Agenda

Expanding on the Poisson regression

1. Administrative
2. Interpreting coefficients from Poisson regressions
3. "Over-dispersed" Poisson regressions
4. Zero-inflated Poisson regressions
5. ***Hands on:*** *Poisson models in R*

Interpreting Poisson coefficients

INTERPRETING COEFFICIENTS

$$H_i \sim \text{Pois}(\lambda_i)$$
$$\log(\lambda_i) = \alpha + \beta_M M_i + \beta_G G_i$$

$$\alpha \sim \text{Norm}(3, 1)$$
$$\beta_M \sim \text{Norm}(0, 0.5)$$
$$\beta_G \sim \text{Norm}(0, 0.3)$$

	Mean	exp(Mean)
α	0.27	1.32
β_M	1.11	3.05
β_G	-0.14	0.87

α (baseline)

A student who is not a boy (M_i = 0) and is in grade 10 (G_i = 0) is predicted to play about 1.32 hours of games per week.

β_M (gender)

Boys (M_i = 1) are expected to spend about 3.05 times more time than non-boys (M_i = 0) playing games.

β_G (grade)

A one-year increase in grade is associated with playing 0.87 times as much. This is a decrease of 13% each year.

Over-dispersion

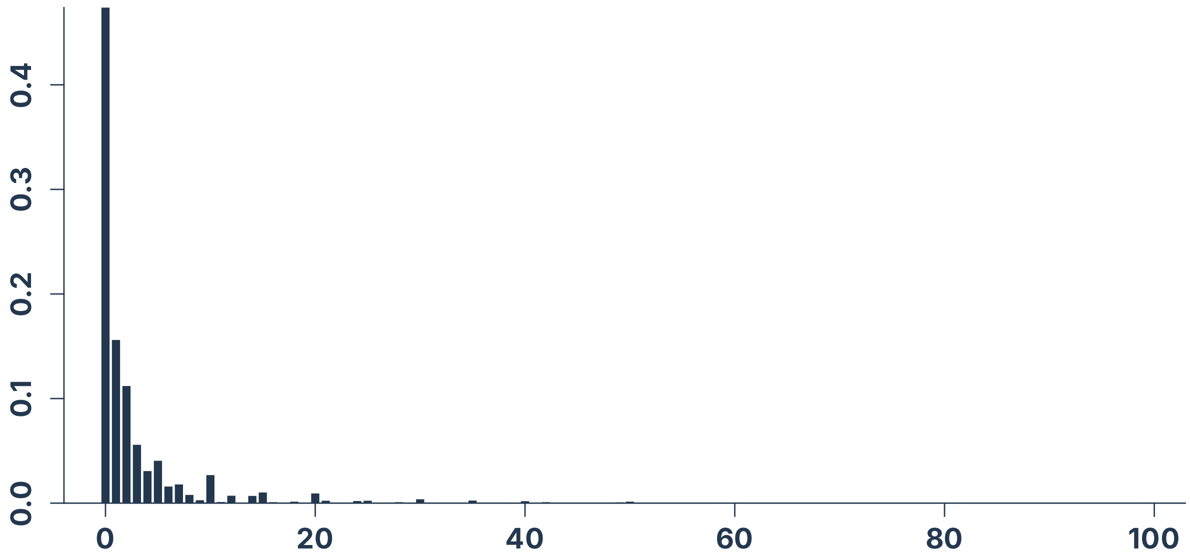
To illustrate, consider the simpler model with only gender as a predictor:

