5-2019

Statistical Modeling of Count Data with Over-Dispersion or Zero-Inflation Problems

Chengxin Zhang
Montclair State University

Abstract

In this study, we will analyze a supply retailing company's data to model the relationship between their customer's past purchase behavior to predict their future online purchase behavior. The data was divided into time periods from 2016: P1-P6(January 31st to July 30th) and P7(July 31st to August 27th). Based on customer's past purchase information from the P1-P6 period, such as money spent, number of cart additions, transactions type, number of unique purchase dates, number of unique purchase skus, number of page views, number browse dates, company size, and number of products purchased, we aim to find if these information could predict the customer's purchase behavior in the P7 period, which is the number of responses the customer responded to emails sent to them during P7. With the response variable as count data, we model the data in R with the Poisson distribution regression with an offset variable. We also model the number of responses out of the number of emails sent using a logistic regression model. For the Poisson model, since there are zero inflation or over-dispersion issues in the response, hurdle model, zero-inflated-poisson (ZIP) model and zero-inflated-negative-binomial (ZINB) model would be used to handle these issues. Model comparisons among the Poisson model with an offset, logistic regression model, hurdle model, ZIP, ZINB is conducted to select the best model to fit the data using the AIC criterion and the cross-validation criterion.

MONTCLAIR STATE UNIVERSITY

Statistical modeling of count data with over-dispersion or zero-inflation problems

by

Chengxin Zhang

A Master's Thesis Submitted to the Faculty of

Montclair State University

In Partial Fulfillment of the Requirements

For the Degree of

Master of Science

May 2019

College/School	The College of Science and Mathematics	Thesis Committee: [redacted]
Department	Mathematical Sciences	Dr. Haiyan Su, Thesis Sponsor
		[redacted]
		Dr. Andrew McDougall, Committee Member
		[redacted]
		Dr. Andrada Ivanescu, Committee Member

Statistical modeling of count data with over-dispersion or zero-inflation problems

A THESIS

Submitted in partial fulfillment of the requirements

for the degree of Master of Science

by

Chengxin Zhang

Montclair State University

Montclair, NJ

May 2019

Acknowledgements

I would like to thank Dr. Haiyan Su for her years of boundless support as an advisor and express my deepest appreciation to my thesis sponsor, for her support throughout this study. Without her guidance and persistent help, this thesis would not have been possible.

I would like to thank my committee members, Dr. Andrew McDougall and Dr. Andrada Ivanescu, for many helpful feedback and encouragement they have given to me during the thesis study.

Finally, I would like to thank my friends and family for their continuous support and love. None of this would have been possible without them.

Contents

Chapter 1

Introduction

1.1 Background

In this study, we aim to study the relationship between customer's past purchasing behavior and their future purchase behavior for a supply retailing company (SRC). The variable of interest is the number of purchases made by the customer in a certain period after receiving emails from the company. In business, psychology, social, and public health related research, it is common that the outcomes are relatively infrequent behaviors and phenomena. Data with abundant zeros are especially frequent in research studies when counting the occurrence of certain behavioral events, such as number of purchases made, number of school absences, number of cigarettes smoked, or number of hospitalizations. These types of data are called count data and their values are usually non-negative with a lower bound of zero. Common issues when dealing with count data are typically zero inflation (excessive zeros), over-dispersion (greater variability than expected) and under-dispersion (less variability than expected).

The classical Poisson regression model for count data is often of limited use in these disciplines because empirical count data sets typically exhibit over-dispersion, under-dispersion or an excess number of zeros. One way to deal with over-dispersion is a negative bino-

mial (NB) regression. The negative binomial model belongs to the family of generalized linear models [13]. However, although negative binomial model typically can capture over-dispersion rather well, it is in many applications not sufficient for modeling excess zeros. In the econometrics and statistics literature, Mullahy [12] and Lambert [10] proposed the zero-augmented models that address this modeling by a second model component capturing zero counts. Hurdle models [12] combine a left-truncated count component with a right-censored hurdle component. Zero-inflation models [10] take a somewhat different approach: they are mixture models that combine a count component and a point mass at zero. An overview of count data models in econometrics, including hurdle and zero-inflated models, is provided in Cameron and Trivedi [4, 5].

The remainder of this thesis is organized as follows: Chapter 2 discusses data cleaning and exploratory analysis of the SRC data. In Chapter 3, we discuss all count regression models and check over-dispersion, under-dispersion or zero-inflation for SRC data. In Chapter 4, we use three methods to select the best model for the SRC data. The summary in Chapter 5 concludes the main part of the thesis.

1.2 Data source

The data analyzed in this thesis for SRC account managed business customers and was collected between January 31, 2016 and September 03, 2016. Data from January 31st through July 30th represents SRC's fiscal calendar periods 1 through 6 (referred to as P1 through P6). July 31st through August 27th represents P7. For the purpose of this thesis we are interested in understanding whether and how a customer's information from P1 through P6 may be used to predict if they will respond to an e-mail sent during P7. The sample size of the SRC data is 34,579.

1.3 Variables used in the data

Rewards is a unique identifier of the customer.

Email is the number of emails sent to the customer during P7.

Responses is the number of emails sent during P7 that a customer responded to (a response is the act of making a purchase within 7 days of receiving the email).

Revenue is the amount of money spent by the customer during P1 through P6.

Units is the number of product units purchased by the customer during P1 through P6.

Purchasedates is the number of unique dates during which the customer made a purchase from P1 through P6.

Purchaseskus is the number of unique product SKUs the customer purchased during P1 through P6.

Carts is the number of product units the customer added to their online cart during P1 through P6.

Productviews is the number of product pages (including repeats) the customer viewed during P1 through P6.

Browsedates is the number of unique dates during which the customer browsed Staples.com from P1 through P6.

Companysize is the alphabetic code used to identify the size of the customer's business (if applicable). "X" represents unknown.

Businesssize is the alphabetic indicator of whether a customer is a small business (Y) or not (N) (if applicable).

Consumercode is the alphabetic indicator of whether a customer is a business (B) or consumer (C). "U" represents unknown.

Coupons is the percent of the customer's transactions from P1 through P6 which included the use of a coupon.

1.4 Data masking

The data has been masked to protect sensitive business information by SRC. Numeric variables (except number of emails and number of responses) have been standardized to z-scores (centered around their mean and divided by their standard deviation).

Chapter 2

Exploratory data analysis

In this section, we did data cleaning and conducted exploratory analysis for the variables involved in this study. Bar charts was provided to display the distribution for categorical variables, means and standard deviations were obtained for continuous variables, and histogram and a frequency table were used to display the distribution of the count response variable.

2.1 Number of responses

The histogram in Figure 2.1 illustrates the distribution of the number of responses made by customers. We can see clearly it exhibits both substantial variation and a large number of zeros in the response variable.

