ARISTOTLE UNIVERSITY
OF THESSALONIKI

Count data Models

Eleni Verykouki, PhD
Biostatistician
Research Associate of the Laboratory of Hygiene,
Social-Preventive Medicine and Medical Statistics, AUTH
everykouki@auth.gr

MSc Health Statistics & Data Analytics

Poisson Regression

- Poisson Regression models are best used for modeling events where the outcomes are counts.
- Counts are non-negative integers. They represent the number of occurrences of an event within a fixed period.

- Count based variables:
  - How many heart attacks or strokes one's had.
  - Number of people visiting a doctor's office per month.
  - How many days from outbreak until infection.

Count data can also be expressed as rate data

Poisson distribution

- The Poisson distribution is given by a probability function:

$$P(Y = y) = \frac{e^{-\lambda}\mu^{\lambda}}{y!}, \qquad y = 0, 1, 2, 3, \ldots$$

$$E(y) = \lambda \text{ and } Var(y) = \lambda.$$

λ: the expected number of events in the interval

Models the probability of event or events y occurring within a specific timeframe, assuming that y occurrences are not affected by the timing of previous occurrences of y.

Poisson Regression

- Poisson regression can be used to describe the independent relationship of variables on the count outcome while holding constant the values of other variables.

- Poisson Regression model using the Log-link function is the most commonly used link function.

$$\log(y) = \beta_0 + \beta_1 x_1 + \cdots$$

- the log of the expected value of y is a linear function of the explanatory variables.

Assumptions

- **Poisson Response** The response variable is a count per unit of time or space, described by a Poisson distribution and has non-negative values.
- **Independence** The observations must be independent of one another.
- **Mean=Variance** By definition, the mean of a Poisson random variable must be equal to its variance.
- **Linearity** The log of the mean rate, log(λ), must be a linear function of x.

Dataset of counts

Characteristics of a typical dataset of count data.

- The data consists of non-negative integers (discrete values).
- Regression techniques such Linear Regression inappropriate for modeling such data as these techniques work best on real numbers (continuous variables).
- The frequency distribution quite skewed because in the data there might be many data points for just a few values.
- There might be sparsity in the data, usually when there are rare events.

Interpretation of the Results

- $\exp(\beta_0)$: effect on the mean of *Y*
- $\exp(\beta_i)$: with every unit increase in *X*, the predictor variable has multiplicative effect of $\exp(\beta_i)$
  - If β_i=0, then $\exp(\beta_i) = 1$, *Y* and *X* are not related
  - If β_i>0, then $\exp(\beta_i) > 1$, and the expected count is $\exp(\beta_i)$ times larger than when *X* = 0
  - If β_i <0, then $\exp(\beta_i) < 1$, and the expected count is $\exp(\beta_i)$ times smaller than when *X* = 0

