Week 6: Event Count Model

POLI803

Howard Liu

Week 6, 2024

University of South Carolina

Outline

Event count models

- General statistical model: A revisit

- Event count models
  - A new probability distribution (actually two distributions)
  - Poisson Model
  - Quasi-poisson Model
  - Negative Binomial Model
  - Zero-Inflated Models

Review: probability distribution

- What is a probability distribution?

Review: probability distribution

- What is a probability distribution?
- Probability distribution = list of probabilities assigned to all possible outcomes
- How do we describe a probability distribution?
- Examples of probability distributions:
  - Bernoulli distribution
  - Normal distribution
  - t distribution
  - Uniform distribution
  - ...

Review: probability distribution

The shape of a probability distribution is determined by parameters.

- Normal distribution (two parameters): mean (μ) and SD (σ)
- Bernoulli distribution (one parameter): probability (p)
- Uniform distribution (two parameters): upper and lower bounds

Notations

- When a variable X follows a Normal distribution with mean μ and SD σ, we write

$$X \sim \mathcal{N}(\mu, \sigma)$$

  - e.g., $X \sim \mathcal{N}(0, 1)$, $Y \sim \mathcal{N}(0, 2)$, $Z \sim \mathcal{N}(2, 2)$

- When a variable X follows a Bernoulli distribution with p, we write

$$X \sim Bernoulli(p)$$

- When a variable X follows a uniform distribution with lower bound l and the upper bound u, we write

$$X \sim \mathcal{U}(l, u)$$

General statistical model

What do probability distributions mean for regression models?

Linear regression model can be represented as

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

General statistical model

What do probability distributions mean for regression models?

Linear regression model can be represented as

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

Or we can write

$$\hat{Y} = \boldsymbol{X}\boldsymbol{\beta}$$

General statistical model

What do probability distributions mean for regression models?

Linear regression model can be represented as

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

Or we can write

$$\hat{Y} = \boldsymbol{X}\boldsymbol{\beta}$$

We can also write

$$Y \sim \mathcal{N}(\mu, \sigma) \quad (1)$$
$$\mu = \boldsymbol{X}\boldsymbol{\beta} \quad (2)$$

- (1) is called the **stochastic component**
- (2) is called the **systematic component**

Logistic regression model

Representation 1 (latent variable)

$$Y^* = \boldsymbol{X}\beta$$
$$\hat{P} = \Lambda(Y^*)$$

Representation 2 (random utility)

$$Y^* = \boldsymbol{X}\beta + \epsilon$$
$$Y = 1 \text{ if } Y^* > 0$$
$$Y = 0 \text{ if } Y^* \leq 0$$

Representation 3 (Stochastic-Systemic)

$$Y \sim Bernoulli(p)$$
$$p = \Lambda(\boldsymbol{X}\beta)$$

General statistical model

- Stochastic component: what kind of probability distribution governs the distribution of Y
  - $Y \sim \mathcal{N}(\mu, \sigma)$
  - $Y \sim Bernoulli(p)$
  - $Y \sim Multinomial(p_1, p_2, ..., p_k)$
- Systematic component: connect the linear predictor with X using a link function
  - Linear link: $\mu = \boldsymbol{X\beta}$
  - Logit link: $p = \Lambda(\boldsymbol{X\beta})$

General statistical model

Linear regression model (Normal-linear)

$$Y \sim \mathcal{N}(\mu, \sigma)$$
$$\mu = \boldsymbol{X}\beta$$

Logistic regression model (Bernoulli-logistic)

$$Y \sim Bernoulli(p)$$
$$p = \Lambda(\boldsymbol{X}\beta)$$

Probit regression Model (Bernoulli-probit)

$$Y \sim Bernoulli(p)$$
$$p = \Phi(\boldsymbol{X}\beta)$$

General statistical model

Ordered logistic regression model (with three categories) → Multinomial-logistic

$$Y \sim Multinomial(p_1, p_2, p_3)$$
$$p_1 = \Lambda(cut_1 - \boldsymbol{X}\beta)$$
$$p_2 = \Lambda(cut_2 - \boldsymbol{X}\beta) - \Lambda(cut_1 - \boldsymbol{X}\beta)$$
$$p_3 = \Lambda(\boldsymbol{X}\beta - cut_2)$$

Multinomial logistic regression model (with three categories) → Multinomial-exp.

