The Open University of Hong Kong

School of Science and Technology

BSc (Hons) in Data Science

STAT S461F Data Science Project

Final Report

2020-2021

Project title: Statistical Modelling of Hemophilia Count Data

by

Hung Chun Kwong,

Lai Siu Kwok (Team leader),

Lo Shi Sam,

Pei Zixuan,

Shan Youming,

Zheng Zequan

Supervisor: Dr. Tony Chan

Date: May 2021

Abstract

Hemophilia is one of the most serious diseases in the world, it is difficult to detect in the early stages of this disease. When the patients discover that they have been infected with the Hemophilia disease, it is too late to be cured in most cases or even dead after a while. From statistically, it can be predicted by different factors for early discovery.

However, the existing literature is often mixed with medical research or social research mathematics, and rarely only statistical modelling analysis. Thus, those reasons give us an opportunity for this time of studies.

This project is a study of a sample dataset of 500 groups of patients with hemophilia which focuses on the number of deaths from each group and the other four factors that may affect the risk of death, including HIV status, clotting medicine dose, age group, the time of participation in diseases. We aim to investigate what are the factors that affect the number of deaths of patients with hemophilia modelling, between the death number of patients and affected factors.

Accordingly, each of the 500 groups is an independent count data, “death” would be the dependent variable, and the others belong to independent variables. The major methodologies would be i) using the appropriate regression model for the dataset and ii) discovering the relationship between the number of deaths and the factors. The results could i) determine the significance of which factors from a set of which factors and ii) determine and appropriate the statistical model to model the dataset. Finally, the limitations and future research directions will be contoured.

Contents

1 Introduction 5

1.1 Research Background ····· 5
1.2 Research Objectives ····· 6
1.3 The Dataset ····· 7

2 Multiple Linear Regression Models 8

2.1 Multiple Linear Regression Models ····· 8
2.2 Model Building and Parameter Estimation ····· 9
2.2.1 Statistical Analysis ····· 12
2.3 Diagnostic Checking and Model Selection ····· 15
2.4 Concluding Remarks ····· 23

3 Poisson Log-linear Regression Models 24

3.1 Literature Review ····· 24
3.2 Poisson Log-linear Regression Models ····· 25
3.3 Model Building and Parameter Estimation ····· 26
3.3.1 Statistical Analysis ····· 28
3.4 Diagnostic Checking and Model Selection ····· 29
3.5 Concluding Remarks ····· 33

4 Negative Binomial Regression Models 34

4.1 Literature Review ····· 34
4.2 Negative Binomial Regression Models ····· 34
4.3 Model Building and Parameter Estimation ····· 35
4.3.1 Statistical Analysis ····· 37
4.4 Diagnostic Checking and Model Selection ····· 39
4.5 Concluding Remarks ····· 41

5 Zero-inflated Poisson Regression Models 44

5.1 Literature Review ····· 45
5.2 Zero-inflated Poisson Regression Models ····· 46
5.3 Model Building and Parameter Estimation ····· 47
5.3.1 Statistical Analysis ····· 49
5.4 Diagnostic Checking and Model Selection ····· 51
5.5 Concluding Remarks ····· 52

6 Zero-inflated Negative Binomial Regression Models 54

6.1 Literature Review ····· 54
6.2 Zero-inflated Negative Binomial Regression Models ····· 54
6.3 Model Building and Parameter Estimation ····· 56
6.3.1 Statistical Analysis ····· 58
6.4 Model Selection ····· 60
6.5 Concluding Remarks····· 61

7 Model Evaluation 62

7.1 Model Comparison ····· 62

8 Discussion 67

8.1 Research Questions Revisited ····· 67

9 Conclusion 68

9.1 Overall Summary ····· 68
9.2 Limitations and Future Directions ····· 68

Appendix 69

References 76

1. Introduction

1.1 Research Background

Hemophilia can be serious, it makes the patient's blood unable to clot like normal people, once the patient gets injured, the patient may cause serious consequences and even death. (“What is Hemophilia”, 2020). Every year, there are about three hundred thousand people with hemophilia in the world (Stonebraker et al., 2020).

An article written by Ferreira et al. (2013) studied the health-related quality of life in hemophilia from the Brazilian blood center. The researchers used a statistical analysis with Spearman correlation analyses which is conducted in 60 male’s cases, which is reflected in the health-related quality and has a correlation to hemophilia.

A research paper from Payal et al. (2016) studied a clinical profile of hemophilia patients in Jodhpur. A total of 51 child cases were conducted with the cross-sectional, results from post-traumatic bleeding and gum bleeding are the main features of these hemophilia patients.

Although there have been many studies about hemophilia in the past just like the above two article reviews, we found that most of these studies are complex and extensive such as mathematics mixed medical or social studies. Those studies rarely analyze from statistical modelling only.

Statistical modelling, which is a series of processes by applying statistical analysis in order to infer any hidden relationships between variables. A statistical model is a mathematical product from the dataset, including expression and graphical solution. (Stobierski et al. 2020). Statistical modelling is complex, the more factors, the higher the degree of difficulty will be.

In this project, we would like to decrease the scope of research such as the factors and apply different statistical modelling methods for studying hemophilia dataset to find the relationship among specific factors. Hopefully, we could find out the cause of death from hemophilia and build a model to reduce the risk of patient death.

1.2 Research Objectives and Questions

Based on the dataset, we have obtained the death number of patients caused by hemophilia, the dosage of coagulant used by patients, whether they have HIV, age and study time. We would like to explore the relationship between these factors and death so that effective preventive measures can be taken to reduce the death of patients. The occurrence and control measures provide a scientific reference for better promoting the recovery of patients and helping medical staff to judge hemophilia.

In view of this, the research objectives should be:

- Determine the significant which factors from a set of which factors
- Determine and appropriate the statistical model to model the dataset

Therefore, we can conclude that our research questions should be:

- How do we choose the appropriate regression model for the dataset?
- What is the relationship between the number of deaths and the factors?

1.3 Dataset

Variables Description

A hemophilia study was conducted with 500 groups of patients with hemophilia. The study focuses on the number of deaths from each group and four factors which may affect the risk of death, including HIV status, clotting medicine dose, age group and time of participation. This project is to study what are the factors that affect the number of deaths of patients with hemophilia. Therefore, “death” would be the dependent variable, and the others belong to independent variables. We would introduce the description of five attributes which respectively are “death”, “hiv”, “factor”, “age” and "p_year" in this hemophilia dataset in **TABLE 1.1**.

TABLE 1.1

	Variable name	Property	Description
Dependent	death	Numeric	The number of deaths in that particular age group
Independent	hiv	Category	HIV status of ALL Hemophilia patients in that particular age group. 1 = negative 2 = positive
	factor	Category	The dose of blood clotting preparation (1-5). 1 = high 2 = moderate 3 = low 4 = unknown 5 = none
	age	Category	The range of age groups (1 - 14). 1 = 0 - 4 2 = 5 - 9 . . . 13 = 60 - 64 14 = 65 or above (Range of each age group is equal)
	py	Numeric	The total number of years of ALL patients from a particular age group in a particular year participated in the study.

2. Multiple Linear Regression Models

Multiple linear regression models are used when the number of independent variables is more than one. As the model has lots of independent variables, the multiple linear regression model is using the matrix form into the model to perform the model analysis more efficiently. Multiple linear regression models have to fulfil a few assumptions. Firstly, a linear relationship exists between the dependent variable and the independent variables. Secondly, the dependent variable should follow the normal distribution the same as the residual also should follow the normal distribution. Thirdly, there should be no relationship and no high correlation between each of the independent variables. Finally, the residuals should have a constant variance.

2.1 Multiple Linear Regression Models

A multiple linear regression model is in the following matrix form:

$$Y = X\beta + \varepsilon \tag{2.1}$$

where, Y is the dependent variable vector, $dim(Y) = n \times 1$;

X is the design matrix, $dim(X) = n \times k$;

β is the parameter vector, $dim(\beta) = k \times 1$;

ε is the random error vector, $dim(\varepsilon) = n \times 1$.

or,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \tag{2.2}$$

where, $i = 1, 2, \cdots, n$

Y_i is the dependent variable;

$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ are (p+1) unknown regression coefficients;

$X_{i1}, X_{i2}, \cdots, X_{ip}$ are p independent variables;

ε_i is the unknown random error term.

2.2 Model Building and Parameter Estimation

Based on previous findings, the dependent variable “deaths” is a count data, then the multiple linear regression model may not be used. However, we will still employ the multiple linear regression model in order to see whether using multiple linear regression can fit the dataset as well. Moreover, if the multiple linear regression model is not fit, we may use another model to deal with count data.

A multiple linear regression model of the Hemophilia dataset is in the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \qquad (2.3)$$

A multiple linear regression model often uses quantitative variables as independent variables. However, it may include some qualitative variables in the models, such as categorical variables, we use dummy variables to identify them by using the values 0 and 1. By using dummy variables in the regression model, we can use one regression equation to represent different subgroups of categorical variables. If the variable includes k categorical, we will set k-1 dummy variables. It is because if k-1 dummy variables are equal to zero, then the last category will be employed. The last category which is omitted, it is called the "reference category". The number of dummy variables cannot be introduced up to k, otherwise, multicollinearity will occur.

In this data set, the three independent variables "hiv", "factor" and "age" are qualitative variables, we need to convert these three variables as dummy variables.

The original multiple linear regression model without dummy variables from formula (2.3):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

For the independent variable X_1 for “hiv” which has two categories, we only set one dummy variable as X_1. For the reference category will be employed when X_1 equals to zero.

$$X_1 = 1, if\ "hiv"\ =\ negative;\ X_1 = 0, otherwise$$

For the independent variable X_2 for “factor” which has 5 categories, we only set 4 dummy variables as $X_{2,1},\ X_{2,2}, X_{2,3}, X_{2,4}$. For the reference category will be employed when all dummy variables equal to zero.

$$X_{2,1} = 1, if\ "factor"\ =\ high;\ X_{2,1} = 0, otherwise$$
$$X_{2,2} = 1, if\ "factor"\ =\ moderate;\ X_{2,2} = 0, otherwise$$
$$X_{2,3} = 1, if\ "factor"\ =\ low;\ X_{2,3} = 0, otherwise$$
$$X_{2,4} = 1, if\ "factor"\ =\ unknown;\ X_{2,4} = 0, otherwise$$

For the independent variable X_3 for "age" which has 14 categories, we only set 13 dummy variables as $X_{3,1}, X_{3,2}, X_{3,3}, X_{3,4}, X_{3,5}, X_{3,6}, X_{3,7}, X_{3,8}, X_{3,9}, X_{3,10}, X_{3,11}, X_{3,12}, X_{3,13}$. For the reference category will be employed when all dummy variables equal to zero.

$$X_{3,1} = 1, if\ "age"\ =\ "0-4";\ X_{3,1} = 0, otherwise$$
$$X_{3,2} = 1, if\ "age"\ =\ "5-9";\ X_{3,2} = 0, otherwise$$
$$X_{3,3} = 1, if\ "age"\ =\ "10-14";\ X_{3,3} = 0, otherwise$$
$$X_{3,4} = 1, if\ "age"\ =\ "15-19";\ X_{3,4} = 0, otherwise$$
$$X_{3,5} = 1, if\ "age"\ =\ "20-24";\ X_{3,5} = 0, otherwise$$
$$X_{3,6} = 1, if\ "age"\ =\ "25-29";\ X_{3,6} = 0, otherwise$$
$$X_{3,7} = 1, if\ "age"\ =\ "30-34";\ X_{3,7} = 0, otherwise$$
$$X_{3,8} = 1, if\ "age"\ =\ "35-39";\ X_{3,8} = 0, otherwise$$
$$X_{3,9} = 1, if\ "age"\ =\ "40-44";\ X_{3,9} = 0, otherwise$$
$$X_{3,10} = 1, if\ "age"\ =\ "45-49";\ X_{3,10} = 0, otherwise$$
$$X_{3,11} = 1, if\ "age"\ =\ "50-54";\ X_{3,11} = 0, otherwise$$
$$X_{3,12} = 1, if\ "age"\ =\ "55-59";\ X_{3,12} = 0, otherwise$$
$$X_{3,13} = 1, if\ "age"\ =\ "60-64";\ X_{3,13} = 0, otherwise$$

After employing dummy variables into the original multiple linear regression model, we have a model including dummy variables:

$$Y = \beta_0 + (dummy\ variable\ "hiv")\ +\ (dummy\ variables\ "factor") + (dummy\ variables\ "age") + \beta_4 X_4 + \varepsilon$$

where, $dummy\ variable\ "hiv" = \beta_1 X_1$;

$dummy\ variables\ "factor" = \beta_{2,1} X_{2,1} + \beta_{2,2} X_{2,2} + \beta_{2,3} X_{2,3} + \beta_{2,4} X_{2,4}$;

$dummy\ variables\ "age" = \beta_{3,1} X_{3,1} + \beta_{3,2} X_{3,2} + \cdots + \beta_{3,13} X_{3,13}$.

TABLE 2.1

Independent variables							Dummy variable (factor)			
obs	hiv	factor	age	pv	deaths		factor1	factor2	factor3	factor4
1	1	1	1	1	0		1	0	0	0
2	1	2	1	2	0		0	1	0	0
3	1	1	2	3	0		1	0	0	0
4	1	3	2	5	0		0	0	1	0
5	1	1	3	7	0		1	0	0	0
6	1	4	3	9	0		0	0	0	1
7	1	1	4	4	0		1	0	0	0
8	1	3	4	5	0		0	0	1	0
9	1	1	5	8	0		1	0	0	0
10	1	1	5	12	5		1	0	0	0
:	:	:	:	:	:		:	:	:	:
:	:	:	:	:	:		:	:	:	:
449	2	5	14	1	0		0	0	0	0
500	2	5	14	1	5		0	0	0	0

TABLE 2.1 We have 500 observations from the dataset, the table shows part of the dummy variable combinations about “factor”. If “factor” = 1, then “factor” will equal to “factor1”, then “factor2”, “factor3” and “factor4” will equal to zero. If “factor” = 5, like the observations 499 and 500, then “factor1”, “factor2”, “factor3” and “factor4” will all equal to zero.

Method of Least Squares

We will use the method of least squares to estimate the regression parameter, β.

A multiple linear regression model from formula (2.1):

$$Y = X\beta + \varepsilon$$

From the multiple linear regression model:

$$\varepsilon = Y - X\beta$$

The sum of squared error:

$$\varepsilon^T\varepsilon = [\varepsilon_1\ \varepsilon_2 \dots \varepsilon_n][\varepsilon_1\ \varepsilon_2\ \vdots\ \varepsilon_n] = \sum_{i=1}^{n} {\varepsilon_i}^2$$

$$\varepsilon^T\varepsilon = Y^TT - 2\beta^TX^TY + \beta^TX^TX\beta$$

Differentiate the sum of squared error:

$$\frac{\partial(\varepsilon^T\varepsilon)}{\partial\beta}|_{\beta=\hat{\beta}} = -2X^TY + 2X^TX\hat{\beta} = 0$$

$$X^TX\hat{\beta} = X^TY$$

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

The least square estimator for β **is:**

$$LSE(\hat{\beta}) = (X^TX)^{-1}X^TY$$

2.2.1 Statistical Analysis

Selection Criteria

For the statistical analysis, we would use the SAS program to test whether the variables and coefficients are significant. Moreover, we will use the criteria of AIC and BIC.

There are some criteria that evaluate the goodness and compare the different models for finding out which one is the best fit for the data. The first criteria we will use in this project is the Akaike information criterion, also known as AIC, which was introduced by Hirotugu Akaike between 1972 and 1973. The AIC formula is:

$$AIC = 2k - 2ln(\hat{L}) \quad (2.4)$$

where k is the number of parameters in the model and $\hat{L}$ is the largest value of the likelihood function for the model.

The other criteria we will use in this project is the Bayesian information criterion, also known as BIC, which was introduced by Gideon E. Schwarz in 1978. The formula is:

$$BIC = k\ ln(n) - 2ln(\hat{L}) \quad (2.5)$$

where k is the number of parameters in the model, n is the number of observations and $\hat{L}$ is the largest value of the likelihood function for the model.

After calculating the AIC and BIC for the models, we will select the best model that has the minimum values of both AIC and BIC.

For the statistical analysis, we would use the SAS program to test whether the variables and coefficients are significant.

TABLE 2.2

Parameter Estimates				
Variable	**Parameter Estimate**	**Standard Error**	**t Values**	**Pr > |t|**
(Intercept)	0.23043	0.03944	5.84	<.0001
py	0.00203	0.00182	1.11	0.2657
hiv1	-0.44660	0.06440	-6.93	<.0001
factor1	-0.00005864	0.10184	-0.00	0.9995
factor2	-0.13883	0.10328	-1.34	0.1795
factor3	-0.13221	0.10152	-1.30	0.1934
factor4	-0.05148	0.10341	-0.50	0.6188
age1	-0.26912	0.19288	-1.40	0.1636
age2	-0.21156	0.18047	-1.16	0.2484
age3	-0.21156	0.18828	-1.12	0.2617
age4	-0.18329	0.18165	-1.01	0.3135

age5	-0.10022	0.18462	-0.54	0.5875
age6	-0.18307	0.18679	-0.98	0.3275
age7	-0.18122	0.17960	-1.01	0.3135
age8	0.15026	0.18363	0.82	0.4136
age9	-0.16971	0.18141	-0.94	0.3500
age10	-0.26941	0.17769	-1.52	0.1301
age11	-0.10538	0.18942	-0.56	0.5783
age12	-0.15741	0.18362	-0.86	0.3917
age13	-0.27838	0.19165	-1.45	0.1470
Fit Statistics				
AIC	1093.4233			
BIC	1181.9301			

TABLE 2.2 The above table shows the summary by using the multiple linear regression model with dummy variables. Only the p-value for the "hiv1" is smaller than the 5% level of significance. The values of AIC and BIC are seen very high, which are 1093.4233 and 1181.9301 respectively.

