COUNT DATA: POISSON AND ZERO-INFLATED POISSON REGRESSION

Hyun Woo Lim

November 1, 2018

WUSTL, Political science

Outlines

MOTIVATION

Motivation

- Poisson and zero-inflated Poisson regression are used to model count variables.
- Application in political science
  - Poisson: Reeves (2011): the number of presidential disaster declarations
  - Zero-inflated negative binomial: Barber, Canes-Wrone and Thrower (2016), Barber, Canes-Wrone and Thrower (submitted 2018), Auter and Fine (2018): binned $ amount.

Overview

- Poisson model
- Zero-inflated Poisson regression model (ZIP)
- MLE
- How to use it: poisson vs. ZIP
- How to interpret it
- Example

MODELS FOR COUNT DATA

Models for count data

- Poisson regression is a form of Generalized Linear Model (GLM) for count data.
- Zero-inflated Poisson regression a form of Poisson regression which can handle the excess of zeros in the data.

Poisson distribution

- Dependent variable, y is count data. (Some data can be changed as count data.)
- $y \geq 0$
- Poisson distribution has the property that its means and variance are equal

$$P(y|x;\theta) = \frac{e^{-\mu}\mu^{y}}{y!}(y = 0, 1, 2, ..)$$

Poisson regression model

In Poisson regression, we think that by Poisson rate of f μ (y) is decided by k independent variables (x).

$$\mu = exp(\beta_1 x_1 + ...\beta_k x_k)$$

That is

$$\mu_i = exp(\beta_1 x_1 i + ...\beta_k x_k i) = t_i \mu(x_i' \beta)$$

MLE for Poisson regression

So now we estimate coefficients of Poisson regression by MLE. We need likelihood and Log likelihood function for that.

$$L(\beta|y_i, x) = \prod_{i=1}^{n} p(y_i|x, \beta)$$

$$= \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

$$\propto \prod_{i=1}^{n} e^{-\lambda_i} \lambda_i^{y_i}$$

where $\lambda_i = e^{x_i' \beta}$

MLE for Poisson regression

Now get Log likelihood function.

$$lnL(\beta|y_i, x) = ln(\prod_{i=1}^{n} e^{-\lambda_i} \lambda_i^{y_i})$$

$$= \sum_{i=1}^{n} (-\lambda_i + y_i ln\lambda_i)$$

$$= \sum_{i=1}^{n} (-e^{x_i'\beta} + y_i ln e^{x_i'\beta})$$

$$= \sum_{i=1}^{n} (-e^{x_i'\beta} + y_i x_i'\beta)$$

where $\lambda_i = e^{x_i'\beta}$

To finally get MLE of β, we need to find value maximize the Log likelihood function, and R can do this computation.

Estimating β using R

In R, the command glm() has default Iteratively Reweighted least squares (IRWLS) to find MLE of βs.

- Start to find σ_i
- Get weighted least squares using weights, and find estimated regression $\hat{\beta}^0$.
- Estimate variance, $\hat{\sigma_i^2}$, and redo weighted least squares with weights. Then we get $\hat{\beta}^1$.
- Iterate this process again until the estimates are practically the same.

More formally, it is:

$$\beta^{t+1} = \min \sum_{i=1}^{n} w_i(\beta^t)|y_i - f_i(\beta)|^2$$

APPLICATION

Count data: Poisson regression I

Research questions:

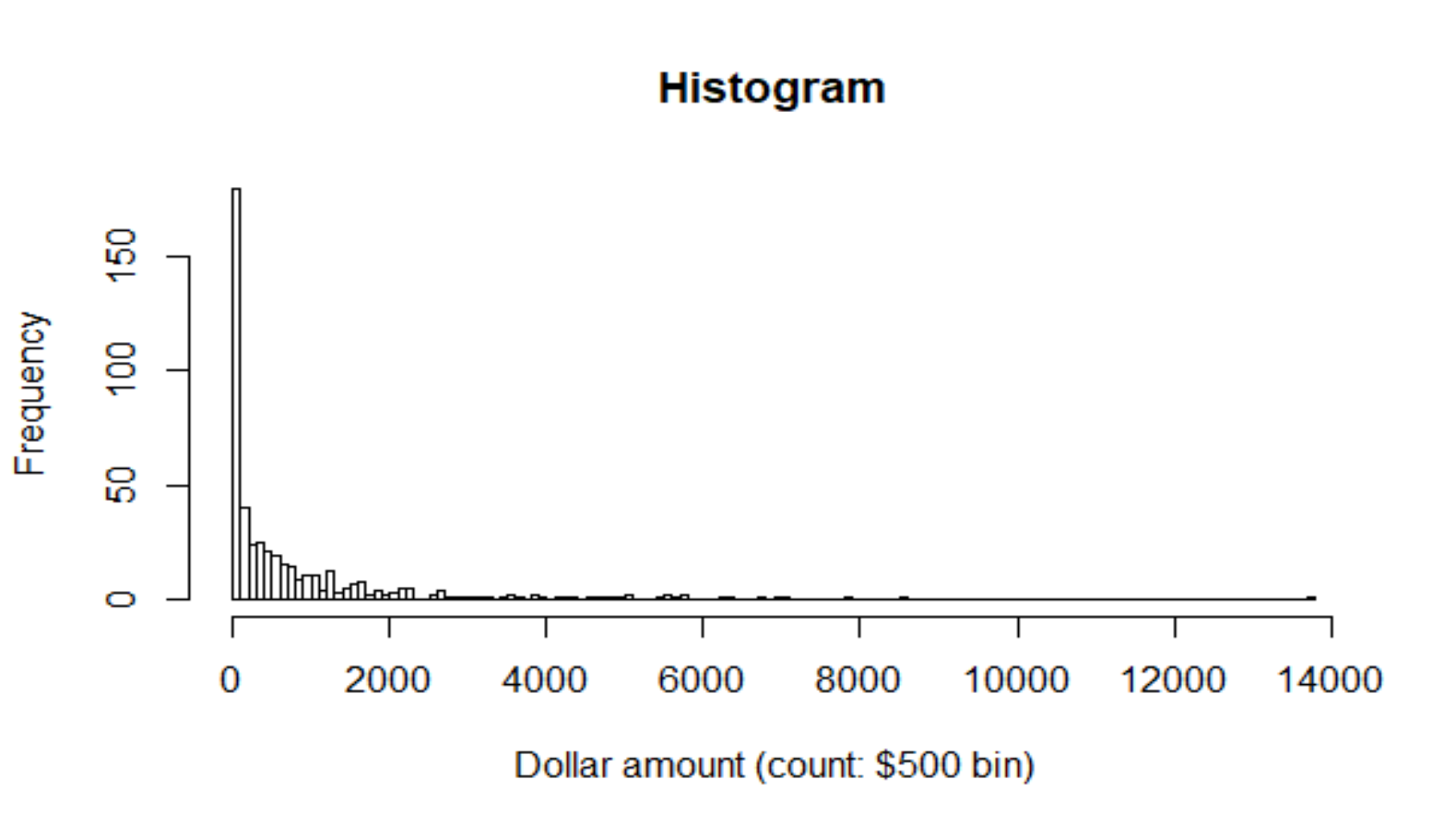
Do the ideological preferences of corporate boards constrain the giving of corporate PACs?

Motivation:

- Why there is so little money in politics?
- How can group's elites be constrained by themselves when they make a collective decision for a group?
- how do partisan preferences and other internal constraints shape interest group strategy?

Count data: Poisson regression II

Check the data: Over-dispersion & excess of zeros



Poisson regression

variables:

- **sum12_count**: Corporate PAC spending in 2012
- **var.dime.cfscore**: variance of corporate boardroom members' ideology
- **sector**: categorical variable of sector of industry

Model:

```
library(MASS)
library(pscl)
m1 <- glm(formula = sum12_count ~ var.dime.cfscore + sector,
              family  = poisson(link = "log"),
              data    = datt)
```

Poisson regression: outcome

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 6.002087   0.009495 632.157   <2e-16 ***
var.dime.cfscore           -0.243072   0.007990 -30.424   <2e-16 ***
sectorCapital Goods         1.399573   0.010057 139.166   <2e-16 ***
sectorConsumer Cyclical    -0.586722   0.016594 -35.358   <2e-16 ***
sectorConsumer Goods        0.989935   0.018189  54.424   <2e-16 ***
sectorConsumer/Non-Cyclical 0.161784   0.012147  13.319   <2e-16 ***
sectorEnergy                0.861043   0.010691  80.536   <2e-16 ***
sectorFinancial             1.051715   0.009423 111.607   <2e-16 ***
sectorHealthcare            1.187664   0.009926 119.648   <2e-16 ***
sectorIndustrial Goods      0.298269   0.033049   9.025   <2e-16 ***
sectorServices              0.356169   0.009820  36.270   <2e-16 ***
sectorTechnology            0.842441   0.010080  83.579   <2e-16 ***
sectorTransportation        1.932137   0.010496 184.082   <2e-16 ***
sectorUtilities             1.129851   0.010231 110.434   <2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How to read the outcome

GLM includes a link function that relates the expected value of the response to the linear predictors in the model. In case of poisson regression, link function, $g(x\beta)$ is log. More formally, When Poisson regression follows $Y_i \sim Poisson(\lambda)$,

$$g(E(Y_i)) = g(\lambda) = ln(\lambda)$$

Where $\lambda = x'\beta$

This means that we need to exponentiate the linear component, $x'\beta$ to find $E(Y_i|x)$. Then, let us go back to the outcome of our model.

Reading Incidence rate ratio

```
> tableone::ShowRegTable(model1)
                     exp(coef) [confint]           p
(Intercept)          404.27    [396.80, 411.85] <0.001
var.dime.cfscore     0.78      [0.77, 0.80]     <0.001
```

My key variable is var.dime.cfscore (diversity within boardroom). We interpret it as follow:
for a change of 1 in diversity score the expected count of spending increases by 0.78, holding all other variables constant. More formally IRR is:

$$exp(\beta_k) = \frac{E(Y_i|x, x_k + 1)}{E(Y_i|x, x_k)}$$

Average marginal effect

In Poisson regression, the better way to read the outcome using R is calculating average marginal effect which gives us the average effect of x on y holding all other variables constant
The basic process in R is:

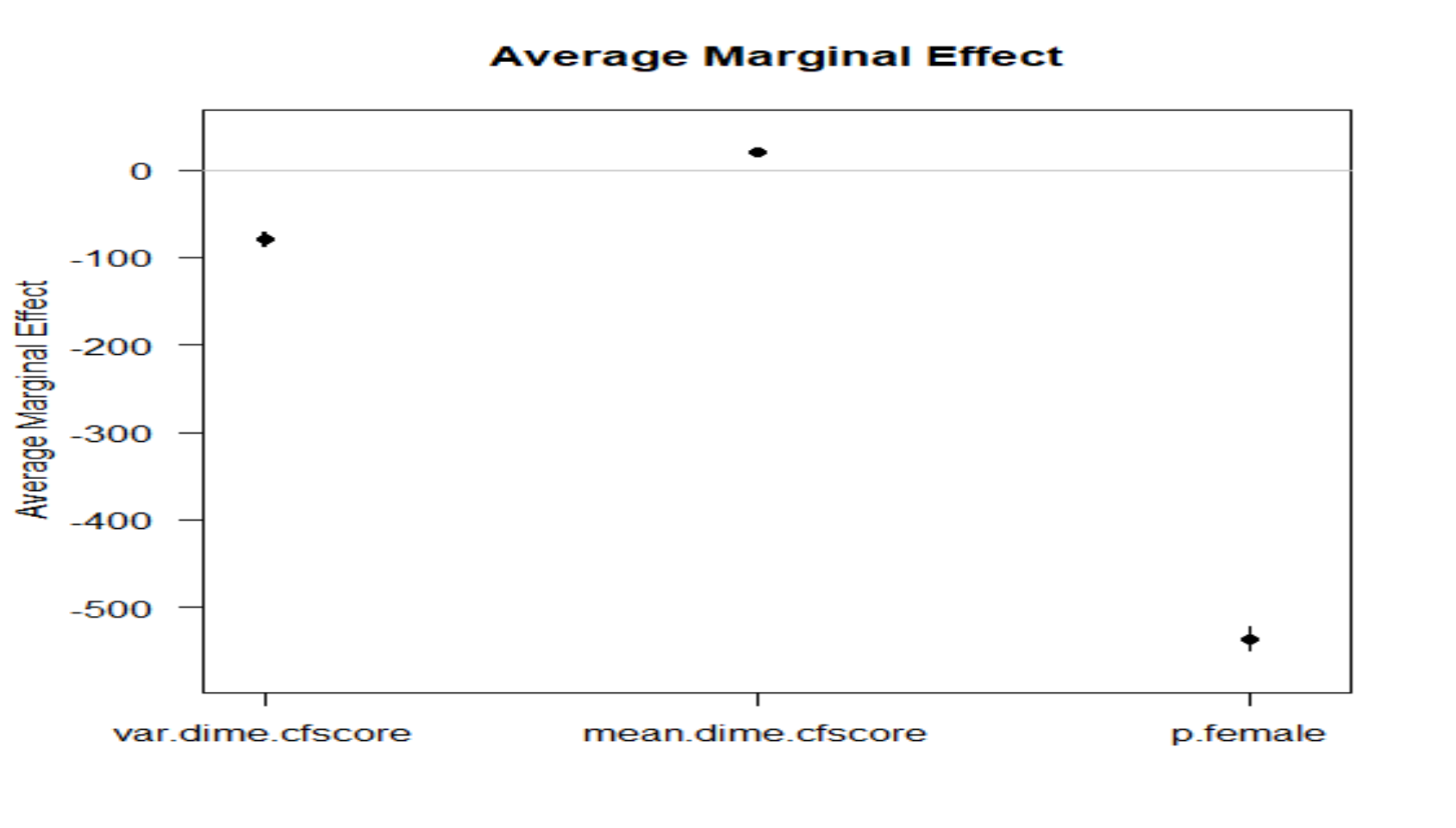
- Calculate expected value of y_i
- Add $\text{sd}(x_i)/1000$ to x_i
- Recalculate expected value of y_i
- Then we calculated $E(y_i)_2 - E(y_i)_1$
- Then we divide it by $\text{d}(x_i)/1000$ again and take mean

In short, average marginal effect gives me the average additive effect on the expected count based on standard deviation. This way is more commonly used in actual papers.

Poisson: average marginal effect

```
summary(m1 <- glm(sum12_count ~
    var.dime.cfscore+mean.dime.cfscore+p.female,
    family="poisson", data=p))
var.sd <- sd(model.frame(m1)$var.dime.cfscore)
predict1 <- predict(m1,type='response')
predict2 <- predict(m1,
                    newdata = data.frame(
                    var.dime.cfscore =
                        model.frame(m1)$var.dime.cfscore+
                        var.sd/1000,
                    mean.dime.cfscore =
                        model.frame(m1)$mean.dime.cfscore,
                    p.female = model.frame(m1)$p.female),
                    type ="response")
print("Average Marginal Effect of Ideological diversity in
    baordroom")
[1] "Average Marginal Effect of Ideological diversity in
    baordroom"
mean((predict2-predict1)/(var.sd/1000))
[1] -535.9046
```