$$H_i \sim \text{Pois}(\lambda_i)$$
$$\log(\lambda_i) = \alpha + \beta M_i$$

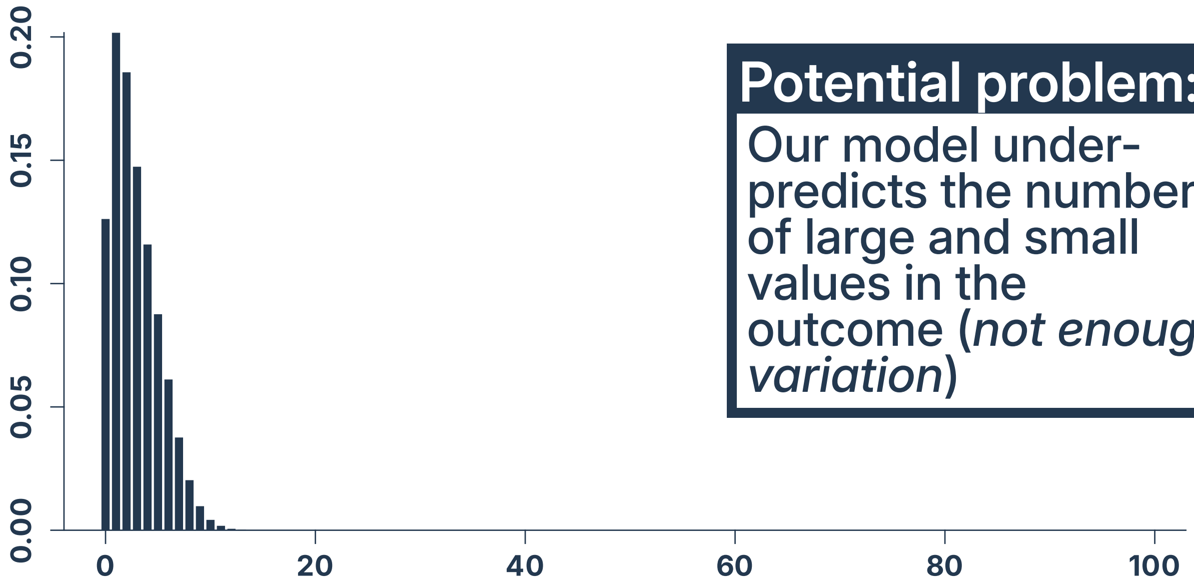
$$\alpha \sim \text{Norm}(3, 1.0)$$
$$\beta \sim \text{Norm}(0, 0.7)$$

	Mean	exp(Mean)
α	0.38	1.46
β	1.10	3.00

Actual distribution



Posterior predicted distribution (Poisson)



Potential problem:

Our model under-predicts the number of large and small values in the outcome (*not enough variation*)

Gamma-Poisson regression:

AKA ***negative binomial***
AKA ***over-dispersed Poisson***

$$H_i \sim \text{Pois}(\lambda_i)$$
$$\lambda_i \sim \text{Gamma}(\mu_i, \theta)$$
$$\log(\mu_i) = \alpha + \beta M_i$$

$$\alpha \sim \text{Norm}(3, 1.0)$$
$$\beta \sim \text{Norm}(0, 0.7)$$
$$\theta \sim \text{Unif}(0, 20)$$

Extra "dispersion" from gamma

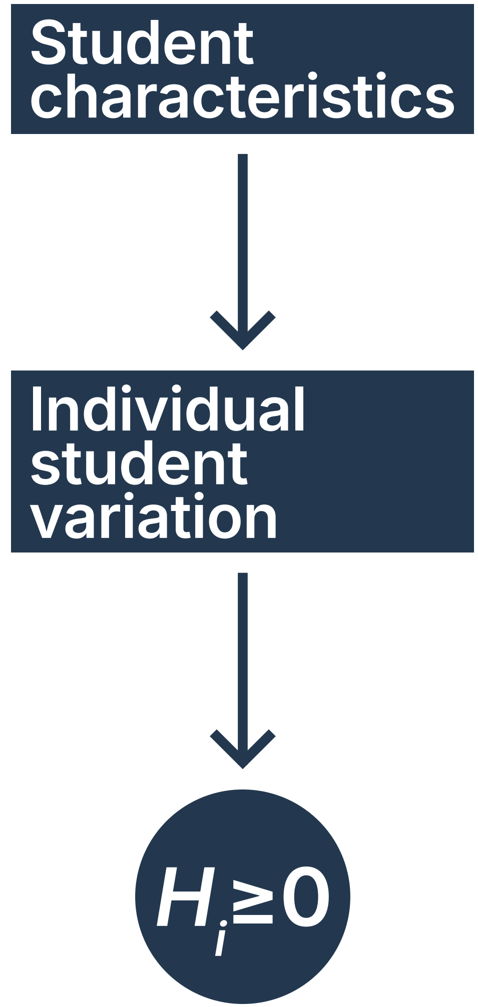
Two students who look identical based on covariates can have different Poisson rates λ_i.

