Integrating Poisson regression into the undergraduate curriculum

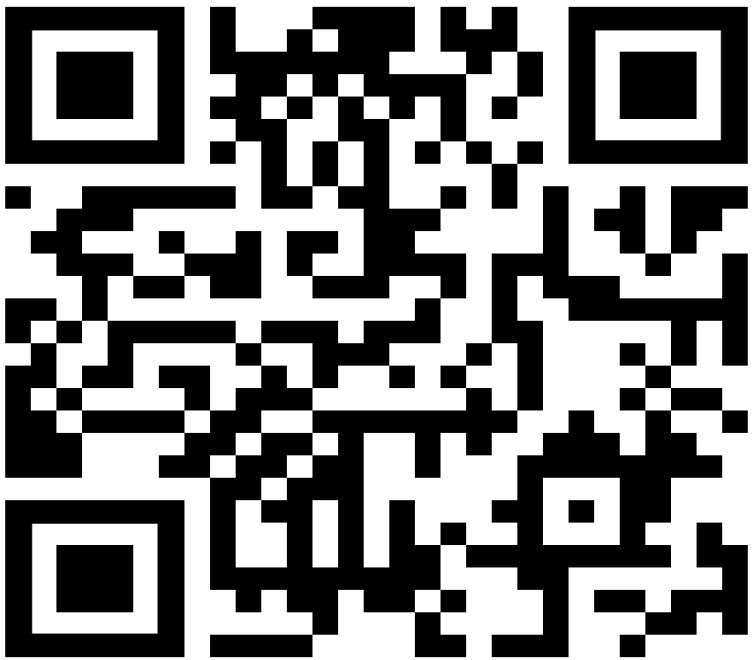
USCOTS25 Breakout Session B3H

Laura Boehm Vock and Paul Roback

ST. OLAF COLLEGE

A quick initial survey!

Please click here or use the following QR code:



Poisson regression at St. Olaf

- 02-04: Not taught. Statistics concentration required Prob Theory and Math Stat plus 2 electives.
- 04-18: Taught as part of Advanced Statistical Modeling (Stat 316). Concentration required Statistical Modeling (Stat 272) and 316 plus 2 electives.
- 18-24: Still taught in Stat 316. Concentration renamed "Statistics and Data Science" and required 272 and Intro to Data Science plus 2 electives. Stat 316 now counts as an upper level elective.
- 24-current: Still taught in Stat 316. Concentration became a major. Stat 316 counts as a "Level 3 Stats Depth" elective course.

Advanced Statistical Modeling at St. Olaf

- Covers generalized linear models (Poisson regr, binomial regr, negative binomial regr, zero-inflated models, hurdle models, etc.) and multilevel modeling
- Prerequsites: Intro Stats and Stat Modeling (nothing else – calculus, linear algebra, computing, ...)
- Applied focus using R
- Uses Beyond Multiple Linear Regression: Applied Generlized Linear Models and Multilevel Models in R by Roback and Legler. Second edition by Roback, Boehm Vock, and Legler expected by Fall 2026.