TABLE 2.3

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	19	32.00457	1.68445	3.37	<.0001
Error	480	239.74543	0.49947	/	/
Corrected Total	499	271.75000	/	/	/

TABLE 2.3 For the F-test, the p-value is less than 0.0001, which is < 5% significant level. Therefore, at least one coefficient (beta) is not equal to zero, which means at least one coefficient(beta) is significant.

2.3 Diagnostic Checking and Model Selection

Constant Variance of Residual:

FIGURE 2.4

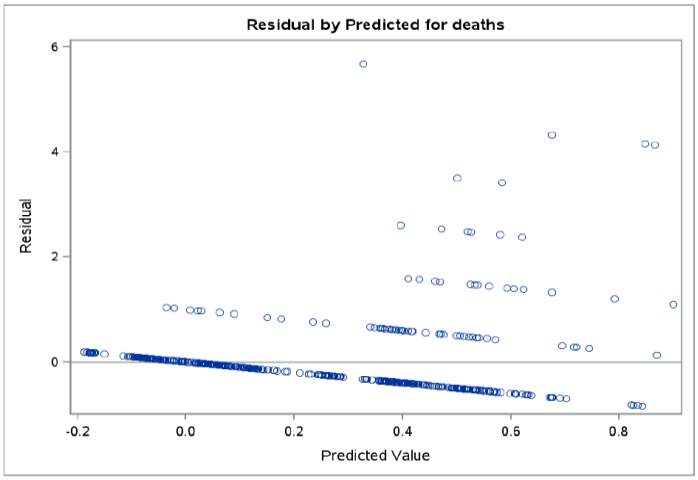

FIGURE 2.4 For the graph above, if the residuals fluctuate in a random fashion inside the band, the variance of residuals is constant. Otherwise, the variance of residuals is non-constant. For the result, there are lots of residuals fluctuating is not in a random fashion inside the band and the variance of residuals is non-constant.

Outliers:

FIGURE 2.5

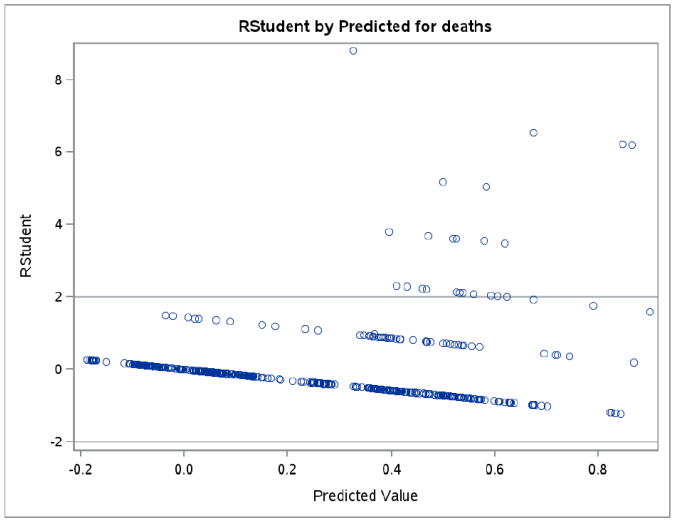

FIGURE 2.5 For the graph above, the points outside the +/- 2 bands (horizontal lines) are outliers. For the result, about 20 points or above are outliers.

Outliers and Leverage:

FIGURE 2.6

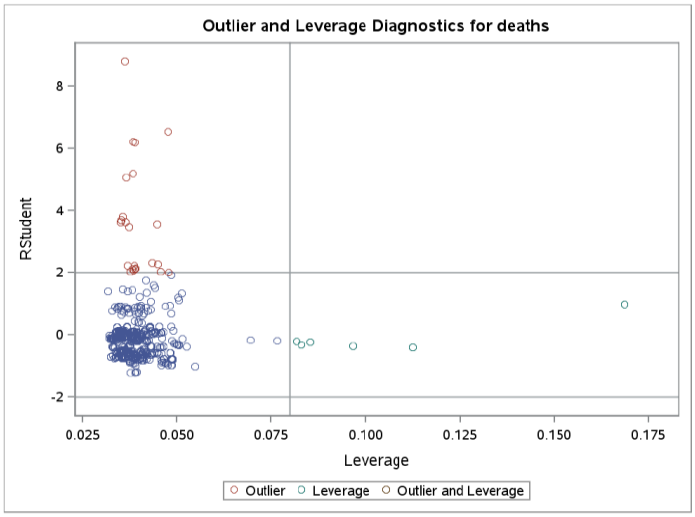

FIGURE 2.6 For the graph above, the points higher than the reference line (vertical line) are leverage points. As a result, the points outside the +/- 2 bands (horizontal lines) are outliers, 6 points are leverage points, about 20 points or above are outliers and no point belongs to outliers and leverage points.

Normal Distribution Checking:

FIGURE 2.7

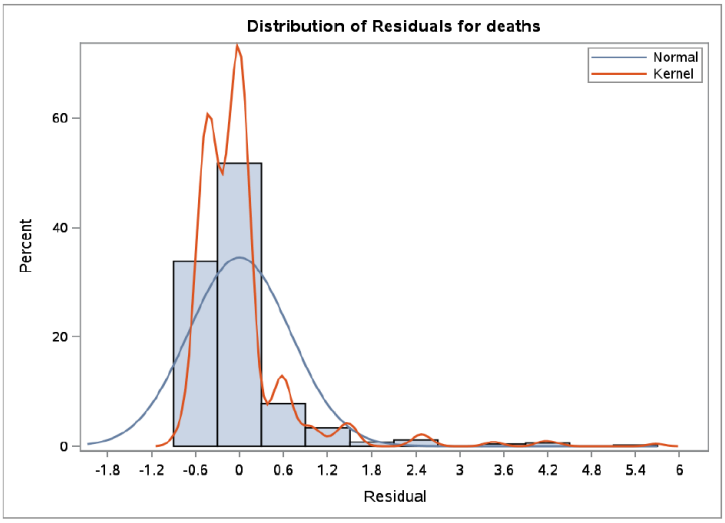

FIGURE 2.7 The graph is about the distribution of residuals for death. The blue curve is the normal curve, the orange curve is the kernel curve. Compared with the two curves, the Kernel curve does not follow the normal curve. Therefore, the residual of death does not follow a normal distribution.

FIGURE 2.8

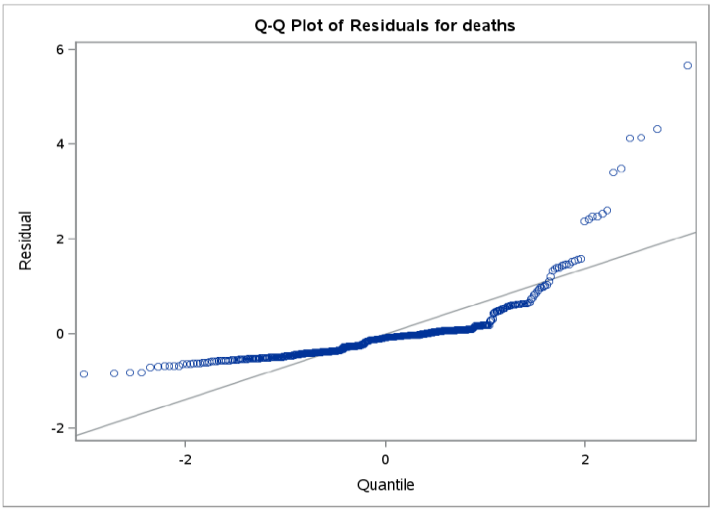

FIGURE 2.8 For the graph above, If the residuals are normally distributed, then the points should be very close to the line of the perfect fit (straight line). For the result, the normality of residuals may not hold and the points do not lie on the line of the perfect fit.

Shapiro-Wilk Normality Test:

FIGURE 2.9

W	p-value
0.65916	<2.2e-16

FIGURE 2.9 We use the Shaprio-Wilk normality test to perform the hypothesis testing if the data follows the normal distribution. The null hypothesis is the data follows the normal distribution. The alternative hypothesis is the data does not follow the normal distribution. The result shows that the p-value is less than 5% significance level. Therefore, we will reject the null hypothesis and the data does not obey the normal distribution.

Proportion of Variation:

FIGURE 2.10

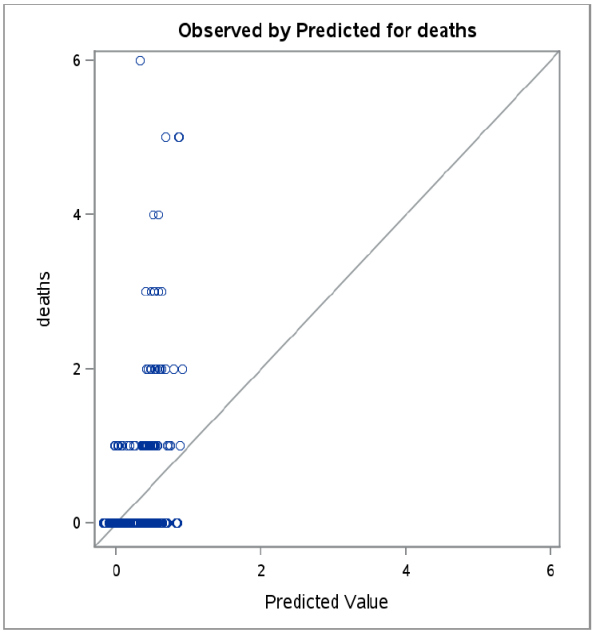

FIGURE 2.10 The graph indicates the proportion of total variation in Y that can be explained by the fitted model. If the fitted model explained a large proportion of total variation in Y, the points will be close to the line of perfect fit (diagonal). For the result, the fitted model explained a very small proportion of total variation in Y.

Cook Distance:

FIGURE 2.11

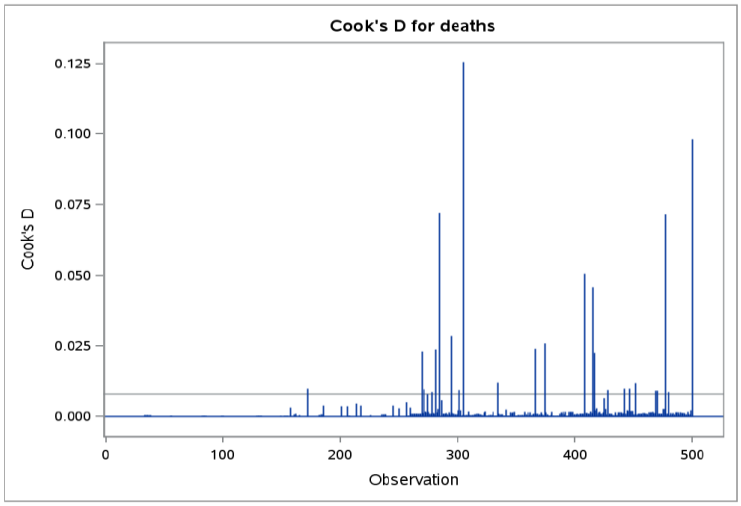

FIGURE 2.11 Cook's Distance is used to Figure out the influential outliers in a set of independent variables. For the graph, the Y-axis (Cook's D) means if the leverages and residuals are higher than the distance will be higher. For the result, some observations are higher than the reference line (horizontal line), around 20 outliers are existing.

Spread of the Residual:

FIGURE 2.12

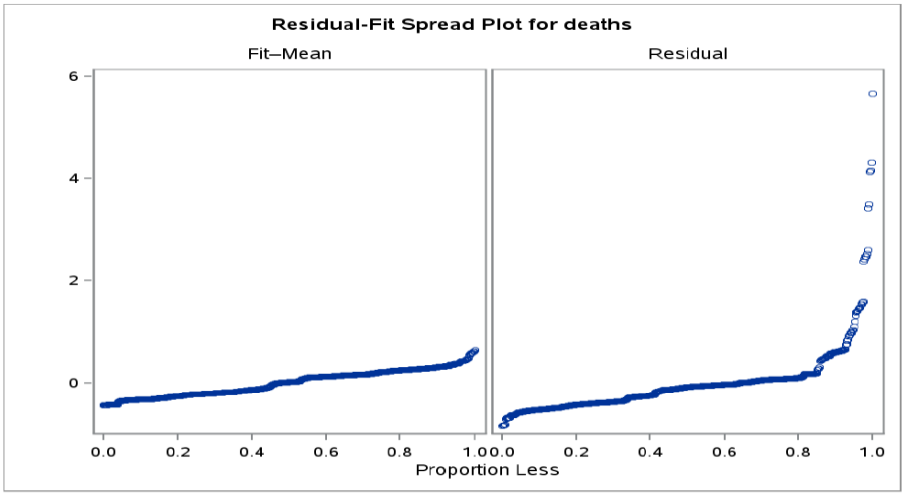

FIGURE 2.12 For the graph above, if the spread of the residual (Right) is larger than the FIt-Mean (Left), then the fitted regression model cannot capture the trend in Y. For the result, the spread of the residual larger than Fit-Mean, and the fitted regression model cannot capture the trend in Y.

Correlation

TABLE 2.13

Root MSE	0.70673	**R-Square**	0.1178
Dependent Mean	0.25000	**Adj R-Square**	0.0829
Coeff var	282.69267		

TABLE 2.13 For the coefficient of determination, R-Square, it indicates that only 11.78% of the total variation in Y can be explained by the fitted regression model. For the Adjusted R-Square, it reflects that only 8.29% of the total variation in Y can be explained by the fitted regression model.

In section 2.2.1, we find that the F-value is significant. However, we observe that the values of R-squared and Adjusted R-squared are very low. From these results, we can conclude that there is a linear relationship between the dependent variable and the independent variables, but some important independent variables are missing in the model to explain the total variation in Y.

Multicollinearity Checking

A model has multicollinearity when there is a high correlation between independent variables. We are using the Variance Inflation Factor (VIF) to check if multicollinearity exists in the model. If the VIF is equal to one, there is no multicollinearity in the model. If the VIF is bigger than one, multicollinearity exists in the model but is not severe. If the VIF is bigger than ten, there is severe multicollinearity in the model.

TABLE 2.14

Parameter Estimates					
Variable	**Parameter Estimate**	**Standard Error**	**t Values**	**Pr > |t|**	**Variance Inflation**
(Intercept)	0.23043	0.03944	5.84	<.0001	0
py	0.00203	0.00182	1.11	0.2657	1.50118
hiv1	-0.44660	0.06440	-6.93	<.0001	1.03670
factor1	-0.00005864	0.10184	-0.00	0.9995	1.64859
factor2	-0.13883	0.10328	-1.34	0.1795	1.58963
factor3	-0.13221	0.10152	-1.30	0.1934	1.62583
factor4	-0.05148	0.10341	-0.50	0.6188	1.83709
age1	-0.26912	0.19288	-1.40	0.1636	2.16595
age2	-0.21156	0.18047	-1.16	0.2484	2.50860
age3	-0.21156	0.18828	-1.12	0.2617	2.24893
age4	-0.18329	0.18165	-1.01	0.3135	2.54147
age5	-0.10022	0.18462	-0.54	0.5875	2.33804
age6	-0.18307	0.18679	-0.98	0.3275	2.48450
age7	-0.18122	0.17960	-1.01	0.3135	2.31318
age8	0.15026	0.18363	0.82	0.4136	2.25747

age9	-0.16971	0.18141	-0.94	0.3500	2.37921
age10	-0.26941	0.17769	-1.52	0.1301	2.08884
age11	-0.10538	0.18942	-0.56	0.5783	2.19725
age12	-0.15741	0.18362	-0.86	0.3917	2.19725
age13	-0.27838	0.19165	-1.45	0.1470	2.00892

TABLE 2.14 From the result, none of the VIF is greater than 10 and close to 1, we can assume that multicollinearity does not exist.

Subset Selection

Stepwise selection is a step-by-step technique that combines forward and backward selection. It is built on the regression model to help for selecting independent variables as the final regression model. When we select the best model from these selection methods, we will compare some of the criteria that are calculated from the model to find the best fit for the data, such as conducting the partial F-test to find which variables are significant or comparing the value of Mallows's Cp for each selection step. For Mallows's Cp, which was introduced by Colin Lingwood Mallows in 1964. The formula is:

$$C_p = \frac{SSE_p}{MSE} + 2p - n \tag{2.6}$$

where SSE_p is the sum of squares error with the number of p independent variables of the model, MSE is the mean squares error of the model and n is the number of sample sizes.

