

# Generalized Linear Models

## Lecture 10. Count data models III. Models with excess zeros

# Data with excess zeros

There are too many zeros (probability distribution suggests less), thus there are so-called '**false**' zeros and '**true**' zeros

## Motivating examples

- Small claims not reported in non-life insurance
- Defective products in a manufacturing process (Lambert, 1992)
- Absent days because of sick-leave (Lam *et al*, 2006)
- Domestic violence cases (Famoye, Singh, 2006)
- Shark counting via bycatch

Tools and examples about how to deal with zero-modified data:

Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M. (2009). *Mixed Effects Models and Extensions in Ecology with R*

<http://www.springer.com/life+sciences/ecology/book/978-0-387-87457-9>

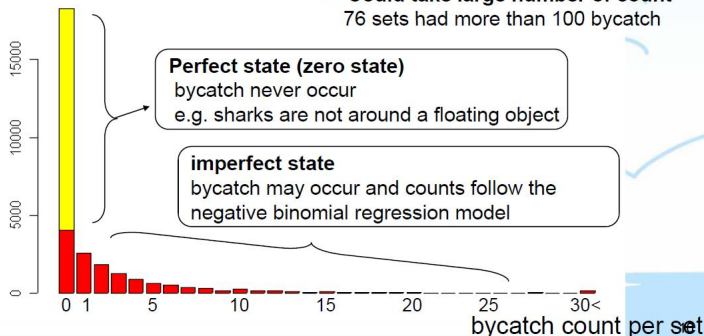
# Example 1. Zero inflation in estimating the shark count

Australia-Japan Workshop on Data Science 2009

## Modeling shark bycatch: The Zero-Inflated Negative Binomial (ZINB) Regression Model (with Smoothing)

Histogram of silky shark bycatch counts

- **Large Proportion of Zeros**  
16375 zero count / 32148 sets = 50.9%
- **Could take large number of count**  
76 sets had more than 100 bycatch



## Example 2: Zero inflation in teeth data

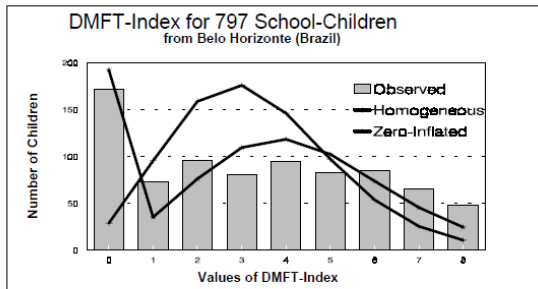
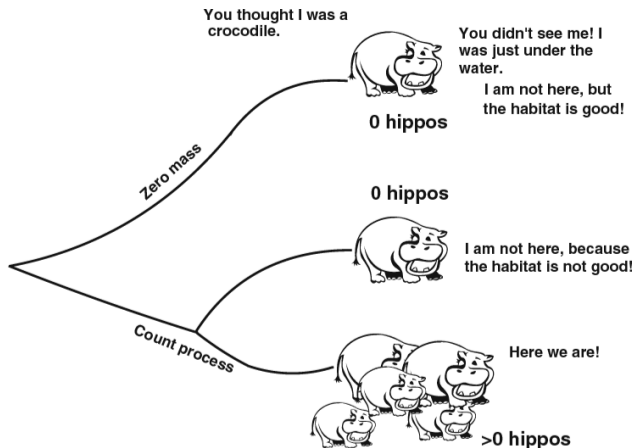


Figure 1: DMFT distribution at begin of study

DMFT index - *Decay, Missing, Filled Teeth* – counts problematic teeth

Source: Böhning, Dietz, Schlattmann (1997). *Zero inflated Count Models and their Applications in Public Health and Social Science*. In: *Applications of Latent Trait and Latent Class Models in the Social Sciences*, Ch 32, p 334.

