**10. Introduction to Generalized Linear Models II**

We learnt in the last lecture that Poisson regression is not an appropriate model when data shows signs of over-dispersion.

**Causes of Over-dispersion**

* Subject heterogeneity. If some relevant predictors are not in the model, then the unexplained heterogeneity among the subjects will cause greater variation in the response than the Poisson model predicts.
* Outliers in the data.
* Positive correlation between responses in clustered data. Examples of naturally occurring clusters are families, households, litters, colonies, and neighbourhoods.

**Negative Binomial Regression**

When there are signs of over-dispersion, alternatively a Negative Binomial pdf can be considered that permits the variance to exceed the mean.

The Binomial distribution counts the number of successes in a fixed number of Bernoulli trials. On the other hand, the Negative Binomial Distribution counts the number of Bernoulli trials required to get a fixed number of successes, *r*. The distribution, mean and variance are:



, 

This shows the interesting property of this distribution that



That is, the variance is a quadratic function of the mean.

Also, If  and  such that , then

 and 

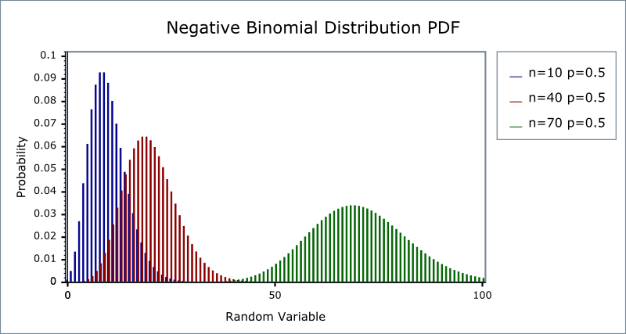
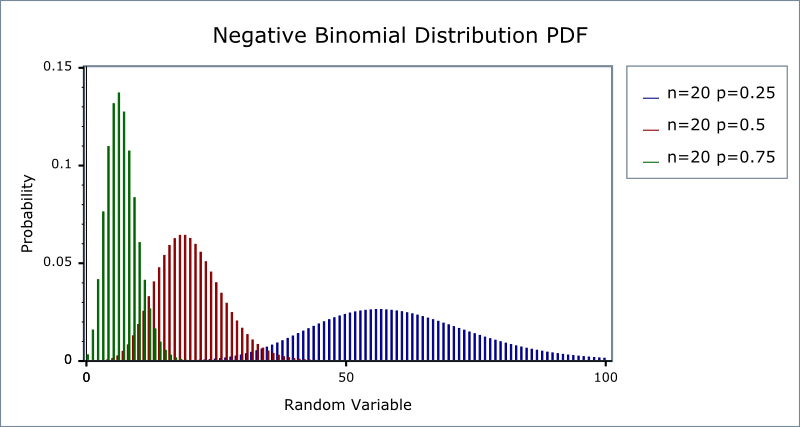
That is, the Negative Binomial distribution includes the Poisson distribution as a limiting case.

**Note:** The non-canonical link function for Negative Binomial distribution is natural log.

The density can be re-written as:



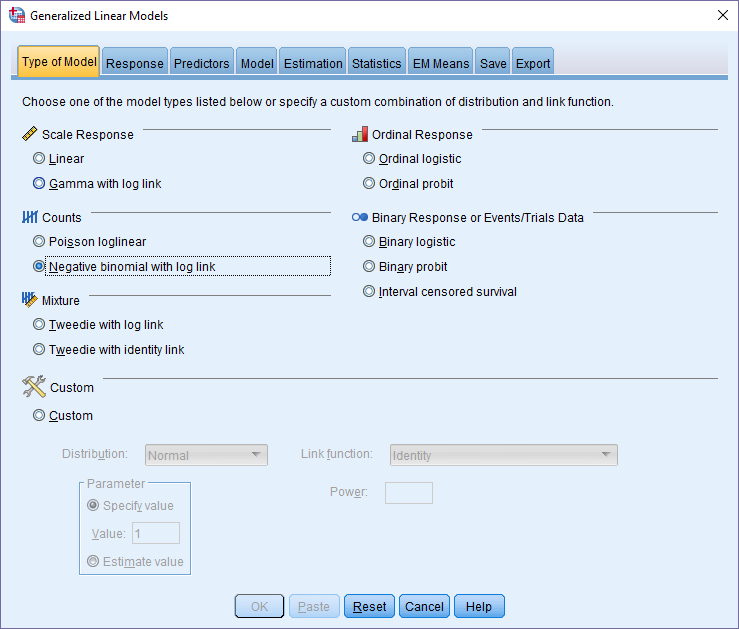
This explains the name Negative Binomial.