$$Y \sim Multinomial(p_1, p_2, p_3)$$
$$p_1 = \frac{\exp(\boldsymbol{X}\beta_1)}{\exp(\boldsymbol{X}\beta_1) + \exp(\boldsymbol{X}\beta_2) + \exp(\boldsymbol{X}\beta_3)}$$
$$p_2 = \frac{\exp(\boldsymbol{X}\beta_2)}{\exp(\boldsymbol{X}\beta_1) + \exp(\boldsymbol{X}\beta_2) + \exp(\boldsymbol{X}\beta_3)}$$
$$p_3 = \frac{\exp(\boldsymbol{X}\beta_3)}{\exp(\boldsymbol{X}\beta_1) + \exp(\boldsymbol{X}\beta_2) + \exp(\boldsymbol{X}\beta_3)}$$

General statistical model

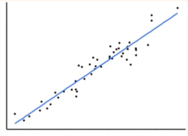
Generalized Linear Models

The approach: allow dependent variable to follow a different distribution

Linear regression

Response variable:
Normal distribution

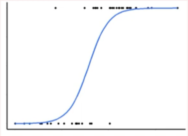
Link function:
Identity



Logistic regression

Response variable:
Bernoulli distribution

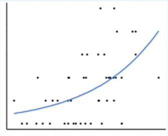
Link function:
Logit



Poisson regression

Response variable:
Poisson distribution

Link function:
Log



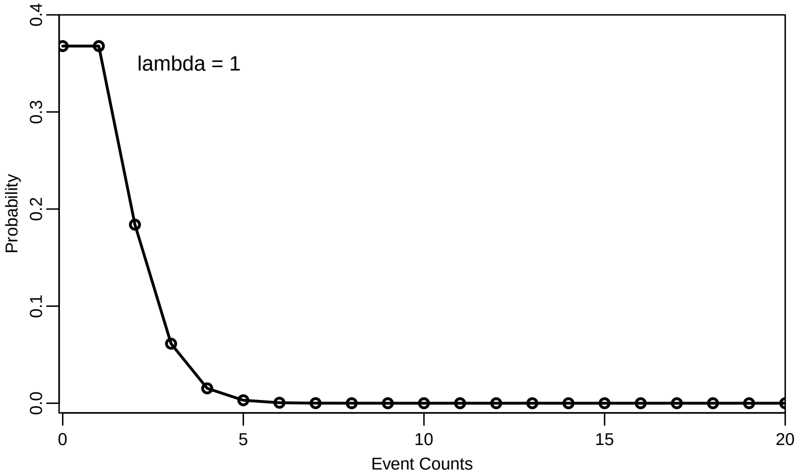
Event count models

Let's say we are interested in Y = the number of times some event happens (0, 1, 2, 3, ...)

- Normal distribution not appropriate
- Bernoulli distribution not appropriate, either
- We can use **Poisson distribution** to describe such process

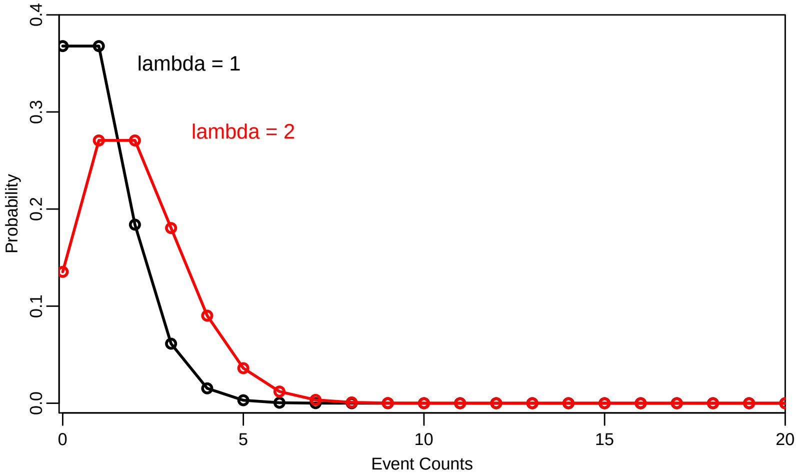
$$Y \sim Poisson(\lambda)$$

Poisson distribution



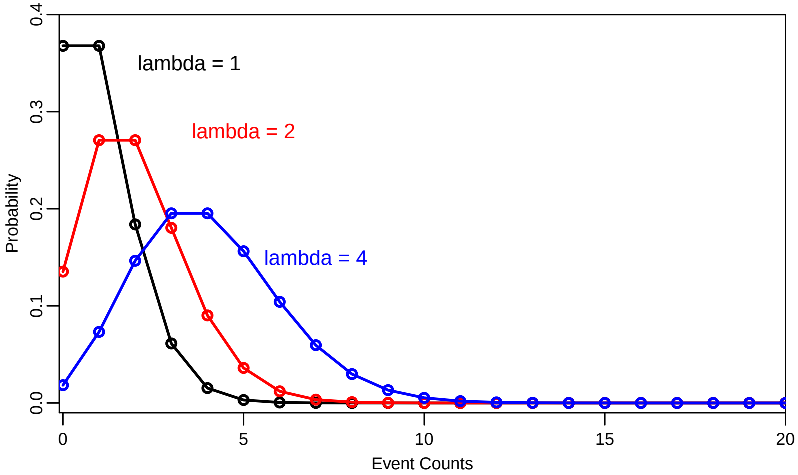
When we set $\lambda = 1$ (e.g., average one attack per year), then the prob. of seeing 4 attacks is 0.01 and 5 attackes is 0.

