A Poisson mixed modeling approach to longitudinal multinomial data of varying cluster sizes

by

Zheng Fu

Bachelor of Science, Northwestern Polytechnical University, 1999
Master of Science, Northwestern Polytechnical University, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

In the Graduate Academic Unit of Mathematics and Statistics

Supervisor(s):	Guohua Yan, Ph.D, Statistics
	Renjun Ma, Ph.D, Statistics
Examining Board:	James Watmough, Ph.D, Mathematics, Chair
	M. Tariqul Hasan, Ph.D, Statistics
	Elaine Perunovic, Ph.D, Psychology
External Examiner:	

This thesis is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

May, 2020

© Zheng Fu, 2020

Abstract

Longitudinal studies and categorical data analysis are commonly used in many areas, such as medicine, public health, and psychology. Researchers have developed many analytic methods for longitudinal data and categorical data separately. However, the methods for analysing data with both longitudinal and categorical properties are sparse. This thesis proposes a baseline-category model for nominal data and a continuation-ratio logit model for ordinal data, which are both constructed from a set of Poisson mixed models. Three levels of random effects are introduced to account for the effects of the subjects, the time and the categories. Since the models are rooted from Poisson mixed models, they can give both proportion and count inference. Additionally, the 3 levels of random effects are very flexible to handle different data sets and to fit different research interests.

Dedication

This thesis is dedicated to my wife, Jin Qiu, who makes great effort to take care of the family, to my son who gives me support and love, and to my daughter who makes me happy every day.

I want to use this thesis to record my first 3 years of life in Canada. Academic achievements, excitement of travelling, the challenges of fitting into the new culture, and the happyness and difficulties of family life are the meaningful memories from this period.

Acknowledgements

I'd like to express my sincerest gratitude to my supervisor, Dr. Guohua Yan. His help and patience were consistent throughout my entire graduate experience, from the degree application to career guidance. His timely replies and weekly reviews are the most important things that helped me to move the thesis forward. Without this support my statistics degree would not have been completed.

I would like to express my sincere gratitude to my other supervisor, Dr. Renjun Ma. His deep understanding and clear instruction make all his classes inspiring. This thesis is motivated by the data set given to me by him, and the model was developed step by step under his encouragement. Without these, the thesis would not be possible.

I'd like to thank Dr. Jeff Picka for all his efforts spent on my seminar. His critical thinking and presentation skills demonstrated during my seminar rehearsals helped me not only for the seminar but also for my future career. I also want to thank Dr. Hasan for his kindness and support while I gave my graduate seminar.

Finally, I want to thank the University of New Brunswick and the Department of Mathematics and Statistics for allowing me to take the program. The excellent environment here made my study enjoyable.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is motivated by the data set of the crash counts for over one thousand major signalized intersections in the city of Chicago from 2004 to 2010, which were provided by [25]. For each intersection, vehicular crashes were counted by crash severity levels, including fatal (FAT), injury Types A (IA), B (IB), and C (IC) for major, moderate, and minor injury levels, property damage only (PDO), and unknown. We can see that the data set demonstrates both longitudinal and categorical properties.

Longitudinal studies are commonly used in many areas, such as medicine, public health and psychology. Longitudinal data, in contrast with cross-sectional data, observes the same sample at different time points. It has some advantages over cross-sectional studies, which are introduced in section 2.1. Researchers have spent a great deal of time deriving methods for analyzing longitudinal data. Several general approaches have been described to model longitudinal data [3][4][5], like ANOVA, the generalized estimating equations (GEE), and the generalized linear mixed model (GLMM).

Categorical data analysis is another widely used method in the biomedical and social sciences[8]. Categorical variables represent types of data which may be divided

into groups. They have two main types of measurement scales: ordered scales and unordered scales. Categorical variables having ordered scales are called ordinal variables, like the appraisal of a company's inventory level (low, about right, high). The variables with unordered scales are called nominal variables; for example a favorite type of music (classical, country, folk, jazz, rock). Many methods have been developed to analyze these data, like the logistic regression model and its extension for binary and multicategory responses, and the contingency tables for analyzing associations.

Numerous studies exist on categorical data and longitudinal data respectively. However, the analysis of this kind of data with both properties is not adequately addressed. Currently, the majority of the research methods analyze data from the angle of categorical data in the longitudinal setup, the analysis is then based on the logit models. The methods developed, basically, can be separated into two groups. The first group is transitional model[8] that includes past observations as predictors in the logit models. For example, the Markov chain model[2], in which only the past observation is included as a predictor, is a transitional model. The second group is the logistic generalized linear mixed models[8] which involves random effects in the logit models to account for the correlation of the responses for the same subject.

We can see that, since these two groups of models are rooted from logit models, they can only get the results regarding to the proportions rather than the counts. For the transitional model, as the past observation is included in the covariates, the effects of the rest covariates of interest are attenuated. For the logistic generalized linear mixed model, which is a kind of generalized linear mixed model, the distribution assumption of the random effects imposes restrictions about this model. Furthermore, it is not always simple to fit. As for the usual maximum likelihood estimation algorithm, in order to get the likelihood function for the parameters the random effects need to be integrated out. The calculus-based integral used to average with respect

to the distribution of the random effects does not have a closed form. Numerical methods for approximating it can be computationally intensive[8].

Therefore, we propose a baseline-category model for nominal data and a continuation-ratio logit model for ordinal data. These are constructed from a set of Poisson mixed models for the data set with both longitudinal and categorical properties. These models could give us the results regarding both proportion and count. Furthermore, the Poisson mixed models are rooted from the dual-Poisson model introduced in [10], which is a GLMM. However, an orthodox best linear unbiased prediction (BLUP) approach[17] is used to do the parameter estimation and the random effects prediction for this model. We obtain an optimal estimating equation based on the orthodox BLUP, which is solved by a modified Newton Algorithm. One advantage of this approach is that only the first and second moment structures of the random effects are needed, which gives greater flexibility to the model. And this approach leads to improved computational efficiency since all the formulas are explicit in each iterative step.

According to the nature of the data set, the subject-specific random effects, the time-specific random effects, and the category-specific random effects are introduced to account for the effects of the intersection, the year an observation was taken, and the category. These all impact the crash count. Meanwhile, these random effects account for the correlation within the responses, which is the key property of the longitudinal data. Finally, a set of Poisson models, which are conditional on the random effects are proposed to model the counts. By organizing the responses and covariates in certain ways [26], the equivalent multicategory logit models are constructed by the set of the conditional Poisson models[11].

The results of the analysis tells us that, over time, the number of property damage only crashes will increase while other categories will be reduced and the possibility of getting into a severe accident is lower. By analyzing the subject-specific random

effects, certain intersections, which could have other effects that may dramatically impact the number of accidents, are identified.

The remaining part of this thesis is organized as follows: Chapter 2 presents an introductory overview of statistical knowledge and models for this thesis including: the longitudinal studies, the multicategory logit models, the Poisson mixed model, and the scatterplot smoothing which is used to pretreat the covariates.

The first part of Chapter 3, from section 3.1 to section 3.7, introduces the Poisson mixed model. In this part, four key assumptions are proposed for the model. The moment structures of the responses and the random effects are derived. The orthodox best linear unbiased predictors of the random effects, as well as the iterative equations for estimating regression and random effects parameters are also developed. In section 3.7, the iteration algorithm is introduced to give a step by step instruction of the estimation process, which could guide the code development. The second part includes section 3.8 and 3.9, which introduces how to construct the equivalent multicategory logit models by the Poisson mixed models. In this part, the baseline-category logit model and the continuation-ratio logit model are introduced. As 3 levels of random effects are proposed in our model, it can be adjusted to fit in different data sets and research interests. This model expansion is discussed in section 3.10.

To demonstrate the proposed model, we analyze the vehicular crashes data set in chapter 4. Section 4.1 gives an initial statistics analysis and points out how to select the covariates. In section 4.2, the baseline-category logit model and continuation-ratio logit model are constructed by the Poisson mixed models to do the analysis. Analysis results, along with relevant graphs and figures, are presented and discussed. In addition, we conduct a brief simulation study to assess the behaviour of our estimating algorithm in section 4.3.

We conclude the thesis in Chapter 5 with a brief discussion of our proposed model

and a few aspects for further research.

Chapter 2

Preliminaries

In this chapter, the basic knowledge and statistics methods which are used in this thesis are introduced. The crashes data set used in this thesis contains 2 major properties: longitudinal and multinomial properties. Section 2.1 introduces longitudinal studies which list out a set of methods used to analyze longitudinal data. Section 2.2 focuses on logit models, which are broadly used in multinomial data analysis. Two sub logit models: the baseline-category logit model and the continuation-ratio logit model, are introduced in this section to handle nominal and ordinal data respectively. Then, in section 2.3, the Poisson mixed model is introduced. It shows a way to combine longitudinal analysis and logit models to analyze the data set with both longitudinal and multinomial properties using hierarchical random effects and conditional Poisson distribution. In the last section, locally weighted scatterplot smoothing (LOWESS) is introduced, which is used to pretreat the raw data. It uses a smoothing spatial surface to create an explanatory variable which represents location parameters (latitude and longitude) that otherwise are not feasible to be included in the mixed Poisson model.

2.1 Longitudinal study

Longitudinal studies are commonly used in many areas, such as medicine, public health and psychology. Longitudinal data, in contrast with cross-sectional data, observes the same sample at different time points. It has some advantages over cross-sectional studies. First, because the repeated measurements from the same subject are not perfectly correlated, the net result is that these measurements provide more independent information than a single measurement obtained from a single subject, which makes longitudinal studies more powerful than cross-sectional studies for a fixed number of subjects.

Second, in a longitudinal study, each subject can serve as its own control. Because intra-subject variability is usually smaller than inter-subject variability, excluding inter-subject variability from measurement errors makes the statistics test more powerful.

Finally, longitudinal data can provide information about individual change, and then the statistical estimates of individual trends can be used to understand heterogeneity in the population and the determinants of change at the individual level[3].

However, due to the correlation between measurements on the same subject, longitudinal data not only brings statistical power, but also analysis complexity. Different methods and models are discussed by [4], like generalized mixed-effects regression models which can handle missing data and irregularly spaced measurement occasions, and marginal models that use Generalized Estimating Equations(GEE) [5].

To give a basic illustration of longitudinal studies, the usual longitudinal data layout is as follows:

Table 2.1: Longitudinal data set layout

ID	Time point	Response	Explanatory variables
1	1	y_{11}	x_{111}.....x_{11p}
.	.	.	.
1	t_1	y_{1t_1}	x_{1t_11}.....x_{1t_1p}
2	1	y_{21}	x_{211}.....x_{21p}
.	.	.	.
2	t_2	y_{2t_2}	x_{2t_21}.....x_{2t_2p}
.	.	.	.
.	.	.	.
n	1	y_{n1}	x_{n11}.....x_{n1p}
.	.	.	.
n	t_n	y_{nt_n}	x_{nt_n1}.....x_{nt_np}

y_{it}: the response of t^{th} time point of the i^{th} subject;
x_{itk}: the k^{th} explanatory variable for the t^{th} time point measurement of the i^{th} subject;
the index of subjects i=1…m;
the index of time points t=1...t_i;
the index of explanatory variables k=1...p.

2.1.1 Analysis approaches

As mentioned above, there are many advantages to longitudinal data analysis. Thus, many analysis methods have been developed. In [6], Student did the analysis of longitudinal data via the t-test, which reduces the repeated measurements into a single summary variable. This summary variable cannot fully represent the repeated measurements in terms of the statistical power. Analysis of variance (ANOVA) is another historical way to handle longitudinal data. However, the assumption made in this approach: constant variances and covariances over time, make it not feasible for the primary portion of longitudinal data. Details of these methods and some other extensions are introduced by [7].

In the later paragraphs, two widely accepted methods, generalized estimating equations (GEE) and generalized mixed-effects regression model, will be introduced. Both of these can be treated as extensions of generalized linear models by accounting for

the within subject correlation via specification of the correlation structure and random effects respectively.

A basic feature of GEE models is that only the marginal distribution of Y_{it} at each timepoint need to be specified. For example, if there are two timepoints and the responses are continuous normal, GEE would require us to assume that the distribution of Y_{i1} and Y_{i2} are two univariate normals, rather than assuming that they are from a bivariate normal distribution. A related feature of GEE models is that the covariance structure is treated as a nuisance, and the focus is on the regression of Y on X. Then, there are four specifications used for GEE models:

(1). the marginal expectation of the response, $E(Y_{it}|X_{it})=\mu_{it}$ depends on the covariates X_{it},

(2). the correlation between Y_{it} and the covariates X_{it} follows the equation $g(\mu_{it})=x_{it}^T\beta$, where $g(.)$ is a known function,

(3). the variance of each response $V(Y_{it})=\phi\nu(\mu_{it})$, where ν is a known variance function and ϕ is a scale parameter that may be known or estimated,

(4). the correlation structure of the repeated measurement R is an n x n matrix which depends on a vector of association parameters, denoted as α.

GEE models the mean and correlation structure separately. In other words, the mean response depends on the explanatory variables only rather than the random effects or previous responses. This implies that misspecification of R has limited impact on the estimation of the regression coefficients, which is one advantage of GEE.

Another widely used approach to the analysis of longitudinal data is the generalized linear mixed model (GLMM). They are considered an extension of two modelling methods: linear mixed models and generalized linear models. Linear mixed models present a useful extension of linear regression models by including both fixed effects

and random effects since observations are dependent on different kinds of correlation. Random effects can be treated as the reflection of natural heterogeneity for the reason that many factors have not been measured. On the other hand, the generalized linear models not only cater to the need of continuous response variables, but also are capable of handling the discrete response variables. In contrast to GEE, GLMM accounts for the within subject correlation by incorporating subject-specific random effects with the fixed effects. Here are the specifications used for GLMM models:
(1) the conditional mean of each response is $E(Y_{it}|b_i)=\mu_{it}$ where b_i is the random effect for subject i,
(2) the conditional mean of Y_{it} is assumed to depend on fixed and random effects via the following linear predictor:

$$g(\mu_{it}) = x_{it}^T\beta + z_{it}^T b_i,$$

where x_{it} and z_{it} are the vectors of the covariates with fixed and random effects, and g(.) is a known link function,
Regression estimates from this specification are termed "subject-specific", in contrast with the GEE model, to emphasize that they are conditional on random effects.

2.2 Multicategory logit models

Logistic regression is used to model binary response variables. Generalizations of it can be used to analyze categorical responses with more than two categories. The traffic accident data set used in this thesis has an ordinal property for the accident categories. And meanwhile, the ordinal data set can be analyzed as nominal data, even though this may result in a certain loss of power because of not using the information about that ordering [8]. So, two multicategory logit models, baseline-category logit model and continuation-ratio logit model, are introduced here to handle nomi-

nal responses and ordinal responses respectively.

2.2.1 Baseline-category logit model

Let J denote the number of categories for Y. Let π_1, ..., π_J denote the response probabilities, satisfying $\sum_1^J \pi_j = 1$. With n independent observations, the probability distribution for the number of outcomes of the J types is the multinomial. When the last category J is the baseline, the baseline-category logit model with the predictors x is

$$\log \frac{\pi_j}{\pi_J} = \alpha_j + x^T \beta_j, \ j = 1, ..., J-1$$

Given that the response falls in category j or category J, the log odds suggest that the response will be j. There are J-1 equations with different parameters for each, and the effects vary according to the category paired with the baseline.
To estimate each category's respective probability, the following equation is used:

$$\pi_j(x) = \frac{\exp(\alpha_j + x^T \beta_j)}{1 + \sum_{h=1}^{J-1} \exp(\alpha_h + x^T \beta_h)}, j = 1, ..., J-1$$

Where α_J=0 and β_J=0. The denominator is the same for each probability, and the numerators for various j sum to the denominator. So, the sum of all the probabilities is 1. Since the response variables are nominal, and the model setup does not have specific requests for the baseline, we can select any category as the baseline-category.

