data published by BJS each year; this is called the bounded version of NCVS data. Beginning in 2006, BJS started to include data from all interviews (1st through 7th waves) in the public-use NCVS data, i.e., the unbounded version of the NCVS data. For research purposes, in 2006 BJS published the unbounded version of NCVS data for the years 1999-2005, which enables researchers to study the bounding effect in the NCVS. We note that there was a redesign of the sample starting in January 2005 based on the 2000 Census results. Under the new sampling design, NCVS samples addresses of housing units from the 2000 Decennial Census instead of the 1990 Decennial Census, and at the same time includes a larger portion of housing units in rural areas and small cities than before. As a result of using unbounded data to provide estimates and the sample redesign, the 2006 estimates of crime rates “differ from the 2005 estimates in a sense that might not be caused by actual changes in the level of crime” (Technical Notes, Criminal Victimization, 2006).

Without any adjustment to account for the effects of telescoping, the unbounded interviews would result in an over-reporting of crime. In 2004 a BJS and Census Bureau methodological working group began the process of developing a “weighting adjustment factor” to adjust for the over-reporting of crime in unbounded interviews to estimate crime rates for the year of 2006. As a first step, they developed adjustment factors for the bounding effect caused by replacement households (i.e., households that move into units already in the sample). Table 1.2 lists the adjustment factors for the downward adjustment of crime reports to account for bias in unbounded interviews from replacement households in the 2006 data. These factors are expected to account for both forward-telescoping (bounding effects) and rotation group bias. For example, if a replacement household was interviewed in 2006 for the first time but entered the
sample as a household in the 5th wave of interviews, we would multiply the sample weight for this household by 0.65 when computing a household crime rate.

The adjustment factors in Table 1.2 are calculated from the 2005 NCVS data by taking the ratios of crime counts reported in the 2nd-7th interviews to the crime counts in the 1st wave interviews using data from all the households that were interviewed in the year 2005 (Technical Notes, Criminal Victimization, 2006). These factors are naive adjustments ignoring the variations among the reported crime counts caused by factors other than the time effect. Furthermore, even for the households that rotate into the sample by design, we still need to adjust for the bounding and time-in-sample effects to obtain more realistic estimates of annual crime rates.

We will first investigate and quantify the bounding and time-in-sample effects through a modeling procedure which can separate these effects from other covariates such as the geographical and household factors. Then we will make adjustments to the reported crime counts with the model fitting results. The modeling procedure will provide us with both a better understanding of the NCVS crime count data and more accurate adjustments than the naive adjustment factors in Table 1.2.

<table>
<thead>
<tr>
<th>Period</th>
<th>Household adjustment factor</th>
<th>Person adjustment factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005 2:1</td>
<td>0.78</td>
<td>0.81</td>
</tr>
<tr>
<td>2005 3:1</td>
<td>0.69</td>
<td>0.73</td>
</tr>
<tr>
<td>2005 4:1</td>
<td>0.69</td>
<td>0.68</td>
</tr>
<tr>
<td>2005 5:1</td>
<td>0.65</td>
<td>0.58</td>
</tr>
<tr>
<td>2005 6:1</td>
<td>0.67</td>
<td>0.77</td>
</tr>
<tr>
<td>2005 7:1</td>
<td>0.62</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 1.2: 2005 bounding adjustment factors, by time-in-sample.
1.3 Organization of this thesis

The remainder of this thesis is organized as follows.

Chapter 2 introduces the format of the NCVS data and the procedure for constructing longitudinal data files for each panel and rotation group from the yearly published survey data.

Chapter 3 describes the crime count data for each household in each interview and fits Generalized Linear Mixed Models (GLMMs) to the crime count data. In particular, zero-inflated Bayesian count models are implemented and applied to fit the property crime count data. The model fitting results are also discussed.

Chapter 4 discusses the correlations over interview sequences in the crime count data and views the seven interviews’ data from the same group of households as longitudinal data. It also presents results from fitting multinomial models to the observed domain level crime count longitudinal data. We also address the problem of missing data in the multinomial model. The fit of these models is also evaluated.

Chapter 5 uses the model fitting results from previous chapters to adjust for bounding and time-in-sample effects in the annual household-level property crime rates. We compare these adjusted crime rates and corresponding confidence intervals to the unadjusted rates to show the effect of our adjustment.

Chapter 6 summarizes the results of the thesis, and points to some directions for future research to extend the work in this thesis.
Chapter 2: The NCVS data

This Chapter introduces the format of the yearly published NCVS data. It also describes in careful detail how to extract the reported crime counts for each household in each interview from the data in years 1999 to 2004. We show the procedure of creating a longitudinal data file for all the 36 panel and rotation groups that had all the seven waves of interviews during the period of 1999 to 2004. Our final cleaned data includes only those households that are in the sample for all seven interviews, i.e., not newly moved in and never moved out during the whole period of 3\frac{1}{2} years. However, these households might have some missing interviews.

2.1 NCVS data structure

The NCVS data are publicly available through the National Archive of Criminal Justice Data (NACJD) on the Inter-university Consortium for Political and Social Research (ICPSR) website. We used the annual unbounded NCVS data files from 1999 to 2004 (collection year format) to construct a longitudinal file of household responses by matching on the scrambled unit identifiers. We tracked all seven waves of interviews for each of the 36 panel and rotation groups. The unbounded data from 1999 to 2004 cover all seven interviews for rotation groups 3 to 6 from sample J20, and for rotation groups 1 and 2 from sample J21 as shown in Table 1.1. The unbounded
data are only available in NACJD for years starting from 1998. However, after 2006 the BJS only published unbounded data due to the reasons stated in Chapter 1.

As mentioned in the National Crime Victimization Survey Resource Guide, NACJD, ICPSR website: “There are four types of records in the NCVS files: 1) address ID record, 2) household record, 3) person record, and 4) incident record. The address ID and household records contain information about the household as reported by the household respondent and characteristics of the surrounding area as computed by the Bureau of Census. The person record contains information about each household member age 12 years and older as reported by that person or a proxy, with one record for each qualifying individual. Finally, the incident record contains information drawn from the incident report completed for each household incident or person incident mentioned during the interview.”

Figure 2.1 shows a typical record in the hierarchical data set. In the record shown in Figure 2.1, there are 3 persons age 12 years or older in this household, corresponding to 3 person-level records, and one incident was reported in this interview (one incident-level record).

```
Address level  →  1 6693 522251027609698262999999151200511162006
Household level  →  2 6693 522251027609698262999999151 011 19981122 ...
                  →  3 6693 522251027609698262999999151 1 12 15353111 ...
Person level     →  3 6693 522251027609698262999999151 2 2211434311 ...
                  →  3 6693 522251027609698262999999151 3 32 31717551 ...
Incident level   →  4 6693 522251027609698262999999151 3 340 19 32005 ...
```

Figure 2.1: Example for NCVS annual hierarchical data structure.
2.2 Creation of longitudinal property crime data by panel and rotation group

We begin with household-level property crimes to construct our longitudinal crime count data. The reason we focus on household-level property crime rates in this research is that there are a manageable number of households in the sample and it is relatively easy to deal with the property crimes. Furthermore, the household-level property crime count data can fulfill our purpose for this study. The household-level crimes are more stable and the effects of bounding interviews and time-in-sample effects are more clear in the property crimes than in the violent crimes (Addington, 2005).

2.2.1 Household-level identification information

In this section, we describe data extraction from the household-level records only. For each panel and rotation group, our goal is to identify those households that are in our sample for all seven interviews although they might miss some interviews. In other words, we begin our analyses by deleting the replacement households (about 1/4 of the sample) from our data because mobility is not the main concern in our current study. These replacement households entered the sample sometime between the 2\textsuperscript{nd} and 7\textsuperscript{th} interviews of a panel and rotation group thus are not in the sample for all seven interviews.

Take panel 1 rotation group 1 in sample J21 as an example. This panel and rotation group entered the sample in January of 2001. We first filtered the 2001 unbounded data by the panel and rotation group number. The variables in the NCVS codebook are named as “v” plus a number, for example, v2008 represents the panel
and rotation group number and v2004 represents the sample number (see “National Crime Victimization Survey: Unbounded Data Codebook”, Bureau of Justice Statistics, 2004). So we select cases with v2008 being 11 (panel 1 in rotation group 1) and sample number v2004 being J21. Then we select the records with the variable for quarter and year, v2003, to 011, i.e., year 01 quarter 1. Since this is the first interview for this panel and rotation group, we want to keep only those households that are in the sample for the first time, i.e., household status v2011=9 (not previously interviewed household), unit status v2010=2 or 9 (unit in sample for the first time this period or missing information for unit status), and household number v2006=1 (original household). We then record and store for later use the household scrambled control number (v2005) which is uniquely assigned to each household, and the household weights (v2116) for all the households remaining after the previous filtering.

For the second interview of panel 1 rotation group 1, we again filter the 2001 unbounded data with panel and rotation group number (v2008=11), sample number (v2004=J21), and quarter and year (v2003=013, i.e., year 01 quarter 3, since the second interview for panel 1 and rotation group 1 is in July of 2001). As this is the second interview for this panel and rotation group, we only keep those households with household number v2006=1 (original household, not a replacement household) to exclude all the replacement households that entered the sample at the 2nd interview of panel 1 rotation group 1. Again, we record and store the household scrambled control number (v2005) and household weights (v2116) for the remaining households. Here we do not need to filter by household status (v2011) since by having household number v2006=1 we have already ruled out those households with household status v2011=2 (replacement household since the previous enumeration).
We next repeat the same process for the 3rd through 7th interviews of panel 1 rotation group 1 to extract the household scrambled control number (v2005) and household weights (v2116) for original households for each of the interview periods. Finally, we merge the household records for all seven interviews of panel 1 rotation group 1 using the R function “merge()” on the unique household identifier–household scrambled control number (v2005). This will only keep the households whose scrambled control number (v2005) occurs in all seven interviews, and their corresponding household weights (a total of seven weights for each household, one for each interview).

The same procedure is repeated for the other 35 panel and rotation groups. Thus, we have 36 datasets corresponding to 36 panel and rotation groups (part of the data for panel 1 rotation group 1 is shown in Table 2.1). Each of the data sets has 8 variables: the household scrambled control number (v2005) and seven household weights. The households in each dataset are only those that remain in the sample for all seven interviews. The zero weights in Table 2.1 indicate a missing interview for the corresponding household at that particular interview time. The household could have been there and not responded, but no new replacement household was interviewed.

2.2.2 Incident-level property crime counts

We now describe the extraction of the incident-level data and counts of the number of property crimes reported for each household. The types of the crimes in NCVS are labeled by variable v4529. The property crimes are labeled as v4529>30. Thus, we count the number of crimes reported for each household with v4529>30 to obtain the property crime counts. The resulting file contains only records for those households
Table 2.1: Panel 1 rotation group 1 household weights over the seven interviews.

<table>
<thead>
<tr>
<th>Scrambled control#</th>
<th>hh_weights_1</th>
<th>hh_weights_2</th>
<th>hh_weights_3</th>
<th>hh_weights_4</th>
<th>hh_weights_5</th>
<th>hh_weights_6</th>
<th>hh_weights_7</th>
</tr>
</thead>
<tbody>
<tr>
<td>17073056293</td>
<td>1657.04618</td>
<td>1702.27630</td>
<td>1819.74837</td>
<td>1999.32204</td>
<td>0</td>
<td>0</td>
<td>2113.30050</td>
</tr>
<tr>
<td>17073093393</td>
<td>1946.01015</td>
<td>1919.68255</td>
<td>2124.13802</td>
<td>2342.97001</td>
<td>2234.34264</td>
<td>2147.56799</td>
<td>2105.24523</td>
</tr>
<tr>
<td>11786455782</td>
<td>2109.92285</td>
<td>2948.66794</td>
<td>2501.26515</td>
<td>2255.23123</td>
<td>2588.43572</td>
<td>2127.35106</td>
<td>2970.03173</td>
</tr>
<tr>
<td>175060906604</td>
<td>3080.66450</td>
<td>3042.90627</td>
<td>3104.46881</td>
<td>3112.58816</td>
<td>3120.16381</td>
<td>3291.31203</td>
<td>3292.24908</td>
</tr>
<tr>
<td>214474958218</td>
<td>1915.50403</td>
<td>2037.75075</td>
<td>2089.34619</td>
<td>2084.87950</td>
<td>2112.42157</td>
<td>2089.84741</td>
<td>2017.20590</td>
</tr>
<tr>
<td>289689000132</td>
<td>1624.35082</td>
<td>1672.65108</td>
<td>1829.25331</td>
<td>1818.38375</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>347862952982</td>
<td>2867.86673</td>
<td>2959.54603</td>
<td>2950.93851</td>
<td>3564.58899</td>
<td>3312.05846</td>
<td>3490.27980</td>
<td>3525.97634</td>
</tr>
<tr>
<td>480021955793</td>
<td>2946.27351</td>
<td>2894.51866</td>
<td>2903.04888</td>
<td>3188.17116</td>
<td>3242.77154</td>
<td>3357.28316</td>
<td>3298.63170</td>
</tr>
<tr>
<td>794194130829</td>
<td>1775.84272</td>
<td>1888.42916</td>
<td>1897.80462</td>
<td>1834.58313</td>
<td>1847.74892</td>
<td>1903.45619</td>
<td>1794.94581</td>
</tr>
<tr>
<td>951461052087</td>
<td>1974.96779</td>
<td>1966.85682</td>
<td>1809.55524</td>
<td>1948.04463</td>
<td>0</td>
<td>1900.00750</td>
<td>1923.63083</td>
</tr>
</tbody>
</table>

that have reported at least one incident (otherwise, there will be no incident-level data for the households). The result (shown partly in Table 2.2) of this operation will be a property crime count file for each year containing three variables: the unique household scrambled control number, the property crime count (at least one), and the year and quarter information for the corresponding interview. We then add into this property crime count data set those households that have a zero crime count but are in the household-level identification data set described in the previous section.

2.2.3 Combining household weights and property crime count data

To create the final property crime count data for each household in each panel and rotation group, we need to merge the household weights data file described in
<table>
<thead>
<tr>
<th>scrambled control #</th>
<th>property crime counts</th>
<th>year and quarter</th>
</tr>
</thead>
<tbody>
<tr>
<td>86288814958202</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>5626415980490</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>47625356206532</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>4873272120567</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>5689363368888</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>65334867475333</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>96971793035337</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>45842500439212</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>83058570193756</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>76294898919258</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>4860318905308</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>93255071558221</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>73259927105323</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>70991798430486</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>52465536190485</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>40827094118223</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>93047049362901</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>91954947684294</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>19737544465323</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>3718839906227</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 2.2: A portion of the property crime count data file.
Section 2.2.1 with the incident counts data described in Section 2.2.2 by the unique household scrambled control number. As an example, we merge the household weight in the first interview for panel 1 rotation group 1 with the property crime count data for year 2001 1st quarter (year and quarter=11) to get a list of households in panel 1 rotation group 1 that reported at least one property crime in the 1st interview. For the households that reported no property crime in the first interview, we record zero for the corresponding crime count.

Similarly, we merge the household weight in the second interview for panel 1 rotation group 1 with the property crime count data for year 2001 3rd quarter (year and quarter=13) to obtain a list of household property crime counts in panel 1 rotation group 1 in the 2nd interview. We then, in a similar way, create the property crime count data with weights for the 3rd to 7th interviews of panel 1 rotation group 1.

The same procedure was followed for the remaining 35 panel and rotation group combinations to get a total of 36 data sets of seven crime counts over the seven interview periods with weights. The total of 252 (36 by 7) lists of interviews are then combined to create the final property crime count data that we use for our study. As the reported property crimes are on the household level, we also included other information about each household as potential covariates for our models. For a list of our covariates, refer to Table 2.3. Among the covariates in Table 2.3, the explanatory variable named “Missing” is an indicator of whether the interview previous to the current interview is missing. This “Missing” variable addresses the situation when the current interview is unbounded because a household was not interviewed in the previous interview period.
2.3 Preliminary descriptive data analysis

The household property crime count data set we created in Section 2.2.3 may be used to calculate the weighted property crime rates. Taken together, the seven crime rates for each of the 36 panel and rotation group combinations form a longitudinal data set. In other words, for each of the seven interviews, we have 36 crime rates corresponding to the 36 panel and rotation groups. Figure 2.2 shows all the 252 (36 by 7) weighted crime rates ordered by sequence of the interviews (1 to 7). The red line in Figure 2.2 is the curve of the average crime rates for the 36 panel and rotation groups by sequence of the interviews. As we expected, although there is considerable variability in crime rates at each interview wave, we can still see very clearly the evidence of bounding effect (forward-telescoping bias) and time-in-sample effect (rotation group bias) from Figure 2.2.

<table>
<thead>
<tr>
<th>Name</th>
<th>Explanation</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>income</td>
<td>total household annual income</td>
<td>categorical</td>
</tr>
<tr>
<td>region</td>
<td>region of household</td>
<td>categorical</td>
</tr>
<tr>
<td>MSA</td>
<td>Metropolitan Statistical Area</td>
<td>categorical</td>
</tr>
<tr>
<td>quarter</td>
<td>quarter of the interview</td>
<td>categorical</td>
</tr>
<tr>
<td>Missing</td>
<td>whether previous interview is missing</td>
<td>indicator</td>
</tr>
<tr>
<td>sequence</td>
<td>numeration of the interview</td>
<td>numerical</td>
</tr>
<tr>
<td>hh_sex</td>
<td>gender of the household reference person</td>
<td>indicator</td>
</tr>
<tr>
<td>hh_num</td>
<td># of people older than 12 in the household</td>
<td>numerical</td>
</tr>
</tbody>
</table>

Table 2.3: List of household-level covariates.
Figure 2.2: Property crime rates for the 36 panel and rotation groups that completed all seven interviews between 1999 and 2004.

We combined the data to create a bar graph and a table of household property crime counts of all interviews for all 36 panel and rotation groups (Figure 2.3) ignoring the interview time. We have a total of 26759 households in our data with 174567 interviews. If all the households in the sample responded in all seven interview periods, we should have $26759 \times 7 = 187313$ interviews. That means we have $187313 - 174567 = 12746$ missing interviews. The bar graph and table in Figure 2.3 show that most households in our data reported no crime in an interview.

Possible GLMMs for count data include the Poisson GLMM model and the Negative Binomial model. An important characteristic of our NCVS property crime count data is that it exhibits an excessive number of zero counts that cannot be accounted
Figure 2.3: Bar graph of NCVS property crime counts for the 26759 households that completed all seven interviews between 1999 and 2004.

for by a Poisson or Negative Binomial model. This characteristic of the data leads us to a set of zero-inflated models which allow for extra zeros in the Poisson or Negative Binomial count models.

Due to the size of our property crime count data set (174567 counts for 26759 unique households), fitting zero-inflated GLMMs using R packages is very computationally intensive and thus time consuming or sometimes impossible. An alternative way of fitting our zero-inflated count models is to fit the Bayesian version of these models using an MCMC method. We will consider Bayesian models in Chapters 3 and 4.

To begin our modeling, we first convert every covariate that is categorical into a set of indicator variables. Among our covariate candidates (Table 2.3), the number of people age 12 or older in each household and the sequence of each interview (1 to 7) can be viewed as numerical variables. The gender of the household reference person (1
for male, 2 for female) and whether the previous interview is missing (1 for missing, 0 for not missing) are already indicator variables. For the “region” and “quarter” variables, we created 3 indicators each (see Table 2.4). For the “MSA” variable, two indicators were created. Finally, for the total income of a household, Figure 2.4 shows the bar graph and a table for the 14 different levels of income information provided in the NCVS household-level data. To avoid including too many indicator explanatory variables in our count model, we grouped these 14 levels of total household income into only 3 levels: high income, low income and middle income. As the division of income levels remains controversial, there is no commonly accepted way to define the income as high, middle or low. Purely by observing the bar graph of the household income in our data (Figure 2.4), we try to separate the households evenly into the 3 categories by the household income. Thus, we created the following categories: low income (< $20,000), middle income ($20,000 to $50,000), and high income (> $50,000). We also deleted the 3620 households that never reported income information reported during any of the seven interviews. Thus we have 23139 households in our data set for use in modeling household crime counts. Our final indicator variables for the categorical covariates are summarized in Table 2.4.

Table 2.5 shows the unweighted crime rates computed for each category of the potential covariates. The unweighted crime rates are calculated by dividing the actual crime count in each category by the total number of households in each category. From Table 2.5, we can get some general idea of how each of these covariates affects the household property crime rates. For example, the gender of the household reference person and the quarter of the interview may not have any significant effect on crime rates since the rates are fairly similar across categories for each of these two covariates.
Table 2.4: Indicators for categorical covariates.

<table>
<thead>
<tr>
<th>Name</th>
<th>indicator</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>income</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>region</td>
<td>northeast</td>
<td>midwest</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSA</td>
<td>Central city</td>
<td>Not MSA</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>quarter</td>
<td>1st</td>
<td>2nd</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 2.4: Bar graph and table of household income categories.

For the number of people in each household, in general, there are higher crime counts for those households that have more people in them. We also note that households with higher income report more crimes. For the region factor, households in the midwest and south have basically the same crime reporting rate in the NCVS, but households in the northeast are less likely to report a property crime committed against them while households in the west are more likely to report a property crime. This is consistent with common understanding of property crimes in the literature (Fay et al., 2013). For the MSA factor, bigger cities have a higher property crime rate while the rural areas (not in MSA) have the lowest crime rate. In the next Chapter, we will use these variables to model the household-level property crime counts. We expect the estimated coefficients of these covariates to be consistent with the exploratory data analysis.
Table 2.5: Unweighted household property crime rates by each covariate.

<table>
<thead>
<tr>
<th># of people in the household</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted crime rates</td>
<td>.048</td>
<td>.063</td>
<td>.117</td>
<td>.161</td>
<td>.204</td>
<td>.22</td>
<td>.341</td>
<td>.304</td>
<td>.071</td>
<td>.375</td>
</tr>
<tr>
<td>region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>northeast</td>
<td>.060</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>midwest</td>
<td></td>
<td>.079</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>south</td>
<td></td>
<td></td>
<td>.078</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>west</td>
<td></td>
<td></td>
<td></td>
<td>.101</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>.068</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>middle</td>
<td></td>
<td>.083</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high</td>
<td></td>
<td></td>
<td>.093</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quarter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>.083</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd</td>
<td></td>
<td>.077</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd</td>
<td></td>
<td></td>
<td>.077</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td></td>
<td>.080</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central city</td>
<td>.101</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small city</td>
<td></td>
<td>.074</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not MSA</td>
<td></td>
<td></td>
<td>.059</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender of household reference person</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.080</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.079</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.5: Unweighted household property crime rates by each covariate.
Chapter 3: Zero-inflated Bayesian GLMM count models

The primary goal of this thesis is to model the bounding and time-in-sample effects in NCVS data so that crime-rate estimates can be adjusted for these biases. For this purpose, we will need to study how the first interviews are different from other interviews and how the interview results change over time. Statistical models are useful tools for analyzing the relationships between different variables in the data. In this Chapter, we fit different zero-inflated count models to the property-crime data to quantify the effect of each of the potential covariates on the crime counts in addition to estimating the bounding and time-in-sample effects.

When considering appropriate models to describe the property crime count data, there are three primary features of the data that we need to address in the design of the models. First, the desired model needs to be able to handle the excessive number of zeros in the crime count data. Standard count models such as Poisson and Negative Binomial Generalized Linear Models (GLMs) result in a very small value for the mean parameter for data with excessive zeros since the average count in the data is very small due to the large majority of zeros in the data. As a result, the model tends to ignore the larger counts because they are too few compared to the zeros. Thus, the excessive zeros need to be separated from the larger counts so that the model can capture the trends in the larger counts. As previously mentioned in the exploratory...
data analysis, we will address this problem by using zero-inflated count models. We start with the regular count models and add the zero-inflation component to account for the excessive zeros making the GLMs zero-inflated GLMs. In NCVS data, the reported zero counts can have the following possible interpretations: 1. true zero count meaning no crime happened to the interviewed household in the previous six months; 2. respondent failed to report any crimes but in fact there was at least one crime committed against the interviewed household in the previous six months. The reason for not reporting a crime may be: a. the respondent did not recall the crime; b. the respondent did not realize that it was a crime; c. the respondent did not want to report an actual crime because he/she did not want to answer further questions about that crime. The probability of not reporting a crime generally increases as the number of interviews increases, which is consistent with the time-in-sample effect. In all the zero-inflated models described in this thesis, we do not take the source of zeros into consideration, but instead model all the zero counts together as a single group.

