Zero inflated models

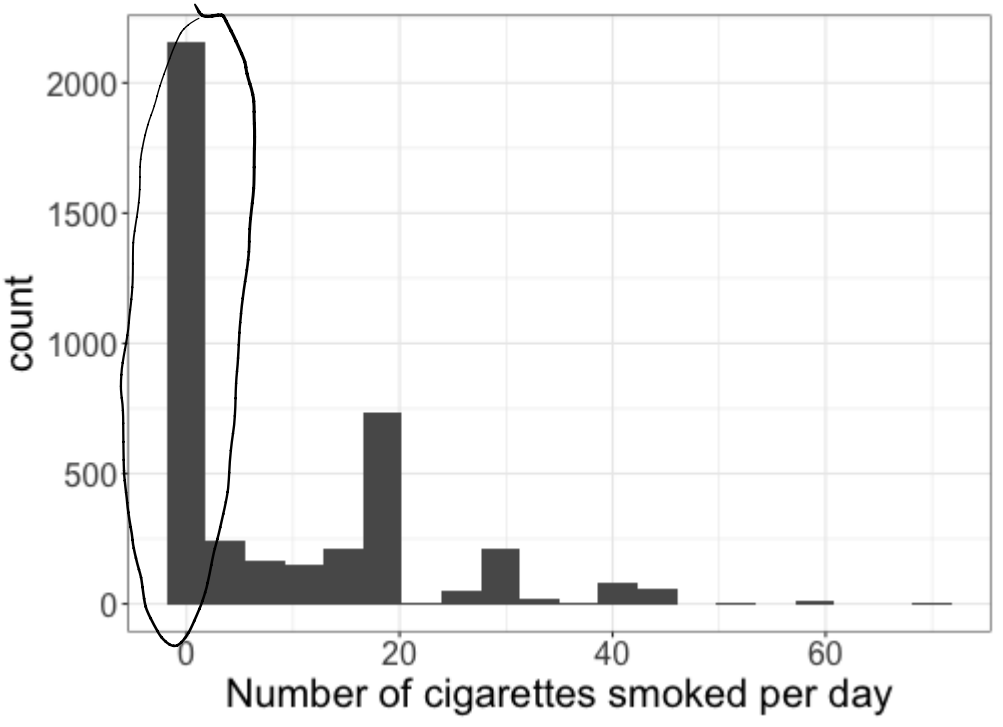
Data: Framingham heart study

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.

> Why might we see zero inflation in the number of cigarettes smoked per day?

EDA: number of cigarettes smoked



Class activity, Part I

https://sta214-s23.github.io/class_activities/ca_lecture_26.html

Class activity

$\alpha_i = P(Z_i = 1)$ = P(person i is a nonsmoker)
↑ latent variable
λ_i = average # cigarettes smoked by smokers

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

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

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

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

(Among smokers)
Having diabetes is associated w/ a decrease in the log-mean # of cigarettes per day by 0.046 (holding education fixed)

Class activity

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

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

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

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

$\hat{\alpha}_i = \dfrac{e^{-2.51 + 0.051(50)}}{1 + e^{-2.51 + 0.051(50)}} \approx 0.51$

Class activity

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

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

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

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

$\hat{\lambda}_i = \exp\{2.93 - 0.022 - 0.046\} \approx 17.5$

Class activity

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

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

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

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

Making predictions

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

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

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

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

(45 years old)

If $Z_i = 1$ (nonsmoker) expected # cigarettes $= 0$

If $Z_i = 0$ (smoker) expected # cigarettes $= \lambda_i$

Overall: $0(\alpha_i) + \lambda_i(1-\alpha_i) = \lambda_i(1-\alpha_i)$

$\hat{\lambda}_i(1-\hat{\alpha}_i) = 17.5(1-0.45) = 9.6$

A new question

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

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

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

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

Hypothesis testing (e.g., LRT)

$H_0: \beta_5 = 0$ $H_A: \beta_5 \neq 0$

Wald test

function to fit ZIP model (pscl package)
response variable (~)
Poisson (count) component
logistic component

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

(output for Poisson component)

```
...
##                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 ***
...
```

p-value

$H_0: \beta_S = 0 \qquad H_A: \beta_S \neq 0$

$$Z = \frac{\hat{\beta}_S - 0}{SE(\hat{\beta}_S)} = \frac{-0.0056}{0.00067} = -8.338$$

$G = 2(\log L_{full} - \log L_{reduced})$

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
```

$G = 2(-14023.42 + 14058.41)$

$\sim \chi^2_1$

Class activity, Part II

https://sta214-s23.github.io/class_activities/ca_lecture_26.html

Assessing the shape assumption