Zero-inflation

Ben Bolker

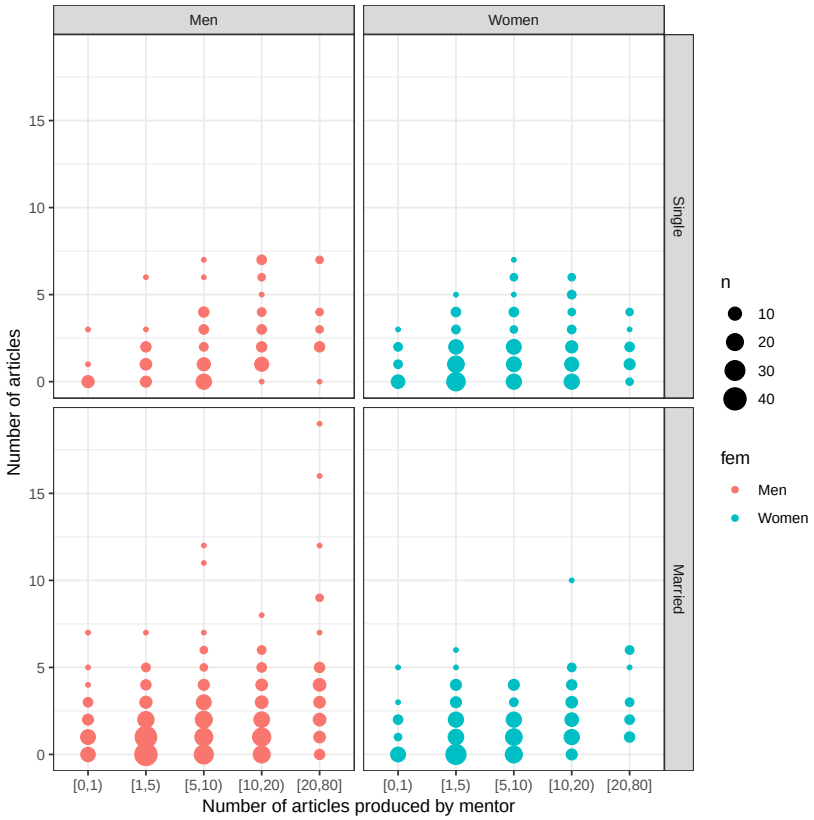
October 23, 2018

Zero-inflation

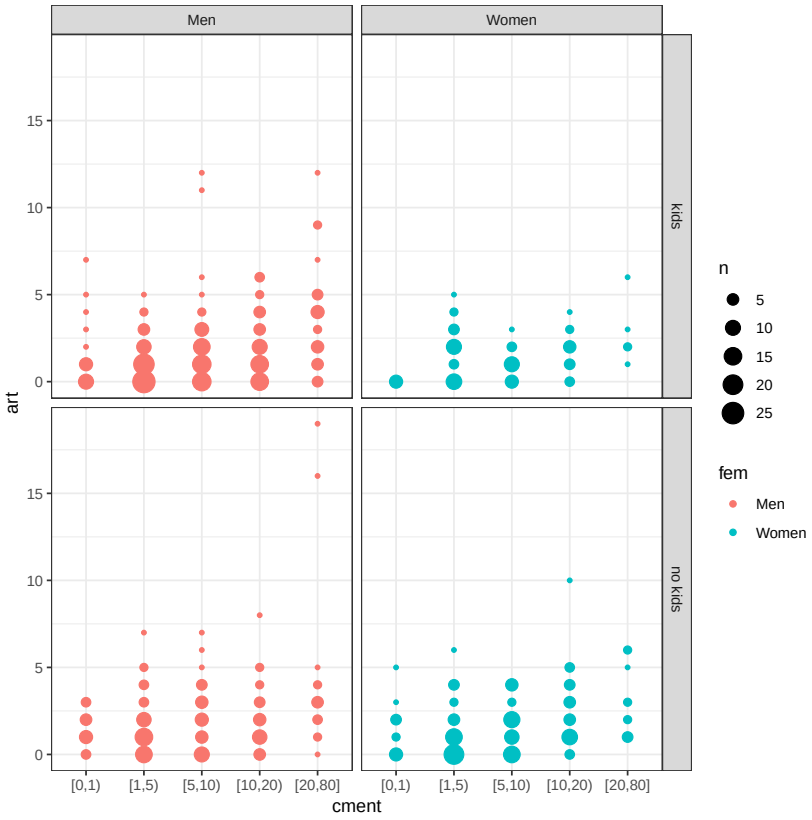
- Models with "too many zeros".
- "Lots of zeros" $\neq$ "zero-inflated" — could just be small mean / large variance (overdispersion)
- Mode at zero plus mode away from zero is *definitely* zero-inflated, however
- Zero-inflated Poisson and negative binomial most common, although zero-inflated binomial is possible
- Zero-inflated *continuous* distributions typically best dealt with as binary + conditional continuous model (or censored model)
- Simplest version, zero-inflation: mixture model. Probability p of *structural* zero, probability $1 - p$ that the variable follows the *conditional* distribution (e.g. if conditional distribution is Poisson, the probability of of a *sampling* zero is $(1 - p)\exp(-\lambda)$. **Please** don't call them "true" and "false" zeros.
- Alternative: *hurdle* model. Zeros lumped together, so we have probability p of zero plus a *truncated* Poisson model (i.e. zeros removed).
- ZI, hurdle models identical for a single sample, but differ in how the covariates act
- Can fit both, but best to use *a priori* reasoning: how do we think zeros are generated?
- Can have separate models (i.e. different subsets of predictors) for the zero-inflation component and the *conditional* distribution
- `pscl` package for simple zero-inflation (ZIP/ZINB); can use `glmmTMB` for mixed models, fancier distributions (e.g. ZINB1)

```
library(pscl)

## Classes and Methods for R developed in the
  ## Political Science Computational Laboratory
  ## Department of Political Science
  ## Stanford University
  ## Simon Jackman
  ## hurdle and zeroinfl functions by Achim Zeileis

data("bioChemists", package = "pscl")
```

```
bioCh2 <- bioChemists
bioCh2$cment <- cut(bioCh2$ment,
                    c(0,1,5,10,20,80),
                    right=FALSE,include.lowest=TRUE)
bioCh2$anykids <- factor(bioCh2$kid5==0,labels=c("kids","no kids"))

library(ggplot2); theme_set(theme_bw())

## Registered S3 methods overwritten by 'ggplot2':
  ##  method         from
  ##  [.quosures     rlang
  ##  c.quosures     rlang
  ##  print.quosures rlang
## Registered S3 method overwritten by 'dplyr':
  ##  method              from
  ##  as.data.frame.tbl_df tibble

ggplot(bioCh2,
       aes(y=art,x=cment,colour=fem))+
  facet_grid(mar~fem)+
  stat_sum(aes(group=cment))+
    labs(x="Number of articles produced by mentor",
         y="Number of articles")
```




```
ggplot(subset(bioCh2,mar=="Married"),
       aes(y=art,x=cment,colour=fem))+facet_grid(anykids~fem)+
  stat_sum(aes(group=cment))
```