The second feature of the NCVS count data is that it includes repeated measurements for each household in the sample. It is common sense to think that the same household tends to report similar numbers of crimes in the seven sequential interviews over the period of $3\frac{1}{2}$ years. In other words, each household’s responses correlate with each other. Ignoring this correlation may lead to incorrect decisions on the importance of other potential covariates in our model. Thus, we must find a way in our zero-inflated count models to account for the correlations between the responses during interviews of the same household. Under a GLM framework, it is a commonly accepted practice to include a random effect for each individual in the
GLM to account for the effect of each subject. These kinds of GLMs are called GLMM (Generalized Linear Mixed-effect Models).

Finally, there are two types of missing-value problems in the property crime count data. One is that for some households, the household income information is never recorded during the whole period when they are in the NCVS sample. The reason for missing income information can be either on the interviewer’s side or the respondent’s side as household income information is somewhat private and personal. However, we do want to include income as one of the covariates for some of our zero-inflated GLMMs, which means that we have to exclude the whole record of those households with missing income information from the property crime count data when we fit the models. Imputation of income could be done but we do not attempt that because it is not the focus of this research. Among the 26759 households in the data set we constructed in Chapter 2, there are 3620 households with missing income information. Thus, for the models including income as one of the covariates, we are left with 23139 households in the data to estimate the model parameters. Another missing value problem is that some households are missing one or more interviews during the period of $3\frac{1}{2}$ years when they are in the NCVS sample. In the NCVS data, a missed interview is automatically marked as having a weight of zero, meaning that we do not use these data in estimating the annual crime rates. Although these missing interviews are not included in the calculation of crime rates, it does not mean that they have no effect on the estimated annual crime rates. Similar to the bounding effect, where the respondents are likely to over-report in their first interviews, the households missing an interview tend to report more crimes in the subsequent interview than the true number of crimes since they may recall a crime as happening within the past six
months while it, in fact, happened earlier. This is because when a household misses the previous interview, the current interview essentially becomes unbounded. We will need to also account for this unboundedness caused by missing interviews in the zero-inflated GLMMs. A way to account for the bounding effect in our model is to set up an indicator variable as one of the covariates indicating whether the current interview is the first interview for a household or whether the interview before the current interview is missing. We expect this indicator to have a positive coefficient corresponding to the over-reporting effect of unbounded interviews.

Other than these three features of the data, we have to take computing time into consideration when developing a reasonable model given the large size of the property crime count data. For this reason, Bayesian versions of the zero-inflated GLMM count models are considered. The remainder of Chapter 3 discusses the zero-inflated Bayesian count models that we have fitted to the property crime count data along with the model-fitting results.

3.1 Overview of Bayesian GLMMs for zero-inflated and overdispersed count data with repeated measures

Zero-inflated count models deal with the excessive number of zero counts in the data by allowing excessive zero counts on top of the traditional Poisson or Negative Binomial model. Lambert (1992), in an application to modeling defects in manufacturing, proposed a count model that is now known as the zero-inflation Poisson (ZIP) model. The ZIP model proposed by Lambert mixes a Poisson distribution with a point mass of probability at zero. This model is shown in Equations 3.1 and 3.2, where $y_{ij}$ is the observed data for individual $i$ at the $j^{th}$ repeated measurement and $Y_{ij}$ is the corresponding random variable. $Y_{ij}$ follows a mixture distribution of a point
mass at zero with probability \( \pi_{ij} \) and a Poisson distribution with probability \( 1 - \pi_{ij} \). \( \mu_{ij} \) is the mean value for the Poisson part of the mixture distribution.

\[
P(Y_{ij} = 0) = \pi_{ij} + (1 - \pi_{ij}) \exp(-\mu_{ij}) \quad (3.1)
\]

\[
P(Y_{ij} = k) = (1 - \pi_{ij})\frac{\mu_{ij}^k \exp(-\mu_{ij})}{k!}, \quad k = 1, 2, ... \quad (3.2)
\]

ZIP models may be used for count data in the case where the incidence of zero is greater than expected for the Poisson distribution. Zero-inflation is a manifestation of overdispersion because, from Equations 3.1 and 3.2, we have:

\[
E(Y_{ij}) = (1 - \pi_{ij})\mu_{ij} \quad (3.3)
\]

\[
\text{var}(Y_{ij}) = (1 - \pi_{ij})\mu_{ij} + \left[\frac{\pi_{ij}}{1 - \pi_{ij}}\right]((1 - \pi_{ij})\mu_{ij})^2 = E(Y_{ij}) + \left[\frac{\pi_{ij}}{1 - \pi_{ij}}\right](E(Y_{ij}))^2. \quad (3.4)
\]

From Equation 3.4, we can see that the variance of a zero-inflated Poisson distribution is larger than the variance of a regular Poisson distribution, \( E(Y_{ij}) \).

Lambert (1992) considered models in which

\[
\log(\mu_{ij}) = X_{ij}\beta \quad \text{and} \quad \text{logit}(\pi_{ij}) = Z_{ij}\gamma \quad (3.5)
\]

where \( X_{ij} \) and \( Z_{ij} \) are vectors of covariates and \( \beta \) and \( \gamma \) are vectors of parameters. The two vectors of covariates, \( X_{ij} \) and \( Z_{ij} \), may share the same covariates.

As described by Lambert (1992), “the ZIP model can be fit using maximum likelihood estimation and asymptotic variance-covariance matrices for the parameter estimates can be estimated using the inverse of the observed Fisher information matrix, and inference can be performed using likelihood ratio tests and confidence intervals”. Lambert did demonstrate through simulation results that when the sample size is
large (which is the case for our count data), normal theory tests and confidence intervals still work well for $\beta$ but do not work as well for $\gamma$. Lambert also discussed computation issues associated with obtaining maximum likelihood estimates with the EM algorithm, such as the choice of initial values and convergence of the algorithm.

Other than the zero-inflation in the NCVS crime count data, the dependency structure among counts from repeated interviews on the same households is another feature that must be addressed in models for the NCVS property count data. Min and Agresti (2005) introduced a random effect into a set of models called Hurdle models to capture the correlation between repeated measurements for each subject. The Hurdle model, which is another type of zero-inflated count model, is basically a mixture of a point mass at zero and a truncated (non-zero) Poisson or other count distribution. A variation of the Hurdle model is called the zero-altered Poisson (ZAP) model, which is just a special case of the Hurdle model where both the zero part and non-zero part of the data are assumed to come from Poisson/truncated-Poisson distributions and the mean parameter in the two Poisson distributions can be different from each other. Min and Agresti proposed a Hurdle random-effects model which has random effects in both the point mass component and the truncated (positive) count component of the model to account for the dependence in longitudinal data. They also discussed the maximum likelihood model-fitting procedure for both the normal and nonparametric forms of random-effects Hurdle models.

Compared to ZIP models, which do require that the two probability components (point mass at zero and Poisson($\mu_{ij}$)) be fit simultaneously, the Hurdle model allows the two components to be fit separately (Neelon et al., 2010). This means that Hurdle models are more convenient in the sense of model selection and mixing with other
non-Poisson distributions than ZIP models. Furthermore, Min and Agresti (2005) demonstrated through simulation studies that a ZIP model may encounter fitting problems if there is zero deflation at any setting of the explanatory variables, while Hurdle models are suitable for both zero-inflation and zero-deflation. In our study, we choose to use ZIP models because of their interpretability, and also because our property crime count data are clearly not zero-deflated as shown in the exploratory data analysis (Figure 2.3).

While several frequentist approaches (eg. the maximum likelihood methods mentioned above) have been proposed to fit zero-inflated models for repeated measures, Neelon et al. (2010) described a practical Bayesian approach using an MCMC method for fitting ZIP, Hurdle, and ZAP models. Bayesian inference has a number of well-known advantages compared to frequentist approaches (Gelman et al., 2004; Congdon, 2005). A Gibbs sampling algorithm was implemented by Neelon et al. in WinBUGS software using the “zeros trick” (details in Section 3.3) to sample from non-standard distributions (for example, truncated Poisson and mixed distributions). Neelon et al. discussed the choice of prior distributions and the computation of posterior distributions. While model selection using the Deviance Information Criterion (DIC) and pseudo-marginal likelihoods were discussed by Neelon et al., they do point out that once an optimal model has been determined, its fit to the data should also be evaluated.

The ZIP, Hurdle and ZAP models in Neelon et al. (2010) contain random effects in both the probability of zero portion of the model and the distribution for the positive count portion of the model. However it is also noted by Ghosh et al. (2012) that in the case of a high occurrence of zeros (higher than 80%), which is the case
in our NCVS property crime count data, including random effects in the occurrence model (the probability of zero portion) leads to highly unstable model fitting and computational intractability. Thus, we only include random effects in the Poisson component of our ZIP model as described later.

The DIC is a commonly accepted criterion for Bayesian model selection. Millar (2009) examined, by both simulation and application to data, the effectiveness of using DIC for model selection among six types of generalized linear count models: Poisson, Poisson-Gamma (or Negative Binomial), mixed-effects Poisson model (Poisson Log-linear model), and the zero-inflated versions of the above three models. Three types of DICs were examined: the traditional (conditional) DIC, the group-level marginalized DIC and the top-level marginalized DIC. The difference between these DICs lies in the likelihood function, either conditioned on parameters or a (partially) marginalized likelihood function, that was used in computing the DICs. Millar found that the traditional DIC does not perform as well as the group-level marginalized and the top-level marginalized DICs for evaluating the multilevel hierarchical models listed above. Millar also discussed another type of model selection criterion under the Bayesian framework, the Bayes’ factor, which is the ratio of the marginal densities of the data under two competing models. Bayes’ factors usually give reliable model selection results, but the method also suffers from the disadvantages of being computationally intensive and sensitive to prior specification (Millar, 2009). In the case of the NCVS property crime count data, we use the traditional DIC since the number of replicate observations per group in our data - up to seven interviews per household - may not be sufficient for the group-level and top-level marginalized DICs to perform well.
Further examination of the validity of using traditional DIC for model selection in our data may be a good candidate for future study.

3.2 Zero-inflated Bayesian GLMMs for household-level property crime count data

3.2.1 Basic setup

Let $y_{ij}$ be the property crime count for household $i$ in its $j^{th}$ interview, where $j = 1, \ldots, 7$. Let $X_{ij} = [1, X_{ij}^M, X_{ij}^S, \ldots]$ be the covariate vector for household $i$’s $j^{th}$ interview, where $X_{ij}^M = 1$ when the interview prior to the $j^{th}$ interview for household $i$ is missing and $= 0$ otherwise, and $X_{ij}^S = j$ indicates the sequence of the current interview. We set $X_{i1}^M = 1$ for all the households to account for the bounding effects meaning that the previous interview is always missing for the 1st wave interviews. Other possible covariates in $X_{ij}$ are listed in Table 2.3 and Table 2.4.

Let $\beta = (\beta_0, \beta_1, \beta_2, \ldots)' \sim N(0, \sigma_\beta^2 I)$ be the regression coefficients, and let $\alpha = \{\alpha_i : i = 1, \ldots, n\}$ be the random effects vector in the mean model of a GLMM model, where $n = 23139$ is the number of unique households for the data set we are using here.

Under a Bayesian framework, we have: $p(y|\alpha, \beta, \sigma_\beta^2) = \prod_{i=1}^{n} \prod_j p(y_{ij}|\alpha_i, \beta, \sigma_\beta^2)$, where $p(y_{ij}|\alpha_i, \beta, \sigma_\beta^2)$ is determined by the assumed data distribution.

3.2.2 Poisson and zero-inflated Poisson Bayesian GLMM models

A Bayesian version of the Poisson count model may be given as follows:

$$y_{ij}|\alpha_i, \beta, \sigma_\beta^2 \sim Poisson(\mu_{ij}),$$  \hspace{1cm} (3.6)$$

$$log(\mu_{ij}) = \alpha_i + \beta_0 + \beta_1X_{ij}^M + \beta_2X_{ij}^S + \ldots,$$  \hspace{1cm} (3.7)
\[ p(\alpha|\sigma_\alpha^2) = \prod_{i=1}^{n} p(\alpha_i|\sigma_\alpha^2), \quad p(\beta, \sigma_\beta^2, \sigma_\alpha^2) = p(\beta|\sigma_\beta^2)p(\sigma_\beta^2)p(\sigma_\alpha^2) \] (3.8)

where

\[ \alpha_i|\sigma_\alpha^2 \sim N(0, \sigma_\alpha^2). \] (3.9)

The zero-inflated Poisson (ZIP) model basically assumes that the count data follows a mixed distribution of a point mass at zero and a Poisson distribution, with \( \pi \) being the probability of the point mass at zero.

A Bayesian version of the ZIP model is as follows:

\[ y_{ij} = y_{ij}^* \psi_{ij} \] (3.10)

where \( \psi_{ij} \) is the realization of random variable \( \Psi_{ij}|\gamma_1 \sim Bernoulli(1 - \pi) \) indicating whether the current data is from the point mass portion of the model (\( \psi_{ij} = 0 \)) versus the Poisson component of the model (\( \psi_{ij} = 1 \)). \( y_{ij}^* \) is the realization of random variable \( Y_{ij}^*|\alpha_i, \beta, \sigma_\beta^2 \sim Poisson(\mu_{ij}) \), which is the Poisson component of the ZIP model. \( y_{ij}^* \) is the imagined observed data as if there are no excessive zeros and the observed counts follow a standard Poisson distribution.

We further have:

\[ \log(\mu_{ij}) = \alpha_i + \beta_0 + \beta_1 X_{ij}^M + \beta_2 X_{ij}^S + ..., \] (3.11)

\[ \logit(\pi) = \gamma_1, \quad \gamma_1 \sim N(0, \sigma_\gamma^2) \] (3.12)

\[ p(\alpha|\sigma_\alpha^2) = \prod_{i=1}^{n} p(\alpha_i|\sigma_\alpha^2) \] (3.13)

\[ p(\beta, \sigma_\beta^2, \sigma_\alpha^2, \sigma_\gamma^2) = p(\beta|\sigma_\beta^2)p(\sigma_\beta^2)p(\sigma_\alpha^2)p(\sigma_\gamma^2) \] (3.14)

where

\[ \alpha_i|\sigma_\alpha^2 \sim N(0, \sigma_\alpha^2), \quad \beta \sim N(0, \sigma_\beta^2 I). \] (3.15)
While the above ZIP model assumes that the probability of the point mass at zero, \( \pi \), is the same for all seven waves of interviews for all of the households, a more general model can be established by introducing different \( \pi \)'s for the different interview sequences. In this case instead of having just one \( \pi \), we have \( \pi_j \), where \( j = 1, 2, \ldots, 7 \) and \( \pi_j \) is the probability of point mass at zero for the \( j^{th} \) interview. Here, we still assume that all households in our data share the same set of \( \pi_j \)'s. At the same time, we eliminate the “Sequence” covariate in the mean component (\( \mu_{ij} \)) of our ZIP model. In other words, we include the “Sequence” effect in the point mass portion of the model instead of including it in the Poisson portion of our ZIP model. Thus this second version of the ZIP model, which is called zero-inflated Poisson2 (ZIP2) here, has the following format:

\[
y_{ij} = y_{ij}^* \psi_{ij} \tag{3.16}
\]

where

\[
Y_{ij}^*|\alpha_i, \beta, \sigma_\beta^2 \sim \text{Poisson}(\mu_{ij}) \tag{3.17}
\]

\[
\Psi_{ij}|\gamma_{1,j} \sim \text{Bernoulli}(1 - \pi_j) \tag{3.18}
\]

where

\[
\log(\mu_{ij}) = \alpha_i + \beta_0 + \beta_1 X_{ij}^M + \ldots, \tag{3.19}
\]

\[
\text{logit}(\pi_j) = \gamma_{1,j}, \quad \gamma_{1,j} \sim \text{N}(0, \sigma_{\gamma}^2) \tag{3.20}
\]

\[
p(\alpha|\sigma_\alpha^2) = \prod_{i=1}^n p(\alpha_i|\sigma_\alpha^2), \tag{3.21}
\]

\[
p(\beta, \sigma_\beta^2, \sigma_\alpha^2, \sigma_\gamma^2) = p(\beta|\sigma_\beta^2)p(\sigma_\beta^2)p(\sigma_\alpha^2)p(\sigma_\gamma^2) \tag{3.22}
\]

where

\[
\alpha_i|\sigma_\alpha^2 \sim \text{N}(0, \sigma_\alpha^2), \quad \beta \sim \text{N}(0, \sigma_\beta I). \tag{3.23}
\]
We will present the fits of the above models to the NCVS data in Section 3.4.

### 3.2.3 Negative Binomial count model and zero-inflated Negative Binomial count model

The Negative Binomial (NB) count model assumes that the count data follow a NB distribution instead of a Poisson distribution. The Negative Binomial count model works better than a Poisson count model when there is overdispersion in the data, but it suffers from a penalization of having more parameters to estimate than in the Poisson count model.

For our property crime count data and the ZIP models described previously, we can similarly use a Negative Binomial distribution to replace the Poisson in the mixed distribution in Equations 3.1 and 3.2. This yields the following mixture model which will be referred to as the zero-inflated Negative Binomial (ZINB) count model from now on:

\[
P(Y_{ij} = 0) = \pi + (1 - \pi)\left(\frac{r}{r + \lambda_{ij}}\right)^r
\]

\[
P(Y_{ij} = k) = (1 - \pi)\left(\frac{k + r - 1}{k}\right)\left(\frac{\lambda_{ij}}{r + \lambda_{ij}}\right)^k\left(\frac{r}{r + \lambda_{ij}}\right)^r, \quad k = 1, 2, ...
\]

where \(\lambda_{ij}\) is the expected value of the property crime count for household \(i\) in its \(j^{th}\) interview and \(r\) is the parameter for number of failures (non-crime in NCVS data) in the NB distribution. Under the ZINB model, we are assuming that ignoring the excessive zeros in the data, the probability of reporting \(k\) crimes by a household in one interview is the same as the probability of observing \(k\) successes in a sequence of independent and identically distributed Bernoulli trials before \(r\) failures occur. In other words, if we toss coins to determine the number of crimes reported in an...
interview by a household, we will count and report the number of tails we have before
the \( r^{th} \) head occurs.

### 3.3 Zero-inflated models in \textit{rjags}

For a ZIP model, the sampling distribution for our property crime counts — the
mixed distribution in Equations 3.1 and 3.2 — is not included in the list of standard
distributions in the R package "\textit{rjags}" (Plummer, 2015). Therefore, it is impossible
to directly code the above ZIP models in \textit{rjags}. However, the literature on Bayesian
ZIP models (Spiegelhalter et al., 2002) provides a way to sample from an arbitrary
distribution in the MCMC procedure, called the “zeros trick”. The “zeros trick” uses
the fact that a \( \text{Poisson}(\phi) \) observation of zero has likelihood function of \( \exp(-\phi) \).
Now, imagine that we have a set of observed data \( z_{ij} \) that is a set of zeros with the
对应的 set of random variables \( Z_{ij} \)s following \( \text{Poisson}(\phi_{ij}) \) distributions. Then
we set \( \phi_{ij} \) to \(-LL_{ij}\) with \( LL_{ij} \) being the loglikelihood function of the ZIP distribution.
We then obtain the correct likelihood function as \( \exp(-\phi_{ij}) = \exp(LL_{ij}) = ll_{ij} \), where
\( ll_{ij} \) is the portion of the likelihood function from the \( i^{th} \) household in the \( j^{th} \) interview,
which is assumed to be the likelihood function of the ZIP distribution. Thus, we are
actually defining the likelihood function for a Poisson variable \( Z_{ij} \) by the portion of
likelihood function from the \( i^{th} \) household in the \( j^{th} \) interview when the realization of
\( Z_{ij} \) happen to be zero. Here, it is important to note that \( \phi_{ij} \) should always be greater
than zero as it is a Poisson mean. Therefore, we may need to add a suitable constant
to \( \phi_{ij} = -LL_{ij} \) to ensure that it is positive.

Using the “zeros trick”, the Bayesian version of our ZIP model becomes:

\[
 z_{ij} = 0, \quad Z_{ij} \sim \text{Poisson}(-LL_{ij}),  \tag{3.26}
\]
\[ LL_{ij} = I(y_{ij} = 0) \log(\pi + (1 - \pi) \exp(-\mu_{ij})) \]
\[ + (1 - I(y_{ij} = 0)) (\log(1 - \pi) - \mu_{ij} + y_{ij} \log(\mu_{ij}) - \log(y_{ij}!)), \]

where the \( Z_{ij} \)'s are Poisson random variables with parameters \(-LL_{ij}\) and the \( z_{ij} \)'s are one set of realizations of the \( Z_{ij} \)'s which happen to be all zero.

We further have:

\[ \text{logit}(\pi) = \gamma_1, \quad \gamma_1 \sim N(0, \sigma_\gamma^2) \]  
(3.28)

\[ \log(\mu_{ij}) = \alpha + \beta_0 + \beta_1 X_{ij}^M + \beta_2 X_{ij}^S + ..., \]  
(3.29)

\[ p(\alpha | \sigma_\alpha^2) = \prod_{i=1}^{n} p(\alpha_i | \sigma_\alpha^2), \]  
(3.30)

\[ p(\beta, \sigma_\beta^2, \sigma_\alpha^2, \sigma_\gamma^2) = p(\beta | \sigma_\beta^2) p(\sigma_\beta^2) p(\sigma_\alpha^2) p(\sigma_\gamma^2) \]  
(3.31)

where

\[ \alpha_i | \sigma_\alpha^2 \sim N(0, \sigma_\alpha^2), \quad \beta \sim N(0, \sigma_\beta^2 I). \]  
(3.32)

For the ZIP2 model, the Bayesian version with the "zeros trick" is:

\[ z_{ij} = 0, \quad Z_{ij} \sim \text{Poisson}(-LL_{ij}), \]  
(3.33)

\[ LL_{ij} = I(y_{ij} = 0) \log(\pi_j + (1 - \pi_j) \exp(-\mu_{ij})) \]
\[ + (1 - I(y_{ij} = 0)) (\log(1 - \pi_j) - \mu_{ij} + y_{ij} \log(\mu_{ij}) - \log(y_{ij}!)), \]

where

\[ \text{logit}(\pi_j) = \gamma_{1,j}, \quad \gamma_{1,j} \sim N(0, \sigma_\gamma^2) \]  
(3.35)

\[ \log(\mu_{ij}) = \alpha_i + \beta_0 + \beta_1 X_{ij}^M + ..., \]  
(3.36)

\[ p(\alpha_i | \sigma_\alpha^2) = \prod_{i=1}^{n} p(\alpha_i | \sigma_\alpha^2), \]  
(3.37)

\[ p(\beta, \sigma_\beta^2, \sigma_\alpha^2, \sigma_\gamma^2) = p(\beta | \sigma_\beta^2) p(\sigma_\beta^2) p(\sigma_\alpha^2) p(\sigma_\gamma^2) \]  
(3.38)
where
\[ \alpha_i | \sigma^2_\alpha \sim N(0, \sigma^2_\alpha), \quad \beta \sim N(0, \sigma^2_\beta I). \] (3.39)

To find the model that fits our data best and also to compare the various models (according to the diagnostic criteria described in the next section), we fit not only the Poisson and ZIP models, but also the Negative Binomial and ZINB count models.

A Bayesian version of the Negative Binomial count model is:
\[ y_{ij} | \alpha_i, \beta, \sigma^2_\beta \sim NB \left( \frac{r}{r + \lambda_{ij}}, r \right), \] (3.40)

where we choose \( r \sim unif(0, 50) \) as a non-informative prior because we expect that the value for \( r \) will be much smaller than 50.