# Approach 1. Zero inflated model

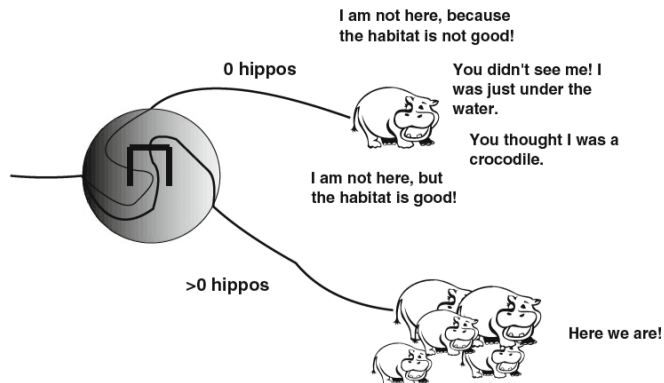


**Fig. 11.5** Sketch of the underlying principle of mixture models (ZIP and ZINB). In counting hippos at sites, one can measure a zero because the habitat is not good (the hippos don't like the covariates), or due to poor experimental design and inexperienced observers (or experienced observers but difficult to observe species)

# Approach 2. Zero altered (*hurdle*) model

11.3 Too Many Zeros

273



**Fig. 11.4** Sketch of a two-part, or hurdle model. There are two processes; one is causing zeros versus non-zeros, the other process is explaining the non-zero counts. This is expressed with the hurdle in the circle; you have to cross it to get non-zero counts. The model does not make a distinction between the different types of zeros

# Zero Inflated (ZI) model. Setup (1)

**Problem:** too many zeros

**Idea:** we divide the data in two imaginary groups:

- First group – only zeros (the **false** zeros). This group is also called the observations with zero mass
- Second group – the count data, which may produce zeros (**true** zeros) and values larger than zero.

Note that

- We are not actively splitting the data in two groups; it is just an assumption that we have these two groups.
- We do not know which of the observations with zeros belong to a specific group.

# Zero Inflated (ZI) model. Setup (2)

Assume now that the data comes from

- the process of (false) zeros with probability  $\pi$ ,  $0 \leq \pi < 1$
- the counting process with probability  $1 - \pi$

Then we have

$$\mathbf{P}\{Y = 0\} = \pi + (1 - \pi)p(0) \quad (*)$$

$$\mathbf{P}\{Y = y\} = (1 - \pi)p(y), \quad y = 1, 2, \dots \quad (**)$$

$p(y)$  – Poisson or NB pmf

The resulting model is a certain mixture of models:

- for the binary part (Bernoulli process) we use *logit* or *probit* link
- for the counting process we apply Poisson or NB model



# ZIP model (*Zero Inflated Poisson*)

Let us start with Poisson model with pmf

$$p(y_i; \mu_i) = \frac{\exp(-\mu_i) \mu_i^{y_i}}{y_i!},$$

i.e. we estimate counts with the Poisson model  $\mu_i = \mu_i(\beta) = \exp(\mathbf{x}_i^T \beta)$

Let us also have a process generating zeros:  $\pi_i = \pi_i(\gamma)$

Now the relations (\*) and (\*\*) from previous slide allow us to write

ZIP model

$$\begin{aligned}\mathbf{P}\{Y_i = 0\} &= \pi_i(\gamma) + (1 - \pi_i(\gamma)) \exp(-\mu_i(\beta)) \\ \mathbf{P}\{Y_i = y_i\} &= (1 - \pi_i(\gamma)) \frac{\exp(-\mu_i(\beta)) [\mu_i(\beta)]^{y_i}}{y_i!}, \quad y_i = 1, 2, \dots\end{aligned}$$

$\beta$  – parameter vector of the counting process

$\gamma$  – parameter vector the process generating zeros

# ZIP model: mean and variance

Since  $\mathbf{P}(\mathbf{Y}_i = y_i) = (1 - \pi_i)p(y_i)$ , the mean of the response is:

$$\mathbf{E}(Y_i) = \mu_i(1 - \pi_i)$$

and to find the variance, we use

$$\mathbf{D}(Y_i) = \mathbf{E}(Y_i^2) - (\mathbf{E}Y_i)^2 = (1 - \pi_i)(\mu_i + \mu_i^2) - (1 - \pi_i)^2 \mu_i^2$$

Thus the variance of the response is

$$\mathbf{D}(Y_i) = \mu_i(1 - \pi_i)(1 + \mu_i\pi_i)$$

Variance is greater than the mean:  $\mathbf{D}(Y_i) > \mathbf{E}(Y_i)$

$\Rightarrow$  excessive number of (false) zeros causes overdispersion!