The forward selection begins from the null model which means the model contains intercept only and none of the independent variables. Then calculate the value of Mallow's Cp when adding each variable to the model, or do not select the variable that is not significant in the partial F-test. Repeat the process until the minimum value of Mallow's Cp of the model is obtained or the variables are selected enough in the partial F-test.

The backward selection is like an invert of the forward selection. It starts with all independent variables, deletes one that is not significant in the partial F-test at a time or calculates the value of Mallow's Cp when deleting each variable from the model. Repeat the process until the minimum value of Mallow's Cp of the model is obtained or the variables are deleted enough in the partial F-test.

For stepwise selection, it is formed by both adding and deleting the independent variable. Adding or deleting one independent variable at a time depends on the partial F-test or calculates the value of Mallow's Cp. Repeat the process until the minimum value of Mallow's Cp of the model is obtained or the variables are selected and deleted enough in the partial F-test.

For the subset selection, we would use the SAS program to perform the stepwise selection.

TABLE 2.15

Stepwise Selection				
Parameter	**Estimates**	**Std. Error**	**Chi-Square**	**Pr > ChiSq**
(Intercept)	0.257076	0.031313	67.4002	<.0001
hiv1	-0.440813	0.062733	49.3761	<.0001
age8	0.333900	0.119746	7.7751	0.0053
Model Diagnostics with 5% level of significance				
AIC	1069.86			
BIC	1086.72			

TABLE 2.15 We have used the stepwise selection method to do the subset selection. The independent variables "hiv1" and "age8" are selected because their p-values are less than 5% level of significance. The values of AIC and BIC of the model with independent variables “hiv1” and “age8” are 1069.86 and 1086.72 respectively. The independent variables of "hiv1", and “age8” would be chosen in the multiple linear regression models.

2.4 Concluding Remarks

From the above result, we find that the dataset does not fulfil the multiple linear regression assumptions since the dependent variable of the dataset does not follow the normal distribution, and the variance of residuals is not constant. The dataset only does not have multicollinearity. The multiple linear regression model required the dependent variable as a continuous variable.

FIGURE 2.16

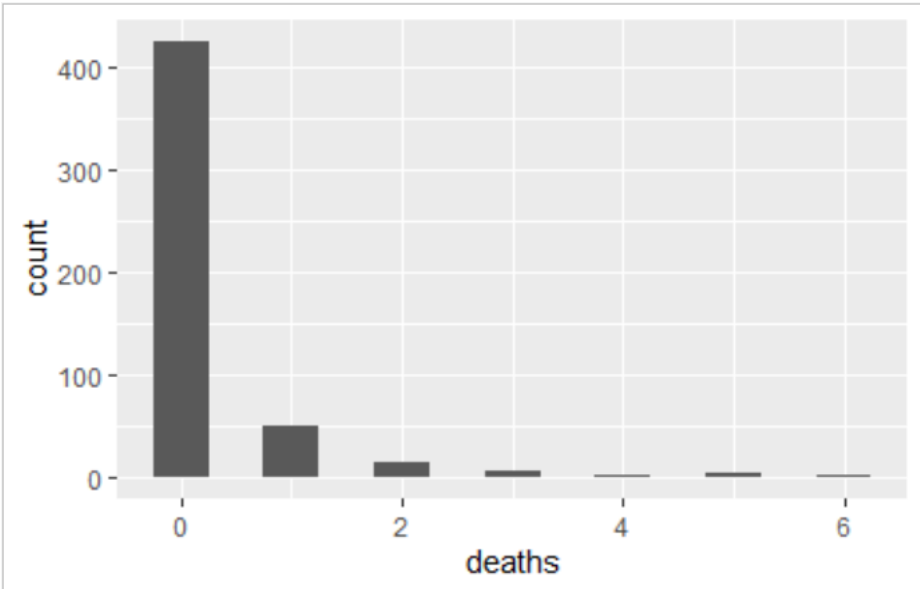

FIGURE 2.16 Most of the “deaths” count zero, then significantly decrease with the amount of death in each age group.

We notice that the dependent variable of this dataset, "death" belongs to the discrete variable. Therefore, we cannot employ the multiple linear regression model to get a well-fitting model for this dataset. About the regression model for count data, we may need to employ the regression model for count data like the Poisson log-linear regression model, it should be more appropriate.

3. Poisson Log-linear Regression Models

Poisson log-linear regression models are the same as the Poisson regression models. We adopt the former instead of the latter because we would like to specify the models using the log-linear as the regression model.
Poisson log-linear regression is a very useful tool when predicting count data through a series of continuous variables or categorical variables. We usually use the Poisson log-linear regression model when we want to model the average number of occurrences per unit of time or space. Moreover, Poisson log-linear regression requires that the dependent variable follows the Poisson distribution. Poisson distribution is suitable to describe some situations within the period, such as the number of patients' illnesses in a week, the number of inferior products produced by the factory in a month, and the number of customers arriving in a certain period of time in the queue. In this data set, the described object is the number of deaths in a particular age group, which is the dependent variable.

3.1 Literature Review

Chou et al. (2010) claims that the Poisson regression model in the effect of climate can be adopted in the effect of uncertainty on diarrhoea-related diseases in Taiwan in 1996 to 2007. As a result, the model indicated it really has a strong correlation between climate uncertainty and diarrhoea related.

A research article written by Doyle and Bottomley (2019), who studied Relative Age Effects (RAE) in European elite soccer by Poisson Regression Modelling (PRM). In this article, the argument about age-banding confers advantages to older members than young members. In the end, the publisher tested the competitive interpretation of RAE, controlled unnecessary covariance, and established an interactive effect model.

Poisson regression is very common in modelling, but it still has its shortcomings. There is a study paper written by Andrzejczak, Mlynczak and Selech (2018) who indicates a simple failure modelled from the Poisson process. Researchers claim that the problem is about predicting the costs of maintenance of the fleet of urban transport vehicles. However, when the vehicles during operation are subject to unexpected failures at random times, then it will affect the data. To solve this problem, the author applied the stochastic process in order to reduce the random data.

3.2 Poisson Log-linear Regression Models

The Poisson distribution:

$$P(Y = Y_i) = \frac{e^{-\lambda i}\ {\lambda_i}^{Yi}}{Y_i!} \tag{3.1}$$

where, $i = 1, 2, \cdots, n$;

$P(Y = Y_i)$ is the probability of Y_i occurrences;

Y_i is the number of occurrences;

λ_i is the expected value of Y_i;

e is the natural number ($2.71828\cdots$).

The Poisson log-linear regression model is in the following:

$$\lambda_i = exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \tag{3.2}$$

or,

$$ln(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \tag{3.3}$$

where, $i = 1, 2, \cdots, n$;

$\beta_1, \beta_2, \cdots, \beta_k$ are k unknown regression coefficients;

$X_1, X_2, \cdots, X_k$ are k independent variables.

Poisson log-linear regression is a kind of generalized linear model for modelling a positive integer dependent variable with linear independent variables by the link function. It assumes the dependent variable has a Poisson distribution and assumes the independent variables from the linear combination of unknown parameters. The Poisson log-linear regression model has to fulfil a few assumptions. The mean and the variance of the dependent variable should be the same. The dependent variable needs to be the discrete variable which is the count data with the non-negative values and it follows Poisson distribution rather than the normal distribution.

3.3 Model Building and Parameter Estimation

A Poisson log-linear regression model of the hemophilia dataset is in the following form:

$$ln(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \quad (3.4)$$

In section 2.2 about model building and parameter estimation. To avoid multicollinearity, we have discussed that we use dummy variables to identify qualitative variables by using the values 0 and 1. In the Poisson log-linear regression model, therefore, we also convert three independent variables "hiv", "factor" and "age", which are the qualitative variables, as the dummy variables.

The original Poisson log-linear regression model without dummy variables from formula (3.4):

$$ln(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

For the independent variable X_1 for “hiv” which has two categories, we only set one dummy variable as X_1. For the reference category will be employed when X_1 equals to zero.

$$X_1 = 1, if\ "hiv"\ =\ negative;\ X_1 = 0, otherwise$$

For the independent variable X_2 for “factor” which has 5 categories, we only set 4 dummy variables as $X_{2,1},\ X_{2,2}, X_{2,3}, X_{2,4}$. For the reference category will be employed when all dummy variables equal to zero.

$$X_{2,1} = 1, if\ "factor"\ =\ high;\ X_{2,1} = 0, otherwise$$
$$X_{2,2} = 1, if\ "factor"\ =\ moderate;\ X_{2,2} = 0, otherwise$$
$$X_{2,3} = 1, if\ "factor"\ =\ low;\ X_{2,3} = 0, otherwise$$
$$X_{2,4} = 1, if\ "factor"\ =\ unknown;\ X_{2,4} = 0, otherwise$$

For the independent variable X_3 for “age” which has 14 categories, we only set 13 dummy variables as $X_{3,1}, X_{3,2}, X_{3,3}, X_{3,4}, X_{3,5}, X_{3,6}, X_{3,7}, X_{3,8}, X_{3,9}, X_{3,10}, X_{3,11}, X_{3,12}, X_{3,13}$. For the reference category will be employed when all dummy variables equal to zero.

$$X_{3,1} = 1, if\ "age"\ =\ "0-4";\ X_{3,1} = 0, otherwise$$
$$X_{3,2} = 1, if\ "age"\ =\ "5-9";\ X_{3,2} = 0, otherwise$$
$$X_{3,3} = 1, if\ "age"\ =\ "10-14";\ X_{3,3} = 0, otherwise$$
$$X_{3,4} = 1, if\ "age"\ =\ "15-19";\ X_{3,4} = 0, otherwise$$
$$X_{3,5} = 1, if\ "age"\ =\ "20-24";\ X_{3,5} = 0, otherwise$$
$$X_{3,6} = 1, if\ "age"\ =\ "25-29";\ X_{3,6} = 0, otherwise$$
$$X_{3,7} = 1, if\ "age"\ =\ "30-34";\ X_{3,7} = 0, otherwise$$
$$X_{3,8} = 1, if\ "age"\ =\ "35-39";\ X_{3,8} = 0, otherwise$$
$$X_{3,9} = 1, if\ "age"\ =\ "40-44";\ X_{3,9} = 0, otherwise$$
$$X_{3,10} = 1, if\ "age"\ =\ "45-49";\ X_{3,10} = 0, otherwise$$

$$X_{3,11} = 1, if\ "age"\ =\ "50-54";\ X_{3,11} = 0, otherwise$$
$$X_{3,12} = 1, if\ "age"\ =\ "55-59";\ X_{3,12} = 0, otherwise$$
$$X_{3,13} = 1, if\ "age"\ =\ "60-64";\ X_{3,13} = 0, otherwise$$

After employing dummy variables into the original Poisson log-linear regression model, we have a model including dummy variables:

$$ln(\lambda) = \beta_0 + (dummy\ variable\ "hiv")\ +\ (dummy\ variables\ "factor") + (dummy\ variables\ "age") + \beta_4 X_4$$

where, $dummy\ variable\ "hiv" = \beta_1 X_1$;

$dummy\ variables\ "factor" = \beta_{2,1}X_{2,1} + \beta_{2,2}X_{2,2} + \beta_{2,3}X_{2,3} + \beta_{2,4}X_{2,4}$;

$dummy\ variables\ "age" = \beta_{3,1}X_{3,1} + \beta_{3,2}X_{3,2} + \cdots + \beta_{3,13}X_{3,13}$.

Method of Maximum Likelihood

We will use the method of maximum likelihood to estimate the regression parameter, β.

The Poisson distribution from formula (3.1):

$$P(Y = Y_i) = \frac{e^{-\lambda i}{\lambda_i}^{Yi}}{Y_i!}$$

We perform the likelihood:

$$LogL(\beta) = \sum_{i=1}^{n} \{-\ exp\ exp\ (\beta^T X_i)\ + Y_i(\beta^T X_i) - log\ (Y_i!)\}$$

The maximum likelihood estimator for β is:

$$\hat{\beta} = argmax(LogL(\beta))$$

where, $argmax$is the point of the function so that the value of the function is the maximum.

3.3.1 Statistical Analysis

For the statistical analysis, we would use the SAS program to test whether the variables and coefficients are significant.

TABLE 3.1

Coefficients				
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-0.082637	0.337662	-0.245	0.8067
py	0.010433	0.005657	1.844	0.0651
hiv1	-2.496948	0.320152	-7.799	6.23e-15
factor1	-0.080673	0.262250	-0.308	0.7584
factor2	-0.689753	0.315617	-2.185	0.0289
factor3	-0.565790	0.296776	-1.906	0.0566
factor4	-0.255186	0.283817	-0.899	0.3686
age1	-1.200230	0.654473	-1.834	0.0677
age2	-0.718603	0.465616	-1.543	0.1227
age3	-0.719521	0.470086	-1.531	0.1259
age4	-0.718456	0.464325	-1.547	0.1218
age5	-0.447318	0.431719	-1.036	0.3001
age6	-0.719844	0.465800	-1.545	0.1223
age7	-0.644521	0.469388	-1.373	0.1697

age8	0.353657	0.381666	0.927	0.3541
age9	-0.550173	0.443490	-1.241	0.2148
age10	-1.095869	0.501389	-2.186	0.0288
age11	-0.302958	0.461665	-0.656	0.5117
age12	-0.517935	0.476291	-1.087	0.2768
age13	-0.983490	0.535213	-1.838	0.0661
Fit Statistics				
AIC			591.2575	
BIC			675.5496	

TABLE 3.1 The above table shows the parameter estimation by using Poisson log-linear regression model with the dummy variables. From the result, the p-values only for independent variables “hiv1”, “factor2”, and “age10” are smaller than 5% level of significance. Although most of the variables are not satisfied, they are still better than the multiple linear regression model. The values of AIC and BIC are much lower, which are 591.2575 and 675.5496, than multiple linear regression models.

3.4 Diagnostic Checking and Model Selection

Pseudo R-squared is a scale that is similar to the R-squared. It is ranging from 0 to 1. If the value of pseudo R-squared for the model approaches 1, then it is a better model. Moreover, it only measures the goodness-of-fit for the count models such as the Poisson log-linear regression models.

TABLE 3.2

Pseudo-R^2	0.2726175

TABLE 3.2 We calculate the pseudo R-squared is 0.27261756, which is not close to 1. Therefore, the Poisson log-linear regression model is not fit.

TABLE 3.3

Mean	0.25
Variance	0.5445892

TABLE 3.3 We calculated the mean and the variance of the dependent variable, we find that the mean is not equal to the variance.

FIGURE 3.4

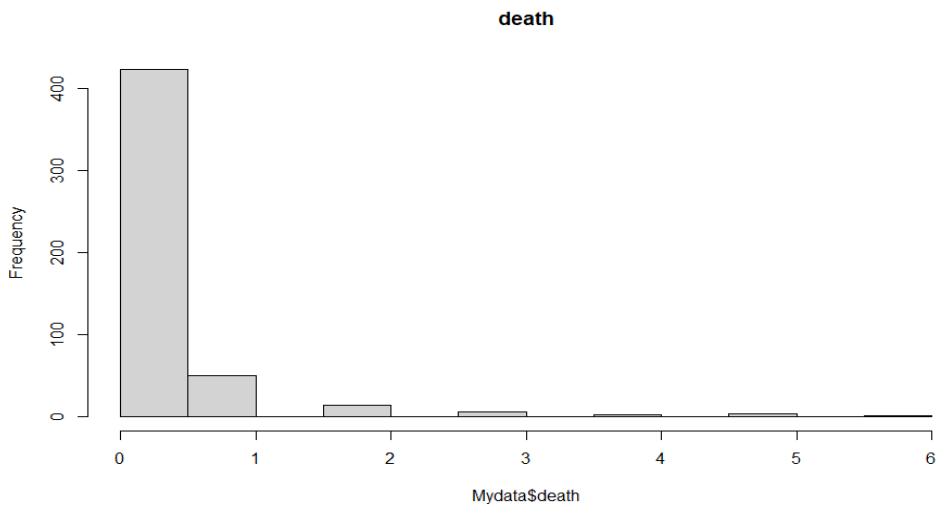

FIGURE 3.4 From the histogram of the dependent variable, we find that it is right-skewed and assume that it does not follow the normal distribution. Moreover, the histogram is showing that the dependent variable is a count data which is a non-negative integer value.

Dispersion Checking

In a Poisson distribution, the variance and mean are equal. However, there are three types of dispersion that occur in the Poisson distribution depending on the variance and the mean. The first dispersion is overdispersion which occurs in Poisson log-linear regression when the observed variance of the dependent variable is larger than the mean of the dependent variable. The second dispersion is underdispersion which occurs when the observed variance of the dependent variables is lower than the mean of the dependent variable. The last dispersion is equal-dispersion which occurs when the model fulfils the Poisson distribution assumption which is that the observed variance of the dependent variables and the mean of the dependent variable are equal.

The result of section 3.4 shows that the variance is larger than the mean. We may assume that the model is overdispersed. As overdispersion is often encountered when dealing with count data and can have a negative impact on the interpretation of the results, we will spend some time discussing it. Overdispersion is suggested if the ratio of the residual deviance to the residual degrees of freedom is much larger than 1.