[](http://www.google.co.uk/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=0ahUKEwiVi47LhufXAhWLA8AKHYM0A0MQjRwIBw&url=http://www.boost.org/doc/libs/1_36_0/libs/math/doc/sf_and_dist/html/math_toolkit/dist/dist_ref/dists/negative_binomial_dist.html&psig=AOvVaw2Io8Red3Gvpr7_MvPAnBsE&ust=1512157006559001)[](http://www.google.co.uk/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwiak4eNhufXAhVIIMAKHYRhDFkQjRwIBw&url=http://www.boost.org/doc/libs/1_36_0/libs/math/doc/sf_and_dist/html/math_toolkit/dist/dist_ref/dists/negative_binomial_dist.html&psig=AOvVaw2Io8Red3Gvpr7_MvPAnBsE&ust=1512157006559001)

**Example 10.1**

We repeat the GLZ regression analysis this time for a Negative Binomial distribution.

Here, only a selection of output is included.



|  |  |
| --- | --- |
| **Model Information** | |
| Dependent Variable | Infections |
| Probability Distribution | Negative binomial (1) |
| Link Function | Log |

|  |  |  |  |
| --- | --- | --- | --- |
| **Goodness of Fita** | | | |
|  | Value | df | Value/df |
| Deviance | 350.247 | 282 | **1.242** |
| Scaled Deviance | 350.247 | 282 |  |
| Pearson Chi-Square | 403.188 | 282 | **1.430** |
| Scaled Pearson Chi-Square | 403.188 | 282 |  |
| Log Likelihoodb | -451.969 |  |  |
| Akaike's Information Criterion (AIC) | 913.939 |  |  |
| Finite Sample Corrected AIC (AICC) | 914.152 |  |  |
| Bayesian Information Criterion (BIC) | 932.236 |  |  |
| Consistent AIC (CAIC) | 937.236 |  |  |
| Dependent Variable: Infections  Model: (Intercept), Swimmer, Location, Sex, Age | | | |
| a. Information criteria are in smaller-is-better form. | | | |
| b. The full log likelihood function is displayed and used in computing information criteria. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Tests of Model Effects** | | | |
| Source | Type III | | |
| Wald Chi-Square | df | Sig. |
| (Intercept) | 5.947 | 1 | .015 |
| Swimmer | 14.736 | 1 | .000 |
| Location | 8.792 | 1 | .003 |
| Sex | .198 | 1 | .656 |
| Age | 3.213 | 1 | .073 |
| Dependent Variable: Infections  Model: (Intercept), Swimmer, Location, Sex, Age | | | |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | | | | | | |
| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
| Lower | Upper | Wald Chi-Square | df | Sig. | Lower | Upper |
| (Intercept) | .430 | .4155 | -.385 | 1.244 | 1.070 | 1 | .301 | 1.537 | .681 | 3.470 |
| [Swimmer=1] | .614 | .1600 | .301 | .928 | 14.736 | 1 | .000 | 1.848 | 1.351 | 2.529 |
| [Swimmer=2] | 0a | . | . | . | . | . | . | 1 | . | . |
| [Location=1] | .488 | .1645 | .165 | .810 | 8.792 | 1 | .003 | 1.629 | 1.180 | 2.249 |
| [Location=2] | 0a | . | . | . | . | . | . | 1 | . | . |
| [Sex=1] | -.077 | .1732 | -.416 | .262 | .198 | 1 | .656 | .926 | .659 | 1.300 |
| [Sex=2] | 0a | . | . | . | . | . | . | 1 | . | . |
| Age | -.033 | .0184 | -.069 | .003 | 3.213 | 1 | .073 | .968 | .933 | 1.003 |
| (Scale) | 1b |  |  |  |  |  |  |  |  |  |
| (Negative binomial) | 1b |  |  |  |  |  |  |  |  |  |
| Dependent Variable: Infections  Model: (Intercept), Swimmer, Location, Sex, Age | | | | | | | | | | |
| a. Set to zero because this parameter is redundant. | | | | | | | | | | |
| b. Fixed at the displayed value. | | | | | | | | | | |

