Zero inflated Poisson

Mark Andrews
Psychology Department, Nottingham Trent University

mark.andrews@ntu.ac.uk

Probabilistic mixture models; latent class models

- Assume our data is n observations $y_1, y_2 \ldots y_n$.
- A non-mixture model of this data might be

$$y_i \sim N(\mu, \sigma^2), \quad \text{for } i \in 1 \ldots n,$$

- A K component mixture model assumes that there is a discrete latent variable $z_1, z_2 \ldots z_n$, where each $z_i \in \{1, 2 \ldots K\}$, and then (e.g., $K = 3$)

$$y_i \sim \begin{cases} N(\mu_1, \sigma_1^2), & \text{if } z_i = 1 \\ N(\mu_2, \sigma_2^2), & \text{if } z_i = 2 \\ N(\mu_3, \sigma_3^2), & \text{if } z_i = 3 \end{cases},$$

$$z_i \sim P(\pi),$$

where $\pi = [\pi_1, \pi_2, \pi_3]$ is a probability distribution of $\{1, 2, 3\}$.

Probabilistic mixture regression; latent class regression

- Assume our data is $\{(y_1, x_1), (y_2, x_2) \dots (y_n, x_n)\}$.
- In non-mixture regression, we assumes

$$y_i \sim (\alpha + \beta x_i, \sigma^2), \quad \text{for } i \in 1 \dots n,$$

- In a mixture of $K = 3$ regressions, we assume that there is a latent variable $z_1, z_2 \dots z_n$, with each $z_i \in K$, and (e.g., $K = 3$)

$$y_i \sim \begin{cases} N(\alpha_1 + \beta_1 x_i, \sigma_1^2), & \text{if } z_i = 1 \\ N(\alpha_2 + \beta_2 x_i, \sigma_2^2), & \text{if } z_i = 2 \\ N(\alpha_3 + \beta_3 x_i, \sigma_3^2), & \text{if } z_i = 3 \end{cases},$$

$$z_i \sim P(\pi),$$

where $\pi = [\pi_1, \pi_2, \pi_3]$ is a probability distribution of $\{1, 2, 3\}$.

Probabilistic mixture regression; latent class regression

- In the previous mixture of regressions, we assume the probability that each z_i take any value in $1, 2 \ldots K$ is constant, i.e. it is given by π.
- However, the value of z_i could also be a function function of the predictor x_i.
- If $K = 2$ for example, the probability that z_i takes on the value of 1 of 2 (equivalently, 0 or 1) could be determined by a logistic regression with x_i as predictor.

$$y_i \sim \begin{cases} N(\alpha_1 + \beta_1 x_i, \sigma_1^2), & \text{if } z_i = 0 \\ N(\alpha_2 + \beta_2 x_i, \sigma_2^2), & \text{if } z_i = 1 \end{cases},$$

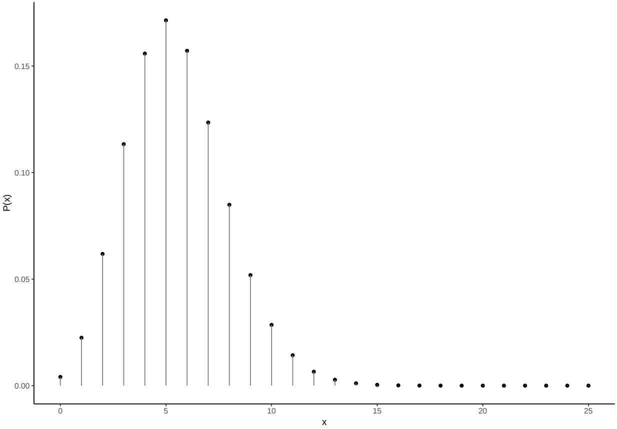
$$\log\left(\frac{P(z_i = 1)}{1 - P(z_i = 1)}\right) = a + b x_i$$

Zero-Inflated Poisson regression

- A zero inflated Poisson regression is $K = 2$ mixture regression model.
- There are two component models, so $K = 2$ and each latent variable $z_i \in \{0, \}$.
- The probability that $z_i = 1$ is a logistic regression function of the predictor(s) x_i.
- The two component of the zero-inflated Poisson model are:
  1. A Poisson distribution.
  2. A zero-valued point mass distribution (a probability distribution with all its mass at zero).

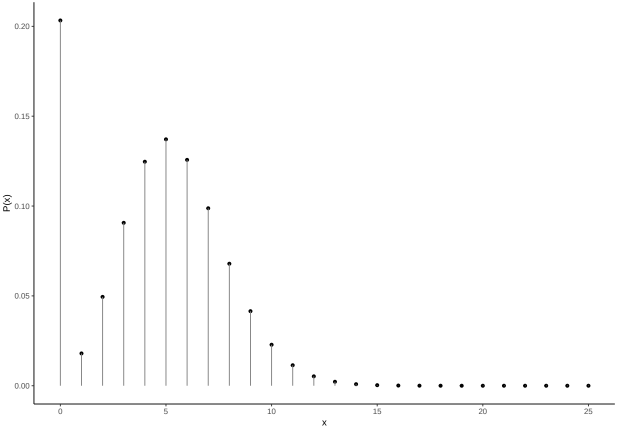
Poisson Distribution

A sample from a Poisson distribution with $\lambda = 5.5$.



Zero inflated Poisson Distribution

A sample from a zero inflated Poisson distribution with $\lambda = 5.5$, with probability of *zero-component* is 0.2.



Poisson regression to Zero-Inflated Poisson regression

- In Poisson regression (with a single predictor, for simplicity), we assume that each y_i is a Poisson random variable with rate λ_i that is a function of the predictor x_i.
- In Zero-Inflated Poisson regression, we assume that each y_i is distributed as a Zero-Inflated Poisson mixture model:

$$y_i \sim \begin{cases} \text{Poisson}(\lambda_i) & \text{if } z_i = 0, \\ 0, & \text{if } z_i = 1 \end{cases}$$

where rate λ_i and $P(z_i = 1)$ are functions of the predictor x_i.

Zero-Inflated Poisson regression

- Assuming data $\{(x_i, y_i), (x_2, y_2) \ldots (x_n, y_n)\}$, Poisson regression models this data as:

$$y_i \sim \begin{cases} \text{Poisson}(\lambda_i) & \text{if } z_i = 0, \\ 0, & \text{if } z_i = 1 \end{cases},$$

$$z_i \sim \text{Bernoulli}(\theta_i),$$

where θ_i and λ_i are functions of x_i.

Zero-Inflated Poisson regression

- The θ_i and λ_i variables are the usual suspects, i.e.

$$\log(\lambda_i) = \alpha + \beta x_i,$$

and

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = a + b x_i.$$

- In other words, λ_i is modelled just as in ordinary Poisson regression and θ_i is modelled in logistic regression.