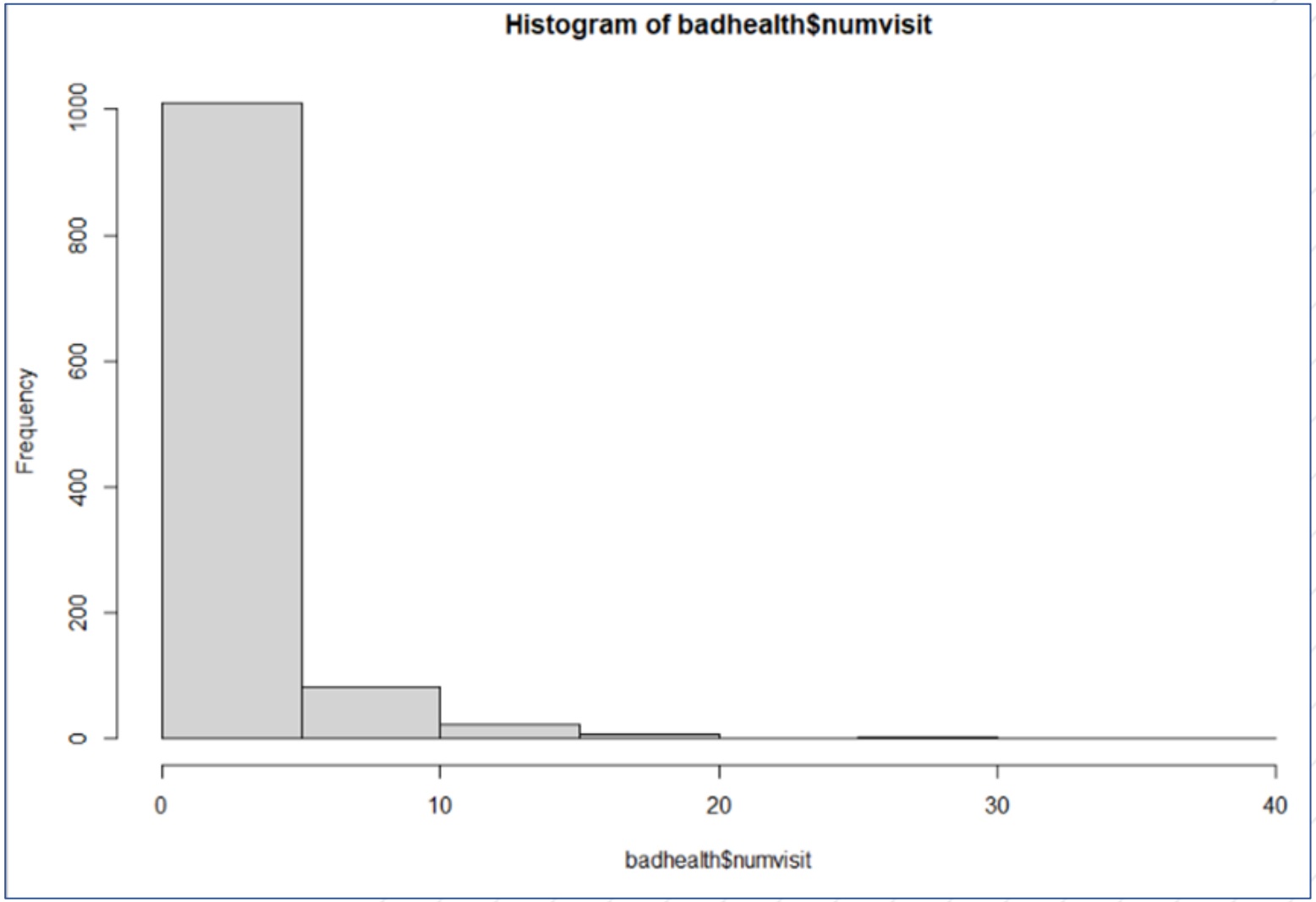
  - If `family = poisson` is kept in glm() then, these parameters are calculated using Maximum Likelihood Estimation MLE.

Example

- Dataset “badhealth” from package “COUNT” in R.
- The dataset consists of 1,127 observations on the following 3 variables:

- `numvisit`: number of visits to doctor during 1998
- `badh`: 1=patient claims to be in bad health; 0=not in bad health
- `age`: age of patient

Histogram of response variable : `numvisit`



`numvisit` clearly does not follow the normal distribution

Univariate Poisson Regression with one continuous explanatory variable

- we will examine whether the number of visits to doctor, `numvisit`, (response variable) is affected by the `age` of the patient (explanatory variable).

R results

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5519  -1.9977  -0.7242   0.4776  11.8941

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.290283   0.071300   4.071 4.68e-05 ***
age         0.014837   0.001759   8.436  < 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: 4020.3  on 1126  degrees of freedom
Residual deviance: 3949.9  on 1125  degrees of freedom
AIC: 6121.2

Number of Fisher Scoring iterations: 6
```

```
> exp(coef(model))
(Intercept)         age
   1.336805    1.014948
> exp(confint(model))
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept) 1.161780 1.536432
age         1.011452 1.018449
```

For every one year increase in age, the number of visits to the doctor increases by a factor of 1.015. [IRR=1.015, 95%CI: (1.0114, 1.018), p<0.001].

Univariate Poisson Regression with one categorical explanatory variable

- we will examine whether the number of visits to doctor, `numvisit`, (response variable) is affected by `badh` of the patient (explanatory variable).

R results

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4974  -1.9687  -0.7434   0.7056  10.4043

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.66162    0.02255   29.34   <2e-16 ***
badhYes      1.14930    0.04436   25.91   <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: 4020.3  on 1126  degrees of freedom
Residual deviance: 3475.5  on 1125  degrees of freedom
AIC: 5646.7

Number of Fisher Scoring iterations: 5

> exp(coef(model))
(Intercept)     badhYes
   1.937931    3.155980
> exp(confint(model))
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept) 1.853545 2.024838
badhYes     2.891341 3.440639
>
```