# Estimation of ZIP model

To estimate parameter  $\gamma$ , often a *logit*-model is used (**to estimate the probability of zeros!**):

$$\ln \frac{\pi_i(\gamma)}{1 - \pi_i(\gamma)} = \mathbf{w}_i^T \gamma, \quad \pi_i(\gamma) = \frac{\exp(\mathbf{w}_i^T \gamma)}{1 + \exp(\mathbf{w}_i^T \gamma)}$$

If the zero-model is *logit*, the mean of ZIP model is

$$\mathbf{E}(Y_i) = \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{w}_i^T \gamma)}$$

Another common choice for estimating  $\gamma$  is *probit*-model  $\pi_i(\gamma) = \Phi(\mathbf{w}_i^T \gamma)$

Now, the arguments of the ZIP model are divided into two groups:

- arguments of the counting process,  $\mathbf{x}$ :

$$\mu_i(\beta) = \mu_i(\mathbf{x}_i^T \beta)$$

- arguments of the zero-generating process,  $\mathbf{w}$ :

$$\pi_i(\gamma) = \pi_i(\mathbf{w}_i^T \gamma)$$

Usually  $\mathbf{x}_i$  and  $\mathbf{w}_i$  differ

# Log-likelihood for ZIP model

Log-likelihood of ZIP model consists of log-likelihood of zero-model and log-likelihood of Poisson model

If zero-model is *logit*:

$$\begin{aligned} l(\mathbf{y}; \beta, \gamma) = & \sum_{i: y_i=0} \ln[\exp(\mathbf{w}_i^T \gamma) + \exp(-\exp(\mathbf{x}_i^T \beta))] \\ & + \sum_{i: y_i>0} [y_i \mathbf{x}_i^T \beta - \exp(\mathbf{x}_i^T \beta) - \ln y_i!] - \sum_{i=1}^n \ln[1 + \exp(\mathbf{w}_i^T \gamma)] \end{aligned}$$

If zero-model is *probit*:

$$\begin{aligned} l(\mathbf{y}; \beta, \gamma) = & \sum_{i: y_i=0} \ln[\Phi(\mathbf{w}_i^T \gamma) + (1 - \Phi(\mathbf{w}_i^T \gamma)) \exp(-\exp(\mathbf{x}_i^T \beta))] \\ & + \sum_{i: y_i>0} [\ln(1 - \Phi(\mathbf{w}_i^T \gamma)) - \exp(\mathbf{x}_i^T \beta) + y_i \mathbf{x}_i^T \beta - \ln y_i!] \end{aligned}$$

## Example. Australian doctor visits

The dataset contains information for approximately 5,000 Australian individuals about the number and possible determinants of doctor visits that were made during a two-week interval.

Variables used for modelling:

- doctorco – response variable, the number of visits
- sex – 0/1 (male/female)
- age – age/100 (people over 72 are coded to age 72)
- illness - number of illnesses during 2 weeks (1, .., 5; over 5 coded to 5)
- income - income (in 1000AUD)
- hscore - health score (bigger score means worse health)

Arguments used:

- In zero-model: only age was used (as empirical studies show that younger people tend to not go to a doctor)
- In counting model: all mentioned arguments

