Finding the Right Distribution for Highly Skewed Zero-Inflated Clinical Data

Cordula Eggerth (00750881)

February 12, 2019

1 Introduction

The paper illustrates the case of count data with a high number of zeroes and provides options to deal with the situation when it comes to statistical modeling. In this particular case, the underlying data stems from a clinical study on children with an electrophysiological (EP) disorder. For the majority of these children no surgical intervention was necessary in line with the treatment of the EP disorder.
The study aims to provide insights on how to address the mentioned data situation and on how to analyze the data in a way that the estimation and inference results are appropriate (i.e. unbiased and reflecting the true distribution of the data). For this reason, several statistical models, namely Poisson model, Negative Binomial model, as well as the respective Zero-Inflated models were fitted and applied in order to find the best fitting model among them.

2 Data

2.1 Data Simulation

The study comprises 286 children with their age ranging from 8 to 18 years. The number of female and male participants was approximately the same. Four clinical variables were used to statistically model the dependent count variable, which refers to the count of hospital stays caused by the EP disorder:

- **Heart-related Hospitalization:** binary variable to model how often a child was hospitalized per year to the EP disorder
  - **0:** 2 or fewer hospital stays annually
  - **1:** more than 2 hospital stays annually
- **Heart Block:** binary variable to model whether heart activity is drastically reduced
  - **0:** no heart block

  - **1:** heart block

- **Prematurity:** binary variable to model whether the child was born prior to the 37th week of the pregnancy
  - **0:** child born in or after the 37th week of the pregnancy
  - **1:** child born prior to the 37th week of the pregnancy
- **Time:** continuous variable to model the time passed since the last hospital stay (measured in months)

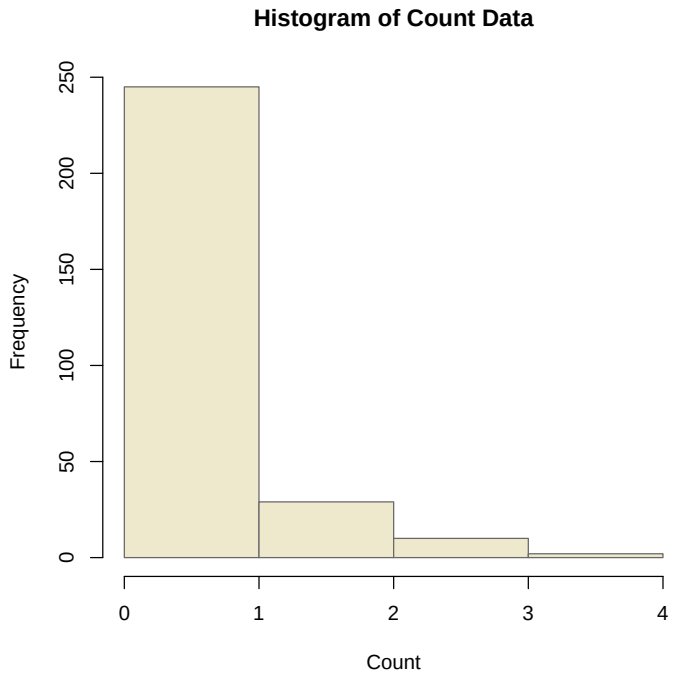
The dependent count variable is simulated using with the help of random numbers stemming from a poisson distribution and extra zeroes are added. The binary variables are simulated using a binomial distribution with n of 1, size of 1 and a 50 percent probability. The continuous time variable is simulated using a negative binomial distribution.

Exemplary simulation data entries:

	y_count	hospitalization	heartblock	prematurity	time
271	4	0	1	0	6
272	0	1	0	1	16
273	0	1	1	0	13
274	0	1	0	0	13
275	2	1	0	1	12
276	2	0	0	0	8
277	2	0	0	0	13
278	1	0	0	0	12
279	2	1	1	1	13
280	2	1	1	0	18
281	0	1	0	0	7
282	1	0	0	0	10
283	1	1	0	0	10
284	0	1	0	0	6
285	0	0	0	1	10
286	0	1	1	0	18

2.2 Descriptive Statistics

This sub-chapter presents the descriptive statistics based on the simulated data. From the histogram below it can be seen that the distribution of the number of hospitalization events that occurred is highly skewed to the occurrence of zero as a count result.