One more prior

Gamma-Poisson regression:

$$
\begin{aligned}
H_i &\sim \text{Pois}(\lambda_i) \\
\lambda_i &\sim \text{Gamma}(\mu_i, \theta) \\
\log(\mu_i) &= \alpha + \beta M_i \\
\\
\alpha &\sim \text{Norm}(3, 1.0) \\
\beta &\sim \text{Norm}(0, 0.7) \\
\theta &\sim \text{Unif}(0, 20)
\end{aligned}
$$

Data story:



Gamma-Poisson and negative-binomial regressions are the same:

$$\begin{aligned} H_i &\sim \text{Pois}(\lambda_i) \\ \lambda_i &\sim \text{Gamma}(\mu_i, \theta) \\ \log(\mu_i) &= \alpha + \beta M_i \end{aligned}$$

$$\begin{aligned} H_i &\sim \text{NegBin}(\lambda_i, \theta) \\ \log(\lambda_i) &= \alpha + \beta M_i \end{aligned}$$

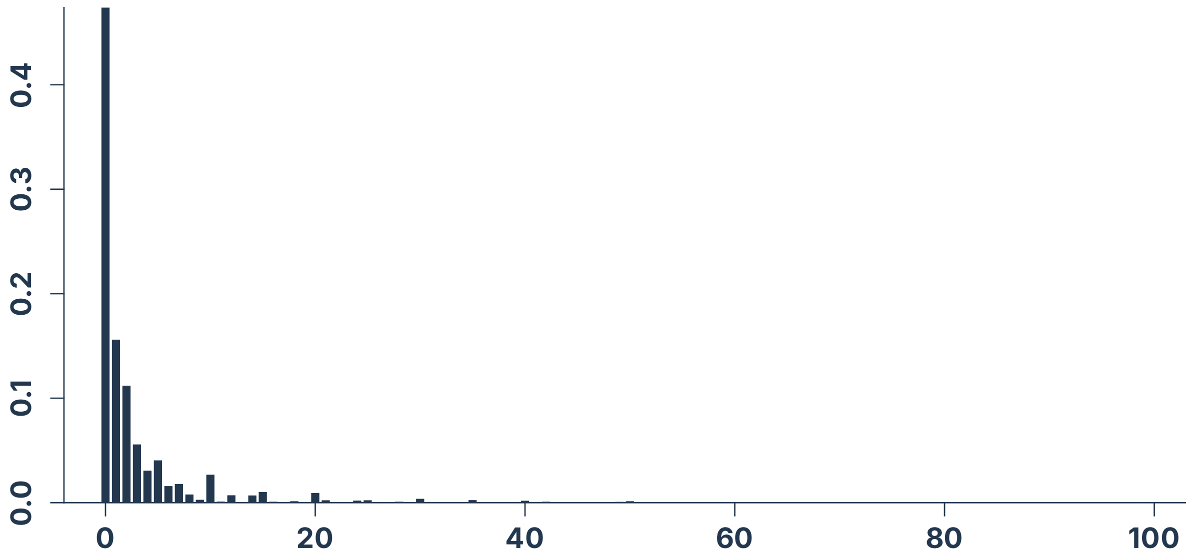
"***Negative binomial regression***" is the typical terminology

$$
\begin{aligned}
H_i &\sim \mathrm{NegBin}(\lambda_i, \theta) \\
\log(\lambda_i) &= \alpha + \beta M_i \\
\\
\alpha &\sim \mathrm{Norm}(3, 1.0) \\
\beta &\sim \mathrm{Norm}(0, 0.7) \\
\theta &\sim \mathrm{Unif}(0, 20)
\end{aligned}
$$