Patients who claim to be in bad health have about 3.156 times more visits to the doctor compared to patients who claim that are not in bad health.

[IRR=3.156,

95%CI: (2.891, 3.441), p<0.001].

Multiple Poisson Regression

- Both `age` and `badh` are included in the model

```
Call:
glm(formula = numvisit ~ badh + age, family = poisson(link = "log"),
    data = badhealth)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6653  -1.9186  -0.6789   0.6292  10.0684

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.447022   0.071428   6.258 3.89e-10 ***
badhYes     1.108331   0.046169  24.006  < 2e-16 ***
age         0.005822   0.001822   3.195   0.0014 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4020.3  on 1126  degrees of freedom
Residual deviance: 3465.3  on 1124  degrees of freedom
AIC: 5638.6

Number of Fisher Scoring iterations: 5

> exp(coef(glmbadp))
(Intercept)     badhYes         age
   1.563648    3.029299    1.005839
> exp(confint(glmbadp))
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept) 1.358589 1.797621
badhYes     2.765629 3.314413
age         1.002249 1.009434
```

Interpretation

- Patients who claim to be in bad health have about 3.029 times more visits to the doctor compared to patients who claim that are not in bad health, adjusted for age [IRR=3.029, 95%CI: (2.765, 3.314), p<0.001].

- For every one-year increase in age, the number of visits to the doctor increases by a factor of 1.006, adjusted for bad health. [IRR=1.006, 95%CI: (1.002, 1.009), p=0.001].

Goodness of fit

- The model's residual deviance can be used to assess the degree to which the predicted values differ from the observed. It is used to check overdispersion
- Model is true: the residual deviance is distributed as a χ^2 random variable with degrees of freedom equal to the model's residual degrees of freedom.

Lack of fit:

- A more comprehensive data set may be needed.
- Extreme observations may cause the deviance to be larger than expected
- Problem with the Poisson model

Deviance

- Given the null hypothesis that the Poisson is the appropriate model, the test of goodness-of-fit for the Poisson model is given by

$$\sum_{i=1}^{n} res_{di} \overset{H_0}{\sim} \chi^2_{n-p}.$$

The goodness of fit in our case:

```
> deviances2 <- residuals(model,type="deviance")
> dev.tvalue <- sum(deviances2^2)
> c(dev.tvalue, 1-pchisq(dev.tvalue,df))
[1] 3465.301    0.000
```

p-value <0.001 meaning that this model is not a good fit.

Overdispersion

- **Overdispersion** suggests that there is more variation in the response than the model implies.

- To estimate dispersion parameter φ
- $\hat{\varphi} = \frac{\text{Deviance}}{df}$

- $\varphi = 1$, no dispersion
- $\varphi > 1$, overdispersion → Quasi-Poisson model or Negative Binomial model

Quasi Poisson model

- a Quasi-Poisson model assumes that the variance is a linear function of the mean.
- Standard errors are inflated by multiplying the variance by φ, so that the standard errors are larger than the likelihood approach would imply

$$SE_Q(\hat{\beta}) = \sqrt{\hat{\varphi}} * \mathrm{SE}(\widehat{\beta})$$

Applying the Quasi-Poisson model to our data

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.447022   0.140217   3.188  0.00147 **
badhYes     1.108331   0.090633  12.229  < 2e-16 ***
age         0.005822   0.003577   1.628  0.10389
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 3.853594)

    Null deviance: 4020.3  on 1126  degrees of freedom
Residual deviance: 3465.3  on 1124  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

