Poisson Regression

Part II

Rebecca C. Steorts (slide adaption from Maria Tacket) and material from Chapter 4 of Roback and Legler (2021).

Announcements

1. Homework 3 due Friday at 5 PM.
2. Quiz 1 will be released on Gradescope later this week along with a deadline. Coverage will be Chapters 1 – 3 of BMLR. This quiz is open note/open book and take home. If you use other resources, cite them and instructions will be on the quiz.
3. This quiz is to be completed individually. You will have one opportunity to submit it, so be careful regarding this.
4. Any questions should be sent through direct message in Canvas (or email). DO NOT make any group postings about quiz information!

Topics

- Define and calculate residuals for the Poisson regression model
- Use Goodness-of-fit to assess model fit
- Identify overdispersion
- Apply modeling approaches to deal with overdispersion
  - Quasi-Poisson
  - Negative binomial

Notes based on Sections 4.4 and 4.9 of Roback and Legler (2021) unless noted otherwise.

The data: 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")
```

Poisson regression model

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

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

Poisson regression model

If $Y_i \sim \mathit{Poisson}$ 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

Model 1: Household vs. Age

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

tidy(hh_age) |>
  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

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

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

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

Determined Model 2 is a better fit than Model 1 based on the drop-in-deviance test.

Add *location* to model?

```
model3 <- glm(total ~ age + age2 + location,
              data = hh_data, family = poisson)
```

Use a drop-in-deviance test to determine if Model 2 or Model 3 (with location) is a better fit for the data.

```
anova(model2, model3, test = "Chisq") %>%
  kable(digits = 3)
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1497	2200.944	NA	NA	NA
1493	2187.800	4	13.144	0.011

The p-value is small ($0.01 < 0.05$), so we conclude that Model 3 is a better fit for the data.

Selected model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.384	0.182	-2.111	0.035	-0.744	-0.031
age	0.070	0.007	10.190	0.000	0.057	0.084
age2	-0.001	0.000	-10.944	0.000	-0.001	-0.001
locationDavaoRegion	-0.019	0.054	-0.360	0.718	-0.125	0.086
locationIlocosRegion	0.061	0.053	1.158	0.247	-0.042	0.164
locationMetroManila	0.054	0.047	1.154	0.248	-0.038	0.147
locationVisayas	0.112	0.042	2.685	0.007	0.031	0.195

Selected model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.384	0.182	-2.111	0.035	-0.744	-0.031
age	0.070	0.007	10.190	0.000	0.057	0.084
age2	-0.001	0.000	-10.944	0.000	-0.001	-0.001
locationDavaoRegion	-0.019	0.054	-0.360	0.718	-0.125	0.086
locationIlocosRegion	0.061	0.053	1.158	0.247	-0.042	0.164
locationMetroManila	0.054	0.047	1.154	0.248	-0.038	0.147
locationVisayas	0.112	0.042	2.685	0.007	0.031	0.195

Does this model sufficiently explain the variability in the mean household size?

Goodness of Fit

- Pearson residuals
- Deviance residuals

Pearson residuals

We can calculate two types of residuals for Poisson regression: Pearson residuals and deviance residuals

$$\text{Pearson residual}_i = \frac{\text{observed} - \text{predicted}}{\text{std. error}} = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

Pearson residuals

We can calculate two types of residuals for Poisson regression: Pearson residuals and deviance residuals

$$\text{Pearson residual}_i = \frac{\text{observed} - \text{predicted}}{\text{std. error}} = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

- Similar interpretation as standardized residuals from linear regression
- Expect most to fall between -2 and 2
- Used to calculate overdispersion parameter (more on this soon)

Deviance residuals

- The **deviance residual** describes how the observed data deviates from the fitted model

$$\text{deviance residual}_i = \text{sign}(Y_i - \hat{\lambda}_i)\sqrt{2\left[Y_i \log\left(\frac{Y_i}{\hat{\lambda}_i}\right) - (Y_i - \hat{\lambda}_i)\right]}$$