Poisson distribution

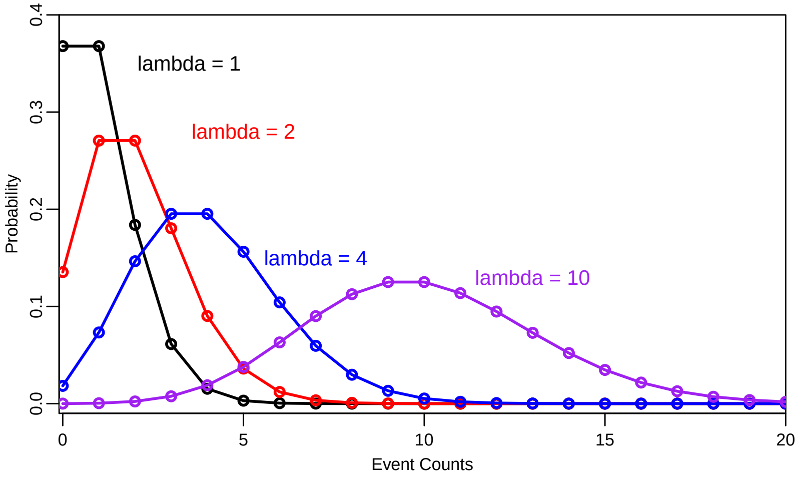


When we set $\lambda = 1$ (e.g., average one attack per year), then the prob. of seeing 4 attacks is 0.1 and 5 attackes is 0.05.

Poisson distribution



Poisson distribution



Poisson regression model

- The stochastic component: Poisson distribution

Poisson regression model

- The stochastic component: Poisson distribution
- The systematic component: connect λ with X

Poisson regression model

- The stochastic component: Poisson distribution
- The systematic component: connect λ with X
- Recall that some parameters have restricted range (e.g., $0 \leq p \leq 1$)

Poisson regression model

- The stochastic component: Poisson distribution
- The systematic component: connect λ with X
- Recall that some parameters have restricted range (e.g., $0 \leq p \leq 1$)
- The parameter of a Poisson distribution, λ, must be positive

$$Y \sim Poisson(\lambda)$$
$$\lambda = \exp(\boldsymbol{X}\beta)$$

where λ **is the mean and the variance**

The Problem of Overdispersion

- This one-to-one relationship (λ **is the mean and the variance**) often fails in real-world data
- Often the variance of the residuals is larger than the mean

$$\text{Poission Assumption: } E[Y] = var(Y)$$
$$\text{Over-dispersion: } E[Y] < var(Y)$$
$$\text{Under-dispersion: } E[Y] > var(Y)$$

R can generate the test of over-dispersion

Quasipoisson regression

- Over-dispersion gives biased coefficient estimates and standard errors
- Need a strategy to disentangle the mean and variance
- We estimate a dispersion parameter ϕ from the residuals

$$\hat{\phi} = \frac{1}{N-k} \sum \left(\frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} \right)$$

$$\text{var} = \phi \text{mean}$$

- So what differs between the poisson and the quasi-poission model?

Quasipoisson regression

- Over-dispersion gives biased coefficient estimates and standard errors
- Need a strategy to disentangle the mean and variance
- We estimate a dispersion parameter ϕ from the residuals

$$\hat{\phi} = \frac{1}{N-k} \sum \left(\frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} \right)$$

$$\text{var} = \phi \text{mean}$$

- So what differs between the poisson and the quasi-poission model? $\rightarrow$ **the s.e. differs but the mean remains unchanged**

Quasipoisson regression

- Over-dispersion gives biased coefficient estimates and standard errors
- Need a strategy to disentangle the mean and variance
- We estimate a dispersion parameter ϕ from the residuals

$$\hat{\phi} = \frac{1}{N-k} \sum \left(\frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} \right)$$

$$\text{var} = \phi \text{mean}$$

- So what differs between the poisson and the quasi-poission model? $\rightarrow$ **the s.e. differs but the mean remains unchanged**
- Quasi-poission will have a larger s.e., but the MLE model is the same

