Zero inflated models

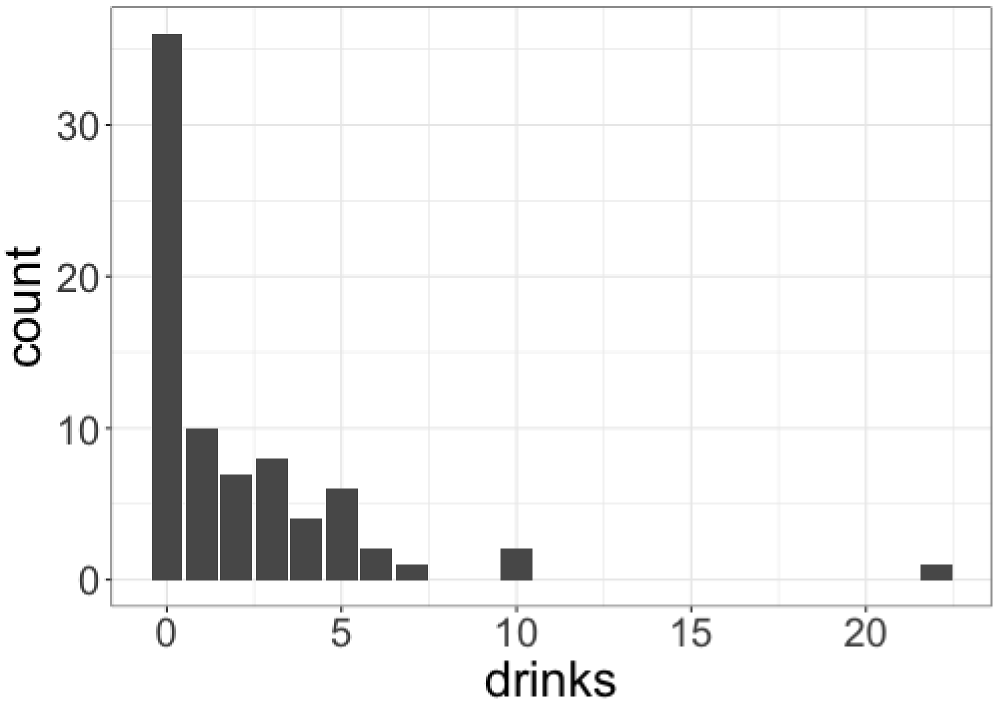
Recap: College drinking

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

> Our goal: model the number of drinks students report consuming.

Recap: EDA



What do you notice about this distribution?

Excess zeros

Why might a student report consuming 0 drinks?

Zero-inflated Poisson (ZIP) model

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

where

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma_0 + \gamma_1 FirstYear_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 OffCampus_i + \beta_2 Male_i$$

> What do α_i and λ_i represent in this model?

Fitting the model in R

```r
library(pscl)
m1 <- zeroinfl(drinks ~ OffCampus + sex | FirstYear,
               data = wdrinks)
summary(m1)
```

```
...
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.7543     0.1440   5.238 1.62e-07 ***
## OffCampusTRUE    0.4159     0.2059   2.020   0.0433 *
## sexm             1.0209     0.1752   5.827 5.63e-09 ***
##
## Zero-inflation model coefficients (binomial with logit l
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.6036     0.3114  -1.938   0.0526 .
## FirstYearTRUE    1.1364     0.6095   1.864   0.0623 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
...
```

Fitted ZIP model

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

$$\log\left(\frac{\widehat{\alpha}_i}{1-\widehat{\alpha}_i}\right) = -0.60 + 1.14 FirstYear_i$$

$$\log(\widehat{\lambda}_i) = 0.75 + 0.42\ OffCampus_i + 1.02\ Male_i$$

> How would I interpret the coefficient 1.14 in the fitted model?

Fitted ZIP model

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

$$\log\left(\frac{\widehat{\alpha}_i}{1-\widehat{\alpha}_i}\right) = -0.60 + 1.14 FirstYear_i$$

