Lecture 22

Last time

Survey data from 77 college students on a dry campus (i.e., alcohol is prohibited) in the US. Survey asks students "How many alcoholic drinks did you consume last weekend?"

- `drinks`: number of drinks the student reports consuming
- `sex`: whether the student identifies as male
- `OffCampus`: whether the student lives off campus
- `FirstYear`: whether the student is a first-year student

Question: Why might students report 0 drinks?

- they might be lying
- Two groups: students who never drink
  students who didn't drink last weekend

Paper de-brief

How did Lambert (1992) address the problem of excess 0s?

- Two sources of zeros:
  - Zeros from a Poisson
  - Zeros (that are always zeros)
    ↑
    capture w/ logistic regression model

Zero-inflated Poisson (ZIP) model

Y_i = # drinks consumed by student i

Two possibilities:

Student i never drinks $Z_i = 1$

Student i sometimes drinks $Z_i = 0$

} Z_i is unobserved (latent variable)

$Y_i \mid (Z_i = 1) = 0$ (point mass at 0)

$Y_i \mid (Z_i = 0) \sim \text{Poisson}(\lambda_i)$

$\log(\lambda_i) = \beta^T X_i$ (λ_i depends on explanatory variables)

$P(Z_i = 1) = p_i$ $\log\left(\frac{p_i}{1-p_i}\right) = \gamma^T X_i$

Problem: Don't get to see Z_i!

Zero-inflated Poisson (ZIP) model

$$P(Y_i = y \mid Z_i = 1) = \begin{cases} 1 & y = 0 \\ 0 & y > 0 \end{cases}$$

$$P(Y_i = y \mid Z_i = 0) = \frac{e^{-\lambda_i} \lambda_i^y}{y!}$$

$$P(Y_i = y) = P(Y_i = y \mid Z_i = 0) P(Z_i = 0) + P(Y_i = y \mid Z_i = 1) P(Z_i = 1)$$

$$P(Y_i = y) = \begin{cases} p_i + (1 - p_i) e^{-\lambda_i} & y = 0 \\ (1 - p_i) \dfrac{e^{-\lambda_i} \lambda_i^y}{y!} & y > 0 \end{cases}$$

=> we can still calculate probabilities & expectations for Y, even if we don't see the latent variable Z

In R

```r
library(pscl)

m1 <- zeroinfl(drinks ~ sex + FirstYear + OffCampus,
               dist = "poisson",
               data = wdrinks)

m1$coefficients
```

```
$count
  (Intercept)          sexm FirstYearTRUE OffCampusTRUE
    0.8010438     0.9834689    -0.1619318     0.3723519

$zero
  (Intercept)          sexm FirstYearTRUE OffCampusTRUE
   -0.3961839    -0.0707907     0.8919687    -1.6913744
```

Paper de-brief

How did Lambert (1992) propose fitting the ZIP model?

Expectation-Maximization (EM) algorithm

Fitting ZIP models

Suppose we can actually observe latent variable Z_i

$Z_i \sim \text{Bernalli}(p_i)$ $\quad \log\left(\frac{p_i}{1-p_i}\right) = \gamma^T X_i$

$Y_i \mid (Z_i = 0) \sim \text{Poisson}(\lambda_i)$ $\quad \log(\lambda_i) = \beta^T X_i$

$$L(\gamma, \beta) = \prod_{i=1}^{n} P(Y_i, Z_i \mid \gamma, \beta) = \prod_{i=1}^{n} P(Z_i \mid \gamma) P(Y_i \mid Z_i, \beta)$$

$$= \prod_{i=1} p_i^{Z_i} (1-p_i)^{1-Z_i} \left(\frac{e^{-\lambda_i} \lambda_i^{Y_i}}{Y_i!} \right)^{1-Z_i}$$

{ if $Z_i = 0 \Rightarrow$ Poisson
if $Z_i = 1 \Rightarrow P(Y_i = 0) = 1$

$$\Rightarrow \ell(\gamma, \beta) = \sum_i (Z_i \log(p_i) + (1-Z_i) \log(1-p_i))$$

$$+ \sum_i \{(1-Z_i)(-\lambda_i + Y_i \log \lambda_i) - (1-Z_i) \log(Y_i!)\}$$

If we see Z_i we can separately maximize

$$\ell(\beta) = \sum_i (1-Z_i)(-\lambda_i + Y_i \log \lambda_i)$$

$$\ell(\gamma) = \sum_i (Z_i \log(p_i) + (1-Z_i) \log(1-p_i))$$

Problem: don't get to see Z_i

Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_22.html