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.

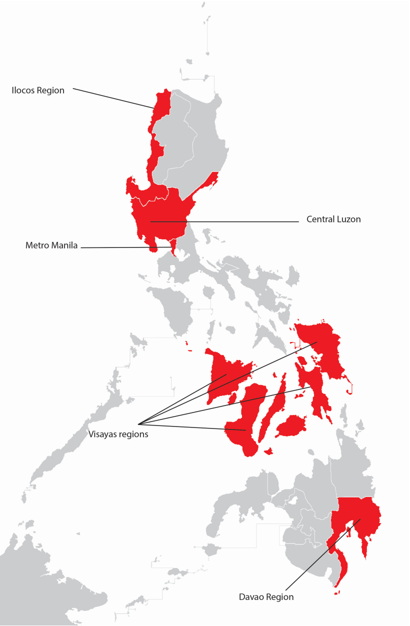
First case study: Philippine households

- 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:

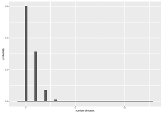
- `location` = region (Central Luzon, Davao, Ilocos, Metro Manila, or Visayas)
- `age` = the age of the head of household
- `total` = the number of people in the household other than the head
- `numLT5` = the number in the household under 5 years of age
- `roof` = the type of roof (stronger material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



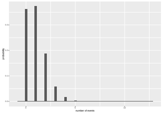
Poisson distribution

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

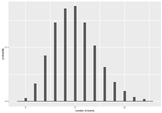
Note that both $E(Y_i) = \lambda_i$ and $Var(Y_i) = \lambda_i$.



(a) $\lambda = 0.5$



(b) $\lambda = 1$



(c) $\lambda = 5$

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

Poisson regression model

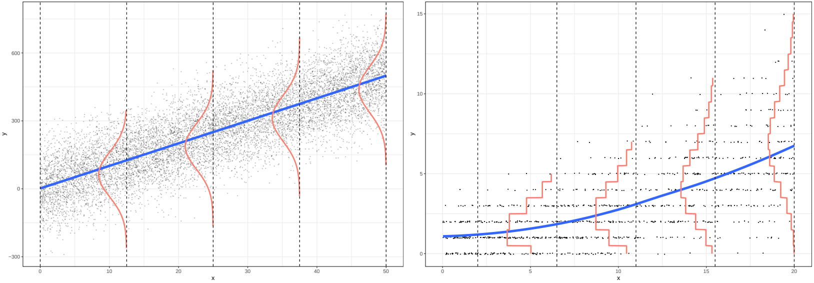
$$log(\lambda_i) = \beta_0 + \beta_1 x_i$$

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.

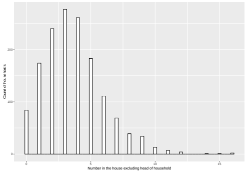
Pause to Ponder

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?

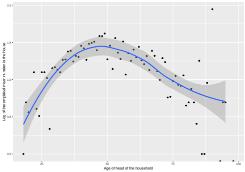
Differences with linear regression (LLSR)

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).

Exploratory data analysis



(a) Distribution of household size across all 5 Philippine regions.



(b) The log of the mean household sizes by age of the head of household, with loess smoother.

Figure 3: EDA: selected plots

Initial model

$$log(\hat{\lambda}) = 1.55 - 0.0047\text{age}$$

```
##                 Estimate    Std. Error   z value      Pr(>|z|)
## (Intercept)  1.549942225 0.0502754106 30.829032 1.070156e-208
## age         -0.004705881 0.0009363388 -5.025832  5.012548e-07
```

How can the coefficient estimates be interpreted?

Interpreting model coefficients

If your students have interpreted coefficients with a transformed response in linear regression, or log odds in logistic regression, this is similar.

Consider how the estimated mean number in the house, λ, changes as the age of the household head increases by an additional year.

$$\begin{aligned}
log(\lambda_X) &= \beta_0 + \beta_1 X \\
log(\lambda_{X+1}) &= \beta_0 + \beta_1 (X+1) \\
log(\lambda_{X+1}) - log(\lambda_X) &= \beta_1 \\
log\left(\frac{\lambda_{X+1}}{\lambda_X}\right) &= \beta_1 \\
\frac{\lambda_{X+1}}{\lambda_X} &= e^{\beta_1}
\end{aligned} \tag{1}$$

These results suggest that by exponentiating the coefficient on age we obtain the *multiplicative* factor by which the mean count changes.

Interpreting model coefficients (continued)

In this case, the mean number in the house changes by a factor of $e^{-0.0047} = 0.995$ or decreases by 0.5% (since $1 - .995 = .005$) with each additional year older the household head is;

Or, we predict a 0.47% *increase* in mean household size for a 1-year *decrease* in age of the household head (since $1/.995 = 1.0047$).

The quantity on the left-hand side of Equation 1 is referred to as a **rate ratio** or **relative risk**, and it represents a percent change in the response for a unit change in X.