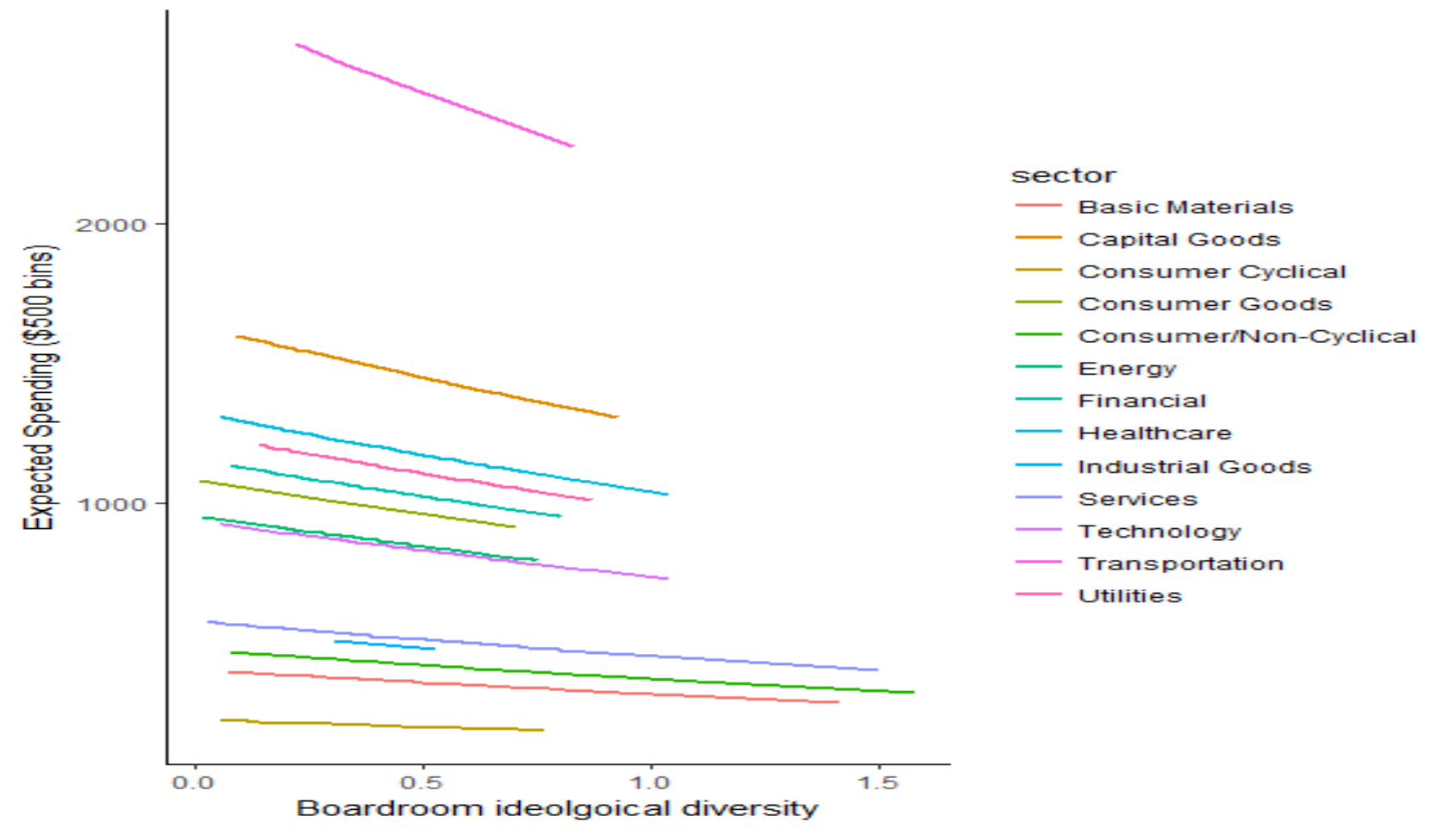
Poisson: average marginal effect graph for boardroom diversity



