STAT S461F - Data Science Project
2020-2021

Final Presentation

Project Title: Statistical Modelling of Hemophilia Count Data

by
Lai Siu Kwok (Team Leader)
Zheng Zequan
Hung Chun Kwong
Shan Youming
Lo Shi Sam
Pei Zixuan

Supervisor: Dr. Tony Chan
Date: May 2021

Research Background

- Hemophilia was one of the serious diseases in the world
  - As of 2018, there were about three hundred thousand people with Hemophilia (Stonebraker et al., 2020)

- Statistical modelling
  - Inferring any relationships between variables
  - Mining out the hidden information
  - A mathematical product and expression from the dataset
  - Visualization technique for presentation (Stobierski et al., 2019)

- In this project, we would like to study Hemophilia dataset in statistical way and try to apply different statistical modelling methods for the data.

Research Objective and Questions

Objectives:

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

Questions:

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

Variables Description

- The study was conducted with 500 groups of patients with Hemophilia

- The study focuses on the number of deaths from each group and the other four factors which may affect the risk of death
  - including HIV status, clotting medicine dose, age group and time of participation

Variables

	Variable name	Property	Description
Dependent	deaths	Numeric	The number of deaths in 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.

Agenda

1. Rejected simple linear regression model from proposal
2. Conduct a **multiple linear regression model**
    - 2.1. Whether the dependent variable is a count data and the assumptions are hold
    - 2.2. Multiple linear regression may not be appropriate
3. Conduct a **Poisson log-linear regression model**
    - 3.1. Whether overdispersion exists and the assumptions are hold
    - 3.2. Poisson log-linear regression may not be appropriate

We would revisit the multiple linear regression model and Poisson log-linear regression model from the previous presentation. Some of the contents are improved.

Agenda

4. Conduct a **negative binomial regression model**
   - 4.1. Whether excessive zeros in the data
   - 4.2. Negative binomial regression may be appropriate
5. Conduct a **zero-inflated Poisson regression model**
   - 5.1. Whether excessive zeros have solved and overdispersion exists
   - 5.2. zero-inflated Poisson regression may be appropriate
6. Conduct a **zero-inflated negative binomial regression model**
   - 6.1. Whether excessive zeros have solved
   - 6.2. zero-inflated negative binomial regression may be appropriate
7. Model Comparison

In addition to discussing the above models, we would like to answer our research questions and make a final conclusion in this presentation

Dummy Variables
&
Multiple Linear Regression Models

Reasons of Using Multiple Linear Regression Model

- Based on the findings
  - The dependent variable “deaths” is a count data
  - Multiple linear regression model may not be used
- Still employ multiple linear regression model
  - To see whether using multiple linear regression can fit the dataset as well
- If multiple linear regression model is not fit
  - We may use another model to deal with count data

Multiple Linear Regression Model - Assumptions

- Normality
  - The dependent variable should follow normal distribution
- No multicollinearity
  - The independent variables have no high correlation with other independent variables
- Heterogeneity of variance for residuals
  - The residuals should have a constant variance

- Model adequacy checking will be shown in the following slides

Dummy Variables

- Regression models
  - Normally, independent variables regarded as quantitative variables
- When the model includes qualitative variables
  - Employ dummy variables to identity qualitative variables
  - Use values 0 and 1
  - Represent different subgroups of categorical independent variables
- A nominal or ordinal variable with a total number of k categories should be introduced to (k-1) dummy variables
- The omitted category is called the "reference category"
- The number of dummy variables cannot be introduced up to k categories
  - Otherwise multicollinearity will occur.

Dummy Variables

- A qualitative variable cannot be directly used as an independent variable for regression analysis
  - Needs to be converted into a dummy variable
- The variable ”factor”, “age” and “hiv” are qualitative variables
  - They need to converted into dummy variables
- The table shows part of the dummy variable combinations
- if “factor” = 1, then “factor” will equal to “factor1”, then “factor2”, “factor3” and “factor4” will equal to 0

							dummy variable (factor)			
obs	hiv	factor	age	py	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
.	.	.	.	.	.		.	.	.	.
.	.	.	.	.	.		.	.	.	.
.	.	.	.	.	.		.	.	.	.

- **We will employ dummy variables to all regression models in this project**
  - **Let show the example of employ the dummy variables to the multiple linear regression model**

Dummy Variables

For the multiple linear regression model:
Original model equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ **(X1: “hiv”, X2: “factor”, X3: “age”, X4: “py”)**

- **For predictor variable X_1 for “hiv”, (2 categories, 1 dummy variable)**

  $X_1 = 1$, if negative; $X_1 = 0$, otherwise.

- **For predictor variable X_2 for “factor”, (5 categories, 4 dummy variables)**

  $X_{2,1} = 1$, if high; $X_{2,1} = 0$, otherwise.

  $X_{2,2} = 1$, if moderate; $X_{2,2} = 0$, otherwise.

  $X_{2,3} = 1$, if low; $X_{2,3} = 0$, otherwise.

  $X_{2,4} = 1$, if unknown; $X_{2,4} = 0$, otherwise.

Dummy Variables

For the multiple linear regression model:
Original model equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ **(X1: "hiv", X2: "factor", X3: "age", X4: "py")**

- **For predictor variable X_3 for "age", (14 categories, 13 dummy variables)**

$X_{3,1} = 1$, if "0 - 4"; $X_{3,1} = 0$, otherwise.

