Zero-Inflated Models

General Principles

Zero-Inflated Regression models are used when the outcome variable is a count variable with an excess of zero counts. These models combine a count model (e.g., Poisson or Negative Binomial) with a separate model for predicting the probability of excess zeros.

Considerations

> Caution
>
> - In Bayesian Zero-Inflated regression, we consider uncertainty in the model parameters and provide a full posterior distribution over them. We need to declare prior distributions for $W_{1\pi}, W_{2\pi}, ..., W_{n\pi}$, $W_{1\lambda}, W_{2\lambda}, ..., W_{n\lambda}$, b_{π}, and b_{λ}.

Example

Below is an example code snippet demonstrating Bayesian Zero-Inflated Poisson regression using the Bayesian Inference (BI) package. The data represent the production of books in a monastery (y), which is affected by the number of days that individuals work, as well as the number of days individuals drink.

Python

```
from BI import bi
from jax.scipy.special import expit
# Setup device ------------------------------------------------
m = bi('cpu')

# Simulated data----------------------------------------------
```

```
prob_drink = 0.2  # 20% of days
rate_work = 1     # average 1 manuscript per day

# Sample one year of production
N = 365
drink = m.dist.binomial(1, prob_drink, shape = (N,), sample = True)
y = (1 - drink) * m.dist.poisson(rate_work, shape = (N,), sample = True)

# Setup device------------------------------------------------
m = bi(platform='cpu')
m.data_on_model = dict(
    y=jnp.array(y)
)

# Define model -----------------------------------------------
def model(y):
    al = dist.normal(1, 0.5, name='al')
    ap = dist.normal(-1.5, 1, name='ap')
    p = expit(ap)
    lambda_ = jnp.exp(al)
    m.zeroinflatedpoisson(p, lambda_, obs=y)

# Run MCMC ---------------------------------------------------
m.fit(model)

# Summary ----------------------------------------------------
m.summary()
```

R

```
library(BI)

# setup platform---------------------------------------------
m=importbi(platform='cpu')

# Simulate data ----------------------------------------------
prob_drink = 0.2  # 20% of days
rate_work = 1     # average 1 manuscript per day
# sample one year of production
N = as.integer(365)
drink = bi.dist.binomial(total_count = as.integer(1), probs = prob_drink, shape = c(N), samp
```

```
y = (1 - drink) *  bi.dist.poisson(rate_work, shape = c(N), sample = T)
data = list()
data$y = y
m$data_on_model = data

# Define model ------------------------------------------------
model <- function(y){
  al = bi.dist.normal(1, 0.5, name='al', shape=c(1))
  ap = bi.dist.normal(-1, 1, name='ap', shape=c(1))
  p = jax$scipy$special$expit(ap)
  lambda_ = jnp$exp(al)
  m$zeroinflatedpoisson(p, lambda_, obs=y)
}

# Run MCMC ----------------------------------------------------
m$run(model) # Optimize model parameters through MCMC sampling

# Summary -----------------------------------------------------
m$summary() # Get posterior distribution
```

Mathematical Details

Frequentist formulation

We model the relationship between the independent variable X and the count outcome variable Y using two components:

1) A logistic regression model to predict the probability of an excess zero.
2) A count model (e.g., Poisson or Negative Binomial) to predict the count outcome.

The overall model can be represented as follows:

$$\begin{aligned}
&\text{logit}(\pi) = \alpha_\pi + \beta_\pi X_i \\
&\log(\lambda) = \alpha_\lambda + \beta_\lambda X_i \\
&Y_i \sim \begin{cases} 0 & \text{with probability } \pi \\ \text{CountModel}(\lambda) & \text{with probability } (1-\pi) \end{cases}
\end{aligned}$$

Where:

- π is the probability of an excess zero.
- λ is the mean rate parameter of the count model.
- α_π and β_π are respectively the intercept and the regression coefficient for the logistic model.
- α_λ and β_λ are respectively the intercept and the regression coefficient for the count model.
- X_i are the independent variables' values for observation i.

Bayesian formulation

In the Bayesian formulation, we define each parameter with priors . We can express the Bayesian regression model accounting for prior distributions as follows:

$$Y \sim ZIPoisson(\pi, \lambda)$$

$$\text{logit}(\pi) = \alpha_\pi + \beta_\pi X$$

$$\log(\lambda) = \alpha_\lambda + \beta_\lambda X$$

$$\alpha_\pi \sim \text{Normal}(0, 1)$$

$$\beta_\pi \sim \text{Normal}(0, 1)$$

$$\alpha_\lambda \sim \text{Normal}(0, 1)$$

$$\beta_\lambda \sim \text{Normal}(0, 1)$$

Where:

- π is the probability of an excess zero.
- λ is the mean rate parameter of the count model.
- α_π and β_π are respectively the intercept and the regression coefficient for the logistic model.

- α_λ and β_λ are respectively the intercept and the regression coefficient for the count model.
- X_i are the independent variables' values for observation i.

Reference(s)

McElreath (2018)

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in r and Stan*. Chapman; Hall/CRC.