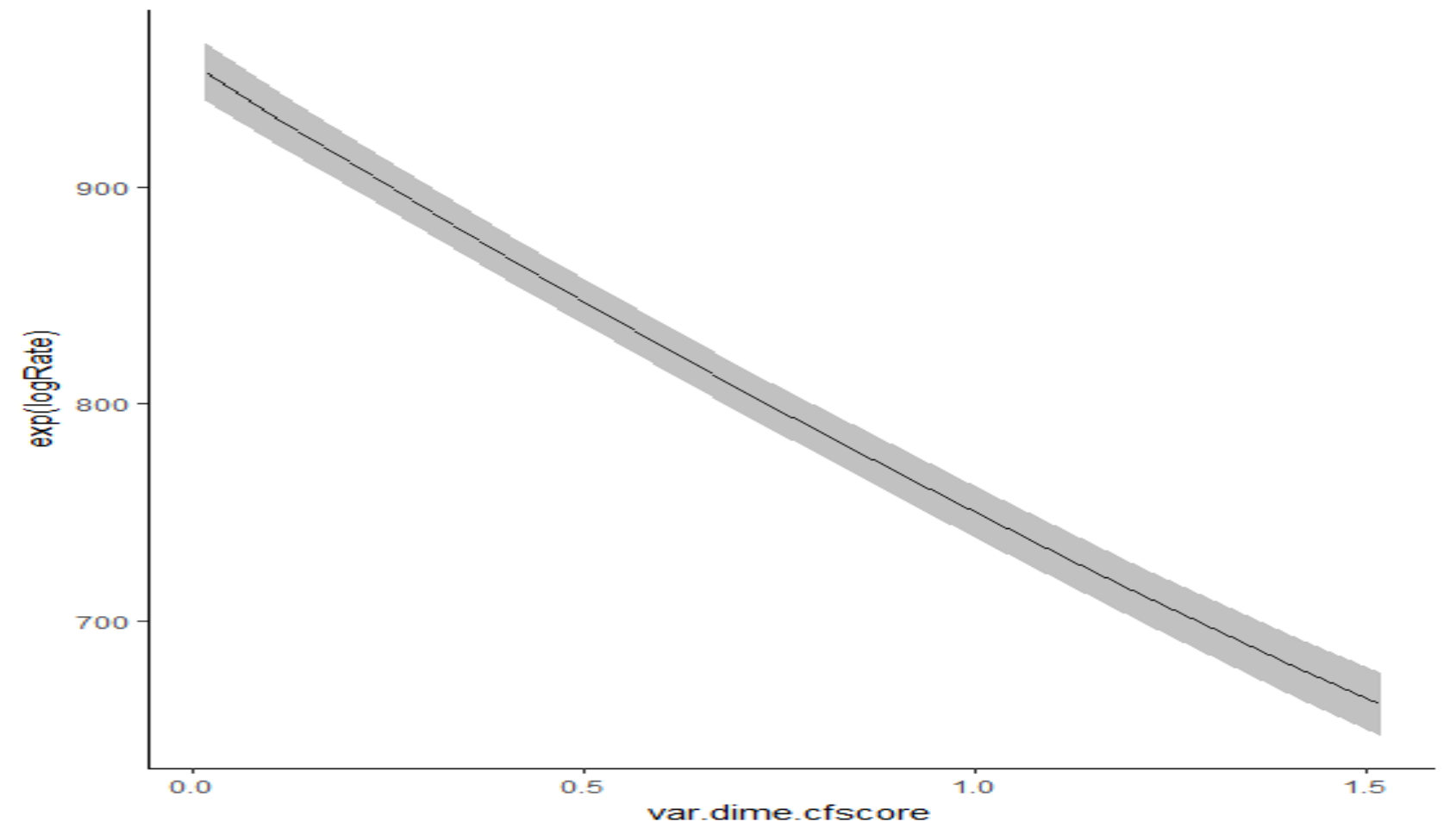
Estimated effect

Lastly, for reading the outcome of Poisson regression, my favorite way of displaying it is using estimated effect plot

Poisson: estimated effects



Poisson: estimated effects: Energy sector only



Zero inflated Poisson regression

In here we assume, data is from two process

- In one process, we model for zeros.
- In the other process, we model nonzeros.

More substantively, when we use zero inflated model, we assume there is something special about zeros in data generating process. (i.e. Some Fortune 500 companies have PAC and do not spend mony and some other PACs do have PAC but did not spend in 2012 election cycle.)

Zero inflated Poisson regression

The Zero inflated Poisson regression can be written as:

$$p(Y_i = y_i|x_i, z_i) = \begin{cases} \theta_i(z_i) + (1 - \theta_i(z_i) poisson(\lambda_i; 0|x_i) \text{ when y =0} \\ (1 - \theta_i(z_i)) poisson(\lambda_i; y_i|x_i) \text{ when y>0} \end{cases}$$

We need to note that in regular Poisson regression model, we assumed the variance and mean are the same. However, in here, Zero inflated Poisson regression, we can relax the assumption and the model can take care of $var(y_i|x_i, z_i) > e(y_i|x_i, z_i)$.

Zero inflated Poisson regression MLE

The Zero inflated Poisson regression MLE can be written as:

$$I_i = \begin{cases} 1 \text{ when y=0} \\ 0 \text{ otherwise} \end{cases}$$

Log-likehood function:

$$logL = \sum_{i=1}^{n} I_i ln(exp(z_i'\gamma) + exp(-exp(x_i'\beta)))$$

$$+ \sum_{i=1}^{n} (1 - I_i)(y_i(x_i)'\beta - exp(x'\beta) - ln(y_i!)) - ln(1 + exp(z'\gamma))$$

This is ugly, but simply what we want is

$$E(y) = (1 - \theta)(\lambda)$$

$$V(y) = \lambda(1 - \theta)(1 + \theta\lambda)$$

Zero inflated Poisson regression in R

In R, we use library pscl.

The default algorithm is BFGS (Broyden–Fletcher–Goldfarb–Shanno algorithm) and this is an iterative method for solving unconstrained nonlinear optimization problems.

ZIP: OUTCOME

```
Count model coefficients (poisson with log link):
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)       6.9076443  0.0050718 1361.962  < 2e-16 ***
var.dime.cfscore -0.1195422  0.0085145  -14.040  < 2e-16 ***
mean.dime.cfscore 0.0351158  0.0049618    7.077 1.47e-12 ***
p.female          0.0126928  0.0001857   68.345  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.01055    0.31634  -3.195 0.001401 **
var.dime.cfscore   1.74103    0.51966   3.350 0.000807 ***
mean.dime.cfscore  0.49180    0.33659   1.461 0.143987
p.female          -0.04722    0.01234  -3.827 0.000130 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 10
Log-likelihood: -2.578e+05 on 8 Df
```

Reading Zero inflated Poisson regression in R

The poisson part (first stage), we interpret the coefficient by exponentiate it as we did in IRR.