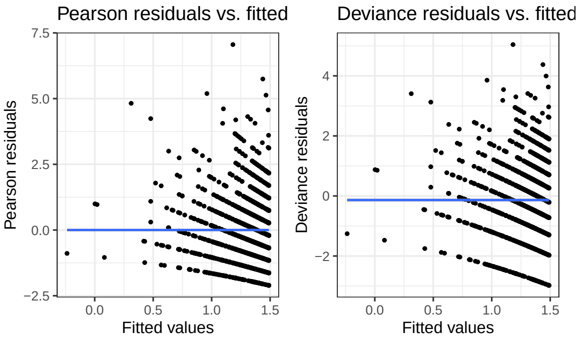
where

$$\text{sign}(Y_i - \hat{\lambda}_i) = \begin{cases} 1 & \text{if } (Y_i - \hat{\lambda}_i) > 0 \\ -1 & \text{if } (Y_i - \hat{\lambda}_i) < 0 \\ 0 & \text{if } (Y_i - \hat{\lambda}_i) = 0 \end{cases}$$

- Good fitting models $\Rightarrow$ small deviances

Selected model: Residual plots

```
model3_aug_pearson <-
  augment(model3, type.residuals = "pearson")
model3_aug_deviance <-
  augment(model3, type.residuals = "deviance")
```



A “good fit” in a residual plot appears as random, evenly spread points around the horizontal axis (zero) without a discernible pattern.

Goodness-of-fit

- **Goal**: Use the (residual) deviance to assess how much the predicted values differ from the observed values.

$$\text{deviance} = \sum_{i=1}^{n} (\text{deviance residual})_i^2$$

- When a model is true, we expect

$$\text{deviance} \sim \chi^2_{df}$$

where df is the model's residual degrees of freedom

Goodness-of-fit

- **Goal**: Use the (residual) deviance to assess how much the predicted values differ from the observed values.

$$\text{deviance} = \sum_{i=1}^{n} (\text{deviance residual})_i^2$$

- When a model is true, we expect

$$\text{deviance} \sim \chi^2_{df}$$

where df is the model's residual degrees of freedom

- **Question to answer**: What is the probability of observing a deviance larger than the one we've observed, given this model sufficiently fits the data?

$$P(\chi^2_{df} > \text{ deviance})$$

Goodness-of-fit calculations

```r
model3$deviance
```

```
## [1] 2187.8
```

```r
model3$df.residual
```

```
## [1] 1493
```

Goodness-of-fit calculations

```r
model3$deviance
```

```
## [1] 2187.8
```

```r
model3$df.residual
```

```
## [1] 1493
```

```r
pchisq(model3$deviance, model3$df.residual,
       lower.tail = FALSE)
```

```
## [1] 3.153732e-29
```

Goodness-of-fit calculations

```
model3$deviance
```

```
## [1] 2187.8
```

```
model3$df.residual
```

```
## [1] 1493
```

```
pchisq(model3$deviance, model3$df.residual,
       lower.tail = FALSE)
```

```
## [1] 3.153732e-29
```

The probability of observing a deviance greater than 2187.8 is ≈ 0, so there is significant evidence of **lack-of-fit**.

Lack-of-fit

There are a few potential reasons for observing lack-of-fit:

- Missing important interactions or higher-order terms
- Missing important variables (perhaps this means a more comprehensive data set is required)
- There could be extreme observations causing the deviance to be larger than expected (assess based on the residual plots)
- There could be a problem with the Poisson model
  - Only one parameter λ to describe mean and variance
  - May need more flexibility in the model to handle **overdispersion**

Overdispersion

- The Poisson model only has one parameter, λ, which must describe both the mean and the variance
- Often, the variance can appear larger than the corresponding means.
- In this case, the response is more variable than assumed by the Poisson model, and the response is said to be overdispersed.

Overdispersion

Overdispersion: There is more variability in the response than what is implied by the Poisson model

Overall

mean	var
3.685	5.534

by Location

location	mean	var
CentralLuzon	3.402	4.152
DavaoRegion	3.390	4.723
IlocosRegion	3.586	5.402
MetroManila	3.707	4.863
Visayas	3.902	6.602

Why overdispersion matters

If there is overdispersion, then there is more variation in the response than what's implied by a Poisson model. This means

The standard errors of the model coefficients are artificially small

$\Rightarrow$ The p-values are artificially small

$\Rightarrow$ Could lead to models that are more complex than what is needed

Why overdispersion matters

If there is overdispersion, then there is more variation in the response than what's implied by a Poisson model. This means

