Modelling Count Variables with R

Zero-inflated Regression models - Summary

- Zero-inflated models attempt to account for excess zeros.
- In other words, two kinds of zeros are thought to exist in the data, "**true zeros**" and "**excess zeros**".

Zero-inflated Regression models

Two Distinct Processes

- The two parts of the a zero-inflated model are a binary model, usually a logit model to model which of the two processes the zero outcome is associated with and a count model, in this case, a negative binomial model, to model the count process.
- In other words, the excess zeros are generated by a separate process from the count values and that the excess zeros can be modelled independently.
- Zero-inflated models estimate two equations simultaneously, one for the count model and one for the excess zeros.
- The expected count is expressed as a combination of the two processes.

Zero-inflated Regression models

Fishing Data Set

- We have data on 250 groups that went to a park.
- Each group was questioned about how many fish they caught (**count**), how many children were in the group (**child**), how many people were in the group (**persons**), and whether or not they brought a camper to the park (**camper**).
- In addition to predicting the number of fish caught, there is interest in predicting the existence of excess zeros, i.e., the probability that a group caught zero fish.
- We will use the variables child, persons, and camper in our model.

Fishing Data Set

- In addition to predicting the number of fish caught, there is interest in predicting the existence of excess zeros, i.e., the probability that a group caught zero fish.
- We will use the variables child, persons, and camper in our model.

```
> head(fish)
nofish livebait camper persons child         xb
1       1        0       0       1       0 -0.8963146
2       0        1       1       1       0 -0.5583450
3       0        1       0       1       0 -0.4017310
4       0        1       1       2       1 -0.9562981
5       0        1       0       1       0  0.4368910
6       0        1       1       4       2  1.3944855
zg count
1  3.0504048     0
2  1.7461489     0
3  0.2799389     0
4 -0.6015257     0
5  0.5277091     1
6 -0.7075348     0
```

What is a Zero-Inflated Model?

The Fishing Example

- A zero-inflated model assumes that zero outcome is due to two different processes.
- For instance, in the example of fishing presented here, the two processes are that a subject has *gone fishing* vs. *not gone fishing*.
- If not gone fishing, the only outcome possible is zero.
- If gone fishing, it is then a count process.

$$E(nfishcaught = k) = P(notgonefishing) \times 0 + P(gonefishing) \times E(y = k|g$$

Zero-inflated Poisson regression

Though we can run a Poisson regression in `R` using the `glm` function in one of the core packages, we need another package to run the zero-inflated poisson model. We use the **pscl** package.

```
summary(m1 <- zeroinfl(count ~ child + camper |
     persons, data = zinb))
```

Zero-inflated Poisson regression

```
##
## Call:
## zeroinfl(formula = count ~ child + camper | persons, dat
##
## Pearson residuals:
##     Min      1Q Median      3Q    Max
## -1.237 -0.754 -0.608 -0.192 24.085
```

Zero-inflated Poisson regression

```
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.5979     0.0855   18.68   <2e-16 ***
## child        -1.0428     0.1000  -10.43   <2e-16 ***
## camper1       0.8340     0.0936    8.91   <2e-16 ***
```

Zero-inflated Poisson regression

```
## Zero-inflation model coefficients (binomial with logit l
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.297      0.374    3.47  0.00052 ***
## persons        -0.564      0.163   -3.46  0.00053 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Number of iterations in BFGS optimization: 12
## Log-likelihood: -1.03e+03 on 5 Df
```

Zero-inflated Poisson regression

- Below the model call, you will find a block of output containing Poisson regression coefficients for each of the variables along with standard errors, z-scores, and p-values for the coefficients.
- A second block follows that corresponds to the inflation model.
- This includes logit coefficients for predicting excess zeros along with their standard errors, z-scores, and p-values.

- All of the predictors in both the count and inflation portions of the model are statistically significant.