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

$X_{3,3} = 1$, if "10 - 14"; $X_{3,3} = 0$, otherwise.

$X_{3,4} = 1$, if "15 - 19"; $X_{3,4} = 0$, otherwise.

$X_{3,5} = 1$, if "20 - 24"; $X_{3,5} = 0$, otherwise.

$X_{3,6} = 1$, if "25 - 29"; $X_{3,6} = 0$, otherwise.

$X_{3,7} = 1$, if "30 - 34"; $X_{3,7} = 0$, otherwise.

$X_{3,8} = 1$, if "30 - 34"; $X_{3,8} = 0$, otherwise.

$X_{3,9} = 1$, if "35 - 39"; $X_{3,9} = 0$, otherwise.

$X_{3,10} = 1$, if "40 - 44"; $X_{3,10} = 0$, otherwise.

$X_{3,11} = 1$, if "45 - 49"; $X_{3,11} = 0$, otherwise.

$X_{3,12} = 1$, if "50 - 54"; $X_{3,12} = 0$, otherwise.

$X_{3,13} = 1$, if "55 - 59"; $X_{3,13} = 0$, otherwise.

Dummy Variables

- For the original multiple linear regression model:
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ **(X1: “hiv”, X2: “factor”, X3: “age”, X4: “py”)**
- After employ dummy variables:
  - $Y = \beta_0 +$ (dummy variable "hiv") + (dummy variables "factor") + (dummy variables "age") $+ \beta_4 X_4 + \varepsilon$

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

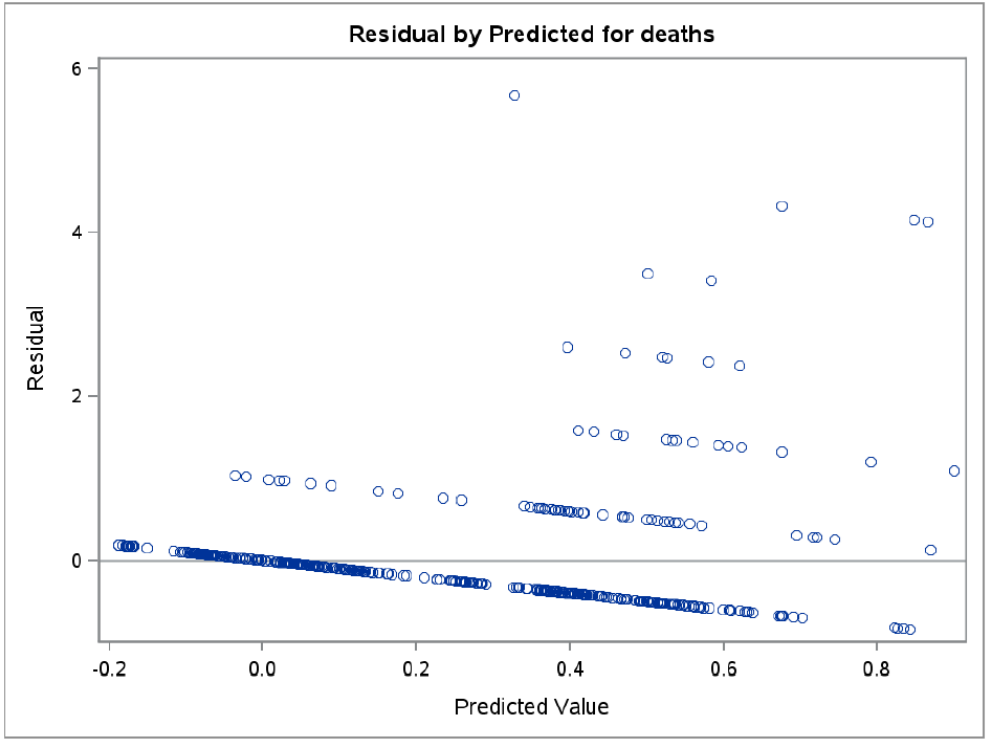
    dummy variables "factor" = $\beta_{2,1} X_{2,1} + \beta_{2,2} X_{2,2} + \beta_{2,3} X_{2,3} + \beta_{2,4} X_{2,4}$;

