Poisson Regression

Rebecca C. Steorts (slide adaption from Maria Tacket) and material from Chapter 4 of Roback and Legler text.

Computing set up

```r
library(tidyverse)
library(tidymodels)
library(knitr)
library(patchwork)
library(viridis)
library(gridExtra)
library(dplyr)

ggplot2::theme_set(ggplot2::theme_bw(base_size = 16))

colors <- tibble::tibble(green = "#B5BA72")
```

Topics

- Describe properties of the Poisson random variable
- Write the mathematical equation of the Poisson regression model
- Describe how the Poisson regression differs from least-squares regression
- Interpret the coefficients for the Poisson regression model
- Compare two Poisson regression models

Notes based on Section 4.4 - 4.5, and 4.9 of Roback and Legler (2021) unless noted otherwise.

Scenarios to use Poisson regression

- Does the number of employers conducting on-campus interviews during a year differ for public and private colleges?
- Does the daily number of asthma-related visits to an Emergency Room differ depending on air pollution indices?
- Does the number of paint defects per square foot of wall differ based on the years of experience of the painter?

Scenarios to use Poisson regression

- Does the **number of employers conducting on-campus interviews during a year** differ for public and private colleges?
- Does the **daily number of asthma-related visits to an Emergency Room** differ depending on air pollution indices?
- Does the number of paint defects per square foot of wall differ based on the years of experience of the painter?

Scenarios to use Poisson regression

- Does the **number of employers conducting on-campus interviews during a year** differ for public and private colleges?
- Does the **daily number of asthma-related visits to an Emergency Room** differ depending on air pollution indices?
- Does the number of paint defects per square foot of wall differ based on the years of experience of the painter?

Each response variable is a **count per a unit of time or space.**

Poisson distribution

Let Y be the number of events in a given unit of time or space. Then Y can be modeled using a **Poisson distribution**

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad y = 0, 1, 2, \ldots, \infty$$

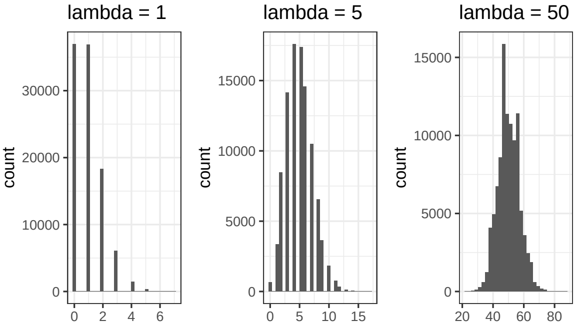
Poisson distribution

Let Y be the number of events in a given unit of time or space. Then Y can be modeled using a **Poisson distribution**

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!} \qquad y = 0, 1, 2, \ldots, \infty$$

- $E(Y) = Var(Y) = \lambda$
- The distribution is typically skewed right, particularly if λ is small
- The distribution becomes more symmetric as λ increases
  - If λ is sufficiently large, it can be approximated using a normal distribution (Click here for an example.)

Simulation



	Mean	Variance
lambda $=$ 1	0.99351	0.9902178
lambda $=$ 5	4.99367	4.9865798
lambda $=$ 50	49.99288	49.8962683

Earthquakes

The annual number of earthquakes registering at least 2.5 on the Richter Scale and having an epicenter within 40 miles of downtown Memphis follows a Poisson distribution with mean 6.5.[1]

[1] Example adapted from [Introduction to Probability Theory Example 28-2](https://online.stat.psu.edu/stat414/lesson/28/28.2).

Earthquakes

What is the probability there will be at 3 or fewer such earthquakes next year?

Earthquakes

What is the probability there will be at 3 or fewer such earthquakes next year?

$$P(Y \leq 3) = P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3)$$

Earthquakes

What is the probability there will be at 3 or fewer such earthquakes next year?

$$P(Y \leq 3) = P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3)$$

$$= \frac{e^{-6.5} 6.5^0}{0!} + \frac{e^{-6.5} 6.5^1}{1!} + \frac{e^{-6.5} 6.5^2}{2!} + \frac{e^{-6.5} 6.5^3}{3!}$$

$$= 0.112$$

Earthquakes

What is the probability there will be at 3 or fewer such earthquakes next year?

$$P(Y \leq 3) = P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3)$$

$$= \frac{e^{-6.5} 6.5^0}{0!} + \frac{e^{-6.5} 6.5^1}{1!} + \frac{e^{-6.5} 6.5^2}{2!} + \frac{e^{-6.5} 6.5^3}{3!}$$

$$= 0.112$$

```r
ppois(3, 6.5)
```

```
## [1] 0.1118496
```

Poisson regression

Poisson regression: Household size in the Philippines

The data fHH1.csv come from the 2015 Family Income and Expenditure Survey conducted by the Philippine Statistics Authority.