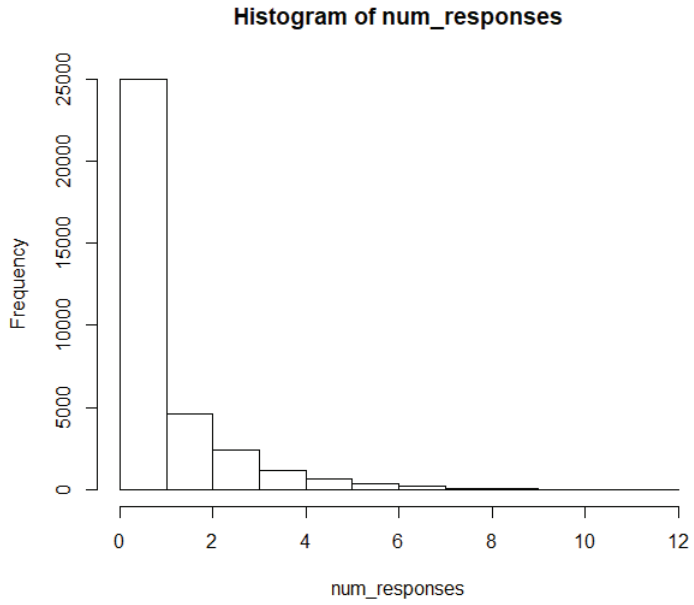

Figure 2.1: Number of responses made by customers

Table 2.1: Frequency of responses made by customers

Responses	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	15327	9651	4619	2393	1195	682	377	199	80	38	11	6	1

From Table 2.1, we found that more than 15,000 customers do not respond when they got the emails from the retailing company. Figure 2.1 indicates that the response might be zero-inflation. Based on the properties of the response variable, we consider the following steps to model the relationships in SRC data:

1. We analyze the SRC data using poisson and logistic models to predict our response variable.

2. We then use Quasi-Poisson to test whether the count data is over-dispersion or under-dispersion. If it is over-dispersion, we can use negative binomial model. If it is under-dispersion and zero-inflation, we can use Hurdle Model and Zero-inflated Model.

3. We use Vuong test [15], Akaike information criterion (AIC) [2] and Cross-Validation [8] to find which model is the best for our data.

2.2 Summary statistics of numerical variables

The second step in the exploratory analysis is to obtain the summary statistics for the numerical variables.

Table 2.2: Summary statistics of numerical variables

Variable Name	Mean	Std	Max	Med	Min
emails	2.309	1.488	12	2	1
responses	1.130	1.474	12	1	0
revenue	0.534	1.619	110.638	0.1475	-0.504
units	0.010	1.629	204.907	-0.056	-0.079
purchasedates	0.727	1.372	22.350	0.346	-1.087
purchaseskus	0.701	1.580	138.919	0.338	-0.914
carts	0.920	1.673	77.875	0.475	-0.442
productviews	0.779	1.728	83.181	0.272	-0.392
browsedates	1.047	1.402	18.077	0.671	-0.572
coupons	0.370	0.736	2.492	0.324	-1.122

Summary statistics of numerical variables are given in Table 2.2. The mean of the responses made by customers is 1.130 and the mean emails received by customers is 2.309; and the minimum of responses made by customers is 0, and the minimum of emails received by customers is 1; The maximum number of responses made by customers is 12 as in the maximum of emails received by customers. The standard deviation is 1.474 for responses made by customers, and the standard deviation of emails received by customers is 1.488.

2.3 Categorical variables

We obtained the distribution of the Categorical variables using bar plots and frequency tables shown below:

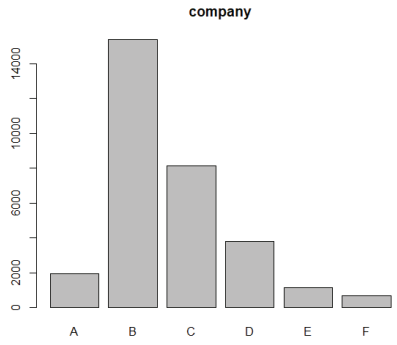

Figure 2.3: company size

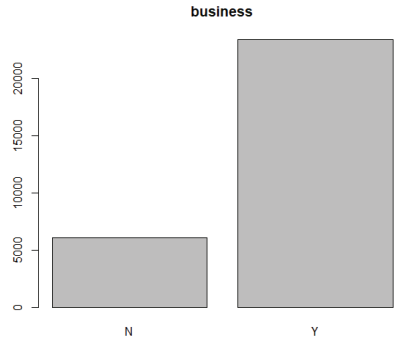

Figure 2.4: customer small or not business

Company size	Frequency	Percentage
A	1943	6.25%
B	15386	49.49%
C	8157	26.24%
D	3808	12.25%
E	1116	3.59%
F	678	2.18%

Table 2.3: Frequency of comany size

Small business	Frequency	Percentage
N	6065	20.61%
Y	23357	79.39%

Table 2.4: Frequency of Customer Small Business

Table 2.3 shows us that there are seven different sizes for companies. In past six months, 6.25% online purchase behaviors come from Company size A, 49.49% come from Company size B, 26.24% come from Company size C, 12.25% come from Company size D, 3.59% come from Company E, 2.18% come from Company size F. Table 2.4 Shows us whether the customer is a small business or not. 79.39% customers are small business.

Figure 2.4: business code

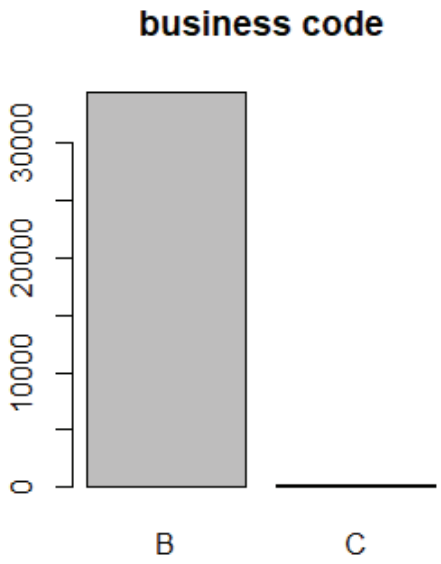

Table 2.5: Frequency of Business Code

Business	Frequency	Percentage
B	34344	99.39%
C	212	0.61%

Table 2.5 tells us there are two kinds of customer code Business and Consumer: 99.39% customers were business customers, only 0.61% customers were consumer customers. When including the customers business code in the models, the models do not converge, so we drop the variable consumer code in the modeling.

Chapter 3

Statistical model of the SRC data

In this section, we fit the SRC data using different statistical models listed in Table 3.1. The classic Poisson and negative binomial models are described in a generalized linear model (GLM) [11] framework; they are implemented in R [14] by the *glm* function in the *stats* package and the *glm.nb* function in $MASS$ package. The hurdle and zero-inflated extensions of these models are provided by the functions *hurdle* and $zeroinfl$ in package *pscl* [7]. Each model is introduced in more details separately in next section.