3 Statistical Models

3.1 Distributions

Due to the fact that there are highly skewed, zero-inflated count data, a normal linear regression would not be appropriate because it could produce negative predicted values (which is theoretically impossible in this case).
Therefore, a Poisson regression model or one of its variants or at least a model that takes these circumstances into account is used. However, this model assumes that the mean and the variance are equal. In practice, the variance appears to be often larger than the mean. Thus, the negative binomial model might constitute an appropriate alternative that can handle the extra variance via a dispersion parameter.
In some cases, there are more zeroes than Poisson or Negative Binomial distribution would predict. Consequently, a Zero-Inflated Poisson (ZIP) or Zero-Inflated Negative Binomial (ZINB) model would be a better choice

The probability distribution of Poisson distribution is given by:

$$P(y_i \mid X = x) = \frac{e^{-\mu} \cdot \mu_i^y}{y_i!};\ y_i = 0, 1, 2, ...$$
$$with\ \mu = e^{x'\beta}$$

where x denotes the covariates; and β denotes the vector of coefficients that need to be estimated.

The probability distribution of Negative Binomial distribution is given by:

$$P(y_i \mid \mu, \alpha) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i + 1)\Gamma(\frac{1}{\alpha})}(\frac{1}{1 + \alpha\mu})^{\frac{1}{\alpha}}(\frac{\mu}{\frac{1}{\alpha} + \mu})^{y_i}$$

where α is the dispersion parameter.

The probability density function of the Zero-inflated Poisson is given by:

$$Pr(Y_i = 0) = \pi_i + (1 - \pi_i)e^{-\mu_i}$$
$$Pr(Y_i = y_i) = (1 - \pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!};\ y_i = 1, 2, ...$$
$$where\ \pi_i = \frac{e^{z_i'\gamma}}{1 + e^{z_i'\gamma}};\ \mu_i = e^{x'\beta}$$

The probability density function of the Zero-inflated Negative Binomial is given by:

$$Pr(Y_i = 0) = \pi_i + (1 - \pi_i)(1 + \alpha\mu_i)^{-\frac{1}{\alpha}}$$
$$Pr(Y_i = y_i) = (1 - \pi_i)\frac{\Gamma(y_i + \frac{1}{\alpha})}{y_i!\Gamma(\frac{1}{\alpha})}(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \mu_i})^{\frac{1}{\alpha}}(\frac{\mu_i}{\frac{1}{\alpha} + \mu_i})^{y_i};\ y_i = 1, 2, ...$$

After simulating the data, these four types of models were used to evaluate model fit.

3.2 Underlying data structure and model selection

In this case, the dependent count variable is the number of surgeries (metric), the explanatory variables are: time (metric), heartblock (nominal), hospitalization (categorical) and prematurity (nominal). Moreover, an interaction between heartblock and prematurity is present.

To evaluate the models and to determine the most appropriate model, some

assessment criteria such as Akaike Information Criterion (AIC) or the Vuong test.
First of all, the four models discussed previously were applied to the clinical data. Afterwards, a Poissonness plot was generated to determine if the data were likely to have come from a Poisson distribution while a Lagrange multiplier (LM) test was used to check for model over-dispersion. Under H_0 of the Poisson model with no overdispersion, the limiting distribution of the LM statistic would follow a χ_1 distribution. Furthermore, Van den Broek score tests were used to formally test for zero inflation in the data. Its statistic is based on a comparison of actual zeros to those predicted by the model to test for zero-inflation relative to a Poisson distribution. Under H_0, i.e. the case of no zero-inflation, the test follows a χ_1 distribution.
Each of the four models was compared using log-likelihood estimates as a measure of model performance. In addition, the Vuong test statistic was used to compare non-nested models (e.g. Poisson versus Zero-Inflated Poisson).
Model fit was examined by comparing the Akaike information criterion (AIC) of the models under concern.