We further have:
\[ \log(\lambda_{ij}) = \alpha_i + \beta_0 + \beta_1 X_{ij}^M + \beta_2 X_{ij}^S + ..., \] (3.41)

\[ p(\alpha | \sigma^2_\alpha) = \prod_{i=1}^{n} p(\alpha_i | \sigma^2_\alpha), \] (3.42)

\[ p(\beta, \sigma^2_\beta, \sigma^2_\alpha, \sigma^2_\gamma) = p(\beta | \sigma^2_\beta)p(\sigma^2_\beta)p(\sigma^2_\alpha)p(\sigma^2_\gamma). \] (3.43)

where
\[ \alpha_i | \sigma^2_\alpha \sim N(0, \sigma^2_\alpha), \quad \beta \sim N(0, \sigma^2_\beta I). \] (3.44)

For the ZINB model (with the “zeros trick”):
\[ z_{ij} = 0, \quad Z_{ij} \sim Poisson(-LL_{ij}), \] (3.45)

\[ LL_{ij} = I(y_{ij} = 0) \log(\pi + (1 - \pi)(\frac{r}{r + \lambda_{ij}})^r) \]
\[ + (1 - I(y_{ij} = 0))(\log(1 - \pi) + \log(\frac{y_{ij} + r - 1}{y_{ij}})) \]
\[ + y_{ij}\log(\frac{\lambda_{ij}}{r + \lambda_{ij}}) + r\log(\frac{r}{r + \lambda_{ij}})), \] (3.46)
where $r \sim \text{unif}(0, 50)$ and

$$
\text{logit}(\pi) = \gamma_1, \quad \gamma_1 \sim N(0, \sigma_{\gamma}^2),
$$

(3.47)

$$
\log(\lambda_{ij}) = \alpha_i + \beta_0 + \beta_1 X_{ij}^M + \beta_2 X_{ij}^S + \ldots,
$$

(3.48)

$$
p(\alpha|\sigma_{\alpha}^2) = \prod_{i=1}^{n} p(\alpha_i|\sigma_{\alpha}^2),
$$

(3.49)

$$
p(\beta, \sigma_\beta^2, \sigma_\alpha^2, \sigma_\gamma^2) = p(\beta|\sigma_{\beta}^2)p(\sigma_{\beta}^2)p(\sigma_{\alpha}^2)p(\sigma_{\gamma}^2),
$$

(3.50)

where

$$
\alpha_i|\sigma_{\alpha}^2 \sim N(0, \sigma_{\alpha}^2), \quad \beta \sim N(0, \sigma_{\beta}^2I).
$$

(3.51)

Similar to the ZIP2 model, we can also have a ZINB2 model which moves the “Sequence” effect from the Negative Binomial component of the model to the point mass portion of our ZINB model. Details about the ZINB2 model are omitted here since it is just a variation of the ZINB model. We will discuss the model fitting results of the ZIP and ZIP2 models in detail in Section 3.4, but we will omit the model fitting results of the ZINB and ZINB2 models since they do not differ very much from those of ZIP and ZIP2 models.

### 3.4 Model fitting results and model comparison

We fit the Poisson, NB and the zero-inflated GLMMs described in the previous two sections (Section 3.2 and Section 3.3) using an MCMC implemented through the “rjags” package in R (Plummer, 2015). We sampled the posteriors every five steps for a total of 40000 samples with 10000 burn-in steps. As for the priors of the random effects ($\alpha_i$s), we assume that $\alpha_i \sim N(0, \sigma_{\alpha}^2)$ for $i = 1, 2, \ldots, 7$ and $\sigma_{\alpha}^2$ follows a Inverse Gamma (0.01, 100) distribution. For the priors of the covariate coefficients ($\beta$s), we assume $\beta \sim N(0, 10^4 I)$. For the prior of the probability of point mass at
zero ($\pi$ or $\pi_j$) in the zero-inflated GLMMs, we assume $\text{logit}(\pi) = \gamma_1 \sim N(0, \sigma^2_{\gamma_1})$ or $\text{logit}(\pi_j) = \gamma_{1,j} \sim N(0, \sigma^2_{\gamma_{1,j}})$ with $\sigma^2_{\gamma_1} = \sigma^2_{\gamma_{1,j}} = 10^4$. In other words, we set all the variances, $\sigma^2_{\alpha}$, $\sigma^2_{\beta}$ and $\sigma^2_{\gamma_1}$, to be very large numbers for the purpose of using very wide Normal distributions as non-informative priors for the model parameters.

For the purpose of sensitivity analysis, we tried a few of other Normal prior distributions for $\gamma_1$ (or $\gamma_{1,j}$) with different values for $\sigma^2_{\gamma_1}$ (or $\sigma^2_{\gamma_{1,j}}$). Figure 3.1 lists the prior distributions that we have used and shows the histograms of samples of $\pi$s drawn from each of the prior distributions. From Figure 3.1 we can see that the shape of the prior distributions that we have tried out for $\pi$ (or $\pi_j$) are quite different from each other. However, the model fitting result turned out to be not sensitive to the choice of the prior distribution.

### 3.4.1 Comparison of Poisson, NB and ZIP models

We first compare the model fitting results from three types of count models: Poisson, Negative Binomial (NB) and zero-inflated Poisson (ZIP). The covariates included in the three models being compared are: $X^M_{ij}$ (indicator of whether the interview previous to the current interview is missing – “Missing”), $X^S_{ij}$ (the numeration of the current interview – “Sequence”) and $X^G_{ij}$ (the gender of the household reference person – “Gender”). Since we did not include the income covariate in these models, we used records from all 26759 households in the data. We did not save any of the posterior distributions for the random effects because saving one random effect posterior distribution for each unique household requires too much memory given that there are a total of 26759 unique households.
Figure 3.1: Histograms of the prior distributions for the $\pi$ in the ZIP or the $\pi_j$s in the ZIP2 model.
The boxplots of the posterior distributions of the coefficients obtained from fitting the three models (Poisson, NB and ZIP) are shown in Figure 3.2. Since the magnitude of the intercept is much larger than those of the other coefficients, we provide the boxplots for the posterior distributions of the coefficients excluding the intercept in Figure 3.3. We also list the posterior percentiles for all of the four coefficients in Table 3.1.

From the table and figures, it is reasonably clear that the gender of the household reference person does not have any great effect on household property crime counts in any of the three models, which is consistent with our exploratory analysis in Section 2.3. Another important finding is that the posterior means of the coefficients corresponding to each covariate from these three models are quite close to each other. In other words, the three count models under examination are consistent with each other in the sense of estimating the coefficients.

(a) Poisson model  (b) NB model  (c) ZIP model

Figure 3.2: Posterior distributions of coefficients (with intercept): Poisson, NB, and ZIP count models.
There is an intercept in all three models around $-3$, which means that the mean of our count model is around $exp(-3) = 0.050$ crimes per interview per household. Although the magnitudes of the coefficients for “Missing” and “Sequence” are much smaller than that of the intercept, the coefficients of those two covariates still have posterior means that are very different from zero. The “Missing” variable has a positive effect on crime counts, meaning that if the interview prior to the current interview is missing for a household, then this household is likely to report a higher crime count in the current interview. This conclusion is consistent with the bounding effect, saying that the respondents are likely to identify a crime as having happened in the past six months when it actually happened earlier. The “Sequence” variable has a negative effect meaning that the crime counts reported by a household are likely to decline as the households have more and more interviews. This is consistent with rotation group bias, saying that the respondents are likely to report fewer and fewer crimes in repeated interviews. All of these results agree with our expectation and intuition.

To choose the best one from these three models (Poisson, NB and ZIP), we compare the DIC (Deviance Information Criterion) of these three models. DIC is defined as:

$$DIC = \overline{D(\theta)} + p_D$$  \hspace{1cm} (3.52)

where $\theta$ is a set of parameters, and $\overline{D(\theta)}$ is the posterior mean of the deviance $D(\theta)$, which can be expressed as:

$$D(\theta) = -2\log(f(y|\theta)).$$  \hspace{1cm} (3.53)
In Equation 3.52, $p_D$ quantifies the “effective number of parameters” in the model, and it has the form of:

$$p_D = \bar{D}(\theta) - D(\bar{\theta})$$

(3.54)

where $\bar{\theta}$ is the posterior mean of parameter $\theta$. Thus, $D(\bar{\theta})$ is the deviance evaluated at $\bar{\theta}$.

So, the DIC in Equation 3.52 can be calculated by:

$$DIC = 2\bar{D}(\theta) - D(\bar{\theta}),$$

(3.55)

which is exactly how the DIC is calculated in the rjags package (Plummer, 2015).

The corresponding DICs for the Poisson, NB and ZIP models are shown in Table 3.2. As we prefer the model that has the smallest DIC, from Table 3.2, we arrive at the conclusion that the ZIP model outperforms the other two models. The Poisson and NB count models do not differ too much from each other in the sense of DIC.
Table 3.1: Posterior percentiles of coefficients: Poisson, NB, and ZIP count models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Poisson</th>
<th>NB</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept 2.5%</td>
<td>-3.012</td>
<td>-2.959</td>
<td>-2.527</td>
</tr>
<tr>
<td>50%</td>
<td>-2.924</td>
<td>-2.865</td>
<td>-2.429</td>
</tr>
<tr>
<td>97.5%</td>
<td>-2.831</td>
<td>-2.775</td>
<td>-2.325</td>
</tr>
<tr>
<td>Missing   2.5%</td>
<td>0.155</td>
<td>0.156</td>
<td>0.153</td>
</tr>
<tr>
<td>50%</td>
<td>0.204</td>
<td>0.210</td>
<td>0.205</td>
</tr>
<tr>
<td>97.5%</td>
<td>0.254</td>
<td>0.265</td>
<td>0.261</td>
</tr>
<tr>
<td>Sequence  2.5%</td>
<td>-0.081</td>
<td>-0.082</td>
<td>-0.082</td>
</tr>
<tr>
<td>50%</td>
<td>-0.071</td>
<td>-0.071</td>
<td>-0.071</td>
</tr>
<tr>
<td>97.5%</td>
<td>-0.060</td>
<td>-0.060</td>
<td>-0.059</td>
</tr>
<tr>
<td>Gender    2.5%</td>
<td>-0.064</td>
<td>-0.064</td>
<td>-0.062</td>
</tr>
<tr>
<td>50%</td>
<td>-0.017</td>
<td>-0.017</td>
<td>-0.017</td>
</tr>
<tr>
<td>97.5%</td>
<td>0.031</td>
<td>0.032</td>
<td>0.031</td>
</tr>
</tbody>
</table>

Table 3.2: Comparison of models by DIC.

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>92246</td>
</tr>
<tr>
<td>NB</td>
<td>92760</td>
</tr>
<tr>
<td>ZIP</td>
<td>85146</td>
</tr>
</tbody>
</table>

Section 3.5 has more details on the overall fit of these models to the property crime count data.

### 3.4.2 Covariate effects

As we can see from the model-fitting results in the previous section, the gender of the household reference person does not have a big impact on property crime, so we exclude the variable “Gender” from our models. To decide whether the other covariates affect the property crime counts or not, we fit a ZIP model with all of
Table 3.3: Posterior percentiles for coefficients in the ZIP(14) model.

<table>
<thead>
<tr>
<th>parameter</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.322</td>
<td>-3.201</td>
<td>-3.085</td>
</tr>
<tr>
<td>Missing</td>
<td>0.154</td>
<td>0.210</td>
<td>0.266</td>
</tr>
<tr>
<td>Sequence</td>
<td>-0.079</td>
<td>-0.067</td>
<td>-0.055</td>
</tr>
<tr>
<td>Income_low</td>
<td>-0.141</td>
<td>-0.075</td>
<td>-0.007</td>
</tr>
<tr>
<td>Income_high</td>
<td>-0.028</td>
<td>0.023</td>
<td>0.075</td>
</tr>
<tr>
<td>centrality</td>
<td>0.324</td>
<td>0.377</td>
<td>0.431</td>
</tr>
<tr>
<td>notMSA</td>
<td>-0.249</td>
<td>-0.172</td>
<td>-0.097</td>
</tr>
<tr>
<td>Northeast</td>
<td>-0.504</td>
<td>-0.429</td>
<td>-0.354</td>
</tr>
<tr>
<td>Midwest</td>
<td>-0.218</td>
<td>-0.150</td>
<td>-0.082</td>
</tr>
<tr>
<td>South</td>
<td>-0.219</td>
<td>-0.155</td>
<td>-0.091</td>
</tr>
<tr>
<td>firstQ</td>
<td>-0.030</td>
<td>0.029</td>
<td>0.088</td>
</tr>
<tr>
<td>secondQ</td>
<td>-0.088</td>
<td>-0.035</td>
<td>0.016</td>
</tr>
<tr>
<td>thirdQ</td>
<td>-0.113</td>
<td>-0.053</td>
<td>0.008</td>
</tr>
<tr>
<td>Num_members</td>
<td>0.380</td>
<td>0.401</td>
<td>0.421</td>
</tr>
</tbody>
</table>

our other potential covariates. This model has 14 coefficients corresponding to the intercept and the 13 explanatory variables in the mean model and we named this model ZIP(14). We correspondingly call the ZIP model described in the previous section ZIP(4) model as it has 4 coefficients corresponding to the intercept and three explanatory variables (Missing, Sequence and Gender) in the mean model.

The posterior percentiles of the coefficients corresponding to the covariates in ZIP(14) are listed in Table 3.3. The boxplots of the posterior distributions of the coefficients in ZIP(14) (with and without the intercept boxplot) are shown in Figures 3.4 and 3.5. From the table and boxplots, it is clear that the factors indicating quarter of the year do not appear to affect the reported crime counts. In addition, the “Income_high” factor does not seem to affect crime count. This means that no matter what time of the year the interview was carried out, respondents tend to report the similar numbers of property crime. Also, high income households are not
Figure 3.4: Coefficients in the ZIP(14) model (including intercept).

Figure 3.5: Coefficients in the ZIP(14) model (excluding intercept).
very different from the households with middle income in the number of reported property crime counts. The factors of number of people in each household, whether the previous interview is missing, and whether the household is in the central city of an MSA have a positive effect on the household property crime counts. This means that those households that have more people in them, have missed the previous interview or are in bigger cities, are likely to report more property crimes. Also, the number of property crimes tends to decrease as the sequence of the interview increases. For those households that have relatively low income or are in a rural area, it is less likely that they will report a property crime committed against them. For the “region” factor, households in the midwest and south have basically the same chance to report a crime, but households in the northeast are less likely to report a property crime committed against them; households in the western region report the most property crimes. These conclusions are highly consistent with our discussions in Chapter 2.

After examining the effect of the covariates in the ZIP(14) model, we decided to exclude covariates “quarter” and “Income_high”. In addition, we combine the midwest and south variables in the “region” variable into a new indicator called “midwest_south” which equals 1 if the interviewed household is in either the midwest region or in the southern region of U.S. After making these modifications in the covariate set, we have 9 coefficients corresponding to the intercept and the eight covariates in the mean model of our ZIP model. We accordingly call this new model ZIP(9).

Table 3.4 compares the DIC values for the three ZIP models (ZIP(4), ZIP(14), and ZIP(9)) examined in this section. The ZIP(9) model has a DIC value that is just slightly larger than that of the DIC of the ZIP(14) model. However, the DICs for
Figure 3.6: Coefficients in the ZIP(9) model (including intercept).

Figure 3.7: Coefficients in the ZIP(9) model (excluding intercept).
ZIP(9) and ZIP(14) are both much smaller than the DIC for ZIP(4). Table 3.5 shows the posterior percentiles of coefficients in the ZIP(9) model. Figure 3.6 and Figure 3.7 show the boxplots of the posterior distributions of the coefficients in the ZIP(9) model with and without the intercept boxplot, respectively. From Table 3.5, we can see that none of the middle 95% posterior percentiles of the nine covariates (including the intercept) in the ZIP(9) model contains zero.

### 3.4.3 Comparison of the ZIP model with the ZIP2 model

In this section, we further compare the ZIP model with the ZIP2 model that we discussed in Section 3.2.2. Here we used the ZIP(9) discussed in the previous section,
and compare ZIP(9) with a ZIP2 model that does not have the “Sequence” variable in the mean model but keeps all the remaining 8 covariates (including the intercept) in the mean model. We name this ZIP2 model ZIP2(8). The ZIP2(8) model has one fewer covariate in the mean model than the ZIP(9) model, but in the ZIP2(8) model the point mass at zero can vary by the sequence number of the current interview. In other words, ZIP2(8) has π_{j}'s for the point mass at zero, where j = 1, 2, ..., 7 is the sequence number of the current interview, while ZIP(9) only has one π for the point mass component of the ZIP model for all of the interviews. The π_{j}s in ZIP2(8) allow different probabilities for the point mass at zero in the ZIP model for interviews from the 1st to 7th waves.

The boxplots of the posterior distributions and posterior percentiles for the coefficients in the ZIP(9) model were provided in Figures 3.6, 3.7 and Table 3.5 of the previous section. Table 3.6 shows the posterior percentiles for the coefficients in the mean model of the ZIP2(8) model. Figures 3.8 and 3.9 show the boxplots of the posterior distributions of the coefficients in the mean model of the ZIP2(8) model with and without the intercept plotted, respectively. The table and boxplots show little difference from those for the ZIP(9) coefficients.

The main difference between the ZIP(9) and ZIP2(8) models is the assumption about the point mass at zero in the ZIP model. To further compare these two models, we summarize the posterior percentiles for π in the ZIP(9) model and the π_{j}s in the ZIP2(8) model in Table 3.7. Figure 3.10 shows the posterior boxplots for these estimated parameters. Note that the posterior means of these parameters are very close to the values of their corresponding posterior medians. All of the π and π_{j}s shown in Table 3.7 have posterior distributions that only contain positive numbers,
Table 3.6: Posterior percentiles for coefficients in the mean of the ZIP2(8) model.

<table>
<thead>
<tr>
<th>parameter</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.513</td>
<td>-3.412</td>
<td>-3.309</td>
</tr>
<tr>
<td>Missing</td>
<td>0.006</td>
<td>0.104</td>
<td>0.192</td>
</tr>
<tr>
<td>Income_low</td>
<td>-0.147</td>
<td>-0.087</td>
<td>-0.025</td>
</tr>
<tr>
<td>centralcy</td>
<td>0.322</td>
<td>0.375</td>
<td>0.427</td>
</tr>
<tr>
<td>notMSA</td>
<td>-0.251</td>
<td>-0.176</td>
<td>-0.101</td>
</tr>
<tr>
<td>Northeast</td>
<td>-0.503</td>
<td>-0.427</td>
<td>-0.353</td>
</tr>
<tr>
<td>Midwest_south</td>
<td>-0.211</td>
<td>-0.152</td>
<td>-0.093</td>
</tr>
<tr>
<td>Num_members</td>
<td>0.382</td>
<td>0.403</td>
<td>0.422</td>
</tr>
</tbody>
</table>

Figure 3.8: Coefficients in the ZIP2(8) model (including intercept).
which confirms our assumption of zero-inflation. Also, the boxplots in Figure 3.10 show the rotation group bias effect directly: the \( \pi_j \)s increase as \( j \) increases. In other words, the trend in the locations of the posterior distributions of the \( \pi_j \)s reflects the rotation group bias, and the magnitude of the posterior means of the \( \pi_j \)s quantifies the zero inflation in each wave of interviews. Since we include the 1\textsuperscript{st} wave interviews in our data and models, these posterior distributions also capture the bounding effects: the point mass at zero for the 1\textsuperscript{st} wave interviews (\( \pi_1 \)) is much smaller than the other point masses at zero for the 2\textsuperscript{nd} to 7\textsuperscript{th} waves of interviews.

In the ZIP(9) model, unlike the ZIP2(8) model, the rotation group bias is modeled by the “Sequence” variable which is included in the mean model of ZIP as a numerical covariate. The ZIP(9) model gives a posterior median of \(-0.067\) for the “Sequence” factor, which means that on average the household property crime count for the next interview is equal to \( \exp(-0.067) = 0.935 \) times the property crime counts in the
Table 3.7: Posterior percentiles for the $\pi$ in the ZIP(9) and the $\pi_j$'s in the ZIP2(8) model.

<table>
<thead>
<tr>
<th>parameter</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi$(ZIP(9))</td>
<td>0.291</td>
<td>0.319</td>
<td>0.352</td>
</tr>
<tr>
<td>$\pi_1$(ZIP2(8))</td>
<td>0.026</td>
<td>0.124</td>
<td>0.187</td>
</tr>
<tr>
<td>$\pi_2$(ZIP2(8))</td>
<td>0.218</td>
<td>0.285</td>
<td>0.311</td>
</tr>
<tr>
<td>$\pi_3$(ZIP2(8))</td>
<td>0.317</td>
<td>0.361</td>
<td>0.402</td>
</tr>
<tr>
<td>$\pi_4$(ZIP2(8))</td>
<td>0.342</td>
<td>0.386</td>
<td>0.428</td>
</tr>
<tr>
<td>$\pi_5$(ZIP2(8))</td>
<td>0.373</td>
<td>0.415</td>
<td>0.455</td>
</tr>
<tr>
<td>$\pi_6$(ZIP2(8))</td>
<td>0.374</td>
<td>0.416</td>
<td>0.455</td>
</tr>
<tr>
<td>$\pi_7$(ZIP2(8))</td>
<td>0.395</td>
<td>0.437</td>
<td>0.475</td>
</tr>
</tbody>
</table>

previous interview. The $\pi$ in the ZIP(9) model has a posterior median of 0.319 while the unweighted average of the posterior medians of $\pi_j$'s, $j = 1, 2, ..., 7$, in the ZIP2(8) model is 0.343. This means on average, these two models capture about the same amount of zero inflation in our property crime count data.

Table 3.8 compares the ZIP(9) and ZIP2(8) models by DIC. Neither of the two models significantly outperforms the other one in the sense of DIC.

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZIP(9)</td>
<td>83563</td>
</tr>
<tr>
<td>ZIP2(8)</td>
<td>83533</td>
</tr>
</tbody>
</table>

Table 3.8: Comparison of the ZIP model and ZIP2 model by DIC.

3.5 Problem: predicted household-level property crime rates

Other than comparing the models by their DIC values, we next assess the goodness of the model fit for any of our ZIP or ZIP2 models by checking the predicted property
crime counts to see whether the weighted household property crime rates predicted by our zero-inflated count models agree with the actual property crime rates (Figure 2.2) for the seven waves of interviews.

To calculate the predicted crime counts, we first sample $n_p$ sets of parameters from the posterior distributions. In other words, we randomly select $n_p$ numbers from 1 to 40000 as we have a total of 40000 samples from the posterior distributions. Each of the $n_p$ numbers corresponds to one record in the 40000 samples from the posterior distributions. The $p^{th}$ number in the $n_p$ numbers corresponds to the parameters $(\hat{\beta}_0^{(p)}, \hat{\beta}_1^{(p)}, \hat{\beta}_2^{(p)}, \ldots)$ and $\hat{\pi}^{(p)}$ or $\hat{\pi}_j^{(p)}$ for $p = 1, \ldots, n_p$. We then use the predicted marginal mean for the Poisson and NB models, which means that

$$\hat{y}_{ij}^{(p)} = \exp(\hat{\beta}_0^{(p)} + \hat{\beta}_1^{(p)} X_{ij}^M + \hat{\beta}_2^{(p)} X_{ij}^S + \ldots),$$

(3.56)
for both Poisson and NB count models, where \((\hat{\beta}_0^{(p)}, \hat{\beta}_1^{(p)}, \hat{\beta}_2^{(p)}, \ldots)\) is the \(p^{th}\) set of posterior coefficients, \(p = 1, \ldots, n_p\). The \(n_p\) is set to be 1000 in our calculation.

For ZIP models, we first draw samples from Bernoulli(1 − \(\hat{\pi}^{(p)}\)) or Bernoulli(1 − \(\hat{\pi}_j^{(p)}\)) (for ZIP2 models) distributions, then multiply the \(\hat{y}_{ij}^{(p)}\) in Equation 3.56 by the Bernoulli outcomes. Equation 3.56 needs to be modified accordingly for ZIP2 models as there is no “Sequence” covariate in the mean model of ZIP2 models and the number of \(\hat{\beta}\)s may differ depending on the ZIP or ZIP2 model. The posterior mean of all \(\hat{y}_{ij}^{(p)}, p = 1, \ldots, n_p\), is taken as the final predicted crime count for household \(i\) on the \(j^{th}\) interview.

After obtaining the predicted crime counts, we calculate seven predicted weighted crime rates for each model (one predicted rate for each wave of interviews) using the household weights. Figure 3.11 shows how the predicted crime rates (from 5 different models) compare to the actual reported crime rates. Figure 3.12 shows how the histograms of the predicted household property crime counts (from our five different models) compare to the actual reported crime counts. Table 3.9 shows the predicted property crime counts (from our five different models) distributed by the number of crimes, and how they compare to the percentage of actually reported property crimes in each category. From these figures, we conclude that all of our count models tend to underestimate the property crime counts for the households. The NB model did slightly better than the Poisson model, probably because of its greater ability to account for overdispersion in the data. The ZIP(4) model did slightly better than the Poisson and NB count models because it accounted for the zero inflation in the data, while the Poisson and NB models completely ignore the zero-inflation effect which is an important characteristic of our data. Although the ZIP(4) model does
not include enough covariates in the mean model to capture all the variation in the response variable, the ZIP(9) and ZIP(14) models outperform the ZIP(4) model only in a limited manner. The ZIP models still seriously underestimate the crime count data. The histograms in Figure 3.12 and results in Table 3.9 show that almost all of our models tend to predict more zero counts than what was reported, and give predictions that are less than actual counts at larger numbers of crime counts.