Table 3.1: Models and Functions

Type	Distribution	Method	Description
GLM	Poisson	ML	Poisson regression: classical GLM
		Quasi	Quasi-Poisson regression: overdispersion GLM
	NB	ML	NB regression: extended GLM
	Logistic	ML	Logistic regression: extended GLM
Zero-augmented	Poisson	ML	Zero-inflated Poisson: ZIP
			Hurdle Poisson: Hurdle
	NB	ML	Zero-inflated negative binomial: ZINB
			Hurdle negative binomial: Hurdle-NB

3.1 Classical Poisson model

3.1.1 Poisson model

The simplest distribution used for modeling count data is the Poisson distribution with probability density function

$$f(y;\mu) = \frac{\exp(-\mu)(\mu)^y}{y!}. \tag{3.1}$$

Poisson distribution is a special case of the generalized linear model (GLM) framework [11]. The canonical link is the *log* function resulting in a log-linear relationship between the mean and the linear predictor. The variance in the Poisson model is identical to the mean, thus the dispersion is fixed at $\phi = 1$ and the variance function is $V(\mu) = \mu$.

Poisson regression models allow researchers to examine the relationship between predictors and count outcome variables. Using these regression models gives much more accurate parameter estimates than trying to fit an least square linear regression model whose assumptions rarely fit count data such as normal residuals and constant variance.

In R, this can easily be specified in the *glm* function by setting $family = poisson$. As a first attempt to capture the relationship between the number of responses made by customers and the regressors, we fit the basic Poisson regression model. The output is given in Table 3.2:

Table 3.2: Poisson model output

Regressor	Estimate	Std Error	Z value	P-value
emails	0.365	0.003	113.590	<0.001
revenue	-0.023	0.004	-5.602	<0.001
purchasedates	0.121	0.007	18.105	<0.001
purchaseskus	0.015	0.005	3.254	0.001
carts	-0.021	0.004	-5.579	<0.001
productviews	-0.012	0.004	-2.801	0.005
browsedates	0.073	0.007	10.384	<0.001
browseskus	-0.039	0.007	-5.619	<0.001
coupons	-0.054	0.009	-6.358	<0.001
companysizeB	-0.009	0.025	-0.373	0.709
companysizeC	0.020	0.026	0.761	0.447
companysizeD	0.053	0.028	1.916	0.055
companysizeE	0.101	0.034	2.989	0.003
companysizeF	0.081	0.039	2.075	0.038
smallbusiness	0.008	0.014	0.587	0.557

In the output in Table 3.2, we can see the P-values for emails received, revenue, purchase dates, carts, productviews, browsedates, browseskus and coupons predictor variables are all statistically significant at 0.05 significance level. We can see the relationships between the number of responses made by customers and purchasedates, purchaseskus, browsedates are all positive. In specific, with one additional email the customer received, the log number of responses made by customers increases 0.365; for one additional unique date when the customer made a purchase in last six month, the log number of responses made by customers increases 0.121; for one additional unique product SKUs the customer purchased in last six months, the log number of responses made by customers increases 0.015. The relationships between the number of responses made by customers and revenue, carts, browsekus are

negative. Specifically, for each additional dollar spent by customers from last six months, the log number of responses made by customers decreases 0.023; for one additional product which the customers added to their online carts in last six months, the log number of responses made by customers decreases 0.021; for one additional product page that the customer viewed in last six months, the log number of responses made by customers decreases 0.012.

3.1.2 Poisson regression with an offset

Poisson model can also handle rates. A rate is just a count per unit time. Poisson models handle exposure variables by using simple algebra to change the dependent variable from a rate into a count. If the rate is $count/exposure$, when both sides of the equation are logged, the final model contains $log(exposure)$ as a term that is added to the regression coefficients. That is $log\left(\frac{Y}{exposure}\right) = log(Y) - log(exposure) = \theta' x$, which implies $log(Y) = log(exposure) + \theta' x$, where $log(exposure)$ is called the offset variable.

In the SRC data, since the response variable, the number of purchases made by the customers is considered only when they they received emails from SRC, modeling the count outcome (number of purchases) should consider the number of emails they received from SRC. Thus, instead of modeling count directly using emails as a covariate, we consider a poisson model with an offset, which is log(number of emails received) in the poisson regression. We fit the SRC data using a poisson model with an offset, the result is given in Table 3.3.

Table 3.3: Output from the poisson model with an offset

Regressor	Estimate	Std Error	Z value	P-value
revenue	-0.014	0.004	-3.388	<0.001
purchasedates	0.123	0.007	18.542	<0.001
purchaseskus	0.011	0.004	2.424	0.015
carts	-0.021	0.004	-5.422	<0.001
productviews	-0.009	0.004	-2.136	0.033
browsedates	0.079	0.007	11.658	<0.001
browseskus	0.004	0.007	0.634	0.527
coupons	-0.065	0.009	-7.545	<0.001
companysizeB	-0.0002	0.025	-0.006	0.995
companysizeC	0.031	0.026	1.179	0.238
companysizeD	0.054	0.028	1.948	0.051
companysizeE	0.090	0.034	2.663	0.008
companysizeF	0.047	0.039	1.218	0.223
smallbusiness	-0.006	0.014	-0.463	0.646

From Table 3.3, we can see the P-values for the revenue, purchasedates, purchaseskus, carts, productviews, browsedates, coupons predictor variables are all statistically significant at 0.05 significance level. Comparing with Poisson model, browseskus is not statistically significant at 0.05 significance level in the poisson offset model. We can see that the relationships between the rate of responses made by customers and purchasedates, purchaseskus, browsedates and browsekus are all positive. For example, with one additional unique date when the customer made a purchase in last six months, the log rate of responses made by customers increases 0.123, it is less than Poisson model in section 3.1.1. The relationships between the response rates made by customers and revenue, carts and productviews are all negative. For example, with one additional dollar spent by customers in last six months, the

log rate of responses made by customers decreases 0.014, it is less than the value in Poisson model in section 3.1.1 too.

3.2 Logistic regression

In modeling the rate of purchases out of the number of emails sent to the customers, a logistic regression can also be fitted to the data. If we use linear regression to model a dichotomous variable (as Y), the resulting model might not restrict the predicted responses within 0 and 1. Besides, other assumptions of linear regression such as normality of errors may get violated.

The logistic regression (or logit model) is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds for the value labeled "1" is a linear combination of one or more explanatory variables ("predictors"); the explanatory variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 and 1.

In logistic regression, we model the log odds of the event $log(\frac{P}{1-P})$, where P is the probability of event.

$$log(\frac{P_i}{1-P_i}) = \beta_0 + \beta_1 x_{1i} + .. + \beta_k x_{ki}, \tag{3.2}$$

where i is the index of i^{th} subjects, k is the index for the k^{th} predictor. The above equation can be modeled using the *glm* function by setting the family argument to "*binomial*". We fit SRC data in logistic regression and the results are given in Table 3.4.