The count logit part (second stage), the exponentiated coefficients are odd ratios. That means for a unit change in x, the odds are expected to chagne by a factor of exp(β), holding all other variables constant.

ZIP: OUTCOME

```
> exp(coef(m2))
      count_(Intercept)  count_var.dime.cfscore
             999.8890026               0.8873265
count_mean.dime.cfscore          count_p.female
               1.0357397               1.0127737
       zero_(Intercept)   zero_var.dime.cfscore
               0.3640201               5.7031934
 zero_mean.dime.cfscore           zero_p.female
               1.6352560               0.9538798
> exp(confint(m2))
                              2.5 %       97.5 %
count_(Intercept)       989.9987334 1009.8780774
count_var.dime.cfscore    0.8726416    0.9022585
count_mean.dime.cfscore   1.0257160    1.0458613
count_p.female            1.0124051    1.0131424
zero_(Intercept)          0.1958210    0.6766927
zero_var.dime.cfscore     2.0596010   15.7925803
zero_mean.dime.cfscore    0.8454270    3.1629724
zero_p.female             0.9310912    0.9772261
```

ZIP: READ THE OUTCOMES

The coefficients of count (2nd stage) and zero inflation (1st stage) part of ZIP model

- For Count part, the model explains the amount of spending given conditional on PAC spends money in the election cycle (2012): This is the 2nd stage.
  - It is based on logit model with odds ratio. Therefore, to interpret the coefficients, we need to exponent the coefficients.
  - It means, when there is one unit changes in x, the odds change by exp(coefficient of x), holding all other variables constant.
- For Zero inflation part, the model explains the probability of spending (0 or not). This is the 1st stage.

ZIP: Interpretation

Count model:
Baseline spending is 989 (x $500 bin) among those who can spend money.
One unit increase in baordroom ideological diversity deceases spending by 0.88 times among those who could have spent money.

Zero-inflation model:
If boardroom is more diverse then, the corporate PAC is morely like to spend money.

ZIP: Interpretation2

Recall:
The first-stage equation estimates the likelihood of spending, and the second-stage equation the amount spending.

It means that:
If boardroom is more diverse, then the PAC is more likely to spend money, but among the PACs that spent money, if boardroom is more diverse, then the PAC spend less.

What model to use?

We can compare ZIP over Poisson regression by **Vuong test**.
The statistic tests the null hypothesis that the two models are equally close to the true data generating process, against the alternative that one model is closer. It cannot make any decision whether the ”closer” model is the true model. (i.e. The test only can tell which one is better)

More formally, the test is using statistic,
$P_m i$ is the probability of an observed count for case i from model 1

$$m_i = ln \frac{P_m 1}{p_m 2}$$

So, Vuong test is formally,

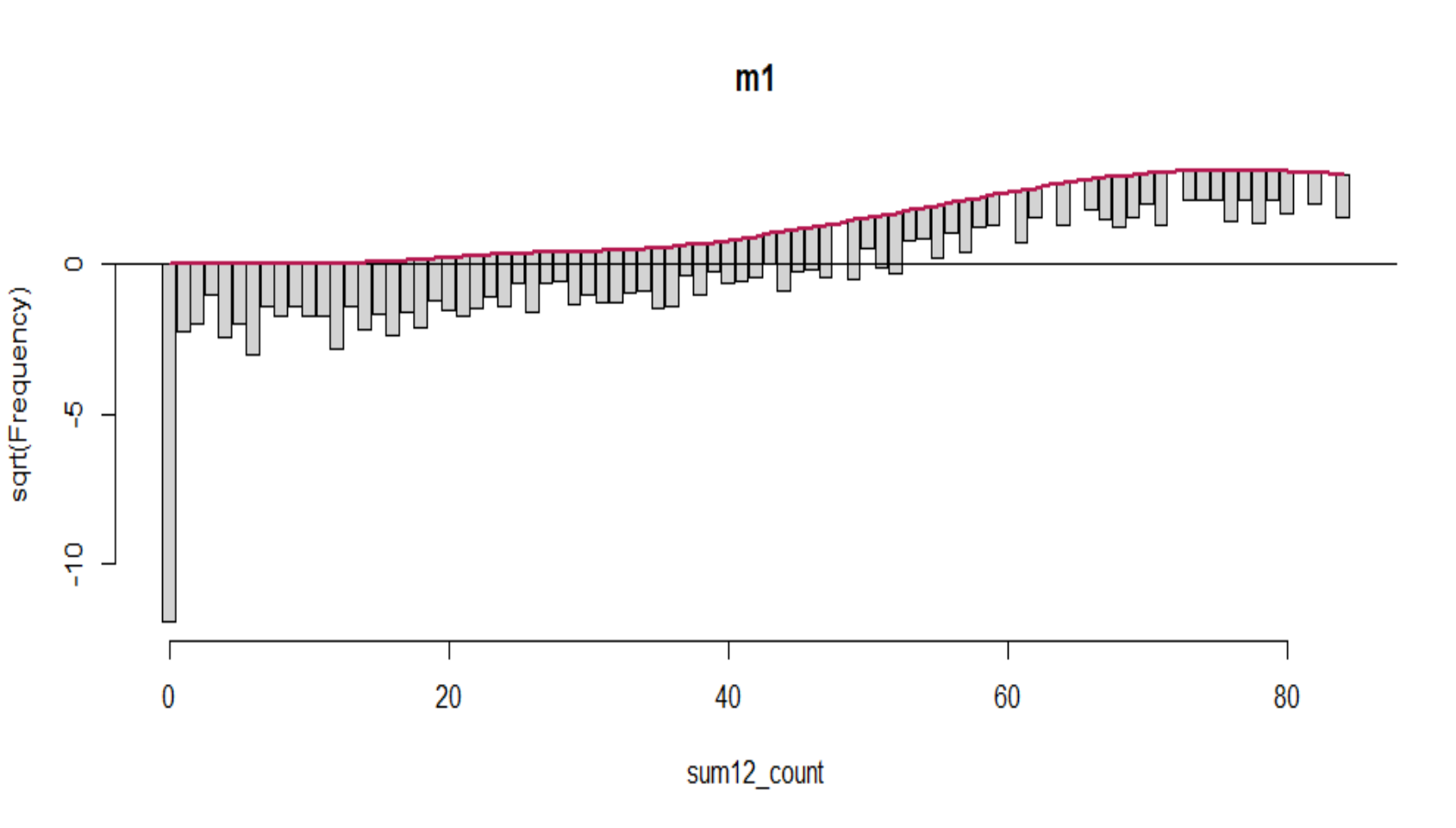
$$V = \frac{\sqrt{n}(n^{-1} \sum mi)}{\sqrt{n^{-1} \sum (mi - \bar{m})^2}}$$