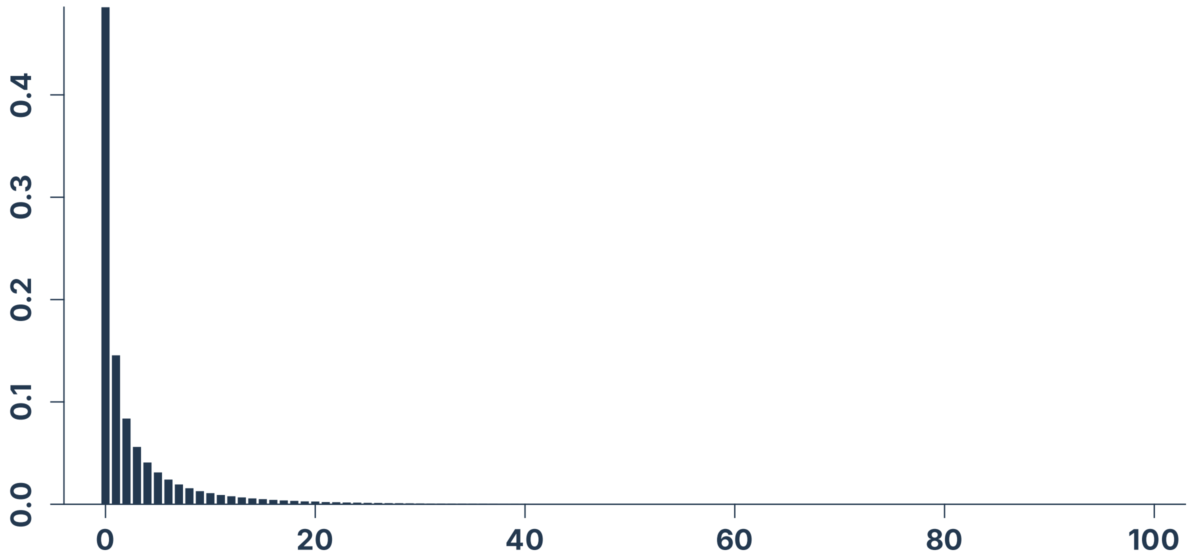
	Mean	95% CI	exp(Mean)
α	0.38	(0.32, 0.45)	1.47
β	1.09	(1.01, 1.18)	2.99
θ	0.35	(0.33, 0.37)	—

OVER-DISPERSION

Actual distribution



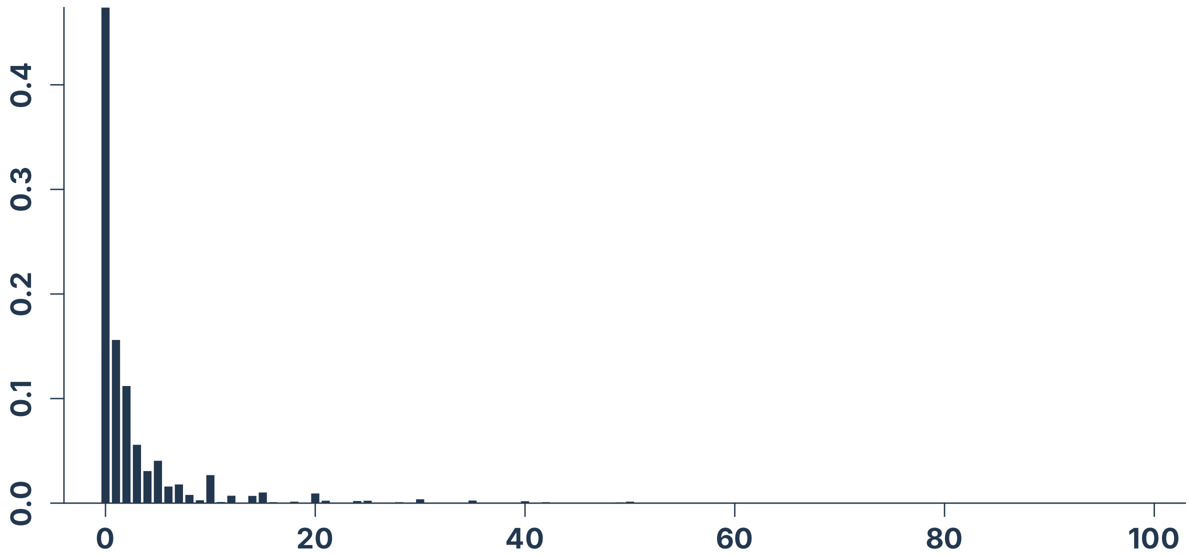
Posterior predicted distribution (Gamma-Poisson)