Table 3.4: Logistic regression output

Regressor	Estimate	Std Error	Z value	P-value
revenue	0.004	0.011	0.337	0.736
purchasedates	0.439	0.015	28.253	<0.001
purchaseskus	0.144	0.019	7.670	<0.001
carts	0.019	0.013	1.480	0.139
productviews	0.025	0.011	2.293	0.022
browsedates	0.150	0.015	10.198	<0.001
browseskus	-0.116	0.022	-5.191	<0.001
coupons	-0.120	0.012	-9.701	<0.001
companysizeB	-0.017	0.038	-0.436	0.667
companysizeC	0.014	0.039	0.363	0.717
companysizeD	0.058	0.043	1.374	0.169
companysizeE	0.137	0.056	2.449	0.014
companysizeF	0.077	0.068	1.128	0.259
smallbusiness	-0.039	0.021	-1.811	0.070

From the output in Table 3.4, we can see the P-values for purchasedates, purchaseskus, productviews, browsedates, browsesuks and coupons are all statistically significant at 0.05 significance level. On the other hand carts is not statistically significant at 0.05 significance level. We can see that the relationships between the odds of responses made by customers and purchasedates, purchaseskus, productviews and browsedates are all positive. For example, with one additional unique date when the customer made a purchase in last six month, the log-odds of responses made by customers increases 0.439; For one additional unique product SKUs which the customer purchased in last six months, the log-odds of responses made by customers increases 0.144. Similarly we can see the odds of responses made by customers and browseskus and coupons are both negative. For example, with one additional product

which the customers added to their online carts in last six months, the log-odds of responses made by customers decreases 0.12.

3.3 Models for dealing with over/under-dispersion

3.3.1 Quasi-Poisson model

We can use two ways to test whether the model is over-dispersion or under-dispersion. One way to confirm with over-dispersion (under-dispersion) is to use the mean and variance functions from the Poisson GLM but to leave the dispersion parameter ϕ unrestricted. Thus ϕ is not fixed at 1 but is estimated from the data. This strategy leads to the same coefficient estimates as the standard Poisson model but inference is adjusted for over-dispersion. This procedure is called Quasi-Poisson procedure and is usually used to confirm the over-dispersion in the data. In R, the quasi-Poisson model with estimated dispersion parameter can also be fitted using the *glm* function by setting $family = quasipoisson$.

We fit SRC data using the Quasi-Poisson regression model and the dispersion estimate is 0.663 from the result. Since the estimator is less than one, it turns out the conditional variance is actually smaller than the conditional mean, which indicates we have under-dispersion in the SRC data in the Poisson model.

3.3.2 Testing for dispersion

The other method to test over-dispersion or under-dispersion is by uisng the *dispersiontest* function in *AER* package by Cameron Trivedi [4]. It follows a simple idea: In a Poisson model, the mean is $E(Y) = \mu$ and the variance is $Var(Y) = \mu$ as well. The test simply tests the null hypothesis $Var(Y) = \mu$ against $Var(Y) = \mu + c * f(\mu)$ as an alternative where the constant $c < 0$ means under-dispersion, $c > 0$ means over-dispersion, and the function $f(\mu)$ is some mononton function (often linear or quadratic; the former is the default). The

resulting test is equivalent to testing $H_0 : c = 0$ vs $H_1 : c \neq 0$ and the test statistic used is asymptotically standard normal under the null.

The *dispersiontest* function in R assesses the hypothesis that the equidispersion assumption holds against the alternative that the variance is of the form:

$$VAR[y] = \mu + \alpha * trafo(\mu), \tag{3.3}$$

where $trafo$ is a specification of the transformation function such as $trafo(\mu) = \mu^2$ which corresponds to a negative binomial (NB) model with quadratic variance function (called NB2 [5]) or $trafo(\mu) = \mu$ which corresponds to a NB model with linear variance function (called NB1 [5]) or Quasi-Poisson model with dispersion parameter, i.e.,

$$VAR[y] = (1 + \alpha)\mu = dispersion * \mu. \tag{3.4}$$

By default, for $trafo = NULL$, the latter dispersion formulation is used in *dispersiontest*. Otherwise, if $trafo$ is specified, the test is formulated in terms of the parameter α. The transformation $trafo$ can either be specified as a function or an integer corresponding to the function x^{trafo}, such that $trafo = 1$ and $trafo = 2$ yield the linear and quadratic formulations respectively.

We used this test on SRC data, and the α estimate is -0.232 when $trafo = 1$. This indicates there is evidence of under-dispersion.

3.3.3 Negative binomial (NB) model

Another way of modeling over-dispersed count data is to assume a negative binomial (NB) distribution for $y_{ij}|x_{ij}$ which arise as a gamma mixture of Poisson distributions. Its probability density function is

$$f(y; \mu, \theta) = \frac{\Gamma(y + \theta)}{\Gamma(\theta)y!} \frac{\mu^y \theta^\theta}{(\mu + \theta)^{y+\theta}} \tag{3.5}$$

with mean μ and shape parameter $\theta, \Gamma()$ is the gamma function. It also has $\phi = 1$ but with variance function $V() = \mu + \frac{\mu^2}{\theta}$.

If θ is not known but to be estimated from the data, the negative binomial model is not a special case of the general GLM–however, an ML fit can easily be computed re-using GLM methodology by iterating estimation of β given θ and vice versa. This leads to ML estimates for both β and θ which are computed using the function *glm.nb* from the package $MASS$. It returns a model of class *negbin* inheriting from *glm* for which appropriate methods to the generic functions described above are again available.

Because we already showed the Poisson model is with under-dispersion, we don't use negative binomial model to fit the data here. We used Hurdle-negative-binomial model and Zero-inflated-negative-binomial model to fit the SRC data in section 3.4.

3.4 Models dealing with zero-inflation

3.4.1 Hurdle model

Besides over-dispersion and under-dispersion, many count data exhibit more zero observations than would be allowed for by the Poisson model. One model class capable of capturing both properties is the hurdle model, originally proposed by Mullahy in 1986 [12] in the econometrics literature. A review article can be refered to Cameron and Trivedi [4, 5]. They are two-component models: A truncated count component, such as Poisson, geometric or negative binomial, is employed for positive counts, and a hurdle component models zero vs. larger counts. For the latter, either a binomial model or a censored count distribution can be employed.