TABLE 3.5

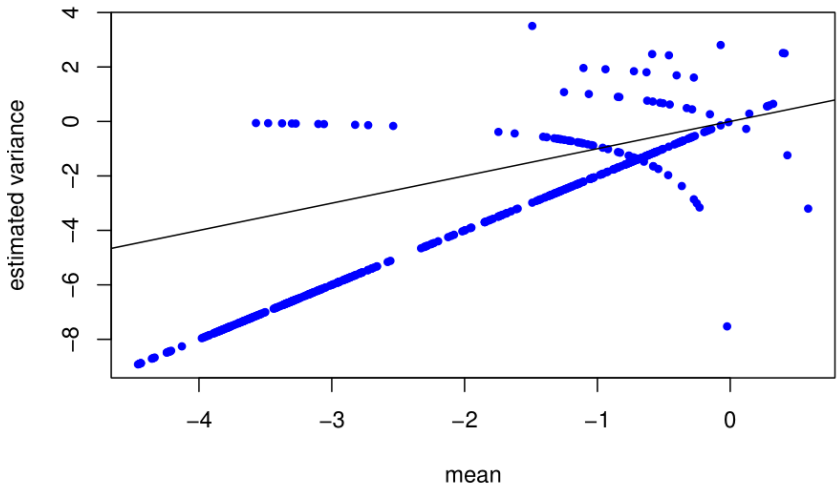

TABLE 3.5 We can see that the majority of the variance is not equal to the mean, which indicates that this dataset does have an overdispersion.

Quantitatively, the dispersion parameter can be estimated using Pearson Chi-squared statistic and the degree of freedom. When the dispersion estimate is larger than 1, the model has an overdispersion problem. When the dispersion estimate is less than 1, the model has an underdispersion problem. When the dispersion estimates approaches 1, the model does not have a dispersion problem. The result gives us 1.474642 and confirms that the Poisson log-linear regression model has a significant overdispersion problem.

Subset Selection

For the subset selection, we would use the SAS program to perform the stepwise selection.

TABLE 3.6

Stepwise Selection				
Parameter	**Estimates**	**Std. Error**	**Chi-Square**	**Pr > ChiSq**
(Intercept)	-2.174639	0.181867	142.9765	<.0001
py	0.010492	0.004268	6.0420	0.0140
hiv1	-2.513017	0.318721	62.1684	<.0001
age8	1.008440	0.249345	16.3568	<.0001
Model Diagnostics with 5% level of significance				
AIC	575.40			
BIC	592.26			

TABLE 3.6 We have used the stepwise selection method to do the subset selection. The independent variables ”py”, "hiv1" and "age8" are selected because their p-values are less than 5% level of significance. The values of AIC and BIC of the model with independent variables ”py”, "hiv1" and "age8" are 575.40 and 592.26 respectively. The independent variables of ”py”, "hiv1", and “age8” would be chosen in the Poisson log-linear regression models.

3.5 Concluding Remark

According to the result, we found that the dataset does not fulfil the assumptions of the Poisson log-linear regression model. The mean and the variance of the dependent variable are not the same, each observation is not independent. We find that the dependent variable is over-dispersed. The Poisson log-linear regression model may not be suitable because of the existence of overdispersion. Therefore, we may adopt some other regression models that can address the overdispersion.

Problem of Overdispersion or Underdispersion

Count data under Poisson assumption are usually overdispersed. The Pearson or deviance residuals are too large and the deviance goodness of fit statistics shows a poor fit.

The expectation and variance of negative binomial distribution are given below:

$$\begin{cases} \mathrm{E(Y)} = \lambda & \lambda = 0 \\ \mathrm{Var(Y)} = \lambda[1 + \alpha\lambda] & \alpha \geq 0, \lambda \geq 0 \end{cases}$$

Since $\alpha \geq 0, \lambda \geq 0$, therefore $Var(Y) > E(Y)$. Thus, it is an example to employ the negative binomial distribution for modeling data with overdispersion. Negative binomial regression models may be an approach to deal with overdispersion.

The consequences of ignoring overdispersion may be serious. First, the size and signs of the covariate estimates in a log-linear or logit regression model may be very similar if overdispersion is not properly handled. Second, biased parameter estimates together with underestimated standard error will be obtained. Third, the precision of parameter estimates will be extremely high, whereas the p-value for testing the significance of covariate will be extremely low. Consequently, incorrect, or misleading inference will result.

4. Negative Binomial Regression Models

When the sample mean and the sample variance of the count data are not the same, we would consider replacing the Poisson regression to another regression model which is negative binomial regression. This situation is known as overdispersion or underdispersion. In the hemophilia dataset study, the variance is greater than the mean, we consider that it might encounter the overdispersion. The negative binomial regression is similar to multiple linear regression, but the dependent variable of negative binomial regression should follow the negative binomial distribution, such that the dependent variable is non-negative. (NCSS Statistical Software, n.d.)

4.1 Literature Review

A research paper written by An, Wu, Fan, Pan and Sun (2016) who studied in fort putting the negative binomial regression model about predicting from the beginning of a hand, foot, mouth disease epidemic in the Dalian, Liaoning province, China. Lastly, the models are chosen by the Akaike Information Criterion (AIC), also the sensitivity can be reached up to 100%.

Besides, a paper written by Byers, Allore, Gill and Peduzzi (2003) who provided about the appliance of the negative binomial modelling in a study about the ageing field. with discrete outcomes. The article mentioned that the data output of the primary survey with the sample variance almost six times greater than the sample mean, and the data distribution was right skewed. In the end, the negative binomial regression model is very effective abnormally in this case.

4.2 Negative Binomial Regression Models

The negative binomial distribution:

$$P(Y = Y_i|\mu_i\alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu^{-1}}{\alpha^{-1} + \mu_i}\right)^{Y_i} \qquad (4.1)$$

where,

$i = 1, 2, \cdots, n$;

$P(Y = Y_i)$ is the probability of Y_i occurrences;

Γ is the gamma function used extend the factorial function;

$\mu_i = t_i\mu$;

μ is the mean incidence rate of y per unit of population size;

t_i is the population size for a specific observation;

$\alpha = \frac{1}{v}$;

α is the dispersion parameter;

v is the scale parameter.

The negative binomial regression model is the following:

$$\mu_i = exp(ln(t_i) + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}) \quad (4.2)$$

where, $i = 1, 2, \cdots, n$;

t_i is an exposure time for a specific observation;

$\beta_1, \beta_2, \cdots, \beta_k$ are k unknown regression coefficients;

$X_1, X_2, \cdots, X_k$ are k independent variables.

4.3 Model Building and Parameter Estimation

A negative binomial regression model of the hemophilia dataset is in the following form:

$$\mu = exp(\ ln(t) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4) \quad (4.3)$$

To avoid multicollinearity, we have discussed that we use dummy variables to identify qualitative variables by using the values 0 and 1. In the negative binomial regression model, therefore, we convert three independent variables "hiv", "factor" and "age", which are the qualitative variables, as the dummy variables.

The original negative binomial regression model without dummy variables from formula (4.3):

$$\mu = exp(\ ln(t) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)$$

For the independent variable X_1 for “hiv” which has two categories, we only set one dummy variable as X_1. For the reference category will be employed when X_1 equals to zero.

$$X_1 = 1, if\ "hiv"\ =\ negative;\ X_1 = 0, otherwise$$

For the independent variable X_2 for “factor” which has 5 categories, we only set 4 dummy

variables as $X_{2,1},\ X_{2,2}, X_{2,3}, X_{2,4}$. For the reference category will be employed when all dummy variables equal to zero.

$$X_{2,1} = 1, if\ "factor"\ =\ high;\ X_{2,1} = 0, otherwise$$
$$X_{2,2} = 1, if\ "factor"\ =\ moderate;\ X_{2,2} = 0, otherwise$$
$$X_{2,3} = 1, if\ "factor"\ =\ low;\ X_{2,3} = 0, otherwise$$
$$X_{2,4} = 1, if\ "factor"\ =\ unknown;\ X_{2,4} = 0, otherwise$$

For the independent variable X_3 for "age" which has 14 categories, we only set 13 dummy variables as $X_{3,1}, X_{3,2}, X_{3,3}, X_{3,4}, X_{3,5}, X_{3,6}, X_{3,7}, X_{3,8}, X_{3,9}, X_{3,10}, X_{3,11}, X_{3,12}, X_{3,13}$. For the reference category will be employed when all dummy variables equal to zero.

$$X_{3,1} = 1, if\ "age"\ =\ "0-4";\ X_{3,1} = 0, otherwise$$
$$X_{3,2} = 1, if\ "age"\ =\ "5-9";\ X_{3,2} = 0, otherwise$$
$$X_{3,3} = 1, if\ "age"\ =\ "10-14";\ X_{3,3} = 0, otherwise$$
$$X_{3,4} = 1, if\ "age"\ =\ "15-19";\ X_{3,4} = 0, otherwise$$
$$X_{3,5} = 1, if\ "age"\ =\ "20-24";\ X_{3,5} = 0, otherwise$$
$$X_{3,6} = 1, if\ "age"\ =\ "25-29";\ X_{3,6} = 0, otherwise$$
$$X_{3,7} = 1, if\ "age"\ =\ "30-34";\ X_{3,7} = 0, otherwise$$
$$X_{3,8} = 1, if\ "age"\ =\ "35-39";\ X_{3,8} = 0, otherwise$$
$$X_{3,9} = 1, if\ "age"\ =\ "40-44";\ X_{3,9} = 0, otherwise$$
$$X_{3,10} = 1, if\ "age"\ =\ "45-49";\ X_{3,10} = 0, otherwise$$
$$X_{3,11} = 1, if\ "age"\ =\ "50-54";\ X_{3,11} = 0, otherwise$$
$$X_{3,12} = 1, if\ "age"\ =\ "55-59";\ X_{3,12} = 0, otherwise$$
$$X_{3,13} = 1, if\ "age"\ =\ "60-64";\ X_{3,13} = 0, otherwise$$

After employing dummy variables into the original negative binomial regression model, we have a model including dummy variables:

$$\mu = exp(\ ln(t) + (dummy\ variable\ "hiv")\ +\ (dummy\ variables\ "factor") + (dummy\ variables\ "age") + \beta_4 X_4)$$

where, $dummy\ variable\ "hiv" = \beta_1 X_1$;

$dummy\ variables\ "factor" = \beta_{2,1} X_{2,1} + \beta_{2,2} X_{2,2} + \beta_{2,3} X_{2,3} + \beta_{2,4} X_{2,4}$;

$dummy\ variables\ "age" = \beta_{3,1} X_{3,1} + \beta_{3,2} X_{3,2} + \cdots + \beta_{3,13} X_{3,13}$.

Method of Maximum Likelihood

We will use the method of maximum likelihood to estimate the regression parameter, β.

The likelihood equations:

$$\sum_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{1 + \mu_i} = 0, \quad j = 1, 2, \cdots, k$$

$$\sum_{i=1}^{n} \left\{ \alpha^{-2} \left(ln(1 + \alpha\mu_i) - \sum_{j=0}^{y_i - 1} \frac{1}{j + \alpha^{-1}} \right) + \frac{y_i - \mu_i}{\alpha(1 + \alpha\mu_i)} \right\} = 0$$

The maximum likelihood estimator for β is:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} \sim N \left[\begin{matrix} \beta \\ \alpha \end{matrix} \middle| \begin{matrix} V(\hat{\beta}) & Cov(\hat{\beta}, \hat{\alpha}) \\ Cov(\hat{\beta}, \hat{\alpha}) & V(\hat{\alpha}) \end{matrix} \right]$$

where, $V(\hat{\beta}) = \left[\sum_{i=1}^{n} \frac{\mu_i}{1+\alpha\mu_i} x_i x'_j \right]^{-1}$;

$V(\hat{\alpha}) = \sum_{i=1}^{n} \left\{ \alpha^{-4} \left(ln(1 + \alpha\mu_i) - \sum_{j=0}^{y_i - 1} \frac{1}{j+\alpha^{-1}} \right)^2 + \frac{\mu_i}{\alpha^2(1+\alpha\mu_i)} \right\}^{-1}$;

$Cov(\hat{\beta}, \alpha) = [0]$.

4.3.1 Statistical Analysis

We have used SAS to test whether the dummy variables and its coefficients are significant, the results as belows:

TABLE 4.1

Coefficients				
	Estimate	**Std. Error**	**Wald Chi-Square**	**Pr > ChiSq**
(Intercept)	-2.2725	0.2190	107.71	<.0001
py	0.0121	0.0071	2.89	0.0891
hiv1	-2.5311	0.3590	49.71	<0.0001
factor1	-0.2698	0.3791	0.51	0.4766

factor2	-0.7861	0.4247	3.43	0.0642
factor3	-0.6864	0.4041	2.88	0.0894
factor4	-0.2793	0.3912	0.51	0.4753
age1	-1.2793	0.8148	2.46	0.1164
age2	-0.6210	0.6431	0.93	0.3342
age3	-0.7878	0.6597	1.43	0.2324
age4	-0.7789	0.6544	1.42	0.2340
age5	-0.5147	0.6146	0.70	0.4023
age6	-0.8739	0.6527	1.79	0.1806
age7	-0.6925	0.6497	1.14	0.2865
age8	0.1260	0.5904	0.05	0.8310
age9	-0.5920	0.6249	0.90	0.3435
age10	-1.1518	0.6604	3.04	0.0811
age11	-0.2740	0.6561	0.17	0.6763
age12	-0.5014	0.6625	0.57	0.4492
age13	-0.9602	0.7017	1.87	0.1712
(Dispersion)	1.8460	0.5327	/	/
Fit Statistics				
AIC			552.5830	
BIC			641.0898	

TABLE 4.1 The above table indicates the parameter estimation by using the negative binomial regression model with the dummy variables. For the result, the p-value only for the independent variable of “hiv1” is smaller than the 5% level of significance. After the analysis, since only the independent variable of “hiv1” is significant, it seems that the performance is no major change compared with the previous regression models. The values of AIC and BIC are low, which are 552.5830 and 641.0898.

4.4 Diagnostic Checking and Model Selection

The assumptions of negative binomial regression models are very similar to Poisson regression models, the reason is these two models have several of the same assumptions which can refer to section 3.2. The difference is negative binomial regression models allow the variance much greater than the mean from the dataset.

FIGURE 4.2

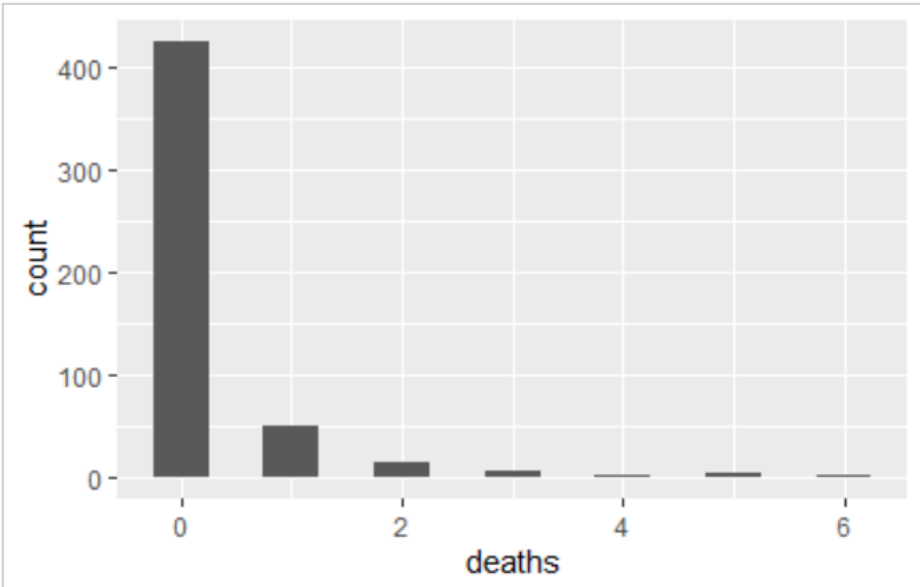

FIGURE 4.2 The figure indicates the distribution of values in dependent variable “deaths”. We can see that there are no negative values in the dependent variable which means the dependent variable is non-negative and follows the negative binomial distribution. The model may be appropriate to the dataset.

Dispersion Checking

TABLE 4.3

Pearson Chi-square dispersion statistic
0.9463652

TABLE 4.3 The Pearson Chi-square dispersion statistic indicates that if variance equals to mean, then the dispersion statistic value should be 1. The dispersion test reports that there is a slight underdispersion problem since the dispersion statistic is less than but approximately to 1.

Comparison with Poisson Regression Models

Using the AIC & BIC criteria, comparing Poisson log-linear regression models and the negative binomial regression models. If the information criterion value is smaller, then the model fitted should be better.

TABLE 4.4

Models	AIC	BIC
Poisson regression models	591.2575	675.5496
Negative binomial regression models	552.5830	641.0898

TABLE 4.4 The result shows that the negative binomial regression models have the smallest AIC and BIC. Therefore, the negative binomial regression models would perform better than Poisson log-linear regression models. In the meantime, this result corresponds with the expectation from the assumptions

.

Subset Selection

For the subset selection, we would use the SAS program to perform the stepwise selection.