Negative binomial regression

- A different solution is to use a even more flexible **model** with two parameters (λ and θ)
- In practice, negative binomial model is more frequently used

$$Y \sim negbin(\lambda, \theta)$$
$$\lambda = \exp(\boldsymbol{X}\boldsymbol{\beta})$$

where λ is the mean and θ captures the variance

When $\theta = 1$, the model reduces to Poisson

Estimation

- To fit a poisson regression in R:

  glm(Y $\sim$ X1 + X2 + X3..., data = data, family = poisson)

- To fit a negative binomial regression in R:

  ```
  library(MASS)
  ```
  glm.nb(Y $\sim$ X1 + X2 + X3..., data = data)

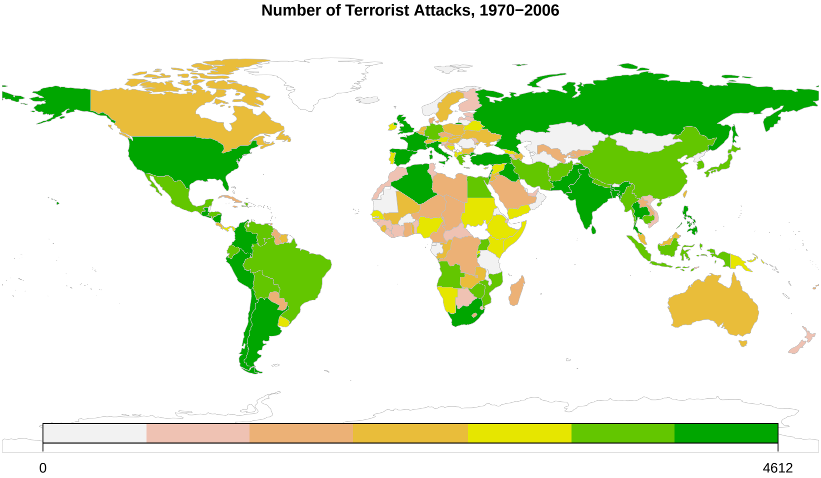
- Note: AIC scores are not comparable across these two models
  - A statistically significant estimate of $\theta \rightsquigarrow$ negative binomial is appropriate (potentially due to an excess of zeros)
  - θ is usually interpreted as a measure of overdispersion in the Negative Binomial distribution

Example: domestic terrorist attacks

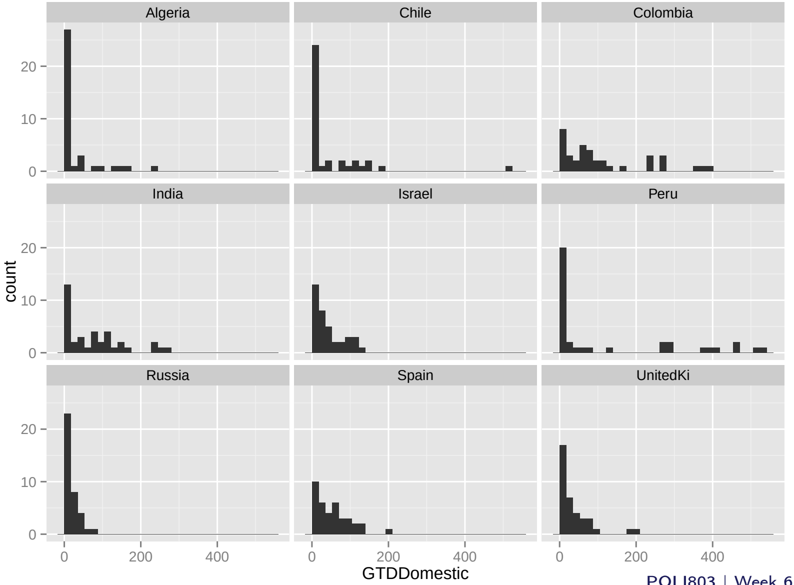
Piazza (2006), *JPR*

- Y: number of domestic terrorist attacks a country experiences per year
- Unit of observation: country-year (172 countries, 1970–2006)