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

Diagnostic Plot - Constant Variance of Residual

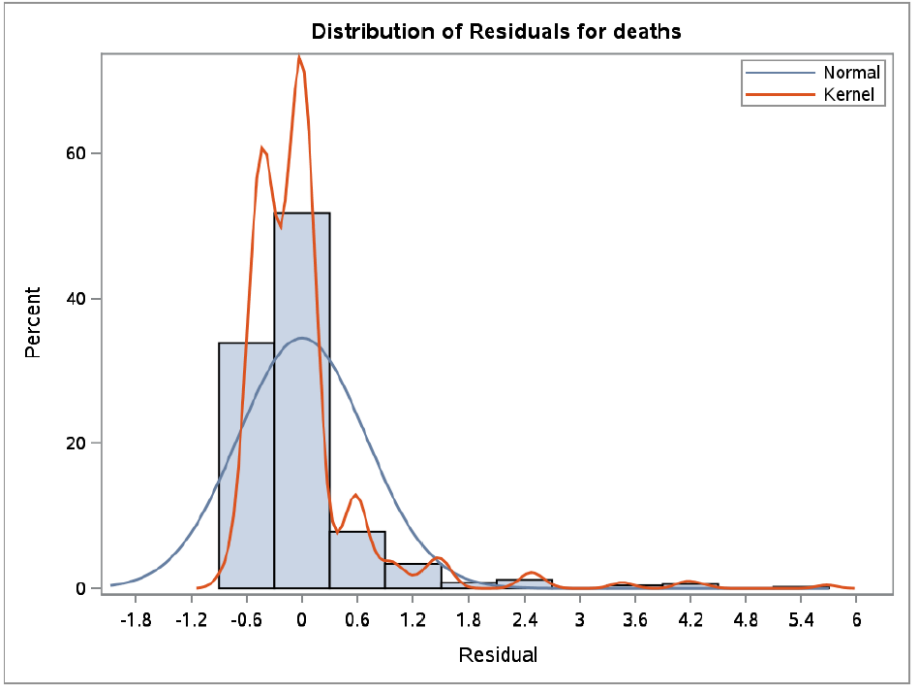
- The graph
  - If the residuals fluctuate in a random fashion inside the band
    - The variance of residuals is constant
  - Otherwise
    - The variance of residuals is non-constant

- Result
  - Lots of residuals fluctuate is not in a random fashion inside the band
    - It is an outward opening funnel
  - The variance of residuals is non-constant



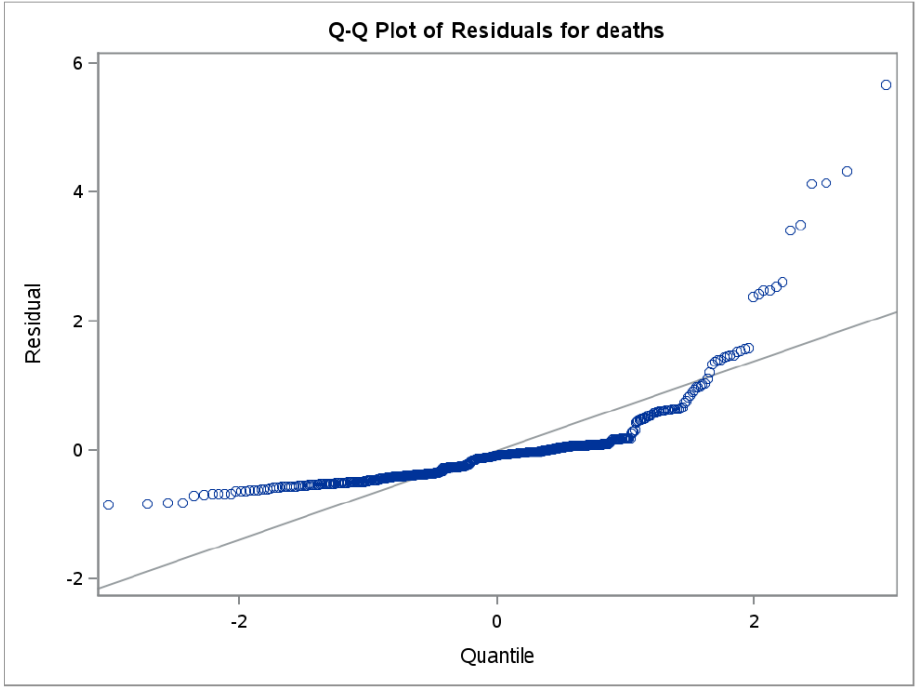
Diagnostic Plot - Normal Distribution Assumption Checking

- The graph
  - About the distribution of residuals for death
  - Blue curve is the normal curve
  - Orange curve is the kernel curve
- Compare with the two curves
  - Kernel curve does not follow the normal curve
- Result
  - The residual is not follow a normal distribution



Diagnostic Plot - Normal Distribution Assumption Checking

- The graph
  - If the residuals is normally distributed the points should very close to the line of the perfect fit (straight line)
- Result
  - The normality of residuals may not hold, the points do not lie on the line of the perfect fit



Normal Distribution Assumption Checking - Shapiro-Wilk Normality Test

Analysis of Shapiro-Wilk normality test:

- H0: The data is follow normal distribution
- H1: The data is not follow normal distribution

```
##  Shapiro-Wilk normality test
##
## data:  residual
## W = 0.65915, p-value < 2.2e-16
```

- W = 0.65915
- p-value = < 2.2e-16 <0.05, which indicates that the data does not obey the normal population.

Multiple Linear Regression Model - Correlation

- For the F-test, the p-value is less than 0.0001, which is < 5% significant level, there is a linear relationship between independent variables and the dependent variable

- For the coefficient of determination, R-Square, which indicates that only 11.78% of the total variation in Y can be explained by the fitted regression model.

- For the Adjusted R-Square, which reflects that only 8.29% of the total variation in Y can be explained by the fitted regression model.