We fit another ZIP2 model that includes the “Sequence” effect in the point mass portion of the model and also includes “Sequence” in the mean model of the Poisson part. Since the only difference between this new model and the ZIP2(8) model is that this model has one more covariate (“Sequence”) in the mean model for the Poisson part, we name this model ZIP2(9). In Figure 3.11, it is hard to distinguish the property crime rate curve predicted by the ZIP2(9) model from those predicted by the ZIP(9) and ZIP2(8) models. The ZIP2(9) model actually weakens the effect of the “Missing” factor in our data, weakening the bounding effect, i.e., making the predicted crime rate for 1st wave interviews smaller than the actual counts.
Figure 3.12: Histograms of the actual and predicted crime rates under five models (binwidth=0.5, left closed and right open intervals).

Table 3.9: Percentages of predicted property crime counts.
As our predicted crime counts are not as good as necessary for our purposes (shown in Figure 3.11 and Table 3.9), we tried a few additional models to find a possible model for better prediction, such as the zero-inflated Negative Binomial (ZINB) model (see Section 3.2.3). Garay et al. (2011) fit a Bayesian ZINB model using the apple cultivar data set (Ridout et al., 2001) and claimed that the ZINB model can fit data better than the ZIP model especially in situations where the zero excess is greater than expected under a ZIP model. In our case, from Table 3.9 we can see that the zero excess in our crime count data is so big that it is very likely to be greater than what is expected under a ZIP model. Thus, it seemed possible that a ZINB model would be a better choice than the ZIP model we previously fit. For the ZINB model, we also tried different combinations of covariates in both the point mass probability model and the NB mean model while only including the random effects in the NB mean model. Still, in contrast to the conclusions in Garay et al. (2011), our ZINB models did not improve the under-estimation in predicted household property crime counts.

### 3.6 Truncated Poisson parameter MLE derivation

To better understand the low bias in estimates from our zero-inflated count models, we derive the MLE of a truncated (non-zero) Poisson distribution and compare the estimated count probabilities to the actual count probabilities from our household property crime count data to see how similar they are to each other. The reason we compare to a truncated Poisson distribution is that the zero-inflated Poisson count models estimate the Poisson parameter mainly depending on the non-zero count data.
For a truncated (non-zero) Poisson, the likelihood function is:

\[
f(x_1, x_2, ..., x_n|\lambda) = \frac{e^{-\lambda \sum x_i}}{\prod x_i!(1 - e^{-\lambda})^n}.
\]

Thus, the loglikelihood function is:

\[
\log f(x_1, x_2, ..., x_n|\lambda) = -n\lambda + (\sum x_i) \log \lambda - \log(\prod x_i!) - n\log(1 - e^{-\lambda}).
\]

After taking the derivative of the log-likelihood function with respect to \(\lambda\) and setting it to zero, we obtained the following equation which may be solved for \(\hat{\lambda}\):

\[
e^{-\hat{\lambda}} + \frac{n}{\sum x_i} \hat{\lambda} - 1 = 0.
\]

We estimate \(\hat{\lambda}\) using the summarized property crime count data based on 23139 households given in Table 3.10.

<table>
<thead>
<tr>
<th># crime count</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td># crime</td>
<td>9451</td>
<td>1147</td>
<td>230</td>
<td>57</td>
<td>21</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3.10: Crime counts summary: max count 6, n=10908 total non-zero reports.

Using the data summarized in Table 3.10, the estimated parameter is: \(\hat{\lambda} = 0.3256\).

We then use this estimate to obtain the expected cell count probabilities. Table 3.11 compares the estimated count probabilities and the actual count probabilities under the truncated Poisson model.

As we expected, the estimated crime count probabilities at high crime counts are much lower than the actual crime count probabilities. This result explains why our
zero-inflated count models underestimate the predicted property crime counts. This result also indicates that the reported property crime count or the sampled households might have two different underlying mechanisms, with some normal households but also some other special households that tend to report extremely high crime counts. These households might have particular reasons for being exposed to more crimes than other households, for example, they might live in special areas or might have had some special events happen to them during the period covered by the interview.

One idea to improve the predictions of our zero-inflated models could be a zero-inflated mixture Poisson model under which, other than the point mass at zero, we also have a mixture Poisson model for the non-zero crime counts. The zero-inflated mixture Poisson model can be represented as follows:

\[
P(Y_{ij} = 0) = (1 - \pi_{ij}^{(1)} - \pi_{ij}^{(2)}) + \pi_{ij}^{(1)} \exp(-\mu_{ij}^{(1)}) + \pi_{ij}^{(2)} \exp(-\mu_{ij}^{(2)}) \tag{3.60}
\]

\[
P(Y_{ij} = k) = \pi_{ij}^{(1)} \left(\frac{\mu_{ij}^{(1)} k}{k!}\right) \exp(-\mu_{ij}^{(1)}) + \pi_{ij}^{(2)} \left(\frac{\mu_{ij}^{(2)} k}{k!}\right) \exp(-\mu_{ij}^{(2)}) \tag{3.61}
\]

where \(\pi_{ij}^{(1)}\) and \(\pi_{ij}^{(2)}\) are the probabilities for the two Poisson distributions and \(\mu_{ij}^{(1)}\) and \(\mu_{ij}^{(2)}\) are the two Poisson means. \(\mu_{ij}^{(1)}\) and \(\mu_{ij}^{(2)}\) may or may not depend on the

<table>
<thead>
<tr>
<th></th>
<th>estimated</th>
<th>actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(x=1)</td>
<td>0.8457</td>
<td>0.8664</td>
</tr>
<tr>
<td>P(x=2)</td>
<td>0.1377</td>
<td>0.1052</td>
</tr>
<tr>
<td>P(x=3)</td>
<td>0.0149</td>
<td>0.0211</td>
</tr>
<tr>
<td>P(x=4)</td>
<td>0.0012</td>
<td>0.0052</td>
</tr>
<tr>
<td>P(x=5)</td>
<td>0.00008</td>
<td>0.00193</td>
</tr>
<tr>
<td>P(x=6)</td>
<td>0.0000043</td>
<td>0.0001834</td>
</tr>
</tbody>
</table>

Table 3.11: Comparison of the estimated count probabilities and the actual count probabilities.
household index $i$ and interview sequence $j$ depending on the assumption we make on the models of the Poisson means.

A main problem with the mixture Poisson model is the instability of parameter estimation which makes the model difficult to use for estimation and prediction. New parameter estimation methods, which can lead to stable parameter estimation, or other forms of count models, may be considered for future studies.

Since the models in the current chapter failed to provide an acceptable fit to the property crime count data, we will consider another type of models in Chapter 4. In Chapter 3, each record in the data contains information of only one interview from one household. The sequence of the interview was taken as one of the covariates and the corresponding estimated coefficient quantifies how much the crime count decreases as the respondents are interviewed each additional time. However, as the NCVS has repeated interviews with the same group of households for up to seven times, the crime counts from the seven interviews should be viewed as longitudinal data with length of seven. The correlations between the seven interviews mainly come from the fact that the same households are likely to have similar experiences from one interview to the next. In other words, the correlations between the seven interviews from the same household may not be fully accounted for by the covariates. This is the reason we include a random effect for each interviewed household in addition to the covariates in our zero-inflated GLMMs in Chapter 3. The random effect in the mean of the zero-inflated count models accounts for the effect of each individual household and remains the same for each household through all seven interviews. In Chapter 4, we take another view of the longitudinal data from the same group of households. Instead of viewing every interview of a household as one record, we will take the
seven interviews from the same group of households as one piece of data and look for models that can properly describe these longitudinal data vectors.
Chapter 4: Multinomial count models

As introduced at the beginning of Chapter 3, one of the important features of the property crime count data is that the count data includes seven repeated measurements for the same group of households. This means that we are, in fact, dealing with longitudinal data vectors of length seven as the observation. Therefore, a more realistic model needs to be able to model data vectors instead of scalar data.

Multinomial count models are frequently used for describing categorical dependent variables with more than two response categories. The data are often aggregated into the number of individuals in each category. Given a fixed sample size, the total number of individuals in the sample is distributed into each category according to a vector of probabilities. We propose models that are based on those in the existing literature. For example, Stasny (1986) provides an example applying multinomial count models to complicated survey data with missing observations. Diggle et al. (2002) and Molenberghs and Verbeke (2005) also illustrate applications of multinomial models on discrete longitudinal data. Yu et al. (2008) use multinomial Bayesian models to study the response bias in NCVS data for estimating rape and domestic violence rates.

In the longitudinal property crime count vector data, we count the number of reported crimes from all households in each of the seven interviews. Each of the
seven interviews can be viewed as one category and there is a probability vector, also of length seven, to tell us how the total number of reported crimes will be distributed into each of the interview times. One problem is that the total number of reported crimes is not known before the interviews are completed. Thus, we are here making an assumption that the models in this chapter are built after completing all seven interviews for the households in the data. We note that for our goal of using the fitted model to adjust for the bounding and time-in-sample effects for future samples in the NCVS, it is reasonable to build models based on the complete seven-interview data.

It is natural to think that households with the similar features tend to have similar crime-count probability vectors. In other words, instead of having a separate set of multinomial probabilities for each interviewed household, we would like to relate the multinomial probabilities to the covariates such as the demographic information of each household. However, in multinomial count models, the probabilities are usually not explicitly expressed by a linear or non-linear formula of the covariates. An alternate method is to have a set of multinomial probabilities for every combination of the covariates. This leads to a small domain multinomial count model with each combination of the important covariates defining one of the small non-overlapping domains (Section 4.1).

To address the problem of missing interviews under the multinomial setting, we further divide the households by the number of reported property crimes in each interview (Section 4.2). Thus, in each interview, we now have 8 sub-categories: reporting zero crimes, reporting one crime, reporting two crimes, reporting three crimes, reporting four crimes, reporting five crimes, reporting six crimes and missing the current
interview. Categorizing the data by both interview sequence and the number of reported crimes will result in having one contingency table for each small domain. In this case, corresponding to each contingency table, we will have a probability table for each domain.

4.1 Reported crimes in each of seven interviews

As previously described, we first divide the household property crime data into small domains defined by different combinations of the covariates to study the effect of covariates on the multinomial probabilities. We divide all the households into 18 groups by their region (midwest or south, north and west), MSA status (central city, small city and rural area) and income level (low and not low). A list of the 18 domains is shown in Table 4.1. The total number of crimes reported by each of the 18 groups is further categorized into 7 interviews making the final data an 18 by 7 matrix. This small domain data only contains those households with all seven interviews’ data (no missing interviews). Each row in our data represents the total crime counts in seven interviews for all responding households in one of the 18 small domains. The observed data are shown in Table 4.2. In Table 4.2, domain 14 has small crime counts since there are only 13 households in that domain. The small sample size in domain 14 may affect the goodness of fit of the models.

4.1.1 Small domain multinomial models and model fitting results

The multinomial crime count model for the 18 small domains by 7 interviews data can be described as below:
Table 4.1: List of small domains.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Region</th>
<th>MSA</th>
<th>Income</th>
<th># of HHs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Northeast</td>
<td>Central city</td>
<td>not low</td>
<td>371</td>
</tr>
<tr>
<td>2</td>
<td>Northeast</td>
<td>Central city</td>
<td>low</td>
<td>144</td>
</tr>
<tr>
<td>3</td>
<td>Midwest or South</td>
<td>Central city</td>
<td>not low</td>
<td>1657</td>
</tr>
<tr>
<td>4</td>
<td>Midwest or South</td>
<td>Central city</td>
<td>low</td>
<td>430</td>
</tr>
<tr>
<td>5</td>
<td>West</td>
<td>Central city</td>
<td>not low</td>
<td>980</td>
</tr>
<tr>
<td>6</td>
<td>West</td>
<td>Central city</td>
<td>low</td>
<td>207</td>
</tr>
<tr>
<td>7</td>
<td>Northeast</td>
<td>small city</td>
<td>not low</td>
<td>902</td>
</tr>
<tr>
<td>8</td>
<td>Northeast</td>
<td>small city</td>
<td>low</td>
<td>68</td>
</tr>
<tr>
<td>9</td>
<td>Midwest or South</td>
<td>small city</td>
<td>not low</td>
<td>2813</td>
</tr>
<tr>
<td>10</td>
<td>Midwest or South</td>
<td>small city</td>
<td>low</td>
<td>388</td>
</tr>
<tr>
<td>11</td>
<td>West</td>
<td>small city</td>
<td>not low</td>
<td>1301</td>
</tr>
<tr>
<td>12</td>
<td>West</td>
<td>small city</td>
<td>low</td>
<td>178</td>
</tr>
<tr>
<td>13</td>
<td>Northeast</td>
<td>rural area</td>
<td>not low</td>
<td>105</td>
</tr>
<tr>
<td>14</td>
<td>Northeast</td>
<td>rural area</td>
<td>low</td>
<td>13</td>
</tr>
<tr>
<td>15</td>
<td>Midwest or South</td>
<td>rural area</td>
<td>not low</td>
<td>639</td>
</tr>
<tr>
<td>16</td>
<td>Midwest or South</td>
<td>rural area</td>
<td>low</td>
<td>245</td>
</tr>
<tr>
<td>17</td>
<td>West</td>
<td>rural area</td>
<td>not low</td>
<td>165</td>
</tr>
<tr>
<td>18</td>
<td>West</td>
<td>rural area</td>
<td>low</td>
<td>39</td>
</tr>
</tbody>
</table>

For each domain $d$, we assume that:

$$(y^d_1, y^d_2, y^d_3, y^d_4, y^d_5, y^d_6, y^d_7) \sim \text{multinomial}((\alpha p^d_2, p^d_2, \beta d^d p^d_2, (\beta d^d)^2 p^d_2, (\beta d^d)^3 p^d_2, (\beta d^d)^4 p^d_2, (\beta d^d)^5 p^d_2, \sum_i y^d_i), \quad (4.1)$$

or

$$(y^d_1, y^d_2, y^d_3, y^d_4, y^d_5, y^d_6, y^d_7) \sim \text{multinomial}((\alpha d^d p^d_2, p^d_2, \beta d^d p^d_2, \beta d^d p^d_2, \beta d^d p^d_2, \beta d^d p^d_2, \beta d^d p^d_2, \sum_i y^d_i), \quad (4.2)$$

where $y^d_i$ is the total number of reported property crimes in the $i^{th}$ interview of domain $d$. 

72
<table>
<thead>
<tr>
<th>Domain</th>
<th>$1^{st}$</th>
<th>$2^{nd}$</th>
<th>$3^{rd}$</th>
<th>$4^{th}$</th>
<th>$5^{th}$</th>
<th>$6^{th}$</th>
<th>$7^{th}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73</td>
<td>58</td>
<td>56</td>
<td>47</td>
<td>45</td>
<td>51</td>
<td>41</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>21</td>
<td>21</td>
<td>19</td>
<td>27</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>323</td>
<td>269</td>
<td>178</td>
<td>238</td>
<td>221</td>
<td>196</td>
<td>232</td>
</tr>
<tr>
<td>4</td>
<td>94</td>
<td>82</td>
<td>53</td>
<td>43</td>
<td>56</td>
<td>60</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>201</td>
<td>165</td>
<td>130</td>
<td>121</td>
<td>116</td>
<td>128</td>
<td>119</td>
</tr>
<tr>
<td>6</td>
<td>34</td>
<td>24</td>
<td>35</td>
<td>33</td>
<td>32</td>
<td>28</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>188</td>
<td>156</td>
<td>135</td>
<td>107</td>
<td>116</td>
<td>115</td>
<td>85</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>10</td>
<td>5</td>
<td>15</td>
<td>6</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>583</td>
<td>465</td>
<td>408</td>
<td>374</td>
<td>326</td>
<td>350</td>
<td>307</td>
</tr>
<tr>
<td>10</td>
<td>73</td>
<td>61</td>
<td>49</td>
<td>44</td>
<td>51</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>11</td>
<td>284</td>
<td>180</td>
<td>185</td>
<td>163</td>
<td>145</td>
<td>167</td>
<td>177</td>
</tr>
<tr>
<td>12</td>
<td>33</td>
<td>19</td>
<td>31</td>
<td>29</td>
<td>17</td>
<td>27</td>
<td>22</td>
</tr>
<tr>
<td>13</td>
<td>19</td>
<td>24</td>
<td>20</td>
<td>10</td>
<td>12</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>125</td>
<td>95</td>
<td>95</td>
<td>78</td>
<td>86</td>
<td>88</td>
<td>72</td>
</tr>
<tr>
<td>16</td>
<td>68</td>
<td>44</td>
<td>31</td>
<td>29</td>
<td>32</td>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>17</td>
<td>52</td>
<td>15</td>
<td>21</td>
<td>21</td>
<td>25</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4.2: Observed small domain crime counts by interview period.

For Model 4.1, we assume that the 18 small domains are independent of each other and each of them has its own parameters: $\alpha^d, p_d^2$ and $\beta^d$, where $d = 1, 2, ..., 18$ is the index of the domain. $\alpha^d$ and $\beta^d$ describe the bounding effect and the time-in-sample effect for domain $d$, respectively. The $p_d^2$ stands for the probability that a reported crime occurred in the second interview for domain $d$. $p_d^2$ is taken as a base probability since we are taking the second interview to be the closest to “truth” and we want to compare probabilities in other interviews with the probability in the second interview. $\alpha^d$ is expected to be greater than 1 as the bounding effect is thought to result in the households over-reporting crime in their first interviews. We expect $\beta^d$ to be less than 1 to be consistent with the fact that the households report fewer and fewer crimes as
they have more and more interviews. Another assumption we have made in Model 4.1 is that the time-in-sample effect results in the probability of a reported crime being in each interview to be decreasing with the same coefficient $\beta^d$ for each domain. In other words, starting from the second interview, the probability of reporting a crime for each interview is $\beta^d$ times the probability in the previous interview and $\beta^d$ is a fixed constant expected to be between 0 and 1 for each domain.

For Model 4.2, we again assume that each of the 18 small domains has its own parameter for the bounding effect, $\alpha^d$, and for the probability of being in the second interview, $p^d_2$. For the time-in-sample effect, instead of having a fixed decreasing rate for each domain, we assume that the decreasing rate can be different from interview to interview but the five decreasing rates are the same for all domains, i.e., instead of having just one $\beta^d$ we now have five $\beta$-parameters: $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$, but the $\beta$-parameters no longer differ by domain, $d$. Next, we will use MLEs to estimate the parameters in Model 4.1 and Model 4.2.

For Model 4.1, we must solve the following system of equations:

\[
\frac{5(N^d - y^d_1)}{a^d} - 1)(\hat{\beta}^d)^5 + \left(\frac{4(N^d - y^d_1)}{a^d} - 1\right)(\hat{\beta}^d)^4 + \left(\frac{3(N^d - y^d_1)}{a^d} - 1\right)(\hat{\beta}^d)^3 \\
+ \left(\frac{2(N^d - y^d_1)}{a^d} - 1\right)(\hat{\beta}^d)^2 + \left(\frac{N^d - y^d_1}{a^d} - 1\right)\hat{\beta}^d - 1 = 0 \quad (4.3)
\]

\[
\hat{p}^d_2 \ast [((\hat{\beta}^d)^5 + (\hat{\beta}^d)^4 + (\hat{\beta}^d)^3 + (\hat{\beta}^d)^2 + \hat{\beta}^d + 1] + \frac{y^d_1}{N^d} - 1 = 0 \quad (4.4)
\]

\[
\hat{\alpha}^d = \frac{y^d_1}{N^d} \ast \frac{1}{\hat{p}^d_2} \quad (4.5)
\]

where $N^d = \sum y^d_i$ and $a^d = y^d_3 + 2y^d_4 + 3y^d_5 + 4y^d_6 + 5y^d_7$ are the sufficient statistics and $d = 1, 2, ..., 18$. 
Model 4.2 has direct (closed-form) MLEs for the parameter estimators. For Model 4.2, the MLEs are:

\[ \hat{\beta}_1 = \frac{\sum_d y_3^d}{\sum_d y_2^d}, \quad \hat{\beta}_2 = \frac{\sum_d y_4^d}{\sum_d y_2^d}, \quad \hat{\beta}_3 = \frac{\sum_d y_5^d}{\sum_d y_2^d}, \quad \hat{\beta}_4 = \frac{\sum_d y_6^d}{\sum_d y_2^d}, \quad \hat{\beta}_5 = \frac{\sum_d y_7^d}{\sum_d y_2^d}, \]

\[ \hat{p}_2^d = \frac{(\sum_i y_i^d - y_1^d) \sum_d y_2^d}{(\sum_d \sum_i y_i^d - \sum_d y_1^d) \sum_i y_1^d}, \]

\[ \hat{\alpha}_1^d = \frac{(\sum_d \sum_i y_i^d - \sum_d y_1^d) y_1^d}{(\sum_i y_i^d - y_1^d) \sum_d y_2^d} \] (4.6)

The estimated values for the parameters and their standard errors for Model 4.1 are shown in Table 4.3. The standard errors of the estimated parameters are calculated as the inverse of the square root of the diagonal elements of the Fisher Information matrix. Also shown in Table 4.3 are the $X^2$ and $G^2$ statistics with 3 degrees of freedom for each domain. Because in Model 4.1 all the domains are assumed to be independent of each other with respect to parameter estimation, we can simply calculate the $X^2$ and $G^2$ statistics for each domain separately. Although the complex survey data do not satisfy the independence assumptions that are required to use the $X^2$ and $G^2$ statistics as measure of goodness of fit (Stasny, 1986), we are only using them in the current study as summary measures to make model comparisons.

From Table 4.3, we can see that domains 13 (Northeast, rural area and not low income) and 14 (Northeast, rural area and low income) have high $\hat{p}_2$ values and low $\hat{\beta}$ and $\hat{\alpha}$ (lower than 1) values, which means small/no bounding effect, large time-in-sample effect and high crime rates. For domain 17 (West, rural and not low income), we have high $\hat{\alpha}$ (greater than 2) and $\hat{\beta}$ (close to 1) values and low $\hat{p}_2$ value, which means small time-in-sample effect, large bounding effect and low crime rate. Domains
2 and 18 also have $\hat{\alpha}$ values lower than 1, which means there does not appear to be a bounding effect in those domains.

