Models for Count Data

Eleni Verykouki (February 2022)

Aim

To describe the relationship between a count variable and one or more continuous or categorical explanatory variables.

Objectives

By the end of this session, you will be able to:

- Understand assumptions of the Poisson regression model and be able to fit it in sample data.
- Understand how to fit the model and interpret the parameter estimates.
- Be able to test for overdispersion and apply the appropriate models to overcome this.
- Understand what zero-inflated models are, be able to fit them to data and interpret the results.

Poisson regression is another special case of the generalized linear model, where the random component is specified by the Poisson distribution. In this case, the response variable is usually a count. Counts are discrete data with non-negative integer values that count something. Examples of count variables can be:

1. The number of heart attacks or strokes one's had within five years.
2. Number of people visiting a doctor's office per month.
3. Number of days from outbreak until infection.

But let's start introducing first the Poisson distribution.

Poisson Distribution

The Poisson distribution is given by a probability function:

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

Poisson distribution 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.

Lambda (λ) can be thought of as the expected number of events in the interval. It is also called rate parameter (in some textbooks is denoted as μ).

The mean and the variance of the Poisson distribution are:

$E(y) = \lambda$ and $Var(y) =$λ.

Equality of the mean and variances is often referred to as the *equidispersion* property of the Poisson distribution.

Properties/assumptions so that a random variable follows a Poisson distribution:

- A random variable that follows a Poisson distribution is always nonnegative and takes only whole number values.
- The rate at which the event occurs remains constant over time. The rate of an event occurring in the time interval we have set does not increase or decrease if we increase or decrease that time interval.
- The occurrence of an event does not change the likelihood that another event will occur. (memoryless property of the Poisson distribution). For example, the occurrence of a seizure now will not affect the risk of a patient having another seizure in the future.
- The shorter the time interval we set, the less events we will expect. For example, if we only follow patients with seizures for one minute instead of 1 week, the number of seizures we would expect for that minute will tend toward 0.

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.

A Poisson regression model assumes that the conditional mean $\lambda_i = E(Y_i|x_i)$ is given by

$$g(\lambda_i) = \beta_0 + \beta_1 x_1 + \cdots$$

Where g is a link function.

The link function of Poisson regression usually involves a nonlinear transformation, which means that the expected value of the dependent variable is a nonlinear function of the independent variables.

The most commonly used link function is the Log Link:

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

Where,

y is the response variable, $\beta_0, \beta_1, \ldots, \beta_k$ are numeric coefficients and $x_1, \ldots x_k$ are the explanatory variables. So, the log of the expected value of y is a linear function of the explanatory variables. Note that the explanatory variables can be categorical continuous or ordinal.

Note that the Poisson regression model contains no separate error term like the ε we usually see in linear regression, because λ determines both the mean and the variance of a Poisson random variable.

Interpretation of a Poisson regression model

Since a log link function is used, results are expressed in a log scale. We need to exponentiate in order to get the real effect of the explanatory variables on the response variable.

The exponentiated estimates are also called Incidence Rate ratios (IRR) and have the following interpretation.

- $\exp(\beta_0)$: 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, no effect
    - 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

Assumptions of Poisson Regression

- The response variable (Y) is a count variable.
- Counts must be positive integers (i.e. whole numbers) equal or greater than zero.
- Explanatory variables can be continuous, dichotomous or ordinal.
- Observations must be independent.
- Counts must follow a Poisson distribution. The mean and variance should be the same.

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.

Example:

Now, let's see how to use simple or multiple Poisson regression through an example.

