binomial-gamma-hurdle

March 19, 2019

0.1 Binomial-Gamma Hurdle Models

0.1.1 Model description

Dependent variable, a number of animals observed per minute (y) is semi-continuous (i.e. a point mass in a single value and a continuous distribution elsewhere). The data generating process for this type of data can be modelled using a gamma distribution. The main problem is however that response variable has a high proportion of zeros (96%), which is more than expected from a gamma distribution with, therefore it cannot be readily applied.

Lets consider the two common methods for dealing with zero-inflated data:

(1) Modelling a zero-inflation parameter that represents the probability a given 0 comes from the main distribution (say the negative binomial distribution) or is an excess 0;
(2) Modelling the zero and non-zero data with one model and then modelling the non-zero data with another. This is often called a hurdle model.

In (1), the response variable is modelled as a mixture of a Bernoulli distribution (a point mass at zero) and a Poisson distribution (or any other count distribution supported on non-negative integers). In (2), the basic idea is that a Bernoulli probability governs the binary outcome of whether a variable has a zero or positive realization. If the realization is positive, the hurdle is crossed, and the conditional distribution of the positives is governed by a truncated-at-zero model. Hurdle models model the zeros and non-zeros as two separate processes and can be useful in that they allow you to model the zeros and non-zeros with different predictors or different roles of the same predictors.

Zero-inflation models may be more elegant and informative if the same predictors are thought to contribute to the extra and real zeros.

Hurdle models can be useful in that they allow you to model the zeros and non-zeros with different predictors or different roles of the same predictors. Maybe one process leads to the zero/non-zero data and another leads to the non-zero magnitude.

Here we shall focus on (2) and model the zeros separately from the non-zeros in a binomial-Gamma hurdle model.

0.1.2 Load libraries

```
In [25]: # library(R.utils)
         library(ggplot2)
         # library(GGally)
         # library(lmtest)
```

```
# library(tidyverse)
library(lme4)
library(effects)
library(optimx)
```

In [26]:
```
set.seed(4322)
Sys.time()
```

```
[1] "2019-03-19 15:19:04 GMT"
```

In [15]:
```
# require(devtools)
# install_version("effects", version = "4.0-0")
```

0.1.3 Read in data

In [31]:
```
dat <- read.csv(file = 'data.csv', row.names=1)
# sunfish <- read.csv('ignore/sunfish.csv')
```

Variable y is a response variable, variables x1 and x2 are explanatory variables. Variable x1 represent a number of observers, variable x2 represent an environmental variable (such as sea surface temperature).

In [29]:
```
head(dat)
```

y	x1	x2	year
0	1	10.40875	1971
0	1	10.40875	1971
0	1	10.40875	1971
0	1	10.40875	1971
0	1	10.40875	1971
0	1	10.40875	1971

0.1.4 Scale data

In [39]:
```
# select variables to scale
cols = c("x1", "x2")
# scale variables and add to a df
dat[, paste0(cols, "_", "sc")] <- scale(dat[ ,cols])
summary(dat)
```

```
       y                  x1               x2              year
 Min.   :0.000000   Min.   : 1.000   Min.   : 9.207   Min.   :1971
 1st Qu.:0.000000   1st Qu.: 2.000   1st Qu.:13.169   1st Qu.:1984
 Median :0.000000   Median : 4.000   Median :15.261   Median :1998
 Mean   :0.002676   Mean   : 4.824   Mean   :14.621   Mean   :1996
 3rd Qu.:0.000000   3rd Qu.: 6.000   3rd Qu.:16.288   3rd Qu.:2009
 Max.   :0.132941   Max.   :40.000   Max.   :18.168   Max.   :2017
                    NA's   :1485     NA's   :61
```

```
     x1_sc              x2_sc        
 Min.   :-0.9718   Min.   :-2.5651  
 1st Qu.:-0.7176   1st Qu.:-0.6882  
 Median :-0.2093   Median : 0.3032  
 Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.: 0.2990   3rd Qu.: 0.7894  
 Max.   : 8.9408   Max.   : 1.6802  
 NA's   :1485      NA's   :61       
```

0.1.5 Binomial model

When relating the sightings to temperature what we are interested in detecting are annual trends over and above seasonal fluctuations that we would expect. So we would expect that within each year as temperature increases during spring and summer and zooplankton blooms occur, sunfish sightings will increase. What we want to know is -- in a year when zooplankton abundance and temperatures are high are sunfish sightings also high.

```
In [43]: summary(glm(ifelse(dat$y>0,1,0) ~
                       x1_sc +
                       x2_sc +
                       year,
                       data = dat,
                   family = binomial(link = logit)))
```

```
Call:
glm(formula = ifelse(dat$y > 0, 1, 0) ~ x1_sc + x2_sc + year, 
    family = binomial(link = logit), data = dat)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.3255  -0.2907  -0.2742  -0.2629   2.6607  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 17.079047  13.167616   1.297    0.195
x1_sc        0.018503   0.089106   0.208    0.836
x2_sc        0.051264   0.094037   0.545    0.586
year        -0.010179   0.006605  -1.541    0.123

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1136.3  on 3487  degrees of freedom
Residual deviance: 1133.6  on 3484  degrees of freedom
  (1487 observations deleted due to missingness)
AIC: 1141.6
```

```
Number of Fisher Scoring iterations: 6
```

We see that observer related variables (x1) is highly significant. We try to isolate its effect for each year. We apply a mixed-effect modeling framework and fit a varying intercept model with lmer. This approach is useful when we are interested explicitly in variation among and by groups. Group level variables are specified using a special syntax: (1|year) to fit a linear model with a varying-intercept group effect using the variable year.

We include 'year' as random effect with noise variables.

```
In [45]: m.bin.full.re <- glmer(ifelse(dat$y>0,1,0) ~
                          x1_sc +
                          (1|year) ,
                        data = dat,
         #               control = glmerControl(optimizer ='optimx', optCtrl=list(method='nlminb'
                        family = binomial(link = logit))

In [46]: summary(m.bin.full.re)

Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: ifelse(dat$y > 0, 1, 0) ~ x1_sc + (1 | year)
   Data: dat

     AIC      BIC   logLik deviance df.resid
  1141.4   1159.9   -567.7   1135.4     3487

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.2532 -0.2034 -0.1948 -0.1898  5.5795

Random effects:
 Groups Name        Variance Std.Dev.
 year   (Intercept) 0.05928  0.2435
Number of obs: 3490, groups:  year, 38

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.24648    0.10363 -31.327   <2e-16 ***
x1_sc        0.04109    0.08446   0.486    0.627
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
      (Intr)
x1_sc -0.044
```