# How do we know that we have excess zeros? (1)

Simplest way is to compare the amount of zeros in data with the amount estimated by model:

```
# Apply Poisson and NB models
modelP=glm(doctorco~sex+age+illness+hscore,family="poisson",data=docvisit)
library(MASS)
modelNB=glm.nb(doctorco~sex+age+illness+hscore,data=docvisit)

data_counts = table(docvisit$doctorco) # actual counts in data

# counts by Poisson model
lambda=fitted(modelP)
modelP_counts = NA
for (i in (0:9)) {
  modelP_counts[i+1] = nrow(docvisit)*mean(dpois(i,lambda=lambda))
}
```

## How do we know that we have excess zeros? (2)

```
# counts by NB model
mu = fitted(modelNB)
k = modelNB$theta
modelNB_counts = NA
for (i in (0:9)) {
  modelNB_counts[i+1] = nrow(docvisit)*mean(dnbinom(i,mu=mu,size=k))
}
```

```
# Comparison of counts
> rbind(data_counts,modelP_counts,modelNB_counts)
```

	0	1	2	3	4
data_counts	4141.000	782.000	174.0000	30.00000	24.000000
modelP_counts	3923.240	1027.489	192.3367	36.83231	7.821768
modelNB_counts	4162.377	711.002	193.1876	66.62437	27.385283

	5	6	7	8	9
data_counts	9.000000	12.000000	12.00000000	5.00000000	1.000000000
modelP_counts	1.768392	0.401649	0.08806722	0.01824272	0.003535122
modelNB_counts	12.849942	6.654180	3.70833122	2.18477570	1.343859220

# Example solution in R. ZIP model (1)

```
> library(pscl)
> modelZIP1 = zeroinfl(doctorco~sex+age+illness+income+hscore | age,
                        dist="poisson", link="logit", data=docvisit)
> summary(modelZIP1)
...
Count model coefficients (poisson with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.92742    0.14339  -6.468 9.95e-11 ***
sex          0.12474    0.06265   1.991  0.0465 *
age         -0.20144    0.20192  -0.998  0.3185
illness      0.23971    0.02013  11.906 < 2e-16 ***
income     -0.16805    0.09118  -1.843  0.0653 .
hscore       0.08775    0.01006   8.723 < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0945    0.1673   6.543 6.03e-11 ***
age          -2.3300    0.3654  -6.376 1.82e-10 ***
...
Number of iterations in BFGS optimization: 17
Log-likelihood: -3500 on 8 Df
```



## Example solution in R. ZIP model (1)

```
> modelZIP2 = zeroinfl(doctorco~sex+illness+hscore | age,  
                        dist="poisson",link="logit", data=docvisit)  
> summary(modelZIP2)
```

...

Count model coefficients (poisson with log link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.13238	0.07611	-14.878	<2e-16 ***
sex	0.14999	0.06029	2.488	0.0129 *
illness	0.24005	0.01991	12.056	<2e-16 ***
hscore	0.08948	0.01002	8.933	<2e-16 ***

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

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.0164	0.1297	7.836	4.64e-15 ***
age	-2.1570	0.2690	-8.019	1.07e-15 ***

...

Number of iterations in BFGS optimization: 14

Log-likelihood: -3502 on 6 Df

# ZINB model (*Zero Inflated Negative Binomial* )

Idea is very similar to ZIP model, the difference is that the Poisson count model is substituted with the NB model

Let us recall the pmf of NB distribution:

$$p(y_i; \mu_i, k) = \frac{\Gamma(k + y_i)}{y_i! \Gamma(k)} \left( \frac{k}{k + \mu_i} \right)^k \left( 1 - \frac{k}{k + \mu_i} \right)^{y_i}$$

We again have

- a counting process, with  $\mu_i = \mu_i(\beta) = \exp(\mathbf{x}_i^T \beta)$
- a process generating zeros  $\pi_i(\gamma)$

## ZINB model

$$\mathbf{P}\{Y_i = 0\} = \pi_i(\gamma) + [1 - \pi_i(\gamma)] \left( \frac{k}{k + \mu_i(\beta)} \right)^k$$

$$\mathbf{P}\{Y_i = y_i\} = [1 - \pi_i(\gamma)] p(y_i; \mu_i, k), \quad y_i = 1, 2, \dots$$