Goal: Understand the association between household size and various characteristics of the household

Response:

- `total`: Number of people in the household other than the head

Predictors:

- `location`: Where the house is located
- `age`: Age of the head of household
- `roof`: Type of roof on the residence (proxy for wealth)

Other variables:

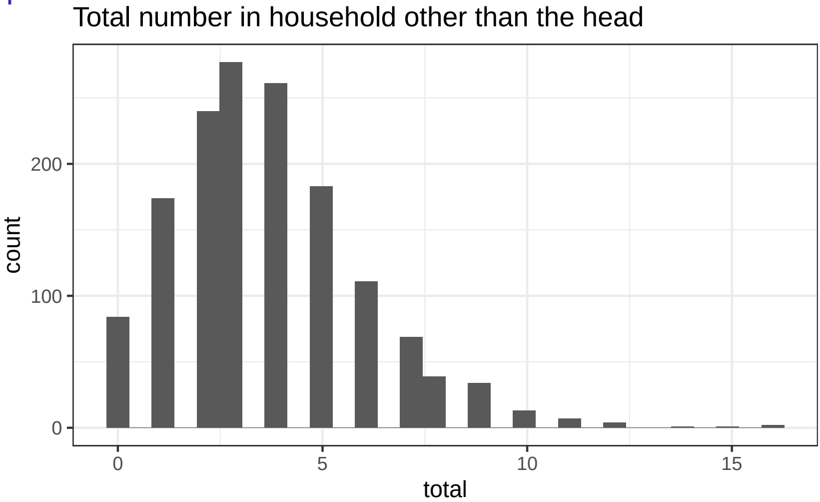
- `numLT5`: Number in the household under 5 years old

The data

```r
hh_data <- read_csv("data/fHH1.csv")
hh_data |> slice(1:5) |> kable()
```

location	age	total	numLT5	roof
CentralLuzon	65	0	0	Predominantly Strong Material
MetroManila	75	3	0	Predominantly Strong Material
DavaoRegion	54	4	0	Predominantly Strong Material
Visayas	49	3	0	Predominantly Strong Material
MetroManila	74	3	0	Predominantly Strong Material

Response variable



mean	var
3.685	5.534

Why the least-squares model doesn't work

The goal is to model λ, the expected number of people in the household (other than the head), as a function of the predictors (covariates)

Why the least-squares model doesn't work

The goal is to model λ, the expected number of people in the household (other than the head), as a function of the predictors (covariates)

We might be tempted to try a linear model

$$\lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Why the least-squares model doesn't work

The goal is to model λ, the expected number of people in the household (other than the head), as a function of the predictors (covariates)

We might be tempted to try a linear model

$$\lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

This model won't work because...

- It could produce negative values of λ for certain values of the predictors
- The equal variance assumption required to conduct inference for linear regression is violated.

Poisson regression model

Poisson regression model

If $Y_i \sim Poisson$ with $\lambda = \lambda_i$ for the given values $x_{i1}, \ldots, x_{ip}$, then

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Poisson regression model

If $Y_i \sim \textit{Poisson}$ with $\lambda = \lambda_i$ for the given values $x_{i1}, \ldots, x_{ip}$, then

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

- Each observation can have a different value of λ based on its value of the predictors $x_1, \ldots, x_p$
- λ determines the mean and variance, so we don't need to estimate a separate error term

Poisson vs. multiple linear regression

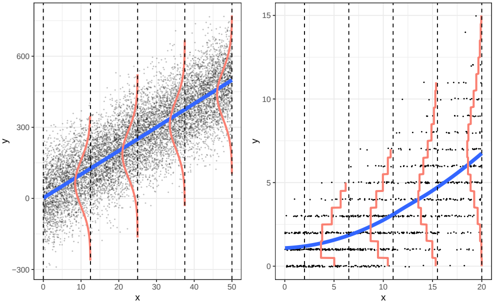


Figure 1: Regression models: Linear regression (left) and Poisson regression (right).