2.2.2 Continuation-ratio logit model

When response categories are ordered, a continuation-ratio logit contrasts each category with a grouping of categories from higher levels of the response scale. We first set the probabilities of the response equal to j as follows, where j corresponds to the j^{th} category:

$$P(Y = j) = \pi_j, \ j = 1, ..., J$$

And, then J - 1 continuation-ratio logits are generated in order to complete a continuation-ratio logit model as follows:

$$\log\left[\frac{\pi_j}{\pi_{j+1}+...+\pi_J}\right] = \alpha_j + x^T\beta_j,\ j = 1, ..., J-1$$

Given responses falls in j, ..., J, a continuation-ratio logit for category j is same as a binary logistic regression model in which category j forms a single category and categories j + 1,...,J form a second category. The parameter β_j represents the effect of x on the log odds of response in category j.

There are other logit models can handle ordinal categorical data, such as the adjacent-categories logit model and the cumulative logit model. [9] gives a detailed introduction. I will not introduce them one by one here.

2.3 Dual-Poisson mixed model

This work's approach is applying GLMM to Poisson distribution. Since GLMM has been introduced previously, this section talks about mixed Poisson modeling starting with the prototype: the dual-Poisson mixed model.

The dual-Poisson mixed modeling method is introduced in [10] to take account of both intra-cluster correlation and extra-variation arising from the random cluster sizes in clustered binary data analysis. The model setup is as follows:

(1) There are m pairs of observations (S_i,N_i), indexed by i=1, ..., m. S_i is the number of successes, and N_i is the number of total trials (the cluster).

(2) U_1, ..., U_m are cluster level positive random effects, which are independent and identically distributed with the two moments below:

$E(U_i)=1$, and $Var(U_i)= \sigma^2$

(3) U_{i1} and U_{i2} are positive random effects for the numbers of successes and failures of cluster i respectively. Given the cluster level random effects (U_1, ..., U_i)=U, U_{11},

U_{12}, ..., U_{m1}, U_{m2} are independent and identically distributed with the following two moments:

$E(U_{ij}|U)=U_i$, and $Var(U_{ij}|U)= \nu_j^2$, where j=1,2

(4) Then set up the dual Poisson mixed models whose responses are $Y_{i1}= S_i$, $Y_{i2}=N_i$-S_i. And given $U_{**}= (U_1,..,U_m,U_{11}, U_{12},...,U_{m1}, U_{m2})$, $Y_{11},Y_{12},...,Y_{m1},Y_{m2}$ are independent and identically distributed with the following distribution:

$$\begin{cases} Y_{i1}|U_{**}\sim \text{Poisson}\{U_{i1}\exp(z_i^T\alpha+x_i^T\beta)\} ; \\ Y_{i2}|U_{**}\sim \text{Poisson}\{U_{i2}\exp(z_i^T\alpha)\} ; \end{cases} \tag{2.1}$$

where z_i and x_i are known covariates vectors of cluster i, and normally they are the same.

According to [11], we obtain

$$Y_{i1}|Y_{i1}+Y_{i2}=N_i,\ U_{**}\sim \text{Binomial}(N_i, \pi_i) \tag{2.2}$$

Where

$$\begin{aligned} \pi_i &= \frac{U_{i1}\exp(z_i^T\alpha + x_i^T\beta)}{U_{i1}\exp(z_i^T\alpha + x_i^T\beta) + U_{i2}\exp(z_i^T\alpha)} \\ &= \frac{U_{i1}\exp(x_i^T\beta)}{U_{i1}\exp(x_i^T\beta) + U_{i2}} \end{aligned}$$

The equivalent binomial model is given by

$$\text{logit } \pi_i = x_i^T\beta + \text{logit}\frac{U_{i1}}{U_{i1}+U_{i2}}$$

This report extends this model to multinomial data with random cluster size by fitting multiple random effects Poisson models following the existing methods [12].

2.4 Scatterplot smoothing

Not every data set or trend can be modeled by traditional parametric techniques. In other words, it is impossible to describe every trend with a formula [13]. The scatterplot smoothing technique is one alternative way to highlight the underlying trend in the data set. Furthermore, the smoothing technique can not only fit a line but also can fit a surface: one response variable vs. two (or more) explanatory variables [14].

The data set used in this thesis is the vehicle crash counts in the city of Chicago over one thousand major signalized intersections from 2004 to 2010. All the intersections are identified by latitude and longitude. When plotting out the crash counts versus latitude and longitude, it impossible to link these variables with one formula. So it is unreasonable to include latitude and longitude directly in the explanatory variables when crash counts are the response in further analysis. However, this link is one major interest of the analysis. Now, scatterplot smoothing is used to summarize the location parameters into one parameter which correlated with the response (crash counts). This will be introduced in details in later sections.

There are many scatterplot smoothing techniques, like penalized spline regression, local polynomial fitting and series-based smoothers[13][15]. Because the smoothing needed here is for pre-treatment of the explanatory variable, and the data set has a longitudinal property(7 years result over one location), this thesis selects the most robust and computationally easy method: the locally weighted scatterplot smoothing (LOWESS)[13].

2.4.1 Locally weighted scatterplot smoothing (LOWESS)

Here I will briefly introduce LOWESS with the basic smoothing line, which can be easily expanded to smoothing surface by increasing explanatory variables to two or more.

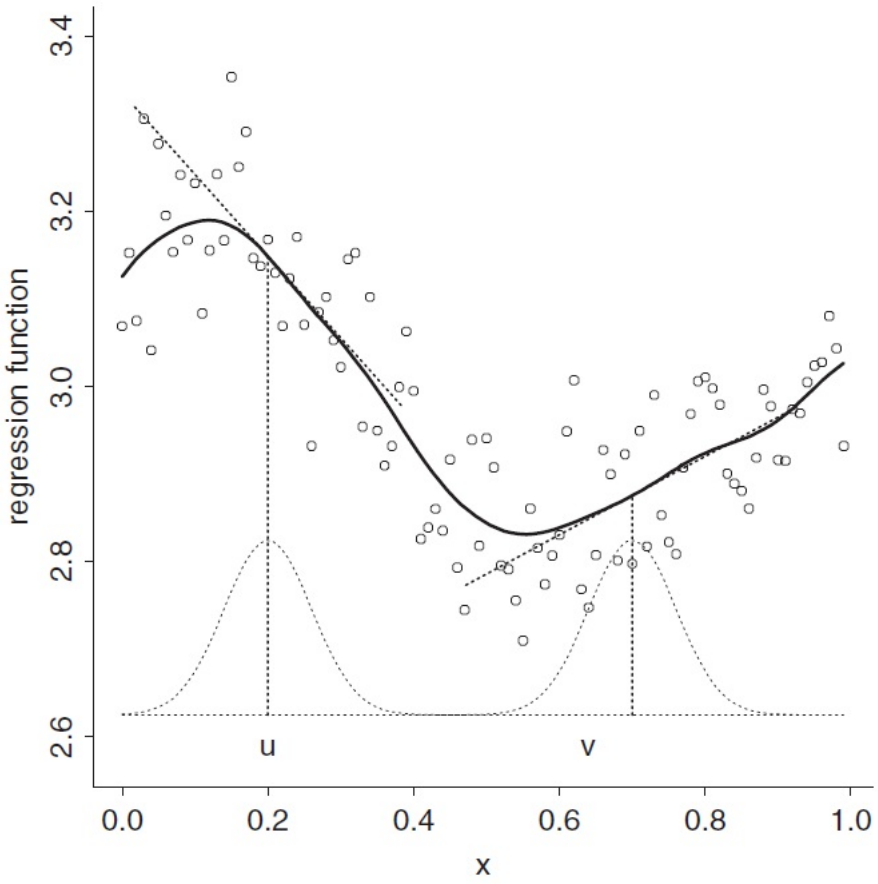

Figure 2.1: An illustration of how LOWESS curve is constructed [13]

The figure above illustrates the basic idea of LOWESS. To get a smooth(expected) point at x=u is to get the weighted least squared line at this point. The weight distribution is the dotted curve shown below the points. For a point at x=v, it is the same. Following this process, get all smoothed values over a grid of x values and connect them to get the solid smooth line. Two things should be noted here.

First, the least squared line p can be any degree, but normally the quadratic line is enough. Second, the weight function K(x) is usually positively symmetric, and its value decreases as |x| increases. The principle smoothing process is as below:
The local line is

$$E(y_i) = \beta_0 + \beta_1(x_i - x) + ... + \beta_p(x_i - x)^p$$

where x_i (i =1,...n) is x value of each observation and y_i (i =1,...n) is the response versus the value.
The weight function is

$$K\{b^{-1}(x_i - x)\}$$

where b>0 is the bandwidth which controls the smoothness.
Then we calculate $\hat{\beta}_0$,...,$\hat{\beta}_p$ which minimizes

$$\sum_{i=1}^{n}\{y_i - \beta_0 - \beta_1(x_i - x) + ... - \beta_p(x_i - x)^p\}^2 K\{b^{-1}(x_i - x)\}.$$

The smooth estimate at point x is $\hat{\beta}_0$.

Chapter 3

A Poisson mixed modeling approach to longitudinal multinomial data of varying cluster sizes

This chapter introduces the Poisson mixed model for longitudinal multinominal data. The model makes four assumptions about the response and random effects and assumes a hierarchical structure for them. The model includes regression parameters, which characterize how the average response behaves in the population, and random effects parameters, which model how individuals, time and category deviate from this population behaviour. There is also a set of correlation parameters since, as longitudinal data, the response is assumed to be correlated with previous measurements[16]. For sections 3.1-3.7, the estimation process is introduced. It begins by introducing the orgnization of the responses and the four assumptions made by the model, then derives the marginal moments of the random effects and response from the previously defined conditional moments. The orthodox best linear unbiased predictor

(BLUP)[17], a predictor for the random effects, is introduced after this and will be critical to parameter estimation and forecasting future values. Next, the method for estimating the regression parameters will be discussed. Then, estimators for the random effects parameters will be introduced based on adjusted Pearson estimators. Finally, in section 3.7, the iteration algorithm, which links sections 3.1-3.6, is introduced.

For sections 3.8 and 3.9, the baseline-category logits model and the continuation-ratio logit model are introduced by their equivalent Poisson mixed models. Finally in section 3.10, how this model can be expanded to handle different data sets and analysis interests[18] are briefly introduced.

3.1 Model specification

This section will lay out the four important assumptions on both the random effects and response of the Poisson mixed model in question. The assumptions made on the random effects concern only the first two moments, which gives some robustness to the model[17]. The response is assumed to be Poisson distributed. Let Y_{itj} represent the response of subject i at time point t for category j (i=1,...m; t=1,...T; j=1,..J). Then all responses for subject i are Y_i=(Y_{i11},...Y_{iT1},...,Y_{i1J},...Y_{iTJ})^T, and all responses are Y=(Y_1^T,...,Y_i^T)^T. There are three levels of random effects introduced here to account for the subject-specific effects, the time-specific effects and the category-specific effects:

(1). the subject-specific random effect for intersection i: U_i, and the vector for all subject-specific random effects: U = (U_1,...U_m)^T;

(2). the time-specific random effect for subject i at time point t: V_{it}, and the vector for all time-specific random effects: V = (V_{11},...V_{1T},...,V_{m1},...V_{mT})^T;

(3). the category-specific random effect for subject i at time point t of category j:

W_{itj}, and the vector for all category-specific random effects:

$W = (W_{111},...,W_{1T1},...,W_{11J},...,W_{1TJ},...,W_{m11},...,W_{mT1},...,W_{m1J},...,W_{mTJ})^T$.

Then, the model is characterized by four assumptions.

The first assumption concerns the subject-specific random effects U_1,..., U_m, which are positive, independently and identically distributed with mean 1 and variance σ^2:

$$E(U_i) = 1 \; and \; Var(U_i) = \sigma^2.$$

The second assumption concerns the conditional distribution of the time-specific random effects, conditioned on the subject-specific random effects. Given the subject-specific random effects U, the moment structure of the positive time-specific random effects V can be expressed as:

$$E(V_{it}|U) = U_i \; and \; Cov(V_{it}, V_{i't'}|U) = \begin{cases} \tau^2\rho^{|t-t'|}, & \text{if } i = i'; \\ 0, & \text{if } i \neq i'; \end{cases}$$

The parameter $\rho^{|t-t'|}$ represents a AR(1)[19] correlation structure.

The third assumption concerns the conditional distribution of the category-specific random effects, conditional on the subject and time specific random effects. Given the subject and time specific random effects (U,V), the moment structure of the positive category-specific random effects W can be expressed as:

$$E[W_{itj}|(U,V)] = V_{it} \; and \; Var[W_{itj}|(U,V)] = w_j^2 V_{it}$$

The last assumption deals with the distribution of the response variable, Y. Given the random effects (U,V,W), the components of Y are conditionally independent, and the conditional distribution of Y_{itj}, given (U,V,W), depends on W_{itj} only, which is:

$$Y_{itj}|(U,V,W) \; \sim \; \text{Poisson}(W_{itj}\mu_{itj})$$

where

$$\mu_{itj} = \exp(x_{it}^T \beta_j)$$

with the vector of covariates x_{it} and the regression parameter vectorβ_j.

3.2 Moment structure

The methods used to estimate the parameters in the model introduced in the later sections require marginal expected values and covariances of the response and random effects. Since only the conditional moments were defined for the response and random effects, these will need to be determined before the model can be estimated from data. The following identities, known as the Law of Total Expectation and the Law of Total Covariance respectively[20], reduce this problem to straight forward algebra.
If X, Y and Z are any random variables, then:

$$\begin{cases} E(Y) = E(E(Y|X)), \\ Var(Y) = E(Var(Y|X)) + Var(E(Y|X)), \\ Cov(Y,Z) = E(Cov(Y,Z|X)) + Cov(E(Y|X), E(Z|X)). \end{cases}$$

provided that the moments exit.

3.2.1 Expectations

The expectation of V_{it}:

$$E(V_{it}) = E(E(V_{it}|U) = E(U_i) = 1$$

The expectation of W_{itj}:

$$E(W_{itj}) = E(E(W_{itj}|(U,V)) = E(V_{it}) = 1$$

The expectation of Y_{itj}:

$$E(Y_{itj}) = E(E(Y_{itj}|(U,V,W))) = E(W_{itj}\mu_{itj}) = \mu_{itj}E(W_{itj}) = \mu_{itj} \tag{3.1}$$

3.2.2 Variance and covariance

The calculation of the unconditional variance of the responses Y_{itj}:

$$\begin{aligned}
Var(Y_{itj}) &= E[Var(Y_{itj}|(U,V,W))] + Var[E(Y_{itj}|(U,V,W))] \\
&= E(\mu_{itj}W_{itj}) + Var(\mu_{itj}W_{itj}) \\
&= \mu_{itj}E(E(W_{itj}|(U,V))) + \mu_{itj}^2 Var(W_{itj}) \\
&= \mu_{itj} + \mu_{itj}^2\{E(Var(W_{itj}|(U,V))) + Var(E(W_{itj}|(U,V)))\} \\
&= \mu_{itj} + \mu_{itj}^2\{E(w_j^2 V_{it}) + Var(V_{it})\} \\
&= \mu_{itj} + \mu_{itj}^2\{E(w_j^2 V_{it}) + E(Var(V_{it}|U)) + Var(E(V_{it}|U))\} \\
&= \mu_{itj} + \mu_{itj}^2(w_j^2 + \tau^2 + \sigma^2)
\end{aligned}$$