What model to use?

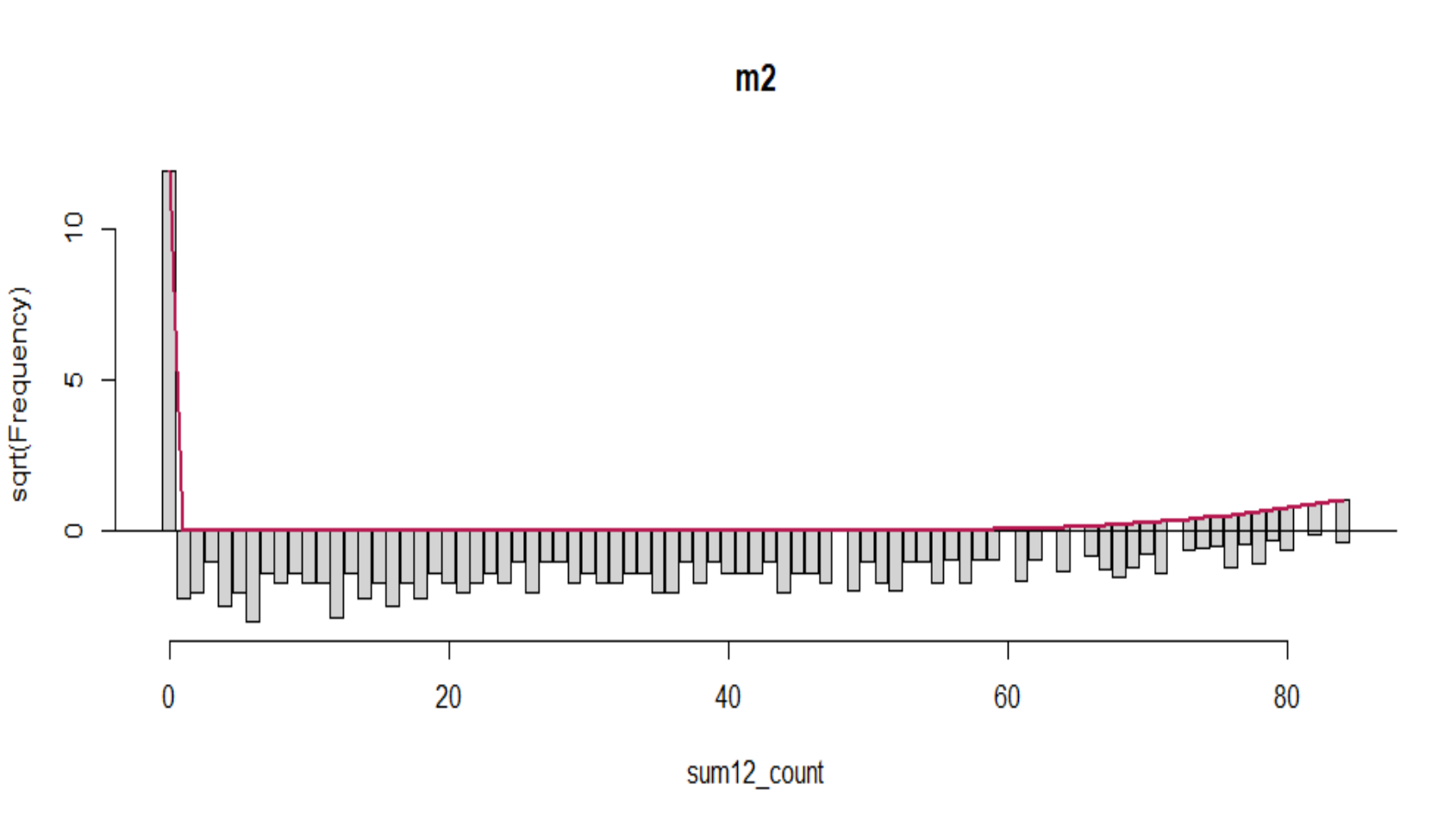
```
> vuong(m1,m2)
NA or numerical zeros or ones encountered in fitted
    probabilities
dropping these 166 cases, but proceed with caution
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                  -4.090300 model2 > model1 2.1541e-05
AIC-corrected        -4.089634 model2 > model1 2.1603e-05
BIC-corrected        -4.088394 model2 > model1 2.1719e-05
```