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			

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

Multiple Linear Regression Model - Parameter Estimates

- Finding
  - For 5% level of p-value
  - Only “hiv1” is statistically significant

- Summary
  - Multiple linear regression model may not be satisfied

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.23043	0.03944	5.84	<.0001
py	py	1	0.00203	0.00182	1.11	0.2657
hiv1		1	-0.44660	0.06440	-6.93	<.0001
factor1		1	-0.00005864	0.10184	-0.00	0.9995
factor2		1	-0.13883	0.10328	-1.34	0.1795
factor3		1	-0.13221	0.10152	-1.30	0.1934
factor4		1	-0.05148	0.10341	-0.50	0.6188
age1		1	-0.26912	0.19288	-1.40	0.1636
age2		1	-0.20855	0.18047	-1.16	0.2484
age3		1	-0.21156	0.18828	-1.12	0.2617
age4		1	-0.18329	0.18165	-1.01	0.3135
age5		1	-0.10022	0.18462	-0.54	0.5875
age6		1	-0.18307	0.18679	-0.98	0.3275
age7		1	-0.18122	0.17960	-1.01	0.3135
age8		1	0.15026	0.18363	0.82	0.4136
age9		1	-0.16971	0.18141	-0.94	0.3500
age10		1	-0.26941	0.17769	-1.52	0.1301
age11		1	-0.10538	0.18942	-0.56	0.5783
age12		1	-0.15741	0.18362	-0.86	0.3917
age13		1	-0.27838	0.19165	-1.45	0.1470

Multiple Linear Regression Model - Multicollinearity Checking

- Multicollinearity
  - A strong correlation between independent variables
- Variance Inflation Factor (VIF)
  - The degree of variance of the estimator inflated by multicollinearity
  - VIF = 1, no multicollinearity
  - 1 < VIF < 10, multicollinearity exists but not severe
  - VIF ≥ 10, multicollinearity exists
- Result
  - None of the VIF ≥ 10
  - And they are all close to 1
  - Multicollinearity does not exist

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	0.23043	0.03944	5.84	<.0001	0
py	py	1	0.00203	0.00182	1.11	0.2657	1.50118
hiv1		1	-0.44660	0.06440	-6.93	<.0001	1.03670
factor1		1	-0.00005864	0.10184	-0.00	0.9995	1.64859
factor2		1	-0.13883	0.10328	-1.34	0.1795	1.58963
factor3		1	-0.13221	0.10152	-1.30	0.1934	1.62583
factor4		1	-0.05148	0.10341	-0.50	0.6188	1.83709
age1		1	-0.26912	0.19288	-1.40	0.1636	2.16595
age2		1	-0.20855	0.18047	-1.16	0.2484	2.50860
age3		1	-0.21156	0.18828	-1.12	0.2617	2.24893
age4		1	-0.18329	0.18165	-1.01	0.3135	2.54147
age5		1	-0.10022	0.18462	-0.54	0.5875	2.33804
age6		1	-0.18307	0.18679	-0.98	0.3275	2.33369
age7		1	-0.18122	0.17960	-1.01	0.3135	2.48450
age8		1	0.15026	0.18363	0.82	0.4136	2.31318
age9		1	-0.16971	0.18141	-0.94	0.3500	2.25747
age10		1	-0.26941	0.17769	-1.52	0.1301	2.37921
age11		1	-0.10538	0.18942	-0.56	0.5783	2.08884
age12		1	-0.15741	0.18362	-0.86	0.3917	2.19725
age13		1	-0.27838	0.19165	-1.45	0.1470	2.00892

Multiple Linear Regression Model - Subset Selection

Stepwise selection is used:

- Independent variables "hiv1" and "age8" are selected
- Their p-values are less than 5% level of significance

Summary:

- The independent variables of "hiv1", and “age8” would be chosen

	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			

Multiple Linear Regression Model - Summary

- Multiple Linear Regression model
  - Does not fulfil all the assumptions
    - Normality, constant residuals variance
  - The dependent variable is not a continuous
- After the analysis
  - Dependent variable, “death” belongs to the discrete variable
  - Cannot use multiple linear regression model
  - About the regression model for count data
    - Poisson regression model should be more appropriate

Poisson Log-linear Regression Model

Poisson Log-linear Regression Model

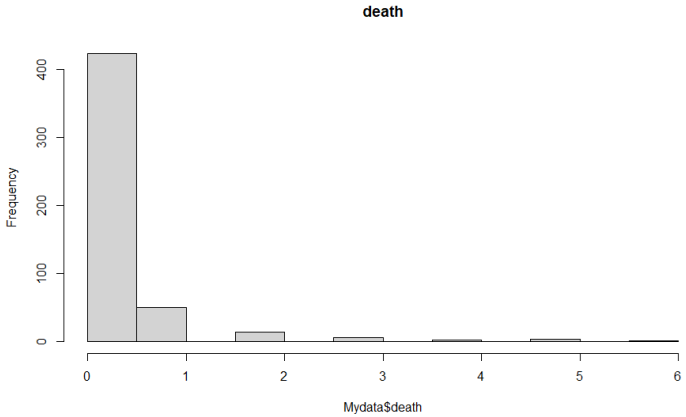
- Poisson regression is very useful when predicting count data through a series of continuous and/or categorical variables.

- Poisson regression model naturally arises when we want to model the average number of occurrences per unit of time or space.

- In this dataset, the described object is the number of deaths in particular age group.

Poisson Log-linear Regression Model - Diagnostic Checking

- Estimated mean is not equals to estimated variance
  - Estimated mean = 0.25
  - Estimated variance = 0.5445892
  - This result shows that the overdispersion occurs
- Dependent variable “deaths” is not follow normal distribution
  - Right skewed
  - “deaths” is a count variable, non-negative integer values



Poisson Log-linear Regression Model - Parameter Estimates

- Using the 5% level of significance, p-values only for predictor variables “hiv_1”, “factor_2”, and “age_10” are statistically significant, which are not satisfied.