The calculation of the unconditional covariance of the responses Y_{itj} and $Y_{i't'j'}$:

$$\begin{aligned}
Cov(Y_{itj}, Y_{i't'j'}) &= E\{Cov(Y_{itj}, Y_{i't'j'}|(U,V,W))\} + Cov\{E(Y_{itj}|(U,V,W)), E(Y_{i't'j'}|(U,V,W))\} \\
&= 0 + Cov(\mu_{itj}W_{itj}, \mu_{i't'j'}W_{i't'j'}) \\
&= \mu_{itj}\mu_{i't'j'}Cov(W_{itj}, W_{i't'j'}) \\
&= \mu_{itj}\mu_{i't'j'}\{E(Cov(W_{itj}, W_{i't'j'}|(U,V)) + Cov\{E(W_{itj}|(U,V)), E(W_{i't'j'}|(U,V))\} \\
&= \mu_{itj}\mu_{i't'j'}Cov(V_{it}, V_{i't'}) \\
&= \mu_{itj}\mu_{i't'j'}\{E(Cov(V_{it}, V_{i't'}|U)) + Cov(E(V_{it}|U), E(V_{i't'}|U))\}
\end{aligned}$$

a. If i≠i',then:

$$Cov(Y_{itj}, Y_{i't'j'}) = 0$$

b. If i=i' and t≠t',then:

$$Cov(Y_{itj}, Y_{i't'j'}) = \mu_{itj}\mu_{it'j'}(\tau^2\rho^{|t-t'|} + \sigma^2)$$

c. If i=i', t=t' and j≠j',then:

$$Cov(Y_{itj}, Y_{i't'j'}) = \mu_{itj}\mu_{itj'}(\tau^2 + \sigma^2)$$

The calculation of the unconditional covariance of the responses U_i and $Y_{i'tj}$:

$$\begin{aligned}
Cov(U_i, Y_{i'tj}) &= E(Cov(U_i, Y_{i'tj}|(U,V,W))) + Cov(E(U_i|(U,V,W)), E(Y_{i'tj}|(U,V,W))) \\
&= 0 + Cov(U_i, W_{i'tj}\mu_{i'tj}) \\
&= \mu_{i'tj}\{E(Cov(U_i, W_{i'tj}|(U,V))) + Cov(E(U_i|(U,V)), E(W_{i'tj}|(U,V)))\} \\
&= \mu_{i'tj}Cov(U_i, V_{i't}) \\
&= \mu_{i'tj}\{E(Cov(U_i, V_{i't}|U)) + Cov(E(U_i|U), E(V_{i't}|U))\} \\
&= \mu_{i'tj}Cov(U_i, U_{i'})
\end{aligned}$$

a. If i=i',then:

$$Cov(U_i, Y_{i'tj}) = \mu_{itj}\sigma^2$$

b. If i≠i',then:

$$Cov(U_i, Y_{i'tj}) = 0$$

The calculation of the unconditional covariance of the responses V_{it} and $Y_{i't'j}$:

$$\begin{aligned}
Cov(V_{it}, Y_{i't'j}) &= E(Cov(V_{it}, Y_{i't'j}|(U,V,W))) + Cov(E(V_{it}|(U,V,W)), E(Y_{i't'j}|(U,V,W))) \\
&= 0 + Cov(V_{it}, W_{i't'j}\mu_{i't'j}) \\
&= \mu_{i't'j}\{E(Cov(V_{it}, W_{i't'j}|(U,V))) + Cov(E(V_{it}|(U,V)), E(W_{i't'j}|(U,V)))\} \\
&= \mu_{i't'j}Cov(V_{it}, V_{i't'}) \\
&= \mu_{i't'j}\{E(Cov(V_{it}, V_{i't'}|U)) + Cov(E(V_{it}|U), E(V_{i't'}|U))\}
\end{aligned}$$

a. If i=i',then:

$$Cov(V_{it}, Y_{i't'j}) = \mu_{it'j}(\tau^2\rho^{|t-t'|} + \sigma^2)$$

b. If i≠i',then:

$$Cov(V_{it}, Y_{i't'j}) = 0$$

The calculation of the unconditional covariance of the responses W_{itj} and $Y_{i't'j'}$:

$$\begin{aligned}
Cov(W_{itj}, Y_{i't'j'}) &= E(Cov(W_{itj}, Y_{i't'j'}|(U,V,W))) + Cov(E(W_{itj}|(U,V,W)), E(Y_{i't'j'}|(U,V,W))) \\
&= 0 + Cov(W_{itj}, W_{i't'j'}\mu_{i't'j'}) \\
&= \mu_{i't'j'}\{E(Cov(W_{itj}, W_{i't'j'}|(U,V))) + Cov(E(W_{itj}|(U,V)), E(W_{i't'j'}|(U,V)))\} \\
&= \mu_{i't'j'}\{E(Cov(W_{itj}, W_{i't'j'}|(U,V))) + Cov(V_{it}, V_{i't'})\}
\end{aligned}$$

a. If i=i', t=t' and j=j', then

$$\begin{aligned} Cov(W_{itj}, Y_{i't'j'}) &= \mu_{itj}\{w_j^2 + Cov(V_{it}, V_{it})\} \\ &= \mu_{itj}\{w_j^2 + E(Var(V_{it}|U)) + Var(E(V_{it}|U))\} \\ &= \mu_{itj}(w_j^2 + \tau^2 + \sigma^2) \end{aligned}$$

b. If i=i', t=t' and j≠j', then:

$$\begin{aligned} Cov(W_{itj}, Y_{i't'j'}) &= \mu_{itj'}\{0 + Cov(V_{it}, V_{it})\} \\ &= \mu_{itj'}\{0 + E(Var(V_{it}|U)) + Var(E(V_{it}|U))\} \\ &= \mu_{itj'}(\tau^2 + \sigma^2) \end{aligned}$$

c. If i=i'and t≠t', then:

$$\begin{aligned} Cov(W_{itj}, Y_{i't'j'}) &= \mu_{it'j'}\{0 + Cov(V_{it}, V_{it'})\} \\ &= \mu_{it'j'}\{0 + E(Cov(V_{it}, V_{it'}|U)) + Cov(E(V_{it}|U), E(V_{it'}|U))\} \\ &= \mu_{it'j'}(\tau^2\rho^{|t-t'|} + \sigma^2) \end{aligned}$$

d. If i≠i',then:

$$Cov(W_{itj}, Y_{i't'j'}) = 0$$

3.3 The best linear unbiased predictors(BLUP) of random effects

In almost every practical situation, the random effects are unknown; however, they can be predicted from the data. This section discusses the orthodox Best Linear Un-

biased Predictor (BLUP) of the random effects U_i, V_{it}, and W_{itj}. This is a predictor of the random effects which is linear in Y, unbiased, and best in the sense that it minimizes the mean square distance between the true random effects and any other linear predictor[17].

Let U_{i**} represent $\underbrace{(U_i, ..., U_i)^T}_{Length\ JT\ vector}$, and U_{***} represent $\underbrace{(U_{1**}^T, ..., U_{m**}^T)^T}_{Length\ mJT\ vector}$.Similarly, let V_{i**} represent $\underbrace{(V_{i1}, ..., V_{iT}, V_{i1}, ..., V_{iT}, ..., V_{i1}, ..., V_{iT})^T}_{Length\ JT\ vector}$, and V_{***} represent $\underbrace{(V_{1**}^T, ..., V_{m**}^T)^T}_{Length\ mJT\ vector}$. Let W_{i**} represent $\underbrace{(W_{i11}, ..., W_{iT1}, W_{i12}, ..., W_{iT2}, ..., W_{i1J}, ..., W_{iTJ})^T}_{Length\ JT\ vector}$, and W_{***} represent $\underbrace{(W_{1**}^T, ..., W_{m**}^T)^T}_{Length\ mJT\ vector}$. Then, the best linear unbiased predictors (BLUP) of random effect can be calculated as:

$$\hat{U}_{***} = E(U_{***}) + Cov(U_{***}, Y)Cov^{-1}(Y)\{Y - E(Y)\}$$
$$\hat{V}_{***} = E(V_{***}) + Cov(V_{***}, Y)Cov^{-1}(Y)\{Y - E(Y)\}$$
$$\hat{W}_{***} = E(W_{***}) + Cov(W_{***}, Y)Cov^{-1}(Y)\{Y - E(Y)\}$$

By doing so, all the random effects can be calculated at once. However, if the data set is big, the calculation efficiency is very low for R code. Furthermore, if the data set is too big, for example 30000 observations, R could overflow when calculating the covariance matrices.

According to the variance and covariance calculation in section 3.2 and after some matrix operations, the best linear unbiased predictors (BLUP) of random effect can be calculated block by block:

$$\hat{U}_{i**} = E(U_{i**}) + Cov(U_{i**}, Y_i)Cov^{-1}(Y_i)\{Y_i - E(Y_i)\} \tag{3.2}$$
$$\hat{V}_{i**} = E(V_{i**}) + Cov(V_{i**}, Y_i)Cov^{-1}(Y_i)\{Y_i - E(Y_i)\} \tag{3.3}$$
$$\hat{W}_{i**} = E(W_{i**}) + Cov(W_{i**}, Y_i)Cov^{-1}(Y_i)\{Y_i - E(Y_i)\} \tag{3.4}$$

All the elements in the above formulas can be explained by w_j^2, σ^2 and τ^2 via the results from section 3.2.2.

3.4 Estimation of regression parameters

This section will go through the details of parameter estimation for the Poisson mixed model. In order to estimate the regression parameters, we first assume known the random effects parameters. And then in the next section, the estimation of the random effects parameters are introduced.

Let X_i represent the covariates matrix for subject i, and p is the number of the covariates; then:

$$X_i = \begin{bmatrix} x_{i10} & x_{i11} & \cdots & x_{i1p} \\ x_{i20} & x_{i21} & \cdots & x_{i2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{iT0} & x_{iT1} & \cdots & x_{iTp} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{i10} & x_{i11} & \cdots & x_{i1p} \\ x_{i20} & x_{i21} & \cdots & x_{i2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{iT0} & x_{iT1} & \cdots & x_{iTp} \end{bmatrix} \tag{3.5}$$

There are J repeating blocks in (3.5).

And then, let $X=(X_1^T, X_2^T, ..., X_m^T)^T$.

One theorem (theorem 5.1) from [17] and the Newton scoring algorithm from [21] are used to estimate the regression parameters.

Theorem 3.4.1. *For Tweedie mixed models, the estimated score function, sensitivity*

and variability matrices can be expressed as follows:

$$1.\psi(\beta) = X^T diag(E(Y))Cov^{-1}(Y)(Y - E(Y)),$$
$$2.V(\beta) = Var(\psi(\beta)) = X^T diag(E(Y))Cov^{-1}(Y)diag(E(Y))X,$$
$$3.J(\beta) = -S(\beta) = V(\beta)$$

The first statement can be treated as a multivariate version of the quasi-score function, first proposed by [22], which gives a global matrix expression for estimated score function. The second and the third statements give the expression for variability matrix and sensitivity matrix respectively. The Poisson distribution used in this thesis belongs to Tweedie mixed models[23]. So all the statements can be applied to this thesis' model.

And the asymptotic covariance matrix[17] of $\hat{\beta}$ can be achieved by the inverse of the variability matrix.

Then the estimates of the regression parameters can be obtained using the Newton scoring algorithm:

$$\beta^* = \beta - S^{-1}(\beta)\psi(\beta)$$

where β^* is the next iteration of β.

According to the structure of Cov(Y), after some matrix operations, the calculation of $\psi(\beta)$ and S(β) can be simplified as:

$$1.\psi(\beta) = \sum_{i=1}^{m} X_i^T diag(E(Y_i))Var^{-1}(Y_i)[Y_i - E(Y_i)]$$
$$2.S(\beta) = -\sum_{i=1}^{m} X_i^T diag(E(Y_i))Var^{-1}(Y_i)diag(E(Y_i))X_i$$

3.5 Estimation of random effects parameters

In this section, moment estimators of random effects parameters are used, supposing that the correlation parameter ρ is known.

3.5.1 Subject-specific random effects parameter

Following the concept raised by [17], the iterative equation for estimating σ^2 can be expressed as:

$$\begin{aligned}\hat{\sigma}_r^2 &= \frac{1}{m}\sum_{i=1}^{m}\{(\hat{U}_i - 1)^2 + \hat{\sigma}_{r-1}^2 - Var(\hat{U}_i - 1)\} \\ &= \frac{1}{m}\sum_{i=1}^{m}\{(\hat{U}_i - 1)^2 + \hat{\sigma}_{r-1}^2 - Var(\hat{U}_i)\}\end{aligned}$$

$\hat{\sigma}_{r-1}^2$ is the estimate from the previous iteration. According to (3.2), the variance of $\hat{U}_i$ is :

$$Var(\hat{U}_i) = Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov(Y_i, U_i)$$

In this equation, Cov(U_i,Y_i) is any row vector of Cov(U_{i**},Y_i) and Cov(Y_i,U_i) is the transformation of Cov(U_i,Y_i).

Then the the iterative equation for estimating subject specific random effects parameter σ^2 can be expressed as:

$$\hat{\sigma}_r^2 = \frac{1}{m}\sum_{i=1}^{m}\{(\hat{U}_i - 1)^2 + \hat{\sigma}_{r-1}^2 - Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov((Y_i, U_i)\}$$

3.5.2 Time-specific random effects parameter

Similarly, the iterative equation for estimating τ^2 can be expressed as:

$$\hat{\tau}_r^2 = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}\{(\hat{V}_{it} - \hat{U}_i)^2 + \hat{\tau}_{r-1}^2 - Var(\hat{V}_{it} - \hat{U}_i)\}$$

where $\hat{\tau}_{r-1}^2$ is the estimate from the previous iteration. And:

$$\begin{aligned} Var(\hat{V}_{it} - \hat{U}_i) &= Var(\hat{V}_{it}) + Var(\hat{U}_i) - 2Cov(\hat{V}_{it}, \hat{U}_i) \\ &= Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov((Y_i, U_i) \\ &+ Cov(V_{it}, Y_i)Cov^{-1}(Y_i)Cov(Y_i, V_{it}) \\ &- 2Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov(Y_i, V_{it}) \end{aligned}$$

In this equation, Cov(V_{it},Y_i) and Cov(Y_i,V_{it}) are the accordance row and transformation vectors of matrix Cov(V_{i**},Y_i), respectively. And then:

$$\begin{aligned} \hat{\tau}_r^2 = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}\{&(\hat{U}_{it} - \hat{U}_i)^2 + \hat{\tau}_{r-1}^2 \\ &- Cov(U_{it}, Y_i)Cov^{-1}(Y_i)Cov(Y_i, U_{it}) \\ &- Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov(Y_i, U_i) \\ &+ 2Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov(Y_i, U_{it})\} \end{aligned}$$

3.5.3 Category-specific random effects parameter

Using the same concept as in the previous 2 subsections, the iterative equation for estimating ω_j^2 can be expressed as:

$$\hat{\omega}_{jr}^2 = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}\{(\hat{W}_{itj} - \hat{V}_{it})^2 + \hat{\omega}_{j(r-1)}^2 - Var(\hat{W}_{itj} - \hat{V}_{it})\}$$

where $\hat{\omega}^2_{j(r-1)}$ is the estimate from the previous iteration. And:

$$\begin{aligned} Var(\hat{W}_{itj} - \hat{V}_{it}) &= Var(\hat{W}_{itj}) + Var(\hat{V}_{it}) - 2Cov(\hat{W}_{itj}, \hat{V}_{it}) \\ &= Cov(V_{it}, Y_i)Cov^{-1}(Y_i)Cov((Y_i, V_{it}) \\ &+ Cov(W_{itj}, Y_i)Cov^{-1}(Y_i)Cov(Y_i, W_{itj}) \\ &- 2Cov(V_{it}, Y_i)Cov^{-1}(Y_i)Cov(Y_i, W_{itj}) \end{aligned}$$

In this equation, Cov(W_{itj},Y_i) and Cov(Y_i,W_{itj}) are the accordance row and its transformation vectors of matrix Cov(W_{i**},Y_i), respectively. And then:

$$\begin{aligned} \hat{\omega}^2_{jr} = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}\{&(\hat{W}_{itj} - \hat{V}_{it})^2 + \hat{\omega}^2_{j(r-1)} \\ &- Cov(V_{it}, Y_i)Cov^{-1}(Y_i)Cov(Y_i, V_{it}) \\ &- Cov(W_{itj}, Y_i)Cov^{-1}(Y_i)Cov(Y_i, W_{itj}) \\ &+ 2Cov(V_{it}, Y_i)Cov^{-1}(Y_i)Cov(Y_i, W_{itj})\} \end{aligned}$$

3.6 Estimation of time-specific random effects correlation parameter for AR(1) correlation structure

Many correlation structures can be used in the model[19], and this thesis use AR(1), which is known as the autoregressive structure of order 1 and has only one parameter ρ. In this section, the correlation parameter ρ is estimated by a moment approach:

$$\hat{\rho}_r = \frac{\sum_{i=1}^{m}\sum_{t=1}^{T-1}\{(\hat{V}_{it} - \hat{U}_i)(\hat{V}_{i(t+1)} - \hat{U}_i) + b_i(t, t+1)\}}{\{\sum_{i=1}^{m}\sum_{t=1}^{T-1}\{(\hat{V}_{it} - \hat{U}_i)^2 + b_i(t,t)\}\}^{1/2}\{\sum_{i=1}^{m}\sum_{t=1}^{T-1}\{(\hat{V}_{i(t+1)} - \hat{U}_i)^2 + b_i(t+1, t+1)\}\}^{1/2}}$$

where

$$
\begin{aligned}
b_i(t,t) &= \tau^2 - Var(\hat{V}_{it}) - Var(\hat{U}_i) + 2Cov(\hat{V}_{it}, \hat{U}_i) \\
&= \tau^2 - Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov((Y_i, U_i) \\
&- Cov(V_{it}, Y_i)Cov^{-1}(Y_i)Cov(Y_i, V_{it}) \\
&+ 2Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov(Y_i, V_{it})
\end{aligned}
$$

and:

$$
\begin{aligned}
b_i(t,t+1) &= \tau^2\hat{\rho}_{r-1} - Cov(\hat{V}_{it} - \hat{U}_i, \hat{V}_{i(t+1)} - \hat{U}_i) \\
&= \tau^2\hat{\rho}_{r-1} - Cov(\hat{V}_{it}, \hat{V}_{i(t+1)}) - Cov(\hat{U}_i, \hat{U}_i) + Cov(\hat{V}_{it}, \hat{U}_i) + Cov(\hat{V}_{i(t+1)}, \hat{U}_i) \\
&= \tau^2\hat{\rho}_{r-1} - Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov((Y_i, U_i) \\
&- Cov(V_{it}, Y_i)Cov^{-1}(Y_i)Cov(Y_i, V_{i(t+1)}) \\
&+ Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov(Y_i, V_{it}) \\
&+ Cov(U_i, Y_i)Cov^{-1}(Y_i)Cov(Y_i, V_{i(t+1)})
\end{aligned}
$$

where Cov($V_{i(t+1)}$,Y_i) and Cov(Y_i,$V_{i(t+1)}$) are the t+1 row vector and its transformation vector of matrix Cov(V_{i**},Y_i) respectively.