<table>
<thead>
<tr>
<th>Domain</th>
<th>$\hat{\beta}^d$</th>
<th>SE($\hat{\beta}^d$)</th>
<th>$\hat{\rho}_2^d$</th>
<th>SE($\hat{\rho}_2^d$)</th>
<th>$\hat{\alpha}^d$</th>
<th>SE($\hat{\alpha}^d$)</th>
<th>$X^2$</th>
<th>$G^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.943</td>
<td>0.022</td>
<td>0.154</td>
<td>0.008</td>
<td>1.275</td>
<td>0.149</td>
<td>1.366</td>
<td>1.349</td>
</tr>
<tr>
<td>2</td>
<td>0.982</td>
<td>0.035</td>
<td>0.151</td>
<td>0.013</td>
<td>0.873</td>
<td>0.200</td>
<td>2.452</td>
<td>2.291</td>
</tr>
<tr>
<td>3</td>
<td>0.981</td>
<td>0.011</td>
<td>0.141</td>
<td>0.003</td>
<td>1.386</td>
<td>0.077</td>
<td>21.356</td>
<td>22.012</td>
</tr>
<tr>
<td>4</td>
<td>0.918</td>
<td>0.021</td>
<td>0.159</td>
<td>0.008</td>
<td>1.371</td>
<td>0.141</td>
<td>10.930</td>
<td>11.060</td>
</tr>
<tr>
<td>5</td>
<td>0.948</td>
<td>0.014</td>
<td>0.151</td>
<td>0.005</td>
<td>1.361</td>
<td>0.096</td>
<td>5.493</td>
<td>5.470</td>
</tr>
<tr>
<td>6</td>
<td>0.964</td>
<td>0.029</td>
<td>0.152</td>
<td>0.011</td>
<td>1.078</td>
<td>0.185</td>
<td>4.532</td>
<td>4.702</td>
</tr>
<tr>
<td>7</td>
<td>0.907</td>
<td>0.014</td>
<td>0.166</td>
<td>0.006</td>
<td>1.254</td>
<td>0.091</td>
<td>4.918</td>
<td>4.948</td>
</tr>
<tr>
<td>8</td>
<td>0.970</td>
<td>0.054</td>
<td>0.135</td>
<td>0.016</td>
<td>1.857</td>
<td>0.450</td>
<td>8.938</td>
<td>8.612</td>
</tr>
<tr>
<td>9</td>
<td>0.925</td>
<td>0.008</td>
<td>0.159</td>
<td>0.003</td>
<td>1.302</td>
<td>0.054</td>
<td>4.797</td>
<td>4.820</td>
</tr>
<tr>
<td>10</td>
<td>1.000</td>
<td>0.022</td>
<td>0.135</td>
<td>0.007</td>
<td>1.390</td>
<td>0.163</td>
<td>4.219</td>
<td>4.192</td>
</tr>
<tr>
<td>11</td>
<td>0.985</td>
<td>0.012</td>
<td>0.135</td>
<td>0.004</td>
<td>1.161</td>
<td>0.096</td>
<td>5.603</td>
<td>5.718</td>
</tr>
<tr>
<td>12</td>
<td>0.989</td>
<td>0.032</td>
<td>0.139</td>
<td>0.010</td>
<td>1.330</td>
<td>0.231</td>
<td>6.563</td>
<td>6.719</td>
</tr>
<tr>
<td>13</td>
<td>0.812</td>
<td>0.042</td>
<td>0.216</td>
<td>0.021</td>
<td>0.837</td>
<td>0.192</td>
<td>2.322</td>
<td>2.504</td>
</tr>
<tr>
<td>14</td>
<td>0.865</td>
<td>0.111</td>
<td>0.214</td>
<td>0.059</td>
<td>0.359</td>
<td>0.359</td>
<td>1.893</td>
<td>2.054</td>
</tr>
<tr>
<td>15</td>
<td>0.958</td>
<td>0.017</td>
<td>0.149</td>
<td>0.006</td>
<td>1.315</td>
<td>0.118</td>
<td>2.292</td>
<td>2.308</td>
</tr>
<tr>
<td>16</td>
<td>0.873</td>
<td>0.029</td>
<td>0.165</td>
<td>0.011</td>
<td>1.685</td>
<td>0.204</td>
<td>4.309</td>
<td>4.418</td>
</tr>
<tr>
<td>17</td>
<td>1.000</td>
<td>0.036</td>
<td>0.114</td>
<td>0.009</td>
<td>2.761</td>
<td>0.383</td>
<td>8.965</td>
<td>10.029</td>
</tr>
<tr>
<td>18</td>
<td>0.920</td>
<td>0.065</td>
<td>0.183</td>
<td>0.029</td>
<td>0.562</td>
<td>0.281</td>
<td>2.939</td>
<td>2.766</td>
</tr>
</tbody>
</table>

Table 4.3: Estimated parameters, standard errors, $X^2$ and $G^2$ statistics for Model 4.1.

The $X^2$ and $G^2$ statistics suggest that the model for domains 3, 4, 8 and 17 is inconsistent with the observed count data. The multinomial Model 4.1 fits different domains differently. The fit of the model is good for most of the domains but not as good for domains 3, 4, 8 and 17.

The estimated values and standard errors of parameters for Model 4.2 are shown in Table 4.4. The estimated parameters values and standard errors for the $\hat{\rho}_2^d$s and
<table>
<thead>
<tr>
<th>Domain</th>
<th>Parameter 1</th>
<th>MLE</th>
<th>SE</th>
<th>Parameter 2</th>
<th>MLE</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\hat{p}_2^d$</td>
<td>0.161</td>
<td>0.008</td>
<td>$\hat{\alpha}^d$</td>
<td>1.220</td>
<td>0.143</td>
</tr>
<tr>
<td>2</td>
<td>0.174</td>
<td>0.015</td>
<td>0.757</td>
<td>0.174</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.162</td>
<td>0.004</td>
<td>1.206</td>
<td>0.067</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.157</td>
<td>0.008</td>
<td>1.394</td>
<td>0.144</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.160</td>
<td>0.005</td>
<td>1.285</td>
<td>0.091</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.168</td>
<td>0.012</td>
<td>0.979</td>
<td>0.168</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.159</td>
<td>0.005</td>
<td>1.312</td>
<td>0.096</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.151</td>
<td>0.018</td>
<td>1.661</td>
<td>0.403</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.159</td>
<td>0.003</td>
<td>1.302</td>
<td>0.054</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.163</td>
<td>0.008</td>
<td>1.154</td>
<td>0.135</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.157</td>
<td>0.004</td>
<td>1.391</td>
<td>0.083</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.164</td>
<td>0.012</td>
<td>1.134</td>
<td>0.197</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.164</td>
<td>0.016</td>
<td>1.101</td>
<td>0.252</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.185</td>
<td>0.051</td>
<td>0.415</td>
<td>0.415</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.161</td>
<td>0.006</td>
<td>1.212</td>
<td>0.108</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.145</td>
<td>0.009</td>
<td>1.914</td>
<td>0.232</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.137</td>
<td>0.011</td>
<td>2.292</td>
<td>0.318</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.180</td>
<td>0.029</td>
<td>0.569</td>
<td>0.285</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All $\hat{\beta}_1$ 0.863 0.023
All $\hat{\beta}_2$ 0.812 0.022
All $\hat{\beta}_3$ 0.778 0.021
All $\hat{\beta}_4$ 0.783 0.021
All $\hat{\beta}_5$ 0.746 0.021

Table 4.4: Estimated parameters and standard errors for Model 4.2.
\( \hat{\alpha}_d \)s in Table 4.4 show little difference from the results for Model 4.1. For Model 4.2, we have: \( X^2 = 123.569 \) and \( G^2 = 125.359 \). With 67 degrees of freedom for Model 4.2, these \( X^2 \) and \( G^2 \) statistics show that our small domain crime count Model 4.2 is not quite a good fit to the observed small domain crime count data.

Figure 4.1 shows the time series plots for the crime count data of the 18 small domains. From Figure 4.1 we can see a lot of zigzag-shaped perturbations across time. The models we have proposed thus far cannot capture this sort of variation in the data. To deal with this kind of fluctuation and also to try to find a better model for our data, we need to have some kind of smoothing over all domains. The Bayesian version of the small domain multinomial model with a common Dirichlet prior for the probabilities in all domains might be promising.
4.1.2 Small domain Bayesian multinomial model and model fitting results

The multinomial Bayesian model can be described as:

\[
(y_1^d, y_2^d, y_3^d, y_4^d, y_5^d, y_6^d, y_7^d) \sim \text{multinomial}(p_1^d, p_2^d, p_3^d, p_4^d, p_5^d, p_6^d, p_7^d, \sum_i y_i^d)
\]

\[
(p_1^d, p_2^d, p_3^d, p_4^d, p_5^d, p_6^d, p_7^d) \sim \text{dirch}((\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7)), 
\]

for \(d = 1, 2, \ldots, 18\). In addition, we assume \(\alpha_t = \beta_0 + \beta_1 t\) for \(t = 1, 2, \ldots, 7\) to measure the time-in-sample effect with parameter \(\beta_1\). Here, we directly model only the time-in-sample effect and leave the bounding effect to the random part of the model as a simple start at developing our model.

The above multinomial Bayesian model assumes that the total crime counts for the seven interviews in each domain follow a multinomial distribution. The crime-reporting multinomial probability vectors for all 18 domains share a common Dirichlet prior with parameters \(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7\). Furthermore, \(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7\) change with time \(t\) at the same rate \((\beta_1)\) for each interview.

We use the MCMC code from the “rjags” package in R (Plummer, 2015) to obtain the posterior distributions for \(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \beta_0, \beta_1\) and all the crime-reporting probabilities for all 18 domains: \((p_1^1, p_2^1, p_3^1, p_4^1, p_5^1, p_6^1, p_7^1)\) to \((p_1^{18}, p_2^{18}, p_3^{18}, p_4^{18}, p_5^{18}, p_6^{18}, p_7^{18})\). We sampled the posteriors every five steps for a total of 40000 samples with 10000 burn-in steps. In the MCMC approach, we used \(N(0, 10^{10})\) as non-informative priors for \(\beta_0\) and \(\beta_1\).

Posterior quantiles for the \(\beta\)'s and \(\alpha\)'s from fitting this multinomial Bayesian model are summarized in Table 4.5. We have omitted listing all the posterior quantiles for the 18 by 7 \(p_t^d\)'s, where \(t = 1, 2, \ldots, 7\) and \(d = 1, 2, \ldots, 18\), for the sake of brevity.
### Table 4.5: Posterior quantiles for parameters in the multinomial Bayesian small domain model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>45.551</td>
<td>80.159</td>
<td>150.598</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>42.753</td>
<td>75.036</td>
<td>140.871</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>39.894</td>
<td>69.953</td>
<td>131.208</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>37.026</td>
<td>64.858</td>
<td>121.597</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>34.064</td>
<td>59.772</td>
<td>112.066</td>
</tr>
<tr>
<td>$\alpha_6$</td>
<td>31.082</td>
<td>54.641</td>
<td>102.316</td>
</tr>
<tr>
<td>$\alpha_7$</td>
<td>28.065</td>
<td>49.576</td>
<td>92.875</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>48.401</td>
<td>85.286</td>
<td>160.439</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-9.778</td>
<td>-5.100</td>
<td>-2.741</td>
</tr>
</tbody>
</table>

From Table 4.5, we can see that there is a strictly decreasing order from $\hat{\alpha}_1$ to $\hat{\alpha}_7$ and the sign of $\hat{\beta}_1$ is negative. Thus, we can conclude that the total crime counts for the seven interviews for all the domains will strictly decrease as the NCVS interview sequence increases from 1 to 7. The posterior quantiles in Table 4.5 strongly suggest that all the parameters that were estimated in this model are different from zero. These results are consistent with our expectation for the time-in-sample effect.

Using the posterior means of the $p_t^d$'s, $t = 1, 2, ..., 7$ and $d = 1, 2, ..., 18$, we can calculate the expected total crime counts for each of the 18 domains in each of the seven interviews. The expected total crime counts under Model 4.7 are listed in Table 4.6.

Similarly to previous models, we calculated the $X^2$ and $G^2$ statistics for the Bayesian small domain model by comparing the observed (Table 4.2) and expected (Table 4.6) crime counts. Although the $X^2$ and $G^2$ statistics are not typically used as measurements for goodness of fit of Bayesian models, we are using them here to
Table 4.6: Expected small domain data by sequence of interviews under Model 4.7.

compare the goodness of fit of the Bayesian model with the previous non-Bayesian models. The resulting $X^2$ and $G^2$ statistics for the Bayesian small domain model are 72.543 and 73.346, respectively, which is much smaller than those of Model 4.1 and Model 4.2. So the Bayesian small domain model did account for more variation in the original observed data. We will use this model for adjusting the yearly estimated property crime rates in the next chapter.

Although the above Bayesian small domain model gives a relatively good fit to our observed total crime count data, one major drawback from the model is that it fails to include the data from those households that have missing interviews, so victimization
information from households with missing interviews is wasted and ignored. In the next section, we will allow for missingness in the multinomial models.

4.2 Modeling numbers of reported crimes in each of the seven interviews including missing interviews

4.2.1 Full data including missing interviews

4730 of the 23139 NCVS households in our data have missing interviews. The missing data patterns of all the households in our data are complicated. In fact, our data shows all the 64 (2⁶, as we only use records that have the first interviews) possible missingness patterns. A summary of the missing data patterns in our data is shown in Table 4.7. In the second column of Table 4.7, we show the missing data patterns by a series of seven numbers consisted of zeros and ones, where the zeros stand for non-missing interviews and ones stand for missing interviews. For example, a missing-pattern of “0000001” means that we have the data for the first six interviews of the household but the seventh interview is missing.

From Table 4.7, we can see that although our data contains all of the 64 possible missing-data patterns, almost 80% of the households have no missing interviews over the period of 3½ years. Another 10% of the households are only missing one out of the seven interviews. Other than the non-missing pattern, only eight of the remaining 63 patterns contain more than 1% of our data. We do want to include the households with missing interviews in our model, however, since ignoring 20% of the data could result in a considerable amount of bias in our results.

<table>
<thead>
<tr>
<th>missing data pattern</th>
<th># of HHs</th>
<th>proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>no missing</td>
<td>18409</td>
</tr>
</tbody>
</table>

Table 4.7 – Continued on next page

82
<table>
<thead>
<tr>
<th>missing data pattern</th>
<th># of HHs</th>
<th>proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000001</td>
<td>822</td>
<td>0.0355245</td>
</tr>
<tr>
<td>00000010</td>
<td>437</td>
<td>0.0188858</td>
</tr>
<tr>
<td>00000111</td>
<td>405</td>
<td>0.0175029</td>
</tr>
<tr>
<td>00001000</td>
<td>364</td>
<td>0.0157311</td>
</tr>
<tr>
<td>00010000</td>
<td>320</td>
<td>0.0138294</td>
</tr>
<tr>
<td>00100000</td>
<td>319</td>
<td>0.0137863</td>
</tr>
<tr>
<td>00001111</td>
<td>298</td>
<td>0.0128787</td>
</tr>
<tr>
<td>01000000</td>
<td>272</td>
<td>0.0117550</td>
</tr>
<tr>
<td>00011111</td>
<td>212</td>
<td>0.0091620</td>
</tr>
<tr>
<td>01111111</td>
<td>201</td>
<td>0.0086867</td>
</tr>
<tr>
<td>00111111</td>
<td>199</td>
<td>0.0086002</td>
</tr>
<tr>
<td>00001110</td>
<td>77</td>
<td>0.0033277</td>
</tr>
<tr>
<td>00011100</td>
<td>51</td>
<td>0.0022040</td>
</tr>
<tr>
<td>00001011</td>
<td>50</td>
<td>0.0021609</td>
</tr>
<tr>
<td>00010011</td>
<td>42</td>
<td>0.0018151</td>
</tr>
<tr>
<td>00110000</td>
<td>41</td>
<td>0.0017719</td>
</tr>
<tr>
<td>00011110</td>
<td>39</td>
<td>0.0016855</td>
</tr>
<tr>
<td>01100000</td>
<td>37</td>
<td>0.0015990</td>
</tr>
<tr>
<td>00010100</td>
<td>30</td>
<td>0.0012965</td>
</tr>
<tr>
<td>00010111</td>
<td>29</td>
<td>0.0012533</td>
</tr>
<tr>
<td>01010000</td>
<td>29</td>
<td>0.0012533</td>
</tr>
<tr>
<td>00101000</td>
<td>28</td>
<td>0.0012101</td>
</tr>
<tr>
<td>00100001</td>
<td>25</td>
<td>0.0010804</td>
</tr>
<tr>
<td>00100100</td>
<td>22</td>
<td>0.0009508</td>
</tr>
<tr>
<td>00101111</td>
<td>22</td>
<td>0.0009508</td>
</tr>
<tr>
<td>00111100</td>
<td>21</td>
<td>0.0009075</td>
</tr>
<tr>
<td>00011101</td>
<td>20</td>
<td>0.0008644</td>
</tr>
<tr>
<td>01000001</td>
<td>20</td>
<td>0.0008643</td>
</tr>
<tr>
<td>01000100</td>
<td>20</td>
<td>0.0008644</td>
</tr>
<tr>
<td>01011111</td>
<td>20</td>
<td>0.0008643</td>
</tr>
<tr>
<td>00111110</td>
<td>19</td>
<td>0.0008211</td>
</tr>
<tr>
<td>01110000</td>
<td>19</td>
<td>0.0008211</td>
</tr>
<tr>
<td>01111110</td>
<td>19</td>
<td>0.0008212</td>
</tr>
<tr>
<td>01000111</td>
<td>16</td>
<td>0.0006914</td>
</tr>
<tr>
<td>01010100</td>
<td>15</td>
<td>0.0006483</td>
</tr>
<tr>
<td>00100111</td>
<td>14</td>
<td>0.0006050</td>
</tr>
<tr>
<td>00110111</td>
<td>12</td>
<td>0.0005186</td>
</tr>
<tr>
<td>01111011</td>
<td>12</td>
<td>0.0005186</td>
</tr>
<tr>
<td>01001000</td>
<td>10</td>
<td>0.0004322</td>
</tr>
<tr>
<td>01101111</td>
<td>10</td>
<td>0.0004322</td>
</tr>
</tbody>
</table>

Table 4.7 – Continued on next page
Table 4.7: Missing patterns of household-level data.

To deal with the missing data problem, we first summarize our data as shown in Table 4.8, then look for appropriate statistical models to describe the data in that form.

In Table 4.8, we organized the data by both the sequence of the interview and the number of reported crimes. Each row in Table 4.8 represents one interview time, so each row of the table sums to a total of 23139 households. For example, in the first interview (the first row in Table 4.8), 20768 out of 23139 households reported zero
Table 4.8: Number of households by number of reported crimes and interview sequence.

<table>
<thead>
<tr>
<th>Seq</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20768</td>
<td>2001</td>
<td>275</td>
<td>68</td>
<td>20</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>20604</td>
<td>1506</td>
<td>194</td>
<td>37</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>783</td>
</tr>
<tr>
<td>3</td>
<td>20546</td>
<td>1290</td>
<td>163</td>
<td>31</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1095</td>
</tr>
<tr>
<td>4</td>
<td>20312</td>
<td>1252</td>
<td>133</td>
<td>30</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1405</td>
</tr>
<tr>
<td>5</td>
<td>20078</td>
<td>1164</td>
<td>132</td>
<td>25</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>1733</td>
</tr>
<tr>
<td>6</td>
<td>19646</td>
<td>1146</td>
<td>135</td>
<td>16</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>2190</td>
</tr>
<tr>
<td>7</td>
<td>19432</td>
<td>1092</td>
<td>115</td>
<td>23</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2473</td>
</tr>
</tbody>
</table>

Table 4.8: Number of households by number of reported crimes and interview sequence.

crimes, and 2001 households reported one crime, 275 households reported two crimes, 68 households reported three crimes, 20 households reported four crimes, 6 households reported five crimes and only one household reported six crimes. The maximum number of crimes reported by a single household in our data is six. No household is missing from the first interview because we are only using those households that are present in the first interview in our modeling. On the remaining interview times, 783 households were missing the second interview, 1095 households were missing the third interview, 1405 households were missing the fourth interview, 1733 households were missing the fifth interview, 2190 households were missing the sixth interview and 2473 households were missing the seventh interview. It is not surprising to see that the number of missing household responses increases by the sequence of the interview. In fact, this phenomenon is also part of the time-in-sample effect resulting from the rotating panel design in NCVS. Another reason for the increased missingness comes from the households that drop out of the sample. Drop-outs are
those households that miss an interview and then never return to the sample. For example, a household in a sampled address that stopped being interviewed after the first couple of interviews will be a drop-out. From Table 4.7, there are 201 households that dropped out starting from the second interview (missing pattern No. 11), 199 households dropped out starting from the third interview (missing pattern No. 12), 212 households dropped out starting from the fourth interview (missing pattern No. 10), 298 households dropped out starting from the fifth interview (missing pattern No. 8) and 405 households dropped out starting from the sixth interview (missing pattern No. 4). These drop-outs account for 27.8% of the total number of households with missing interviews. The drop-outs are an important underlying reason for the increased missingness across the sequence of interviews in Table 4.8. However, it is hard to pick out these drop-outs from the organization of Table 4.8 and our models discussed in this thesis do not specifically target the households that drop out of the sample. Models that can describe the drop-out patterns of the sampled households might be a direction for future studies.

From Table 4.8, we can see that the number of households in each column is generally decreasing as the households’ time-in-sample increases, except in the “missing” column. This is expected since the number of missing households becomes larger as time-in-sample increases, leaving fewer households in the first seven columns. Furthermore, the respondents are less likely to report a crime the longer they are in the sample. Thus, for the second to the seventh columns in Table 4.8, the numbers are almost strictly decreasing from top to bottom. We also note that the majority of households reported zero crimes in all seven interviews (80% to 90% households are
in the first column of Table 4.8), which is consistent with our zero-inflated assumption in the previous chapters. In each interview, for those households that reported a crime, most of them reported only one or two crimes (second and third columns in Table 4.8), and a very small portion of the households reported more than two crimes (fourth to seventh columns in Table 4.8).

We propose a multinomial model for the data in Table 4.8. We assume that the counts in each row of Table 4.8 follow a multinomial distribution with a total number of \(N=23139\) trials. The cell probabilities under our model have the form shown in Table 4.9, where \(\pi_i\) is the probability of missing a survey response in the \(i^{th}\) interview and \(w_i\) is the probability of reporting zero crimes in the \(i^{th}\) interview given all the \(N=23139\) households at each interview, for \(i = 1, 2, ..., 7\). Note that \(\pi_1 = 0\). \(p_1, p_2, ..., p_6\) are the zero-truncated Poisson probabilities for counts 1 to 6 with parameter \(\lambda\): \(p_k = \frac{\lambda^k}{(e^\lambda - 1)k!}\) for \(k = 1, 2, ..., 6\). We further define \(p_+ = \sum_{k=1}^6 p_k\).

By assuming all seven interviews share the same Poisson parameter \(\lambda\), we are ignoring the bounding effect at this point. The probabilities for missing and zero crimes are different for the seven interviews, however, accounting for the time-in-sample effect. We choose to model the exact number of reported crimes instead of collapsing the final cell to, for example, “more than 6 crimes reported”, to make our later estimation of total crime rates easier. In the multinomial model shown in Table 4.9, we need to estimate seven \(\pi_i\)s, seven \(w_i\)s and \(\lambda\) using our 8 by 7 data table with seven constraints on the row sums. Thus, we are left with \((8 \times 7) - 7 - 7 - 1 - 7 = 34\) degrees of freedom.
<table>
<thead>
<tr>
<th>Seq</th>
<th>0 crimes</th>
<th>1 crime</th>
<th>2 crimes</th>
<th>3 crimes</th>
<th>4 crimes</th>
<th>5 crimes</th>
<th>6 crimes</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(w_1)</td>
<td>((1 - w_1))</td>
<td>((1 - w_1))</td>
<td>((1 - w_1))</td>
<td>((1 - w_1))</td>
<td>((1 - w_1))</td>
<td>((1 - w_1))</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>(w_2)</td>
<td>((1 - \pi_2))</td>
<td>((1 - \pi_2))</td>
<td>((1 - \pi_2))</td>
<td>((1 - \pi_2))</td>
<td>((1 - \pi_2))</td>
<td>(\pi_2)</td>
<td>(\pi_2)</td>
</tr>
<tr>
<td></td>
<td>((1 - \pi_2))</td>
<td>((1 - w_2))</td>
<td>((1 - w_2))</td>
<td>((1 - w_2))</td>
<td>((1 - w_2))</td>
<td>(\pi_2)</td>
<td>(\pi_2)</td>
<td>(\pi_2)</td>
</tr>
<tr>
<td>3</td>
<td>(w_3)</td>
<td>((1 - \pi_3))</td>
<td>((1 - \pi_3))</td>
<td>((1 - \pi_3))</td>
<td>((1 - \pi_3))</td>
<td>(\pi_3)</td>
<td>(\pi_3)</td>
<td>(\pi_3)</td>
</tr>
<tr>
<td></td>
<td>((1 - \pi_3))</td>
<td>((1 - w_3))</td>
<td>((1 - w_3))</td>
<td>((1 - w_3))</td>
<td>((1 - w_3))</td>
<td>(\pi_3)</td>
<td>(\pi_3)</td>
<td>(\pi_3)</td>
</tr>
<tr>
<td>4</td>
<td>(w_4)</td>
<td>((1 - \pi_4))</td>
<td>((1 - \pi_4))</td>
<td>((1 - \pi_4))</td>
<td>((1 - \pi_4))</td>
<td>(\pi_4)</td>
<td>(\pi_4)</td>
<td>(\pi_4)</td>
</tr>
<tr>
<td></td>
<td>((1 - \pi_4))</td>
<td>((1 - w_4))</td>
<td>((1 - w_4))</td>
<td>((1 - w_4))</td>
<td>((1 - w_4))</td>
<td>(\pi_4)</td>
<td>(\pi_4)</td>
<td>(\pi_4)</td>
</tr>
<tr>
<td>5</td>
<td>(w_5)</td>
<td>((1 - \pi_5))</td>
<td>((1 - \pi_5))</td>
<td>((1 - \pi_5))</td>
<td>((1 - \pi_5))</td>
<td>(\pi_5)</td>
<td>(\pi_5)</td>
<td>(\pi_5)</td>
</tr>
<tr>
<td></td>
<td>((1 - \pi_5))</td>
<td>((1 - w_5))</td>
<td>((1 - w_5))</td>
<td>((1 - w_5))</td>
<td>((1 - w_5))</td>
<td>(\pi_5)</td>
<td>(\pi_5)</td>
<td>(\pi_5)</td>
</tr>
<tr>
<td>6</td>
<td>(w_6)</td>
<td>((1 - \pi_6))</td>
<td>((1 - \pi_6))</td>
<td>((1 - \pi_6))</td>
<td>((1 - \pi_6))</td>
<td>(\pi_6)</td>
<td>(\pi_6)</td>
<td>(\pi_6)</td>
</tr>
<tr>
<td></td>
<td>((1 - \pi_6))</td>
<td>((1 - w_6))</td>
<td>((1 - w_6))</td>
<td>((1 - w_6))</td>
<td>((1 - w_6))</td>
<td>(\pi_6)</td>
<td>(\pi_6)</td>
<td>(\pi_6)</td>
</tr>
<tr>
<td>7</td>
<td>(w_7)</td>
<td>((1 - \pi_7))</td>
<td>((1 - \pi_7))</td>
<td>((1 - \pi_7))</td>
<td>((1 - \pi_7))</td>
<td>(\pi_7)</td>
<td>(\pi_7)</td>
<td>(\pi_7)</td>
</tr>
<tr>
<td></td>
<td>((1 - \pi_7))</td>
<td>((1 - w_7))</td>
<td>((1 - w_7))</td>
<td>((1 - w_7))</td>
<td>((1 - w_7))</td>
<td>(\pi_7)</td>
<td>(\pi_7)</td>
<td>(\pi_7)</td>
</tr>
</tbody>
</table>

Table 4.9: Underlying probabilities for the observed crime count data under model in Section 4.2.1.
The probabilities in Table 4.9 require the assumption that the seven rows are independent from each other given the common Poisson parameter $\lambda$, i.e., we assume there is no correlation in crime reports across interview sequence. This is a strong assumption as we know that the seven interviews are related to each other since they are summarizing repeated interviews of the same group of households. For example, if a household reported zero crimes in one interview, it is more likely to also report no crimes in the following interviews. We will address this correlation in later sections.