Vuong Testing

- 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.
- We can determine this by running the corresponding standard Poisson model and then performing a Vuong test of the two models.

```
summary(p1 <- glm(count ~ child + camper,
family = poisson, data = fishing))
```

- The Vuong test compares the zero-inflated model with an ordinary Poisson regression model.
- In this example, we can see that our test statistic is significant, indicating that the zero-inflated model is superior to the standard Poisson model.

```
vuong(p1, m1)
## Vuong Non-Nested Hypothesis Test-Statistic: -3.574
## (test-statistic is asymptotically distributed N(0,1)
##  null that the models are indistinguishible)
## in this case:
## model2 > model1, with p-value 0.0001756
```

Zero-Inflated Negative Binomial regression

- We are going to use the variables: **child** and **camper** to model the count in the part of negative binomial model and the variable **persons** in the logit part of the model.
- We use the **pscl** to run a zero-inflated negative binomial regression.
- We begin by estimating the model (called `m1`) with the variables of interest.

```
m1 <- zeroinfl(count ~ child + camper | persons,
  data = fishing, dist = "negbin",
  EM = TRUE)

summary(m1)
```

```
## Call:
## zeroinfl(formula = count ~ child + camper | persons,
##     data = fishing,
##     dist = "negbin", EM = TRUE)
##
## Pearson residuals:
##    Min     1Q Median     3Q    Max
## -0.586 -0.462 -0.389 -0.197 18.013
```

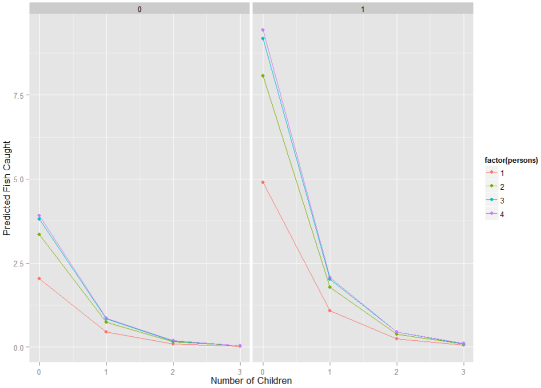
- Below the model call, you will find a block of output containing negative binomial regression coefficients for each of the variables along with standard errors, z-scores, and p-values for the coefficients.
- A second block follows that corresponds to the inflation model. This includes logit coefficients for predicting excess zeros along with their standard errors, z-scores, and p-values.

```
## Count model coefficients (negbin with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.371      0.256    5.35  8.6e-08 ***
## child         -1.515      0.196   -7.75  9.4e-15 ***
## camper1        0.879      0.269    3.26   0.0011 **
## Log(theta)    -0.985      0.176   -5.60  2.1e-08 ***
```

```
## Zero-inflation model coefficients (binomial with logit l
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.603      0.836    1.92    0.055 .
## persons        -1.666      0.679   -2.45    0.014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Theta = 0.373
## Number of iterations in BFGS optimization: 2
## Log-likelihood: -433 on 6 Df
```

Tests of Significance

- All of the predictors in both the count and inflation portions of the model are statistically significant.
- This model will fit the data significantly better than the null model, i.e., the intercept-only model.
- To show that this is the case, we could compare with the current model to a null model without predictors using chi-squared test on the difference of log likelihoods.


0
1
Predicted Fish Caught
7.5
5.0
2.5
0.0
0
1
2
3
Number of Children
factor(persons)
1
2
3
4

- Note that the model output above does not indicate in any way if our zero-inflated model is an improvement over a standard negative binomial regression.
- We can determine this by running the corresponding standard negative binomial model and then performing a Vuong test of the two models.
- We use the MASS package to run the standard negative binomial regression.

```
library(MASS)
summary(m2 <- glm.nb(count ~ child + camper, data = zinb))
.....
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.073      0.242    4.42  9.7e-06 ***
## child         -1.375      0.196   -7.03  2.1e-12 ***
## camper1        0.909      0.284    3.21   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
```

```
vuong(m1, m2)

## Vuong Non-Nested Hypothesis Test-Statistic: 1.702
## (test-statistic is asymptotically distributed N(0,1) unc
##  null that the models are indistinguishible)
## in this case:
## model1 > model2, with p-value 0.0444
```

- The predictors child and camper in the part of the negative binomial regression model predicting number of fish caught (count) are both significant predictors.
- The predictor person in the part of the logit model predicting excessive zeros is statistically significant.
- For these data, the expected change in log(count) for a one-unit increase in child is -1.515255 holding other variables constant.
- A camper (camper $= 1$) has an expected log(count) of 0.879051 higher than that of a non-camper (camper $= 0$) holding other variables constant.

- The log odds of being an excessive zero would decrease by 1.67 for every additional person in the group.
- In other words, the more people in the group the less likely that the zero would be due to not gone fishing.
- Put plainly, the larger the group the person was in, the more likely that the person went fishing.
- The Vuong test suggests that the zero-inflated negative binomial model is a significant improvement over a standard negative binomial model.