The hurdle model combines a count data model $f_{count}(y; x, \beta)$ (left truncated at $y = 1$)

and a zero hurdle model $f_{zero}(y; z; \gamma)$ (right censored at $y = 1$):

$$f_{hurdle}(y; x, z, \beta, \gamma) = \begin{cases} f_{zero}(0; z, \gamma) & \text{if } y = 0 \\ (1 - f_{zero}(0; z, \gamma)) f_{count}(y; x, \beta)/(1 - f_{count}(0; x, \beta)) & \text{if } y > 0, \end{cases} \tag{3.6}$$

where y is the value of the dependent variable, z is a vector denoting the predictor variable in the zero hurdle model, x represents a vector denoting the predictor variable in the count data model, γ is a vector of coefficients related to z, and β denotes a vector of coefficients related to x. f_{zero} is a probability density function of $y = 0$, which is typically modeled with binary logistic regression, where all counts greater than 0 are given a value of 1 and otherwise 0. In the zero part using the SRC data, the probability of zeros are estimated probability of non-responses made by customers. The lower part in equation (3.6) f_{count} is modeled with a left-truncated ($y > 0$) count model.

We now use the hurdle model to fit SRC data, but it is now truncated for responses less than 1 and has an additional hurdle component modeling zero versus count observations. The zero hurdle part targets the odds of non-responses made by customers. The count part models the number of responses made by customers. In R, hurdle count data models can be fitted with the *hurdle* function from the *pscl* package. Both its fitting function and the returned model objects of class "hurdle" are modeled after the corresponding GLM functionality in R. The hurdle model results are given in Table 3.5.

Table 3.5: Estimated coefficients, standard errors and Z value for hurdle model

	Count model			Zero hurdle model		
Regressor	Estimate	Std Error	Z value	Estimate	Std Error	Z value
revenue	-0.004	0.004	-0.945	-0.012	0.024	-0.494
purchasedates	0.080 ***	0.008	10.397	0.484 ***	0.032	15.225
purchaseskus	0.006	0.005	1.115	0.205 ***	0.037	5.571
carts	-0.011 **	0.004	-2.650	0.007	0.031	0.221
productviews	-0.014 **	0.005	-2.942	0.169 ***	0.027	6.130
browsedates	0.077 ***	0.008	9.958	0.196 ***	0.033	5.969
browseskus	0.012	0.008	1.503	-0.227 ***	0.051	-4.445
coupons	-0.024 *	0.012	-1.958	-0.151 ***	0.018	-8.285
companysizeB	-0.003	0.034	-0.074	-0.035	0.058	-0.606
companysizeC	0.029	0.035	0.840	-0.009	0.060	-0.157
companysizeD	0.065 .	0.036	1.787	-0.020	0.068	-0.294
companysizeE	0.099 *	0.042	2.351	0.038	0.095	0.398
companysizeF	0.054	0.048	1.132	0.083	0.120	0.693
smallbusiness	0.006	0.017	0.325	-0.043	0.034	-1.239

Note:*** P-value<0.001, ** P-value<0.01, * P-value<0.05, . P-value<0.1.

In Table 3.5, we get outputs from two different models. The first section of the output is for the positive-count process. The second section is for the zero-count process. Following the result of zero hurdle model, we can see the P-value for purchasedates, purchaseskus, productviews and browsedates, browseskus and coupons are all statistically significant at 0.05 significance level. On the other hand revenue and carts are not statistically significant at 0.05 significance level. This indicates the relationships between the odds of non-responses made by customers and purchasedates, purchaseskus, productviews, browsedates are positive. For example, with one additional date the customer made a purchase, the log-odds of non-

responses made by customers increases 0.484; For one additional product SKUs which the customer purchased, the log-odds of non-reponses made by customers increases 0.205. The relationship between the odds of non-responses made by customers and predictors browsekus and coupons are negative. For example, with one additional using coupons' percentage of customer's transactions, the log-odds of non-responded by customers decreases 0.151. From the output in positive count model, we can see the P-value for purchasedates, carts, productviews and browsedates and coupons are all statistically significant at 0.05 significance level. On the other hand, revenue, purchaseskus and browseskus are not statistically significant at 0.05 significance level. This means the relationships between the number of responses made by customers and purchasedates and browsedates are positive. For example, with one additional date the customer made a purchase in P1 through P6, the log number of responses made by customers increases 0.08 among those who have positive counts. The relationships between the number of responses made by customers and carts, productviews and coupons are negative. For example, with one additional using coupons' percentage of customer's transactions, the log number of responses made by customers decreases 0.024 among those who have positive counts.

3.4.2 Hurdle-Negative-Binomial model

Hurdle models are two-component models with a truncated count component for positive counts and a hurdle component that models the zero counts. The count model is typically a truncated Poisson or negative binomial regression. In this part, we fit the count model of hurdle regression with negative binomial. We get the estimated regression coefficients listed in Table 3.6.

Table 3.6: Estimated coefficients, standard errors and Z value for hurdle-negative binomial

	Count model			Zero hurdle model		
Regressor	Estimate	Std Error	Z value	Estimate	Std Error	Z value
revenue	-0.004	0.004	-0.945	-0.012	0.024	-0.494
purchasedates	0.080 ***	0.008	10.397	0.484 ***	0.032	15.225
purchaseskus	0.006	0.005	1.115	0.205 ***	0.037	5.571
carts	-0.011 **	0.004	-2.650	0.007	0.031	0.221
productviews	-0.014 **	0.005	-2.942	0.169 ***	0.027	6.130
browsedates	0.077 ***	0.008	9.958	0.196 ***	0.033	5.969
browseskus	0.012	0.008	1.503	-0.227 ***	0.051	-4.445
coupons	-0.024 *	0.012	-1.958	-0.151 ***	0.018	-8.285
companysizeB	-0.003	0.034	-0.074	-0.035	0.058	-0.606
companysizeC	0.029	0.035	0.840	-0.009	0.060	-0.157
companysizeD	0.065 .	0.036	1.787	-0.020	0.068	-0.294
companysizeE	0.099 *	0.042	2.351	0.038	0.095	0.398
companysizeF	0.054	0.048	1.132	0.083	0.120	0.693
smallbusiness	0.006	0.017	0.325	-0.043	0.034	-1.239

Note:*** P-value<0.001, ** P-value<0.01, * P-value<0.05, . P-value<0.1.

Following the result of Zero-hurdle model, we can see the P-values for purchasedates, purchaseskus, productviews and browsedates, browseskus and coupons are all statistically significant at 0.05 significance level. On the other hand, revenue and carts are not statistically significant at 0.05 significance level. This means the relationships between the odds of non-responses made by customers and purchasedates, purchaseskus, productviews, browsedates are positive. And the relationship between the odds of non-responses made by customers and browsekus and coupons are negative. From output in positive count model, we can see the P-values for purchasedates, carts, productviews and browsedates and coupons are all

statistically significant at 0.05 significance level. On the other hand, revenue, purchaseskus and browseskus are not statistically significant at 0.05 significance level. The relationships between the number of responses made by customers and predictors purchasedates and browsedates are positive. And the relationships between the number of responses made by customers and predictors carts, productviews and coupons are negative. Comparing with Hurdle mode in section 3.4.1, the estimated regression coefficients for hurdle-negative-binomial are almost the same.