Model output using Poisson family

```
Call:
glm(formula = y_count ~ hospitalization + time + heartblock +
    prematurity + heartblock * prematurity, family = "poisson",
    data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3464  -1.0542  -0.9213   0.5314   3.1112

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.898179   0.281323  -3.193  0.00141 **
hospitalization        0.269558   0.160668   1.678  0.09340 .
time                   0.003731   0.019232   0.194  0.84619
heartblock             0.176365   0.197406   0.893  0.37164
prematurity            0.466947   0.202346   2.308  0.02102 *
heartblock:prematurity -1.061137   0.407569  -2.604  0.00923 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 356.82  on 285  degrees of freedom
Residual deviance: 345.19  on 280  degrees of freedom
AIC: 591.44
```

```
Number of Fisher Scoring iterations: 6
```

Model output using Negative Binomial family

```
Call:
glm.nb(formula = y_count ~ hospitalization + time + heartblock +
    prematurity + heartblock * prematurity, data = sim_data,
    init.theta = 1.43488429, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1839  -0.9717  -0.8610   0.4346   2.3202

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.877552   0.331996  -2.643  0.00821 **
hospitalization        0.278839   0.189946   1.468  0.14211
time                   0.001557   0.022974   0.068  0.94598
heartblock             0.180045   0.232173   0.775  0.43806
prematurity            0.470830   0.244961   1.922  0.05460 .
heartblock:prematurity -1.084270   0.468376  -2.315  0.02062 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.4349) family taken to be 1)

    Null deviance: 264.39  on 285  degrees of freedom
Residual deviance: 255.86  on 280  degrees of freedom
AIC: 581.31

Number of Fisher Scoring iterations: 1


              Theta:  1.435
          Std. Err.:  0.558

 2 x log-likelihood:  -567.310
```

Model output using Zero-Inflated Poisson

```
Call:
zeroinfl(formula = y_count ~ hospitalization + time + heartblock + prematurity +
    heartblock * prematurity, data = sim_data, dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.8580 -0.6509 -0.5525  0.5672  3.9447
```

```
Count model coefficients (poisson with log link):
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.126749   0.431408  -0.294    0.769
hospitalization        -0.256803   0.251743  -1.020    0.308
time                   -0.002074   0.033448  -0.062    0.951
heartblock              0.230708   0.292216   0.790    0.430
prematurity             0.492518   0.308479   1.597    0.110
heartblock:prematurity -0.323145   0.594310  -0.544    0.587

Zero-inflation model coefficients (binomial with logit link):
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.13424    1.03394   0.130   0.8967
hospitalization       -1.46292    0.74069  -1.975   0.0483 *
time                  -0.01437    0.08396  -0.171   0.8641
heartblock             0.15258    0.72393   0.211   0.8331
prematurity            0.07077    0.77676   0.091   0.9274
heartblock:prematurity  1.49728    1.16448   1.286   0.1985
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 18
Log-likelihood:  -278 on 12 Df
```

Model output using Zero-Inflated Negative Binomial

```
Call:
zeroinfl(formula = y_count ~ hospitalization + time + heartblock + prematurity +
    heartblock * prematurity, data = sim_data, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.8579 -0.6510 -0.5525  0.5672  3.9449

Count model coefficients (negbin with log link):
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.126560   0.431473  -0.293    0.769
hospitalization        -0.256320   0.251737  -1.018    0.309
time                   -0.002091   0.033460  -0.063    0.950
heartblock              0.230429   0.292290   0.788    0.430
prematurity             0.492303   0.308568   1.595    0.111
heartblock:prematurity -0.322523   0.594296  -0.543    0.587
Log(theta)             10.010912  72.622998   0.138    0.890

Zero-inflation model coefficients (binomial with logit link):
                      Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)             0.13548    1.03371   0.131   0.8957
hospitalization        -1.46103    0.73973  -1.975   0.0483 *
time                   -0.01444    0.08397  -0.172   0.8635
heartblock              0.15143    0.72387   0.209   0.8343
prematurity             0.06995    0.77672   0.090   0.9282
heartblock:prematurity  1.49778    1.16417   1.287   0.1982
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 22268.1446
Number of iterations in BFGS optimization: 30
Log-likelihood:  -278 on 13 Df
```