We will use the 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

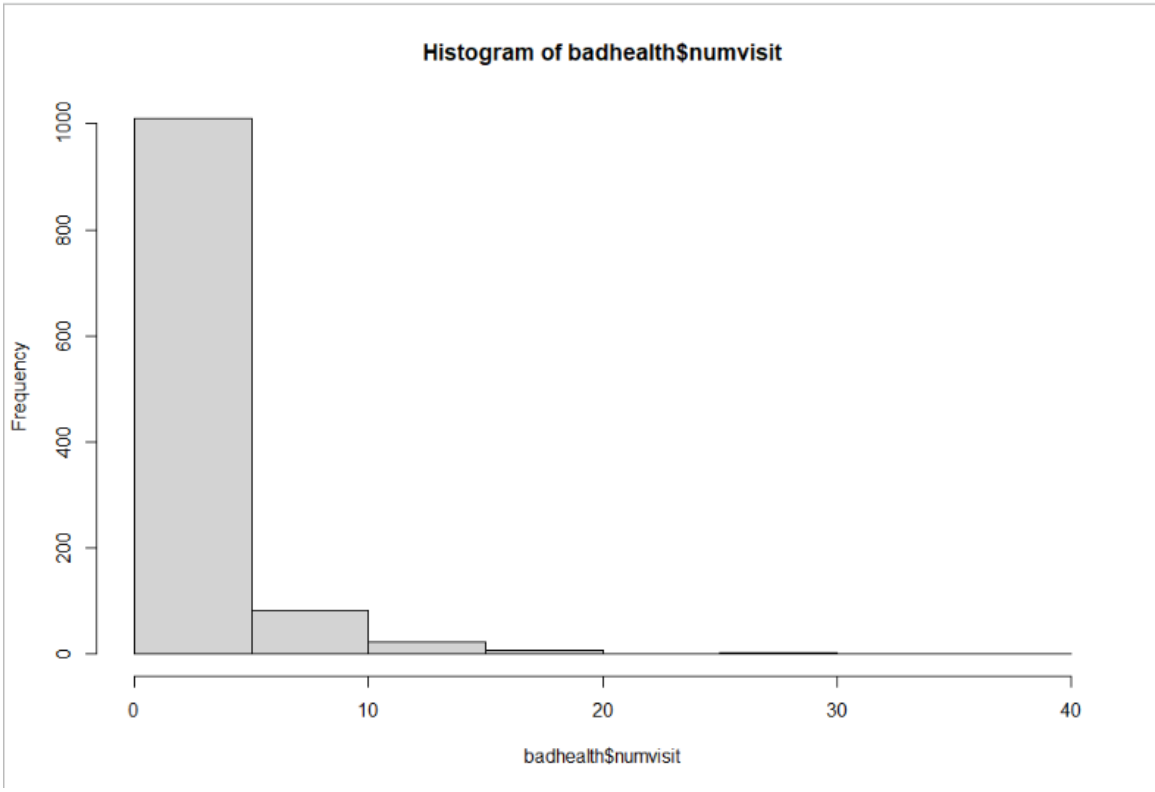
badh: 1=patient claims to be in bad health; 0=not in bad health

age: age of patient

Simple Poisson Regression with a quantitative (continuous) explanatory variable

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

We can view the dependent variable numvisit data continuity by creating a histogram



It can be clearly seen that the data is not in the form of a bell curve like in a normal distribution.

We fit a Poisson model and the output from R looks like this:

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

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
```

Interpretation

The first column in the Coefficients table, named Estimate, is the coefficient values of β_0 (intercept) and β_1. Since we have used a log link function results are expressed in a log scale.

After calculating the exponentiated estimate coefficient values we have:

```
> 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
```

Thus, $\exp(\beta_0) = e^{0.290} = 1.337$ is the effect on the number of visits to the doctor (response variable) when age =0. This is not meaningful here.

For age we have, $\exp(\beta_1) = e^{0.015} = 1.015$.

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].

Simple Poisson Regression with a qualitative (categorical) explanatory variable

Now we move on examining whether the number of visits to the doctor is affected by the patient's health. Here we will use variable badh from the data set which is a binary variable taking value=1 if the patient claims to be in bad health and value=0 if not.

We fit a Poisson model and the output from R looks like this:

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

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
>
```

Interpretation

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].