> exp(coef(model_q))
(Intercept)     badhYes         age
   1.563648    3.029299    1.005839
> exp(confint(model_q))
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept) 1.1852402 2.053887
badhYes     2.5305495 3.610800
age         0.9987946 1.012902
```

Negative Binomial

- Introduces another parameter in addition to λ which gives the model more flexibility.

$$E(y) = \lambda \text{ and } Var(y) = \lambda + \mathrm{D}\lambda^2$$

D: dispersion parameter

- Greater heterogeneity in the Poisson means results in a larger value of D
- The negative binomial model assumes an explicit likelihood model

Applying the Negative Binomial model to our data

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.404116   0.130847   3.088  0.00201 **
badhYes     1.107342   0.111603   9.922  < 2e-16 ***
age         0.006952   0.003397   2.047  0.04070 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.9975) family taken to be 1)

    Null deviance: 1355.7  on 1126  degrees of freedom
Residual deviance: 1217.7  on 1124  degrees of freedom
AIC: 4475.3

Number of Fisher Scoring iterations: 1


              Theta:  0.9975
          Std. Err.:  0.0693

 2 x log-likelihood:  -4467.2850
> exp(coef(model_nb))
(Intercept)     badhYes         age
   1.497977    3.026304    1.006977
> exp(confint(model_nb))
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept) 1.163508 1.929749
badhYes     2.444392 3.777076
age         1.000427 1.013598
```

Zero-inflated Poisson model (ZIP model)

- Empirical data often show more zeroes than would be expected under Poisson or Negative Binomial model.
- The ZIP model is a special case of a more general type of statistical model referred to as a latent variable model.

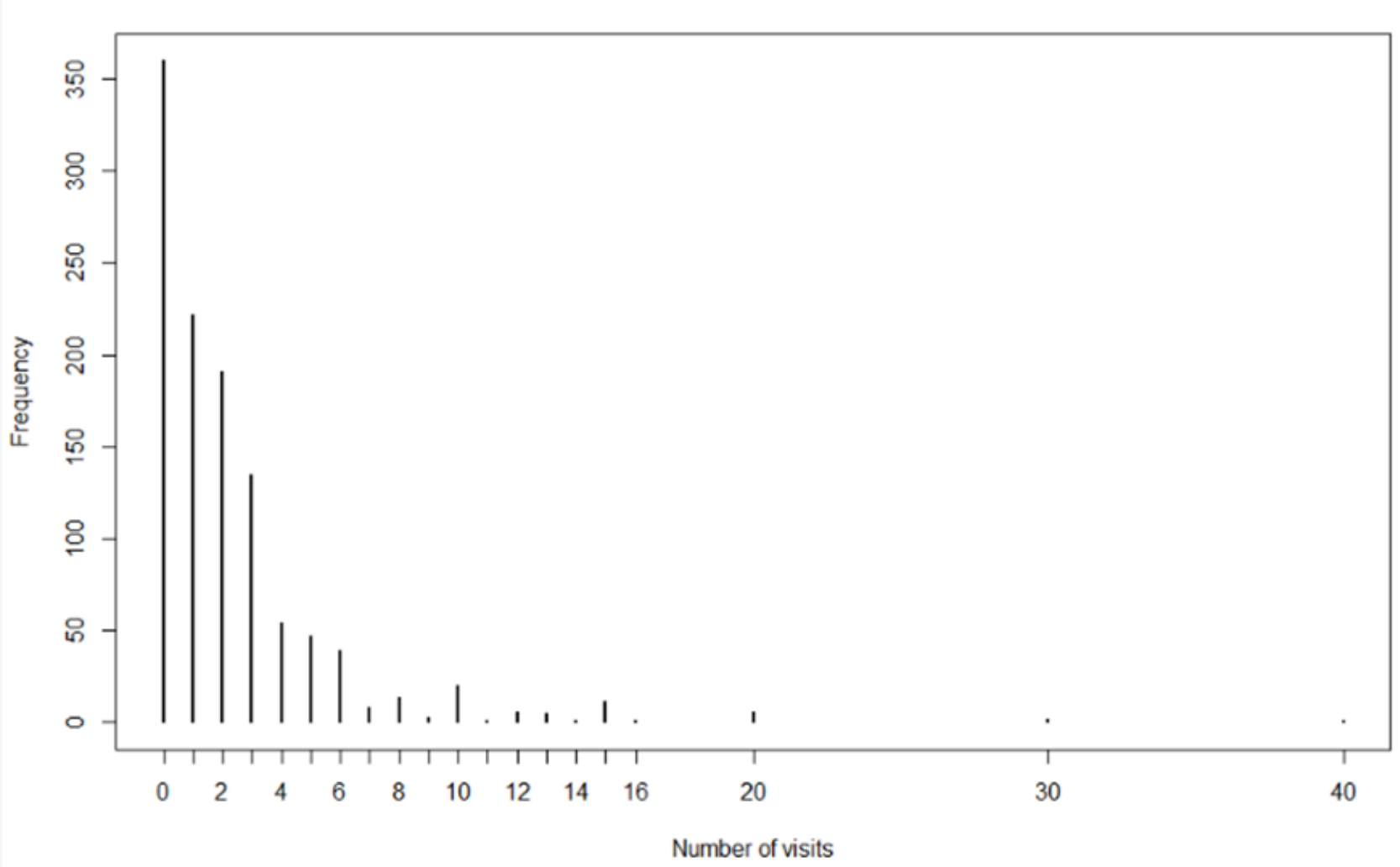
- Examples:
- How many alcoholic drinks did you consume last weekend?
- Modelling the factors associated with dental caries.
- Factors associated with the number of reported cases of malaria among under-5 children.