TABLE 4.5

Stepwise Selection				
Parameter	**Estimates**	**Std. Error**	**Chi-Square**	**Pr > ChiSq**
(Intercept)	-1.953766	0.172182	128.7568	<0.001
hiv1	-2.410334	0.344364	48.9914	<0.001
Dispersion	2.360109	0.610571	/	/
Model Diagnostics with 5% level of significance				
AIC	534.98			
BIC	547.63			

TABLE 4.5 We have used the stepwise selection method to do the subset selection. The independent variable "hiv1" is selected because the p-values are less than 5% level of significance. The values of AIC and BIC of the model with independent variable “hiv1” are 534.98 and 547.63 respectively. The independent variable of "hiv1" would be chosen in the negative binomial regression models.

4.5 Concluding Remarks

Based on the above analysis, the values of AIC and BIC in both parameter estimation and stepwise selection are much lower than before models. The Pearson Chi-square dispersion statistic evinced that the negative binomial regression models can deal with the overdispersion problem but leads to a slight underdispersion problem. After further investigation, this is caused by overmuch zero in the dependent variable which is about 85.5% of the observations (Figure 4.6). At this time, some advanced models could be employed because this model can effectively deal with too many zero problems. In conclusion, the negative binomial regression model may not be the best model since these are not enough to make a final decision. We should go for some advanced models to see if there is a better performance.

Advanced Model Checkings

FIGURE 4.6

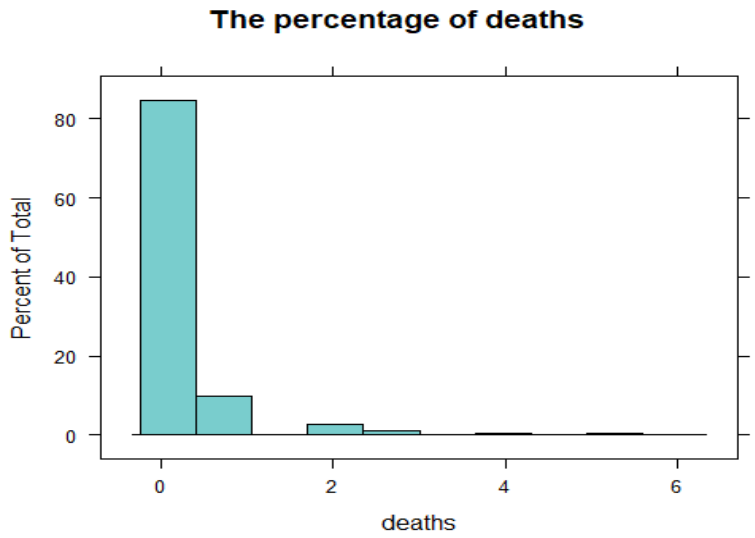

FIGURE 4.6 The plot indicates that nearly 85.5% of the data independent variable “deaths” is zero.

FIGURE 4.7

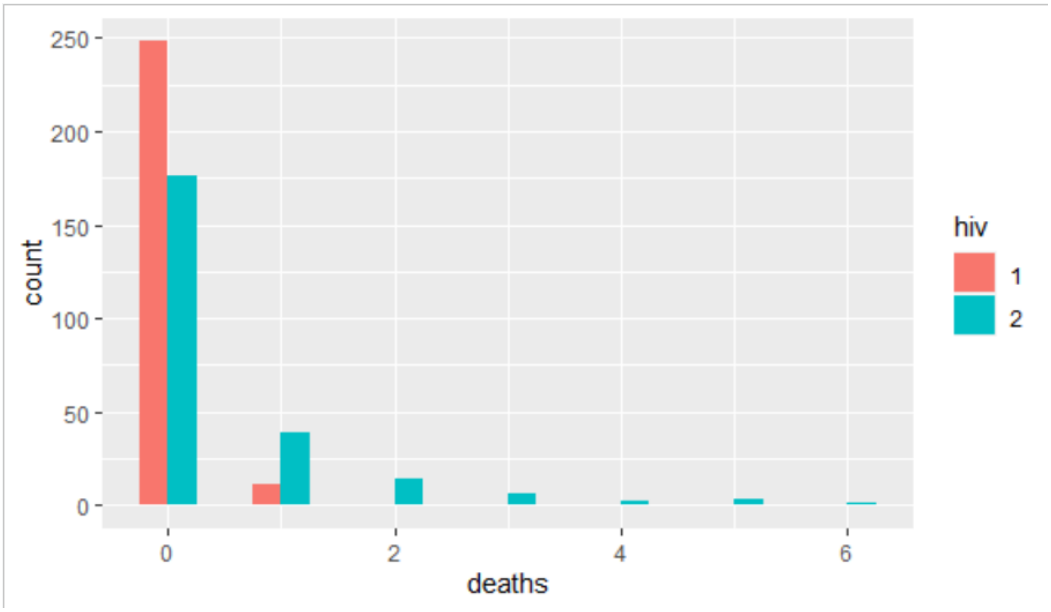

FIGURE 4.7 In the "0" death group, the number of patients who do not suffer HIV is larger than those who suffer HIV. The patients with positive status in HIV have a higher risk of death.

FIGURE 4.8

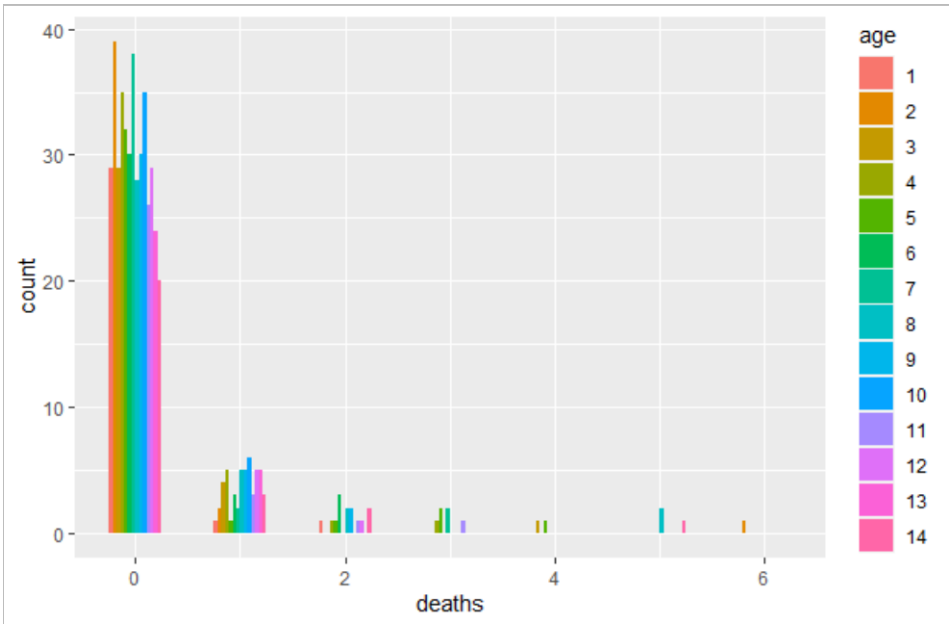

FIGURE 4.8 There are 14 age groups present in this dataset. Regardless of the "age" group, most of the "death" counts zero.

FIGURE 4.9

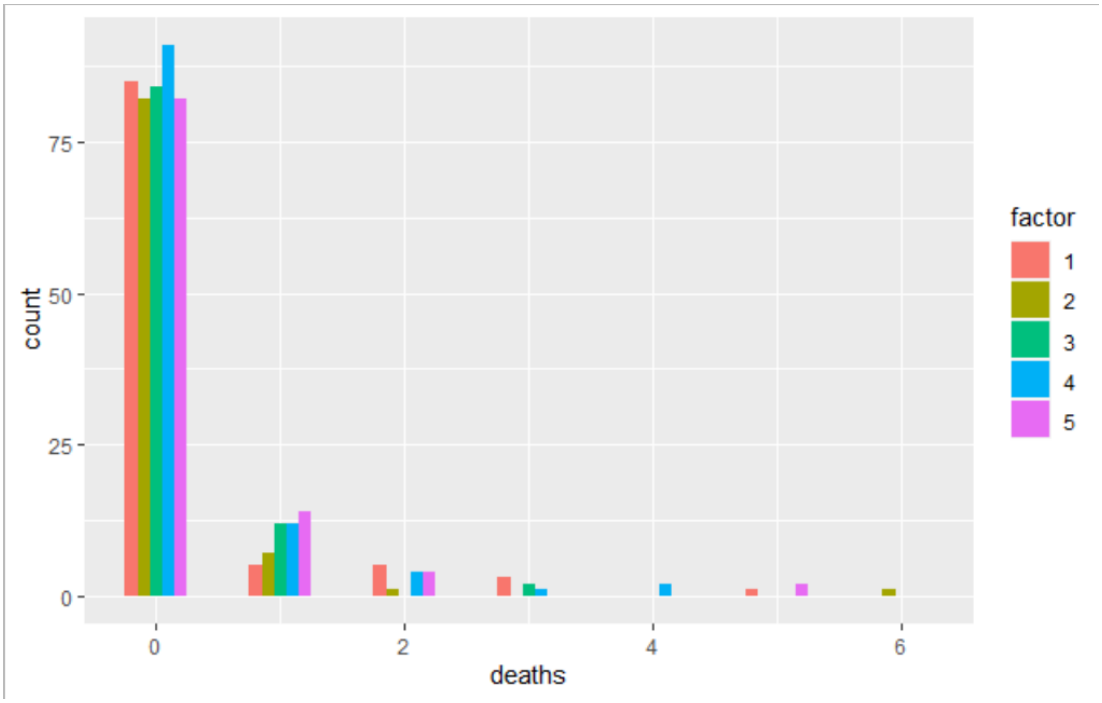

FIGURE 4.9 There are five-factor groups present in this dataset. No matter how many doses of blood clotting preparation each patient, most of the "death" counts zero.

Findings

These bar charts plotted the data distribution of the dependent variable "deaths" with each independent variable. We found that most of the data from independent variables distributed in the dependent variable "deaths" counts zero.

For the problem of excessive zero, two types of advanced models that could handle the problem respectively are zero-inflated models and hurdle models. Zero-inflated models include the zero-inflated Poisson regression models and zero-inflated negative binomial regression models. They are generated by a mixture distribution consisting of the degenerate and non-degenerate parts. On the other hand, hurdle models contain Poisson hurdle models and negative binomial hurdle models. Hurdle models consist of two parts, the first part is the probability for zero values, and the second part is the probability for non-zero values. Both zero-inflated models and hurdle models can also be used to deal with overdispersion and excessive zero data. Although these two kinds of models have their own benefits, we would mainly focus on zero-inflated models in this project for our dataset. For the details, we would like to discuss these models in the next two sections.

5. Zero-inflated Poisson Regression Models

After reviewing the previous classical models, we found that classical models may not be very suitable for the dataset. Therefore, we would like to build-up some other useful models, zero-inflated models would be the more suitable models that can be used for the dataset that addresses the problems of excessive zeros and overdispersion.

Zero-inflated Poisson regression model is one of the zero-inflated models. The model generates two models and combines them at the end. (a logit model and a Poisson model) Assuming two types of zero existing in the data. The first part is when the structure zero is generated by the existence of some special structure in the data; the other part is the sampling zero generated by the Poisson distribution. (Xie et al., 2013)

Zero-inflated models are very popular in various fields, such as public health, biomedicine, economics, actuarial science, road safety, insurance, and agriculture. In general, zero-inflated data are relative to non-zero count data. The number of zeros is more than expected. For example, consider a set of data that contains many zeros. This dataset consists of two parts: Part 1: Zeros and non-zero count data form a Poisson distribution. Part 2: The remaining zeros are additionally obtained. These are the zero-inflated data or structural zero. In practice, these extra zeros can be regarded as the zeros generated by a degenerate population, and the other count data and generated by a non-degenerate population, such as the standard Poisson distribution. The zero-inflated Poisson model uses two components that refer to two different zero generating processes. The first process is controlled by a binary distribution (intercept only model) degenerated at zero that yields excessive zero, whereas the second one is regulated by a Poisson distribution that yields zero and non-zero counts. The degenerate binary distribution can capture the excessive zeros.

In effect, zero-inflated data are generated by a mixture distribution consisting of the degenerate and non-degenerate parts. Zero-inflation can be found in both continuous and discrete datasets.

Lambert (1992) considered the zero-inflated regression models (ZIP). Such a model assumes that the excessive zeros and the count data generated by the Poisson distribution are of proportion ϕ and $(1-\phi)$. This makes up the mixture distribution in equations (6.4). In addition, covariates can be integrated into the excessive zeros part and the count data part generated by the Poisson distribution.

The log-linear model (McCullagh and Nelder, 1989) is generally adopted for the Poisson regression part:

$$log(\lambda) = X^T\beta \tag{5.1}$$

Where X is the covariate vector and β is the regression parameter vector. Because of the proportions ϕ and $(1-\phi)$ for the excessive zeros and the Poisson distribution generated counts, as in the binomial distribution, McCullagh and Nelder (1989) employed the logistic regression model:

$$logit(\phi) = Z^T\gamma \tag{5.2}$$

where $logit(\phi) = log\frac{\phi}{1-\phi}$, Z is the covariates and γ is the regression coefficient.

Thus, the zero-inflated Poisson model can be expressed as:

$$\begin{cases} \text{logit}(\phi) = Z^T\gamma \\ \log(\lambda) = X^T\beta \end{cases} \tag{5.3}$$

In this model, there is a parameter ϕ_i which is the proportion of structural zeros (non-Poisson data). When $0 < \phi_i < 1$, zero inflation exists. The larger the value of ϕ_i, the higher proportion of structural zeros is. When $\phi_i = 0$, the zero-inflated Poisson model reduces to the standard Poisson. When $\phi_i < 0$, insufficient zeros in the data. This is called zero-deflation.

5.1 Literature Review

An article written by Poston and McKibben (2003) who studied the use of zero-inflated Poisson regression models to estimate the count data of the fertility of U. S. women. Researchers highlighted the advantages of the zero-inflated Poisson model and the zero-inflated negative binomial model in processing count data by outputing Poisson regression model and negative binomial regression model.

Besides, Yesilova, Kaydan and Kaya (2010) offered that the zero-inflated Poisson regression could be modelled for Insect-egg data with excess zeros. Researchers established this model for the overdispersion of data had been very effective and tested the performance of this model is better than other Poisson regression.

Moreover, Nie, Wu, Brockman and Zhang (2006) claim that zero-inflated Poisson regression can be used to determine the relationship style between mRNA and protein richness. As a result, researchers found that the predicted protein values are corrected like experimentally detected values by the model. Then it can be estimated for the value of undetected proteins.

5.2 Zero-inflated Poisson Regression Models

The probability distribution of the zero-inflated Poisson:

$$P(Y_i = j) = \begin{cases} \phi_i + (1-\phi_i)e^{-\lambda_i} & \text{if } j = 0 \\ (1-\phi_i)\dfrac{e^{-\lambda_i}\,\lambda_i{}^{Yi}}{Y_i!} & \text{if } j > 0 \end{cases} \tag{5.4}$$

where,

$i = 1, 2, \cdots, n$;

$P(Y = Y_i)$ is the probability of Y_i occurrences;

$\phi_i = \frac{\mu_i}{1+\mu_i}$ is the logistic link function;

$\mu_i = exp(ln(t_i) + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_m Z_{im})$;

$\gamma_1, \gamma_2, \cdots, \gamma_m$ are m unknown regression coefficients;

$Z_1 + Z_2 + \cdots + Z_i$ are m independent variables.

The expectation and variance of zero-inflated Poisson mixture model is given by:

$$E(Y) = (1-\phi)\lambda \tag{5.5}$$

$$Var(Y) = E(Y)[1 + \lambda - E(Y)] \tag{5.6}$$

The zero-inflated Poisson regression model is the following:

$$\lambda_i = exp(ln(t_i) + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) \tag{5.7}$$

where, $i = 1, 2, \cdots, n$;

t_i is an exposure time for a specific observation;

$\beta_1, \beta_2, \cdots, \beta_k$ are k unknown regression coefficients;

$X_1, X_2, \cdots, X_k$ are k independent variables.

The regressor Z_i and X_i in the $\mu_i = exp(ln(t_i) + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_m Z_{im})$ and $\lambda_i = exp(ln(t_i) + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$ of the zero-inflated Poisson model need not be different.

5.3 Model Building and Parameter Estimation

TABLE 5.1

deaths				
deaths	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	248	95.75	248	95.75
1	11	4.25	259	100.00

TABLE 5.1 The table indicates the situation when independent variable "hiv" = 1 and how many zero of the dependent variable "deaths". From the result, we can observe that 95.75% of "deaths" is 0 when "hiv" = 1. In summary, we can almost certain that when "hiv" = 1, "deaths" must be 0. Therefore, the probability of zero count depends on the "hiv". We will put “hiv1” variable to "zeromodel" part of the formula., which is $\phi_i + (1 - \phi_i)e^{-\lambda_i}$ in formula (5.4), refer to section 5.2.

A zero-inflated Poisson regression model of the hemophilia dataset is in the following form:

$$\lambda = exp(ln(t) + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 + \beta_4 X_4) \qquad (5.8)$$

To avoid multicollinearity, we have discussed that we use dummy variables to identify qualitative variables by using the values 0 and 1. In the zero-inflated Poisson regression model, therefore, we convert three independent variables "hiv", "factor" and "age", which are the qualitative variables, as the dummy variables.