4 Discussion

For the purpose of discussing the results of the simulated study, a seed was set for the respective random number generation in the data simulation section to make the results and their interpretation reproducible.
Four models, namely the Poisson model, the Negative Binomial model as well as their zero-inflated counterparts were fitted in order to provide a basis for deciding on the best fitting model for the underlying data and situation. All four models comprised the same covariates, which means that in each model the count variable (regarding hospitalization events) is to be explained by the regressors hospitalization, time, heartblock, prematurity and the interaction between heartblock and prematurity.
When it comes to the standard Poisson model assuming that the expected value equals the mean of the distribution, which was fitted using the glm function, the intercept was significant on the 99 percent level. The variable prematurity was significant on the 95 percent level. The interaction between heartblock and prematurity was significant on the 99 percent level. The AIC of the Poisson model was 591.44.
To fit the Negative Binomial model, the glm.nb function was used. This model took a dispersion parameter into account in order to relax the assumption of equality of the mean and expected value of the underlying distribution. In this scenario, the intercept appeared to be significant on the 99 perecent level and the interaction term on the 95 percent level. There was also a significance on the 90 percent level for the variable of prematurity. The AIC of the Negative Binomial model was 581.31, which means that the error was slightly more favorable than in the standard Poisson model.
With regard to the Zero-Inflated Poisson (ZIP) model, among the count model coefficients, which were determined based on a Poisson distribution with log link, none of the coefficients were significant. Among the zero-inflation model coefficients, which were determined based on a Binomial distribution with logit link, the hospitalization variable was significant on the 95 percent level. The remaining coefficients were not significant. The log-likelihood (LL) of the model

was -278 on 12 degrees of freedom.
Finally, the Zero-Inflated Negative Binomial (ZINB) model was fitted for the coefficients and interaction term named above. Using the ZINB model, none of the coefficients were significant for the count model portion (based on a Negative Binomial distribution with log link). Among the coefficients of the zero-inflation part (based on a Binomial distribution with logit link), the variable of hospitalization was significant on the 95 percent level. The LL of the model was -278 on 13 degrees of freedom.

The below output depicts the coefficient estimates for the Poisson, the NB model and for the count parts of the zero-inflated models as well as the respective deviance. Additionally, the dispersion parameter is 22268.15 for the ZINB model and 1.43 for the Negative Binomial model.
The log-likelihood and the AIC, which are both presented below, served as criteria to compare the four models. In terms of LL, the models appear to be fitting better to the data when moving from Poisson and Negative Binomial to the Zero-Inflated models. In fact, when we merely take LL into account, the ZIP model appears to be the most favorable in view of the underlying data (closely followed by the ZINB model). Regarding the AIC, again the ZIP is the best model and is followed closely by the ZINB model. The Poisson and Negative Binomial model have both a higher AIC, which indicates they do not fit the data so well as the other proposed models do.

```
$Estimate
                        Poisson      NB ZIP-count ZINB-count
(Intercept)             -0.8982 -0.8776   -0.1267    -0.1266
hospitalization          0.2696  0.2788   -0.2568    -0.2563
time                     0.0037  0.0016   -0.0021    -0.0021
heartblock               0.1764  0.1800    0.2307     0.2304
prematurity              0.4669  0.4708    0.4925     0.4923
heartblock:prematurity -1.0611 -1.0843   -0.3231    -0.3225

$Deviance
        Null deviance DF Null Residual deviance DF Residuals
Poisson      356.8228     285          345.1882          280
NB           264.3946     285          255.8596          280

$Teststatistik
                   Poisson        NB       ZIP      ZINB
Log_likelihood   -289.7178 -283.6550 -278.0248  -278.025
AIC               591.4356  581.3101  580.0496   582.050
Dispersion_param        NA    1.4349        NA 22268.145
```