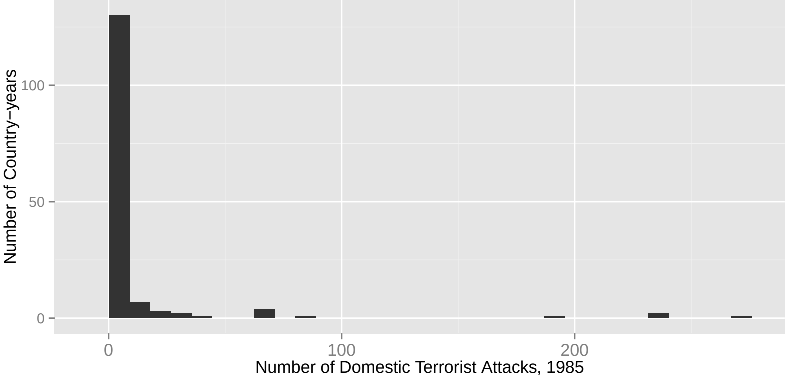
Domestic terrorist attacks: DV



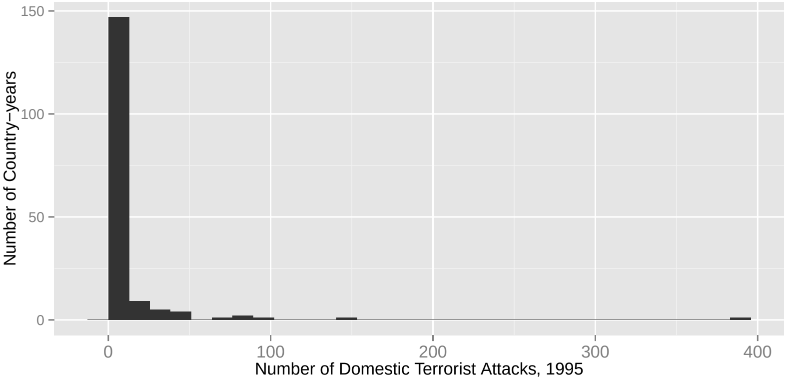
Domestic terrorist attacks: DV



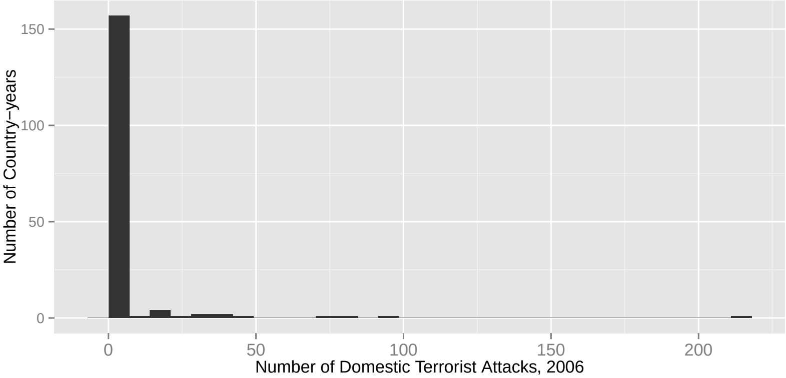
Domestic terrorist attacks: DV



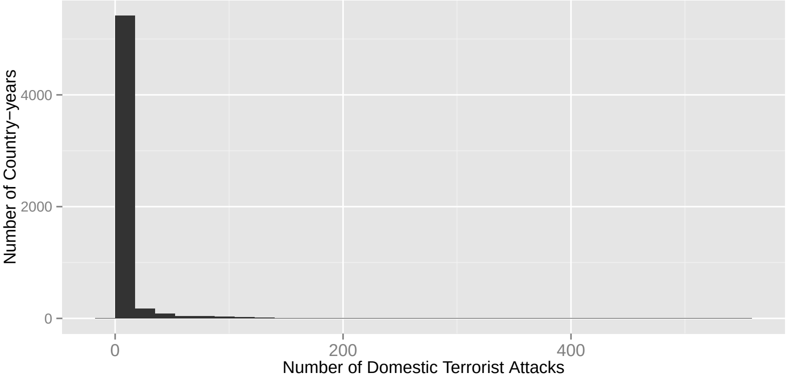
Domestic terrorist attacks: DV



Domestic terrorist attacks: DV

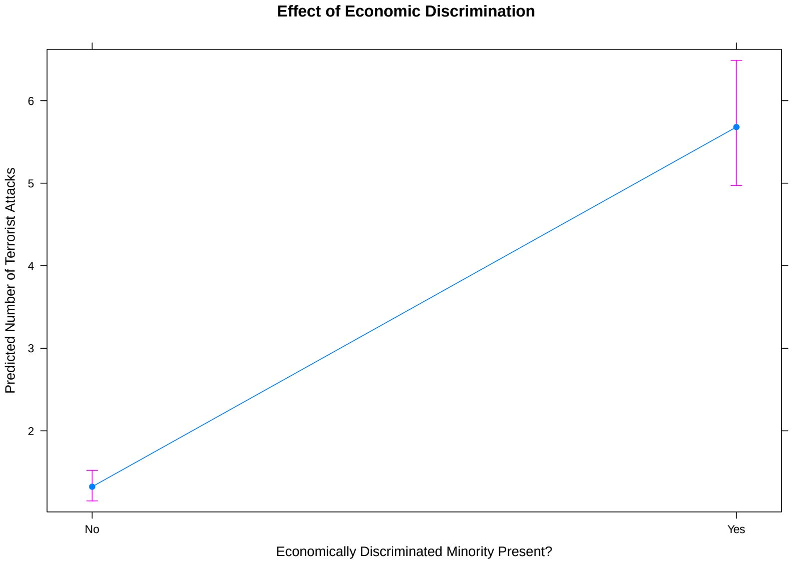


Domestic terrorist attacks: DV

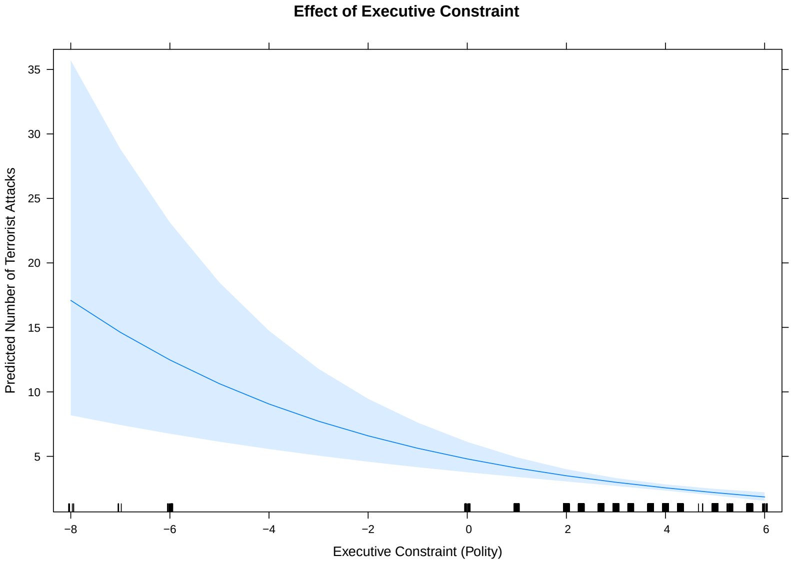


	Poisson	Negative Binomial
ECDIS	1.217***	1.457***
	(0.020)	(0.102)
Population	0.704***	1.082***
	(0.007)	(0.046)
Area	−0.318***	−0.348***
	(0.006)	(0.036)
Durable	−0.003***	−0.009***
	(0.0002)	(0.002)
GNI	0.209***	0.215***
	(0.006)	(0.040)
GINI	0.038***	0.049***
	(0.001)	(0.006)
Partic	−0.222***	−0.299***
	(0.008)	(0.049)