The original zero-inflated Poisson regression model without dummy variables from formula (5.8):

$$\lambda = exp(ln(t) + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 + \beta_4 X_4)$$

For the independent variable X_1 for “hiv” which has two categories, we only set one dummy variable as X_1. For the reference category will be employed when X_1 equals to zero.

$$X_1 = 1, if\ "hiv"\ =\ negative;\ X_1 = 0, otherwise$$

For the independent variable X_2 for “factor” which has 5 categories, we only set 4 dummy variables as $X_{2,1},\ X_{2,2}, X_{2,3}, X_{2,4}$. For the reference category will be employed when all dummy variables equal to zero.

$$X_{2,1} = 1, if\ "factor"\ =\ high;\ X_{2,1} = 0, otherwise$$

$$X_{2,2} = 1, if\ "factor"\ =\ moderate;\ X_{2,2} = 0, otherwise$$
$$X_{2,3} = 1, if\ "factor"\ =\ low;\ X_{2,3} = 0, otherwise$$
$$X_{2,4} = 1, if\ "factor"\ =\ unknown;\ X_{2,4} = 0, otherwise$$

For the independent variable X_3 for "age" which has 14 categories, we only set 13 dummy variables as $X_{3,1}, X_{3,2}, X_{3,3}, X_{3,4}, X_{3,5}, X_{3,6}, X_{3,7}, X_{3,8}, X_{3,9}, X_{3,10}, X_{3,11}, X_{3,12}, X_{3,13}$. For the reference category will be employed when all dummy variables equal to zero.

$$X_{3,1} = 1, if\ "age"\ =\ "0-4";\ X_{3,1} = 0, otherwise$$
$$X_{3,2} = 1, if\ "age"\ =\ "5-9";\ X_{3,2} = 0, otherwise$$
$$X_{3,3} = 1, if\ "age"\ =\ "10-14";\ X_{3,3} = 0, otherwise$$
$$X_{3,4} = 1, if\ "age"\ =\ "15-19";\ X_{3,4} = 0, otherwise$$
$$X_{3,5} = 1, if\ "age"\ =\ "20-24";\ X_{3,5} = 0, otherwise$$
$$X_{3,6} = 1, if\ "age"\ =\ "25-29";\ X_{3,6} = 0, otherwise$$
$$X_{3,7} = 1, if\ "age"\ =\ "30-34";\ X_{3,7} = 0, otherwise$$
$$X_{3,8} = 1, if\ "age"\ =\ "35-39";\ X_{3,8} = 0, otherwise$$
$$X_{3,9} = 1, if\ "age"\ =\ "40-44";\ X_{3,9} = 0, otherwise$$
$$X_{3,10} = 1, if\ "age"\ =\ "45-49";\ X_{3,10} = 0, otherwise$$
$$X_{3,11} = 1, if\ "age"\ =\ "50-54";\ X_{3,11} = 0, otherwise$$
$$X_{3,12} = 1, if\ "age"\ =\ "55-59";\ X_{3,12} = 0, otherwise$$
$$X_{3,13} = 1, if\ "age"\ =\ "60-64";\ X_{3,13} = 0, otherwise$$

After employing dummy variables into the original zero-inflated Poisson regression model, we have a model including dummy variables:

$$\lambda = exp(\ ln(t) + (dummy\ variable\ "hiv")\ +\ (dummy\ variables\ "factor") + (dummy\ variables\ "age") + \beta_4 X_4)$$

where, $dummy\ variable\ "hiv" = \beta_1 X_1$;

$dummy\ variables\ "factor" = \beta_{2,1} X_{2,1} + \beta_{2,2} X_{2,2} + \beta_{2,3} X_{2,3} + \beta_{2,4} X_{2,4}$;

$dummy\ variables\ "age" = \beta_{3,1} X_{3,1} + \beta_{3,2} X_{3,2} + \cdots + \beta_{3,13} X_{3,13}$.

Method of Maximum Likelihood

Logarithm of the likelihood function:

$$L = L1 + L2 - L3$$

where,

$L1 = \sum_{\{i:Y_i = 0\}} ln[\mu_i + exp(-\lambda_i)]$;

$L2 = \sum_{\{i:Y_i = 0\}} \{Y_i ln(\lambda_i) - \lambda_i - ln(Y_i!)\}$;

$L3 = \sum_{i = 0} ln(1 + \mu_i)$.

Maximum likelihood estimates:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} \sim N \left[\begin{matrix} \beta \\ \gamma \end{matrix} \middle| \left(\begin{matrix} -\frac{\partial^2 \mathcal{L}}{\partial \beta_r \, \partial \beta_s} & -\frac{\partial^2 \mathcal{L}}{\partial \beta_r \, \partial \gamma_s} \\ -\frac{\partial^2 \mathcal{L}}{\partial \beta_r \, \partial \gamma_s} & -\frac{\partial^2 \mathcal{L}}{\partial \gamma_r \, \partial \gamma_s} \end{matrix} \right)^{-1} \right]$$

5.3.1 Statistical Analysis

We have used SAS to test whether the dummy variables and its coefficients are significant and the results as below:

TABLE 5.2

Coefficients				
	Estimate	**Std. Error**	**Wald Chi-Square**	**Pr > ChiSq**
(Intercept)	-7.1740	0.6952	106.50	<.0001
py	0.0122	0.0067	3.28	0.0702
hiv1	3.5460	0.3569	98.70	<.0001
factor1	-0.4274	0.3166	1.82	0.1770
factor2	-0.9690	0.3900	6.17	0.0130
factor3	-0.9975	0.3691	7.30	0.0069
factor4	-0.3651	0.3542	1.06	0.3026

age1	-1.3157	0.7576	3.02	0.0825
age2	-0.2360	0.7163	0.11	0.7418
age3	-0.7759	0.5715	1.84	0.1746
age4	-0.9194	0.5472	2.82	0.0929
age5	-0.2604	0.5234	0.25	0.6188
age6	-1.1677	0.5286	4.88	0.0272
age7	-0.4561	0.6071	0.56	0.4524
age8	-0.2329	0.4463	0.27	0.6017
age9	-0.9137	0.5085	3.23	0.0723
age10	-1.4706	0.5624	6.84	0.0089
age11	-0.3709	0.5473	0.46	0.4980
age12	-0.7270	0.5714	1.62	0.2032
age13	-1.2624	0.6092	4.29	0.0382

Zero Inflation Coefficients				
	Estimate	**Std. Error**	**Wald Chi-Square**	**Pr > ChiSq**
(Intercept)	-10.3885	0.2054	2557.77	<.0001
hiv1	-21.5286	0.0000	.	.
Fit Statistics				
AIC			549.1380	
BIC			641.8594	

TABLE 5.2 This table shows the parameter estimation by using the zero-inflated Poisson regression model with the dummy variables. For the result, the p-values for independent variables of “hiv1”, "factor2", "factor3", "age6", "age10", "age13" are smaller than 5% level of significance. After the analysis, we can see that the model is having six independent variables that are significant, therefore, the performance is better than the previous regression models. Still, we can observe some of the independent variables are close to 5% level of significance such as "py", "age1", "age4", "age9", perhaps these variables could be significant after deploying the subset selection.

5.4 Diagnostic Checking and Model Selection

For the zero-inflated Poisson regression models, we would like to check if the zero-inflation problem exists in the model by comparing the number of observed zero in the dependent variable for the dataset and the number of predicted zero in the dependent variable. If the number of observed zero is larger than the number of predicted zero, it indicates the zero-inflation problem exists in the dataset.

TABLE 5.3

Check for zero-inflation	
Observed zeros	424
Predicted zeros	398

TABLE 5.3 From the above results, the dependent variable of the original data set has 424 zeros. The predicted zero value of the zero-inflated Poisson regression model is 398, which is about 94%. Therefore, the model is still insufficient to fit all of the excess zero.

Dispersion Checking

According to section 4.4, we found that the Poisson log-linear regression model has an overdispersion problem, which indicates that the zero-inflated Poisson regression model may also have an overdispersion problem.

TABLE 5.4

Pearson Chi-square dispersion statistic
0.9455683

TABLE 5.4 The Pearson Chi-square dispersion statistic indicates that if variance equals to mean, then the dispersion statistic value should be 1. The dispersion test reports that there is a slight underdispersion problem since the dispersion statistic is less than but approximately to 1.

Subset Selection

TABLE 5.5

Stepwise Selection			
Variable	**Parameter estimates**	**Chi-Square**	**Pr > ChiSq**
(Intercept)	-1.260307	29.3927	<0.001
py	0.013162	7.2428	0.0071
hiv1	-2.568592	52.8896	<0.001
Zero-Inflation Parameter Estimates			
Variable	**Parameter Estimates**	**Chi- Square**	**Pr > ChiSq**
(Intercept)	0.436935	4.6284	0.0314
Model Diagnostics			
AIC	538.97		
BIC	555.82		

TABLE 5.5 There are two estimations for zero-inflated models which are parameter estimates and zero-inflation parameter estimates. We have used stepwise selection for the subset selection. For the parameter estimates, the result of the stepwise selection only selects two independent variables "py" and "hiv1" for the model. For the zero-inflation parameter estimates, the stepwise selection does not select any variables except the intercept term. The values of AIC and BIC for stepwise selection are 538.97 and 555.82 respectively. Therefore, "py" and "hiv1" would be selected in zero-inflated Poisson regression models.

5.5 Concluding Remarks

According to the analysis, we have applied the zero-inflated Poisson regression models in order to control the excess-zero problems in data. The outcome expresses that the models can handle the excess-zero indeed, we can notice that as more independent variables are significant, the performance from estimation and subset selection has been improved. However, we have checked that the models are still insufficient to fit all of the excess zero and exist a slight overdispersion problem, therefore we could continue to another zero-inflated model namely zero-inflated negative binomial regression models to see if it is dealing with the problems of overdispersion and excessive zero clearly.

Problem of Overdispersion or Underdispersion

In many count datasets, zero-inflation and overdispersion (or underdispersion) co-exist. According to section 3.5, count data under Poisson assumption are usually overdispersed. Zero-inflation is one of the sources of overdispersion and traditional statistical methods are inappropriate to handle overdispersion.

From the section 3.5, the expectation and variance of negative binomial distribution are given below:

$$\begin{cases} E(Y) = \lambda \quad \lambda = 0 \\ Var(Y) = \lambda[1 + \alpha\lambda] \quad \alpha \geq 0, \lambda \geq 0 \end{cases}$$

Since $\alpha \geq 0, \lambda \geq 0$, therefore $Var(Y) > E(Y)$. Thus, it is an example to employ the negative binomial distribution for modeling data with overdispersion.

It can be shown that the expectation and variance of the zero-inflated negative binomial distribution are given by:

$$\begin{cases} E(Y) = (1 - \phi)\lambda \\ Var(Y) = E(Y)[1 + \lambda(1 + \alpha) - \mathrm{E(Y)}] \end{cases}$$

Zero-inflated negative binomial regression models may be an approach to deal with excessive zeros and overdispersion. According to section 3.5, the consequences of ignoring overdispersion may cause serious problems.

6. Zero-inflated Negative Binomial Regression Models

The zero-inflated negative binomial regression model (ZINB) is one of the zero-inflated models and also a more suitable model for dealing with overdispersion and excessive zeros than classical models. Similar to the zero-inflated Poisson regression model (Section 5), but there are some differences between them. The model generates two models and combines them at the end. (a logit model and a binomial model). Assuming two types of zero exist in the data. The first part is when the structure zero is generated by the existence of some special structure in the data; the other part is the sampling zero generated by the negative binomial distribution. (Xie et al., 2013)

6.1 Literature Review

A study paper is written by Lee, Park and Choi (2016) who offered to employ zero-inflated negative binomial modelling of the count data. This is about the exercise in periodical exercise and concerning elements in patients with Parkinson's disease. Finally, the Vuong test evidenced that if the count data is many-zero, then it indicates that zero-inflated negative binomial modelling is suitable in this study.

Pittman et al. (2018) appraised different models to analyze zero-inflated with over-dispersed count data from the Cigarette and the Marijuana. Under the goodness-of-fit statistics, several models can be suited for the count data. But the zero-inflated negative binomial model supported a better caption. Except for the statistics, the final choice between models should also depend on the null hypothesis, the research design and the raising of research questions.

On the other hand, Sharma and Landge (2013) claims that zero-inflated negative binomial regression can be applied in modelling critical car accident rates on rural India highways. In the result, researchers through the AIC, then confirmed the goodness of fit with the models.

6.2 Zero-inflated Negative Binomial Regression Models

The probability distribution of the zero-inflated negative binomial:

$$P(Y_i = j) = \begin{cases} \phi_i + (1-\phi_i)g(Y_i = 0) & \text{if } j = 0 \\ (1-\phi_i)g(Y_i) & \text{if } j > 0 \end{cases} \tag{6.1}$$

where,

$i = 1, 2, \cdots, n$;

$P(Y = Y_i)$ is the probability of Y_i occurrences;

$g(Y_i) = P(Y = Y_i \mid \lambda_i, \alpha) = \frac{\Gamma(Y+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(Y+1)}(\frac{1}{1+\alpha\lambda_i})^{\alpha^{-1}}(\frac{\alpha\lambda_i}{1+\alpha\lambda_i})^{Y}$, is the negative binomial distribution;

Γ is the gamma function used extend the factorial function;

$\mu_i = t_i\mu$;

μ is the mean incidence rate of y per unit of population size;

t_i is the population size for a specific observation;

$\alpha = \frac{1}{v}$;

α is the dispersion parameter;

v is the scale parameter.

$\phi_i = \frac{\mu_i}{1+\mu_i}$ is the logistic link function;

$\mu_i = exp(ln(t_i) + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_m Z_{im})$;

$\gamma_1, \gamma_2, \cdots, \gamma_m$ are m unknown regression coefficients;

$Z_1 + Z_2 + \cdots + Z_i$ are m independent variables.

The zero-inflated negative binomial regression model is the following:

$$\lambda_i = exp(ln(t_i) + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) \quad (6.2)$$

where, $i = 1, 2, \cdots, n$;

t_i is an exposure time for a specific observation;

$\beta_1, \beta_2, \cdots, \beta_k$ are k unknown regression coefficients;

$X_1, X_2, \cdots, X_k$ are k independent variables.

6.3 Model Building and Parameter Estimation

Same as zero-inflated Poisson regression models in section 5.2, we have found that the probability of zero count depends on the "hiv". We will also put the variable “hiv” to "zeromodel" part of the formula, which is $\phi_i + (1 - \phi_i)g(Y_i = 0)$ in formula (6.1), refer to section 6.2.

A zero-inflated negative binomial regression model of the hemophilia dataset is in the following form:

$$\lambda = exp(ln(t) + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 + \beta_4 X_4) \quad (6.3)$$

To avoid multicollinearity, we have discussed that we use dummy variables to identify qualitative variables by using the values 0 and 1. In the zero-inflated negative binomial regression model, therefore, we convert three independent variables "hiv", "factor" and "age", which are the qualitative variables, as the dummy variables.

The original zero-inflated negative binomial regression model without dummy variables from formula (6.3):

$$\lambda = exp(ln(t) + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 + \beta_4 X_4)$$

For the independent variable X_1 for “hiv” which has two categories, we only set one dummy variable as X_1. For the reference category will be employed when X_1 equals to zero.

$$X_1 = 1, if\ "hiv"\ =\ negative;\ X_1 = 0, otherwise$$

For the independent variable X_2 for “factor” which has 5 categories, we only set 4 dummy variables as $X_{2,1},\ X_{2,2}, X_{2,3}, X_{2,4}$. For the reference category will be employed when all dummy variables equal to zero.

$$X_{2,1} = 1, if\ "factor"\ =\ high;\ X_{2,1} = 0, otherwise$$
$$X_{2,2} = 1, if\ "factor"\ =\ moderate;\ X_{2,2} = 0, otherwise$$
$$X_{2,3} = 1, if\ "factor"\ =\ low;\ X_{2,3} = 0, otherwise$$
$$X_{2,4} = 1, if\ "factor"\ =\ unknown;\ X_{2,4} = 0, otherwise$$

For the independent variable X_3 for “age” which has 14 categories, we only set 13 dummy variables as $X_{3,1}, X_{3,2}, X_{3,3}, X_{3,4}, X_{3,5}, X_{3,6}, X_{3,7}, X_{3,8}, X_{3,9}, X_{3,10}, X_{3,11}, X_{3,12}, X_{3,13}$. For the reference category will be employed when all dummy variables equal to zero.

$$X_{3,1} = 1, if\ "age"\ =\ "0-4";\ X_{3,1} = 0, otherwise$$
$$X_{3,2} = 1, if\ "age"\ =\ "5-9";\ X_{3,2} = 0, otherwise$$

$$X_{3,3} = 1, if\ "age"\ =\ "10-14";\ X_{3,3} = 0, otherwise$$
$$X_{3,4} = 1, if\ "age"\ =\ "15-19";\ X_{3,4} = 0, otherwise$$
$$X_{3,5} = 1, if\ "age"\ =\ "20-24";\ X_{3,5} = 0, otherwise$$
$$X_{3,6} = 1, if\ "age"\ =\ "25-29";\ X_{3,6} = 0, otherwise$$
$$X_{3,7} = 1, if\ "age"\ =\ "30-34";\ X_{3,7} = 0, otherwise$$
$$X_{3,8} = 1, if\ "age"\ =\ "35-39";\ X_{3,8} = 0, otherwise$$
$$X_{3,9} = 1, if\ "age"\ =\ "40-44";\ X_{3,9} = 0, otherwise$$
$$X_{3,10} = 1, if\ "age"\ =\ "45-49";\ X_{3,10} = 0, otherwise$$
$$X_{3,11} = 1, if\ "age"\ =\ "50-54";\ X_{3,11} = 0, otherwise$$
$$X_{3,12} = 1, if\ "age"\ =\ "55-59";\ X_{3,12} = 0, otherwise$$
$$X_{3,13} = 1, if\ "age"\ =\ "60-64";\ X_{3,13} = 0, otherwise$$

After employing dummy variables into the original zero-inflated negative binomial regression model, we have a model including dummy variables:

$$\lambda = exp(\ ln(t) + (dummy\ variable\ "hiv")\ +\ (dummy\ variables\ "factor") + (dummy\ variables\ "age") + \beta_4 X_4)$$

where, $dummy\ variable\ "hiv" = \beta_1 X_1$;

$dummy\ variables\ "factor" = \beta_{2,1}X_{2,1} + \beta_{2,2}X_{2,2} + \beta_{2,3}X_{2,3} + \beta_{2,4}X_{2,4}$;

$dummy\ variables\ "age" = \beta_{3,1}X_{3,1} + \beta_{3,2}X_{3,2} + \cdots + \beta_{3,13}X_{3,13}$.