Another assumption is that the probability of missing is not related to the number of crimes that would be reported.

In Table 4.8, let $y_{ij}$ be the cell count of the $i^{th}$ row (interview sequence) and $j^{th}$ column, where $i = 1, 2, \ldots, 7$ and $j = 0, 1, \ldots, 6, m$ ("m" stands for "missing"). Then, the likelihood function for our data as displayed in Table 4.8 can be written as:

$$L(w, \lambda, \pi | y) \propto 7 \prod_{i=1}^{7} \left( [(1 - \pi_i)w_i]^{y_{i0}}[(1 - \pi_i)(1 - w_i)p_1/p_+]^{y_{i1}} \right. $$

$$\left. [(1 - \pi_i)(1 - w_i)p_2/p_+]^{y_{i2}}[(1 - \pi_i)(1 - w_i)p_3/p_+]^{y_{i3}} \right. $$

$$\left. [(1 - \pi_i)(1 - w_i)p_4/p_+]^{y_{i4}}[(1 - \pi_i)(1 - w_i)p_5/p_+]^{y_{i5}} \right. $$

$$\left. [(1 - \pi_i)(1 - w_i)p_6/p_+]^{y_{i6} \pi_{im}} \right) $$

$$= 7 \prod_{i=1}^{7} \left( [(1 - \pi_i)\sum_{j=0}^{6} y_{ij} \pi_{im}] \ast \left[ w_{i}^{y_{i0}}(1 - w_{i})^{\sum_{j=1}^{6} y_{ij}} \right] \ast \prod_{j=1}^{6} (p_j/p_+)^{y_{ij}} \right). $$

We can see from the above likelihood function that under our multinomial probability model and our independence assumption, the likelihood function separates into a factor involving the $w$ parameters alone, a factor involving the $\pi$ parameters alone and a factor involving the $\lambda$ (in the $p$-parameters) alone. Furthermore, the estimation (MLE) of $\pi_i$ and $w_i$ is straightforward. Taking the derivatives of the likelihood function in Equation 4.8 with respect to $\pi_i$ and $w_i$ and setting them to zero, we obtain
the following closed-form estimators for $\pi_i$ and $w_i$:

$$\hat{\pi}_i = \frac{y_{im}}{N}, \quad \hat{w}_i = \frac{y_{i0}}{N - y_{im}},$$  \hspace{1cm} (4.9)$$

where $N = 23139$ is the total number of households in each interview period (row).

For the MLE of the Poisson parameter $\lambda$, after taking the derivative of the above likelihood function with respect to $\lambda$ and setting the derivative to zero, we obtain a $5^{th}$ order polynomial equation in $\hat{\lambda}$:

$$a_5 \hat{\lambda}^5 + a_4 \hat{\lambda}^4 + a_3 \hat{\lambda}^3 + a_2 \hat{\lambda}^2 + a_1 \hat{\lambda} + a_0 = 0,$$

where

$$a_5 = -\frac{1}{144} c_1 - \frac{1}{180} c_2 - \frac{1}{240} c_3 - \frac{1}{360} c_4 - \frac{1}{720} c_5,$$

$$a_4 = -\frac{1}{30} c_1 - \frac{1}{40} c_2 - \frac{1}{60} c_3 - \frac{1}{120} c_4 + \frac{1}{120} c_6,$$

$$a_3 = -\frac{1}{8} c_1 - \frac{1}{12} c_2 - \frac{1}{24} c_3 + \frac{1}{24} c_5 + \frac{1}{12} c_6,$$

$$a_2 = -\frac{1}{3} c_1 - \frac{1}{6} c_2 + \frac{1}{6} c_4 + \frac{1}{3} c_5 + \frac{1}{2} c_6,$$

$$a_1 = -\frac{1}{2} c_1 + \frac{1}{2} c_3 + c_4 + \frac{3}{2} c_5 + 2 c_6,$$

$$a_0 = c_2 + 2 c_3 + 3 c_4 + 4 c_5 + 5 c_6,$$

where $c_k = \sum_{i=1}^{7} y_{ik}$ for $k = 1, 2, ..., 6$, are the sums of the $2^{nd}$ column to the $7^{th}$ column in Table 4.8.

Solving the above equation gives us: $\hat{\lambda} = 0.3256$ with the standard error: $SE(\hat{\lambda}) = 0.0026$.

Plugging the estimated parameters into our probability model and multiplying the whole table by the total number of households in our data ($N=23139$), we obtain the expected number of households in each category in each interview as listed in Table 4.10.

As we can see from the expected cell counts in Table 4.10, the expected numbers of households in the first and last columns of our data match the observed data exactly,
Table 4.10: Expected crime counts by interview sequence under model in Table 4.9.

<table>
<thead>
<tr>
<th>Seq</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20768</td>
<td>2005.888</td>
<td>326.581</td>
<td>35.447</td>
<td>2.886</td>
<td>0.188</td>
<td>0.010</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>20604</td>
<td>1482.208</td>
<td>241.320</td>
<td>21.32</td>
<td>1.396</td>
<td>0.139</td>
<td>0.008</td>
<td>783</td>
</tr>
<tr>
<td>3</td>
<td>20546</td>
<td>1267.322</td>
<td>206.334</td>
<td>22.396</td>
<td>1.823</td>
<td>0.119</td>
<td>0.006</td>
<td>1095</td>
</tr>
<tr>
<td>4</td>
<td>20312</td>
<td>1203.025</td>
<td>195.866</td>
<td>21.259</td>
<td>1.731</td>
<td>0.113</td>
<td>0.006</td>
<td>1405</td>
</tr>
<tr>
<td>5</td>
<td>20078</td>
<td>1123.500</td>
<td>182.919</td>
<td>19.854</td>
<td>1.616</td>
<td>0.105</td>
<td>0.006</td>
<td>1733</td>
</tr>
<tr>
<td>6</td>
<td>19646</td>
<td>1102.350</td>
<td>179.475</td>
<td>19.480</td>
<td>1.586</td>
<td>0.103</td>
<td>0.006</td>
<td>2190</td>
</tr>
<tr>
<td>7</td>
<td>19432</td>
<td>1043.975</td>
<td>169.971</td>
<td>18.449</td>
<td>1.502</td>
<td>0.098</td>
<td>0.005</td>
<td>2473</td>
</tr>
</tbody>
</table>

i.e., for the number of households reporting zero crimes and the number of households that missed each interview, we directly used the observed cell counts as the expected cell counts because we are estimating 14 parameters with 14 cell counts yielding zero degrees of freedom for these two columns.

Examining the expected cell counts in Table 4.10 and the observed counts in Table 4.8, we notice that our model over-estimates the number of households that reported two crimes, while under-estimating the number of households that reported three, four and five crimes. This is because our data does not exactly follow a truncated Poisson distribution as we assumed in our probability model.

To further compare the predicted cell counts in Table 4.10 with the observed counts in Table 4.8, $X^2$ and $G^2$ statistics for goodness of fit are calculated. With degrees of freedom being 34, we have the values of $X^2 = 3854.033$ and $G^2 = 352.134$. The reason the $X^2$ and $G^2$ statistics are very different from each other is due to the very small cell counts in the 6th and 7th columns of our data table, under which circumstances the $G^2$ statistic is more suitable for measuring the goodness of fit. This
is because the expected cell counts appear in the denominator of the calculation of the \( X^2 \) statistic, which causes problems when the expected cell counts are close to zero (Bishop et al., 1975). For this reason, we stop reporting the \( X^2 \) statistics for the rest of models in the current chapter. Another important reason for the difference is that the \( X^2 \) and \( G^2 \) statistics have the same asymptotic \( \chi^2 \) distribution only if the observations are independent of each other and the model fits the data. However, for our complex survey data, the independence assumption is not satisfied. Furthermore, the \( X^2 \) and \( G^2 \) statistics show that this model is not a good fit to our observed data in Table 4.8.

To further study the summarized count data in Table 4.8, similar to what we did in Section 4.1, we next build models for the 18 small domains listed in Table 4.1.

4.2.2 Small domains model with missing interviews

In this section, we divide all the households in Table 4.8 into 18 sub-tables with each table representing one of the 18 small domains listed in Table 4.1. Thus, we now have 18 tables of 7 by 8 observed cell counts and the total counts for the \((i,j)\) cell in each of the 18 tables should sum to the corresponding \((i,j)\) cell count in Table 4.8. Note that the households are not evenly divided into small domains. Some of the domains have a relatively large number of households in them, while other domains have relatively few households in them. For example, the largest domain (domain 9 in Table 4.1) has 6252 households; the smallest domain (domain 14 in Table 4.1) has only 104 households.

We extend the model given in Table 4.9 to 18 tables of probability models with a multinomial probability model for each domain. Similar to the probability model
given in Table 4.9, we again assume independence across interviews and independence across households. We model the Poisson parameters as:

\[
\log(\lambda^{(d)}) = \beta_0 + \beta_1 I_{NE}^{(d)} + \beta_2 I_{W}^{(d)} + \beta_3 I_{CC}^{(d)} + \beta_4 I_{SC}^{(d)} + \beta_5 I_{LI}^{(d)} \tag{4.10}
\]

for the \(d = 1, 2, \ldots, 18\) domains. The indicators in Equation 4.10 are defined as in Table 4.11.

<table>
<thead>
<tr>
<th>(I^{(d)})</th>
<th>Southwest region</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I_{NE}^{(d)})</td>
<td>Northeast region</td>
</tr>
<tr>
<td>(I_{W}^{(d)})</td>
<td>West region</td>
</tr>
<tr>
<td>(I_{CC}^{(d)})</td>
<td>Central city</td>
</tr>
<tr>
<td>(I_{SC}^{(d)})</td>
<td>Small city in MSA</td>
</tr>
<tr>
<td>(I_{LI}^{(d)})</td>
<td>Low income household</td>
</tr>
</tbody>
</table>

Table 4.11: Definitions of indicator variables in Section 4.2.2.

The probability of no crimes, \(w_i\), and the probability of missing a response at interview \(i\), \(\pi_i\), are taken to be the same for all 18 domains for the \(i^{th}\) interview, where \(i = 1, 2, \ldots, 7\). Thus, in this small domain model, we are estimating seven \(w_i\)s, seven \(\pi_i\)s and six \(\beta_i\)s with 18 observed 8 by 7 data tables having 7 constraints on the row sums of each of the 18 data tables for the small domains. The degrees of freedom associated with this model are, therefore, \((18 \times 7 \times 8) - 7 - 7 - 6 - (18 \times 7) = 862\). We label this model as Model 4.2.2.1.

The MLEs for the \(w_i\) and \(\pi_i\) parameters are the same as for the model in Section 4.2.1. The portion of the likelihood function involving \(\lambda^{(d)}\) is now:

\[
L(\lambda|y) \propto \prod_{d=1}^{18} \frac{(\lambda^{(d)})^{c_{1}^{(d)}}}{(\lambda^{(d)}) + (\lambda^{(d)})^2 + (\lambda^{(d)})^3 + (\lambda^{(d)})^4 + (\lambda^{(d)})^5 + (\lambda^{(d)})^6 + (\lambda^{(d)})^7)^{c_{2}^{(d)}}} \tag{4.11}
\]
where

\[ C_1^{(d)} = c_1^{(d)} + 2c_2^{(d)} + 3c_3^{(d)} + 4c_4^{(d)} + 5c_5^{(d)} + 6c_6^{(d)}, \]

\[ C_2^{(d)} = c_1^{(d)} + c_2^{(d)} + c_3^{(d)} + c_4^{(d)} + c_5^{(d)} + c_6^{(d)}, \]

and \( c_k^{(d)} = \sum_{i=1}^{7} y_{ik}^{(d)} \) for \( k = 1, 2, ..., 6 \), are the sums of the 2<sup>nd</sup> column to the 7<sup>th</sup> column in the data table for domain \( d \).

Thus, the portion of the loglikelihood function involving the \( \beta \)s is:

\[
\begin{align*}
LL(\beta|y) \propto & \sum_{d=1}^{18} \left[ C_1^{(d)}(\beta_0 + \beta_1 I_{NE}^{(d)} + \beta_2 I_{W}^{(d)} + \beta_3 I_{CC}^{(d)} + \beta_4 I_{SC}^{(d)} + \beta_5 I_{LI}^{(d)}) \\
& - C_2^{(d)} \log(e^{(\beta_0 + \beta_1 I_{NE}^{(d)} + \beta_2 I_{W}^{(d)} + \beta_3 I_{CC}^{(d)} + \beta_4 I_{SC}^{(d)} + \beta_5 I_{LI}^{(d)})} + 2!) \\
& + e^{3(\beta_0 + \beta_1 I_{NE}^{(d)} + \beta_2 I_{W}^{(d)} + \beta_3 I_{CC}^{(d)} + \beta_4 I_{SC}^{(d)} + \beta_5 I_{LI}^{(d)})}/3!} \\
& + e^{4(\beta_0 + \beta_1 I_{NE}^{(d)} + \beta_2 I_{W}^{(d)} + \beta_3 I_{CC}^{(d)} + \beta_4 I_{SC}^{(d)} + \beta_5 I_{LI}^{(d)})}/4!} \\
& + e^{5(\beta_0 + \beta_1 I_{NE}^{(d)} + \beta_2 I_{W}^{(d)} + \beta_3 I_{CC}^{(d)} + \beta_4 I_{SC}^{(d)} + \beta_5 I_{LI}^{(d)})}/5!} \\
& + e^{6(\beta_0 + \beta_1 I_{NE}^{(d)} + \beta_2 I_{W}^{(d)} + \beta_3 I_{CC}^{(d)} + \beta_4 I_{SC}^{(d)} + \beta_5 I_{LI}^{(d)})}/6!)\right].
\end{align*}
\]

After taking derivatives with respect to \( \beta_0, \beta_1, ..., \beta_5 \) and setting the derivatives equal to zero, we obtain a system of non-linear equations to solve for the \( \hat{\beta} \)s. Solving this system of equations using the R function \( BBsolve() \) in package BB (Varadhan and Gilbert, 2009) gives us the MLEs and the estimated standard errors for the \( \beta \)s shown in Table 4.12.

Substituting the parameter estimates into our probability model and multiplying them by the total number of households in each domain, we obtain the expected number of households in each category for the seven interviews in each of the 18 domains to compare to the observed cell counts for the 18 domains. The resulting
Table 4.12: Estimated $\beta$ parameter values and standard errors for Model 4.2.2.1.

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
<th>$\hat{\beta}_3$</th>
<th>$\hat{\beta}_4$</th>
<th>$\hat{\beta}_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>-1.2943</td>
<td>-0.1566</td>
<td>-0.0843</td>
<td>0.3048</td>
<td>0.1641</td>
<td>0.1036</td>
</tr>
<tr>
<td>SE</td>
<td>0.0088</td>
<td>0.0223</td>
<td>0.0172</td>
<td>0.0148</td>
<td>0.0121</td>
<td>0.0214</td>
</tr>
</tbody>
</table>

$G^2$ statistic is 1952.263, with 862 degrees of freedom. Examining the contributions to the $G^2$ statistic of all the cell counts, we found that the major lack of fit of Model 4.2.2.1 comes from the lack of fit in the probabilities of reporting zero crimes and the probabilities of missing an interview. Thus, we will try to improve the fit of the model by introducing additional parameters into the model to seek a better fit in the $w_i$s and $\pi_i$s.

Next we fit another model that is very similar to Model 4.2.2.1 except that we assume the probability for reporting zero crimes and the probability for missing a response are different for each domain. In other words, instead of assuming the $w_i$s and $\pi_i$s are the same for all domains, we now have $w_i^{(d)}$ and $\pi_i^{(d)}$ for domain $d$. Thus, we are estimating $7 \times 18$ $w_i^{(d)}$s and $7 \times 18$ $\pi_i^{(d)}$s instead of seven $w_i$s and seven $\pi_i$s. We keep all other assumptions the same as in Model 4.2.2.1, i.e., we still have

$$\log(\lambda^{(d)}) = \beta_0 + \beta_1 I_{NE}^{(d)} + \beta_2 I_W^{(d)} + \beta_3 I_{CC}^{(d)} + \beta_4 I_{SC}^{(d)} + \beta_5 I_{LI}^{(d)}.$$  

Thus, we now have a model for each domain as shown in Table 4.13 and we number this model as Model 4.2.2.2. In this model, we are estimating $2 \times (7 \times 18 - 7) = 238$ more parameters than for Model 4.2.2.1, leaving us $862 - 238 = 624$ degrees of freedom.
<table>
<thead>
<tr>
<th>Seq</th>
<th>0 crimes</th>
<th>1 crime</th>
<th>2 crimes</th>
<th>3 crimes</th>
<th>4 crimes</th>
<th>5 crimes</th>
<th>6 crimes</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$w_1^{(d)}$</td>
<td>$(1 - w_1^{(d)})$</td>
<td>$(1 - w_1^{(d)})$</td>
<td>$(1 - w_1^{(d)})$</td>
<td>$(1 - w_1^{(d)})$</td>
<td>$(1 - w_1^{(d)})$</td>
<td>$(1 - w_1^{(d)})$</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>$w_2^{(d)}$</td>
<td>$(1 - \pi_2^{(d)})$</td>
<td>$(1 - w_2^{(d)})$</td>
<td>$(1 - w_2^{(d)})$</td>
<td>$(1 - w_2^{(d)})$</td>
<td>$(1 - w_2^{(d)})$</td>
<td>$(1 - w_2^{(d)})$</td>
<td>$\pi_2^{(d)}$</td>
</tr>
<tr>
<td></td>
<td><em>$p_1^{(d)}/p_+$</em></td>
<td><em>$p_2^{(d)}/p_+$</em></td>
<td><em>$p_3^{(d)}/p_+$</em></td>
<td><em>$p_4^{(d)}/p_+$</em></td>
<td><em>$p_5^{(d)}/p_+$</em></td>
<td><em>$p_6^{(d)}/p_+$</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$w_3^{(d)}$</td>
<td>$(1 - \pi_3^{(d)})$</td>
<td>$(1 - w_3^{(d)})$</td>
<td>$(1 - w_3^{(d)})$</td>
<td>$(1 - w_3^{(d)})$</td>
<td>$(1 - w_3^{(d)})$</td>
<td>$(1 - w_3^{(d)})$</td>
<td>$\pi_3^{(d)}$</td>
</tr>
<tr>
<td></td>
<td><em>$p_1^{(d)}/p_+$</em></td>
<td><em>$p_2^{(d)}/p_+$</em></td>
<td><em>$p_3^{(d)}/p_+$</em></td>
<td><em>$p_4^{(d)}/p_+$</em></td>
<td><em>$p_5^{(d)}/p_+$</em></td>
<td><em>$p_6^{(d)}/p_+$</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$w_4^{(d)}$</td>
<td>$(1 - \pi_4^{(d)})$</td>
<td>$(1 - w_4^{(d)})$</td>
<td>$(1 - w_4^{(d)})$</td>
<td>$(1 - w_4^{(d)})$</td>
<td>$(1 - w_4^{(d)})$</td>
<td>$(1 - w_4^{(d)})$</td>
<td>$\pi_4^{(d)}$</td>
</tr>
<tr>
<td></td>
<td><em>$p_1^{(d)}/p_+$</em></td>
<td><em>$p_2^{(d)}/p_+$</em></td>
<td><em>$p_3^{(d)}/p_+$</em></td>
<td><em>$p_4^{(d)}/p_+$</em></td>
<td><em>$p_5^{(d)}/p_+$</em></td>
<td><em>$p_6^{(d)}/p_+$</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$w_5^{(d)}$</td>
<td>$(1 - \pi_5^{(d)})$</td>
<td>$(1 - w_5^{(d)})$</td>
<td>$(1 - w_5^{(d)})$</td>
<td>$(1 - w_5^{(d)})$</td>
<td>$(1 - w_5^{(d)})$</td>
<td>$(1 - w_5^{(d)})$</td>
<td>$\pi_5^{(d)}$</td>
</tr>
<tr>
<td></td>
<td><em>$p_1^{(d)}/p_+$</em></td>
<td><em>$p_2^{(d)}/p_+$</em></td>
<td><em>$p_3^{(d)}/p_+$</em></td>
<td><em>$p_4^{(d)}/p_+$</em></td>
<td><em>$p_5^{(d)}/p_+$</em></td>
<td><em>$p_6^{(d)}/p_+$</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$w_6^{(d)}$</td>
<td>$(1 - \pi_6^{(d)})$</td>
<td>$(1 - w_6^{(d)})$</td>
<td>$(1 - w_6^{(d)})$</td>
<td>$(1 - w_6^{(d)})$</td>
<td>$(1 - w_6^{(d)})$</td>
<td>$(1 - w_6^{(d)})$</td>
<td>$\pi_6^{(d)}$</td>
</tr>
<tr>
<td></td>
<td><em>$p_1^{(d)}/p_+$</em></td>
<td><em>$p_2^{(d)}/p_+$</em></td>
<td><em>$p_3^{(d)}/p_+$</em></td>
<td><em>$p_4^{(d)}/p_+$</em></td>
<td><em>$p_5^{(d)}/p_+$</em></td>
<td><em>$p_6^{(d)}/p_+$</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>$w_7^{(d)}$</td>
<td>$(1 - \pi_7^{(d)})$</td>
<td>$(1 - w_7^{(d)})$</td>
<td>$(1 - w_7^{(d)})$</td>
<td>$(1 - w_7^{(d)})$</td>
<td>$(1 - w_7^{(d)})$</td>
<td>$(1 - w_7^{(d)})$</td>
<td>$\pi_7^{(d)}$</td>
</tr>
<tr>
<td></td>
<td><em>$p_1^{(d)}/p_+$</em></td>
<td><em>$p_2^{(d)}/p_+$</em></td>
<td><em>$p_3^{(d)}/p_+$</em></td>
<td><em>$p_4^{(d)}/p_+$</em></td>
<td><em>$p_5^{(d)}/p_+$</em></td>
<td><em>$p_6^{(d)}/p_+$</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.13: Cell probabilities under Model 4.2.2.2 for the crime count data of domain d.
This model has exactly the same MLEs for the $\beta$ s as in Model 4.2.2.1, since the estimation of the $\beta$ s can be completed separately from the estimation of the $w$s and $\pi$s. The new MLEs for the $w$s and $\pi$s are:

$$\hat{\pi}_i^{(d)} = \frac{y_{im}^{(d)}}{N^{(d)}}, \text{ for } i = 2, 3, ..., 7 \text{ and } d = 1, 2, ..., 18$$

and

$$\hat{w}_i^{(d)} = \frac{y_{i0}^{(d)}}{N^{(d)} - y_{im}^{(d)}}, \text{ for } i = 1, 2, ..., 7 \text{ and } d = 1, 2, ..., 18$$

where $y_{im}^{(d)}$ is the number of households that are missing the $i^{th}$ interview in domain $d$, $N^{(d)}$ is the total number of households in domain $d$ and $y_{i0}^{(d)}$ is the number of households in domain $d$ that reported zero crimes in the $i^{th}$ interview.