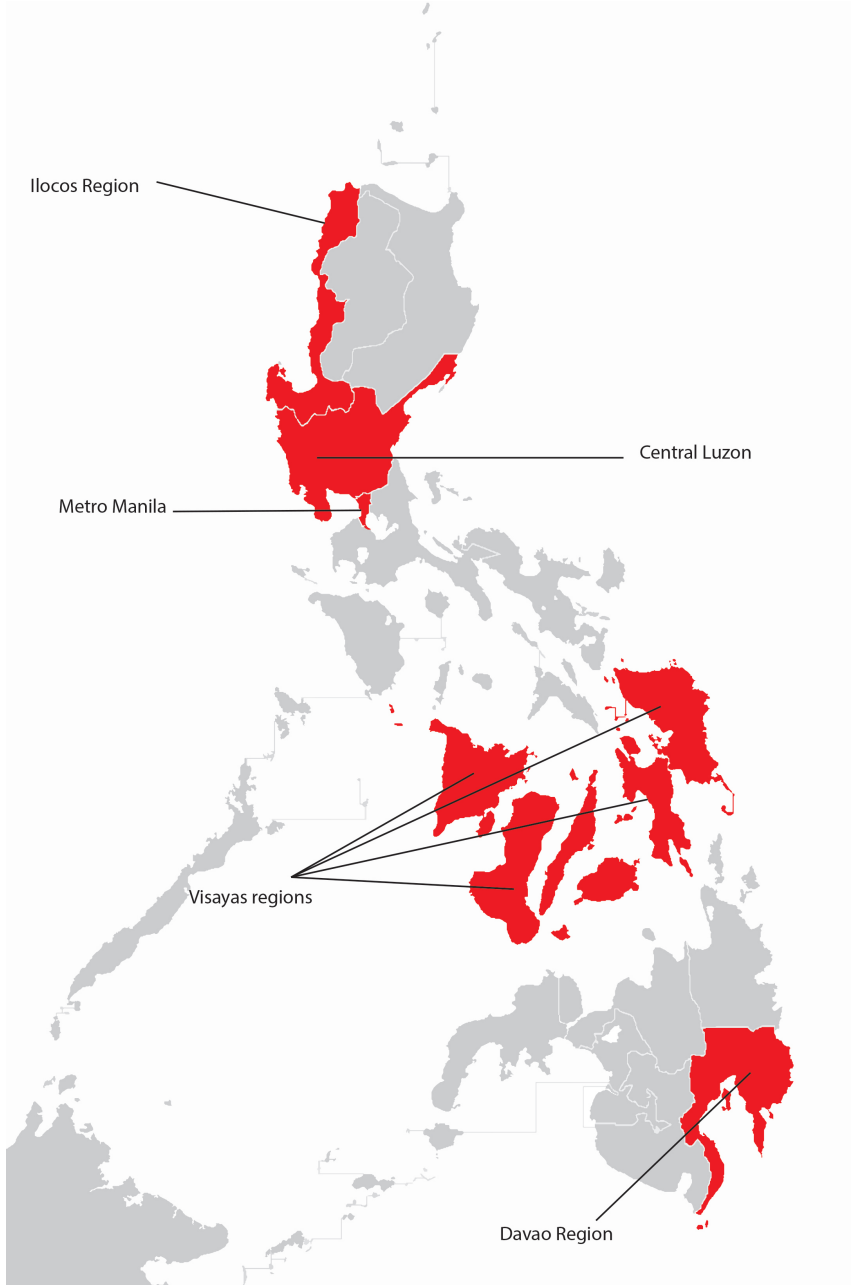
First case study: Philippine households

- International agencies often use household size to determine the magnitude of the household needs
- Want to discern factors associated with larger households
- Data is subset from 2015 Philippine Statistics Authority's Family Income and Expenditure Survey (FIES)
- Primary response is a count, which can make linear regression problematic

Philippine household data

Key variables:

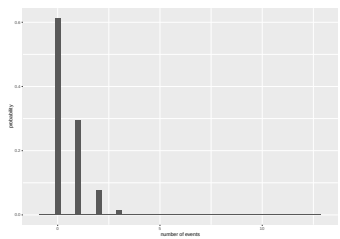
- `location` = region (Central Luzon, Davao, Ilocos, Metro Manila, or Visayas)
- `age` = the age of the head of household
- `total` = the number of people in the household other than the head
- `numLT5` = the number in the household under 5 years of age
- `roof` = the type of roof (stronger material can be used as a proxy for greater wealth)



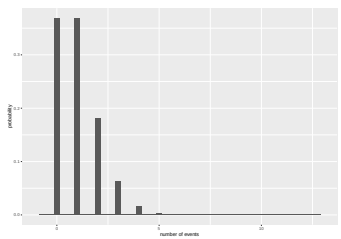
Poisson distribution

$$P(Y_i = y_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \quad \text{for} \quad y_i = 0, 1, \ldots, \infty,$$

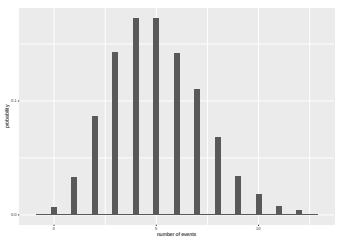
Note that both $E(Y_i) = \lambda_i$ and $Var(Y_i) = \lambda_i$.



(a) $\lambda = 0.5$



(b) $\lambda = 1$



(c) $\lambda = 5$

Figure 1: Poisson distributions with $\lambda = 0.5,\ 1$, and 5.

Poisson regression model

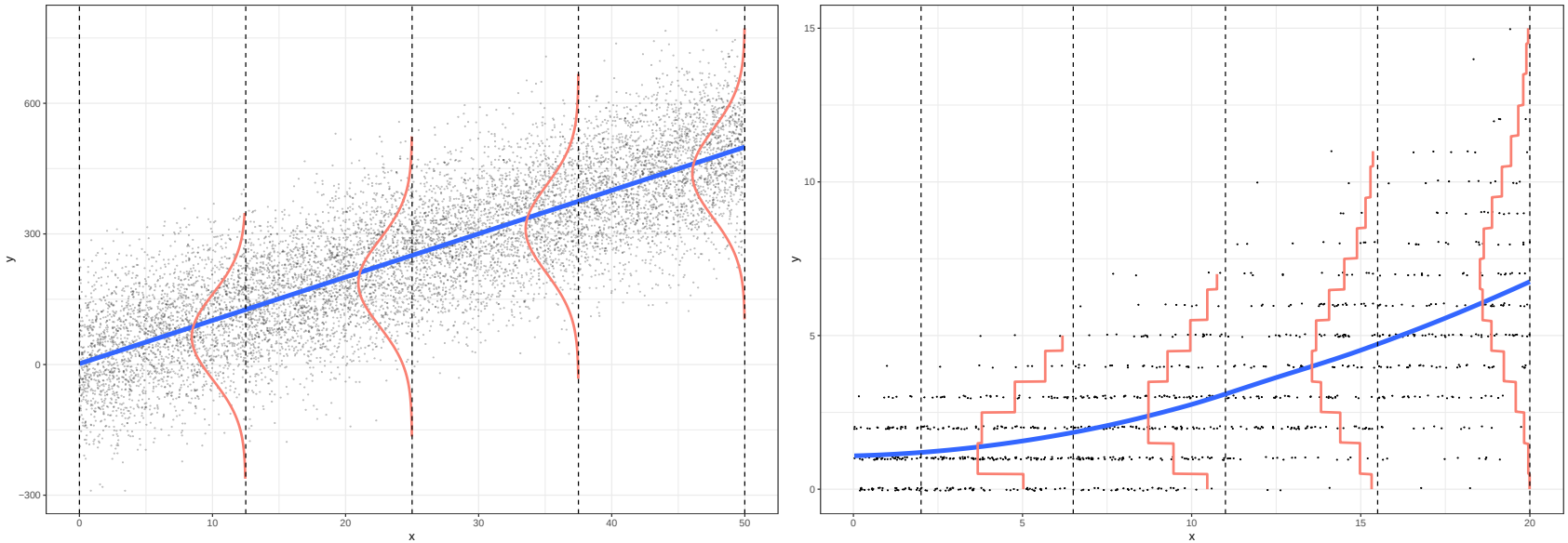
$$log(\lambda_i) = \beta_0 + \beta_1 x_i$$