3.4.3 Zero-Inflated-Poisson regression model

Another model class capable of dealing with excess zero counts is zero-inflated models (ZIP) [12, 10]. They are two-component mixture models combining a point mass at zero with a count distribution such as Poisson, geometric or negative binomial. There are two sources of zeros: zeros may come from both the point mass and from the count component. The zero-inflated density is a mixture of a point mass at zero and a count distribution. The probability of observing a zero count is inflated with probability $\pi = f_{zero}(0; z, \gamma)$:

$$f_{zeroinfl}(y; x, z, \beta, \gamma) = f_{zero}(0; z, \gamma) \cdot I_{\{0\}}(y) + (1 - f_{zero}(0; z, \gamma)) \cdot f_{count}(y; x, \beta), \qquad (3.7)$$

where y is the value of the dependent variable, z is a vector denoting the predictor variable in the zero part, x is a vector denoting the predictor variable in the count part, γ is a vector of coefficients related to z, and β denotes a vector of coefficients related to x. f_{zero} is a probability density function for excess zeros, and f_{count} is a probability density function for Poisson count. $I()$ is the indicator function and the unobserved probability π of belonging to the point mass component is modeled by a binomial GLM $\pi = g^{-1}(z^T\gamma)$. The corresponding regression equation for the mean is

$$\mu_i = \pi_i \cdot 0 + (1 - \pi_i) \exp(x_i^T \beta), \qquad (3.8)$$

using the canonical log link. The vector of regressors in the zero-infation model z_i and the regressors in the count component x_i need not to be distinct. The default link function $g(\pi)$ in binomial GLMs is the logit link [1]. The parameters of β, γ and potentially the dispersion parameter θ (if a negative binomial count model is used) can be estimated by ML.

Unlike in the hurdle model where all zeros are modeled in f_{zero}, in the zero-inflated model only excess zeros are estimated in f_{zero} part. More specifically, individuals with $y = 0$ can be part of two groups: one group (excess zero) is not part of count process, and one group belongs to the count part. The right part in equation (3.7), f_{count} is typically modeled with Poisson model. Comparing with hurdle model, the only difference is in the way that the zero values are modeled.

Table 3.7: Estimated coefficients, standard errors and Z value for ZIP

	Count model			Zero-inflation model		
Regressor	Estimate	Std Error	Z value	Estimate	Std Error	Z value
revenue	-0.010 *	0.004	-2.230	0.157	0.192	0.820
purchasedates	0.112 ***	0.007	16.662	-2.156 ***	0.310	-6.952
purchaseskus	0.008 .	0.004	1.808	-0.512	0.328	-1.562
carts	-0.016 ***	0.004	-4.067	0.660 **	0.217	3.040
productviews	-0.008 .	0.004	-1.946	-1.194 **	0.415	-2.875
browsedates	0.068 ***	0.007	9.855	-0.967 **	0.317	-3.046
browseskus	0.002	0.007	0.321	0.113	0.404	0.281
coupons	-0.048 ***	0.010	-4,914	-0.380 ***	0.095	3.999
companysizeB	-0.005	0.027	-0.196	-0.166	0.312	0.530
companysizeC	0.027	0.028	0.962	0.295	0.327	0.904
companysizeD	0.051 .	0.029	1.728	0.369	0.379	0.974
companysizeE	0.076 *	0.035	2.166	-0.764	0.985	-0.776
companysizeF	0.041	0.040	1.005	-0.736	1.017	-0.724
smallbusiness	-0.006	0.014	-0.419	-0.053	0.212	-0.251

Note:*** P-value<0.001, ** P-value<0.01, * P-value<0.05, . P-value<0.1.

In R, zero-inflated count data models can be fitted using the $zeroinfl$ function from the *pscl* package. Both the fitting function interface and the returned model objects of class "zeroinfl" are almost identical to the corresponding *hurdle* functionality. We fit the ZIP model using the SRC data and the output are given in Table 3.7.

This output is also a two-part model. However, both parts predict zero counts. The count model predicts some zero counts, and on the top of that the zero-inflation binary model part adds zero counts, thus, the name zero-"inflation".

From output in zero-inflation model, we can see the P-values for purchasedates, carts, productviews, browsedates and coupons are all statistically significant at 0.05 significance level. On the other hand, revenue, purchaseskus and browseskus are not statistically significant at 0.05 significance level. We see that the relationship between the odds of non-responses made by customers and predictor carts is positive. For example, with one additional product which the customers added to their online cart in the last six months, the log-odds of non-responses made by customers increases 0.660. And the relationship between the odds of non-responses made by customers and purchasedates, productviews, browsedates and coupons are negative. For example, with one additional unique date the customers browsed on website in the last six months, the log-odds of non-responses made by customers decreases 0.967. For one additional using coupons' percentage of customer's transactions , the log-odds of non-responses made by customers decreases 0.380. From the output of count model, we can see the P-values for revenue, purchasedates, carts, browsedates and coupons are all statistically significant at 0.05 significance level. On the other hand, purchaseskus , productviews and browseskus are not statistically significant at 0.05 significance level. We can see that the relationships between the number of responses made by customers and predictors purchasedates and browsedates are positive. For example, with one additional date the customer made a purchase, the log number of responses made by customers increases 0.112. And we can see that the relationships between the number of responses made by customers and predictors revenue, carts and coupons are negative. For example, with one

additional using coupons' percentage of customer's transactions, the log number of responses made by customers decreases 0.048.

3.4.4 Zero-Inflated-Negative-Binomial model (ZINB)

A Poisson distribution assumes that the variance of the outcome variable equals its mean. When over-dispersion or under-dispersion also comes from the non-zero part of the outcome, the ZIP model can be extended to ZINB model to deal with zero-inflation and over-dispersion (under-dispersion) at the same time. In this section, we fit the SRC data using the ZINB model with the result given in Table 3.8.

From the output shown in Table 3.8, similar conclusions were obtained as those in the ZIP model in section 3.4.3.

Table 3.8: Estimated coefficients, standard errors and Z value for ZINB

	Count model			Zero-inflation model		
Regressor	Estimate	Std Error	Z value	Estimate	Std Error	Z value
revenue	-0.010 *	0.004	-2.230	0.157	0.192	0.820
purchasedates	0.112 ***	0.007	16.662	-2.156 ***	0.310	-6.952
purchaseskus	0.008 .	0.004	1.808	-0.512	0.328	-1.562
carts	-0.016 ***	0.004	-4.067	0.660 **	0.217	3.040
productviews	-0.008 .	0.004	-1.946	-1.194 **	0.415	-2.875
browsedates	0.068 ***	0.007	9.855	-0.967 **	0.317	-3.046
browseskus	0.002	0.007	0.321	0.113	0.404	0.281
coupons	-0.048 ***	0.010	-4,914	-0.380 ***	0.095	3.999
companysizeB	-0.005	0.027	-0.196	-0.166	0.312	0.530
companysizeC	0.027	0.028	0.962	0.295	0.327	0.904
companysizeD	0.051 .	0.029	1.728	0.369	0.379	0.974
companysizeE	0.076 *	0.035	2.166	-0.764	0.985	-0.776
companysizeF	0.041	0.040	1.005	-0.736	1.017	-0.724
smallbusiness	-0.006	0.014	-0.419	-0.053	0.212	-0.251

Note:*** P-value<0.001, ** P-value<0.01, * P-value<0.05, . P-value<0.1.

