hw6

Ben Tankus

5/6/2021

R Question

1. (5 points) Consider data on 4406 individuals, aged 66 and over, who are covered by Medicare, a public insurance program, in the file DT.rda. The objective is to model the number of physician/nonphysician office and hospital outpatient visits using available covariate information for the patients.

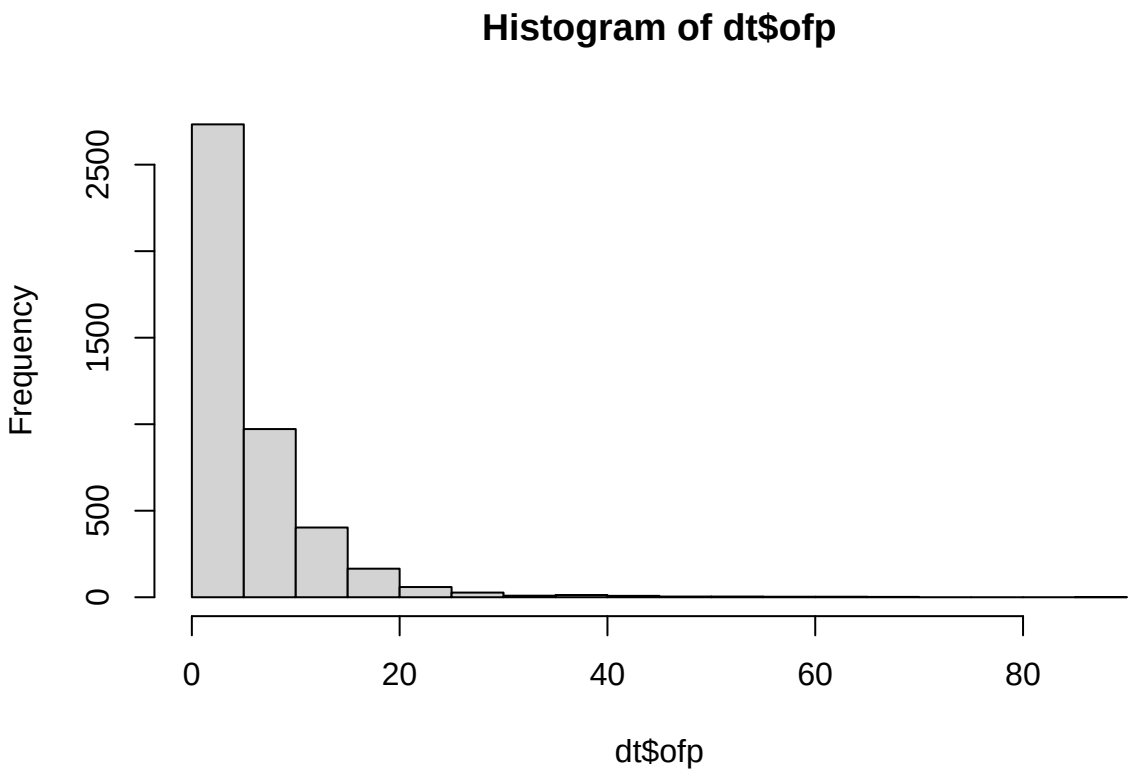
```
load(file = 'DT.rda')
head(dt)
```

```
##   ofp hosp  health numchron gender school privins
## 1   5    1 average        2   male      6     yes
## 2   1    0 average        2 female     10     yes
## 3  13    3    poor        4 female     10      no
## 4  16    1    poor        2   male      3     yes
## 5   3    0 average        2 female      6     yes
## 6  17    0    poor        5 female      7      no
```

The covariates include health status variables hosp (number of hospital stays), health (self-perceived health status), numchron (number of chronic conditions), as well as socioeconomic variables gender, school (number of years of education), and privins (private insurance indicator). Once you download the DR.rda file, you can load it into R using load("DT.rda").

(a) Produce a histogram of the dependent variable. What types of statistical models (.e.g, Poisson regression, hurdle Poisson model, etc.) should you consider for these data? Give a brief justification for your answer.

```
hist(dt$ofp)
```



This very much looks like a zero inflated negative binomial distribution I would *consider* using both a hurdle model and a zero-inflated negative binomial model, but suspect the latter will provide a more accurate fit.