- p-values are not satisfied, but is better than multiple linear regression model

```
1-fit_p$deviance/fit_p$null.deviance
```

```
## [1] 0.2726175
```

- The Pseudo-R^2 is 0.2726175
  - Which is not close to 1, indicating this model is not satisfied

- Result:
  It can be seen from the output results that only three predictors have passed the significance test at 0.05 significance level, Poisson regression model is not satisfied.

- The p-value of “py”, “factor_3”, “age_1” and “age_13” are close to 5% level of significance
  - Building a new model only within significance variables

```
## glm(formula = deaths ~ ., family = poisson(link = log), data = Mydata_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6622  -0.7573  -0.3031  -0.2127   5.2767
##
## 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 .
## hiv_1       -2.496948   0.320152  -7.799 6.23e-15 ***
## factor_1    -0.080673   0.262250  -0.308   0.7584
## factor_2    -0.689753   0.315617  -2.185   0.0289 *
## factor_3    -0.565790   0.296776  -1.906   0.0566 .
## factor_4    -0.255186   0.283817  -0.899   0.3686
## age_1       -1.200230   0.654473  -1.834   0.0667 .
## age_2       -0.718603   0.465616  -1.543   0.1227
## age_3       -0.719521   0.470086  -1.531   0.1259
## age_4       -0.718456   0.464325  -1.547   0.1218
## age_5       -0.447318   0.431719  -1.036   0.3001
## age_6       -0.719844   0.465800  -1.545   0.1223
## age_7       -0.644521   0.469388  -1.373   0.1697
## age_8        0.353657   0.381666   0.927   0.3541
## age_9       -0.550173   0.443490  -1.241   0.2148
## age_10      -1.095869   0.501389  -2.186   0.0288 *
## age_11      -0.302958   0.461665  -0.656   0.5117
## age_12      -0.517935   0.476291  -1.087   0.2768
## age_13      -0.983490   0.535213  -1.838   0.0661 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 516.90  on 499  degrees of freedom
## Residual deviance: 375.99  on 480  degrees of freedom
## AIC: 591.16
##
## Number of Fisher Scoring iterations: 7
```

Analysis

- A new model with these 7 variables
  - At 0.05 significance level
  - “hiv1”, “factor_2”and intercept are statistically significant
  - The p-value of “factor_3” and “age_10” still close to be significant at 5% level of significance
  - The p-value of “py”, “age_1” and “age_13” still are not significant at 5% level of significance

```
glm(formula = deaths ~ py + hiv_1 + factor_2 + factor_3 + age_1 +
    age_10 + age_13, family = poisson(link = log), data = Mydata_new)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3672  -0.8115  -0.3219  -0.2364   4.8649

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.532516   0.143963  -3.699 0.000216 ***
py           0.005224   0.004553   1.147 0.251210
hiv_1       -2.461871   0.316978  -7.767 8.06e-15 ***
factor_2    -0.593996   0.283406  -2.096 0.036089 *
factor_3    -0.495091   0.260760  -1.899 0.057611 .
age_1       -0.757301   0.587206  -1.290 0.197166
age_10      -0.727221   0.422885  -1.720 0.085493 .
age_13      -0.600758   0.463232  -1.297 0.194671
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 516.90  on 499  degrees of freedom
Residual deviance: 391.25  on 492  degrees of freedom
AIC: 582.42

Number of Fisher Scoring iterations: 6
```

Poisson Log-linear Regression Model - Overdispersion Checking

➢ Overdispersion occurs in Poisson regression when the observed variance of the response variable is larger than the mean of the response variable.

```
dispersiontest(fit_p)
```

```
##
##  Overdispersion test
##
## data:  fit_p
## z = 1.7595, p-value = 0.03925
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   1.442064
```

Result:

- The overdispersion test report there is a overdispersion issue within the model but is not significant.
- As we calculated before, the mean is less than the variance, which shows overdispersion.
- These two results are consistent.

Poisson Log-linear Regression Model - Dispersion Checking

- The Pearson Chi-square dispersion statistic is used

- The dispersion test report that there is an overdispersion problem after selection
  - For before and after stepwise selection, the dispersion estimate are larger than 1
  - Poisson log-linear regression model with stepwise selection may not fit the dataset

Pearson Chi-square Dispersion Statistic	
Before Stepwise Selection	1.474147
After Stepwise Selection	1.520205

Poisson Log-linear Regression Model - Summary

- Poisson log-linear regression model
  - Do not fulfill assumptions (Mean < Variance)
- After the analysis
  - Overdispersion exists
  - Performance of the variables are not satisified
- Conclusion
  - Poisson log-linear regression model may not suitable
  - Using regression models that can address the overdispersion
    - Such as negative binomial regression model

Remarks From Previous Presentation

From Previous Presentation

Although classical models may not be suitable for this dataset, we want to have a look on the results.