$\beta$  – parameter vector of the counting process

$\gamma$  – parameter vector the process generating zeros

Main choices to estimate  $\pi_i$  are again *logit* or *probit* model

# Mean and variance of ZINB model

Mean of the response variable is

$$\mathbf{E}(Y_i) = \mu_i(1 - \pi_i)$$

Now, assuming the zero model is *logit*, we can proceed and obtain

$$\mathbf{E}(Y_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{w}_i^T \boldsymbol{\gamma})}$$

Variance of the response variable is

$$\mathbf{D}(Y_i) = \mu_i(1 - \pi_i)\left(\mu_i + \frac{\mu_i^2}{k}\right) + \mu_i^2(\pi_i^2 + \pi_i)$$

Clearly  $\mathbf{D}(Y_i) > \mathbf{E}(Y_i)$

## Example solution in R. ZINB model (1)

```
> modelZINB1=zeroinfl(doctorco~sex+age+illness+income+hscore|age,
  dist="negbin", link="logit", data=docvisit)
> summary(modelZINB1)
...
Count model coefficients (negbin with log link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.91238    0.19185  -9.968  < 2e-16 ***
sex           0.20288    0.07085   2.863  0.004191 **
age           0.27688    0.25984   1.066  0.286614
illness       0.27450    0.02397  11.453  < 2e-16 ***
income       -0.15122    0.10311  -1.467  0.142500
hscore        0.10969    0.01355   8.096  5.66e-16 ***
Log(theta)   -0.38889    0.10685  -3.640  0.000273 ***
Zero-inflation model coefficients (binomial with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.7688     0.8535   0.901   0.3677
age           -8.8293     4.0542  -2.178   0.0294 *
...
Log-likelihood: -3381 on 9 Df
```

## Example solution in R. ZINB model (2)

```
> modelZINB2 = zeroinfl(doctorco~sex+illness+hscore | age,
                        dist="negbin",link="logit",data=docvisit)
> summary(modelZINB2)
...
Count model coefficients (negbin with log link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.85496    0.08453 -21.945  < 2e-16 ***
sex           0.23801    0.06887   3.456 0.000549 ***
illness       0.28089    0.02380  11.800  < 2e-16 ***
hscore        0.11050    0.01351   8.181 2.81e-16 ***
Log(theta)   -0.32524    0.10261  -3.170 0.001526 **
Zero-inflation model coefficients (binomial with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8226     0.4855   1.694 0.09023 .
age           -7.4834     2.2866  -3.273 0.00107 **
...
Theta = 0.7224
Number of iterations in BFGS optimization: 26
Log-likelihood: -3384 on 7 Df
```

# Hurdle model (also called ZA (*Zero Altered*) model)

## Setup:

- too many zeros in data
- zeros in count process are not of interest

Two-step process:

- ① binary process that models the probability of event that the counting process starts
- ② counting process (without zeros)

Hurdle (ZA) model

$$\mathbf{P}\{Y_i = 0\} = f_1(0)$$

$$\mathbf{P}\{Y_i = y_i | Y_i > 0\} = f_2(y_i) \frac{1 - f_1(0)}{1 - f_2(0)}, \quad y_i = 1, 2, \dots \quad (***)$$

where  $f_1$  and  $f_2$  are some pmf-s (corresponding to binary process and count process)

# Deriving the probabilities for Hurdle model

To explain the formula (\*\*), let us look at the setup in more details:

- ① First step: a binary process  $C$ :
  - $C = 0$  – counting is not yet started,  $\mathbf{P}\{C = 0\} = f_1(0)$
  - $C = 1$  – 'hurdle is crossed' and counting is started,  $\mathbf{P}\{C = 1\} = 1 - f_1(0)$
- ② Second step: the conditional distribution of the counting process (given  $C = 1$ , i.e. 'hurdle is crossed') is found:

$$\mathbf{P}\{Y_i = y_i | C = 1\} = \frac{f_2(y_i)}{1 - f_2(0)}, \quad y_i = 1, 2, \dots,$$

where  $f_2$  is the (non-conditional) pmf of the counting process

Now, since

$\mathbf{P}\{Y_i = y_i\} = \mathbf{P}\{Y_i = y_i | C = 1\} \mathbf{P}\{C = 1\} = \mathbf{P}\{Y_i = y_i | C = 1\} (1 - f_1(0))$ ,  
the equation (\*\*) follows:

$$\mathbf{P}\{Y = y_i | Y > 0\} = \frac{f_2(y_i)}{1 - f_2(0)} (1 - f_1(0))$$

# Poisson Hurdle (ZAP) model

Notation and assumptions:

- $f_2$  – pmf of count model (Poisson)
- $f_1$  – pmf of zero model (logistic/normal: *logit/probit*-link for binary model)

## Poisson Hurdle (ZAP) logit model

$$p(y_i, \mu_i | y_i > 0) = \frac{[1 - \mathbf{P}\{Y_i = 0\}] \exp(-\mu_i) \mu_i^{y_i}}{[1 - \exp(-\mu_i)] y_i!} = \frac{\exp(-\mu_i) \mu_i^{y_i}}{[1 + \exp(\eta_{i0})][1 - \exp(-\mu_i)] y_i!}$$

where

- $\mathbf{P}\{Y_i = 0\} = \frac{\exp(\eta_{i0})}{1 + \exp(\eta_{i0})}$
- $\eta_{i0} = \ln \frac{\pi_i}{1 - \pi_i} = \mathbf{w}_i^T \boldsymbol{\gamma}$ ,  $\pi_i = \mathbf{P}\{Y_i = 0\}$ ,  $\mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$
- $\mathbf{w}_i$  – argument vector for zero model
- $\mathbf{x}_i$  – argument vector for count model

The derivation of NB Hurdle model (*ZANB*) is analogous



# Mean and variance of ZA models

ZAP model:

$$\mathbf{E}(Y_i) = \frac{1 - \pi_i}{1 - \exp(-\mu_i)} \mu_i$$

$$\mathbf{D}(Y_i) = \frac{1 - \pi_i}{1 - \exp(-\mu_i)} (\mu_i + \mu_i^2) - \left( \frac{1 - \pi_i}{1 - \exp(-\mu_i)} \mu_i \right)^2$$

ZANB model:

$$\mathbf{E}(Y_i) = \frac{1 - \pi_i}{1 - P_0} \mu_i, \quad \text{where } P_0 = \left( \frac{k}{\mu_i + k} \right)^k$$

$$\mathbf{D}(Y_i) = \frac{1 - \pi_i}{1 - P_0} \left( \mu_i + \mu_i^2 + \frac{\mu_i^2}{k} \right) - \left( \frac{1 - \pi_i}{1 - P_0} \mu_i \right)^2$$

The mean and variance can be used to calculate Pearson residuals

# Difference of ZI and ZA models

NB! An important difference is that in R

- in ZIP and ZINB, the binomial GLM models the probability of a **false zero versus other types of data**,
- in ZAP and ZANB, the binomial GLM models the probability of **presence versus absence**

Hence, the estimated regression parameters obtained by ZAP and ZANB should have opposite signs compared to those obtained by ZIP and ZINB due to the definition of  $\pi_i$ .

