UNIVERSITY *of* WEST FLORIDA

Zero-Inflated Models

Statistics for Data Science II

Introduction

Now that we know how to handle count data, we need to discuss what to do if there are "too many" zeros in the response.

> If we look at a histogram or bar chart of the responses, we may see a "spike" at zero – this spike indicates zero inflation.

The zero-inflated Poisson model is as follows:

$$y_i \sim \begin{cases} 0, & \text{with probability } 1 - \phi_i \\ \text{Poisson}(\lambda_i), & \text{with probability } \phi_i \end{cases}$$

Thus, we are jointly modeling (1) a logit model for predicting excess zeros, (2) the Poisson count model.

Modeling

We will specify this in R using the `zeroinfl()` function.

e.g., `zeroinfl(outcome ~ predCount1 + predCount2 + ...  | predZero1 + predZero2 + ..., data = dataset)`

Note that we do not need the same predictors in each model.

Example:

```
m1 <- zeroinfl(satellites_num ~ width_cm + spine_cond +
                width_cm:spine_cond | width_cm + spine_cond,
              data=data)
```

Modeling

Example:

```
summary(m1)[1]
```

```
## $coefficients
## $coefficients$count
##                        Estimate Std. Error    z value   Pr(>|z|)
## (Intercept)          2.94347218 1.57010559  1.8746970 0.06083443
## width_cm            -0.04492752 0.05687372 -0.7899522 0.42955566
## spine_cond          -1.00620790 0.65491872 -1.5363859 0.12444376
## width_cm:spine_cond  0.03392428 0.02382947  1.4236274 0.15455437
##
## $coefficients$zero
##                Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) 12.50398066  2.7856534  4.4887065 7.165694e-06
## width_cm    -0.50063567  0.1044376 -4.7936333 1.637875e-06
## spine_cond  -0.04284693  0.2240521 -0.1912364 8.483404e-01
```

Modeling

Example:

The resulting models are

$$\ln\left(\hat{Y}_i\right) = 2.94 - 0.04\text{width} - 1.01\text{spine} + 0.03(\text{width} \times \text{spine})$$

$$\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 12.50 - 0.50\text{width} - 0.04\text{spine}$$

Y_i is the number of satellites,

$\pi_i = P[Z = 1]$, where

$$Z_i = \begin{cases} 0 & \text{if } Y_i = 0 \\ 1 & \text{if } Y_i > 0 \end{cases}$$

Interpretations

Because we are really interested in modeling the count data (Y), we will focus on interpreting the Poisson regression.

Like before, we will convert the $\hat{\beta}_i$ to IRR_i and interpret in terms of the multiplicative effect.

See previous lectures for interpretations in models for count data.

Inference

Testing for significant predictors is the same as previously discussed.

Example:

```
summary(m1)[1]
```

```
## $coefficients
## $coefficients$count
##                        Estimate Std. Error    z value   Pr(>|z|)
## (Intercept)          2.94347218 1.57010559  1.8746970 0.06083443
## width_cm            -0.04492752 0.05687372 -0.7899522 0.42955566
## spine_cond          -1.00620790 0.65491872 -1.5363859 0.12444376
## width_cm:spine_cond  0.03392428 0.02382947  1.4236274 0.15455437
##
## $coefficients$zero
##                Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) 12.50398066  2.7856534  4.4887065 7.165694e-06
## width_cm    -0.50063567  0.1044376 -4.7936333 1.637875e-06
## spine_cond  -0.04284693  0.2240521 -0.1912364 8.483404e-01
```

Inference

Constructing confidence intervals is the same as previously discussed.

Example:

```
confint(m1)
```

```
##                                2.5 %      97.5 %
## count_(Intercept)          -0.13387824  6.02082260
## count_width_cm             -0.15639797  0.06654292
## count_spine_cond           -2.28982500  0.27740921
## count_width_cm:spine_cond -0.01278062  0.08062918
## zero_(Intercept)            7.04420026 17.96376106
## zero_width_cm              -0.70532965 -0.29594169
## zero_spine_cond            -0.48198102  0.39628717
```

Modeling - Negative Binomial

Recall that Poisson regression is not always appropriate. When this is the case, we will use the zero-inflated negative binomial.

We are still using the `zeroinfl()` function, but now we specify `dist = "negbin"`.

Example:

```
m2 <- zeroinfl(satellites_num ~ width_cm + spine_cond +
            width_cm:spine_cond | width_cm + spine_cond,
          dist = "negbin", data=data)
```

Modeling - Negative Binomial

Example:

```
summary(m2)[1]
```

```
## $coefficients
## $coefficients$count
##                        Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)          2.98274988 2.28815419  1.303562 1.923831e-01
## width_cm            -0.04717335 0.08283889 -0.569459 5.690447e-01
## spine_cond          -1.10200683 0.94150231 -1.170477 2.418091e-01
## width_cm:spine_cond  0.03734152 0.03427481  1.089474 2.759448e-01
## Log(theta)           1.62758230 0.35623466  4.568849 4.904105e-06
##
## $coefficients$zero
##                Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) 12.89524062  2.9956364  4.3046749 1.672311e-05
## width_cm    -0.51890153  0.1133999 -4.5758536 4.742820e-06
## spine_cond  -0.05614321  0.2419333 -0.2320606 8.164909e-01
```