(b) Fit the models you indicated in part (a), using the identical set of explanatory variables in each, and report the AIC for each.

```
hurdle.mod <- hurdle(ofp ~ ., data = dt, dist = 'negbin')
zip.mod <- zeroinfl(ofp ~ ., data = dt, dist = 'negbin')
bin.nb <- glm.nb(ofp ~ ., data = dt)


LRstats(hurdle.mod, zip.mod, bin.nb)
```

```
## Likelihood summary table:
##              AIC   BIC LR Chisq   Df Pr(>Chisq)
## hurdle.mod 24210 24319    24176 4389  < 2.2e-16 ***
## zip.mod    24215 24324    24181 4389  < 2.2e-16 ***
## bin.nb     24359 24417    24341 4398  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(zip.mod)
```

```
##
## Call:
## zeroinfl(formula = ofp ~ ., data = dt, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.1966 -0.7097 -0.2784  0.3256 17.7661
##
## Count model coefficients (negbin with log link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.193466   0.056737  21.035  < 2e-16 ***
## hosp             0.201214   0.020392   9.867  < 2e-16 ***
## healthpoor       0.287190   0.045940   6.251 4.07e-10 ***
## healthexcellent -0.313540   0.062977  -4.979 6.40e-07 ***
## numchron         0.128955   0.011938  10.802  < 2e-16 ***
## gendermale      -0.080093   0.031035  -2.581  0.00986 **
## school           0.021338   0.004368   4.886 1.03e-06 ***
## privinsyes       0.126815   0.041687   3.042  0.00235 **
## Log(theta)       0.394731   0.035145  11.231  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.06353    0.27668  -0.230  0.81838
## hosp            -0.81761    0.43876  -1.863  0.06240 .
## healthpoor       0.10173    0.44072   0.231  0.81745
## healthexcellent  0.10488    0.30964   0.339  0.73482
## numchron        -1.24629    0.17918  -6.956 3.51e-12 ***
## gendermale       0.64936    0.20046   3.239  0.00120 **
## school          -0.08481    0.02676  -3.169  0.00153 **
## privinsyes      -1.15808    0.22436  -5.162 2.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 1.484
## Number of iterations in BFGS optimization: 61
## Log-likelihood: -1.209e+04 on 17 Df
```

The AIC for hurdle is 24210, and for zip is 24215. I would consider this difference insignificant. Just to compare to a non- ZIP model I also ran a negative binomial fit, which profided an AIC of 24359, some 140 units higher (worse).

(c) Based on parts (a) and (b) and considering interpretation of the models in the context of the data, write a sentence summarizing your findings in terms of which model seems most appropriate for these data.

I think the zero-inflation model is most appropreate to fit these data, considering the study. The ZIP accounts for zeros in a more robust fashion when compared to the hurdle model, and this distribution has many zeros.

Conceptual Question

2. (5 points) For each of the following response variables, indicate whether a zero-inflated model or a hurdle model would be more appropriate. Justify your answer by providing possible processes responsible for generating the excess zeros and the counts that correspond to the model. There may not be a clear answer for you, so simply pick one and try to justify it.

(a) School administrators study the attendance behavior of high school juniors and take the number of days absent as the response variable.

ZERO INFLATED: Similar to the example in the lab, this is likely to have zero-inflated data from when someone forces the juniors to go to school. Therefore a zero inflated model should be used.

(b) Wildlife biologists want to model how many fish are being caught by fishermen at a state park. All park visitors are asked how many fish they caught.

ZERO INFLATED: Because not all visitors fish, there will be many true zeros in the model, therefore we should use a hurdle analysis method.

(c) Researchers are interested in creating a stock trading model for investors. The response is trades per week made by each investor.

ZERO INFLATED: It is unlikely that an investor would go an entire week without investing, so I recommend a zero inflated model for this one.

(d) Researchers want to create a model for loan defaults. They take the number of outstanding payments that exist for each of a random sample of loans

ZERO INFLATED: The researchers are taking a random sample of loans. ASSUMING the loans are *not* paid off in full, I recommend using a zero inflated model because most loans will have at least one payment remaining.