In other words, assuming that  $\pi_i$  corresponds to the probability of zeros, we should interpret the R output for ZA models as either

- $\text{logit}(1 - \pi_i) = \ln \frac{1 - \pi_i}{\pi_i} = \mathbf{w}_i^T \boldsymbol{\gamma}$ , or
- $\text{logit}(\pi_i) = \ln \frac{\pi_i}{1 - \pi_i} = -\mathbf{w}_i^T \boldsymbol{\gamma}$

## Example solution in R. ZAP model (1)

```
> modelZAP1 = hurdle(doctorco~sex+age+illness+income+hscore | age,
  dist="poisson",link="logit",data=docvisit)
> summary(modelZAP1)
...
Count model coefficients (truncated poisson with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.28073      0.16843  -1.667 0.095574 .
sex          -0.13048      0.08908  -1.465 0.142996
age          -0.05724      0.21614  -0.265 0.791133
illness       0.10324      0.02931   3.523 0.000427 ***
income       -0.33740      0.14077  -2.397 0.016539 *
hscore        0.06879      0.01265   5.436 5.44e-08 ***
Zero hurdle model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.16884      0.08337  -26.02  <2e-16 ***
age          1.85287      0.16727   11.08  <2e-16 ***
...
Number of iterations in BFGS optimization: 14
Log-likelihood: -3619 on 8 Df
```

## Example solution in R. ZAP model (2)

```
> modelZAP2=hurdle(doctorco~illness+hscore+income | age,
                    dist="poisson",link="logit",data=docvisit)
> summary(modelZAP2)
...
Count model coefficients (truncated poisson with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.41924    0.10783  -3.888 0.000101 ***
illness      0.10060    0.02864   3.512 0.000445 ***
hscore       0.06991    0.01258   5.557 2.74e-08 ***
income      -0.27016    0.12831  -2.105 0.035250 *
Zero hurdle model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.16884    0.08337  -26.02  <2e-16 ***
age          1.85287    0.16727   11.08  <2e-16 ***
...
Number of iterations in BFGS optimization: 16
Log-likelihood: -3621 on 6 Df
```

# Example solution in R. ZANB model (1)

```
> modelZANB1=hurdle(doctorco~sex+age+illness+income+hscore | age,
  dist="negbin",link="logit",data=docvisit)
> summary(modelZANB1)
...
Count model coefficients (truncated negbin with log link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.83240    37.96271  -0.285  0.7754
sex          -0.11067     0.14721  -0.752  0.4522
age          -0.24472     0.36378  -0.673  0.5011
illness       0.12985     0.04986   2.605  0.0092 **
income       -0.33252     0.20842  -1.595  0.1106
hscore        0.10451     0.02617   3.994 6.5e-05 ***
Log(theta)  -10.75455    37.96275  -0.283  0.7770
Zero hurdle model coefficients (binomial with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.16884     0.08337 -26.02 <2e-16 ***
age          1.85287     0.16727  11.08 <2e-16 ***
...
Theta: count = 0
Number of iterations in BFGS optimization: 32
Log-likelihood: -3490 on 9 Df
```

## Example solution in R. ZANB model (2)

```
> modelZANB2=hurdle(doctorco~illness+hscore | age,
                     dist="negbin",link="logit",data=docvisit)
> summary(modelZANB2)
...
Count model coefficients (truncated negbin with log link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.65889    47.28034  -0.247  0.80522
illness      0.13020     0.04832   2.695  0.00705 **
hscore       0.10722     0.02609   4.110 3.96e-05 ***
Log(theta)  -11.22227    47.28088  -0.237  0.81238
Zero hurdle model coefficients (binomial with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.16884     0.08337 -26.02  <2e-16 ***
age           1.85287     0.16727  11.08  <2e-16 ***
...
Theta: count = 0
Number of iterations in BFGS optimization: 61
Log-likelihood: -3491 on 6 Df
```

# Which models to choose?

Different aspects that need to be considered while choosing models for count data:

- overdispersion in data (including non-zero part): Poisson vs NB
- too many zeros (overdispersion in zero part), depending on situation: *ZI* or *ZA* models

5 different types of zeros are discussed in (Zuur et al., 2009, p 270)

- ① *structural* zeros – true zeros
- ② *design errors* – e.g., wrong area or wrong season for counting
- ③ *observer errors*
- ④ 'errors' of the subject of counting – the habitat is suitable, but site is not used
- ⑤ *naughty naughts; bad zeros* – obvious errors (need to be deleted)