where the observed values $Y_i \sim$ Poisson with $\lambda = \lambda_i$ for a given x_i.

Poisson model conditions:

1. **Poisson Response** The response variable is a count per unit of time or space, described by a Poisson distribution.
2. **Independence** The observations must be independent of one another.
3. **Mean=Variance** By definition, the mean of a Poisson random variable must be equal to its variance.
4. **Linearity** The log of the mean rate, log(λ), must be a linear function of x.

Poisson regression conditions: A graphical look



(a) Linear Regression

(b) Poisson Regression

Figure 2: Comparison of regression models.

Pause to Ponder

With your neighbor(s), compare the Poisson regression conditions to the usual LINE conditions in linear regression. List similarities and differences. What implications might the differences have for modeling and checking conditions?

Differences with linear regression (LLSR)

1. For each level of X, the responses follow a Poisson distribution (Condition 1). For Poisson regression, small values of λ are associated with a distribution that is noticeably skewed with lots of small values and only a few larger ones. As λ increases the distribution of the responses begins to look more and more like a normal distribution.
2. In the LLSR model, the variation in Y at each level of X, σ^2, is the same. For Poisson regression the responses at each level of X become more variable with increasing means, where variance=mean (Condition 3).
3. In the case of LLSR, the mean responses for each level of X, $\mu_{Y|X}$, fall on a line. In the case of the Poisson model, the mean values of Y at each level of X, $\lambda_{Y|X}$, fall on a curve, not a line, although the logs of the means should follow a line (Condition 4).

Sheena Easton and Game Theory

Sheena Easton describes the following scenario for her baby:

1. Takes the morning train
2. Works from nine 'til five
3. Takes another train home again
4. Finds Sheena Easton waiting for him

A Total Conflict Game Between Sheena Easton and Her Baby

	Stays Home	Goes to Work
Baby Home Again	-100, **100**	**100**, 0
Baby Stays at Work	**50**, 0	-100, **100**

Sheena Easton and her baby are playing a **zero-sum (total conflict) game**.

- Akin to Holmes-Moriarty game (see: von Neumann and Morgenstern)
- Solution: **mixed strategy**

Rick Astley's Re-election Platform

Rick Astley's campaign promises:

- Never gonna give you up.
- Never gonna let you down.
- Never gonna run around and desert you.
- Never gonna make you cry.
- Never gonna say goodbye.
- Never gonna tell a lie and hurt you.

Are these promises (if credible) sufficient to secure re-election?

Rick Astley and Median Voter Theorem

Whereas these pledges conform to the preferences of the **median voter**, we expect Congressman Astley to secure re-election.

Caribbean Queen and Operation Urgent Fury

Billy Ocean released "Caribbean Queen" in 1984.

- Emphasized sharing the same dream
- Hearts beating as one

"Caribbean Queen" is about the poor execution of Operation Urgent Fury.

- Echoed JCS chairman David Jones' frustrations with military establishment.

Billy Ocean is advocating for what became the Goldwater-Nichols Act.

- Wanted to take advantage of **economies of scale**, resolve **coordination problems** in U.S. military.

The Good Day Hypothesis

We know the following about Ice Cube's day.

1. The Lakers beat the Supersonics.
2. No helicopter looked for a murder.
3. Consumed Fatburger at 2 a.m.
4. Goodyear blimp: "Ice Cube's a pimp."

The Good Day Hypothesis

This leads to two different hypotheses:

- H_0: Ice Cube's day is statistically indistinguishable from a typical day.
- H_1: Ice Cube is having a good (i.e. greater than average) day.

These hypotheses are tested using archival data of Ice Cube's life.