In Model 4.2.2.2, we are estimating $18 \times 6 \ \hat{\pi}_i^{(d)}$ s using the $18 \times 6$ cell counts in our data, and estimating the $18 \times 7 \ \hat{w}_i^{(d)}$ s using the $18 \times 7$ cell counts in our data. In other words, the predicted number of households for both those that reported zero crimes and those that missed the interview are going to be exactly the same as the observed corresponding counts in each domain for all seven interviews. Thus, there are no degrees of freedom remaining from the first and last columns of each domain; that portion of the model fits the data perfectly. We are only estimating the $\beta$ parameters in our model. This is the same situation as for the aggregated data model in Section 4.2.1.

Using the same MLEs for the $\beta$ s as in Model 4.2.2.1 and the above estimators for the $\pi$s and $w$s, we can calculate the expected cell counts for each domain under Model 4.2.2.2 and compare them with the observed data table. The $\chi^2$ statistic is 291.309 with 624 degrees of freedom. We further examine the goodness of fit of Model 4.2.2.2 by looking at the absolute residuals and contributions to $\chi^2$ statistic from the
cells in each domain. Large absolute residuals occur when the number of households in a cell is large, for example, in the \(2^{nd}\) and \(3^{rd}\) columns in the data table (number of households reporting one or two crimes). Large contributions to the \(G^2\) statistic occur mostly in the \(2^{nd}\) to \(5^{th}\) columns of the table, where the main lack of fit of the model occurs. The reason for the lack of fit is that the observed counts in each row in the data table decrease too fast to allow a single Poisson distribution to fit the data well (recall that we used a zero-truncated Poisson to describe the probabilities of reporting 1 to 6 crimes). This causes the model to over-estimate the \(3^{rd}\) column (number of HHs reporting two crimes) in the data table and under-estimate the \(2^{nd}\) column (number of HHs reporting one crime) to accommodate fitting the rest of the columns. There is also a trend that domains with a particularly large or small number of households tend to have larger contributions to the \(G^2\) statistic.

The difference between the \(G^2\) statistics of Models 4.2.2.1 and 4.2.2.2 is 1660.954, which is large compared to the difference of degrees of freedom between these two models (238). The major difference between Models 4.2.2.1 and 4.2.2.2 lies in the modeling of the probabilities of reporting zero crimes (the \(w_s\)) and the probabilities of missing an interview (the \(\pi_s\)). Thus, we can conclude that the major lack of fit in Model 4.2.2.1 comes from the lack of fit in the probabilities of reporting zero crimes (the \(w_s\)) and the probabilities of missing an interview (the \(\pi_s\)).

Although the fit of Model 4.2.2.2 is good, it models the seven probabilities of reporting zero crimes and the seven probabilities of missing an interview in each domain separately. To account for the correlation caused by repeated interviews on the same group of households in each domain for seven interview times, in the next
section, we turn to Bayesian models to account for the underlying relationship among these probabilities.

4.2.3 Small domain Bayesian models with missing interviews

Models in Sections 4.2.1 and 4.2.2 all require the assumption that the seven interviews are independent from each other, which is not a realistic assumption as discussed previously. We next try to improve our models by accounting for the correlations among the seven interviews. The correlations between interviews for the same household are accounted for in two aspects of the models. One is in the $\lambda$ parameter of the truncated Poisson distribution, which is the mean reported crime count in each interview if the household did not miss the interview and reported at least one crime. One possible way to relate the mean reported crime count in one interview to that of the next interview is to assume that, in the $\lambda$ parameter, in addition to the $\beta$ coefficients, we also have a random effect $\alpha_i$ for each of the $i = 1, 2, ..., 7$ interviews, since it is common to use random effects in GLMMs to account for correlation among observations.

Another way to introduce correlations among different sequences of interviews into the model is through the probabilities of reporting no crimes (the $w$s) and the probabilities of missing an interview (the $\pi$s). For a fixed domain of households, the proportions of households reporting zero crimes and proportions of households missing an interview over the seven interview times are likely to be similar, although not exactly the same because of the time-in-sample effect. To help develop appropriate models for these cell probabilities, we first plot the observed relative frequencies of $w$s and $\pi$s by interview sequence for all the 18 domains as shown in Figure 4.2.
As discussed at the end of Section 4.2.2, these cells are the major source of lack of fit in the models in Model 4.2.2.1. From Figure 4.2, we can see that the observed proportions of reporting no crimes and the proportions of missing an interview are very different from domain to domain. However, they both show an upward trend and non-linear relationship over the interview times.

Figure 4.2 shows a lot of variation in the data from domain to domain, which calls for some kind of smoothing over all domains. Thus, we use Bayesian models in the following discussion of the current section to try to find a model with better fit for our data. Also, it is convenient to relate parameters to each other through the sharing of the prior distribution.

Under a Bayesian framework, in the model of the mean reported crime count (Equation 4.10), we add a random effect ($\alpha$) for each of the seven interviews. The original coefficient parameters ($\beta$s) are kept the same for all the seven interviews
in all 18 domains. However, the seven random effects ($\alpha$s) are assumed to come from the same prior distribution to accommodate the effect of interviewing the same groups of households in the seven interviews. In other words, the seven random effects ($\alpha$s) are the same for all 18 domains, but within each domain, they are different for the various interview sequences. Thus, the model of the mean reported crime count (the parameters of the truncated Poisson distributions) for interview $i$ in domain $d$ becomes:

$$\log(\lambda_i^{(d)}) = \alpha_i + \beta_1 I_{NE}^{(d)} + \beta_2 I_{W}^{(d)} + \beta_3 I_{CC}^{(d)} + \beta_4 I_{SC}^{(d)} + \beta_5 I_{LI}^{(d)}$$

(4.13)

for $d = 1, 2, ..., 18$ and $i = 1, 2, ..., 7$.

As a result of the random effects in the truncated Poisson mean model, the probabilities of reporting one crime, two crimes, ..., six crimes also differ by interview sequence. That means, instead of having $p_1^{(d)}$, $p_2^{(d)}$, ..., $p_6^{(d)}$, we now have $p_{i1}^{(d)}$, $p_{i2}^{(d)}$, ..., $p_{i6}^{(d)}$ as the zero-truncated Poisson probabilities for crime counts 1 to 6, where

$$p_{ik}^{(d)} = \frac{\lambda_i^{(d)}^k}{(e^\lambda_i^{(d)} - 1)k!}$$

for $k = 1, 2, ..., 6$, and

$$p_{i+}^{(d)} = \sum_{k=1}^{6} p_{ik}^{(d)}.$$

For the probabilities of reporting no crimes and the probabilities of missing an interview, we also want to model these probabilities as changing by interview sequence to account for the correlation caused by repeated interviews on the same group of households in each domain for seven interview times. In the following, we propose three different models to try to capture this relationship.

The first, Model 4.2.3.1, assumes that the probabilities of reporting zero crimes and the probabilities of missing a response are the same across all 18 domains. However, these probabilities change over the various interview sequences. In other words,
we have $\pi_i$s and $w_i$s in Model 4.2.3.1, where $w_i$ and $\pi_i$ are the probabilities of reporting zero crimes and missing a response in the $i^{th}$ interview, for $i = 1, 2, ..., 7$. Furthermore, the probabilities of reporting no crimes (the $w_i$s) and the probabilities of missing an interview (the $\pi_i$s) are assumed to come from the same prior distributions except for $\pi_1$, which is fixed to be zero. In Model 4.2.3.1, all the parameters ($w_i$s, $\pi_i$s, $\beta$s and $\alpha$s in $\lambda$) are the same for all 18 domains but, within each domain, they are different for the various interview sequences.

The small domains Bayesian Model 4.2.3.1 can be summarized as:

$$
(y^{(d)}_{i0}, y^{(d)}_{i1}, y^{(d)}_{i2}, y^{(d)}_{i3}, y^{(d)}_{i4}, y^{(d)}_{i5}, y^{(d)}_{i6}, y^{(d)}_{im}) \sim
\text{multinomial}((w_i(1 - \pi_i), (1 - w_i)(1 - \pi_i)p_{i1}^{(d)}/p_{i+}^{(d)},
(1 - w_i)(1 - \pi_i)p_{i2}^{(d)}/p_{i+}^{(d)}, (1 - w_i)(1 - \pi_i)p_{i3}^{(d)}/p_{i+}^{(d)},
(1 - w_i)(1 - \pi_i)p_{i4}^{(d)}/p_{i+}^{(d)}, (1 - w_i)(1 - \pi_i)p_{i5}^{(d)}/p_{i+}^{(d)},
(1 - w_i)(1 - \pi_i)p_{i6}^{(d)}/p_{i+}^{(d)}, \pi_i, N^{(d)}).
$$

We fit Model 4.2.3.1 using the MCMC code from the “rjags” package in R (Plummer, 2015). We sampled the posteriors every five steps for a total of 40000 samples with 10000 burn-in steps. For $\pi_2$ to $\pi_7$ and the $w_i$s, we first assume a $Beta(1,1)$ ($\text{uniform}(0,1)$) prior for each of these probabilities. We then change to a $Beta(2,2)$ prior for the $\pi_i$s and $w_i$s for a sensitivity analysis. As our prior distributions for other parameters, we chose to use Normal distributions with very large variance for the parameters as non-informative priors. We assume $\alpha_i \sim N(0, 10^{10})$, for $i = 1, 2, ..., 7$ and $\beta_k \sim N(0, 10^{10})$, for $k = 1, 2, ..., 5$ for the random effects and coefficient parameters in the Poisson mean model.
Figure 4.3: Boxplots of posterior distributions of random effects (\(\alpha\)) in the small domain Bayesian multinomial Model 4.2.3.1.

Figure 4.4: Boxplots of posterior distributions of probabilities of missing an interview (\(\pi\)) in the small domain Bayesian multinomial Model 4.2.3.1.
Figure 4.5: Boxplots of posterior distributions of probabilities of reporting zero crimes \( (w) \) in the small domain Bayesian multinomial Model 4.2.3.1.

Figure 4.6: Boxplots of posterior distributions of the covariate coefficients \( (\beta) \) in the small domain Bayesian multinomial Model 4.2.3.1.
The boxplots of the posterior distributions for the parameters in Model 4.2.3.1 based on a $Beta(1, 1)$ prior for the $\pi_2$ to $\pi_7$ and $w_i$s are shown in Figures 4.3, 4.4, 4.5 and 4.6. Estimation results based on a $Beta(2, 2)$ prior for the $\pi_2$ to $\pi_7$ and $w_i$s are not much different from those with the $Beta(1, 1)$ prior, thus we are omitting those plots to avoid repetition.

From these graphs, we can see that the random effects ($\alpha$s, Figure 4.3) in $\lambda$ clearly show a downward trend as households have more and more interviews. This decrease in the mean predicted crime count is consistent with the time-in-sample effect in the NCVS data. The probabilities for missingness ($\pi_i$s, Figure 4.4) are increasing from the second interview to the seventh interview. The probabilities for reporting zero property crimes among all responding households ($w_i$s, Figure 4.5) are increasing and show a clear bounding effect, meaning respondents in the first interviews are reporting fewer zeroes than at the other waves of interviews. At the same time Figure 4.5 shows a time-in-sample effect as we can see that the probabilities for reporting zero crimes are consistently increasing as time increases, meaning more and more households are reporting zero property crimes as they respond to more and more interviews.

Finally, the posterior boxplots of coefficients for the covariates ($\beta$s) in the model for $\lambda$ are shown in Figure 4.6. The middle 95% of the posterior distributions for $\beta_2$ (West region) and $\beta_5$ (Income_low) cover zero in Figure 4.6, while the middle 95% posterior quantiles of the other parameters do not. The results indicate that households in the Northeast region ($\beta_1$) generally have more property crimes than those in the Midwest and South regions (baseline), while households in the West region ($\beta_2$) do not suffer significantly more property crimes than in the Midwest and South regions. This result is not quite the same as that obtained in Chapter 3’s GLMM fitting results.
The reason for this discrepancy may be that the way we aggregate our data in this chapter happened to balance out the effect in the West region. For the MSA factor, central cities in a MSA ($\beta_3$) generally have a higher probability for household-level property crimes than small cities ($\beta_4$), while they both have more property crimes than in the rural areas (baseline). For the income factor ($\beta_5$), domains with different income levels do not show much difference from each other in the sense of the chance for experiencing a property crime.

We calculated the expected crime count tables for all 18 domains and compared them with the observed tables by the $G^2$ statistics, which serves as a summary statistic for comparison with previous models. The expected count tables for all 18 domains are calculated using the posterior means of cell probabilities multiplied by the total number of households in each domain. The calculated $G^2$ values are 2150.07 with $Beta(1,1)$ priors for the $w_i$s and $\pi_i$s and 2150.08 with $Beta(2,2)$ priors for the $w_i$s and $\pi_i$s. These $G^2$s are comparable to our previous results for Model 4.2.2.1 in Section 4.2.2, where we have $G^2 = 1952.263$ with 862 degrees of freedom. In Model 4.2.2.1, we also have different $\pi_i$s and $w_i$s for each interview while keeping them the same for all 18 domains. In the current Model 4.2.3.1, we have, in addition, added a random effect into the $\lambda$ model, so $\lambda$ is not only different for different domains but can also differ from interview to interview even for the same domain. Thus we are estimating six more parameters in the current model with 18 more constraints on the $\pi_1$s. However, the goodness of fit result for Model 4.2.3.1 is not as good as for Model 4.2.2.1. In general, the goodness of fit for both Models 4.2.2.1 and 4.2.3.1 are not good due to the lack of fit in probabilities of reporting zero crimes and probabilities of missing an interview.
Although the seven probabilities of reporting zero crimes and six probabilities of missing an interview in Model 4.2.3.1 share the same prior, they are not directly related to each other. We decided to fit a second order polynomial to model the logit transformed sequence of seven zero probabilities and the sequence of six missing probabilities (Model 4.2.3.2). In other words, we now assume:

\[ \text{logit}(w_i) = s_0 + s_1 i + s_2 i^2, \text{ for } i = 1, 2, ..., 7 \]

\[ \text{logit}(\pi_i) = r_0 + r_1 i + r_2 i^2 \text{ for } i = 2, ..., 7 \]

where the \( \pi_i \)'s and \( w_i \)'s are the probabilities of reporting zero crimes and missing a response in the \( i^{th} \) interview for all domains and \( \pi_1 = 0 \). \( s_0, s_1, s_2 \) and \( r_0, r_1, r_2 \) are the coefficients of the polynomial functions.

Model 4.2.3.2 has four fewer parameters in the \( w \)s and three fewer parameters in the \( \pi_2 \) to \( \pi_7 \) compared to Model 4.2.3.1. We keep all other model assumptions the same as in Model 4.2.3.1. The model-fitting procedure and prior distributions for Model 4.2.3.2 are also the same as for Model 4.2.3.1, except for the priors on the \( w \)s and \( \pi \)s. Instead of assuming prior distributions directly on the \( w \)s and \( \pi \)s, we assume \( N(0, 10^{10}) \) prior distributions for the \( s_0, s_1, s_2 \) and \( r_0, r_1, r_2 \) in Model 4.2.3.2. We also tried Normal priors with variances different than the \( N(0, 10^{10}) \), such as \( N(0, 10^4) \), \( N(0, 100) \) and \( N(0, 1) \) for a sensitivity analysis. We use the same approach to fit the Bayesian Model 4.2.3.2 as we did for Model 4.2.3.1. The model fitting results show that the results are not sensitive to the priors of parameters.

We calculate the posterior means of probabilities of reporting zero crimes and the probabilities of missing an interview under Model 4.2.3.2 and plot the posterior means of these probabilities as functions of interview sequence (red curve) on top of
Figure 4.7: Posterior means of probabilities from Model 4.2.3.1 (green curve) and Model 4.2.3.2 (red curve) and the observed relative frequencies (black curves) for reporting zero crimes and missing a response by interview sequence.

the observed relative frequencies (black curves) as shown in Figure 4.7. We also provide in Figure 4.7 the posterior means of the same probabilities from Model 4.2.3.1 (green curve). As we can see from Figure 4.7, the posterior means of both types of probabilities from Model 4.2.3.1 and Model 4.2.3.2 are generally increasing, capturing the general time-in-sample effects in all domains. Compared to the predicted probabilities from Model 4.2.3.1, the predicted probabilities from Model 4.2.3.2 are smoothed across various interview sequences. (We omitted the boxplots for the $\alpha$ and $\beta$ parameters as they are very similar to those from Model 4.2.3.1.)

We then compared the expected cell counts from fitting Model 4.2.3.2 with the observed counts using the $G^2$ statistics. The $G^2$ value for Model 4.2.3.2 is 2168.069. Compared to the $G^2$ values around 2150 in Model 4.2.3.1, the difference of around
18 in $G^2$ is reasonable with 7 fewer parameters in Model 4.2.3.2, but the model fit is still poor.

The third Model, 4.2.3.3, assumes that the sequence of probabilities of reporting zero crimes and the sequence of probabilities of missing an interview for each of the 18 domains are both modeled as second order polynomial functions of the interview sequence. This Bayesian small domain Model 4.2.3.3 assumes, for domain $d$, $w_i^{(d)}$ s and $\pi_i^{(d)}$ s instead of $w_i$s and $\pi_i$s. In other words, the probabilities of reporting zero crimes and the probabilities of missing a response change not only by interview sequence, but also by the domain. We therefore assume:

$$\logit (w_i^{(d)}) = s_0^{(d)} + s_1^{(d)} i + s_2^{(d)} i^2, \text{ for } i = 1, 2, ..., 7,$$
$$\logit (\pi_i^{(d)}) = r_0^{(d)} + r_1^{(d)} i + r_2^{(d)} i^2, \text{ for } i = 2, ..., 7,$$

where the $\pi_i^{(d)}$ s and $w_i^{(d)}$ s are the probabilities of reporting zero crimes and missing a response, respectively, in the $i^{th}$ interview for domain $d$, for $d = 1, 2, ..., 18$. $\pi_1^{(d)}$ s are assumed to always be zero. $s_0^{(d)}$, $s_1^{(d)}$, $s_2^{(d)}$ and $r_0^{(d)}$, $r_1^{(d)}$, $r_2^{(d)}$ are the coefficients of the polynomial models in domain $d$.

The small domains Bayesian Model 4.2.3.3 described above can be summarized as:

$$\begin{align*}
(y_{i0}^{(d)}, y_{i1}^{(d)}, y_{i2}^{(d)}, y_{i3}^{(d)}, y_{i4}^{(d)}, y_{i5}^{(d)}, y_{i6}^{(d)}, y_{im}^{(d)}) & \sim \\
\text{multinomial}((w_i^{(d)}(1 - \pi_{i}^{(d)}), (1 - w_i^{(d)})(1 - \pi_{i}^{(d)})p_{i1}^{(d)}/p_{i+}^{(d)}), (1 - w_i^{(d)})(1 - \pi_{i}^{(d)})p_{i2}^{(d)}/p_{i+}^{(d)}), (1 - w_i^{(d)})(1 - \pi_{i}^{(d)})p_{i3}^{(d)}/p_{i+}^{(d)}), (1 - w_i^{(d)})(1 - \pi_{i}^{(d)})p_{i4}^{(d)}/p_{i+}^{(d)}), (1 - w_i^{(d)})(1 - \pi_{i}^{(d)})p_{i5}^{(d)}/p_{i+}^{(d)}), (1 - w_i^{(d)})(1 - \pi_{i}^{(d)})p_{i6}^{(d)}/p_{i+}^{(d)}, \pi_i^{(d)}, N^{(d)}))
\end{align*}$$

(4.15)

for $d = 1, 2, ..., 18$. 

109
We fit Model 4.2.3.3 using the same MCMC code from the “rjags” package in R as we did for Model 4.2.3.1 and Model 4.2.3.2. In the MCMC procedure, we again used 10000 steps to burn-in the samplers and sampled the posteriors every five steps for a total of 40000 samples. We used the same priors for the $\alpha$s and $\beta$s as in Model 4.2.3.1. For the probabilities of reporting zero crimes, $w_i^{(d)}$s, and the probabilities of missing an interview, $\pi_i^{(d)}$s, we assume the same $N(0,10^{10})$ prior distribution for all the coefficients $s_0^{(d)}$, $s_1^{(d)}$, $s_2^{(d)}$ and $r_0^{(d)}$, $r_1^{(d)}$, $r_2^{(d)}$ in the $\pi_i^{(d)}$ s and $w_i^{(d)}$ s, respectively. Compared to Model 4.2.2.2, we are estimating six more parameters by adding the random effects in the Poisson mean model ($\lambda$), while estimating $7 \times 18 = 126$ fewer parameters in the models for the $w_i^{(d)}$ s and $\pi_i^{(d)}$ s. Thus, compared to 624 degrees of freedom in Model 4.2.2.2, we have 120 fewer parameters in Model 4.2.3.3.

We calculated the posterior means of the cell probabilities for all 18 domains. To examine the model fit for the probabilities of reporting zero crimes and the probabilities of missing an interview, in Figure 4.8 we plot the posterior means of these probabilities for each of the 18 domains as functions of interview sequence (red curves) along with the observed relative frequencies (black curves). As we can see from Figure 4.8, the posterior means of both types of probabilities in most domains show a strictly increasing trend. The fitted model captures the general time-in-sample effects in these probabilities while at the same time accounting for the individual characteristics of each domain.

Figure 4.9 shows the differences between the observed relative frequencies and the posterior means of probabilities from Model 4.2.3.3 for reporting zero crimes and missing a response by interview sequence. We can see from Figure 4.9 that the differences for most domains are very small, with all falling between $-0.04$ and $0.04$. 

110
Figure 4.8: Posterior means of probabilities from Model 4.2.3.3 (red curves) and the observed relative frequencies (black curves) for reporting zero crimes and missing a response by interview sequence.

Figure 4.9: Differences between the observed relative frequencies and the posterior means of probabilities from Model 4.2.3.3 for reporting zero crimes and missing a response by interview sequence.
The polynomial function for each domain is able to capture the general trend in the observed relative frequencies of reporting zero crimes and the proportion of missing an interview.

We again do not present the results for the $\alpha$ and $\beta$ parameters from Model 4.2.3.3 as they are very similar to those from Model 4.2.3.1. We then compared the expected cell counts with the observed counts using the $G^2$ statistics as before. The calculated $G^2$ value for Model 4.2.3.3 is 932.272, which is reasonably good with 744 free parameters in the model. Compared to the $G^2$ of 291.309 in Model 4.2.2.2 with 624 degrees of freedom, the difference of 641 in $G^2$ is a big amount with only 120 fewer parameters in Model 4.2.3.3. However, Model 4.2.2.2 directly uses the observed relative frequencies of reporting zero crimes and missing an interview as the estimates of these probabilities, which makes less sense for the modeling of the correlation among interviews at different times.

We summarize the $G^2$ values for all three models in the current section in Table 4.14. Also shown in Table 4.14 are the DIC values for the models in the current section for the purpose of comparison. The DIC value of Model 4.2.3.3 is much smaller than the DIC values of Models 4.2.3.1 and 4.2.3.2. Model 4.2.3.3 has a $G^2$ statistic that is less than half of the $G^2$ values of Models 4.2.3.1 and 4.2.3.2. The difference between the $G^2$ values of Model 4.2.3.3 and those of Models 4.2.3.1 and 4.2.3.2 is more than 1200 with only 100 more parameters in Model 4.2.3.3. The improvement of model fit in Model 4.2.3.3 comes from the improvement of fit in the probabilities of reporting zero crimes and the probabilities of missing a response. Models 4.2.3.1 and 4.2.3.2 assume that the 18 small domains share the same probabilities of reporting zero crimes and missing an interview, which provide us with a general idea about how
these probabilities change over time. As shown in Figure 4.2, the probabilities of reporting zero crimes and the probabilities of missing a response are very different from domain to domain. Model 4.2.3.3 takes into consideration the differences among these probabilities for different domains and thus is a better fit to the observed data.