ZIP model cont.

- Zero-inflated models are two-component mixture models combining a point mass at zero with a count distribution such as Poisson or negative binomial.
- two processes may be at work:
  - one that determines whether or not an event happens at all and
  - another that determines how many times the event happens when it does.

We will check whether a zero-inflated Poisson model will be even a better fit to the data instead of Poisson model.

We notice that there are a lot of individuals with zero visits to the doctor.



- The model combines a logit model that predicts which of the two latent classes a person belongs, with a Poisson model that predicts the outcome for those in the second latent class.
- First part (Poisson model):

$$\log(\lambda) = \beta_0 + \beta_1 age + \beta_2 badh$$

λ is the mean number of cases last week

- Second part (Logistic model)

$$\log(a) = \beta_0 + \beta_1 numvisit$$

α is the probability of having no visits to the doctor.

Results from R

```
Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.685152   0.076694   8.934  < 2e-16 ***
badhYes     0.876665   0.047951  18.283  < 2e-16 ***
age         0.008859   0.001938   4.571 4.85e-06 ***

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.402925   0.276958  -5.065 4.07e-07 ***
badhYes     -1.099601   0.294666  -3.732  0.00019 ***
age          0.014191   0.006978   2.034  0.04198 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 10
Log-likelihood: -2549 on 6 Df
> exp(coef(zip_model))
count_(Intercept)     count_badhYes         count_age  zero_(Intercept)      zero_badhYes          zero_age
        1.9840728         2.4028722         1.0088984         0.2458768         0.3330040         1.0142926
> exp(confint(zip_model))
                        2.5 %    97.5 %
count_(Intercept) 1.7071638 2.3058975
count_badhYes     2.1873335 2.6396500
count_age         1.0050735 1.0127379
zero_(Intercept)  0.1428797 0.4231211
zero_badhYes      0.1869085 0.5932938
zero_age          1.0005145 1.0282604
```

Model Fit / Model Comparison

Vuong Test to compare non nested models

We cannot use the drop-in-deviance test we discussed earlier because Poisson model and ZIP model are not nested

```
> vuong(model1, zip_model1)
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                     -7.554618 model2 > model1 2.1005e-14
AIC-corrected           -7.469807 model2 > model1 4.0156e-14
BIC-corrected           -7.256621 model2 > model1 1.9844e-13
```

References/ Further Reading:

1. Rodríguez, G. (2007). Lecture Notes on Generalized Linear Models. URL: https://data.princeton.edu/wws509/notes/ (chapters 4 & 5).
2. Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression Models for Count Data in R. Journal of Statistical Software, 27(8), 1–25. https://doi.org/10.18637/jss.v027.i08

Thank you