We had discussed these regression models in interim presentation:

- **Multiple Linear Regression**
- **Poisson Log-linear Regression**

From Previous Presentation

- Multiple linear regression model is not very suitable model for this dataset
  - Do not fulfil the model assumptions
    - Normality
    - Variance of residuals
  - Dependent variable is not continuous variable
- Poisson log-linear regression model is not very suitable model for this dataset
  - Do not fulfil the model assumptions
    - Mean = Variance
  - **Overdispersion existed but not significant**

Negative Binomial Regression Models

Negative Binomial Regression Models

- We would adopt the negative binomial regression models when the sample mean and the sample variance of the count data are not the same.
  - This situation is known as overdispersion or underdispersion.
- The expectation and variance of the negative binomial distribution are given:
  - Since $\alpha \geq 0$, $\lambda \geq 0$, therefore Var(Y) > E(Y)
  - May be an approach to dead with overdispersion

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

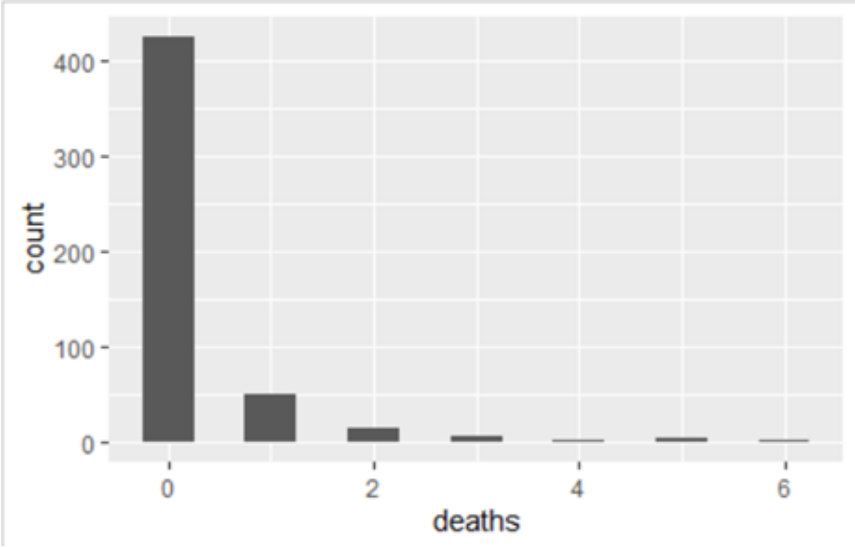
- In this dataset, the variance is greater than the mean, we consider that it might encounter the overdispersion.

Negative Binomial Regression Models - Assumptions

- Negative binomial regression is similar to multiple linear regression
- However, the dependent variable should follow the negative binomial distribution
  - Such that, the dependent variable is non-negative (NCSS Statistical Software, n.d.)

Negative Binomial Regression Models - Diagnostic Checking

- The figure indicates the distribution of values in dependent variable “deaths”
  - There are no negative values in the dependent variable
- Dependent variable is non-negative
  - Follows the negative binomial distribution
  - The model may appropriate to the dataset



Negative Binomial Regression Models - Parameter Estimates

The table indicates the parameter estimation by using the negative binomial regression model with the dummy variables

For the result:

- Only the p-value for independent variable of “hiv1”
  - Less than 5% level of significance

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	/	/

Negative Binomial Regression Models - Subset Selection

- Stepwise selection method is used
  - Perform both adding and deleting independent variables
  - Get a better result than forward selection and backward selection

Summary:

- The independent variable "hiv1" is selected for the models
- The values of AIC and BIC are 534.98 and 547.63 respectively

	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			

Negative Binomial Regression Models - Dispersion Checking

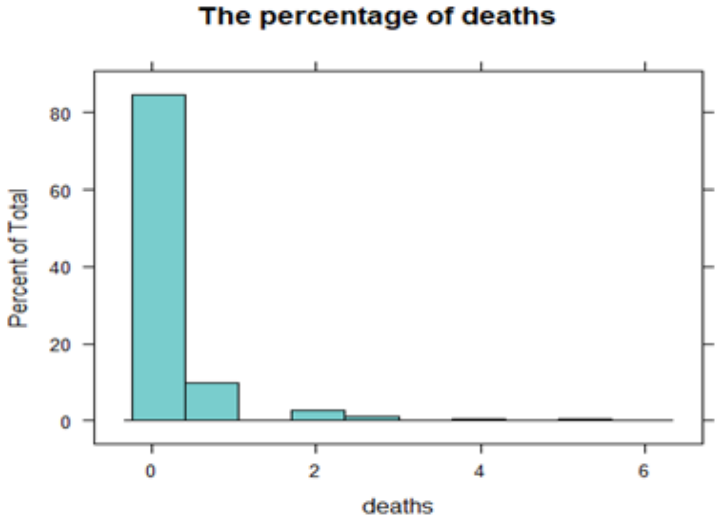
- The Pearson Chi-square dispersion statistic is used
- The dispersion test report that there is no dispersion problem after selection
  - For before and after stepwise selection, both dispersion estimate are less than but approximately to 1
  - Negative binomial regression model with stepwise selection may fit the dataset

Pearson Chi-square Dispersion Statistic	
Before Stepwise Selection	0.9463652
After Stepwise Selection	0.948298

Negative Binomial Regression Models - Summary

Analysis:

- Based on the above analysis:
  - The values of AIC and BIC in stepwise selection are much lower
  - Can deal with the overdispersion problem
  - But not enough to have final decision



Finding:

- Excessive-zero problem in the dependent variable
  - About 85.5% of the observations are zero
  - This model may not be suitable

Summary:

- Some advanced models could be employed because these models can effectively deal with an excessive-zero problem

Advanced Models

After the literature review:

- Two types of advanced models that could handle the excessive-zero problem:
  - Zero-inflated models
  - Hurdle models
- We would focus on zero-inflated models
- Two type of zero-inflated models includes:
  - Zero-inflated Poisson regression models
  - Zero-inflated negative binomial regression models

Zero-inflated Models

- Zero-inflated models are very popular in various fields:
  - Such as public health, biomedicine, economics
- In general, zero-inflated data are relative to non-zero count data.
- For example, consider a set of data that contains many zeros and consists of two parts:
  - Part 1: Zeros and non-zero count data form a Poisson distribution
  - Part 2: The remaining zeros are additionally obtained
    - Zero-inflated data or structural zero
- Thus, two types of zero-inflated models exist:
  - Zero-inflated Poisson regression models
  - Zero-inflated negative binomial regression models

Zero-inflated Poisson Regression Models

Zero-inflated Poisson Regression Models

- Zero-inflated Poisson regression model is one of the zero-inflated models
  - The model generates two models and combines them
  - A logit model and a Poisson model
- Assuming two types of zero existing in the data
  - The first part is when the structural zero (non-Poisson data) 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 Poisson Regression Models - Parameter Estimates

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 “hiv1”, "factor2", "factor3", "age6", "age10", "age13"
  - Less than 5% level of significance

After the analysis:

- Six independent variables are significant

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	.	.

Zero-inflated Poisson Regression Models - Subset Selection

Analysis:

- Only selects two independent variables
  - "py" and "hiv1"
- No variables except the intercept term included in zero-inflation parameter estimates
- The values of AIC and BIC for stepwise selection are 538.97 and 555.82 respectively

Summary:

- "py" and "hiv1" would be selected in the models

Stepwise Selection			
Parameter Estimates			
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		

Zero-inflated Poisson Regression Models - Diagnostic Checking

From the results:

- The dependent variable of the original data set has 424 zeros
- Before stepwise selection, the number of predicted zero of the zero-inflated Poisson regression model is 398
  - Which is about 94%
  - The model is insufficient to fit all of the excess zero
- After stepwise selection, the number of predicted zero of the zero-inflated Poisson regression model is 399
  - Which is about 94%
  - The model is still insufficient to fit all of the excess zero

Check for zero-inflation			
Stepwise selection	Observed zeros	Predicted zeros	Ratio
Before	424	398	0.94
After	424	399	0.94

Zero-inflated Poisson Regression Models - Dispersion Checking

- The Pearson Chi-square dispersion statistic is used
- The dispersion test report that there may be an overdispersion problem after selection
  - Before stepwise selection , the dispersion estimate is less than but approximately to 1
  - After stepwise selection , the dispersion estimate is larger than 1, overdispersion problem
  - Both models may not fit the dataset

Pearson Chi-square Dispersion Statistic	
Before Stepwise Selection	0.9455683
After Stepwise Selection	1.117586

Zero-inflated Poisson Regression Models - Summary

Analysis:

- Zero-inflated Poisson regression models are applied
  - To control the excess-zero problem in data
- The models may handle the excessive-zero
  - Value of dispersion checking is closer to 1 compare with Possion models
- More independent variables are significant from the parameter estimation

Finding:

- The model with stepwise selection also have an overdispersion problem
- Still insufficient to fit all of the excessive-zero

Summary:

- Could continue to another zero-inflated model namely zero-inflated negative binomial regression models to see if it is dealing with excess zeros clearly

Zero-inflated Negative Binomial Regression Models

Zero-inflated Negative Binomial Regression Models

- The second type from zero-inflated models
- More suitable for dealing with overdispersion and excessive zeros than classical models
  - Similar to the zero-inflated Poisson regression models
- The model generates two models and combines them at the end
  - Logit model and a binomial model
- Assuming two types of zero exist in the data
  - First: when the structure zero is generated by the existence of some special structure in the data
  - Secound: Sampling zero generated by the negative binomial distribution (Xie et al., 2013)

Zero-inflated Negative Binomial Regression Models - Parameter Estimates

This table shows the parameter estimation by using the zero-inflated negative binomial regression model with the dummy variables

For the result:

- The p-values for independent variables "hiv1", "factor2", "factor3", "age6", "age10" and "age13"
  - Less than 5% level of significance
  - Same as the result of the zero-inflated Poisson regression models
- AIC and BIC respectively are 551.1381 and 648.0741
  - Similar performance as the zero-inflated Poisson regression models

After the analysis:

❖ Overall, the performance are highly similar to the zero-inflated Poisson regression models

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

Zero-inflated Negative Binomial Regression Models - Subset Selection

Analysis:

- Only selects three independent variables
  - "py", "hiv1" and “age8”
- No variable except the intercept term included in zero-inflation parameter estimates
- The values of AIC and BIC for stepwise selection are 533.16 and 558.45 respectively

Summary:

- "py", "hiv1" and “age8” would be selected in the models

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		

Zero-inflated Negative Binomial Regression Models - Diagnostic Checking

From the results:

- The dependent variable of the original data set has 424 zeros
- Before stepwise selection, the number of predicted zero of the zero-inflated negative binomial regression model is 398
  - Which is about 94%.
  - The model is insufficient to fit all of the excess zero
- After stepwise selection, the number of predicted zero of the zero-inflated negative binomial regression model is 419
  - Which is about 99%.
  - The model is sufficient to fit all of the excess zero