Multivariable Poisson regression

For multivariable Poisson regression we include both explanatory variables (age and badh) in the model and thus the R output is the following:

```
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].

Now let's check model's adequacy. To assess the adequacy of the Poisson model we should first look at the basic descriptive statistics for the event count data. If the count mean and variance are very different (should be equivalent in a Poisson distribution) then the model is likely to be over-dispersed. This means that the estimates are correct, but the standard errors (standard deviation) are wrong and unaccounted for by the model. Without adjusting for overdispersion, we use incorrect, artificially small standard errors leading to artificially small p-values for model coefficients.

Calculating the mean and variance of numvisit we get: mean(numvisit) = 2.353 and var(numvisit)= 11.982. We notice that mean and variance are quite different and variance > mean, which is an indication of overdispersion (if mean > variance there is under dispersion).

Another way to check over-dispersion is by dividing residual deviance (from the R output) by the degrees of freedom, $\hat{\varphi} = \frac{\text{Deviance}}{df}$. If $\hat{\varphi}$is around 1 then there is no dispersion. If it is greater than 1 then over-dispersion exists and a value less than 1 is an indication of under-dispersion.

From the R output: resid.Deviance / degrees of freedom = 3465.3/1124 = 3.08 indicating that there is over-dispersion in the data.

The Null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean) whereas residual with the inclusion of independent variables.

To assess goodness of fit, we look at the deviance residuals, and apply a χ2 test based on the degrees of freedom. In this example we have

```
> ## Test for GOF: Using deviance residuals
> df=1124 # from model's output
> deviances2 <- residuals(model,type="deviance")
> dev.tvalue <- sum(deviances2^2)
> c(dev.tvalue, 1-pchisq(dev.tvalue,df))
[1] 3465.301    0.000
>
```

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

Quasi-Poisson regression models and Negative Binomial models.

Quasi Poisson Model

The quasi-Poisson model and the negative binomial model can both account for overdispersion.

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 using R we get:

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

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

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
```

We notice that the parameter estimates did not change much (if any at all) compared to the Poisson model but the confidence intervals and p-values are different.

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: (2.530, 3.611), p<0.001].

Patient's age does not affect the number of visits to the doctor [IRR=1.006, 95%CI: (0.998, 1.013), p=0.104].

Negative Binomial model

An alternative approach to modeling over-dispersion in count data is to start from a Poisson regression model and add a multiplicative random effect D to represent unobserved heterogeneity. This leads to the negative binomial regression model. The negative binomial model assumes an explicit likelihood model

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

Applying the Negative Binomial model to our data using R we get:

```
Call:
glm.nb(formula = numvisit ~ badh + age, data = badhealth, init.theta = 0.9974812528,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0304  -1.4361  -0.4152   0.3180   3.9516

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
```

The parameter estimates are a slightly different compared to the Poisson model or Quasi-Poisson model as well as confidence intervals and p-values.

Interpretation

Patients who claim to be in bad health have about 3.026 times more visits to the doctor compared to patients who claim that are not in bad health, adjusted for age [IRR=3.026, 95%CI: (2.444, 3.777), 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: (1.000, 1.013), p=0.041].

Negative binomial models assume the conditional means are not equal to the conditional variances. This inequality is captured by estimating a dispersion parameter that is held constant in a Poisson model. The Poisson model can be considered as nested in the negative binomial model since in negative binomial model an extra dispersion parameter is considered. Thus, we can use a likelihood ratio test to compare these two models.

In our example, we use the likelihood ratio test to compare the negative binomial and the Poisson model as shown below.

```
> library(lmtest)
> lrtest(model_nb, model)
Likelihood ratio test

Model 1: numvisit ~ badh + age
Model 2: numvisit ~ badh + age
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   4 -2233.6
2   3 -2816.3 -1 1165.3  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Warning message:
In modelUpdate(objects[[i - 1]], objects[[i]]) :
  original model was of class "negbin", updated model is of class "glm"
```

From the test p<0.001 and thus the null hypothesis that both models fit equally well the data is rejected. The negative binomial model fits better than the Poisson model.

AIC for model comparison.

Note that AIC can also be used for model comparison. The smaller AIC the better the model.

Zero inflated models

Another common problem with count data models, including both Poisson and negative binomial models, is that empirical data often show more zeroes than would be expected under either model. Zero-inflated models (either Poisson or Negative Binomial) attempt to account for excess zeroes.

Zero-inflated models are two-component mixture models combining a point mass at zero with a count distribution such as Poisson or negative binomial. In other words, two kinds of zeros are thought to exist in the data, "true zeros" and "excess zeros". Zero-inflated models estimate two equations simultaneously, one for the count model and one for the excess zeros.