Fit logit-Poisson model: `art ~ .` is the same as `art ~ . | .`, or equivalently `art ~ fem + mar + kid5 + phd + ment | fem + mar + kid5 + phd + ment`, i.e. include all terms in both the zero-inflation model and the hurdle model.

```
fm_hp <- hurdle(art ~ ., data = bioChemists)
fm_hnb <- hurdle(art ~ ., data = bioChemists, dist="negbin")
summary(fm_hnb)

##
## Call:
## hurdle(formula = art ~ ., data = bioChemists, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.2581 -0.8036 -0.2497  0.4745  6.2753
##
## Count model coefficients (truncated negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.355125   0.196832   1.804  0.07120 .
## femWomen    -0.244672   0.097218  -2.517  0.01184 *
## marMarried   0.103417   0.109430   0.945  0.34463
```

```
## kid5        -0.153260   0.072229  -2.122  0.03385 *
## phd         -0.002933   0.048067  -0.061  0.95134
## ment         0.023738   0.004287   5.537 3.07e-08 ***
## Log(theta)   0.603472   0.224995   2.682  0.00731 **
## Zero hurdle model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.23680    0.29552   0.801   0.4230
## femWomen    -0.25115    0.15911  -1.579   0.1144
## marMarried   0.32623    0.18082   1.804   0.0712 .
## kid5        -0.28525    0.11113  -2.567   0.0103 *
## phd          0.02222    0.07956   0.279   0.7800
## ment         0.08012    0.01302   6.155 7.52e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta: count = 1.8285
## Number of iterations in BFGS optimization: 15
## Log-likelihood: -1553 on 13 Df
```

```
fm_zip <- zeroinfl(art ~ ., dist="poisson", data = bioChemists)
fm_zinb <- zeroinfl(art ~ ., dist="negbin", data = bioChemists)
library(bbmle)
```

```
## Loading required package:  stats4
```

```
AICtab(fm_zip,fm_zinb,fm_hp,fm_hnb)
```

```
##         dAIC  df
## fm_zinb   0.0 13
## fm_hnb    5.2 13
## fm_zip  107.6 12
## fm_hp   108.6 12
```

Should consider interactions?
To fit the same models in `glmmTMB`,

```
library(glmmTMB)
fm2_zip <- glmmTMB(art ~ fem + mar + kid5 + phd + ment,
                   zi = ~ .,  ## i.e., parameters same as conditional model
                   family=poisson, data = bioChemists)
```

Expectation-maximization:

- fit GL(M)M for zero-inflated part and conditional part of model; latter is with weights $(1 - z)$
- expectation: set zero probability to $u/(u + (1 - u) * exp(-v))$

where u is the zero-inflation probability and v is the Poisson mean