3.7 Iteration algorithm

In this section, all previous sections in Chapter 3 are linked in one iteration algorithm which will converge to the asymptotic mean of parameters of interest.[17].

3.7.1 Initialization and iteration

In order to improve the computation efficiency, a reasonable initial value is necessary. We use the regression parameters from the generalized linear model as the initial

value. Furthermore, if the data set is big, it is impossible to guarantee no singular matrix will be generated in thousands of iterations for thousands of $\mathrm{Cov}(\mathrm{Y}_i)$ matrices. So, benefiting from the convergence property of the algorithm, all matrix inverse functions are replaced by generalized inverse functions[24].

Step 1. The initial regression variables estimates $\hat{\beta}^{(0)}$ are obtained from the standard generalized linear models while treating all responses independently. And then the expectation of Y_{itj} is calculated as follows:

$$E(Y_{itj}) = \mu_{itj} = exp(x_{it}^{'}\beta_j)$$

Step 2. The initial random effects predictions are:

$$\hat{U_i}^{(0)} = \frac{\frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}\{Y_{itj}\}}{\frac{1}{mTJ}\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{j=1}^{J}\{Y_{itj}\}}$$

$$\hat{V_{it}}^{(0)} = \frac{\frac{1}{J}\sum_{j=1}^{J}\{Y_{itj}\}}{\frac{1}{mTJ}\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{j=1}^{J}\{Y_{itj}\}}$$

$$\hat{W_{itj}}^{(0)} = \frac{Y_{itj}}{\frac{1}{mTJ}\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{j=1}^{J}\{Y_{itj}\}}$$

Step 3. These predictors of the random effects are used for estimating the initial random effects parameters:

$$\hat{\sigma}^{2(0)} = \frac{1}{m}\sum_{i=1}^{m}\{\hat{U}_i^{(0)} - 1\}^2$$

$$\hat{\tau}^{2(0)} = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}\{\hat{V}_{it}^{(0)} - \hat{U}_i^{(0)}\}^2$$

$$\hat{\omega}_j^{2(0)} = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}\{\hat{W}_{itj}^{(0)} - \hat{V}_{it}^{(0)}\}^2$$

And then the initial correlation parameter $\rho^{(0)}$ is set to be 0.5.

Step 4. Use the results from step 1-3 and sections 3.2 and 3.3 to calculate the covariance matrices.

Step 5. Use the results from step 1-4 and sections 3.4 - 3.6 to update all the interested parameters.

Step 6. Use the updated regression parameters to update the expectation of Y_{itj}:

$$E(Y_{itj}) = \mu_{itj} = \exp(x_{it}^{'}\beta_j)$$

Then use the results from steps 5-6 rather than steps 1-3 to start the iteration from step 4.

3.8 Baseline-category logit model

In the previous sections of this chapter, the mixed Possion model is introduced including model assuptions, parameter estimation, and random effects prediction. In this section, baseline-catagory logit models are constructed by organizing a series of

mixed Poisson models while setting the regression part in certain way.

For those responses which are introduced in section 3.1 and have J nominal categories, according to section 2.2.1, the baseline-category logit models consists of J-1 baseline logits, each of which, following the instruction of section 2.3, is constructed from one of the J-1 Poisson mixed models and the Poisson mixed model which describes the baseline.

Thus, if setting the J^{th} category as the baseline categoty, the baseline-category model is equivalent to the multiple Poisson mixed models shown below. And all the assumptions and estimation methods introduced in previous sections apply here.

$$\begin{cases} Y_{it1}|(U,V,W) \ \sim \ \text{Poisson}\{W_{it1} \ exp[x_{it}^T(\beta_1+\alpha)]\}, \\ ... \\ Y_{it(J-1)}|(U,V,W) \ \sim \ \text{Poisson}\{W_{it(J-1)} \ \exp[x_{it}^T(\beta_{J-1}+\alpha)]\}, \\ Y_{itJ}|(U,V,W) \ \sim \ \text{Poisson}\{W_{itJ} \ \exp(x_{it}^T\alpha)\}, \end{cases}$$

And those can be written in equivalent J-1 baseline logits as follows, which give the log odds that the response is j, given that the response falls in category j or category J.

$$\log\frac{\pi_{itj}}{\pi_{itJ}} = \log\frac{W_{itj}}{W_{itJ}} + x_{it}^T\beta_j, \ j = 1, ..., J-1$$

where

$$\pi_{itj} = \frac{W_{itj} \ \exp[x_{it}^T(\beta_j+\alpha)]}{\sum_1^{J-1} W_{itj} \ \exp[x_{it}^T(\beta_j+\alpha)] + W_{itJ}\exp(x_{it}^T\alpha)} \tag{3.6}$$

Using the estimation methods introduced before, all the regression parameters and the random effects can be estimated. And the parameter β_j represents the effect of x on the log odds that the response is j, given that the response falls in category j or category J.

3.9 Continuation-ratio logit model

For those responses which are introduced in section 3.1 and have J ordinal categories, according to section 2.2.2, the continuation-ratio model consists of J-1 continuation-ratio logits, each of which, following the concept of section 2.3, is constructed from the dual-Poisson mixed models as follows. Which is different from the baseline-category logit models which fit all these Poisson model simultaneously, here, the dual-Poisson mixed models for each logit are fitted one by one.

$$\begin{cases} Y_{itj}|(U^{(j)}, V^{(j)}, W^{(j)}) \sim \text{Poisson}\{W_{it1}^{(j)} \exp[\alpha_j + x_{it}^T\beta_j + \delta_j + x_{it}^T\gamma_j]\}, \\ Y_{it(j+1)} + ... + Y_{itJ}|(U^{(j)}, V^{(j)}, W^{(j)}) \sim \text{Poisson}\{W_{it2}^{(j)} \exp[\delta_j + x_{it}^T\gamma_j]\}. \end{cases}$$

where j=1, ..., J-1.

All of these J-1 pairs of dual-Possion mixed models construct the equivalent continuation-ratio model as follows:

$$\log\left[\frac{\pi_j}{\pi_{j+1} + ... + \pi_J}\right] = \log\left[\frac{W_{it1}^{(j)} \exp[\alpha_j + x_{it}^T\beta_j + \delta_j + x_{it}^T\gamma_j]}{W_{it2}^{(j)} \exp[\delta_j + x_{it}^T\gamma_j]}\right] = \log\left[\frac{W_{it1}^{(j)}}{W_{it2}^{(j)}}\right] + \alpha_j + x_{it}^T\beta_j$$

where j=1,...,J-1.

Since the J-1 dual-Poisson models are estimated separately, the upper index (j) of the random effects shows that every dual-Poisson model has its own random effects. The parameter β_j represents the effect of x on the log odds of a response falling in category j, given that response falls in j,...J categories.

3.10 Model expansion

The model introduced in Section 3.1 has 3 levels of random effects (U,V,W). By reducing 1 level of random effect[18] and/or eliminating the random effect correlation ρ[12], it can be easily expanded to other models which can fit different research

interests. And all the calculation method in this Chapter can be reapplied to these models. Below 2 simplifications are listed as examples. And in Chapter 4, the data set is also fitted to these 2 models.

3.10.1 Model 1

This model is one specific case of Tweedie model introduced by [17]. If the main purpose of the research is subject-specific and time-specific effects and the data set shows unobvious category effects, then the model can be simplified to the setup as below.

Assumption 1: Subject-specific random effects U_1,..., U_m are positive, independent and identically distributed with mean 1 and variance σ^2:

$$E(U_i) = 1 \; and \; Var(U_i) = \sigma^2.$$

Assumption 2: Given the subject-specific random effects U, the moment structure of the positive time-specific random effects V can be expressed as:

$$E(V_{it}|U) = U_i \; and \; Cov(V_{it}, V_{i't'}|U) = \begin{cases} \tau^2 \rho^{|t-t'|}, & \text{if } i = i'; \\ 0, & \text{if } i \neq i'; \end{cases}$$

Here, we assume the correlation structure is AR(1).

Assumption 3: Given the random effects (U,V), the components of Y are conditionally independent, and the conditional distribution of Y_{itj}, given (U,V), depends on V_{it} only, which is:

$$Y_{itj}|(U,V) \; \sim \; \text{Poisson}(V_{it}\mu_{itj})$$

where

$$\mu_{itj} = \exp(x_{it}^T \beta_j)$$

with vector of covariates x_{it} and regression parameter β_j

3.10.2 Model 2

If the main purpose of the research is the time-specific effects and the category-specific effects, then the model can be simplified to the setup as below.

Assumption 1: The moment structure of the positive time-specific random effects V can be expressed as:

$$E(V_{it}) = 1 \; and \; Cov(V_{it}, V_{i't'}) = \begin{cases} \tau^2 \rho^{|t-t'|}, & \text{if } i = i'; \\ 0, & \text{if } i \neq i'; \end{cases}$$

Here, we assume the correlation structure is AR(1).

Assumption 2: Given the subject and time specific random effects V, the moment structure of the positive category-specific random effects W can be expressed as:

$$E[W_{itj}|V] = V_{it} \; and \; Var[W_{itj}|V] = w_j^2 V_{it}$$

Assumption 3: Given the random effects (V,W), the components of Y are conditionally independent, and the conditional distribution of Y_{itj}, given (V,W), depends on W_{itj} only, which is:

$$Y_{itj}|(V,W) \; \sim \; \text{Poisson}(W_{itj}\mu_{itj})$$

where

$$\mu_{itj} = \exp(x_{it}^{'} \beta_j)$$

with vector of covariates x_{it} and regression parameter vectorβ_j.

Chapter 4

Application

The data set used to illustrate how the model works is the crash counts over one thousand major signalized intersections in the city of Chicago from 2004 to 2010[25]. For each intersection, vehicle crashes were counted by crash severity levels, including fatal(FAT); injury types A(IA), B(IB), and C(IC) for major, moderate, and minor injury levels; property damage only (PDO); and unknown. The intersections are in dscending order of the total crashes in these sever years. In the later analysis, the "unknown" category is not included. And to fit with the data set, we will use intersection-specific random effects to substitute the subject-specific random effects.

4.1 Data exploration

Table 4.1 shows a few beginning rows of the data set. Columns 1 and 2 are the location of the intersection. The third column is the index of the intersection. Column 4 (YEAR) is the year of taking the observation. The remaining columns are the counts of each category.

There are 1001 intersections observed from 2004 to 2010 with one observation each year for each category. So the data set shows both longitudinal and multinomial properties. And the data set is big, containing 35035 observations if only FAT, IA,

IB, IC and PDO categories are counted. For the entire data set, there is no missing value and there are seven same observation time points for each intersection.

As the model proposed in this thesis is Poisson, to make a initial data exploration,

Table 4.1: The layout of the raw data

	LONG	LAT	INT	YEAR	TOTAL	FAT	IA	IB	IC	PDO	UNKWN
1	-87.59	41.75	1	2004	106	0	0	5	3	55	43
2	-87.59	41.75	1	2005	85	0	1	9	2	39	34
3	-87.59	41.75	1	2006	96	0	1	14	5	44	32
4	-87.59	41.75	1	2007	65	0	0	8	5	30	22
5	-87.59	41.75	1	2008	70	0	0	10	5	40	15
6	-87.59	41.75	1	2009	69	0	0	9	4	53	3
7	-87.59	41.75	1	2010	68	0	0	5	3	57	3
8	-87.71	41.94	2	2004	77	0	0	7	5	31	34
9	-87.71	41.94	2	2005	68	0	1	9	4	26	28
10	-87.71	41.94	2	2006	58	0	1	3	2	16	36
11	-87.71	41.94	2	2007	53	0	0	4	1	20	28
12	-87.71	41.94	2	2008	58	0	1	5	4	32	16
13	-87.71	41.94	2	2009	65	0	0	7	2	47	9
14	-87.71	41.94	2	2010	62	0	1	3	8	46	4
15	-87.59	41.72	3	2004	55	0	1	4	3	26	21
16	-87.59	41.72	3	2005	44	0	0	1	2	21	20
17	-87.59	41.72	3	2006	58	0	0	8	1	19	30
18	-87.59	41.72	3	2007	40	0	1	0	2	13	24
19	-87.59	41.72	3	2008	47	0	0	8	1	29	9
20	-87.59	41.72	3	2009	45	0	1	3	4	31	6
21	-87.59	41.72	3	2010	49	0	0	3	3	41	2
22	-87.75	41.97	4	2004	62	0	2	5	2	26	27
23	-87.75	41.97	4	2005	54	0	0	3	3	21	27

the data set is pretreated by taking the logarithm of each category's count+0.5. Adding 0.5 is to prevent infinity if the count is 0. (Mathematically, adding 0.5 to count will not impact correlation direction, trend of each category versus "YEAR" and the LOWESS smoothing surface shape). In this section, all analyses are based on the pretreated data. To ease reading, in the rest of this section, PDO represents log (PDO+0.5) of the original data set. And IA, IB, IC, FAT and TOTAL follow the same logic.

To get the overall sense of the data set, a correlation matrix is calculated over the

entire data set. Although the correlation tells linear relation between variables, its direction (positive or negative) and scale could give certain kinds of basic understanding between variables.

Table 4.2: The correlation matrix of the raw data

	YEAR	TOTAL	FAT	IA	IB	IC	PDO	UNKWN
YEAR	1.0000	-0.1303	-0.0442	-0.0723	-0.0721	-0.0599	0.3118	-0.5579
TOTAL	-0.1303	1.0000	0.0259	0.2465	0.4731	0.4314	0.7735	0.6664
FAT	-0.0442	0.0259	1.0000	0.0174	0.0039	0.0247	-0.0115	0.0365
IA	-0.0723	0.2465	0.0174	1.0000	0.1449	0.1167	0.1194	0.1265
IB	-0.0721	0.4731	0.0039	0.1449	1.0000	0.2236	0.2781	0.1825
IC	-0.0599	0.4314	0.0247	0.1167	0.2236	1.0000	0.2682	0.1730
PDO	0.3118	0.7735	-0.0115	0.1194	0.2781	0.2682	1.0000	0.0972
UNKWN	-0.5579	0.6664	0.0365	0.1265	0.1825	0.1730	0.0972	1.0000

Some highlights of table 4.2 are listed here:

(1). YEAR is negatively correlated with all the categories except POD. This can be explained in this way: that as time moves on, traffic is busier and then car speed is slower, which causes an increased number of small collisions. And there might be some intervention taken to smooth the traffic in the passing years, which reduced the number of severe accidents like FAT, IA, IB, and IC

(2). TOTAL is positively correlated with all the other categories. At first glance, following common sense, this should be the obvious result: high risk intersections generate more of all kinds of accidents. However, according to the nature of traffic, if there are lots of accidents happening in one intersection which means the traffic is busy there, then the vehicles run slow here which may reduce the number of severe accidents while dramatically increasing low severity accidents like POD. This sce-

nario aligns with the correlation matirx : TOTAL is highly positively correlated with PDO and the correlation is reducing along with the accident severity increasing.
(3). FAT has no obvious correlation with all the other items. Following the concept of (1) and (2), this result can be explained for "busy" and "leisure" intersections separately. In busy intersection, because of the intensive traffic, the possibility of getting people hurt is high, which could increase the FAT rate. However, intensive traffic normally leads to low vehicle speed which, on the other hand, lowers the FAT rate. Vice versa, for "leisure" intersections, low traffic could lower the FAT rate. However, low traffic may lead to high vehicle speed, which may increase the FAT rate. Because of the contradictions, the correlations between FAT and other variables are very small.
The correlation matrix gives the brief relation between variables. To illustrate the longitudinal property of the data set, 100 intersections are randomly selected from these 1001 intersections, and the trends of each category on each intersection over time are illustrated in figure 4.1.

