Modelling Count Variables with R

Zero Truncated Poisson Distribution

Zero-Truncated Poisson Regression

- Zero-truncated Modelling is used to model count data for which the value zero cannot occur.

Examples of Zero-Truncated Model

Example 1.

- A study of length of hospital stay, in days, as a function of age, kind of health insurance and whether or not the patient died while in the hospital.
- Length of hospital stay is recorded as a minimum of at least one day.

Examples of Zero-Truncated Model

Example 2.

- A study of the number of journal articles published by tenured faculty as a function of discipline (fine arts, science, social science, humanities, medical, etc).
- To get tenure faculty must publish, therefore, there are no tenured faculty with zero publications.

Examples of Zero-Truncated Model

Example 3.

- A study by the county traffic court on the number of tickets received by teenagers as predicted by school performance, amount of driver training and gender.
- Only individuals who have received at least one citation are in the traffic court files.

Examples of Zero-Truncated Model

Example 4.

- Consider for example the random variable of the number of items in a shopper's basket at a supermarket checkout line.
- Presumably a shopper does not stand in line with nothing to buy (i.e. the minimum purchase is 1 item), so this phenomenon may follow a ZTP distribution

Zero-Truncated Poisson regression

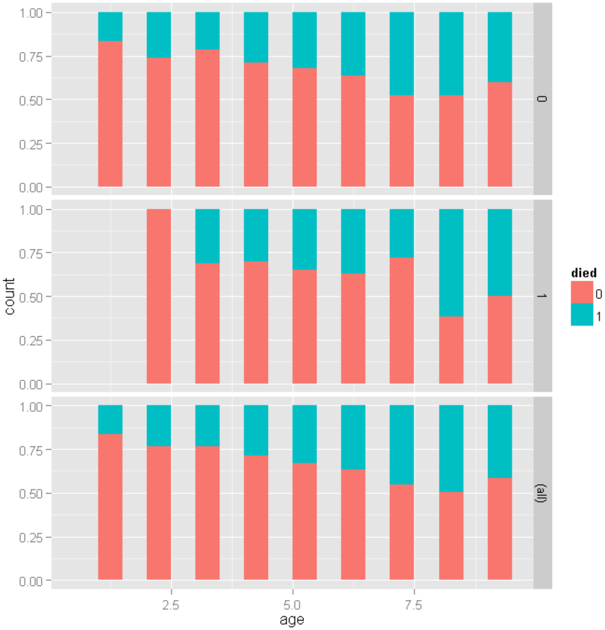
Data Set : hospitalstay

- We have a hypothetical data file, **hospitalstay** with 1,493 observations.
- The length of hospital stay variable is **stay**.
- The variable **age** gives the age group from 1 to 9 which will be treated as interval in this example.
- The variables **hmo** and **died** are binary indicator variables for HMO insured patients and patients who died while in the hospital, respectively.

Zero-Truncated Poisson regression

Data Set : hospitalstay

```
##       stay             age           hmo        died
##  Min.    : 1.00   Min.    :1.00   0:1254   0:981
##  1st Qu.: 4.00   1st Qu.:4.00   1: 239   1:512
##  Median : 8.00   Median :5.00
##  Mean   : 9.73   Mean    :5.23
##  3rd Qu.:13.00   3rd Qu.:6.00
##  Max.   :74.00   Max.    :9.00
```


1.00
0.75
0.50
0.25
0.00
count
0
1
(all)
2.5
5.0
7.5
age
died
0
1

Zero-Truncated Poisson regression

Data Set : hospitalstay

- For the lowest ages, a smaller proportion of people in HMOs died, but for higher ages, there does not seem to be a huge difference, with a slightly higher proportion in HMOs dying if anything.
- Overall, as age group increases, the proportion of those dying increases, as expected.

Zero-Truncated Poisson regression

- To fit the zero-truncated Poisson model, we use the `vglm` function in the VGAM package.
- This function fits a very flexible class of models called **vector generalized linear models** to a wide range of assumed distributions.
- In our case, we believe the data are Poisson, but without zeros.
- Thus the values are strictly positive Poisson, for which we use the positive Poisson family via the `pospoisson` function passed to `vglm`.

