

# Zero inflated models

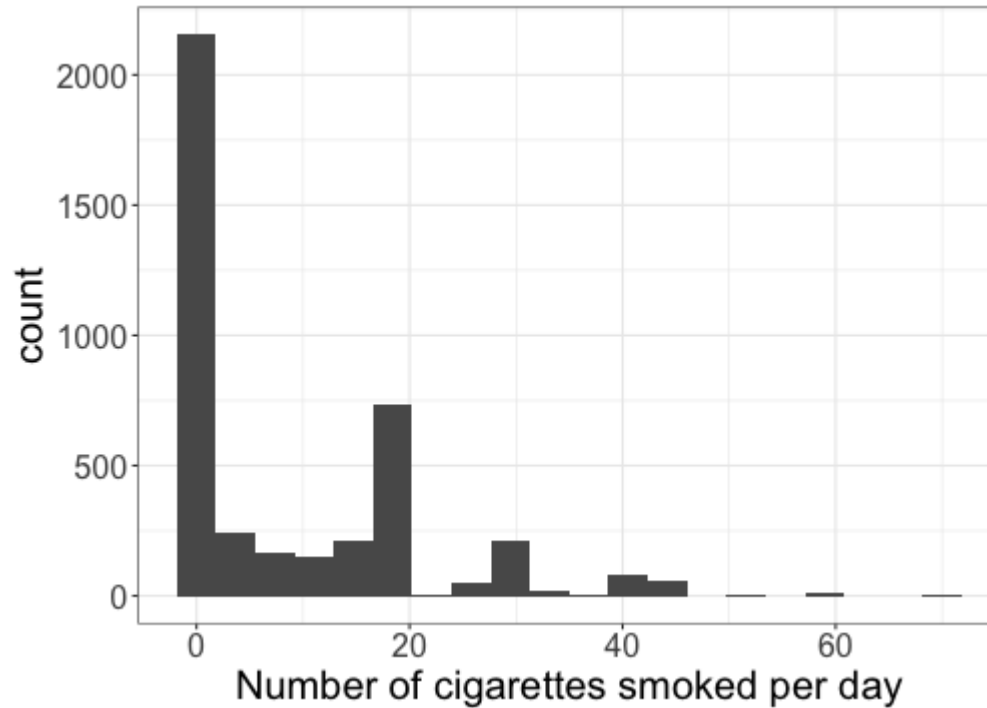
# Data: Framingham heart study

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

## Class activity

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

$$\log\left(\frac{\hat{\alpha}_i}{1 - \hat{\alpha}_i}\right) = -2.51 + 0.051 \text{Age}_i$$

$$\log(\hat{\lambda}_i) = 2.93 - 0.022 \text{EducationSome}_i - 0.067 \text{EducationCollege}_i + \\ 0.009 \text{EducationAdv}_i - 0.046 \text{Diabetes}_i$$

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

## Class activity

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

$$\log\left(\frac{\hat{\alpha}_i}{1 - \hat{\alpha}_i}\right) = -2.51 + 0.051 \text{Age}_i$$

$$\log(\hat{\lambda}_i) = 2.93 - 0.022 \text{EducationSome}_i - 0.067 \text{EducationCollege}_i + \\ 0.009 \text{EducationAdv}_i - 0.046 \text{Diabetes}_i$$

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

## Class activity

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

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

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

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

## Class activity

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

$$\log\left(\frac{\hat{\alpha}_i}{1 - \hat{\alpha}_i}\right) = -2.51 + 0.051 \text{Age}_i$$

$$\log(\hat{\lambda}_i) = 2.93 - 0.022 \text{EducationSome}_i - 0.067 \text{EducationCollege}_i + \\ 0.009 \text{EducationAdv}_i - 0.046 \text{Diabetes}_i$$

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



# Making predictions

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

$$\log\left(\frac{\hat{\alpha}_i}{1 - \hat{\alpha}_i}\right) = -2.51 + 0.051 \text{Age}_i$$

$$\log(\hat{\lambda}_i) = 2.93 - 0.022 \text{EducationSome}_i - 0.067 \text{EducationCollege}_i + \\ 0.009 \text{EducationAdv}_i - 0.046 \text{Diabetes}_i$$

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

## A new question

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

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

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{EducationSome}_i + \beta_2 \text{EducationCollege}_i + \beta_3 \text{EducationAdv}_i + \beta_4 \text{Diabetes}_i$$

New research question: for smokers, does the number of cigarettes smoked per day depend on age?

How would we answer this research question?

## Wald test

```
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
```

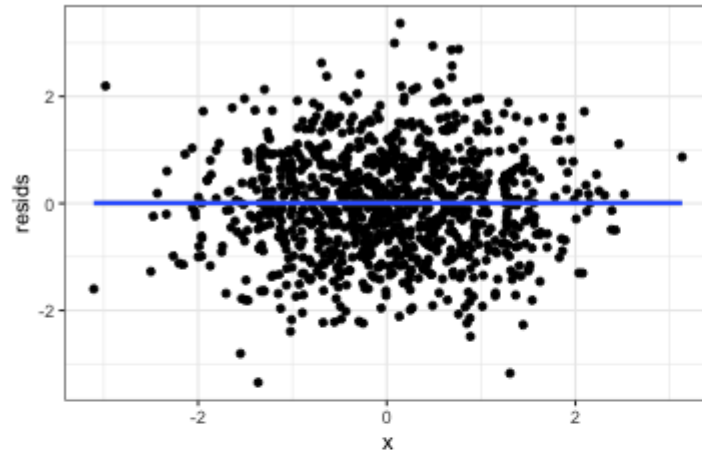
```
## [1] -14058.41
```

# Class activity, Part II

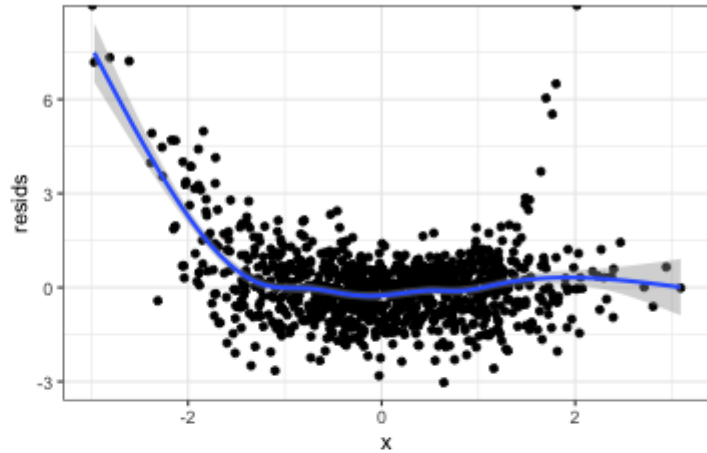
[https://sta214-s23.github.io/class\\_activities/ca\\_lecture\\_26.html](https://sta214-s23.github.io/class_activities/ca_lecture_26.html)

# Assessing the shape assumption

All assumptions satisfied



Poisson shape assumption violated



Logistic shape assumption violated