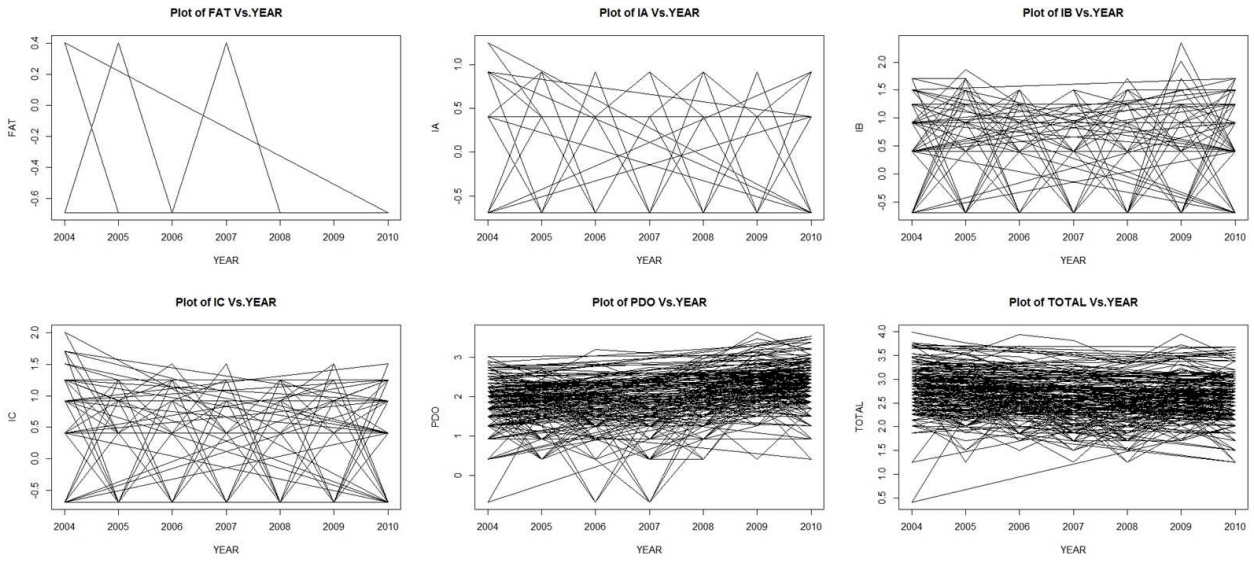

Figure 4.1: The trends of the crashes count over YEAR

Actually, different random samples were taken, and the trends exhibit no big differences. The trends shown in Figure 4.1 are in compliance with the correlation matrix. As mentioned in section 2.3 and the correlation analysis, to link the intersection location with the counts, LOWESS smoothing surface of TOTAL over latitude and longitude is used; the spatial surface is plotted in Figure 4.2.

According to the spatial surface, there is no clear trend and it is difficult to link them with equations. And going back to the original data, the variation between nearby intersections is also big. So, using the estimated value of LOWESS surface as one covariate to link intersection location with category counts is more reasonable than using latitude and longitude directly.

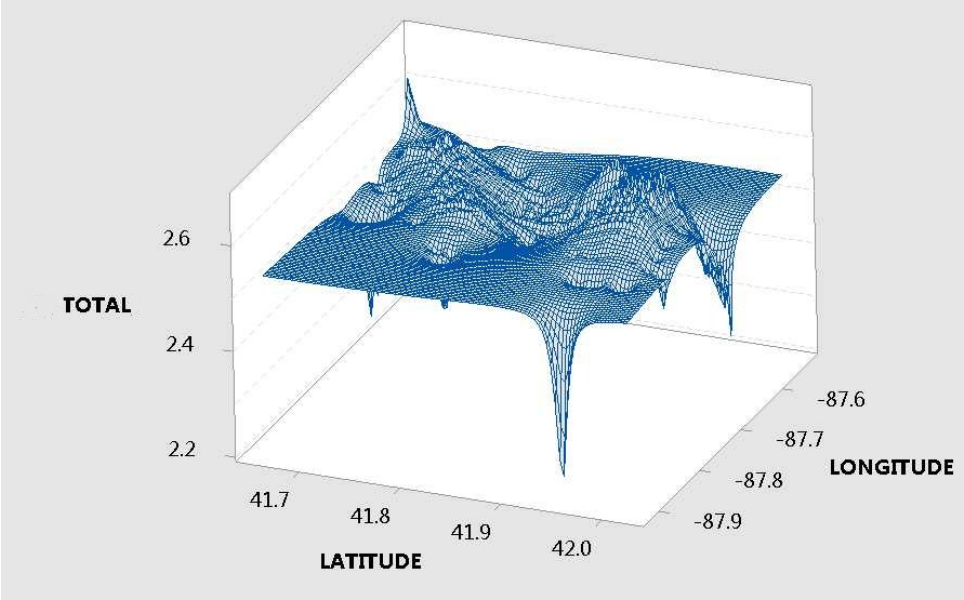

Figure 4.2: The LOWESS smoothing surface of TOTAL over latitude and longitude

4.2 Data analysis using different models

From this section on, different from section 4.1, the notations of "TOTAL", "FAT", "IA", "IB", "IC" and "PDO" represent the original value.

The categories of the data set show ordinal property(FAT, IA, IB, IC, and PDO are ordered by the crash severity), and it can be analyzed as nominal data as well[8]. And the model proposed in Chapter 3 can be applied to both nominal and ordinal data by organizing responses and covariates following section 3.8 and 3.9. To give a basic statistic understanding of this data set and highlight the model performance, this thesis detailedly investigates the nominal side by baseline-category logit model. And following this, the continuation-ratio logit model is used to leverage the ordinal property of the data set.

4.2.1 Baseline-category logit model

Learning from section 4.1, this model selects YEAR and LOWESS estimates of TOTAL over latitude and longitude as the covariates. Without doing any data re-organization and maintaining assumptions 1 to 3 unchanged, for assumption 4, we assume the regression parameters as below:

$$Y_{itj}|(U,V,W) \ \sim \ \text{Poisson}(W_{itj}\mu_{itj})$$

where

$\mu_{itj} = \exp[x_{it}^T(\beta_j + \alpha)]$ for j=1,2,3,4;

$\mu_{itj} = \exp(x_{it}^T\alpha)$ for j=5;

β_j=$(\beta_{j0}, \beta_{j1}, \beta_{j2})^T$;

α=$(\alpha_0, \alpha_1, \alpha_2)^T$;

x_{it}=(1, YEAR, LOWESS(Total))^T, where LOWESS(Total) represents LOWESS estimates of TOTAL over latitude and longitude.

The indexes for the model are as follows:

intersection index i=1,...,1001;

YEAR index t=1,..7, which represent year 2004 to 2010;

category index j=1,...5, which represent FAT, IA, IB, IC, and PDO.

4.2.1.1 Model comparison

Four models are applied on the data set:

(1). Full model: the model with the intersection-specific random effects, the time-specific random effects and the category-specific random effects;

(2). Model1: the model introduced in section 3.8.1, which includes the intersection-specific random effects and the time-specific random effects ;

(3). Model2: the model introduced in section 3.8.2, which includes the time-specific random effects and the category-specific random effects;

(4). GLM: the standard generalized linear model.

The analysis results are shown in table 4.3. The regression parameter estimates under all four models are very similar with slight differences. The standard errors of the full model are all slightly bigger than those of the GLM model execpt β_{10}. For model 1 and 2, the standard errors are slightly different from those of the GLM. Furthermore, the standard errors for α_0, α_1, and α_2 of the full model, model 1 and model 2 are all much bigger than those of the GLM in their own scale. However, all of these differences do not affect the significance of the regression parameters.

In table 4.3, σ^2 is the variance of the intersection-specific random effects. τ^2 is the variance of the time-specific random effects. ω_j^2 (j=1,..., 5) is the variance of the category-specific random effects for category j. For the full model and model 1, which all introduce intersection-specific and time-specific random effects, the parameters of these two random effects are similar. And the sum of these two parameters is similar to model 2's time-specific random effects parameter which does not introduce intersection-specific random effects. For full model and model 2, which both suppose different category-specific variations, the results confirm the supposition.

Overall, all these four models give consistent results. However, according to the

Table 4.3: The results of the different models

Categories	Covariates	Full.model		Model.1		Model.2		GLM	
		EST.	SE	EST.	SE	EST.	SE	EST.	SE
FAT	Intercept β_{10}	-3.3686	0.909	-3.7737	0.91	-3.3661	0.8955	-3.3048	0.9076
	Year β_{11}	-0.4395	0.0975	-0.4412	0.0976	-0.4412	0.0962	-0.4625	0.0984
	Lowess β_{12}	-0.1677	0.0637	-0.168	0.0642	-0.1674	0.0628	-0.1663	0.0637
IA	Intercept β_{20}	-2.2186	0.1231	-2.2194	0.1211	-2.2199	0.1201	-2.1746	0.1216
	Year β_{21}	-0.1803	0.0123	-0.1798	0.0124	-0.1807	0.012	-0.191	0.0122
	Lowess β_{22}	-0.0257	0.0077	-0.0258	0.0076	-0.0255	0.0075	-0.0257	0.0076
IB	Intercept β_{30}	-0.8897	0.0628	-0.893	0.0613	-0.8882	0.0613	-0.8554	0.0597
	Year β_{31}	-0.1465	0.0062	-0.1464	0.006	-0.1468	0.0061	-0.1555	0.0059
	Lowess β_{32}	-0.0192	0.0039	-0.019	0.0039	-0.0192	0.0038	-0.019	0.0037
IC	Intercept β_{40}	-1.3565	0.0734	-1.3587	0.0728	-1.355	0.0714	-1.3211	0.0712
	Year β_{41}	-0.1449	0.0073	-0.145	0.0071	-0.1451	0.0071	-0.1539	0.0071
	Lowess β_{42}	-0.0156	0.0045	-0.0154	0.0043	-0.0156	0.0044	-0.0154	0.0044
PDO	Intercept α_0	0.4464	0.0622	0.4483	0.0631	0.4528	0.0586	0.4134	0.023
	Year α_1	0.1061	0.0032	0.1058	0.0037	0.1056	0.0035	0.1159	0.0023
	Lowess α_2	0.0745	0.0042	0.0746	0.0042	0.0758	0.0039	0.0728	0.0014
	σ^2	0.1103		0.1084					
	τ^2	0.0723		0.0888		0.1645			
	ω_1^2	1.301				1.389			
	ω_2^2	0.1174				0.2605			
	ω_3^2	0.0781				0.0934			
	ω_4^2	0.0675				0.1107			
	ω_5^2	0.0269				0.0316			
	ρ	0.5258		0.4979		0.8575			

nature of the data set, the full model, which accounts for intersection-specific, time-specific and category-specific random effects, could fit the data set the best. Thus, the rest of the analysis will focus on the full model while using the GLM as a comparison.

4.2.1.2 Model estimation performance

For this model, given (U,V,W), Y_{itj} are following Poisson distribution and independent with each other for all i, t and j . So, given (U,V,W), the sum of all Y_{itj} follows a Poisson distribution with mean $\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{j=1}^{J}\omega_{itj}\mu_{itj}$. And this derivation is confirmed by this data set with$\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{j=1}^{J}y_{itj}$=65172 and $\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{j=1}^{J}\omega_{itj}\mu_{itj}$=65256,

while $\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{j=1}^{J}\mu_{itj}$=66138.

The overall estimation performance can be examined by the correlations of the conditional mean $\omega\exp(x^T\beta)$ and the overall mean μ= exp ($\mathrm{x}^T\beta$) versus the observations. The overall correlation matrix is shown in table 4.4.

According to the correlation matrix, the full model's conditional mean are very

Table 4.4: The correlation matrix of the fitted values and observations

	Observation	Conditional mean
Conditional mean	0.955	
Overall mean	0.795	0.884

closely aligned with the observations. On the other hand, according to Table 4.3, since the regression parameters for the full model are very similar to those for the GLM, the correlation of overall mean and observation can somehow represent the estimation performance of GLM for this data set. And this number is less than the correlation of conditional mean and observation, which proves the full model's advantages.

The data can be analyzed from the angle of proportion as mentioned in section 2.2. There are 50 intersections randomly selected and the proportion of each category versus YEAR for estimation and observation has been plotted. Except for those intersections which had no accident for all categories, all of them show that the estimated proportions are in line with the observation proportions and the trends are more or less the same for every intersection. 3 intersections which had the highest total counts for these 7 years are ploted in figure 4.3. All of these results show that this model fits the data set well.

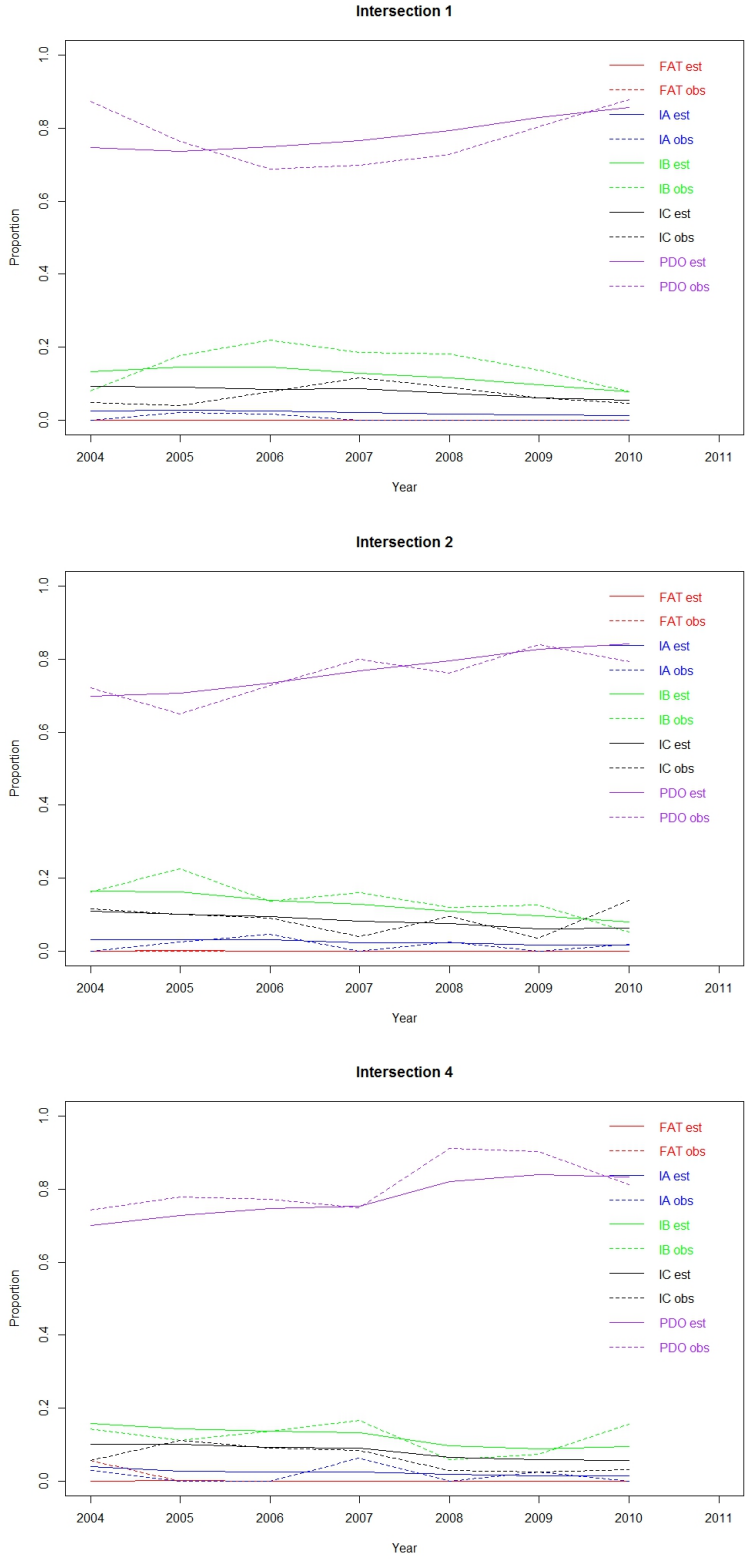

Figure 4.3: The proportion trends for model estimates and observations of each category

4.2.1.3 Interpretation of the model estimates

According to table 4.3, all the coefficients are significant based on significance level of 0.05. And then, the following can be said about this data set:

(1). The coefficients of LOWESS surface are statistically significant. The significance suggests that we should not ignore the nonlinear association between the accident counts in different type and the spatial location. This means that the intersections with different fitted values of the LOWESS surface have different accident counts.