$$\log(\widehat{\lambda}_i) = 0.75 + 0.42\, OffCampus_i + 1.02\, Male_i$$

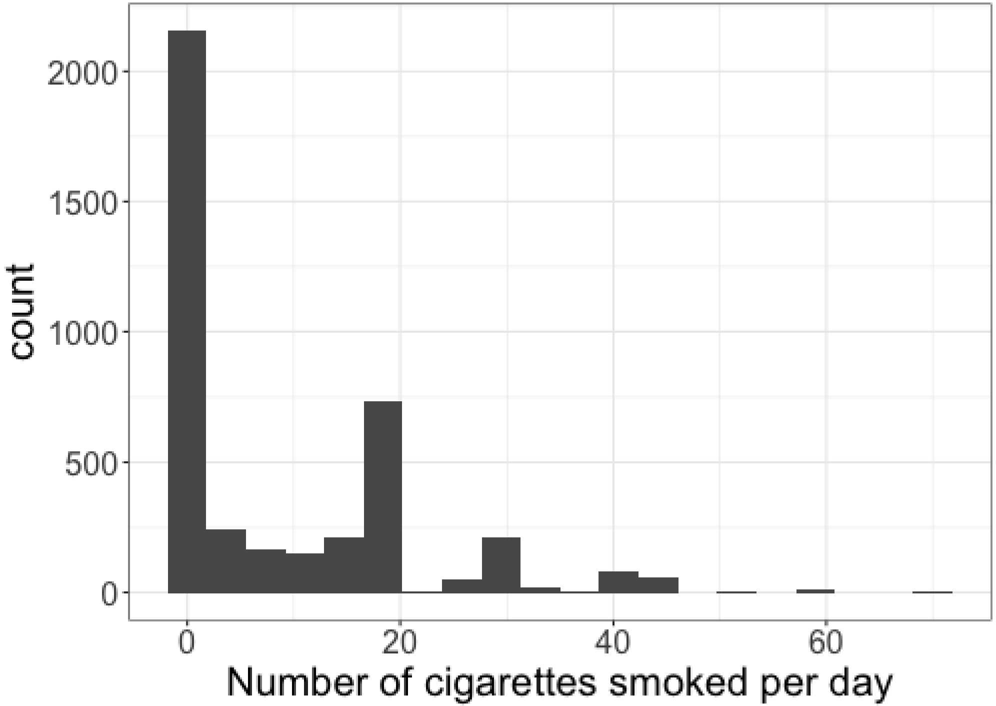
> How would I interpret the coefficient 0.42 in the fitted model?

Data: Framingham heart study

Data collected on residents of Framingham, MA over a long period of time, to study variables related to heart health. We will work with a subset of the data, containing

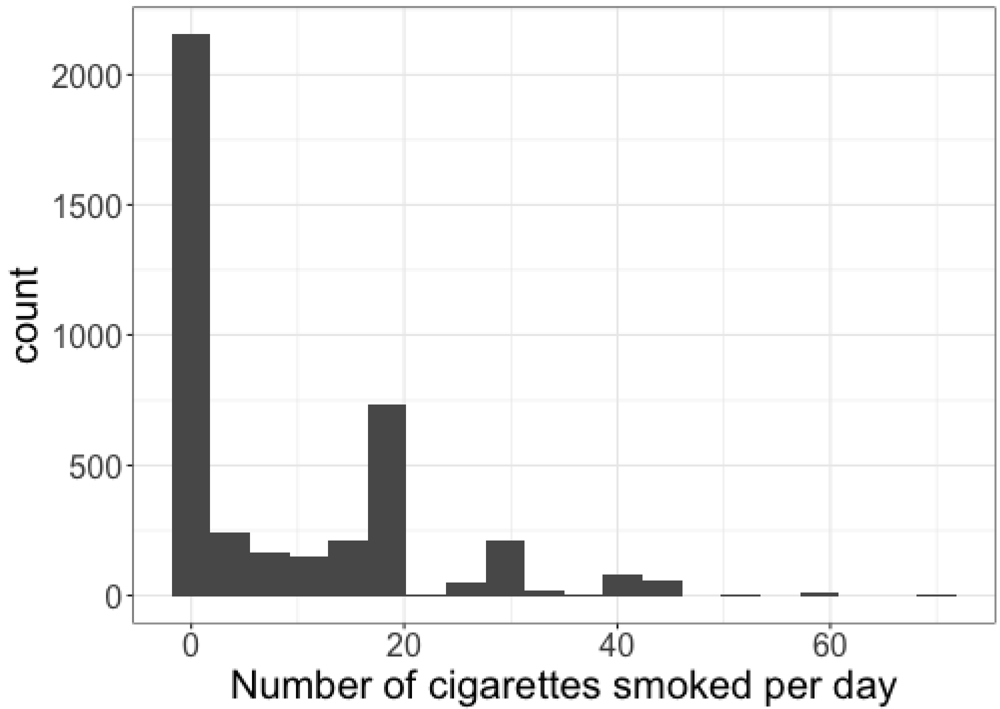
- `cigsPerDay`: The number of cigarettes smoked per day during the study period.
- `education`: 1 = High School, 2 = Some College, 3 = College Degree, 4 = Advanced Degree.
- `male`: 1 = Male, 0 = Female.
- `age`: The age of the individual in years.
- `diabetes`: 1 if the individual has diabetes, 0 otherwise.