Method of Maximum Likelihood

Logarithm of the likelihood function:

$$L = L1 + L2 + L3 - L4$$

where,

$L1 = \sum_{\{i:Y_i = 0\}} ln[\mu_i + (1 + \alpha\lambda_i)^{-\alpha^{-1}}]$;

$L2 = \sum_{\{i:Y_i > 0\}} \sum_{j=0}^{Y_i - 1} ln(j + \alpha^{-1})$;

$L3 = \sum_{\{i:Y_i > 0\}} \{-ln(Y_i!) - (Y_i + \alpha^{-1}) ln(1 + \alpha\lambda_i) + Y_i ln(\alpha) + Y_i ln(\lambda_i)\}$;

$L4 = \sum_{i=1}^{n} ln(1 + \mu_i)$.

Maximum likelihood estimates:

$$\begin{bmatrix}\hat{\beta}\\ \hat{\gamma}\\ \alpha\end{bmatrix} \sim N\left[\begin{matrix}\beta\\ \gamma\\ \alpha\end{matrix}\left|\left(\begin{matrix}-\frac{\partial^2 \mathcal{L}}{\partial\beta_r\,\partial\beta_s} & -\frac{\partial^2 \mathcal{L}}{\partial\beta_r\,\partial\gamma_s} & -\frac{\partial^2 \mathcal{L}}{\partial\beta_r\,\partial\alpha}\\ -\frac{\partial^2 \mathcal{L}}{\partial\beta_r\,\partial\gamma_s} & -\frac{\partial^2 \mathcal{L}}{\partial\gamma_r\,\partial\gamma_s} & -\frac{\partial^2 \mathcal{L}}{\partial\gamma_s\,\partial\alpha}\\ -\frac{\partial^2 \mathcal{L}}{\partial\beta_r\,\partial\alpha} & -\frac{\partial^2 \mathcal{L}}{\partial\gamma_s\,\partial\alpha} & -\frac{\partial^2 \mathcal{L}}{\partial\alpha^2}\end{matrix}\right)^{-1}\right.\right]$$

6.3.1 Statistical Analysis

We have used SAS to test whether the dummy variables and its coefficients are significant and the results as belows:

TABLE 6.1

Coefficients				
	Estimate	**Std. Error**	**Wald Chi-Square**	**Pr > ChiSq**
(Intercept)	-1.8555	0.2271	66.75	<.0001
py	0.0122	0.0068	3.22	0.0726
hiv1	-3.5449	0.3751	89.29	<.0001
factor1	-0.4270	0.3192	1.79	0.1809
factor2	-0.9686	0.3923	6.09	0.0136
factor3	-0.9967	0.3767	7.00	0.0081
factor4	-0.3645	0.3600	1.03	0.3113
age1	-1.3160	0.7585	3.01	0.0827
age2	-0.2374	0.7312	0.11	0.7454
age3	-0.7764	0.5751	1.82	0.1770
age4	-0.9192	0.5484	2.81	0.0937
age5	-0.2613	0.5310	0.24	0.6227

age6	-1.1674	0.5300	4.85	0.0276
age7	-0.4574	0.6208	0.54	0.4612
age8	-0.2328	0.4472	0.27	0.6027
age9	-0.9132	0.5118	3.18	0.0743
age10	-1.4702	0.5646	6.78	0.0092
age11	-0.3703	0.5513	0.45	0.5018
age12	-0.7269	0.5722	1.61	0.2040
age13	-1.2619	0.6117	4.26	0.0391
(Dispersion)	0.0019	0.1634	/	/

Zero Inflation Coefficients				
	Estimate	**Std. Error**	**Wald Chi-Square**	**Pr > ChiSq**
(Intercept)	-5.3478	196.6732	0.00	0.9783
hiv1	-11.4440	393.3460	0.00	0.9768
Fit Statistics				
AIC			551.1381	
BIC			648.0741	

TABLE 6.1 The table above shows the parameter estimation by using the zero-inflated negative binomial regression model with the dummy variables. The result shows that the p-value of the independent variables "hiv1", "factor2", "factor3", "age6", "age10" and "age13" are smaller than 5% level of significance. These 6 independent variables are significant, as same as the result of the zero-inflated Poisson regression model in section 5.3.1. This regression model may have the same performance as the zero-inflated Poisson regression model.

6.4 Model Selection

Subset Selection

TABLE 6.2

Stepwise Selection			
Parameter Estimates			
Variable	**Parameter estimates**	**Chi-Square**	**Pr > ChiSq**
(Intercept)	-2.184731	110.8611	<0.0001
py	0.011664	4.1205	0.0424
hiv1	-2.500253	50.1846	<.0001
age8	0.804120	4.0293	0.0447
(Dispersion)	2.026588	/	/
Zero-Inflation Parameter Estimates			
Variable	**Parameter Estimates**	**Chi-Square**	**Pr > ChiSq**
Intercept	-11.516508	0.0002	0.9884
Model Diagnostics			
AIC	533.16		
BIC	558.45		

TABLE 6.2 There are two estimations for zero-inflated models which are parameter estimates and zero-inflation parameter estimates. We have used stepwise selection for the subset selection. For the parameter estimates, the result of the stepwise selection only selects two independent variables "py", "hiv1" and "age8" for the model. For the zero-inflation parameter estimates, the stepwise selection does not select any variables except the intercept term. The values of AIC and BIC for stepwise selection are 533.16 and 558.45 respectively. Therefore, "py", "hiv1" and "age8" would be selected in zero-inflated negative binomial regression models.

6.5 Concluding Remarks

According to our analysis, we have applied the zero-inflated negative binomial regression models in order to handle the excess-zero problem in zero-inflated models. For the result, the performance of AIC and BIC for zero-inflated negative binomial regression models are slightly better, whatever in estimation and subset selection. Although we can see that there is no major change after performing the zero-inflated negative binomial regression model models, the ability to deal with the excess-zero problems of models is still important, which is really beneficial for count data.

7. Model Evaluation

7.1 Model Comparison

Referring to the previous sections, we have totally conducted five regression models including traditional and advanced models. In this section, we would like to have a final comparison of these five models' results. To have a clearer picture for comparison, the selection criteria for AIC, BIC, the ratio of observed over predicted zeros and the Pearson Chi-square dispersion statistic would be adopted.

Testing with Proposed Selection Criteria

According to section 2.2.1, we calculate the AIC and BIC to evaluate and compare the models for finding out which one is the best fit for the data.

We would also use the Root Mean Square Error (RMSE) as the model comparison criteria. The RMSE is the square root of the Mean Square Error (MSE) which measures the average of the squares of the difference between observed values and estimated values. The formula is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (7.1)$$

where n is the number of observations, y_i and y_i are the observed values and estimated values for observation i. However, we often use the RMSE rather than MSE because RMSE is easier to interpret due to its value being lower than MSE.

According to table 2.15 from section 2.3, we conduct the stepwise selection method to select the best fit for the data.

According to dispersion checking from section 3.4, we perform the Pearson Chi-square dispersion statistic to see if the models have dispersion problems.

The following table is a model comparison of the fitting results among the models analysed in section 2 to section 6 by using the criteria of AIC, BIC, RMSE, the ratio of observed over predicted zeros and Pearson Chi-square dispersion statistic. For AIC, BIC and RMSE, if the information criterion value is smaller, then the model fitted should be better.

TABLE 7.1

Model	AIC	BIC	RMSE	Ratio of Observed / Predicted Zeros	Pearson Chi-square Dispersion Statistic
Multiple Linear Regression	1093.42	1181.93	0.629	/	/
Stepwise Selection	1069.86	1086.72	0.700	/	/
Poisson Log-linear Regression	591.26	675.55	0.677	0.95	1.474147
Stepwise Selection	575.40	592.26	0.687	0.95	1.520205
Negative Binomial Regression	552.58	641.09	0.683	1	0.9463652
Stepwise Selection	534.98	547.63	0.705	1	0.948298
Zero-inflated Poisson Regression	549.14	641.86	0.709	0.94	0.9455683
Stepwise Selection	538.97	555.82	0.705	0.94	1.117586
Zero-inflated Negative Binomial Regression	551.14	648.07	0.709	0.94	0.9470355
Stepwise Selection	533.16	558.45	0.683	0.99	1.05837

TABLE 7.1 The result shows that the negative binomial regression model with stepwise selection and zero-inflated negative binomial regression model with stepwise selection is the best out of all models, and the performance of these two models are very similar. About the details, first, we can see both of the models in the ratio of observed over predicted zeros are approaching. For negative binomial regression with stepwise selection is 1 which completely predicted the observed zeros, and for the zero-inflated negative binomial regression with stepwise selection is 0.99 which approximately predicted all the observed zeros. Second, the negative binomial regression model with stepwise selection provides the lowest value in BIC, but the zero-inflated negative binomial regression model with stepwise selection has the lowest

value in both AIC and RMSE. Third, both models have similar results from the Pearson Chi-square dispersion statistic. The dispersion test of the negative binomial regression model with stepwise selection is less than but approximately to 1. The dispersion test of the zero-inflated negative binomial regression with stepwise selection is larger than but approximately to 1. Both results indicate that they have no dispersion problem. Summarize the five selection criteria, the zero-inflated negative binomial regression model with stepwise selection should be the most fitted model.

Vuong Test

Vuong, Quang H. introduced the Vuong test in 1989 (Wikipedia contributors, 2020). The Vuong test is a hypothesis test that evaluates the best fit model between two non-nested models. Vuong test is similar to the likelihood ratio test which is also the same use as the Vuong test but is evaluating for two nested models. Two nested models mean that one model includes all terms from the other model and also includes some additional terms that the other model does not have. Two non-nested models mean that neither one of the models includes the terms from the other. In the procedures of the Vuong test, two non-nested models perform the hypothesis testing with the variance test statistic to test if both models are different. If the p-value is less than 5% level of significance, the two models are no different. Otherwise, we can proceed to the next step which performs another hypothesis testing with the non-nested likelihood ratio test statistic to test which model is the better fit. There are two alternative hypotheses, one indicates the first model is a better fit, another indicates the second model is a better fit. If either p-value of alternative hypotheses is less than 5% level of significance, the corresponding model is a better fit.

In the following **TABLE 7.2**, the Vuong non-nested test results for the count data. ML is the multiple linear regression model, Poisson is the Poisson regression model, NB is the negative binomial regression model, ZIP is the zero-inflated Poisson regression model, ZINB is the zero-inflated negative binomial regression model.

TABLE 7.2
Variance Test

Model (1) vs. Model (2)	Variance Test Statistic	P-value	Distinguishable / Indistinguishable
ML vs. Poisson	2.064	< 2e-16	Distinguishable
ML vs. NB	3.142	< 2e-16	Distinguishable
ML vs. ZIP	2.893	< 2e-16	Distinguishable
ML vs. ZINB	3.187	3.14e-09	Distinguishable

Poisson vs. NB	0.194	4.53e-07	Distinguishable
Poisson vs. ZIP	0.169	1.88e-06	Distinguishable
Poisson vs. ZINB	0.173	5.85e-07	Distinguishable
NB vs. ZIP	0.021	0.0398	Distinguishable
NB vs. ZINB	0.021	0.207	Indistinguishable
ZIP vs. ZINB	0.045	0.0226	Distinguishable

Distinguishable and Non-nested Likelihood Ratio Test

Model (1) vs. Model (2)	**Alternative Hypothesis**	**Non-nested Likelihood Ratio Test Statistic**	**P-value**	**Preferable Model**
ML vs. Poisson	ML fits better	-7.696	1	Poisson
	Poisson fits better	-7.696	7.014e-15	
ML vs. NB	ML fits better	-6.770	1	NB
	NB fits better	-6.770	6.425e-12	
ML vs. ZIP	ML fits better	-7.003	1	ZIP
	ZIP fits better	-7.003	1.257e-12	
ML vs. ZINB	ML fits better	-6.841	1	ZINB
	ZINB fits better	-6.841	3.936e-12	
Poisson vs. NB	Poisson fits better	-2.145	0.984	NB
	NB fits better	-2.145	0.01599	

Poisson vs. ZIP	Poisson fits better	-2.081	0.981	Both are equal fit
	ZIP fits better	-2.081	0.1872	
Poisson vs. ZINB	Poisson fits better	-2.777	0.997	ZINB
	ZINB fits better	-2.777	0.00274	
NB vs. ZIP	NB fits better	0.624	0.266	Both are equal fit
	ZIP fits better	0.624	0.7337	
NB vs. ZINB	/	/	/	/
	/	/	/	
ZIP vs. ZINB	ZIP fits better	-1.409	0.921	Both are equal fit
	ZINB fits better	-1.409	0.07947	

TABLE 7.2 From the variance test, it denotes that NB vs. ZINB is indistinguishable, thus we have no result at this time, but we can infer the result at the distinguishable and non-nested likelihood ratio test. Then, in the test, we can conclude that these three models, NB, ZIP, ZINB are relatively superior because they are all better fit than ML and Poisson. From NB vs. ZIP, and ZIP vs. ZINB, we can see they are both equally fitted but the p-values of NB and ZINB is smaller in both cases which are 0.266 and 0.07947, therefore, this shows NB and ZINB are better than ZIP. At last, we still do not have any final results.

However, regarding the result of variable selection of table 4.5 from section 4.4, the subset selection for the negative binomial regression models, we can find that only one independent variable "hiv1" is selected in the result. Compare this with the subset selection for zero-inflated negative binomial regression models in table 6.2 from section 6.4, which resulted in three independent variables "py", "hiv1" and "age8". Based on the principle of parsimony, the least independent variable, the better performance of the models. We can notice that only one independent variable "hiv1" is selected to the negative binomial regression model with stepwise selection, which is the smallest number of independent variables in all models, then it would be the best model fitting for the hemophilia dataset.

8. Discussion

8.1 Research Questions Revisited

Regarding to section 1.2, our research questions are:

1. How do we choose the appropriate regression model for the dataset?

2. What is the relationship between the number of deaths and the factors?

For the first question, we have noticed that the dependent variable has too many zeros, and then discovered that it actually has the characteristics of count data from the beginning. Since there are three categorical independent variables, according to the literature, we also decided to use dummy variables to help the research. It can be seen that we always select models from the features of the dataset and test them one by one. From classic models to advanced models to testing until we find the most appropriate regression model for the dataset.

For the second question, we can obtain the relationship from the final model (negative binomial regression model):

$$deaths = exp(-1.953766 - 2.410334(hiv1))$$

We can see that only the dummy variable of "hiv1" has a significant negative correlation with death in the result of the model above. It indicates that only the "hiv1" variable has a better relationship with "deaths". That is to say the other independent variables may not be important for the number of deaths. In other words, if the patient is HIV-positive, then the chance of death will be very high, so that the number of deaths will have a great chance of increasing.

9. Conclusion

9.1 Overall Summary

From the beginning of the project, we first tried to fit the dataset into the multiple linear regression models, but the result shows that the multiple linear regression models might not be good for the hemophilia dataset. After reading some relevant literature, we recognized our characteristics of the hemophilia dataset and started adopting different regression models that are suitable for our dataset. According to the model comparison, the negative binomial regression model with the stepwise selection is the best model. We removed most of the original independent variables from the model. We only keep the independent variable "hiv1" to perform the best model.

9.2 Limitations and Future Directions

This project is limited by several aspects. The first limitation has been limited by the number of testing models. Except for the zero-inflated models, we do need to try different advanced statistical models (e.g. Hurdle models) because there may be a more suitable model we do not know. However, the time is limited, we need more time to become familiar with the advanced statistical model's theory and coding, we only performed zero-inflated models in the end.