Examples of zero-inflated models are:

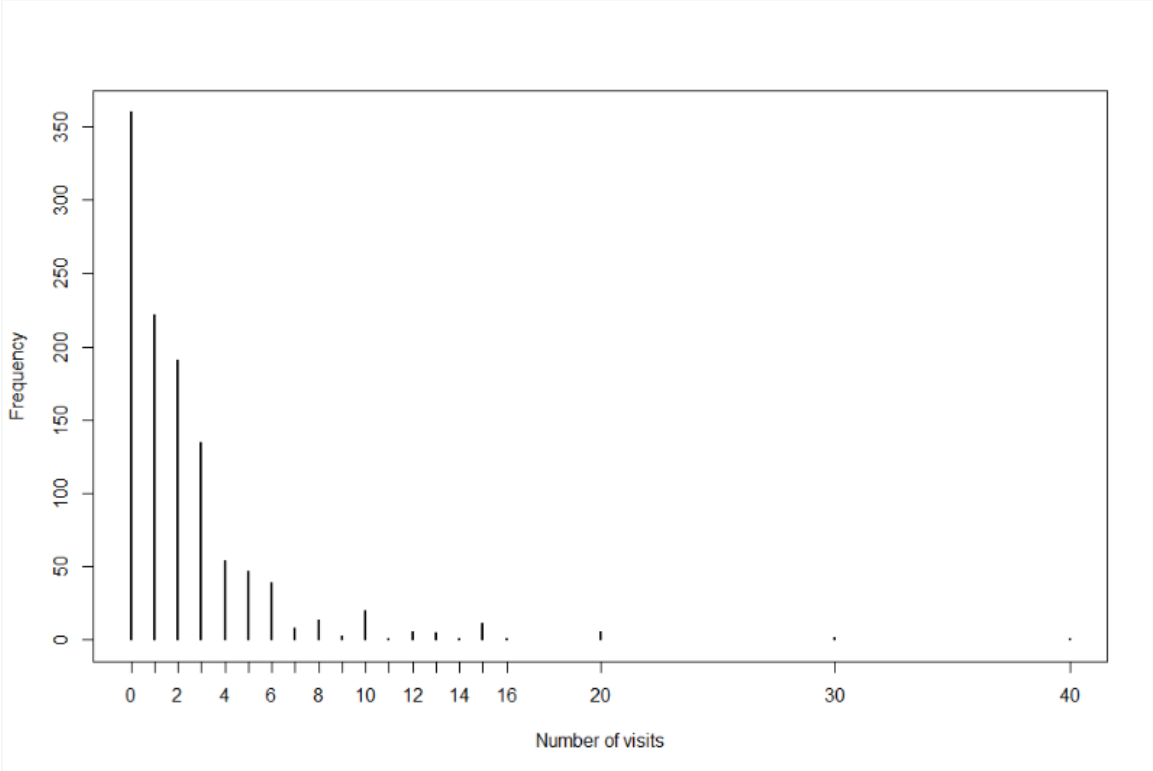
- How many alcoholic drinks did you consume last weekend?
- How many heart attacks or strokes one's had.
- Modelling the factors associated with dental caries.
- Factors associated with the number of reported cases of malaria among under-5 children.

Example

We continue with the same data set and we will check whether a zero-inflated Poisson model will be even a better fit to the data.

In zero-inflated models there are always two categories. The "always zero", which in our example would be individuals who never went to the doctor, and the rest, or "not always zero", for whom the number of visits to the doctor has a Poisson distribution with mean and variance $\lambda > 0$. 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.

Firstly, let's look at the plot of the response variable numvisit. We notice that there are a lot of individuals with zero visits to the doctor.



In R we can check whether there are excess of zeroes in the data with the `performance` package.

```
> library(performance)
> check_zeroinflation(model, tolerance = 0.05)
# Check for zero-inflation

   Observed zeros: 360
  Predicted zeros: 147
            Ratio: 0.41

Model is underfitting zeros (probable zero-inflation).
```

It is suggested that the initial Poisson model might underfit zeroes and thus a zero-inflated model may be a better fit.