(2). Given the fitted values of the LOWESS surface been constant, one year increase will change the number of: fatal crash to $\exp(\hat{\beta}_{11}+\hat{\alpha}_1)$ =exp(-0.44+0.10)=0.71 times of current year in average, major injury to $\exp(\hat{\beta}_{21}+\hat{\alpha}_1)$ =exp(-0.18+0.10)=0.92 times of current year, moderate injury to $\exp(\hat{\beta}_{31}+\hat{\alpha}_1)$ =exp(-0.15+0.10)=0.95 times of current year, minor injury to $\exp(\hat{\beta}_{41}+\hat{\alpha}_1)$ =exp(-0.14+0.10)=0.96 times of current year, and property damage only crash to $\exp(\hat{\alpha}_1)$ =exp(0.10)=1.1 times of current year. Over time, the number of property damage only crashs will be increased while other categories will be reduced. And, along with the injury severity increase, the speed of reducing will be increased. The reason for this might be: as time goes by, the traffic is busier which causes more small collisions. And meanwhile, because of the busier traffic, the speed of the car is slower which makes the severe injury less possible. And there might be other reasons like car safety technology and highway code updates to protect people.

(3) According to equation 3.6, the negative sign of $\hat{\beta}_{11}$, $\hat{\beta}_{21}$, $\hat{\beta}_{31}$, and $\hat{\beta}_{41}$ in table 4.3 suggest that the proportion of all categories except property damage only crash increase as time goes by.

(4) Given the accident falls in categories fatal crash and property damage only crash, for 1 year increase, the estimated odds that the accident type is fatal crash rather than property damage only crash equals $\exp(\hat{\beta}_{11})$=exp(-0.44) = 0.644 times the estimated odds at current year. In other words, given the accident falls in categories

fatal crash and property damage only crash, as time passes, the possibility of getting fatal crash is lower. Given the accident falls in categories major injury and property damage only crash, for 1 year increase, the estimated odds that the accident type is major injury rather than property damage only crash equals $\exp(\hat{\beta}_{21})$=exp(-0.18) = 0.84 times the estimated odds at current year. In other words, given the accident falls in categories major injury and property damage only crash, as time passes, the possibility of getting major injury is lower. Given the accident falls in categories moderate injury and property damage only crash, for 1 year increase, the estimated odds that the accident type is moderate injury rather than property damage only crash equals $\exp(\hat{\beta}_{31})$=exp(-0.15) = 0.86 times the estimated odds at current year. In other words, given the accident falls in categories moderate injury and property damage only crash, over time, the possibility of getting moderate injury is lower. Given the accident falls in categories minor injury and property damage only crash, for 1 year increase, the estimated odds that the accident type is minor injury rather than property damage only crash equals $\exp(\hat{\beta}_{41})$=exp(-0.14) = 0.87 times the estimated odds at current year. In other words, given the accident falls in categories minor injury and property damage only crash, over time, the possibility of getting minor injury is lower.

(5) The estimate of the variance of the intersection-specific random effects is $\hat{\sigma}^2$=0.11, which implies that there are unknow effects that will impact the number of crashes in these intersections. Figure 4.4 plots out all these 1001 intersections' standardized values of predicted intersection-specific random effects. In this figure, the X-axis is the index of the intersection, and the Y-axis is $(\hat{u}_i\text{-}1)/\hat{\sigma}$. Because the intersections are in descending order of the total number of the crashes, the predicted values of the random effects show reduction trend as well. There are two lines shown in this figure. The points over the red line are the intersections of which the predicted values are over 3 times the estimated standard deviation above the mean. These intersections

are number 1, 7, 2, 17, 14, 5, 39, in descending order of the deviation from the mean. The points over the yellow line are the intersections of which the predicted values are over 2 times the estimated standard deviation above the mean. There are 27 intersections of this kind. These may suggest that there are other effects which may dramatically impact the number of accidents. And if the local government wants to improve the circumstances, they may spend more effort in these intersections to identify the causes of these effects.

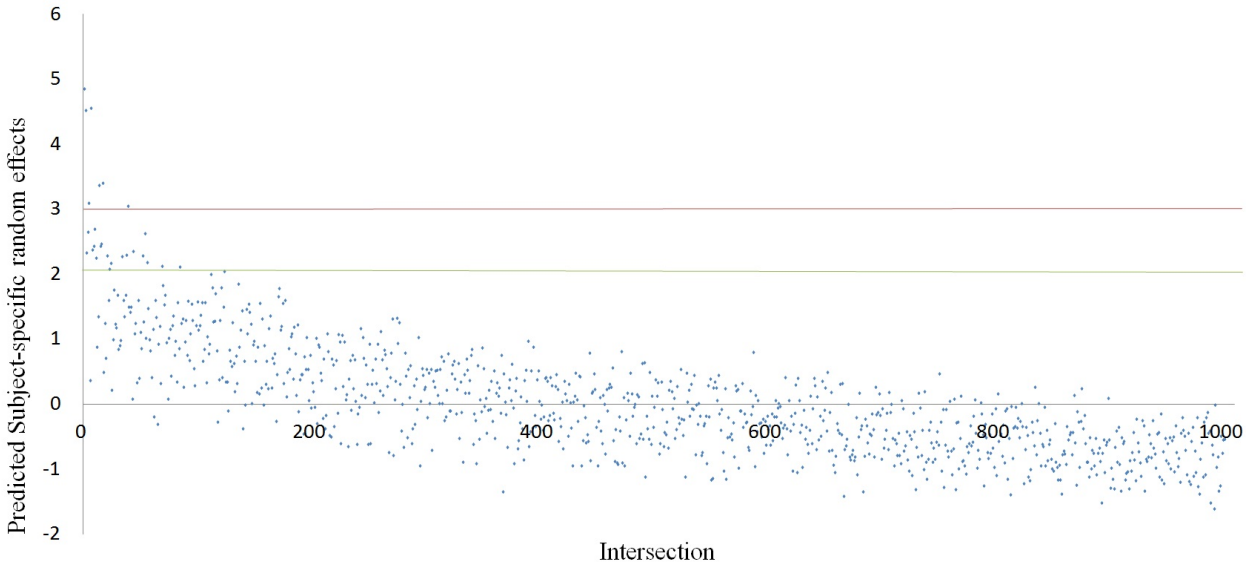

Figure 4.4: Standardized values of predicted intersection-specific random effects

In [25], the top 200 intersections with the highest number of crash occurrences identified based on crash frequency- and severity-based scenarios are shared. In this thesis, the intersections are ranked by the predicted intersection-specific random effects. This ranking gives the risk of the intersections in terms of tendency to have higher than average crash frequencies in all categories and in all years. Figure 4.5 plots all the 1001 intersections with the top 200 intersections highlighted in red. And the list of the top 200 intersections is attached to the end of the thesis.

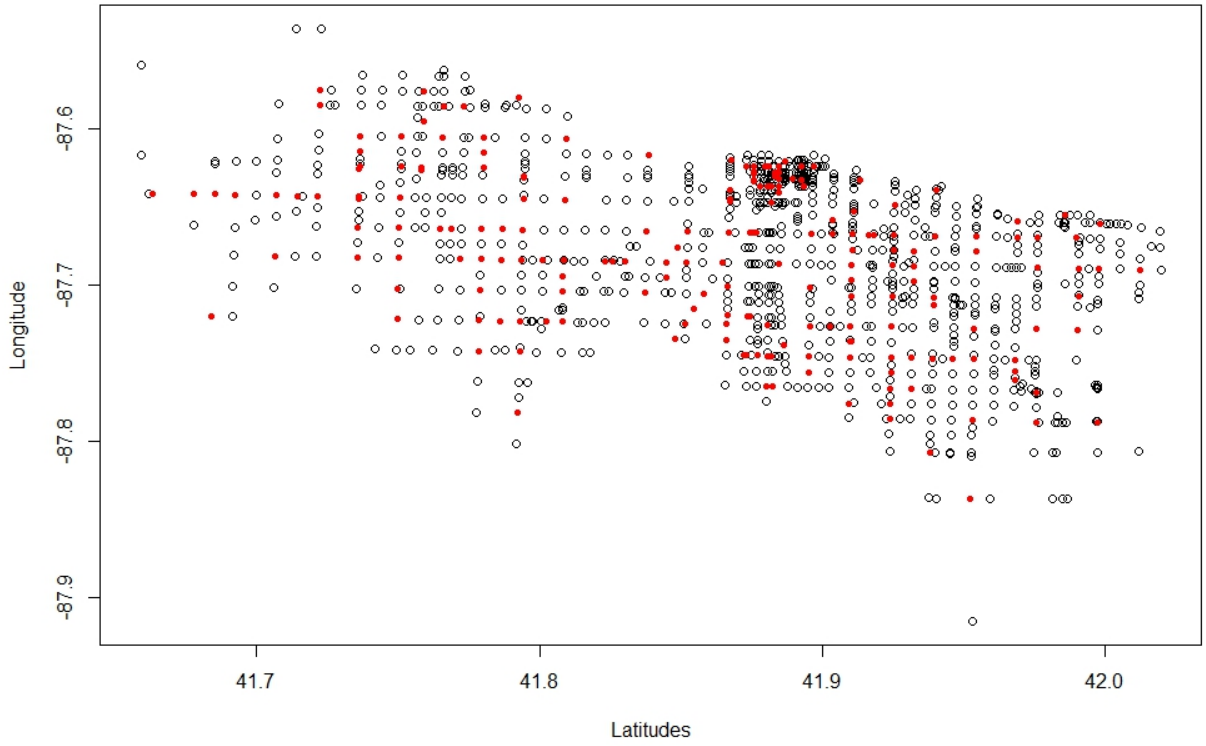

Figure 4.5: Locations of top 200 intersections with highest tendency to have higher than average crash frequencies

Following the ranking of the intersections, the time-specific random effects of 5 high risk intersections (intersection: 1, 7, 2, 17, 14), 5 moderate risk intersections (intersection: 624, 728, 366, 412, 692) and 5 low risk intersections (intersection: 813, 667, 989, 893, 992) are ploted in figure 4.6. The spread of the time-specific random effects is following the model specification: the high risk intersections have high time-specific random effects, the moderate risk intersections have moderate time-specific random effects and the low risk intersections have low time-specific random effects. And there is no obvious temporal trend showed overall.

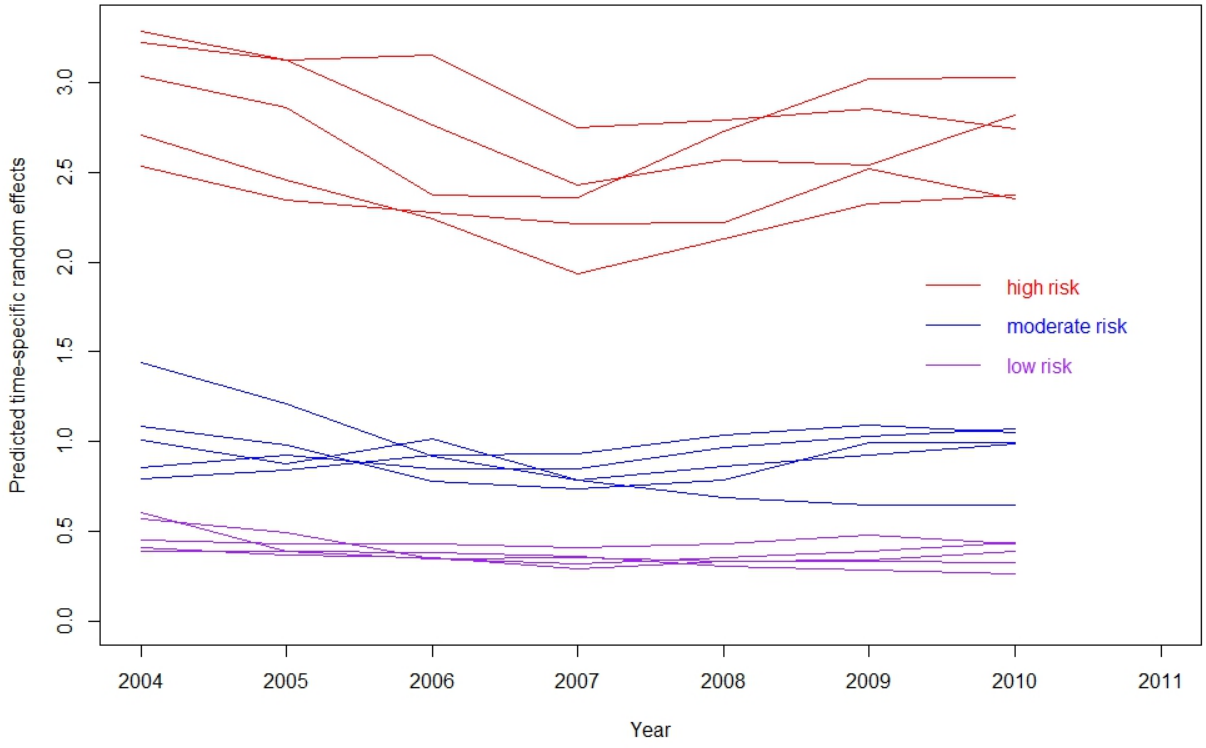

Figure 4.6: Predicted time-specific random effects of 5 high risk intersections, 5 moderate risk intersections and 5 low risk intersections

(6) The estimate of the variance of category-specific random effects for fatal injury is $\hat{\omega}_1^2$=1.30, for major injury is $\hat{\omega}_2^2$=0.12, for moderate injury is $\hat{\omega}_3^2$=0.08, for minor injury is $\hat{\omega}_4^2$=0.07, and for property damage only crash is $\hat{\omega}_5^2$=0.03. This means that beside intersection-specific and time-specific random effects there are category-specific random effects which impact the number of crashes. And all the variances are different with each other, which implies that the variances of accident counts are different from category. In all of these, the variation of category-specific random effects decreases as the severity of the category lessens . And the variance for fatal injury is dramatically bigger than the others, which means that the number of fatal injury cases varies a lot.

4.2.2 Continuation-ratio logit model

According to section 3.9, the continuation-ratio logit model is applied on the data set. The equivalent dual-Poisson models are set up one by one as follows.