Executive	−0.138***	−0.159***
	(0.004)	(0.032)
Constant	1.288***	0.276
	(0.077)	(0.555)
Observations	2,964	2,964
Log Likelihood	−39,985.850	−5,626.221
θ		0.185*** (0.007)
Akaike Inf. Crit.	79,989.700	11,270.440
Note:		*p<0.1; **p<0.05; ***p<0.01

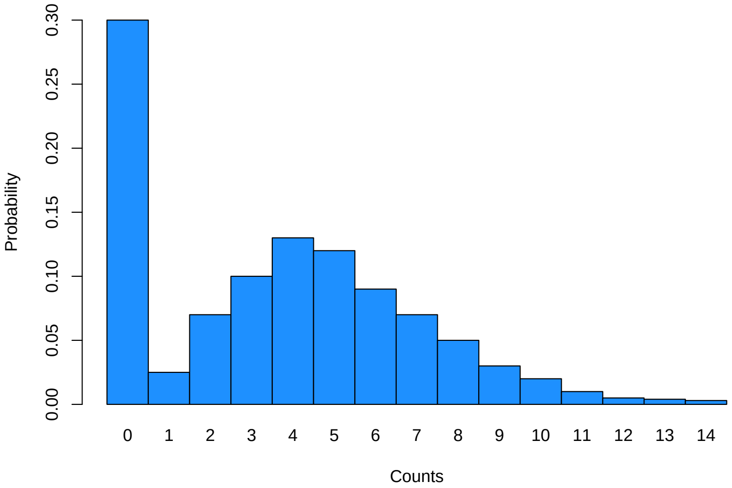
Domestic terrorist attacks: effects



Domestic terrorist attacks: effects



A New Problem: Too many zeros



Zero-inflation

One reason why you might see overdispersion is that there are **too many zeroes** in the count data.

- Empirical reason: by separately accounting for the zeroes, we can do a **better job with standard errors**. $\rightarrow$ more precision

Zero-inflation

One reason why you might see overdispersion is that there are **too many zeroes** in the count data.