Multiple Poisson regression

Most significant ideas from multiple linear regression are reinforced by Poisson regression:

- hypothesis testing
- confidence intervals
- indicator variables
- categorical variables (reference level / Tukey HSD)
- squared terms
- interaction terms
- control for / adjusting for / holding constant
- checking violations of model conditions
- maximum likelihood estimates (equals least squares in linear regr)

Potential final model

```r
modela2L <- glm(total ~ age + age2 + location, family = poisson, data = fHH1)
```

```
##                           Estimate    Std. Error     z value     Pr(>|z|)
## (Intercept)           -0.3843337714 1.820919e-01  -2.1106581 3.480171e-02
## age                    0.0703628330 6.905067e-03  10.1900292 2.196983e-24
## age2                  -0.0007025856 6.420019e-05 -10.9436677 7.125764e-28
## locationMetroManila    0.0544800704 4.720116e-02   1.1542104 2.484139e-01
## locationDavaoRegion   -0.0193872310 5.378273e-02  -0.3604732 7.184933e-01
## locationVisayas        0.1121091959 4.174960e-02   2.6852758 7.246998e-03
## locationIlocosRegion   0.0609819668 5.265981e-02   1.1580362 2.468493e-01
```

For example, $\hat{\beta}_6 = -0.0194$ indicates that, after controlling for the age of the head of household, the log mean household size is 0.0194 lower for households in the Davao Region than for households in the reference location of Central Luzon.

In more interpretable terms, mean household size is $e^{-0.0194} = 0.98$ times "higher" (i.e., 2% lower) in the Davao Region than in Central Luzon, when holding age constant.

Maximum estimated additional number in the house occurs when the head of the household is around 50 years old, after adjusting for location.

Potential final model (continued)

To test for the effect of location, use a drop-in-deviance test (analogous to an extra-sum-of-squares F test in linear regression):

```
##   ResidDF ResidDev Deviance Df       pval
## 1    1497 2200.944       NA NA         NA
## 2    1493 2187.800 13.14369  4 0.01059463
```

Adding the four terms corresponding to location to the quadratic model with age produces a statistically significant improvement ($\chi^2 = 13.144, df = 4, p = 0.0106$), so there is significant evidence that mean household size differs by location, after controlling for age of the head of household.

Lack of fit!

When a model is true, we can expect the residual deviance to be distributed as a χ^2 random variable with degrees of freedom equal to the model's residual degrees of freedom.

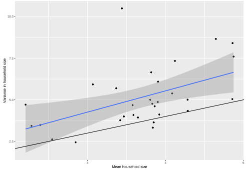
Our final model has a residual deviance of 2187.8 with 1493 df. The probability of observing a deviance this large if the model fits is essentially 0, saying that there is significant evidence of lack-of-fit.

There are several reasons why **lack-of-fit** may be observed:

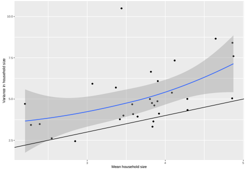
- We may be missing important covariates or interactions.
- There may be extreme observations that cause the deviance to be larger than expected.
- There may be a problem with the Poisson model. In particular, the Poisson model has only a single parameter, λ, for each combination of the levels of the predictors which must describe both the mean and the variance.

Overdispersion

Often in Poisson models the variances in the response are larger than the corresponding means at different levels of the predictors. The response is then considered to be **overdispersed**



(a) Linear fit



(b) Loess smoother

Figure 4: Mean and variance of predicted household sizes.

Quasi-Poisson models

Without adjusting for overdispersion, we use incorrect, artificially small standard errors leading to artificially small p-values for model coefficients.

The simplest way to take overdispersion into account is to use an estimated dispersion factor to inflate standard errors.

$\hat{\phi} = \frac{\sum(\text{Pearson residuals})^2}{n-p}$ where p is the number of model parameters.

$$SE_Q(\hat{\beta}) = \sqrt{\hat{\phi}} * SE(\hat{\beta})$$

Quasi-Poisson models (continued)

```
##                         Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)         -0.3843337714 0.2166025174 -1.7743735 7.620514e-02
## age                  0.0703628330 0.0082137357  8.5664837 2.616622e-17
## age2                -0.0007025856 0.0000763676 -9.2000473 1.168513e-19
## locationMetroManila  0.0544800704 0.0561468673  0.9703136 3.320474e-01
## locationDavaoRegion -0.0193872310 0.0639757901 -0.3030401 7.619015e-01
## locationVisayas      0.1121091959 0.0496621109  2.2574392 2.412461e-02
## locationIlocosRegion  0.0609819668 0.0626400545  0.9735299 3.304477e-01

##  Residual deviance =  2187.8  on  1493 df
##  Dispersion parameter =  1.414965
```

In the absence of overdispersion, we expect the dispersion parameter estimate to be 1.0. The estimated dispersion parameter here is larger than 1.0 (1.415).