Chapter 4

Model selection

4.1 Using AIC criterion

The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. AIC is given by: $AIC = -2 \log L(\theta) + 2k$, where $L(\theta)$ is the maximized likelihood function for the estimated model and k is the number of free parameters in the model. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value [2]. AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, because increasing the number of parameters in the model almost always improves the goodness of the fit. In this section, we will use the AIC criterion to compare all the fitted models to select the best model for the SRC data. AIC and the result of coefficient is given in Table 4.1.

Table 4.1: Coefficient and AIC for the count and zero Models (Logistic regression has binary responses, it doesn't have zero model part in the table. Poisson and Poisson with an offset models do not consider zero-inflation, thus have no zero model part in the table either)

Count model	Poisson-offset	Logistic	Poisson	Hurdle	Hurdle-NB	ZIP	ZINB
revenue	-0.014 **	0.004	-0.023 ***	-0.004	-0.004	-0.010 *	-0.010 *
purchasedates	0.123 ***	0.439 ***	0.121 ***	0.080 ***	0.080 ***	0.112 ***	0.112***
purchaseskus	0.011 *	0.144 ***	0.015 ***	0.006	0.006	0.008.	0.008.
carts	-0.021 ***	0.019	-0.021 ***	-0.011 **	- 0.011 **	-0.016 ***	-0.016 ***
productviews	-0.009 *	0.025 *	-0.012 *	-0.014 **	-0.014 **	-0.008.	-0.008.
browsedates	0.079 ***	0.150 ***	0.073 ***	0.077 ***	0.077 ***	0.068 ***	0.068***
browseskus	0.004	-0.116 ***	-0.039 ***	0.012	0.012	0.002	0.002
coupons	-0.065	-0.120 ***	-0.054 ***	-0.024 *	-0.024 *	-0.048 ***	-0.048 ***
companyB	-0.0002	-0.017	-0.009	-0.003	-0.003	-0.005	-0.005
companyC	0.031	0.014	0.020	0.029	0.029	0.027	0.027
companyD	0.054.	0.058	0.053.	0.065.	0.065.	0.051.	0.051.
companyE	0.090 **	0.137 *	0.101 **	0.099 *	0.099 *	0.076 *	0.075 *
companyF	0.048	0.077	0.081 *	0.054	0.054	0.041	0.040
smallY	-0.006	-0.039	0.008	0.006	0.006	-0.006	-0.006
Zero model	Poisson-offset	Logistic	Poisson	Hurdle	Hurdle-NB	ZIP	ZINB
revenue				-0.012	-0.012	0.157	0.157
purchasedates				0.484 ***	0.484 ***	-2.156 ***	-2.156 ***
purchaseskus				0.205 ***	0.205 ***	-0.512	-0.512
carts				0.007	0.007	0.660 **	0.660 **
productviews				0.169 ***	0.169 ***	-1.194 **	-1.195 **
browsedates				0.196 ***	0.196 ***	-0.967 **	0.968 **
browseskus				-0.227 ***	-0.227 ***	0.113	0.113
coupons				-0.151 ***	-0.151 ***	0.380 ***	0.380 ***
companyB				-0.035	-0.035	0.166	0.168
companyC				-0.009	-0.009	0.295	0.297
companyD				-0.020	-0.020	0.369	0.370
companyE				0.038	0.038	-0.764	-0.764
companyF				0.083	0.083	-0.736	-0.740
smallY				-0.047	-0.047	-0.053	-0.053
AIC	64978	59259	66796	64651	64653	64643	64645

Note:*** P-value<0.001, ** P-value<0.01, * P-value<0.05, . P-value<0.1.

Six models described in section 3 were used to fit the data. AIC values for all the model are presented in Table 4.1. The Poisson regression model had the largest AIC value,

demonstrating the worst fit to the data. For the other four models, the Hurdle-NB and ZINB models had larger AIC values comparing with Hurdle and ZIP models. Among all the poisson models the Zero-inflated Poisson Model had the smallest AIC value, so ZIP is the best choice for the SRC data. Comparing with ZIP model, AIC of logistic regression is smaller, we can say that logistic regression is the best model for SRC data in this study. Because the response variable of logistic regression is binomial, it means that we can choose the logistic regression if we want to know whether the customers would make purchases based on the information given in P1 through P6. If we want to know how many responses made by the customers (count data), the ZIP model is the best choice among all the models we tried in dealing with count in the study.

4.2 Using Vuong Test

The Vuong non-nested test [15] is based on a comparison of the predicted probabilities of two models that do not nest. Examples include comparisons of zero-inflated count models with their non-zero-inflated analogs (e.g., zero-inflated Poisson versus ordinary Poisson, or zero-inflated negative-binomial versus ordinary negative-binomial). A large, positive test statistic provides evidence of the superiority of model 1 over model 2, while a large, negative test statistic is evidence of the superiority of model 2 over model 1. Under the null that the models are indistinguishable, the test statistic is asymptotically standard normal.

Let p_1 be the predicted probabilities from model 1, evaluated conditional on the estimated MLEs. Let p_2 be the corresponding probabilities from model 2. Then the Vuong statistic is $\frac{\sqrt{N}m}{S_m}$ where $m = log(p_1) - log(p_2)$ and s_m is the sample standard deviation of m, N is sample size. We compare all the poisson models using Vuong test in SRC data, the comparison result is given in table 4.2.

Table 4.2: Vuong non-nested tests results for the count data. Poisson = Poisson regression model, ZIP = zero-inflated Poisson model, ZINB = zero-inflated negative binomial model, Hurdle = Hurdle model, Hurdle-NB = Hurdle negative binomial model.