- Empirical reason: by separately accounting for the zeroes, we can do a **better job with standard errors**. → more precision
- Theoretical reason: but there is a substantively important reason why we might want to model the extra zeroes. It may be the case that **the zeroes come from a different data generating process than the nonzeroes**.

Zero-inflation

One reason why you might see overdispersion is that there are **too many zeroes** in the count data.

- Empirical reason: by separately accounting for the zeroes, we can do a **better job with standard errors**. → more precision
- Theoretical reason: but there is a substantively important reason why we might want to model the extra zeroes. It may be the case that **the zeroes come from a different data generating process than the nonzeroes**.

Zero-inflation

One reason why you might see overdispersion is that there are **too many zeroes** in the count data.

- Empirical reason: by separately accounting for the zeroes, we can do a **better job with standard errors**. $\rightarrow$ more precision
- Theoretical reason: but there is a substantively important reason why we might want to model the extra zeroes. It may be the case that **the zeroes come from a different data generating process than the nonzeroes**.

Two groups of observations:

Zero-inflation

One reason why you might see overdispersion is that there are **too many zeroes** in the count data.

- Empirical reason: by separately accounting for the zeroes, we can do a **better job with standard errors**. → more precision
- Theoretical reason: but there is a substantively important reason why we might want to model the extra zeroes. It may be the case that **the zeroes come from a different data generating process than the nonzeroes**.

Two groups of observations:

1. Always-zeros: the group that must have a count of 0 (= immune from the event), or

Zero-inflation

One reason why you might see overdispersion is that there are **too many zeroes** in the count data.

- Empirical reason: by separately accounting for the zeroes, we can do a **better job with standard errors**. $\rightarrow$ more precision
- Theoretical reason: but there is a substantively important reason why we might want to model the extra zeroes. It may be the case that **the zeroes come from a different data generating process than the nonzeroes**.

Two groups of observations:

1. Always-zeros: the group that must have a count of 0 (= immune from the event), or
2. Maybe-zeros: the group that can have a count of 0, but might have a nonzero count.

Zero-inflation

These two groups are only **partially observed**.

Zero-inflation

These two groups are only **partially observed**.

- If the count is nonzero $\rightsquigarrow$ we **know** that the observation is in the second group

Zero-inflation

These two groups are only **partially observed**.

- If the count is nonzero $\rightsquigarrow$ we **know** that the observation is in the second group
- If the count is zero $\rightsquigarrow$ we can only **estimate** which group it comes from

Zero-inflation

These two groups are only **partially observed**.

- If the count is nonzero $\rightsquigarrow$ we **know** that the observation is in the second group
- If the count is zero $\rightsquigarrow$ we can only **estimate** which group it comes from

We model two things at the same time:

1. the probability that each observation could have been in each group, and
2. the expected count for observations in the nonzero-count group.

Zero-inflation

These two groups are only **partially observed**.

- If the count is nonzero $\rightsquigarrow$ we **know** that the observation is in the second group
- If the count is zero $\rightsquigarrow$ we can only **estimate** which group it comes from

We model two things at the same time:

1. the probability that each observation could have been in each group, and
2. the expected count for observations in the nonzero-count group.

This model leads to zero-inflated Poisson and to zero-inflated negative binomial.

Let **group** A be the group that must always be 0.
Let **group** B be the group with potentially nonzero counts.

Let **group** A be the group that must always be 0.
Let **group** B be the group with potentially nonzero counts.

For group A, the count must be zero, so the PMF (Prob. mass function, for discrete variables) is:

$$f_A(y_i = 0) = 1, \; f_A(y_i > 0) = 0.$$

Let **group** A be the group that must always be 0.
Let **group** B be the group with potentially nonzero counts.

For group A, the count must be zero, so the PMF (Prob. mass function, for discrete variables) is:

$$f_A(y_i = 0) = 1, \; f_A(y_i > 0) = 0.$$

For group B, the count is Poisson distributed:

$$f_B(y_i|\lambda_i) = \frac{\lambda_i^y}{y_i!} e^{-\lambda_i}$$

Let **group** A be the group that must always be 0.
Let **group** B be the group with potentially nonzero counts.