Furthermore, the data can be analyzed by the means of a Vuong test. The **Vuong test** intends to compare 2 statistical models referring to the exact same dataset with the help of the maximum-likelihood (ML) theory. In this statistical

test, the H_0 assumes that the statistical models under concern fit the observed data equally well, whereas the H_1 assumes the contrary. Regarding the models, they do not need to be nested models and do not have to be specified in a certain way. In line with the Vuong test, the Kullback-Leibler divergence (KLD) is used to measure the deviance between the true model and the one created from the data (g_t). The H_0 of the Vuong test is stated as:

$$H_0 : D_{KL}(g_t \mid\mid g_1) = D_{KL}(g_t \mid\mid g_2)$$

The Vuong test is commonly used in scenarios of zero inflation or suspected zero inflation. It tests if the observed difference between the LL contribution to the zero-inflation model and to the standard count model is (on average) larger than 0. Under these circumstances, the test statistic of the Vuong test can be stated as:

$$Vuong = (s_{dl}\sqrt{n})^{-1} \sum_{i=1}^{n} dl_i$$
$$dl_i = ln(l(y_i \mid x_i, z_i, \hat{\beta}, \hat{\gamma})) - ln(f(y_i \mid x_i'\tilde{\beta})),$$

whereby $\hat{\beta}$ is the estimated β in the zero-inflated model,$\tilde{\beta}$ the estimated β if the zero-inflation component is not contained in the model. Moreover, dl is a vector of length N, such that the ith element is the ith individual log-likelihood difference, and s_{dl} is the standard deviation of dl. Under H_0, the Vuong test statistic is asymptotically normally distributed in view of the central limit theorem (CLT).

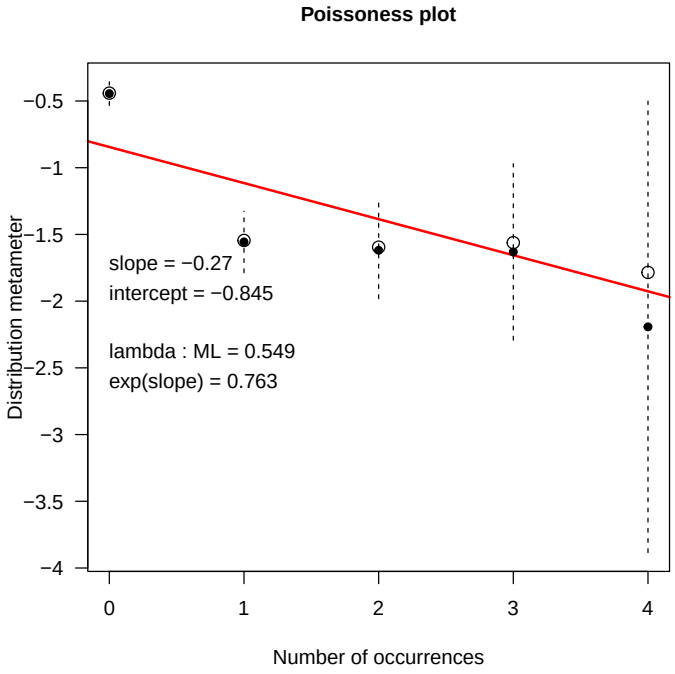
The below output shows the results of the Vuong test for the underlying simulation data. As already pointed out in the course of the model evaluation based on LL and AIC above, the results of the Vuong test also point into the direction of the zero-inflated models when it come to choosing on of the analyzed models. In this analysis, the Poisson model was compared with the ZIP model and the Negative Binomial model with the ZINB model. Thus, the standard models were compared with their zero-inflated counterparts respectively.

```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A   p-value
Raw                  -2.379053 model2 > model1 0.0086786
AIC-corrected        -1.158296 model2 > model1 0.1233717
BIC-corrected         1.073244 model1 > model2 0.1415809

Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
```

```
------------------------------------------------------------
               Vuong z-statistic             H_A    p-value
Raw                  -1.9167795 model2 > model1   0.027633
AIC-corrected         0.1259558 model1 > model2   0.449883
BIC-corrected         3.8600678 model1 > model2 5.6678e-05
```

The **Poissonness Plot** serves to explore whether the simulation data actually corresponds to a Poisson distribution. In this study, the simulated data deviates from a standard Poisson distribution as it can be seen from the below plot that the points are not located on the line.