Zero-Truncated Poisson regression

Fitting the Model with `R`

We will use the *hospitalstay* data.

```
m1 <- vglm(stay ~ age + hmo + died,
    family = pospoisson(),
    data = hospitalstay)
summary(m1)
```

Zero-Truncated Poisson regression

Fitting the Model with R

Model Summary

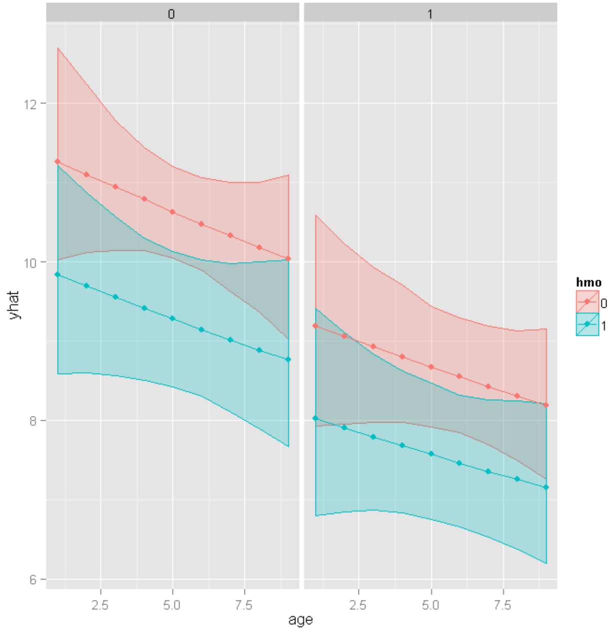
```
## Coefficients:
##               Estimate Std. Error z value
## (Intercept)      2.436      0.027    89.1
## age             -0.014      0.005    -2.9
## hmo1            -0.136      0.024    -5.7
## died1           -0.204      0.018   -11.1
```

Zero-Truncated Poisson regression

- The value of the coefficient for age, -0.0144 suggests that the log count of stay decreases by 0.0144 for each year increase in age.
- The coefficient for hmo, -0.1359 indicates that the log count of stay for HMO patient is 0.1359 less than for non-HMO patients.
- The log count of stay for patients who died while in the hospital was 0.2038 less than those patients who did not die.
- Finally, the value of the constant 2.4358 is the log count of the stay when all of the predictors equal zero.

Zero-Truncated Poisson regression

- Can compute CIs using **boot** package
- Age does not have a significant effect, but hmo and died both do.


0
1
12
10
8
6
yhat
2.5
5.0
7.5
2.5
5.0
7.5
age
hmo
0
1

Zero-truncated negative binomial regression

- Zero-truncated negative binomial regression is used to model count data for which the value zero cannot occur and for which over dispersion exists.

Zero-truncated negative binomial regression

- To fit the zero-truncated negative binomial model, we use the `vglm` function in the VGAM package.
- This function fits a very flexible class of models called vector generalized linear models to a wide range of assumed distributions.
- In our case, we believe the data come from the negative binomial distribution, but without zeros.
- Thus the values are strictly positive poisson, for which we use the positive negative binomial family via the `posnegbinomial` function passed to `vglm`.

Zero-truncated negative binomial regression

Fitting the Model with R

We will use the *hospitalstay* data again.

```
m1 <- vglm(stay ~ age + hmo + died,
   family = posnegbinomial(),
   data = hospitalstay)
```

Zero Truncated Negative Binomial Regression

```
summary(m1)
##
## Call:
## vglm(formula = stay ~ age + hmo + died,
##      family = posnegbinomial(),
##      data = hospitalstay)
##
## Pearson Residuals:
##                Min    1Q Median   3Q Max
## log(munb)   -1.4 -0.70  -0.23 0.45 9.8
## log(size) -14.1 -0.27   0.45 0.76 1.0
```

Zero Truncated Negative Binomial Regression

```
## Coefficients:
##                 Estimate Std. Error z value
## (Intercept):1      2.408      0.072    33.6
## (Intercept):2      0.569      0.055    10.4
## age               -0.016      0.013    -1.2
## hmo1              -0.147      0.059    -2.5
## died1             -0.218      0.046    -4.7
```

Zero Truncated Negative Binomial Regression

-
- The first intercept is what we know as the typical intercept.
- The second is the **over dispersion parameter**, α.
- The number of linear predictors is two, one for the expected mean λ and one for the over dispersion.
- Next the dispersion parameter is printed, assumed to be one after accounting for overdispersion.

Zero Truncated Negative Binomial Regression

- The value of the coefficient for age, -0.0157 suggests that the log count of stay decreases by 0.0157 for each year increase in age.
- The coefficient for hmo, -0.1471 indicates that the log count of stay for HMO patient is 0.1471 less than for non-HMO patients.
- The log count of stay for patients who died while in the hospital was 0.2178 less than those patients who did not die.

Zero Truncated Negative Binomial Regression

- The value of the constant 2.4083 is the log count of the stay when all of the predictors equal zero.
- The value of the second intercept, the over dispersion parameter, α is 0.5686.
- To test whether we need to estimate over dispersion, we could fit a zero-truncated Poisson model and compare the two. (Not Covered).