For group A, the count must be zero, so the PMF (Prob. mass function, for discrete variables) is:

$$f_A(y_i = 0) = 1,\ f_A(y_i > 0) = 0.$$

For group B, the count is Poisson distributed:

$$f_B(y_i|\lambda_i) = \frac{\lambda_i^y}{y_i!}e^{-\lambda_i}$$

Suppose an observation belongs to group A with probability π_i and group B with probability $1 - \pi_i$.
Then any observation has the average of these two distributions:

$$\begin{aligned} f(y_i|\pi_i, \lambda_i) &= \pi_i f_A(y_1 = 0) + (1 - \pi_i) f_B(y_i|\lambda_i) \\ &= \pi_i + (1 - \pi_i)\frac{\lambda_i^y}{y_i!}e^{-\lambda_i} \end{aligned}$$

$$f(y_i|\pi_i, \lambda_i) = \pi_i + (1 - \pi_i)\frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}$$

$$f(y_i|\pi_i,\lambda_i) = \pi_i + (1-\pi_i)\frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}$$

If $y_i = 0$, then this PMF becomes

$$f(y_i = 0|\pi_i,\lambda_i) = \pi_i + (1-\pi_i)\frac{\lambda^0}{0!}e^{-\lambda_i}$$
$$= \pi_i + (1-\pi_i)e^{-\lambda_i}$$

$$f(y_i|\pi_i, \lambda_i) = \pi_i + (1 - \pi_i)\frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}$$

If $y_i = 0$, then this PMF becomes

$$f(y_i = 0|\pi_i, \lambda_i) = \pi_i + (1 - \pi_i)\frac{\lambda^0}{0!}e^{-\lambda_i}$$
$$= \pi_i + (1 - \pi_i)e^{-\lambda_i}$$

If $y_i > 0$, then this PMF becomes

$$f(y_i > 0|\pi_i, \lambda_i) = (1 - \pi_i)\frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}$$

$$f(y_i|\pi_i, \lambda_i) = \pi_i + (1 - \pi_i)\frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}$$

If $y_i = 0$, then this PMF becomes

$$\begin{aligned} f(y_i = 0|\pi_i, \lambda_i) &= \pi_i + (1 - \pi_i)\frac{\lambda^0}{0!}e^{-\lambda_i} \\ &= \pi_i + (1 - \pi_i)e^{-\lambda_i} \end{aligned}$$

If $y_i > 0$, then this PMF becomes

$$f(y_i > 0|\pi_i, \lambda_i) = (1 - \pi_i)\frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}$$

Define a dummy variable I_{0i} to indicate whether $y_i = 0$, then the whole stochastic component is

$$f(y_i|\pi_i, \lambda_i) = \Bigg(\underbrace{\pi_i + (1 - \pi_i)e^{-\lambda_i}}_{\text{Bernoulli}}\Bigg)^{I_{0i}} \Bigg(\underbrace{(1 - \pi_i)\frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}}_{\text{Count}}\Bigg)^{1-I_{0i}}$$

Zero-inflated Poisson

We are going to fit both π_i and λ_i with linear aggregators, so that we can predict which observations have a count, and the count for those that do.

$$y^*_{i\lambda} = \alpha_1 + \beta_1 x_{1i}$$
$$y^*_{i\pi} = \alpha_2 + \beta_2 x_{2i}$$

Zero-inflated Poisson

We are going to fit both π_i and λ_i with linear aggregators, so that we can predict which observations have a count, and the count for those that do.

$$y^*_{i\lambda} = \alpha_1 + \beta_1 x_{1i}$$
$$y^*_{i\pi} = \alpha_2 + \beta_2 x_{2i}$$

and two link functions

$$\lambda_i = e^{y^*_{i\lambda}}$$
$$\pi_i = \frac{1}{1 + e^{-y^*_{i\pi}}}$$

Note: we have different coefficients and we can have different x variables for each part.

Zero-inflated Poisson

Quantities of interest:

- Probability of not being "at risk" (immune to events):

$$\pi_i$$

Zero-inflated Poisson

Quantities of interest:

- Probability of not being "at risk" (immune to events):

$$\pi_i$$

- Average count, conditional on having a count at all:

$$\lambda_i$$

Zero-inflated Poisson

Quantities of interest:

- Probability of not being "at risk" (immune to events):

$$\pi_i$$

- Average count, conditional on having a count at all:

$$\lambda_i$$

- Average count:

$$(1 - \pi_i)\lambda_i$$

Zero-inflated Poisson

Quantities of interest:

- Probability of not being "at risk" (immune to events):

$$\pi_i$$

- Average count, conditional on having a count at all:

$$\lambda_i$$

- Average count:

$$(1 - \pi_i)\lambda_i$$

Note: when a particular x_k appears both in the y^*_π equation and in the y^*_λ equation, the sign of β_k can be really misleading. Interpret them carefully.

Estimation: pscl package

- To fit a zero-inflated poisson regression in R:

```
library(pscl)
zeroinfl(Y ~ X1 + X2 + X3... | Z1 + Z2 + Z3 ..., data = data,
dist = "poisson")
```

- To fit a zero-inflated negative binomial regression in R:

```
library(pscl)
zeroinfl(Y ~ X1 + X2 + X3... | Z1 + Z2 + Z3 ..., data = data,
dist = "negbin")
```