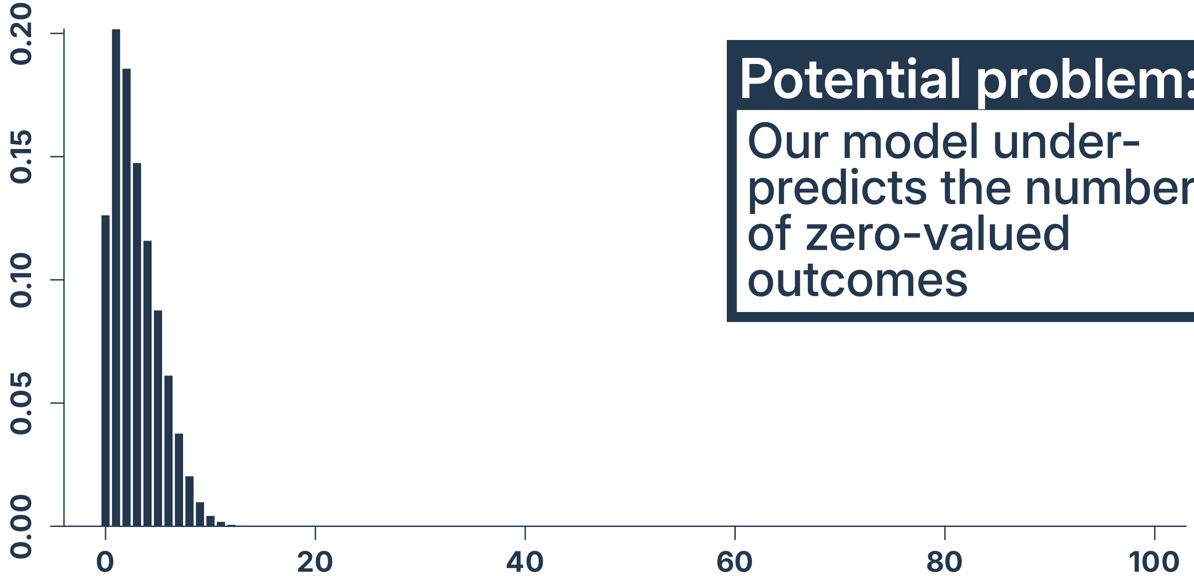
Zero inflation

ZERO INFLATION

Actual distribution



Posterior predicted distribution (Poisson)



Potential problem:

Our model under-predicts the number of zero-valued outcomes

ZERO INFLATION

Outcome variable is the result of *one of two processes*:

Either the student is structurally constrained to play zero hours per week–e.g. they do not own a game console ($C_i = 1$)

Or the student *is* able to play games and does so at a rate λ_i ($C_i = 0$)

$$H_i \begin{cases} = 0 & \text{if } C_i = 1 \\ \sim \text{Pois}(\lambda_i) & \text{if } C_i = 0 \end{cases}$$