EDA: number of cigarettes smoked



What do you notice about this distribution?

EDA: number of cigarettes smoked



> What latent (unobserved) binary variable would impact the number of cigarettes smoked?

Class activity

https://sta214-f22.github.io/class_activities/ca_lecture_23.html

Class activity

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

$$\log\left(\frac{\widehat{\alpha}_i}{1 - \widehat{\alpha}_i}\right) = -2.51 + 0.051 Age_i$$

$$\begin{aligned} \log(\widehat{\lambda}_i) = 2.93 - 0.022 EducationSome_i - 0.067 EducationColleg \\ 0.009 EducationAdv_i - 0.046 Diabetes_i \end{aligned}$$

> How do we interpret the coefficient -0.046 in the fitted model?

Class activity

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

$$\log\left(\frac{\widehat{\alpha}_i}{1 - \widehat{\alpha}_i}\right) = -2.51 + 0.051 Age_i$$

$$\log(\widehat{\lambda}_i) = 2.93 - 0.022 EducationSome_i - 0.067 EducationColleg \\ 0.009 EducationAdv_i - 0.046 Diabetes_i$$

> What is the estimated probability that a 50 year old does not smoke?

Class activity

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

$$\log\left(\frac{\widehat{\alpha}_i}{1 - \widehat{\alpha}_i}\right) = -2.51 + 0.051 Age_i$$

$$\log(\widehat{\lambda}_i) = 2.93 - 0.022 EducationSome_i - 0.067 EducationColleg$$
$$0.009 EducationAdv_i - 0.046 Diabetes_i$$

> What is the expected number of cigarettes smoked per day, for a smoker with diabetes and some college education?

Class activity

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

$$\log\left(\frac{\widehat{\alpha}_i}{1-\widehat{\alpha}_i}\right) = -2.51 + 0.051 Age_i$$

$$\log(\widehat{\lambda}_i) = 2.93 - 0.022 EducationSome_i - 0.067 EducationColleg$$
$$0.009 EducationAdv_i - 0.046 Diabetes_i$$

> What is the probability that a 45 year old college graduate without diabetes smokes one cigarette per day?

Making predictions

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

$$\log\left(\frac{\widehat{\alpha}_i}{1-\widehat{\alpha}_i}\right) = -2.51 + 0.051 Age_i$$

$$\log(\widehat{\lambda}_i) = 2.93 - 0.022 EducationSome_i - 0.067 EducationColleg$$
$$0.009 EducationAdv_i - 0.046 Diabetes_i$$

> How would I estimate the expected number of cigarettes smoked per day, by a college graduate without diabetes?

A new question

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

$$\log\left(\frac{\alpha_i}{1 - \alpha_i}\right) = \gamma_0 + \gamma_1 Age_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 EducationSome_i + \beta_2 EducationCollege_i + \beta_3 EducationAdv_i + \beta_4 Diabetes_i$$

> New research question: for smokers, does the number of cigarettes smoked per day depend on age?
>
> How would we answer this research question?

Inference

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma_0 + \gamma_1 Age_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 EducationSome_i + \beta_2 EducationCollege_i + \beta_3 EducationAdv_i + \beta_4 Diabetes_i + \beta_5 Age_i$$

> Research question: for smokers, does the number of cigarettes smoked per day depend on age?
>
> What are the null and alternative hypotheses?

Wald test

```r
m2 <- zeroinfl(cigsPerDay ~ education +
                 diabetes + age | age,
               data = heart_data)
summary(m2)
```

```
...
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.2063437  0.0342290  93.673  < 2e-16 ***
## education2  -0.0441195  0.0124809  -3.535 0.000408 ***
## education3  -0.0820388  0.0158604  -5.173 2.31e-07 ***
## education4  -0.0062453  0.0171640  -0.364 0.715965
## diabetes    -0.0241419  0.0386336  -0.625 0.532042
## age         -0.0056183  0.0006738  -8.338  < 2e-16 ***
...
```

Likelihood ratio test

```
m2 <- zeroinfl(cigsPerDay ~ education +
                 diabetes + age | age,
               data = heart_data)
m2$loglik
```

```
## [1] -14023.42
```

```
m1 <- zeroinfl(cigsPerDay ~ education +
                 diabetes | age,
               data = heart_data)
m1$loglik
```

```
## [1] -14058.41
```