Furthermore, the **Van den Broek Score Test** was applied to the simulated data. This test has the null hypothesis that no zero inflation is prevalent in the data. Its test statistic is distributed with a Chi-Squared distribution with one degree of freedom. The testing procedure compares the actual number of observed zeroes with the number of zeroes that would theoretically predicted by the model by the means of log-likelihood estimates. Since the p-value of the test statistic is significant, the null hypothesis (of no zero inflation) is rejected.

```
Score test for zero inflation

        Chi-square = 20.3433
        df = 1
        pvalue: 6.4719e-06
```

5 Appendix

```
> ### SIMULATION ###
> set.seed(2345)
> ### ------------------ ----- ###
> ### study participants       ###
> ### ------------------------ ###
> n <- 286
> ### ------------------------ ###
> ### dependent count variable ###
> ### ------------------------ ###
> set.seed(153)
> y_count <- rpois(n,lambda=1)
>  # set extra zeroes
> y_which_is1 <- which(y_count==1)
> length_y_which_is1 <- length(y_which_is1)
> y_which_is1_toSetTo0 <- y_which_is1[1:round((length_y_which_is1/3), digits=0)]
> y_count[y_which_is1_toSetTo0] <- 0
> y_which_isGreater1 <- which(y_count>1)
> length_y_which_isGreater1 <- length(y_which_isGreater1)
> y_which_isGreater1_toSetTo0 <- y_which_isGreater1[1:round((length_y_which_isGreater1/2),
+                                                           digits=0)]
> y_count[y_which_isGreater1_toSetTo0] <- 0
> ### ------------------------ ###
> ### regressors               ###
> ### ------------------------ ###
>
> # REGRESSOR 1: "hospitalization"
> #              0 ... <=2 hospital stays per year
> #              1 ... >2 hospital stays per year
> set.seed(27895)
> hospitalization <- rbinom(n=n, size=1, prob=0.5)
> # REGRESSOR 2: "heartblock"
> #              0 ... no heart block
> #              1 ... heart block
> set.seed(1890)
> heartblock <- rbinom(n=n, size=1, prob=0.4)
> # REGRESSOR 3: "prematurity"
> #              0 ... born prior to 37th week of pregnancy
> #              1 ... born in or after 37th week of pregnancy
> set.seed(33)
> prematurity <- rbinom(n=n, size=1, prob=0.3)
> # REGRESSOR 4: "time"
> #              continuous variable ... months passed since last hospital stay
> set.seed(5)
> time <- rnbinom(n=n, size=24, mu=12)
```

```
> ### ------------------------ ###
> ### simulated data           ###
> ### ------------------------ ###
>
> # create data.frame of simulated data
> sim_data <- data.frame(y_count=y_count, hospitalization=hospitalization,
+                        heartblock=heartblock, prematurity=prematurity,
+                        time=time)
> # show 15 exemplary data.frame entries
> sim_data[(nrow(sim_data)-15):nrow(sim_data), ]

    y_count hospitalization heartblock prematurity time
271       4               0          1           0    6
272       0               1          0           1   16
273       0               1          1           0   13
274       0               1          0           0   13
275       2               1          0           1   12
276       2               0          0           0    8
277       2               0          0           0   13
278       1               0          0           0   12
279       2               1          1           1   13
280       2               1          1           0   18
281       0               1          0           0    7
282       1               0          0           0   10
283       1               1          0           0   10
284       0               1          0           0    6
285       0               0          0           1   10
286       0               1          1           0   18

> ### ------------------------ ###
> ### statistical models       ###
> ### ------------------------ ###
>
> library(MASS)
> library(pscl)
> pois <- glm(y_count ~hospitalization +heartblock +prematurity +time
+             +heartblock*prematurity, data = sim_data, family = poisson)
> summary(pois)

Call:
glm(formula = y_count ~ hospitalization + heartblock + prematurity +
    time + heartblock * prematurity, family = poisson, data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3464  -1.0542  -0.9213   0.5314   3.1112
```