(1) Given the accident falls in categories fatal crash and less severe categories, the the continuation-ratio logit model is:

$$\begin{cases} Y_{it1}|(U^{(1)}, V^{(1)}, W^{(1)}) \sim \text{Poisson}\{W_{it1}^{(1)} \exp[\alpha_1 + x_{it}^T\beta_1 + \delta_1 + x_{it}^T\gamma_1]\}, \\ Y_{it2} + ... + Y_{it5}|(U^{(1)}, V^{(1)}, W^{(1)}) \sim \text{Poisson}\{W_{it2}^{(1)} \exp[\delta_1 + x_{it}^T\gamma_1]\}. \end{cases}$$

(2) Given the accident falls in categories major injury and less severe categories, the the continuation-ratio logit model is:

$$\begin{cases} Y_{it2}|(U^{(2)}, V^{(2)}, W^{(2)}) \sim \text{Poisson}\{W_{it1}^{(2)} \exp[\alpha_2 + x_{it}^T\beta_2 + \delta_2 + x_{it}^T\gamma_2]\}, \\ Y_{it3} + ... + Y_{it5}|(U^{(2)}, V^{(2)}, W^{(2)}) \sim \text{Poisson}\{W_{it2}^{(2)} \exp[\delta_2 + x_{it}^T\gamma_2]\}. \end{cases}$$

(3) Given the accident falls in categories moderate injury and less severe categories, the the continuation-ratio logit model is:

$$\begin{cases} Y_{it3}|(U^{(3)}, V^{(3)}, W^{(3)}) \sim \text{Poisson}\{W_{it1}^{(3)} \exp[\alpha_3 + x_{it}^T\beta_3 + \delta_3 + x_{it}^T\gamma_3]\}, \\ Y_{it4} + ... + Y_{it5}|(U^{(3)}, V^{(3)}, W^{(3)}) \sim \text{Poisson}\{W_{it2}^{(3)} \exp[\delta_3 + x_{it}^T\gamma_3]\}. \end{cases}$$

(4) Given the accident falls in categories minor injury and property damage only crash, the the continuation-ratio logit model is:

$$\begin{cases} Y_{it4}|(U^{(4)}, V^{(4)}, W^{(4)}) \sim \text{Poisson}\{W_{it1}^{(4)} \exp[\alpha_4 + x_{it}^T\beta_4 + \delta_4 + x_{it}^T\gamma_4]\}, \\ Y_{it5}|(U^{(4)}, V^{(4)}, W^{(4)}) \sim \text{Poisson}\{W_{it2}^{(4)} \exp[\delta_3 + x_{it}^T\gamma_3]\}. \end{cases}$$

where

x_{it}=("YEAR", LOWESS(Total))^T;

Y_{itj} is the crash count of category j;

β_j=(β_{j1}, β_{j2})^T;

γ_j=(γ_{j1}, γ_{j2})^T;

and j=1, 2, 3, 4, 5 which represent FAT, IA, IB IC and PDO.

As mentioned in section 3.9, these dual-Poisson models are estimated separately. The estimates of the variance of the random effects are generated from each estimation. So there are four sets of estimates for the variance of the random effects. The results are shown in table 4.5.

All of the regression parameters are statistically significant. And the result can be interpreted as follows:

(1). The coefficients of LOWESS surface are statistically significant. The significance suggests that we should not ignore the nonlinear association between the accident count in different types and the spatial location.

(2). Given the accident falls in categories fatal crash and less severe categories, for 1 year increase, the estimated odds that the accident type is fatal crash rather than less severe categories equals exp($\hat{\beta}_{11}$)=exp(-0.40) = 0.67 times the estimated odds at current year. In another word, given the accident falls in categories fatal crash and less severe categories, as time goes, the possibility of getting fatal crash is lower. Given the accident falls in categories major injury and less severe categories, for 1 year increase, the estimated odds that the accident type is major injury rather than less severe categories equals exp($\hat{\beta}_{21}$)=exp(-0.15) = 0.86 times the estimated odds at current year. In other words, given the accident falls in categories major injury and less severe categories, over time, the possibility of getting major injury is lower. Given the accident falls in categories moderate injury and less severe categories, for 1 year increase, the estimated odds that the accident type is moderate injury rather

Table 4.5: The results of the continuation-ratio logit model

	j=1		j=2		j=3		j=4	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
α_1	-3.8717	0.9092						
β_{11}	-0.4041	0.0976						
β_{12}	-0.1628	0.0637						
δ_1	0.9469	0.0607						
γ_{11}	0.0699	0.0029						
γ_{12}	0.07	0.0041						
α_2			-2.6712	0.1275				
β_{21}			-0.1481	0.0127				
β_{22}			-0.0214	0.008				
δ_2			0.893	0.0608				
γ_{21}			0.074	0.0029				
γ_{22}			0.0705	0.0041				
α_3					-1.0837	0.0647		
β_{31}					-0.1318	0.0064		
β_{32}					-0.0177	0.004		
δ_3					0.6423	0.0615		
γ_{31}					0.0915	0.0031		
γ_{32}					0.0727	0.0042		
α_4							-1.3562	0.0768
β_{41}							-0.1446	0.0076
β_{42}							-0.0157	0.0047
δ_4							0.4449	0.0621
γ_{41}							0.1061	0.0032
γ_{42}							0.0745	0.0042
σ^2	0.1103		0.1104		0.111		0.1087	
τ^2	0.0649		0.0649		0.0622		0.0713	
ω_1^2	1.4249		0.4197		0.1403		0.1805	
ω_2^2	0.0264		0.0271		0.0351		0.0331	
ρ	0.5465		0.5443		0.5401		0.5236	

than less severe categories equals exp($\hat{\beta}_{31}$)=exp(-0.13) = 0.88 times the estimated odds at current year. In other words, given the accident falls in categories moderate injury and less severe categories, as time goes by, the possibility of getting moderate injury is lower. Given the accident falls in categories minor injury and property damage only crash, for 1 year increase, the estimated odds that the accident type is minor injury rather than property damage only crash equals exp($\hat{\beta}_{41}$)=exp(-0.14) = 0.87 times the estimated odds at current year. In other words, given the accident falls in categories minor injury and property damage only crash, over time, the possibility

of getting minor injury is lower. Overall, as time goes, the possibility getting severe accident is lower. This result is same as what we get using baseline-category logit models.

(3). The estimates of the variances of the intersection-specific and the time-specific random effects ($\hat{\sigma}^2$ and $\hat{\tau}^2$) for these four separate logits are more or less the same. This follows common sense: the intersections and the years of taking the observations are not changed, so the random effects caused by them should have no change.

(4). The estimates of the variances of the category-specific random effects ($\hat{\omega}_1^2$ and $\hat{\omega}_2^2$) for these four separate logits are different. This implies that besides intersection-specific and time-specific random effects there are category-specific random effects which impact the number of crashes. And all the variances are different with each other, which implies that the variances of accident counts are different by category. In all of these, the variance for fatal injury is dramatically bigger than the others, which means that the number of fatal injury cases varied a lot. This result is the same as the baseline-category logit model results.

4.3 Simulation

In this section, the performance of the proposed models and the estimation mothods are evaluated by simulation studies. Because, in most scenarioes, the category-specific variations are different, the full model is evaluated.

4.3.1 Simulation study for the full model

To evaluate the performance, 500 independent runs are deployed. For each run, a simulation data with 100 subjects, 4 time points and 3 categories is generated. And the run stops when the sum of all parameters' differences is less than 0.00001. Then, the parameters average versus set value (true value) are compared. The simulation

data is generated as below:

(1). Construct covariance matrix[4x4] following: entry[ij]=$\sigma^2+\tau^2\rho^{|i-j|}$;

(2). Generate 4 random numbers following log-normal distrubution with mean 1 and covariance matrix got from (1). And repeat this prcess 100 times to get 100 sets of 4 correlated data. These data are V.

(3). Generate 3 random numbers following Gamma distribution with mean equals V_{it} and variances equal $\omega_1^2 V_{it}$, $\omega_2^2 V_{it}$ and $\omega_3^2 V_{it}$ respectively. And repeat this process for each V_{it} with i =1,...,100 and t=1, 2, 3, 4. These data are W.

(4). Generate 1 random number following Poisson distribution with mean and variance $W_{itj}\mu_{itj}$. And get 1200 random numbers for i=1,...100, t=1,2,3,4 and j=1,2,3. These data are Y.

Note: if following common sense, (1) and (2) should be generating U following Gamma distribution with mean 1 and variance σ^2 and then generate V following Log-normal distribution with mean U and variance τ^2. However, this process (mathematically) generates too many extreme numbers which lead to bigger variance than the true value.

The average of the regression and dispersion estimates over 500 simulations are showed as follows. The regression parameters are very similar as the true value and

Table 4.6: The simulation results

	Estimated Value	True Value	Bias	Simulated SE	Mean SE
β_{10}	-3.6267	-3.5000	0.1267	0.5916	0.5865
β_{11}	-0.1832	-0.1800	0.0032	0.1438	0.1441
β_{12}	0.0114	0.0100	-0.0014	0.0063	0.0066
β_{20}	-3.5118	-3.5000	0.0118	0.5202	0.5205
β_{21}	-0.1814	-0.1800	0.0014	0.1472	0.1478
β_{22}	0.0100	0.0100	0.0000	0.0058	0.0055
α_0	0.5148	0.5000	-0.0148	0.1111	0.1108
α_1	0.0939	0.1000	0.0061	0.0256	0.0254
α_2	0.0099	0.0100	0.0001	0.0015	0.0015
σ^2	0.0783	0.1000	0.0217	0.0255	
τ^2	0.0851	0.0700	-0.0151	0.0246	
ω_1^2	1.2151	1.3000	0.0849	0.0998	
ω_2^2	0.1155	0.1000	-0.0155	0.0385	
ω_3^2	0.0316	0.0300	-0.0016	0.0186	
ρ	0.4262	0.5000	0.0738	0.0456	

the random effects parameters are reasonably estimated through the model. And, the random effects parameter σ^2 is underestimated and τ^2 is overestimated, which makes the sum of the estimated values the same as the sum of the true values.

The simulated SE in the simulation result is the standard error of these 500 estimated values for each parameter. The Mean SE is the mean of 500 standard errors generated from these 500 simulation runs. They are quite similar as well.

All of these results prove that the proposed model and the estimation procedure are capable of handling the traffic data set used in this Chapter.

Chapter 5

Discussion

5.1 Conclusion

In this thesis, the baseline-category logit model and continuation-ratio logit model, which are constructed by the Poisson mixed models, are proposed to handle the data set of the crash counts over one thousand major signalized intersections in the city of Chicago from 2004 to 2010, which shows both longitudinal and categorical properties. And 3 levels of random effects including subject-specific, time-specific and category-specific random effects are introduced to account for the intersection, the year of taking the observation and the categories impacts to the crash counts. These models fit the vehicular crashes data set well in terms of the fitted values.

In these models, the correlation is handled by the 3 levels of random effects. The random effects are predicted using the best linear unbiased predictor (BLUP). The BLUP depends only on the first two moments of random effects, which gives the proposed model the flexibility to handle the data set with unknown random effect distribution and the advantage of computational efficiency[19]. Moreover, these 3 levels of random effects can be reorganized by eliminating either the intersection-specific random effects or category-specific random effects to fit different date set or

research interests.

For the crash data set itself, the results of the analysis tell us that, over time, the number of property damage only crashs will be increased while other categories will be reduced and the possibility of getting the severer accident is lower. And by analyzing the intersection-specific random effects, certain intersections, which could have other effects that may dramatically impact the number of accidents, are identified.

5.2 Discussion

During the results analysis, we observe that across all the crash categories, the variance of fatal crash is the biggest. And by detailed exam of the raw data set, we find only 1 and 0 are present in this category's count for each intersection in each year. This makes it difficult for our model to fit this category's data. This proposes that model check may be another scope to this model. In this thesis, all the evaluations are done by comparing the fitted value. Other evaluation methods like cross-validation and residual analysis could give us some other understandings of the model.

In section 4.2.1.1, the models with different levels of random effects are used to analyze the crash data set. The variances for the random effects of each model are estimated. We can see that the sums of the variances for each model show certain pattens:

(1). the sum of the intersection-specific random effects variance and the time-specific random effects variance are more or less the same: for the full model $\sigma^2+\tau^2=0.18$, for model 1 $\sigma^2+\tau^2=0.20$, and for model 2 which has only time-specific random effects $\tau^2=0.16$;

(2). whether or not including the category-specific random effects will not impact

the variances of the intersection-specific random effects variance and the time-specific random effects variance. For model 1 which do not have the category-specific random effects, the sum of $\sigma^2+\tau^2$ is the same as for the other two models which include the category-specific random effects.

These pattens suggest that the model could separate the category-specific random effects from the intersection-specific random effects and the time-specific random effects which are the same for the different categories. The sum of the intersection-specific random effects variance and the time-specific random effects variance could account for these two levels of random effects. However, there are mix-ups between the intersection-specific random effects variance and the time-specific random effects variance. This is also shown by the simulation results in section 4.3. However, our estimation algorithm only gives point estimates for the variances without any significance analysis. This may lead to two further study areas: to modify the variances estimation process to seperate these two random effect variances and to give interval estimates for these variances as in the bootstrap method[27] and Bayesian method[28].

For the regression part, we select year and the fitted value of the LOWESS curve of total crash counts over latitude and longitude as the covariates based on the basic statistics analysis of the data set in section 4.1. However there are other organizations of the covariates like adding year^2 as another covariate. This suggests that model-selection criteria might be helpful, like the Akaike Information Criterion (AIC) for GEE models.

Bibliography

[1] Chowdhury, Rafiqul I., 2012, *Inferences in longitudinal multinomial fixed and mixed models*, Doctoral (PhD) thesis, Memorial University of Newfoundland.

[2] Yenpeng Li, Wenyaw Chan, 2006, *Analysis of Longitudinal Multinomial Outcome Data*, Biometrical Journal, Volume48, Issue2.

[3] Donald Hedeker, Robert D. Gibbons, 2006, *Longitudinal Data Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey.

[4] Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. 2009, *Longitudinal Data Analysis*, Chapman & Hall/CRC Taylor Francis Group, Boca Raton.

[5] Throp, J., 2009, *Joint mixed-effects models for longitudinal data analysis: an application for the metabolic syndrome*, Virginia Commonwealth University, Master thesis.

[6] Student, 1908, *The Probable Error of a Mean*, Biometrika, Vol. 6, No. 1 (Mar., 1908), pp. 1-25.

[7] Peter J. Diggle, Patrick Heagerty, Kung-Yee Liang, Scott L. Zeger, 2002, *Analysis of Longitudinal Data* , Oxford University Press Inc., New York.

[8] Alan Agresti, 2007, *An Introduction to Categorical Data Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey.

[9] Alan Agresti, 2010, *Analysis of Ordinal Categorical Data: Second Edition*, John Wiley & Sons, Inc., Hoboken, New Jersey.

[10] R. Ma, B. Jørgensen, and J. D. Willms, 2009, *Clustered binary data with random cluster sizes: A dual Poisson modelling approach*, Statist. Model. 9 (2009), pp. 137–150.

[11] McCullagh P, Nelder JA, 1989, *Generalized Linear Models: 2nd edition*, p.101, Chapman and Hall, London New York.

[12] Guohua Yan, M.Tariqul Hasan, Renjun Ma, 2015, *Modeling proportions and marginal counts simultaneously for clustered multinomial data with random cluster sizes*, Journal of Applied Statistics, 43(6):1-14.

[13] David Ruppert, M. P. Wand, R. J. Carroll, 2003, *Semiparametric Regression*, Cambridge University Press, New York.

[14] Zilin Wang, Renjun Ma, Shudong Li, 2013, *Assessing area-specific ralative risks from large forest fire size in Canada*, Environmental and Ecological Statistics, Vol 20, Number 2.

[15] Chong Gu, 2013, *Smoothing Spline ANOVA Models*, Springer, New York.

[16] Alex Snow, 2019, *Tweedie Nonlinear Mixed Models for Longitudinal Data*, Master's thesis, University of New Brunswick.

[17] Renjun Ma, 1999, *An Orthodox BLUP Approach to Generalized Linear Models*, Ph.D. thesis, University of British Columbia.

[18] Renjun Ma, Guohua Yan, Tariqul Hasan, 2018, *Tweedie family of generalized linear models with distribution-free random effects for skewed longitudinal data: Tweedie mixed models for skewed longitudinal data*, Statistics in Medicine, vol. 37, no. 24, 3519-3532.

[19] Jiaxiu Li, 2018, *Joint Tweedie Mixed Models for Longitudinal Data of Mixed Types*, Master's thesis, University of New Brunswick.

[20] George Casella, Roger L. Berger, 2002, *Statistical Inference: Second Edition*, Thomson Press (India) Ltd, Delhi, India.

[21] Ma, R , Jrgensen, B, 2007, *Nested generalized linear mixed models: An orthodox best linear unbiased predictor approach*, Journal of the Royal Statistical Society, Series B (Statistical Methodology), vol. 69, no. 4, 625-641.

[22] R. W. M. Wedderburn, 1974, *Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method*, Biometrika, Vol. 61, No. 3 (Dec., 1974), pp. 439-447.