Notice that now, and compared to the Poisson model, Values/df are closer to one, log Likelihood is larger and Information Criteria are smaller, all indicating a better fit. On the other hand, Age is now marginally insignificant.

**Zero-Inflated Poisson Models**

In modelling count data, there may be circumstances that the variance is too large because there are many zeros as well as a few very high values. As we just discussed the negative binomial model is an alternative model in such cases.

But sometimes it is just a matter of having too many zeros than a Poisson model would predict.  In this case, a better solution is often the Zero-Inflated Poisson (ZIP) model.  (And when extra variation occurs too, Zero-Inflated Negative Binomial model).

ZIP models assume that some zeros occurred by a Poisson process, but others were not even eligible to have the event occur.  So, there are two processes at work—one that determines if the individual is even eligible for a non-zero response, and the other that determines the count of that response for eligible individuals.

The problem is either process can result in a zero count.   Since you cannot tell which zeros were eligible for a non-zero count, you cannot tell which zeros were results of which process.  The ZIP model fits, simultaneously, two separate regression models.  One is a logistic or probit model that models the probability of being eligible for a non-zero count.  The other models the size of that count.

Both models use the same predictor variables, but estimate their coefficients separately.  So, the predictors can have vastly different effects on the two processes.

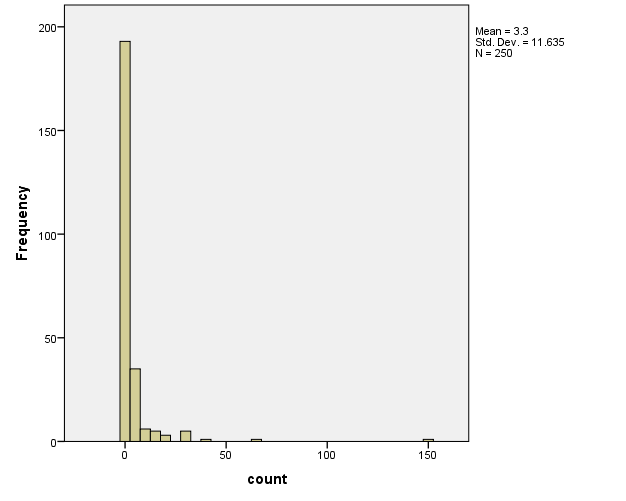
However, ZIP model requires it be theoretically plausible that some individuals are ineligible for a count.  For example, consider modelling the number of alcoholic drinks consumed in a day, which could plausibly be fit with a ZIP model.  Some participants do drink alcohol, but will have consumed zero that day, by chance.   But others just do not drink alcohol, so will never have a non-zero response.  The ZIP model can determine which predictors affect the probability of being an alcohol consumer and which predictors affect how many drinks the consumers consume.  They may not be the same predictors for the two models, or they could even have opposite effects on the two processes.

**Example 10.2**

(Partly adopted from <https://stats.idre.ucla.edu/r/dae/zip/>)

The state wildlife biologists want to model how many fish are being caught by fishermen at a state park. Visitors are asked how long they stayed, how many people were in the group, were there children in the group and how many fish were caught. Some visitors do not fish, but there is no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish.

The data file (<https://stats.idre.ucla.edu/stat/data/fish.csv>) consists of 250 groups that went to a park. Each group was questioned about how many fish they caught (count), how many children were in the group (child), how many people were in the group (persons), and whether or not they brought a camper to the park (camper).