Figures recreated from BMLR Figure 4.1

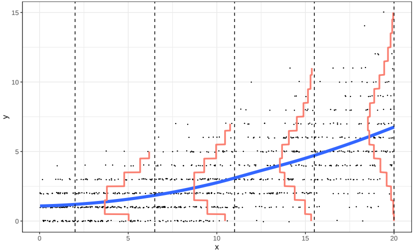
Assumptions for Poisson regression

Poisson response: The response variable is a count per unit of time or space, described by a Poisson distribution, at each level of the predictor(s)

Independence: The observations must be independent of one another

Mean = Variance: The mean must equal the variance

Linearity: The log of the mean rate, $\log(\lambda)$, must be a linear function of the predictor(s)



Model 1: Household vs. Age

```r
model1 <- glm(total ~ age,
              data = hh_data, family = poisson)

tidy(model1) |>
  kable(digits = 4)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.5499	0.0503	30.8290	0
age	-0.0047	0.0009	-5.0258	0

$$\log(\hat{\lambda}) = 1.5499 - 0.0047 \ age$$

Interpretation of coefficient estimates

Consider a comparison of two models – one for a given age (x) and another for age ($x + 1$).

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

Exponentiating the coefficient on age provides the multiplicative factor by which the mean count changes.

Interpretation of coefficient estimates

The mean household size is predicted to decrease by 0.47% (or multiply by a factor of $e^{-0.0047}$) for each year older the head of the household is.

Interpretation of coefficient estimates

1. The mean number in the house changes by a factor of $e^{-0.0047} = 0.995$ with each additional year older the household head is.
2. The mean number in the houses decreases by 0.5 percent with each additional year older the household head is. (Because 1 - 0.995 = 0.005)
3. We predict a 0.47 percent increase in mean household size for a 1-year decrease in age of the household head (because $1/0.995 = 1.0047$).
4. We predict a 0.47 percent decrease in mean household size for a 1-year increase in age of the household head (because $1/0.995 = 1.0047$).

Is the coefficient of age statistically significant?

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.5499	0.0503	30.8290	0	1.4512	1.6482
age	-0.0047	0.0009	-5.0258	0	-0.0065	-0.0029

Is the coefficient of age statistically significant?

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.5499	0.0503	30.8290	0	1.4512	1.6482
age	-0.0047	0.0009	-5.0258	0	-0.0065	-0.0029

1. $H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0$
2. $Z = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-0.0047 - 0}{0.0009} = -5.026$ (using exact values)
3. $P(|Z| > |-5.026|) = 5.01 \times 10^{-7} \approx 0.$
4. Yes, it is statistically significant.

What are plausible values for the coefficient of age?

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.5499	0.0503	30.8290	0	1.4512	1.6482
age	-0.0047	0.0009	-5.0258	0	-0.0065	-0.0029

95% confidence interval for the coefficient of age

$$\hat{\beta}_1 \pm z^* \times SE(\hat{\beta}_1)$$

where $z^* \sim N(0,1)$

$$-0.0047 \pm 1.96 \times 0.0009 = (-\mathbf{.0065}, -\mathbf{0.0029})$$

Interpret the interval in terms of the change in mean household size.

What are plausible values for the coefficient of age?

Interpret the interval $(-\mathbf{.0065}, -\mathbf{0.0029})$ in terms of the change in mean household size.

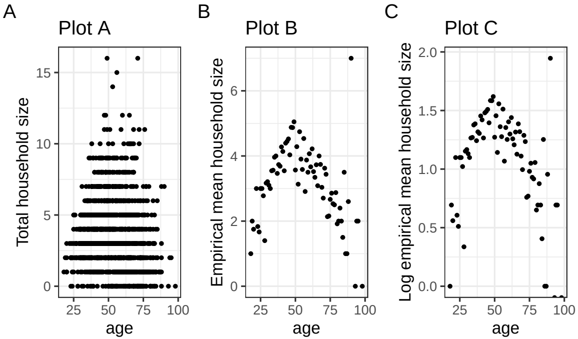
Recall: Exponentiating the endpoints yields a confidence interval for the relative risk; i.e., the percent change in household size for each additional year older.

Thus,

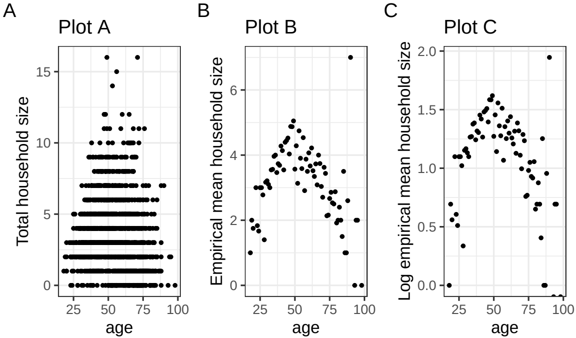
$$(e^{-0.0065}, e^{-0.0029}) = (0.993, 0.997).$$

suggests that we are 95% confident that the mean number in the house decreases between 0.7% and 0.3% for each additional year older the head of household is.

Which plot can best help us determine whether Model 1 is a good fit?



Which plot can best help us determine whether Model 1 is a good fit?



Solution: Plot C. Observe a curvi-linear relationship between age and the log of the mean household size, implying that adding a quadratic term should be considered.

Model 2: Add a quadratic effect for age