For example, the standard error for the Visayas region term from a likelihood based approach is 0.0417, whereas the quasi-likelihood standard error is $\sqrt{1.415} * 0.0417$ or 0.0497. This term is still statistically significant at the 0.05 level under the quasi-Poisson model, but the evidence is not as strong (quasi-Poisson p-value of .024 vs. Poisson p-value of .007).

Quasi-Poisson models (continued)

We can take another look at our final model:

```
##   ResidDF ResidDev        F Df       pval
## 1    1497 2200.944       NA NA         NA
## 2    1493 2187.800 2.322264  4 0.05477322
```

Here, after adjusting for overdispersion, we find that there is *not* statistically significant evidence at the 0.05 level ($F = 2.32, p = .055$) that mean household size differs among regions after adjusting for age.

Negative Binomial models

A negative binomial model introduces another parameter in addition to λ, which gives the model more flexibility and, as opposed to the quasi-Poisson model, the negative binomial model assumes an explicit likelihood model.

Mathematically, you can think of the negative binomial model as a Poisson model where λ is also random, following a gamma distribution.

These results are very similar to the quasi-Poisson model in terms of estimated coefficients (which can change), standard errors, test statistics, and p-values.

Pause to Ponder

Check in with your neighbor(s). What questions do you have at this point? What can we clarify or discuss more fully?

Second case study: bald eagles

Every year in late December, since 1921, birdwatchers in the Hamilton area of Ontario, Canada, have counted and recorded all the birds they see or hear in a day.

The data was made available by the Bird Studies Canada website and distributed through the R for Data Science TidyTuesday project.

We are particularly interested in how the Bald Eagle population has changed over time.

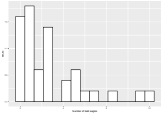
Bald eagle data

Each row of `bald_eagles.csv` contains information about bald eagles counts in Hamilton, Ontario, for one year. There are 37 rows covering 1981 through 2017. The variables include:

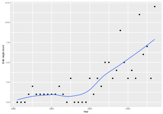
- `year` = year of data collection
- `count` = number of birds observed
- `hours` = total person-hours of observation period
- `count_per_hour` = count divided by hours
- `count_per_week` = count_per_hour multiplied by 168 hours per week

Credit: © Ron Niebrugge/wildnatureimages

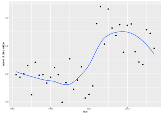
Exploratory data analysis



(a) Histogram of number of bald eagle sitings by year.



(b) Bald eagle counts vs. year



(c) Hours of observation vs. year

Figure 5: EDA: selected plots

Sampling effort

Poisson random variables are often used to represent counts (e.g., number of bald eagles) per unit of time or space (e.g., one year). But what if observation (sampling) effort (as measured by the number of `weeks` people observed birds) is changing over time?

We cannot directly compare the 2 eagles observed in 1985 to the 7 eagles observed in 2015 when there were only 143 person-hours (0.85 weeks) of observation in 1985 compared with 221 person-hours (1.32 weeks) in 2015.

We should examine time trends in the *rate* of bald eagles sightings; for example, we will calculate the bald eagle counts per week ($\frac{\text{number of bald eagles}}{\text{hours of observation}} \cdot (168 \text{ hours/week})$).

But Poisson models must be based on *counts*...

Offsets

One approach is to include a term on the right side of the model called an **offset**, which is the log of the weeks of observation.

Adjusting the yearly count by observation time is equivalent to adding $log(\text{weeks})$ to the right-hand side of the Poisson regression equation—essentially adding a predictor with a fixed coefficient of 1:

$$
\begin{aligned}
log(\frac{\lambda}{\text{weeks}}) &= \beta_0 + \beta_1(\text{type}) \\
log(\lambda) - log(\text{weeks}) &= \beta_0 + \beta_1(\text{type}) \\
log(\lambda) &= \beta_0 + \beta_1(\text{type}) + log(\text{weeks})
\end{aligned}
$$

Thus, modeling $log(\lambda)$ and adding an offset is equivalent to modeling rates, and coefficients can be interpreted in terms of rates.

Modeling results

We are interested primarily in trends over time in eagle sightings. We have no control variables other than sampling effort, so we simply fit a model with year (centered at 1981) and our offset.

```
model_eagles <- glm(count ~ year_1981, family = poisson,
                    offset = log(weeks), data = eagles)
```

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.79971104 0.31969178 -2.501506 1.236662e-02
## year_1981    0.07566483 0.01155782  6.546634 5.884818e-11

##  Residual deviance =  42.39031  on  35 df
##  Dispersion parameter =  1
```

Bald eagle counts are significantly increasing over time ($Z = 6.55$, $p < .001$), even after adjusting for observation time. The average eagle sighting rate per week has grown about 7.9% per year (since $e^{0.0757} = 1.0786$) in Hamilton, Ontario.

Adjustments for potential overdispersion using either quasi-Poisson or negative binomial regression provide minimal changes to model coefficients and tests.