**ZIP Output**

library(pscl)

model1 = zeroinfl(count ~ child + camper | persons, data = fish)

summary(model1)

Pearson residuals:

Min 1Q Median 3Q Max

-1.2369 -0.7540 -0.6080 -0.1921 24.0847

Count model coefficients (poisson with log link):

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.59789 0.08554 18.680 <2e-16 \*\*\*

child -1.04284 0.09999 -10.430 <2e-16 \*\*\*

camper 0.83402 0.09363 8.908 <2e-16 \*\*\*

Zero-inflation model coefficients (binomial with logit link):

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.2974 0.3739 3.470 0.000520 \*\*\*

persons -0.5643 0.1630 -3.463 0.000534 \*\*\*

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 12

Log-likelihood: -1032 on 5 Df

**Standard Poisson Model**

summary(model2 <- glm(count ~ child + camper, family = poisson, data = fish))

Deviance Residuals:

Min 1Q Median 3Q Max

-3.7736 -2.2293 -1.2024 -0.3498 24.9492

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.91026 0.08119 11.21 <2e-16 \*\*\*

child -1.23476 0.08029 -15.38 <2e-16 \*\*\*

camper 1.05267 0.08871 11.87 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2958.4 on 249 degrees of freedom

Residual deviance: 2380.1 on 247 degrees of freedom

AIC: 2723.2

Number of Fisher Scoring iterations: 6

**Comparing Two Models**

The Vuong test compares the zero-inflated model with an ordinary Poisson regression model. In this example, we can see that our test statistic is significant, indicating that the zero-inflated model is superior to the standard Poisson model.

vuong(model2, model1)

Vuong Non-Nested Hypothesis Test-Statistic:

(test-statistic is asymptotically distributed N(0,1) under the

null that the models are indistinguishable)

-------------------------------------------------------------

Vuong z-statistic H\_A p-value

Raw -3.574254 model2 > model1 0.00017561

AIC-corrected -3.552392 model2 > model1 0.00019087

BIC-corrected -3.513900 model2 > model1 0.00022079

**ZINB Output**

ibrary(pscl)

model3 = zeroinfl(count ~ child + camper | persons,dist="negbin", data = fish)

summary(model3)

Pearson residuals:

Min 1Q Median 3Q Max

-0.5861 -0.4617 -0.3886 -0.1974 18.0135

Count model coefficients (negbin with log link):

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.3710 0.2561 5.353 8.64e-08 \*\*\*

child -1.5153 0.1956 -7.747 9.41e-15 \*\*\*

camper 0.8791 0.2693 3.265 0.0011 \*\*

Log(theta) -0.9854 0.1760 -5.600 2.14e-08 \*\*\*

Zero-inflation model coefficients (binomial with logit link):

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.6031 0.8365 1.916 0.0553 .

persons -1.6666 0.6793 -2.453 0.0142 \*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Theta = 0.3733

Number of iterations in BFGS optimization: 22

Log-likelihood: -432.9 on 6 Df

Similarly, this model can be compared with the corresponding standard Negative Binomial model.

**Standard Negative Binomial**

library(MASS)

summary(model4 <- glm.nb (count ~ child + camper, data = fish))

Call:

glm.nb(formula = count ~ child + camper, data = fish, init.theta = 0.2552931119,

link = log)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.3141 -1.0361 -0.7266 -0.1720 4.0163

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.0727 0.2425 4.424 9.69e-06 \*\*\*

child -1.3753 0.1958 -7.025 2.14e-12 \*\*\*

camper 0.9094 0.2836 3.206 0.00135 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.2553) family taken to be 1)

Null deviance: 258.93 on 249 degrees of freedom

Residual deviance: 201.89 on 247 degrees of freedom

AIC: 887.42

Number of Fisher Scoring iterations: 1

Theta: 0.2553

Std. Err.: 0.0329

2 x log-likelihood: -879.4210

vuong(model3, model4)

Vuong Non-Nested Hypothesis Test-Statistic:

(test-statistic is asymptotically distributed N(0,1) under the

null that the models are indistinguishable)

-------------------------------------------------------------

Vuong z-statistic H\_A p-value

Raw 1.7017116 model1 > model2 0.044405

AIC-corrected 1.2026316 model1 > model2 0.114559

BIC-corrected 0.3238863 model1 > model2 0.373012

The test suggests zero-inflated negative binomial model is a significant improvement over a standard negative binomial model.