```
hh_data <- hh_data |>
  mutate(age2 = age*age)

model2 <- glm(total ~ age + age2,
              data = hh_data, family = poisson)
tidy(model2, conf.int = T) |>
  kable(digits = 4)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.3325	0.1788	-1.8594	0.063	-0.6863	0.0148
age	0.0709	0.0069	10.2877	0.000	0.0575	0.0845
age2	-0.0007	0.0001	-11.0578	0.000	-0.0008	-0.0006

Model 2: Add a quadratic effect for age

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.3325	0.1788	-1.8594	0.063	-0.6863	0.0148
age	0.0709	0.0069	10.2877	0.000	0.0575	0.0845
age2	-0.0007	0.0001	-11.0578	0.000	-0.0008	-0.0006

We can determine whether to keep age^2 in the model in two ways:

1. Use the p-value (or confidence interval) for the coefficient (since we are adding a single term to the model). This is known as a Wald-type statistic.
2. Conduct a drop-in-deviance test

Wald-type statistic

Observe that $Z = -11.058$ with p-value approximately 0.

This supports the alternative hypothesis that the quadratic term is statistically significant in the model.

Deviance

A **deviance** is a way to measure how the observed data differs (deviates) from the model predictions.

- It's a measure unexplained variability in the response variable (similar to SSE in linear regression)
- Lower deviance means the model is a better fit to the data

Deviance

A **deviance** is a way to measure how the observed data differs (deviates) from the model predictions.

- It's a measure unexplained variability in the response variable (similar to SSE in linear regression)
- Lower deviance means the model is a better fit to the data

We can calculate the "deviance residual" for each observation in the data (more on the formula later). Let (deviance residual$_i$ be the deviance residual for the i^{th} observation, then

$$\text{deviance} = \sum (\text{deviance residual})_i^2$$

Note: Deviance is also known as the "residual deviance"

Drop-in-Deviance Test

We can use a **drop-in-deviance test** to compare two models. To conduct the test

1. Compute the deviance for each model
2. Calculate the drop in deviance

drop-in-deviance = Deviance(reduced model) - Deviance(larger model)

3. Given the reduced model is the true model (H_0 true), then

$$\text{drop-in-deviance} \sim \chi^2_d$$

where d is the difference in degrees of freedom between the two models (i.e., the difference in number of terms)

Summary of the Drop-in-Deviance

- To use the drop-in-deviance test, the models must be nested
- This means the terms in the smaller model must appear in the larger model
- When the reduced (or smaller model) is true, the drop-in-deviance $\approx \chi^2_d$
- A large drop-in-deviance favors the larger model

Refer to Section 4.4.4 for more details.