From R we get the following results

```
> library(pscl)
> zip_model <- zeroinfl(numvisit ~ badh + age, data=badhealth)
> summary(zip_model)

Call:
zeroinfl(formula = numvisit ~ badh + age, data = badhealth)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.8143 -1.0198 -0.4870  0.4895 12.9158

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
```

There are two models accounting for zero visits. A Poisson model (Count model) and a Logistic Model (Zero-inflation model).

Interpretation

Poisson model

Patients who claim to be in bad health have about 2.403 times more visits to the doctor compared to patients who claim that are not in bad health, adjusted for age [IRR=2.403, 95%CI: (2.187, 2.639), p<0.001].

For every one-year increase in age, the number of visits to the doctor increases by a factor of 1.008, adjusted for bad health. [IRR=1.008, 95%CI: (1.005, 1.013), p<0.001].

Logistic model

The odds that a person who claims to be in bad health will not have a doctor's visit is reduced by 67% compared to those who claim that they are not in bad health, adjusted for age [OR=0.333, 95%CI: (0.187, 0.593), p<0.001].

For every one year increase in age the odds that a person will not visit the doctor is increased by 1.4%, adjusted for bad health. [OR= 1.014, 95%CI: (1.000, 1.028), p=0.042].

Model Fit

The Vuong's test can be used to check whether the zero-inflated model is better than the Poisson model.

In Vuong's test the null hypothesis is that the two models are equally close to the true data generating process, against the alternative that one model is closer.

Applying the Vuong's test in R we have

```
> vuong(model, zip_model)
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
```

It is clear the zero-inflated model fits significantly better than the Poisson model ($p<0.001$).

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

Appendix

Related R code

```
library(COUNT)
library(ggplot2)
library(performance)
library(MASS)
library(lmtest)
library(pscl)

#load data

data(badhealth)
str(badhealth)

# Plot numvisit
plot(table(badhealth$numvisit), xlab = "Number of visits", ylab = "Frequency")

badhealth$badh <- factor(badhealth$badh, levels=c(0,1), labels=c("No", "Yes"))


model1 <- glm(numvisit ~ age, family=poisson(link="log"), data=badhealth)
summary(model1)
exp(coef(model1))
exp(confint(model1))


model2 <- glm(numvisit ~ badh, family=poisson(link="log"), data=badhealth)
summary(model2)
exp(coef(model2))
exp(confint(model2))

model <- glm(numvisit ~ badh + age, family=poisson(link="log"), data=badhealth)
summary(model)
exp(coef(model))
exp(confint(model))

mean(badhealth$numvisi)
var(badhealth$numvisi)

## Test for GOF: Using deviance residuals
df=1124 # from model's output
deviances2 <- residuals(model,type="deviance")
dev.tvalue <- sum(deviances2^2)
c(dev.tvalue, 1-pchisq(dev.tvalue,df))

#Quasi Poisson
model_q <- glm(numvisit ~ badh + age, family=quasipoisson(link = "log"),
data=badhealth)
summary(model_q)
exp(coef(model_q))
exp(confint(model_q))

## Test for GOF: Using deviance residuals
df=1124 # from model_q's output
deviances2 <- residuals(model_q,type="deviance")
dev.tvalue <- sum(deviances2^2)
c(dev.tvalue, 1-pchisq(dev.tvalue,df))

#Negative Binomial
model_nb <- glm.nb(numvisit ~ badh + age, data=badhealth)
summary(model_nb)
```

```
exp(coef(model_nb))
exp(confint(model_nb))

## Test for GOF: Using deviance residuals
df=1124 # from model_nb's output
deviances2 <- residuals(model_nb,type="deviance")
dev.tvalue <- sum(deviances2^2)
print(c(dev.tvalue, 1-pchisq(dev.tvalue,df)), digits=2)

#likelihood ratio test for Poisson and Negative Binomial
lrtest(model_nb, model)

# check zero-inflation
check_zeroinflation(model, tolerance = 0.05)

zip_model <- zeroinfl(numvisit ~ badh + age, data=badhealth)
summary(zip_model)

exp(coef(zip_model))
exp(confint(zip_model))

#test models
vuong(model, zip_model)
```