Note that the model output above does not indicate in any way if our zero-inflated model is an improvement over a standard Poisson regression. In this case, the Vuong test suggests that the zero-inflated negative binomial model is a significant improvement over a standard negative binomial model.
(Recall m1 is ZIP and m2 is Poisson regression)

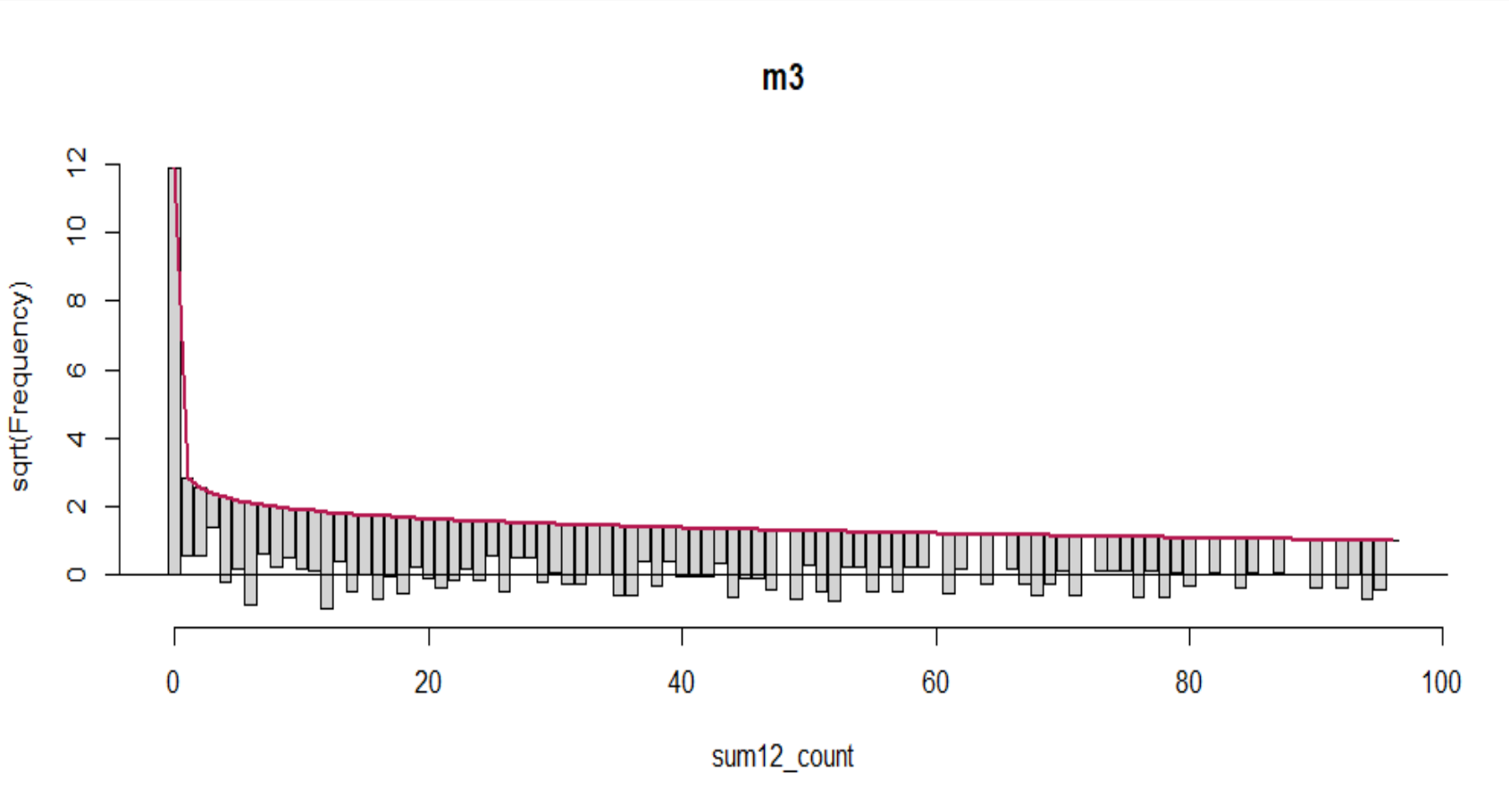
Poisson: rootgram



ZIP: ROOTGRAM



ZINB: ROOTGRAM



Expected counts are shown by the thick red line and observed counts are shown as bars. In conclusion, It looks like ZINB has the best fit. (note: in here I used $5000 bin to make it clear). However,....

Last remark: my thoughts on selecting a model

However, when we think of my data, it is better to use Negative binomial regression or Zero inflated negative binomial regression because the data is over-dispersed and have excessive zeros. Of course variance is much larger than the mean.

Other methods to check fit of model are using rootgram() function in R and checking Chi-squared distribution with deviances from models. However, none of method tells us which one is a true model, so I think the best way is just see the data with researcher's intuition and use the model.