Check for zero-inflation			
Stepwise Selection	Observed zeros	Predicted zeros	Ratio
Before	424	398	0.94
After	424	419	0.99

Zero-inflated Negative Binomial Regression Models - Dispersion Checking

- The Pearson Chi-square dispersion statistic is used
- The dispersion test report that there is no dispersion problem after selection
  - Before subset selection, the dispersion estimate is lower than but approximately to 1
  - After subset selection, the dispersion estimate is larger than but approximately to 1
  - Zero-inflated negative binomial regression model with stepwise selection may fit the dataset

Pearson Chi-square Dispersion Statistic	
Before Stepwise Selection	0.9470355
After Stepwise Selection	1.05837

Zero-inflated Negative Binomial Regression Models - Summary

Analysis:

- Zero-inflated negative binomial regression models are applied
  - To control overdispersion and excess-zero problem in data
- The models can handle the excess-zero indeed
  - Value of dispersion checking is close to 1
- More independent variables are significant

Finding:

- The model with stepwise selection do not have an overdispersion problem
- Still insufficient to fit all of the excessive-zero

Summary:

- Compare to the other models to see which model is the better fit

Model Evaluation

Comparison

Best result:

- Negative binomial regression model (NB) with stepwise selection
- Zero-inflated negative binomial regression model (ZINB) with stepwise selection

The performance of these two models are similar:

- NB with stepwise selection:
  - Lowest in BIC
  - Highest in Ratio of predicted zeros
- ZINB with stepwise selection:
  - Lowest in AIC
  - Lowest in RMSE
- Similar result in the dispersion test (difference <0.001)

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

Vuong Test

A hypothesis test to evaluate the best-fitted model between two non-nested models

Steps:

1. variance test to test if distinguishable
2. non-nested likelihood ratio test for best-fitted model

Variance test results:

ML = multiple linear regression model,
Poisson = Poisson regression model,
NB = negative binomial regression model,
ZIP = zero-inflated Poisson regression model,

ZINB = zero-inflated negative binomial regression model

- ❖ Almost all of the testing are distinguishable
- ❖ Only NB vs. ZINB is indistingguishable

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

Vuong Test

Non-nested likelihood ratio test result:

- NB, ZIP, ZINB are relatively better
  - They are all better than ML and Poisson
- For {NB vs. ZIP}, {ZIP vs. ZINB}
  - Both equally fitted
  - But the p-values of NB and ZINB are smaller in both cases
  - Which are 0.266 and 0.07947
    - Shows NB and ZINB are better than ZIP
- Can be inferred NB or ZINB is the fittest model

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	

Model Evaluation Result - NB or ZINB

From the section, proposed selection criteria:

- The performances of NB or ZINB are Similar

From the section, Vuong test:

- NB or ZINB should be the fittest model

Based on the principle of parsimony:

- The least independent variable, the better performance of the models

From stepwise selection:

- One independent variable "hiv1" is selected in NB
- Three independent variables "py", "hiv1" and "age8" are selected in ZINB

Final result:

- Then, NB would be the best-fitted model for the hemophilia dataset

Discussion

Research Questions Revisited

Q1. How do we choose the appropriate regression model for the dataset?

- We noticed the features of the dependent variable:
  - belongs to count data
  - has too many zeros
- We noticed the features of the four independent variables:
  - 3/4 independent variables are categorical
  - identify by using dummy variables

Based on these features, we decided to:

- Testing the dataset from classical models to advanced models, and choose the best one

Q2. What is the relationship between the number of deaths and the factors?

- We can obtain the relationship from the final model (negative binomial regression model with stepwise selection):

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

- The dummy variable "hiv1" has a significant negative correlation with “deaths”
- The “hiv1” variable has a better relationship with “deaths”
- If the patient is HIV-positive, then the chance of death will be very high

Conclusion

Conclusion - Overall Summary

From the beginning of the project, we tried to fit the dataset into the traditional regression models:

1. Multiple linear regression models
2. Poisson regression models
3. Negative binomial regression models

After fitting the traditional models:

1. We recognized the characteristics of the dataset deeply
2. Started to adopt the advanced model - zero-inflated models
    1. Zero-inflated Negative Binomial Regression models
    2. Zero-inflated Poisson Regression models

After that, we performed a model comparison:

1. Five proposed selection criteria are used
2. Vuong test is conducted

Finally, with the principle of parsimony:

- ❖ Negative binomial regression model with the stepwise selection is the best model

Conclusion - Limitations and Future Directions

- Limitation
  - Too few statistical models have performed
    - There may be a more suitable model we do not know
  - Only four independent variables in the dataset
    - The more variables, the higher the probability of getting significant correlated variables
    - Sensitive information cannot be disclosed in the medical documents
      - For example, the R-square of the multiple linear regression models is pretty low, some variables affect the dependent variable but not included in the dataset
- Future Directions
  - Perform more statistical models for the dataset
    - More opportunities to find a better model
    - e.g. Poisson-Inverse Gaussian regression model, similar to the negative binomial model
    - e.g. Hurdle models, similar to the zero-inflated models
  - Include more observations of the patient and relatable factors in the dataset
    - Observations: Time during the patient's illness
    - Factors: Gender of the patient
  - Trying to put variable "age" reassign into four categories
    - “Children”, “Teenagers”, “Adults” and “Elderly”

Reference

Reference

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 16:42, May 20, 2021, 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.

The End

Thank You