<table>
<thead>
<tr>
<th>Model</th>
<th>$G^2$</th>
<th># of free parameters</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.3.1 ($Beta(1, 1)$)</td>
<td>2150.07</td>
<td>839</td>
<td>4206</td>
</tr>
<tr>
<td>4.2.3.1 ($Beta(2, 2)$)</td>
<td>2150.08</td>
<td>839</td>
<td>4206</td>
</tr>
<tr>
<td>4.2.3.2</td>
<td>2168.07</td>
<td>850</td>
<td>4217</td>
</tr>
<tr>
<td>4.2.3.3</td>
<td>932.27</td>
<td>744</td>
<td>2964</td>
</tr>
</tbody>
</table>

Table 4.14: $G^2$ statistics and DIC for Bayesian multinomial small domain models with missing interviews.

The three multinomial Bayesian small domains model in Section 4.2.3 show different levels of goodness-of-fit to our small domain property crime count data with missing interviews. As shown in this analysis, the major problem in modeling is the difficulty in fitting the probabilities of zero crimes and the probabilities of missingness. However, when calculating the crime rates we are focusing on the households that are interviewed and have reported at least one crime. In Chapter 5, we will use the model fitting results to adjust the yearly property crime rate estimates for bounding and time-in-sample effects. In the adjustment procedure, we will use the probabilities of reporting one, two, ..., up to six crimes in each of the interviews other than the second interview and compare them with the corresponding probabilities in the second interview. Thus, the lack of fit in probabilities of reporting zero crimes
and probabilities of missing a response will not greatly affect the reliability of the adjustment in Chapter 5.
Chapter 5: Adjusting household property crime rates using model estimates for time-in-sample and bounding effects

The ultimate goal of all our proposed modeling procedures is to provide adjustments to the yearly crime rate estimates that are published in the Criminal Victimization reports by the Bureau of Justice Statistics each year.

As mentioned at the end of Chapter 3, all of the zero-inflated GLMMs tend to seriously underestimate the household property crime counts. Thus, we concluded that the zero-inflated GLMMs in Chapter 3 are not suitable for adjusting the annual property crime rates. The fits of the multinomial models in Chapter 4 are more reasonable, although not completely satisfactory. We make our adjustments to the annual property crime rates using the multinomial models in Chapter 4.

As previously discussed, the 2nd wave interviews are thought to be the least affected by bounding and time-in-sample bias, thus are believed to give crime rates that are closest to the truth. Therefore, we adjust the household property crime counts reported in each interview in the data for a specific year to a value predicted by our model as if it were in the 2nd wave interview. Similar to the adjustment factors described at the end of Section 1.2.3, which are used by BJS to adjust for replacement households, we use the ratio of the probability of a reported crime being in the 2nd interview to the probabilities of the crime being in the other interviews to account for
the differences between interviews. We are not only adjusting for replacement households but also for every interview of every household in the sample in each year’s data.

In our adjustments, the sequences of interviews for replacement households are counted separately from the previous households that lived in the same housing units. Also, any interview following a missing one is considered to be the first interview to account for the bounding effect. In general, adjusting for the bounding effect will lower the annual property crime rates, while adjusting for the time-in-sample effect for the 3rd-7th interviews will shift the annual crime rates up. It is, therefore, hard to predict what the overall results will be for adjusting for these two effects at the same time. Other than the adjusted annual property crime rates, we also calculate credible intervals for the adjusted annual property crime rates to give a general idea of how the adjusted crime rates compare to the originally reported crime rates.

5.1 Adjustment method

First, we use the multinomial small domain Bayesian model of Section 4.1.2 to adjust the annual property crime rates. We choose this model both because of its fit and the ease of obtaining posterior credible intervals with Bayesian models. The model fitting results in Section 4.1.2 provide the posterior distributions for \((p_{d1}^d, p_{d2}^d, p_{d3}^d, p_{d4}^d, p_{d5}^d, p_{d6}^d, p_{d7}^d)\), the multinomial probability vector for each domain. These probabilities specify how the total number of crimes reported in each domain is expected to be distributed across the sequence of interviews. In other words, \(p_{d2}^d\) is the probability a reported crime in domain \(d\) is reported in the 2nd interview. We sample
from the posterior distributions of \((p_1^d, p_2^d, p_3^d, p_4^d, p_5^d, p_6^d, p_7^d)\) 1000 sets of the probability vectors: \((p_1^{d(k)}, p_2^{d(k)}, p_3^{d(k)}, p_4^{d(k)}, p_5^{d(k)}, p_6^{d(k)}, p_7^{d(k)})\) for \(k = 1, \ldots, 1000\). Thus, the adjustment factor for adjusting a \(t^{th}\) interview to the second interview in domain \(d\) is \(p_2^{d(k)}/p_t^{d(k)}\) for \(t = 1, 2, \ldots, 7\) and \(k = 1, \ldots, 1000\). We will record the reported property count for the \(i^{th}\) record in the data as \(count_i\), the domain of the current household as \(domain_i\), the sequence of the current interview as \(Sequence_i\) and the household-level weight of the current household as \(hh\_weight_i\). Then, the adjusted weighted property crime rate is:

\[
rate^{(k)} = \frac{\sum_i [count_i \times \frac{p_{domain_i}^{d(k)}}{p_{Sequence_i}^{d(k)}}, \times hh\_weight_i]}{\sum_i hh\_weight_i}
\]  

for \(k = 1, \ldots, 1000\). The mean and the percentiles of the 1000 \(rate^{(k)}\)s are then used to calculate the adjusted property crime rate and the corresponding credible intervals.

In Equation 5.1, we are actually adjusting the household survey weights so that when calculating the annual crime rates, all the interviews recorded in the yearly NCVS data can be viewed as if they are the second interview for the corresponding household. The goal is that, after the weight adjustment, the adjusted crime counts have been corrected to account for bounding and rotation group biases.

To prepare our data for making adjustments, we identify the sequence of interviews for each record in the data. The household-level weights are directly available in the NCVS data. The domain index for each household can be easily assigned using the region, MSA and income information. The results for adjustments using the model fitting results from Section 4.1.2 will be shown in Section 5.2.

Next, we will adjust the property crime counts by the multinomial small domain Bayesian models with missing interviews that are described in Section 4.2.3. Among
the three models in Section 4.2.3, Model 4.2.3.1 fits a common model among the 18
domains and captures the general trend of the probabilities of reporting zero crimes
and the probabilities of missing a response. Thus, we decide to use the results from
Model 4.2.3.1 to adjust the property crime rates although the fit of Model 4.2.3.1 is
not the best among the models in Section 4.2.3. Of course, the procedure of adjusting
the crime rates would be the same no matter one of the three models in Section 4.2.3
is used.

In the small domain multinomial Model 4.2.3.1, we have multinomial probabilities
for the \(i^{th}\) interview as follows, describing how the households in domain \(d\) are dis-
tributed into each of the eight count categories (reporting zero crimes, reporting one
crime, reporting two crimes, reporting three crimes, reporting four crimes, reporting
five crimes, reporting six crimes and missing):

\[
(1 - \pi_i) w_i, \quad (1 - \pi_i)(1 - w_i) p_{i1}^{(d)} / p_{i+}^{(d)}, \quad (1 - \pi_i)(1 - w_i) p_{i2}^{(d)} / p_{i+}^{(d)}, \\
(1 - \pi_i)(1 - w_i) p_{i3}^{(d)} / p_{i+}^{(d)}, \quad (1 - \pi_i)(1 - w_i) p_{i4}^{(d)} / p_{i+}^{(d)}, \\
(1 - \pi_i)(1 - w_i) p_{i5}^{(d)} / p_{i+}^{(d)}, \quad (1 - \pi_i)(1 - w_i) p_{i6}^{(d)} / p_{i+}^{(d)}, \quad \pi_i). \quad (5.2)
\]

We here abbreviate the above cell probabilities as \((P0_i, P1_i^{(d)}, P2_i^{(d)}, P3_i^{(d)}, P4_i^{(d)},
P5_i^{(d)}, P6_i^{(d)}, P\text{missing}_i)\). Section 4.2.3 has the details of the notation in the cell
probabilities in Equation 5.2.

From the model fitting results of Model 4.2.3.1, we have samples of posterior
distributions for all the \(w_i\)s, \(\pi_i\)s, \(\beta\)s and \(\alpha\)s. We can use the posterior distributions of
the \(\beta\)s and \(\alpha\)s to calculate the posterior percentiles of the \(\lambda_i^{(d)}\)s and \(p_{ik}^{(d)}\)s. Thus, we
may obtain the samples of the posterior distribution for the cell probability vector
\((P0_i, P1_i^{(d)}, P2_i^{(d)}, P3_i^{(d)}, P4_i^{(d)}, P5_i^{(d)}, P6_i^{(d)}, P\text{missing}_i)\). We then sample from the posterior
distribution of \((P_0, P_1^d, P_2^d, P_3^d, P_4^d, P_5^d, P_6^d, P_{\text{missing}})\) 1000 probability vectors: \((P_{0i}^{(k)}, P_{1i}^{d(k)}, P_{2i}^{d(k)}, P_{3i}^{d(k)}, P_{4i}^{d(k)}, P_{5i}^{d(k)}, P_{6i}^{d(k)}, P_{\text{missing}})\) for \(k = 1, \ldots, 1000\). We here again adjust every record to the 2\textsuperscript{nd} interview by a ratio of the multinomial probabilities. So, the adjustment factors for adjusting a \(t\)\textsuperscript{th} interview to the 2\textsuperscript{nd} interview in domain \(d\) are: \(\frac{P_{0i}^{(k)}}{P_{0i}^{(k)}}, \frac{P_{1i}^{d(k)}}{P_{1i}^{d(k)}}, \frac{P_{2i}^{d(k)}}{P_{2i}^{d(k)}}, \ldots, \text{ and } \frac{P_{6i}^{d(k)}}{P_{6i}^{d(k)}}\) for the number of reported household property crime counts of zero, one, two, ..., six, respectively.

We then use the same procedure to calculate the weighted adjusted crime rates and the corresponding credible intervals as we did when adjusting by Model 4.1.2, i.e., we use the above adjustment factors to substitute for the \(\frac{P_{\text{domain}}^{(k)}}{P_{\text{Sequence}}^{(k)}}\) in Equation 5.1 to calculate the adjusted property crime rate and the corresponding credible intervals.

### 5.2 Adjusting the simple-weighted property crime rate

Before we describe our adjustment procedure, we must introduce the weighting procedure that BJS uses to calculate the annual crime rates published in the “Criminal Victimization” report each year. The published annual crime rates are weighted crime rates with a complicated weighting procedure. For example, the final household weights are the product of five components: base weight, weighting control factor, household non-interview adjustment, first-stage and second-stage ratio estimate factors (for the definition of these weighting factors, see BJS, 2004, “National Crime Victimization Survey: Unbounded Data Codebook”). Before 2006, the first interviews were only used as the bounding interviews, i.e., they were not included in the data when the BJS calculated the annual crime rates. After 2006, as the first interviews were included in the calculation of crime rates, BJS added a sixth adjustment
factor into the weighting procedure to adjust for the bounding effect in the first interviews. Thus, the crime rates published for the years 2006 and later are already adjusted for the bounding effect (but not for the time-in-sample effect).

Our goal is to use the five-factor adjusted weights and adjust for the bounding and time-in-sample effects simultaneously using estimates from our model. Then we could compare the crime rates adjusted by our model to the six-factor adjusted crime rates published by BJS. However, the NCVS data published on the ICPSR website for the years 2006 and later provide the household base weight and the final adjusted household weights (six-factor adjusted weights), but not the five-factor adjusted weights. Since the final adjusted weights are already adjusted for the bounding effect, the NCVS data after 2006 are not suitable for the purpose of our comparison. We also excluded the data for the transition period of the resampling procedure for the NCVS in 2005 to avoid that complexity. For the years 1999-2004, the original public-use data are the bounded versions, i.e., the first interviews are not included. In 2006, BJS made available the unbounded version of the NCVS data for the years 1999-2004 for academic study purposes. The unbounded version of the NCVS data includes first interviews but only with the household base weights. In other words, BJS published the adjusted weights (adjusted for five factors) for the bounded data but not for the unbounded data. For the purpose of this study, however, we will need the unbounded version of the weights to explore the effect of adjustment for the bounding effect. Thus, there is no way for us to make a comparison with the five-factor adjusted crime rates for the unbounded data from 1999 to 2004.

Because of the difficulties in obtaining the proper weights for the data sets of interest to us, we decided to use the base weight only to calculate the simple-weighted
crime rates. Then we adjust the simple-weighted crime rates for the bounding effect and time-in-sample effect by our model to obtain the adjusted crime rates. We compare the adjusted crime rates to the simple-weighted crime rates to show the performance of the adjustment. Since all the work in this thesis so far is based on the 1999-2004 NCVS data, we calculate the adjusted simple-weighted property crime rates and corresponding credible intervals for the years 1999-2004. These crime rates are different from those published by BJS in the “Criminal Victimization” reports for 1999-2004 because our rates used only the base weights instead of the five-factor adjusted weights.

As described in the previous section, we use the multinomial small domain Bayesian model in Section 4.1.2 (Model 4.1.2) and Model 4.2.3.1 to adjust for the bounding and time-in-sample effects in the property crime count data. To do this, we simply calculate the adjusted simple-weighted property crime rates as shown in Equation 5.1. The property crime rates and credible intervals adjusted using estimates from Model 4.1.2 and Model 4.2.3.1 are shown in Figure 5.1. These adjusted crime rates are comparable to those in Figure 2.2 but are lower than the property crime rates published by BJS in Criminal Victimization report each year due to the use of base weights instead of five-factor adjusted weights in our calculation. However, our predicted/adjusted crime rates and credible intervals show the same trends as in the crime rates from BJS.

In Figure 5.1, we can see that the mean crime rates adjusted by the Model 4.1.2 (red solid line) estimates are similar to those adjusted by the Bayesian small domain Model 4.2.3.1 (blue solid line) estimates. The credible intervals for adjustments by
Figure 5.1: Adjusted base-weighted property crime rates and credible intervals for years 1999-2004.
Model 4.1.2 (red dashed lines) and Model 4.2.3.1 (blue dashed lines) both do not cover the unadjusted property crime rates as shown in Figure 5.1.

In Figure 5.1, the adjusted crime rates based on Model 4.1.2 and Model 4.2.3.1 are about 5%-10% higher than the unadjusted household property crime rates (green line) for the years 1999-2004. The adjustments do not change the overall trend of the crime rates over the years. Instead, the overall effect of adjusting for bounding and time-in-sample effects at the same time is to move the estimates of crime rates upward. The 5%-10% increase in the annual property crime rates is substantial compared to the small change in crime rates every year. According to our adjusted results, the bounding and time-in-sample errors do have important impacts on the annual crime rates.

As previously discussed, adjusting for the bounding effect will lower the annual property crime rates, while adjusting for the time-in-sample effect for the 3rd-7th interviews will shift the annual crime rates up. Our adjustment results indicate that the impact of adjustment for the time-in-sample error is greater than the impact of adjustment for the bounding effect. The adjusted results strongly suggest that when using unbounded data to calculate annual crime rates, it is important to adjust for the bounding and time-in-sample effects. Even when using bounded data to calculate the annual crime rates, it is also important to adjust for the time-in-sample effect to obtain more accurate estimates for the annual crime rates. Thus, we suggest that similar adjustment procedures should be carried out in producing the annual crime rates published by BJS in the Criminal Victimization report each year.
Chapter 6: Conclusions and future work

6.1 Conclusions

In this thesis, we studied the bounding effect and time-in-sample effect in the NCVS data. Both of these biases are caused by the survey design and have to be removed to get precise estimates for the annual crime rates in the U.S. The BJS and Census Bureau currently use adjustment factors to account for the bounding effect caused by replacement households; they do not adjust for the time-in-sample effect. We propose adjusting for these two biases simultaneously through a modeling procedure. The modeling analyses provided in this thesis include details on fitting GLMMs and multinomial models to the property crime count data from 36 panel and rotation groups to explicitly or implicitly quantify the effects of bounding and time-in-sample in NCVS.

One of the main challenges in modeling the NCVS crime count data is the excessive number of zero crime counts (98%). To address this modeling problem, we use zero-inflated GLMMs in Chapter 3 to model the NCVS property crime count data. The advantage of zero-inflated GLMMs is their ability to explicitly quantify the effect of each covariate to provide an intuitive view on both the biases (bounding and time-in-sample effects) and the other factors of interest. However, the zero-inflated
GLMMs in Chapter 3 also suffer from underestimating the property crime counts due to the fact that the reported non-zero crime counts are distributed very differently from a Poisson distribution. Multinomial models described in Chapter 4 are based on the aggregated data for each domain instead of the crime counts for individual households, thus we obtain a general idea about how the total number of reported crimes are distributed into each of the seven interview sequences. The small domain models give insight into how the covariates affect the crime counts.

In this thesis, we fit two types of multinomial models depending on how our data are aggregated. The small domains multinomial models in Section 4.1 are developed to model the total number of reported crimes in the seven interviews by 18 domains data. These models have explicit parameters for both the bounding effect and time-in-sample effect, but do not provide a way to deal with missing interviews. In Section 4.2, we include a category for missing data in our summary tables of the number of reported crimes by seven interviews for each of the small domains and build a series of models that have various assumptions on the parameters for probabilities of reporting zero crimes and probabilities of missing an interview.

Finally, in Chapter 5, we provide adjusted estimates of the annual household-level property crime rates by using the model-fitting results from the multinomial Bayesian models to account for the bounding and time-in-sample effects. We compare the adjusted annual property crime rates to the unadjusted property crime rates to evaluate the performance of our adjustments. The adjusted crime rates are about 10% higher than the unadjusted household property crime rates for the years 1999-2004. We conclude that it is critical to adjust for the bounding and time-in-sample effects to obtain more accurate estimates for the annual crime rates.
We modeled property crimes in this thesis as the household-level crimes are more stable and the effects of bounding interviews and time-in-sample effects are more clear in the property crimes than in the violent crimes (Addington, 2005). However, the methods described in this thesis can be easily extended to personal crimes and violent crimes.

6.2 Future work

6.2.1 Fitting additional models

The zero-inflated GLMMs in Chapter 3 assume a zero-truncated Poisson distribution or a distribution (NB) similar to the Poisson for the non-zero property crime counts. However, as our analyses in Section 3.6 shows, the non-zero part of our property crime count data does not quite follow a Poisson distribution. Thus, we may expect that the improvement from other zero-inflated models such as Hurdle models and ZAP models will be limited as they all assume a zero-truncated Poisson distribution for the property crime counts. In this sense, the zero-inflated mixture Poisson model mentioned at the end of Section 3.6 might be promising if we can somehow resolve the instability problem of the parameter estimation. New parameter estimating methods/algorithms which can lead to stable parameter estimation in mixture Poisson models may be a direction for further studies.

Another possible idea for future modeling studies relates to the households that drop out of the sample during the period of $3^{1/2}$ years in the NCVS. As mentioned in Section 4.2.1, these households are an important underlying reason for the increased missingness across the sequence of interviews. However, it is hard to pick out the drop-out households from the organization of our data in Section 4.2.1 (Table 4.8).
From the missing-data patterns in Table 4.7, there are 201 households that dropped out starting from the second interview (missing pattern No. 11), 199 households dropped out starting from the third interview (missing pattern No. 12), 212 households dropped out starting from the fourth interview (missing pattern No. 10), 298 households dropped out starting from the fifth interview (missing pattern No. 8) and 405 households dropped out starting from the sixth interview (missing pattern No. 4). These drop-outs account for 27.8% of the total number of households with missing interviews. Separating these drop-out households from the other households that miss a certain interview (the last column in Table 4.8), we can re-summarize our property crime count data with missing interviews as in Table 6.1.

<table>
<thead>
<tr>
<th>Seq</th>
<th># of crimes reported</th>
<th>cumulative drop-out</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20768 2001 275 68 20 6 1 0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>20604 1506 194 37 10 5 0 582</td>
<td>201</td>
</tr>
<tr>
<td>2</td>
<td>20546 1290 163 31 12 1 1 695</td>
<td>400</td>
</tr>
<tr>
<td>3</td>
<td>20312 1252 133 30 6 1 0 793</td>
<td>612</td>
</tr>
<tr>
<td>4</td>
<td>20078 1164 132 25 4 3 0 823</td>
<td>910</td>
</tr>
<tr>
<td>5</td>
<td>19646 1146 135 16 2 4 0 875</td>
<td>1315</td>
</tr>
<tr>
<td>6</td>
<td>19432 1092 115 23 3 1 0 2473-</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 6.1: Number of households by number of reported crimes and interview sequence.

The only difference between Table 6.1 and Table 4.8 is that we are splitting the last column in Table 4.8 into the last two columns in Table 6.1. The last column in Table 6.1 is the cumulative number of households that have dropped out of the sample until the current interview. The “missing” column in Table 6.1 is the number of
households that are not drop-outs but have missed the current interview. The major problem in Table 6.1 is that it is almost impossible to separate those households that responded for the first six interviews and missed the last interview from those that dropped out of the sample in the last interview. The NCVS data provide a code for the type of missingness which could be helpful in identifying some of the drop-outs. However, it is also very hard for the interviewer to decide the type of missingness for some households. Since we do not have any more interviews following the seventh interview, there is no way to tell whether these 822 households (missing pattern No. 2 in Table 4.7) would return to the sample if they were contacted to be interviewed more than seven times. Thus, it is very difficult to determine the values in the last two columns of the seventh row in Table 6.1. The only thing we know is that these two cells should sum to 2473 so that the sum of each row in Table 6.1 equals the total number of households in our sample (23139). In fact, we cannot really separate the last two cells even for the 2nd to 6th interviews in Table 6.1. However, a household that missed all the interviews starting from the 2nd interview is much more likely to be a drop-out than a household that missed all the interviews starting from the 6th interview. Thus, we are more comfortable with separating the last two cells in Table 6.1 in an earlier interview than in a later interview.

From Table 6.1 we can see that the number of drop-out households increases as the interview sequence increases. From the second to the last column in Table 6.1, the increasing trend for the number of households that missed an interview does not change even after excluding the drop-out households. Summarizing our property crime count data as in Table 6.1 allows us to consider models that target the drop-outs. Taking Table 6.1 as a start, future studies on the drop-out households may include
finding a possibly better way to separate the drop-out households from the rest of the sample and looking for appropriate statistical models to describe the summarized data in Table 6.1. Multinomial count models similar to those described in Chapter 4 might be a good place to start. Diggle et al. (2002) and Molenberghs and Verbeke (2005) also provide references on drop-out models for longitudinal data.

6.2.2 Model diagnostics

As discussed in Chapter 3, DIC and predicted crime rates are used to assess the goodness of model fit of our Bayesian zero-inflated GLMM count models. Other possible diagnostics for Bayesian GLMMs include the Bayes’ factor, EAIC (Expected Akaike Information Criterion, see Brooks, 2002) and EBIC (Expected Bayesian Information Criterion, see Carlin and Louis, 1996). We can also check the over-dispersion index and the sample skewness (Neelon et al., 2010) to determine the model’s fit to the data. Further investigation of our models using these diagnostic criteria is a potential part of the future work. Our goal is to find enough evidence for the fit of our crime count models.

For multinomial models discussed in Chapter 4, we used $X^2$ and $G^2$ statistics as measures of fit to compare our predicted data tables to the observed crime count tables. However, complex survey data such as the NCVS data we used in this study usually do not satisfy the independence assumption required so that the $X^2$ and $G^2$ statistics have asymptotic $\chi^2$ distributions. Also, these statistics are not typical measures of fit for Bayesian models. Thus, other methods of measuring the goodness of fit for Bayesian multinomial models on complex survey data may be of interest for future studies.
In summary, studies about bounding and time-in-sample biases in the NCVS can help BJS obtain more realistic crime-rate estimates. Moreover, the methods discussed in this thesis are also applicable to other complex surveys with rotating panel design to adjust the weights of the interviewed individuals to obtain better estimates from the survey data. With the continued greater use of complex survey data, this kind of study is of general interest and should be continued in the future.
References