Type 1 (*true negative*), is taken into account by *ZI* models

Types 2–4 are false negatives, which actually are not of interest

# Which model fits best?

Different tools are available to compare the models:

- AIC, BIC
- Likelihood ratio test – *for nested models*  
Poisson vs NB; ZIP vs ZINB
- **Vuong's** test (*closeness test*) – *for nonnested models*  
ZIP vs Poisson; ZINB vs NB



# Vuong's test

Hypothesis that 2 models  $f : F_\theta$  and  $g : G_\gamma$  are close:  $H_0 : E(\ln \frac{f(y)}{g(y)}) = 0$

The corresponding test statistic is a difference of weighted log-likelihoods:

$$V = \frac{\ln LR_n(\hat{\theta}_n, \hat{\gamma}_n)}{n^{1/2}\omega_n} \xrightarrow{a} N(0, 1)$$

$\ln LR_n = l_n^f(\hat{\theta}_n) - l_n^g(\hat{\gamma}_n)$  (log-likelihoods),  $\omega_n$  – weights

Think, e.g.:  $F_\theta$  : ZIP, ZINB     $G_\gamma$  : Poisson, NB

Decision rules (significance level  $\alpha = 0.05$ ):

If  $V > 1.96 \Rightarrow F_\theta$  is better than  $G_\gamma$

If  $V < -1.96 \Rightarrow G_\gamma$  is better than  $F_\theta$

If  $|V| \leq 1.96 \Rightarrow$  models are equally good

Weights:

$$\omega_n^2 = \frac{1}{n} \sum_{t=1}^n \left[ \ln \frac{f(y_t|z_t, \hat{\theta}_n)}{g(y_t|z_t, \hat{\gamma}_n)} \right]^2 - \left[ \frac{1}{n} \sum_{t=1}^n \ln \frac{f(y_t|z_t, \hat{\theta}_n)}{g(y_t|z_t, \hat{\gamma}_n)} \right]^2$$

## Example continued. Comparison of models (1)

```
> library(lmtest)
> lrtest(modelNB,modelP)
```

Likelihood ratio test

Model 1: doctorco ~ sex + age + illness + hscore

Model 2: doctorco ~ sex + age + illness + hscore

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	6	-3385.9			
2	5	-3650.5	-1	529.11	< 2.2e-16 ***

```
> vuong(modelP,modelZIP2)
```

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	-5.481430	model2 > model1	2.1095e-08
AIC-corrected	-5.444507	model2 > model1	2.5974e-08
BIC-corrected	-5.323503	model2 > model1	5.0894e-08

## Example continued. Comparison of models (2)

```
> vuong(modelNB,modelZIP2)
```

```
-----  
              Vuong z-statistic              H_A      p-value  
Raw              5.500703 model1 > model2 1.8914e-08  
AIC-corrected    5.548084 model1 > model2 1.4441e-08  
BIC-corrected    5.703362 model1 > model2 5.8733e-09
```

```
> vuong(modelNB,modelZINB2)
```

```
-----  
              Vuong z-statistic              H_A      p-value  
Raw              -0.5144592 model2 > model1 0.30347  
AIC-corrected    -0.3001882 model2 > model1 0.38202  
BIC-corrected    0.4020304 model1 > model2 0.34383
```

```
> vuong(modelNB,modelZANB2)
```

```
-----  
              Vuong z-statistic              H_A      p-value  
Raw              6.329138 model1 > model2 1.2327e-10  
AIC-corrected    6.329138 model1 > model2 1.2327e-10  
BIC-corrected    6.329138 model1 > model2 1.2327e-10
```

## Example continued. Comparison of models (3)

```
> AIC(modelP,modelZIP2,modelZAP2,modelNB,modelZINB2,modelZANB2)
```

	df	AIC
modelP	5	7310.941
modelZIP2	6	7016.026
modelZAP2	6	7253.176
modelNB	6	6783.834
modelZINB2	7	6781.033
modelZANB2	6	6994.115

**Final decision?**