Besides that, the second limitation comes from the dataset. There are only four independent variables in the dataset. Perhaps, there is some sensitive information that cannot be disclosed in the medical documents. In multiple linear regression models, the R-square is pretty low, only 11.78% variation of the dependent variable is explained by the independent variables. It indicates that there are some more variables that affect the dependent variable but not included in the dataset. This is also one of the limitations that we faced. Due to the above limitations, we have deliberated some ideas of improvement in the future.

For the first limitation, we have tried to understand and test all the proposed models, from traditional simple linear regression models to advanced statistical models. Since the complexity of the models, which consumed a lot of time in the process. Limited by our knowledge and investigation, in fact, the result we have obtained probably is not the best, there might be some other models that are more suitable for the dataset such as the Poisson-Inverse Gaussian (PIG) regression model. This model is similar to the negative binomial model that we have used because both of them are mixture models. Moreover, they can handle the distribution from the right skew and the count data. Thus, we should try to conduct Poisson-Inverse Gaussian models if possible.

For the second limitation, the dataset should include more observations of the patient and also include other relatable factors to the study, such as the gender of the patient and the time during the disease of the patient. Moreover, we notice that one of the factors "age" has too many groups which may cause the problem of excessive 0 and dummy variables are not significant for the model fitting, this factor can reassign into four categories which contain "Children", "Teenagers", "Adults" and "Elderly", we may get a better result for the model fitting. We will study more theory and applications about the advanced regression models in the future and try to apply that knowledge in this report.

Appendix

SAS code:

```
/* Read file  */
options validvarname=v7;
libname hemo xlsx "C:\Users\user\hemo_excel_r.xlsx";
ods graphics on;

data work.hemo;
	set hemo.Sheet1;
run;

/* Generates dummy variable */
data work.dum;
	set work.hemo;

if hiv = 1 then hiv1 = 1/2; else hiv1 = -1/2;
if factor = 1 then factor1  = 4/5; else factor1 = -1/5;
if factor = 2 then factor2  = 4/5; else factor2 = -1/5;
if factor = 3 then factor3  = 4/5; else factor3 = -1/5;
if factor = 4 then factor4  = 4/5; else factor4 = -1/5;
if age = 1 then age1 = 13/14; else age1 = -1/14;
if age = 2 then age2 = 13/14; else age2 = -1/14;
if age = 3 then age3 = 13/14; else age3 = -1/14;
if age = 4 then age4 = 13/14; else age4 = -1/14;
if age = 5 then age5 = 13/14; else age5 = -1/14;
if age = 6 then age6 = 13/14; else age6 = -1/14;
if age = 7 then age7 = 13/14; else age7 = -1/14;
if age = 8 then age8 = 13/14; else age8 = -1/14;
if age = 9 then age9 = 13/14; else age9 = -1/14;
if age = 10 then age10 = 13/14; else age10 = -1/14;
if age = 11 then age11 = 13/14; else age11 = -1/14;
if age = 12 then age12 = 13/14; else age12 = -1/14;
if age = 13 then age13 = 13/14; else age13 = -1/14;

run;

/* plot, vif, collin */
proc reg data=work.dum plots=all;
model deaths = py hiv1
                                factor1 factor2 factor3 factor4
                                age1 age2 age3 age4 age5 age6 age7
                                age8 age9 age10 age11 age12 age13 /clb spec collin vif;
run;
```

```
/* Multiple linear regression model with dummy */
proc genmod data=work.dum plots=all;
  model deaths = py hiv1
                        factor1 factor2 factor3 factor4
                        age1 age2 age3 age4 age5 age6 age7
                        age8 age9 age10 age11 age12 age13 /dist=normal vif collin;
quit;

/* Subset selection for Multiple linear regression model with dummy*/
proc genselect data=work.dum;
model deaths = py hiv1
                              factor1 factor2 factor3 factor4
                              age1 age2 age3 age4 age5 age6 age7
                              age8 age9 age10 age11 age12 age13
        /Distribution=normal;
  selection method=stepwise details=all;

run;

/* Poisson regression model with dummy */
proc genmod data=work.dum;
  model deaths = py hiv1
                        factor1 factor2 factor3 factor4
                        age1 age2 age3 age4 age5 age6 age7
                        age8 age9 age10 age11 age12 age13 /dist=poisson link=log;
quit;

/* Subset selection for Poisson regression model with dummy*/
proc genselect data=work.dum;
model deaths = py hiv1
                              factor1 factor2 factor3 factor4
                              age1 age2 age3 age4 age5 age6 age7
                              age8 age9 age10 age11 age12 age13
        /Distribution=Poisson;
  selection method=stepwise details=all;

run;

/* Negative binomial regression model with dummy */
proc genmod data=work.dum;
model deaths = py hiv1
                        factor1 factor2 factor3 factor4
                        age1 age2 age3 age4 age5 age6 age7
                        age8 age9 age10 age11 age12 age13 /link=log dist=negbin;
run;
/*Subset selection for Negative binomial regression model with dummy*/
proc genselect data=work.dum;
model deaths = py hiv1
```

```
                                    factor1 factor2 factor3 factor4
                                    age1 age2 age3 age4 age5 age6 age7
                                    age8 age9 age10 age11 age12 age13
        /Distribution=negativebinomial link=log;
  selection method=stepwise details=all;

run;

/* Zero-inflated Poisson regression model with dummy */
proc genmod data=work.dum;
        model deaths = py hiv1 factor1 factor2 factor3 factor4
                                    age1 age2 age3 age4 age5 age6 age7
                                    age8 age9 age10 age11 age12 age13 /dist=zip link=log;
        zeromodel hiv1 /link=logit;
run;

/* Subset selection for Zero-inflated Poisson regression model with dummy */
proc hpgenselect data=work.dum;
        model deaths = hiv1 py
                factor1 factor2 factor3 factor4
                age1 age2 age3 age4 age5 age6 age7
                age8 age9 age10 age11 age12 age13 /dist=zip link=log;
                zeromodel hiv1 /link=logit;
  selection method=stepwise details=all;
run;


/* Zero-inflated binomial regression model with dummy */
proc genmod data=work.dum;
        model deaths = py hiv1 factor1 factor2 factor3 factor4
                                    age1 age2 age3 age4 age5 age6 age7
                                    age8 age9 age10 age11 age12 age13 /dist=zinb link=log;
        zeromodel hiv1 /link=logit;
run;

/*  Subset selection for Zero-inflated binomial regression model with dummy */

proc hpgenselect data=work.dum;
        model deaths = py hiv1 factor1 factor2 factor3 factor4
                                    age1 age2 age3 age4 age5 age6 age7
                                    age8 age9 age10 age11 age12 age13 /dist=zinb link=log;
                zeromodel hiv1  /link=logit;
  selection method=stepwise details=all;
run;
```

R code:

Library
```{r load pack}
library(fastDummies)
library(readxl)
library(VIF)
library(car)
library(pedometrics)
library(ggpubr)
library(ISLR)
library(ggplot2)
theme_set(theme_pubr())
library(vcd)
library(pscl)
library(epiDisplay)
library(nonnest2)
library(spdep)
library(AER)
library(performance)
```

Data fetching
```{r data}

Mydata = read_excel("C:\\Users\\LSK\\OneDrive - The Open Unviersity of Hong Kong\\DS project\\FYP_program\\hemo_excel_r.xlsx", sheet = "Sheet1")

head(Mydata)
str(Mydata)
dim(Mydata)
```

Dummy
```{r dummy}
Mydata_dummy <- dummy_cols(Mydata, select_columns = c('hiv', 'factor', 'age'))
str(Mydata_dummy)
Mydata_new = Mydata_dummy[,c(4, 5, 6, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25)]
str(Mydata_new)
```

Percent of deaths
```{r Percent}
require(lattice)
require(openintro)
```

```
histogram(Mydata_new$deaths,col = "darkslategray3", xlab = "deaths", main = "The
percentage of deaths")
```
multiple linear regression
```{r multiple linear regression}
MLR <- lm(deaths ~ . , data = Mydata_new)
summary(MLR)
```
poisson regression
```{r poisson regression}

summary(Mydata_new$deaths)
mean(Mydata_new$deaths)
var(Mydata_new$deaths)

PR <- glm(formula = deaths ~ ., data = Mydata_new, family = poisson(link = "log"))
summary(PR)
check_overdispersion(PR)
```
NB
```{r nb}
NB = glm.nb(deaths ~ . , data = Mydata_new, link=log)
summary(NB)
```

ZIP
```{r zip}
zip = zeroinfl(deaths ~ . | hiv_1,
               dist = 'poisson',
               link = 'logit',
               EM = FALSE,
               data = Mydata_new)
summary(zip)
```
ZINB
```{r zinb}
zinb = zeroinfl(deaths ~ . | hiv_1,
               dist = 'negbin',
               data = Mydata_new)
summary(zinb)
```

Fit model using SAS stepwise output
```{r fitmodel}

MLR.fit <- lm(deaths ~ hiv_1 + age_8 , data = Mydata_new)
summary(MLR.fit)
```

```
PR.fit <- glm(formula = deaths ~ py + hiv_1  + age_8, data = Mydata_new, family =
poisson(link = "log"))
summary(PR.fit)

NB.fit = glm.nb(deaths ~ py + hiv_1, link=log, data = Mydata_new)
summary(NB.fit)

zip.fit = zeroinfl(deaths ~ py + hiv_1,
            dist = 'poisson',
            link = 'logit',
            data = Mydata_new)
summary(zip.fit)

zinb.fit = zeroinfl(deaths ~ py + hiv_1 + age_8,
            dist = 'negbin',
            link = 'logit',
            data = Mydata_new)
summary(zinb.fit)
```

Comparison
```{r comparing}
#compare_performance(MLR,PR,NB,zip,zinb)
#compare_performance(MLR.fit,PR.fit,NB.fit,zip.fit, zinb.fit)

#RMSE
rmse(MLR.fit)
rmse(PR.fit)
rmse(NB.fit)
rmse(zip.fit)
rmse(zinb.fit)

#LRT
lrtest(MLR.fit, PR.fit)
lrtest(MLR.fit, NB.fit)
lrtest(MLR.fit, zip.fit)
lrtest(MLR.fit, zinb.fit)
lrtest(PR.fit, NB.fit)
lrtest(PR.fit, zip.fit)
lrtest(PR.fit, zinb.fit)
lrtest(NB.fit, zip.fit)
lrtest(NB.fit, zinb.fit)
lrtest(zip.fit, zinb.fit)

# Voung test
vuongtest(MLR.fit, PR.fit)
vuongtest(MLR.fit, NB.fit)
vuongtest(MLR.fit, zip.fit)
```

```
vuongtest(MLR.fit, zinb.fit)
vuongtest(PR.fit, NB.fit)
vuongtest(PR.fit, zip.fit)
vuongtest(PR.fit, zinb.fit)
vuongtest(NB.fit, zip.fit)
vuongtest(NB.fit, zinb.fit)
vuongtest(zip.fit, zinb.fit)

# check zero
check_zeroinflation(PR)
check_zeroinflation(NB)
check_zeroinflation(zip)
check_zeroinflation(zinb)
check_zeroinflation(PR.fit)
check_zeroinflation(NB.fit)
check_zeroinflation(zip.fit)
check_zeroinflation(zinb.fit)
```

References

1. Andrzejczak, K., Mlynczak, M., Selech, J. (2018). poisson-distributed failures in the predicting of the cost of corrective maintenance: Semantic scholar. Retrieved from https://www.semanticscholar.org/paper/Poisson-distributed-failures-in-the-predicting-of-Andrzejczak-Mlynczak/01d826ca78b015c6553c1f50178fc396df5e175c
2. An, Q., Wu, J., Fan, X., Pan, L., Sun, W. (2016). Using a negative binomial regression model for early warning at the start of a Hand Foot mouth disease epidemic in Dalian, Liaoning Province, China. Retrieved from https://pubmed.ncbi.nlm.nih.gov/27348747/
3. Byers, A., Allore, H., Gill, T., Peduzzi, P. (2003) Application of Negative Binomial Modeling for Discrete Outcomes. Retrieved from https://www.researchgate.net/publication/6508211_Application_of_Negative_Binomial_Modeling_for_Discrete_Outcomes
4. Centers for Disease Control and Prevention (2020). What is Hemophilia?. Retrieved from https://www.cdc.gov/ncbddd/hemophilia/facts.html
5. Chou, W. C., Wu, J. L., Wang, Y. C., Huang, H., Sung, F. C., Chuang, C. Y. (2010). Modeling the impact of climate variability on diarrhea-associated diseases in Taiwan (1996–2007). Retrieved from http://www.climateknowledge.org/Food_Water_Illness_Models/Chou_Model_Diarrhea_Taiwan_SciTotalEnviron_2010.pdf
6. Doyle, J., Bottomley, P. (2019). The relative age effect in European Elite soccer: A practical guide to Poisson REGRESSION MODELLING. Retrieved from https://pubmed.ncbi.nlm.nih.gov/30943241/
7. Ferreira, A. A., Leite, I. C. G., Bustamante-Teixeira, M. T., Corrêa, C. S. L., Cruz, D. T. D., Rodrigues, D. D. O. W., & Ferreira, M. C. B. (2013). Health-related quality of life in hemophilia: results of the Hemophilia-Specific Quality of Life Index (Haem-a-Qol) at a Brazilian blood center. Revista brasileira de hematologia e hemoterapia, 35(5), 314-318.
8. Hassett, M., McGee, G. (2017). Negative binomial hurdle models to estimate flower production for native and nonnative northeastern shrub taxa. Retrieved from https://academic.oup.com/forestscience/article/63/6/577/4772564?login=true
9. Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), 1-14.
10. Lee, J., Park, C. G., & Choi, M. (2016). Regular exercise and related factors in patients with Parkinson's disease: Applying zero-inflated negative binomial modeling of exercise count data. Applied nursing research : ANR, 30, 164–169. https://doi.org/10.1016/j.apnr.2015.08.002
11. McCullagh, P., Nelder, J. A. (1989). Generalized Linear Models. London: Chapman & Hall / CRC.
12. Nie, L., Wu, G., Brockman, F. J., & Zhang, W. (2006). Integrated analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: zero-inflated Poisson regression models to predict abundance of undetected proteins. *Bioinformatics (Oxford, England)*, *22*(13), 1641–1647. https://doi.org/10.1093/bioinformatics/btl134

13. Payal, V., Sharma, P., Goyal, V., Jora, R., Parakh, M., & Payal, D. (2016). Clinical profile of hemophilia patients in Jodhpur Region. Asian journal of transfusion science, 10(1), 101.
14. Pittman, B., Buta, E., Krishnan-Sarin, S., O'Malley, S., Liss, T., Gueorguieva, R. (2018, April 18). Models for analyzing zero-inflated and OVERDISPERSED count data: An application to cigarette and marijuana use. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7364829/
15. Poston, D., McKibben, S. (2003). Using Zero-inflated Count regression models to estimate the fertility of U. S. WOMEN. Retrieved from https://digitalcommons.wayne.edu/jmasm/vol2/iss2/10/
16. Sarul, L., Sahin, S. (2015). AN APPLICATION OF CLAIM FREQUENCY DATA USING ZERO INFLATED AND HURDLE MODELS IN GENERAL INSURANCE. Retrieved from https://dergipark.org.tr/tr/download/article-file/374499
17. Sharma, A. K., Landge, V. S. (2013). ZERO INFLATED NEGATIVE BINOMIAL FOR MODELING HEAVY VEHICLE CRASH RATE ON INDIAN RURAL HIGHWAY. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.436.7501&rep=rep1&type=pdf
18. Somo-Aina, O., Gayawan, E. Structured additive distributional hurdle Poisson modelling of individual fertility levels in Nigeria. Genus 75, 20 (2019). https://doi.org/10.1186/s41118-019-0067-9
19. Stonebraker, J., Bolton-Maggs, P., Brooker, M., Evatt, B., Iorio, A., Makris, M., . . . Tootoonchian, E. (2020). The World Federation of Hemophilia Annual Global Survey 1999-2018. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/hae.14012
20. Voss, T. S., Elm, J. J., Wielinski, C. L., Aminoff, M. J., Bandyopadhyay, D., Chou, K. L., Sudarsky, L. R., Tilley, B. C., & Falls Writing Group NINDS NET-PD Investigators (2012). Fall frequency and risk assessment in early Parkinson's disease. Parkinsonism & related disorders, 18(7), 837–841. https://doi.org/10.1016/j.parkreldis.2012.04.004
21. Wikipedia contributors. (2020, May 16). Vuong's closeness test. In Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/w/index.php?title=Vuong%27s_closeness_test&oldid=956971934
22. Xie, H., Tao, J., McHugo, G. J., & Drake, R. E. (2013). Comparing statistical methods for analyzing skewed longitudinal count data with many zeros: An example of smoking cessation. Journal of substance abuse treatment, 45(1), 99-108.
23. Yesilova, A., Kaya, Y., Kaki, B., & Kasap, İ. (2010). Analysis of plant protection studies with excess zeros using zero-inflated and negative binomial hurdle models. Gazi University Journal of Science, 23(2), 131-136.