ZERO INFLATION

Each student's probability of not owning a console is modeled with p_i

$$H_i \begin{cases} = 0 & \text{if } C_i = 1 \\ \sim \text{Pois}(\lambda_i) & \text{if } C_i = 0 \end{cases}$$

$$C_i = \text{Bernoulli}(p_i)$$

ZERO INFLATION

The probability p_i is modeled with a linear function of family income

$$H_i \begin{cases} = 0 & \text{if } C_i = 1 \\ \sim \text{Pois}(\lambda_i) & \text{if } C_i = 0 \end{cases}$$

$$C_i = \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \alpha_p + \beta_p W_i$$

ZERO INFLATION

The rate λ_i is modeled with a linear function of gender

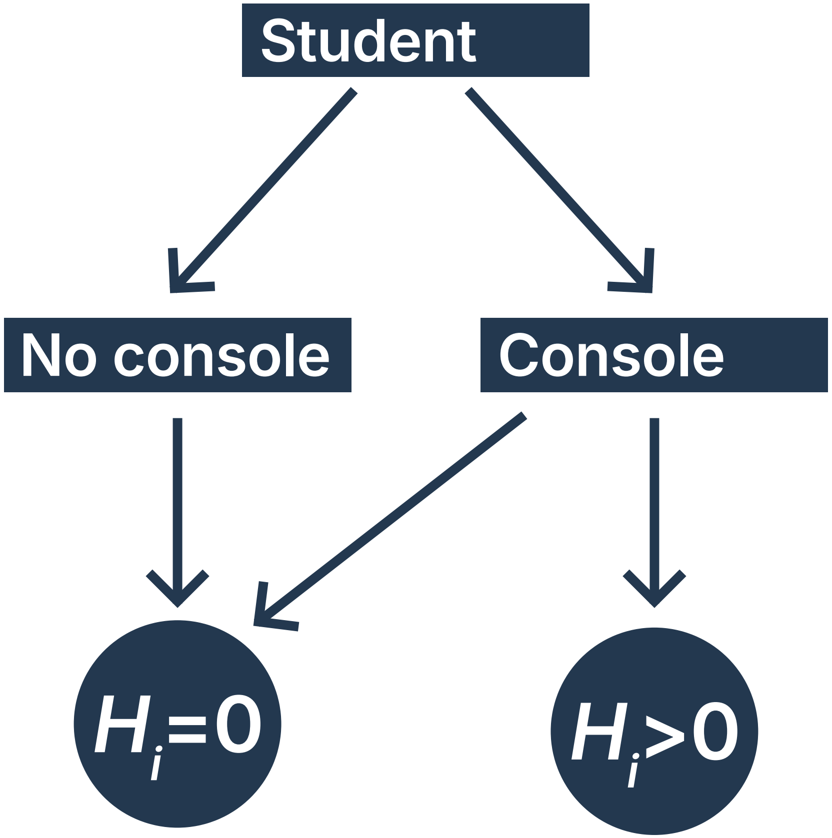
$$H_i \begin{cases} = 0 & \text{if } C_i = 1 \\ \sim \text{Pois}(\lambda_i) & \text{if } C_i = 0 \end{cases}$$

$$C_i = \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \alpha_p + \beta_p W_i$$

$$\log(\lambda_i) = \alpha_\lambda + \beta_\lambda M_i$$

Data story:



$$H_i \begin{cases} = 0 & \text{if } C_i = 1 \\ \sim \text{Pois}(\lambda_i) & \text{if } C_i = 0 \end{cases}$$

$$C_i = \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \alpha_p + \beta_p W_i$$
$$\log(\lambda_i) = \alpha_\lambda + \beta_\lambda M_i$$

$$\alpha_p \sim \text{Norm}(0, 1)$$
$$\beta_p \sim \text{Norm}(0, 2)$$
$$\alpha_\lambda \sim \text{Norm}(3, 1)$$
$$\beta_\lambda \sim \text{Norm}(0, 0.7)$$