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.898179   0.281323  -3.193  0.00141 **
hospitalization         0.269558   0.160668   1.678  0.09340 .
heartblock              0.176365   0.197406   0.893  0.37164
prematurity             0.466947   0.202346   2.308  0.02102 *
time                    0.003731   0.019232   0.194  0.84619
heartblock:prematurity -1.061137   0.407569  -2.604  0.00923 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 356.82  on 285  degrees of freedom
Residual deviance: 345.19  on 280  degrees of freedom
AIC: 591.44

Number of Fisher Scoring iterations: 6

> nb <- glm.nb(y_count ~hospitalization +heartblock +prematurity +time
+              +heartblock*prematurity, data = sim_data)
> summary(nb)

Call:
glm.nb(formula = y_count ~ hospitalization + heartblock + prematurity +
    time + heartblock * prematurity, data = sim_data, init.theta = 1.43488429,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1839  -0.9717  -0.8610   0.4346   2.3202

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.877552   0.331996  -2.643  0.00821 **
hospitalization         0.278839   0.189946   1.468  0.14211
heartblock              0.180045   0.232173   0.775  0.43806
prematurity             0.470830   0.244961   1.922  0.05460 .
time                    0.001557   0.022974   0.068  0.94598
heartblock:prematurity -1.084270   0.468376  -2.315  0.02062 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.4349) family taken to be 1)

    Null deviance: 264.39  on 285  degrees of freedom
```

```
Residual deviance: 255.86  on 280  degrees of freedom
AIC: 581.31

Number of Fisher Scoring iterations: 1


              Theta:  1.435
          Std. Err.:  0.558

 2 x log-likelihood:  -567.310
```

```
> zip<- zeroinfl(y_count ~hospitalization +heartblock +prematurity +time
+                +heartblock*prematurity, data = sim_data, dist = "poisson")
> summary(zip)
```

```
Call:
zeroinfl(formula = y_count ~ hospitalization + heartblock + prematurity +
    time + heartblock * prematurity, data = sim_data, dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.8580 -0.6509 -0.5525  0.5672  3.9447

Count model coefficients (poisson with log link):
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.126749   0.431408  -0.294    0.769
hospitalization        -0.256803   0.251743  -1.020    0.308
heartblock              0.230708   0.292216   0.790    0.430
prematurity             0.492518   0.308479   1.597    0.110
time                   -0.002074   0.033448  -0.062    0.951
heartblock:prematurity -0.323145   0.594310  -0.544    0.587

Zero-inflation model coefficients (binomial with logit link):
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.13424    1.03394   0.130   0.8967
hospitalization       -1.46292    0.74069  -1.975   0.0483 *
heartblock             0.15258    0.72393   0.211   0.8331
prematurity            0.07077    0.77676   0.091   0.9274
time                  -0.01437    0.08396  -0.171   0.8641
heartblock:prematurity  1.49728    1.16448   1.286   0.1985
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 18
Log-likelihood:  -278 on 12 Df
```

```
> zinb <- zeroinfl(y_count ~hospitalization +heartblock +prematurity +time
```

```
+                   +heartblock*prematurity, data = sim_data, dist = "negbin")
> summary(zinb)

Call:
zeroinfl(formula = y_count ~ hospitalization + heartblock + prematurity +
    time + heartblock * prematurity, data = sim_data, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.8579 -0.6510 -0.5525  0.5672  3.9449

Count model coefficients (negbin with log link):
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.126560   0.431473  -0.293    0.769
hospitalization       -0.256320   0.251737  -1.018    0.309
heartblock             0.230429   0.292290   0.788    0.430
prematurity            0.492303   0.308568   1.595    0.111
time                  -0.002091   0.033460  -0.063    0.950
heartblock:prematurity -0.322523   0.594296  -0.543    0.587
Log(theta)            10.010904  72.623322   0.138    0.890