Drop-in-deviance to compare Model1 and Model2

```r
anova(model1, model2, test = "Chisq") |>
  kable(digits = 3)
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1498	2337.089	NA	NA	NA
1497	2200.944	1	136.145	0

a. Write the null and alternative hypotheses.
b. What does the value 2337.089 tell you?
c. What does the value 1 tell you?
d. What is your conclusion?

Drop-in-deviance to compare Model1 and Model2

a.

$$\text{Null (reduced) Model} : \mathsf{log}(\lambda) = \beta_0 + \beta_1 \mathsf{age}$$

$$\text{Larger (full) Model} : \mathsf{log}(\lambda) = \beta_0 + \beta_1 \text{age} + \beta_2 \mathsf{age}^2$$

b. What does the value 2337.089 tell you?

The value 2337.1 (with 1498 df.) provides the residual deviance for the null model.

c. There is only 1 degree of freedom difference between the two models.

Drop-in-deviance to compare Model1 and Model2 (continued)

d. The drop-in-deviance is $2337.089 - 2200.94 = 136.15 \approx \chi^2_1$.

The p-value is 0, indicating there is statistically significant evidence that average household size decreases as age of the head of household increases. This provides significant support for rejecting the null hypothesis (in favor of the alternative) and including the quadratic term.

Add `location` to the model?

Suppose we want to add `location` to the model, so we compare the following models:

Model A: $\lambda_i = \beta_0 + \beta_1\ age_i + \beta_2\ age_i^2$

Model B: $\lambda_i =$
$\beta_0 + \beta_1\ age_i + \beta_2\ age_i^2 + \beta_3\ Loc1_i + \beta_4\ Loc2_i + \beta_5\ Loc3_i + \beta_6\ Loc4_i$

Which of the following are reliable ways to determine if `location` should be added to the model? (See Section 4.5, regarding comparison of linear versus Poisson models)

1. Drop-in-deviance test
2. Use the p-value for each coefficient
3. Likelihood ratio test
4. Nested F Test
5. BIC
6. AIC

Add `location` to the model?

- See Section 4.4.7 regarding adding location to the model. (Pages 109 – 110).

Supplementary Material

Let's consider the connection between the Drop in deviance test and LRT for Poisson regression.

Drop in deviance and LRT

In Poisson regression, the deviance is defined as -2 times the log-likelihood ratio of a fitted model relative to the saturated model.

The saturated model is the model that has one free parameter per observation, so it fits the data perfectly.

When comparing two nested models, the saturated model term cancels in the difference of deviances, leaving exactly the likelihood ratio test statistic.

Proof

Proof: Let $M_0 \subset M_1$ be two nested Poisson regression models. The deviance of a model M is defined as

$$D(M) = 2\left\{\ell\left(\hat{\lambda}^{\mathsf{sat}}\right) - \ell\left(\hat{\lambda}^{(M)}\right)\right\},$$

where the saturated model satisfies $\hat{\lambda}_i^{\mathsf{sat}} = y_i$.

Because both M_0 and M_1 are compared to the same saturated model,

$$\begin{aligned} D(M_0) - D(M_1) &= 2\left\{\ell\left(\hat{\lambda}^{(1)}\right) - \ell\left(\hat{\lambda}^{(0)}\right)\right\} \\ &= -2\log\frac{L\left(\hat{\lambda}^{(0)}\right)}{L\left(\hat{\lambda}^{(1)}\right)}, \end{aligned}$$

which is exactly the likelihood ratio test statistic for testing M_0 against M_1.

Since M_0 is nested in M_1, the statistic is asymptotically χ^2 with degrees of freedom equal to the number of additional parameters in M_1.

Formal treatment

Is the LRT a special case of the Drop-in-deviance test for GLMs?

- Yes! Please see http://users.stat.umn.edu/~helwig/notes/generalized-linear-models.html#example-2-poisson-regression

Interpretation of GLM regression coefficients

A generalized linear model (GLM) is defined by

$$g(\mathbb{E}[Y \mid X]) = \eta = X\beta,$$

where $g(\cdot)$ is the link function and η is the linear predictor.

GLM coefficients describe how predictors affect the ***mean of the response*** *through the link function.*

Coefficients describe how predictors change the linear predictor η and therefore how they change the mean of Y through the inverse link.

Linear versus Poisson regression

- **Identity link (linear regression):**

$$\mathbb{E}[Y \mid X] = X\beta,$$

so β_j is the additive change in the mean/median of Y.

- **Log link (Poisson):**

$$\log \mathbb{E}[Y \mid X] = X\beta,$$

so

$$\begin{aligned} \mathbb{E}[Y \mid X] &= \exp(X\beta), \\ \beta_j &\text{ is the additive change in the log-mean} \\ e^{\beta_j} &= \text{multiplicative change in the mean.} \end{aligned}$$

Looking ahead

- For next time - Chapter 4 - Poisson Regression
  - Sections 4.6, 4.10

References

Roback, Paul, and Julie Legler. 2021. *Beyond multiple linear regression: applied generalized linear models and multilevel models in R*. CRC Press.