[23] Jorgensen, Bent, 1987, *Exponential dispersion models (with discussion)*, J. Roy. Statist. Soc. Ser. B 49, pp127-162.

[24] Ben-Israel, Adi, Greville, Thomas N.E., 2003, *Generalized inverses: Theory and applications: Second Edition*, Springer, New York.

[25] Yi Liu, Zongzhi Li, Jingxian Liu, 2016, *Vehicular crash data used to rank intersections by injury crash frequency and severity*, Elsevier, Data in Brief 8 (2016) 930–933.

[26] Jason Buss, 2009, *Mixed models for clustered categorical data with varying cluster sizes*, Master's report, University of New Brunswick.

[27] Lele, S., 1991, *Resampling using estimating equations. Theory of Estimating Equations*, (ed. Godambe, V.P.) Oxford: Clarendon Press.

[28] William J. Browne, and David Draper, 2006, *A comparison of Bayesian and likelihood-based methods for fitting multilevel models*, Bayesian Analysis (2006) 1, Number 3, pp. 473–514.

Appendix A

Top 200 high risk intersections in terms of tendency to have higher than average crash frequencies in all categories and in all years

Ranking	X_CORD	Y_CORD	LONG	LAT	Intersection No.	$\hat{U}_i$
1	1188267.86	1852925.67	-87.59	41.75	1	2.61
2	1160996.80	1878757.02	-87.68	41.82	7	2.51
3	1154467.53	1921103.21	-87.71	41.94	2	2.50
4	1161159.94	1873448.62	-87.68	41.81	17	2.13
5	1165388.02	1912729.61	-87.67	41.92	14	2.12
6	1162614.71	1915951.05	-87.68	41.93	5	2.03
7	1180003.37	1863326.72	-87.62	41.78	39	2.01
8	1149011.14	1926281.36	-87.73	41.95	10	1.89
9	1143506.63	1931483.81	-87.75	41.97	4	1.88
10	1177343.93	1863238.77	-87.63	41.78	54	1.87

Ranking	X_CORD	Y_CORD	LONG	LAT	Intersection No.	$\hat{U}_i$
11	1177456.38	1847282.85	-87.63	41.74	16	1.82
12	1159963.77	1915885.23	-87.69	41.92	9	1.81
13	1168605.77	1938053.88	-87.66	41.99	15	1.81
14	1159231.62	1939785.71	-87.69	41.99	8	1.79
15	1127522.01	1920435.75	-87.81	41.94	44	1.78
16	1188504.43	1842299.32	-87.59	41.72	3	1.77
17	1132569.60	1934189.98	-87.79	41.98	38	1.76
18	1173912.27	1901257.20	-87.64	41.88	52	1.76
19	1156418.25	1852064.73	-87.70	41.75	21	1.76
20	1150639.44	1867824.70	-87.72	41.79	34	1.75
21	1144002.10	1915529.45	-87.75	41.92	11	1.74
22	1167128.07	1942648.56	-87.66	42.00	56	1.72
23	1157229.59	1918486.82	-87.70	41.93	24	1.72
24	1144390.08	1899592.07	-87.75	41.88	69	1.71
25	1189598.28	1867846.95	-87.58	41.79	85	1.70
26	1170542.91	1916168.88	-87.65	41.93	23	1.69
27	1155522.21	1883904.49	-87.70	41.84	124	1.68
28	1177657.48	1852637.49	-87.62	41.75	112	1.66
29	1175037.39	1904162.64	-87.63	41.89	136	1.61
30	1177324.39	1899960.95	-87.62	41.88	70	1.61
31	1164793.26	1931948.16	-87.67	41.97	121	1.59
32	1161234.86	1870796.40	-87.68	41.80	113	1.59
33	1166869.28	1858961.70	-87.66	41.77	172	1.59
34	1151093.58	1851927.29	-87.72	41.75	27	1.58
35	1133351.74	1915254.45	-87.79	41.92	116	1.56
36	1167189.62	1847025.33	-87.66	41.74	72	1.56
37	1172649.70	1841845.60	-87.64	41.72	37	1.56
38	1146770.72	1910270.46	-87.74	41.91	30	1.55
39	1172337.90	1852478.52	-87.64	41.75	171	1.55
40	1172072.46	1847159.25	-87.65	41.74	35	1.53
41	1166613.61	1868242.31	-87.66	41.79	177	1.53
42	1162740.52	1910648.44	-87.68	41.91	22	1.53
43	1144209.65	1904877.11	-87.75	41.90	63	1.53
44	1154413.17	1939711.75	-87.71	41.99	48	1.53
45	1174919.27	1911704.95	-87.63	41.91	92	1.53
46	1167046.71	1852332.72	-87.66	41.75	100	1.52
47	1150543.51	1871144.92	-87.72	41.80	107	1.52
48	1171214.40	1894928.46	-87.65	41.87	104	1.52
49	1178296.95	1902192.92	-87.62	41.89	83	1.52
50	1182784.75	1858113.20	-87.61	41.77	155	1.51

Ranking	X_CORD	Y_CORD	LONG	LAT	Intersection No.	$\hat{U}_i$
51	1166170.82	1884149.15	-87.67	41.84	175	1.51
52	1119351.99	1925509.22	-87.84	41.95	145	1.51
53	1166052.33	1889485.94	-87.67	41.85	95	1.51
54	1160118.43	1910573.20	-87.69	41.91	71	1.51
55	1165087.15	1921316.33	-87.67	41.94	42	1.50
56	1132979.85	1925907.07	-87.79	41.95	40	1.50
57	1173045.88	1828593.19	-87.64	41.69	123	1.50
58	1191171.55	1842370.02	-87.58	41.72	57	1.49
59	1182646.42	1863411.24	-87.61	41.78	143	1.49
60	1144505.07	1896598.69	-87.74	41.87	139	1.48
61	1188077.68	1860882.96	-87.59	41.77	146	1.47
62	1177360.28	1898101.72	-87.62	41.88	41	1.47
63	1139963.92	1931404.66	-87.76	41.97	103	1.46
64	1138644.67	1915399.33	-87.77	41.92	79	1.45
65	1151357.34	1897096.50	-87.72	41.87	164	1.45
66	1159878.28	1918537.02	-87.69	41.93	13	1.45
67	1177384.29	1897167.88	-87.62	41.87	36	1.45
68	1173817.53	1904421.57	-87.64	41.89	64	1.44
69	1171909.65	1868385.22	-87.65	41.79	275	1.44
70	1177137.29	1855257.25	-87.63	41.76	271	1.44
71	1173973.55	1898933.73	-87.64	41.88	157	1.44
72	1161750.96	1852199.44	-87.68	41.75	87	1.44
73	1173311.53	1894981.33	-87.64	41.87	96	1.43
74	1160981.69	1879828.98	-87.68	41.83	90	1.43
75	1149579.53	1905046.96	-87.73	41.90	120	1.43
76	1173005.19	1921539.95	-87.64	41.94	115	1.42
77	1165339.24	1913386.89	-87.67	41.92	53	1.42
78	1155831.00	1873295.88	-87.70	41.81	114	1.42
79	1155303.50	1891462.62	-87.71	41.86	277	1.42
80	1161882.03	1846888.94	-87.68	41.74	131	1.42
81	1173948.59	1899859.52	-87.64	41.88	86	1.41
82	1143810.23	1920843.67	-87.75	41.94	46	1.41
83	1145506.22	1867725.63	-87.74	41.79	19	1.41
84	1154636.74	1915778.55	-87.71	41.92	147	1.41
85	1185523.48	1855582.53	-87.60	41.76	28	1.41
86	1165583.38	1905422.99	-87.67	41.90	188	1.40
87	1172701.26	1901238.64	-87.64	41.88	99	1.40
88	1141514.74	1931436.52	-87.76	41.97	78	1.40
89	1161454.67	1862806.40	-87.68	41.78	102	1.40
90	1175899.23	1868512.77	-87.63	41.79	174	1.40
91	1149317.16	1915655.63	-87.73	41.92	67	1.40
92	1167994.26	1908147.09	-87.66	41.90	185	1.39
93	1143904.21	1918186.79	-87.75	41.93	29	1.39

Ranking	X_CORD	Y_CORD	LONG	LAT	Intersection No.	$\hat{U}_i$
94	1176354.74	1900824.22	-87.63	41.88	243	1.39
95	1160578.90	1893985.05	-87.69	41.86	163	1.39
96	1176004.76	1898054.72	-87.63	41.88	76	1.38
97	1172728.42	1839206.74	-87.64	41.71	61	1.38
98	1158494.12	1873375.83	-87.69	41.81	173	1.38
99	1141323.42	1915463.13	-87.76	41.92	101	1.38
100	1145458.81	1862404.25	-87.74	41.78	138	1.38
101	1147235.24	1894340.43	-87.73	41.87	261	1.37
102	1176150.65	1898057.34	-87.63	41.88	50	1.37
103	1159154.62	1942419.86	-87.69	42.00	45	1.36
104	1139056.96	1899464.30	-87.76	41.88	214	1.36
105	1166750.43	1862937.59	-87.66	41.78	183	1.36
106	1177323.64	1899497.03	-87.62	41.88	144	1.36
107	1164586.40	1939534.45	-87.67	41.99	94	1.36
108	1157447.33	1910515.50	-87.70	41.91	224	1.35
109	1149728.99	1899712.71	-87.73	41.88	227	1.35
110	1167449.41	1932022.95	-87.66	41.97	182	1.35
111	1150785.84	1862524.96	-87.72	41.78	109	1.34
112	1171051.50	1900256.37	-87.65	41.88	245	1.34
113	1166907.58	1857634.56	-87.66	41.76	294	1.34
114	1173133.65	1825945.56	-87.64	41.68	196	1.34
115	1165268.62	1916017.22	-87.67	41.93	55	1.34
116	1164940.59	1926633.56	-87.67	41.95	75	1.34
117	1173309.48	1820637.69	-87.64	41.66	204	1.34
118	1144089.53	1910204.65	-87.75	41.91	26	1.33
119	1176445.90	1898062.65	-87.63	41.88	58	1.33
120	1159413.52	1934463.38	-87.69	41.98	33	1.32
121	1150052.64	1889081.19	-87.72	41.85	150	1.32
122	1143739.86	1923502.98	-87.75	41.95	81	1.32
123	1156198.73	1905179.24	-87.70	41.90	390	1.32
124	1175555.05	1868498.89	-87.63	41.79	229	1.32
125	1177806.57	1847296.14	-87.62	41.74	89	1.32
126	1174801.03	1898031.78	-87.63	41.88	225	1.32
127	1132520.28	1942097.10	-87.79	42.00	73	1.31
128	1175069.16	1902989.36	-87.63	41.89	276	1.31
129	1152683.12	1890237.15	-87.72	41.85	258	1.31
130	1172498.06	1847169.50	-87.64	41.74	66	1.31
131	1160781.33	1886715.12	-87.69	41.84	251	1.31
132	1165823.62	1897809.75	-87.67	41.88	111	1.31
133	1161382.37	1865470.71	-87.68	41.79	206	1.30
134	1173935.99	1900330.14	-87.64	41.88	149	1.30
135	1177329.78	1904235.63	-87.62	41.89	32	1.30
136	1136064.68	1910015.45	-87.78	41.91	159	1.30

Ranking	X_CORD	Y_CORD	LONG	LAT	Intersection No.	$\hat{U}_i$
137	1134648.92	1867404.94	-87.78	41.79	153	1.29
138	1160377.43	1901214.65	-87.69	41.88	395	1.29
139	1178436.26	1895117.95	-87.62	41.87	12	1.29
140	1172795.83	1836548.09	-87.64	41.71	137	1.29
141	1182931.38	1852800.34	-87.61	41.75	207	1.29
142	1176339.78	1901287.90	-87.63	41.88	350	1.29
143	1165433.29	1910726.60	-87.67	41.91	51	1.28
144	1148639.00	1939575.65	-87.73	41.99	84	1.28
145	1149902.67	1894409.91	-87.73	41.87	180	1.28
146	1177577.85	1855268.45	-87.62	41.76	341	1.28
147	1148787.20	1934254.87	-87.73	41.98	106	1.28
148	1143656.43	1926160.53	-87.75	41.95	31	1.28
149	1158120.12	1886641.82	-87.70	41.84	264	1.28
150	1149473.97	1910337.22	-87.73	41.91	128	1.27
151	1183083.43	1847467.75	-87.60	41.74	117	1.27
152	1162530.29	1918595.60	-87.68	41.93	59	1.27
153	1160933.88	1881411.44	-87.69	41.83	215	1.27
154	1139035.19	1900192.32	-87.76	41.88	472	1.27
155	1182383.45	1874021.89	-87.61	41.81	588	1.26
156	1151439.63	1894448.32	-87.72	41.87	270	1.26
157	1151343.16	1897437.27	-87.72	41.87	282	1.26
158	1156543.34	1894561.23	-87.70	41.87	444	1.26
159	1141543.95	1904834.20	-87.76	41.89	320	1.26
160	1161308.86	1867985.32	-87.68	41.79	186	1.26
161	1150484.87	1873172.89	-87.72	41.81	168	1.26
162	1179417.33	1884525.34	-87.62	41.84	110	1.26
163	1164725.16	1934606.18	-87.67	41.98	265	1.25
164	1135998.51	1915325.66	-87.78	41.92	199	1.25
165	1172962.78	1831243.24	-87.64	41.69	236	1.25
166	1151766.85	1828025.68	-87.72	41.68	367	1.25
167	1175107.78	1901271.26	-87.63	41.88	340	1.25
168	1165913.31	1894780.04	-87.67	41.87	91	1.25
169	1162300.99	1926564.62	-87.68	41.95	80	1.25
170	1169402.41	1910834.12	-87.65	41.91	134	1.25
171	1162206.36	1836268.50	-87.68	41.71	314	1.24
172	1165809.84	1898237.41	-87.67	41.88	194	1.24
173	1177301.76	1905757.54	-87.62	41.90	166	1.24
174	1161527.82	1860169.19	-87.68	41.77	246	1.23
175	1153163.41	1921074.95	-87.71	41.94	20	1.23
176	1144435.14	1898247.68	-87.75	41.88	256	1.23
177	1137842.81	1934235.07	-87.77	41.98	218	1.23
178	1166682.40	1865590.83	-87.66	41.79	336	1.23
179	1150712.09	1865176.75	-87.72	41.79	211	1.23
180	1163413.88	1888082.56	-87.68	41.85	262	1.23

Ranking	X_CORD	Y_CORD	LONG	LAT	Intersection No.	$\hat{U}_i$
181	1180435.02	1847392.14	-87.61	41.74	238	1.23
182	1144369.63	1900072.34	-87.75	41.88	222	1.22
183	1165507.34	1908072.73	-87.67	41.90	93	1.22
184	1174958.35	1898032.85	-87.63	41.88	234	1.22
185	1160708.92	1889336.01	-87.69	41.85	316	1.22
186	1188149.79	1858262.04	-87.59	41.77	161	1.22
187	1147438.60	1887659.85	-87.73	41.85	129	1.22
188	1165833.87	1897453.88	-87.67	41.87	141	1.22
189	1177281.74	1901288.61	-87.62	41.88	151	1.22
190	1146787.71	1910373.18	-87.74	41.91	105	1.22
191	1144484.47	1896911.54	-87.74	41.87	325	1.22
192	1146327.99	1901671.41	-87.74	41.89	492	1.21
193	1138557.48	1918057.60	-87.77	41.93	132	1.21
194	1171760.48	1873690.67	-87.65	41.81	490	1.21
195	1159015.89	1947721.31	-87.69	42.01	167	1.21
196	1175527.72	1900814.76	-87.63	41.88	329	1.21
197	1156137.35	1862678.38	-87.70	41.78	376	1.20
198	1171694.33	1894941.24	-87.65	41.87	248	1.20
199	1154745.89	1910458.44	-87.71	41.91	219	1.20
200	1190840.11	1855634.65	-87.58	41.76	324	1.20

Vita

Candidate's full name: Zheng Fu

University attended :

Bachelor of Science, 1999, Northwestern Polytechnical University, Xi'an China

Master of Science, 2003, Northwestern Polytechnical University, Xi'an China

Master of Science, 2020, University of New Brunswick, Fredericton, NB, Canada

Publications: None

Conference Presentations: None