Zero-inflation model coefficients (binomial with logit link):
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.13548    1.03371   0.131   0.8957
hospitalization       -1.46103    0.73973  -1.975   0.0483 *
heartblock             0.15143    0.72387   0.209   0.8343
prematurity            0.06995    0.77672   0.090   0.9282
time                  -0.01444    0.08397  -0.172   0.8635
heartblock:prematurity  1.49778    1.16417   1.287   0.1982
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 22267.9582
Number of iterations in BFGS optimization: 30
Log-likelihood:  -278 on 13 Df

> ### ------------------------ ###
> ### Results                ###
> ### ------------------------ ###
>
> library(stats)
> Estimate <- cbind(pois$coefficients,nb$coefficients,zip$coefficients$count,
+                   zinb$coefficients$count)
> colnames(Estimate) <- c("Poisson", "NB", "ZIP-count","ZINB-count")
> Deviance <- rbind(cbind(pois$null.deviance, pois$df.null,
+                         pois$deviance, pois$df.residual),
```

```
+                   cbind(nb$null.deviance,nb$df.null,
+                         nb$deviance,nb$df.residual))
> colnames(Deviance) <- c("Null_dev","DF_Null","Residual_dev","DF_Residuals")
> rownames(Deviance) <- c("Poisson","NB")
> teststatistik <- cbind(rbind(logLik(pois),pois$aic,NA),
+                        rbind(logLik(nb),nb$aic,nb$theta),
+                        rbind(logLik(zip),AIC(zip),NA),
+                        rbind(logLik(zinb),AIC(zinb),zinb$theta))
> colnames(teststatistik) <- c("Poisson", "NB", "ZIP", "ZINB")
> rownames(teststatistik) <- c("Log_likelihood", "AIC","Dispersion_param")
> Summary <- list("Estimate"=Estimate, "Deviance"= Deviance,
+                 "Teststatistik" = teststatistik)
> Summary_round <- lapply(Summary, round, 4)
> print(Summary_round)

$Estimate
                      Poisson      NB ZIP-count ZINB-count
(Intercept)           -0.8982 -0.8776   -0.1267    -0.1266
hospitalization        0.2696  0.2788   -0.2568    -0.2563
heartblock             0.1764  0.1800    0.2307     0.2304
prematurity            0.4669  0.4708    0.4925     0.4923
time                   0.0037  0.0016   -0.0021    -0.0021
heartblock:prematurity -1.0611 -1.0843   -0.3231    -0.3225

$Deviance
        Null_dev DF_Null Residual_dev DF_Residuals
Poisson 356.8228     285     345.1882          280
NB      264.3946     285     255.8596          280

$Teststatistik
                   Poisson        NB       ZIP      ZINB
Log_likelihood   -289.7178 -283.6550 -278.0248  -278.025
AIC               591.4356  581.3101  580.0496   582.050
Dispersion_param        NA    1.4349        NA 22267.958

> ### ------------------------ ###
> ### vuong test               ###
> ### ------------------------ ###
>
> vuong(pois, zip)

Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A   p-value
Raw                   -2.379053 model2 > model1 0.0086786
```

```
AIC-corrected         -1.158296 model2 > model1 0.1233717
BIC-corrected          1.073244 model1 > model2 0.1415809

> vuong(nb, zinb)

Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                  -1.9167795 model2 > model1   0.027633
AIC-corrected         0.1259558 model1 > model2   0.449883
BIC-corrected         3.8600678 model1 > model2 5.6678e-05

> ### ------------------------ ###
> ### poissoness plot          ###
> ### ------------------------ ###
>
> library(vcd)
> distplot(sim_data$y_count, "poisson")
> ### ------------------------ ###
> ### van den broek score test ###
> ### ------------------------ ###
> #install.packages("vcdExtra")
> library(vcdExtra)
> zero.test(sim_data$y_count)

Score test for zero inflation

         Chi-square = 20.3433
         df = 1
         pvalue: 6.4719e-06
```

6 Bibliography

The following source was used as a basis for the whole report:

Gupta, R., et al. 2013. "Finding the Right Distribution for Highly Skewed Zero-Inflated Clinical Data." In *Epidemiology Biostatistics and Public Health.* Vol. 10 No. 1, S. e8732-1 - e8732-15.