MODEL COMPARISONS

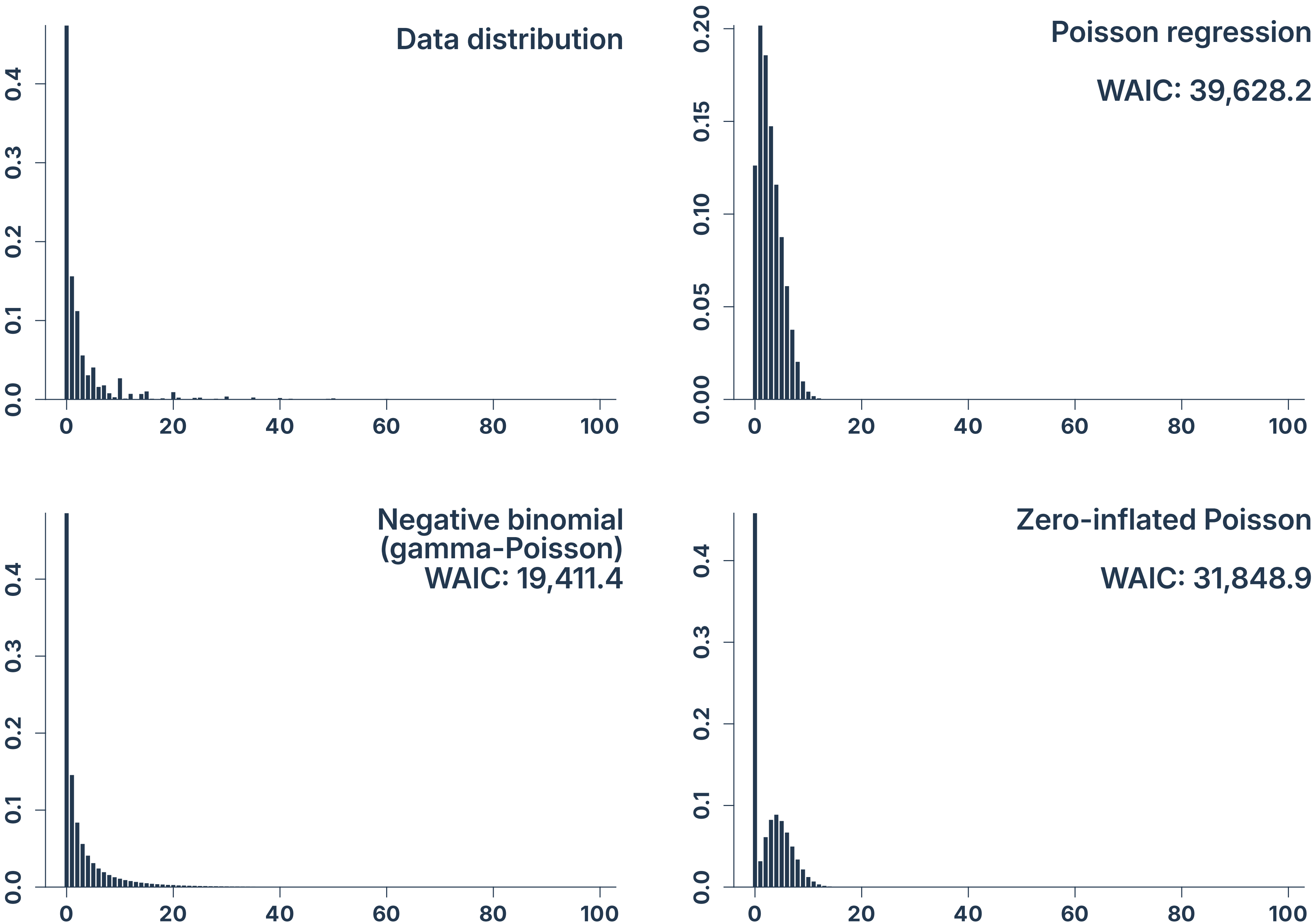
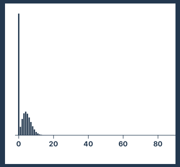


Image credit



Figures by Peter McMahan (source code)

Still from National Treasure (2004)

Still from Brazil (1985)

Merchandise from The Nightmare Before Christmas (1993)