Model Comparison	Vuong Test Statistic	P-value	Preferable Model
ZIP vs. Poisson	8.462	<.0001	ZIP
Hurdle vs. Poisson	4.656	<.0001	Hurdle
ZINB vs. Poisson	8.461	<.0001	ZINB
Hurdle-NB vs. Poisson	4.656	<.0001	Hurdle-NB
ZINB vs. ZIP	-0.223	0.412	ZIP
ZIP vs. Hurdle-NB	0.122	0.452	ZIP
Hurdle vs. Hurdle-NB	0.649	0.258	Hurdle
ZINB vs. Hurdle-NB	0.122	0.452	ZINB
ZINB vs. Hurdle	0.122	0.452	ZINB
ZIP vs. Hurdle	0.122	0.452	ZIP

From the results given in Table 4.2, the comparison between the ZIP model and Poisson model had a Vuong test statistic of 8.462 with P-value<.0001, indicating the ZIP model was preferred. The similar result was obtained when comparing ZINB model with Poisson model, the Vuong test statistic is 8.461 with P-value<.0001. For comparing Hurdle model with Poisson model, the Vuong test statistic is 4.656 with P-value<.0001. When comparing Hurdle-NB with Poisson model, the Vuong test statistic is 4.656 with P-value<.0001. All the results indicate, the Hurdle model, Hurdle-NB model, ZIP model and ZINB model are all better than Poisson model. Using Vuong test, we didn't find the significant difference between any two models among Hurdle model, Hurdle-NB model, ZIP model and ZINB model in Table 4.2. But we still can see that the most preferable model is the ZIP model, because it has the smallest P-value in Vuong test, and is the best model we can choose.

4.3 Using five-fold Cross-validation (CV)

Suppose we have a model with one or more unknown parameters, and a data set to which the model can be fit (the training data set). The fitting process optimizes the model parameters to make the model fit the training data as well as possible. If we then take an independent sample of validation data from the same population as the training data, it will generally turn out that the model does not fit the validation data as well as it fits the training data. The size of this difference is likely to be large especially when the size of the training data set is small, or when the number of parameters in the model is large. Cross-validation [9] is a way to estimate the size of this effect.

We will randomly divide the set of observations into five approximately equal size groups or folds. The first fold is treated as a validation set, and the method is fit on the remaining four folds. The mean squared error, MSE_1, is then computed on the observations in the held-out fold. This procedure is repeated five times; each time, a different group of observations is treated as a validation set [6]. This process results in five estimates of the test error, $MSE_1, MSE_2, \ldots, MSE_5$. The five-fold CV estimate is computed by averaging these values, $CV_{(5)} = \frac{1}{5}\sum_{i=1}^{5} MSE_i$.

Five fold cross-validation procedure was applied to every poisson model in Chapter 3 to obtain the Cross-Validation *errors* as below.

Table 4.3: Five-fold Cross-Validation MSE results

Model	Average MSE
Poisson	1.461
Hurdle	0.847
Hurdle-NB	0.909
ZIP	0.880
ZINB	0.885

After we obtained the five-fold cross-validation MSE for the five models in Table 4.3, we order the models by increasing MSE order: Hurdle, ZIP, ZINB, Hurdle-NB, Poisson. The smallest value of MSE was obtained for the Hurdle model. Based on the Cross-Validation criterion, the Hurdle model is the best choice.

4.4 Summary of Model selection

Model selection is a process of seeking the model in a set of candidate models that gives the best balance between model fit and complexity [3]. In this paper, we use three difference methods to select the best model for the count data. Based on AIC criterion, we found the best model is ZIP model. Using the Vuong Test method to compare five models, we found ZIP, ZINB, Hurdle and Hurdle-NB are all better than Poisson model, and the best model is the ZIP model because the P-value is the smallest when comparing with Poisson model. This result is the same as using the AIC method . When using five-fold Cross-Validation to estimate MSE.CV for each model, we found that Hurdle model has the smallest cross validation error, so Hurdle model is the best choice in this case.

Chapter 5

Conclusion

In this thesis, the Poisson model, logistic regression model, hurdle model, zero-inflated-poisson and zero-inflated-negative-binomial model have been applied to model the customers' online purchase behavior which deal with under-dispersion and zero-inflation problems, based on customers' past purchase information in the last six months.

We used the AIC criterion, Vuong test and five-fold cross-validation criterion to select the best model for modeling the count data. The results show that ZIP and Hurdle models have better performance than Poisson model for count data. We also use logistic regression to fit the binary response using the same set of predictors.

Using ZIP model in SRC data, we find that the number of responses made by customers are positively related with purchasedates and browsedates, and negative related with revenue, carts and coupons. If we use Hurdle model, the number of responses made by customers are positively related with purchasedates and browsedates, and negatively related with carts, productviews and coupons. If we want to know whether the customers make a purchase or not, the logistic regression is the best choice. The odds of responses made by customers and purchasedates, purchaseskus, productviews and browsedates are positively correlated. The odds of responses made by customers and browseskus and coupons are negatively correlated.

In this thesis, the models that we built include all the predictors regardless whether

they are significant or not significant. In the future, we may use some variable selection techniques to build the best model including only significant predictors. For example, we can use stepwise selection procedure to eliminate all the insignificant predictors from the current models. We can also try the subset selection procedures to select significant predictors in all the models or use the lasso (least absolute shrinkage and selection operator)-based techniques [16] for variable selection in Poisson model and ZIP models. With the variable selection procedures, not only we can find all the statistical significant predictors to the response variable, we may also remove possible collinearity problem among the predictors.

Bibliography

[1] Achim Zeileis, Christian Kleiber, Simon Jackman. *Regression Models for Count Data in R*, ISSN 1548-7660, 2008.

[2] Bozdogan, H. Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44(1), 62-91, 2000.

[3] Burnham, K.P. and Anderson, D.R. *Model Selection and Inference: A Practical Information-Theoretic Approach.* 2nd Edition, Springer-Verlag, New York, 2002.

[4] Cameron, AC, Trivedi, PK. *Regression Analysis of Count Data.* Cambridge University Press, Cambridge, 1998.

[5] Cameron, AC, Trivedi, PK. *Microeconometrics: Methods and Applications.* Cambridge University Press, Cambridge, 2005.

[6] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with applications in R*, Springer, New York, 2013.

[7] Jackman S (2008). *pscl: Classes and Methods for R Developed in the Political Science Com- putational Laboratory, Stanford University*, Department of Political Science, Stanford University, Stanford, California. R package version 0.95.

[8] Kohavi, Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection.*Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, San Mateo, CA: Morgan Kaufmann. 2 (12): 1137-1143, 1995.

[9] Max Kuhn, Kjell Johnson *Applied Predictive Modeling* 3rd Edition, 2018.

[10] Diane Lambert. Zero-inflated Poisson Regression, with an Application to Defects in-Manufacturing. *Technometrics*, 34, 1-14. 1992.

[11] P. McCullagh, J.A. Nelder. *Generalized Linear Models*. 2nd edition. Chapman & Hall, 1989. London.

[12] Mullahy, J. Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*, 33, 341-365, 1986.

[13] J.A. Nelder, R.W.M. Wedderburn. Generalized Linear Models *Journal of the Royal Statistical Society*. Series A, 135, 370-384. 1972.

[14] R Core Team (2018) *R: A Language and Environment for Statistical Computing.*, R Foundation for Statistical Computing, Vienna. https://www.R-project.org

[15] Vuong, Quang H. Likelihood Ratio Tests for Model Selection and non-nested Hypotheses. *Econometrica*, 57 (2): 307-333, 1989.

[16] H. Zou, The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418-1429, 2006.