The standard errors of the model coefficients are artificially small

$\Rightarrow$ The p-values are artificially small

$\Rightarrow$ Could lead to models that are more complex than what is needed

We can take overdispersion into account by

- inflating standard errors by multiplying them by a dispersion factor
- using a negative-binomial regression model

Quasi-Poisson

Dispersion parameter

The **dispersion parameter** is represented by ϕ

$$\hat{\phi} = \frac{\sum_{i=1}^{n}(\text{Pearson residuals})^2}{n - p}$$

where p is the number of terms in the model (including the intercept)

Dispersion parameter

The **dispersion parameter** is represented by ϕ

$$\hat{\phi} = \frac{\sum_{i=1}^{n} (\text{Pearson residuals})^2}{n - p}$$

where p is the number of terms in the model (including the intercept)

- If there is no overdispersion $\hat{\phi} = 1$
- If there is overdispersion $\hat{\phi} > 1$

Accounting for dispersion

- We inflate the standard errors of the coefficient by multiplying the variance by $\hat{\phi}$

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

- "Q" stands for **quasi-Poisson**, since this is an ad-hoc solution
- The process for model building and model comparison is called **quasilikelihood** (similar to likelihood without exact underlying distributions)

Quasi-Poisson model

```
hh_age_loc_q <- glm(total ~ age + I(age^2) + location,
                    data = hh_data, family = quasipoisson)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.3843	0.2166	-1.7744	0.0762	-0.8134	0.0358
age	0.0704	0.0082	8.5665	0.0000	0.0544	0.0866
I(age^2)	-0.0007	0.0001	-9.2000	0.0000	-0.0009	-0.0006
locationDavaoRegion	-0.0194	0.0640	-0.3030	0.7619	-0.1451	0.1058
locationIlocosRegion	0.0610	0.0626	0.9735	0.3304	-0.0620	0.1837
locationMetroManila	0.0545	0.0561	0.9703	0.3320	-0.0552	0.1649
locationVisayas	0.1121	0.0497	2.2574	0.0241	0.0156	0.2103

Poisson vs. Quasi-Poisson models

	Poisson		Quasi-Poisson	
term	estimate	std.error	estimate	std.error
(Intercept)	-0.3843	0.1821	-0.3843	0.2166
age	0.0704	0.0069	0.0704	0.0082
age2	-0.0007	0.0001	-0.0007	0.0001
locationDavaoRegion	-0.0194	0.0538	-0.0194	0.0640
locationIlocosRegion	0.0610	0.0527	0.0610	0.0626
locationMetroManila	0.0545	0.0472	0.0545	0.0561
locationVisayas	0.1121	0.0417	0.1121	0.0497

Quasi-Poisson: Inference for coefficients

term	estimate	std.error
(Intercept)	-0.3843	0.2166
age	0.0704	0.0082
I(age^2)	-0.0007	0.0001
locationDavaoRegion	-0.0194	0.0640
locationIlocosRegion	0.0610	0.0626
locationMetroManila	0.0545	0.0561
locationVisayas	0.1121	0.0497

Test statistic

$$t = \frac{\hat{\beta} - 0}{SE_Q(\hat{\beta})} \sim t_{n-p}$$

Quasi-Poisson model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.3843	0.2166	-1.7744	0.0762	-0.8134	0.0358
age	0.0704	0.0082	8.5665	0.0000	0.0544	0.0866
I(age^2)	-0.0007	0.0001	-9.2000	0.0000	-0.0009	-0.0006
locationDavaoRegion	-0.0194	0.0640	-0.3030	0.7619	-0.1451	0.1058
locationIlocosRegion	0.0610	0.0626	0.9735	0.3304	-0.0620	0.1837
locationMetroManila	0.0545	0.0561	0.9703	0.3320	-0.0552	0.1649
locationVisayas	0.1121	0.0497	2.2574	0.0241	0.0156	0.2103

Negative binomial regression model

Negative binomial regression model

Another approach to handle overdispersion is to use a **negative binomial regression model**

- This has more flexibility than the quasi-Poisson model, because there is a new parameter in addition to λ

Negative binomial regression model

Another approach to handle overdispersion is to use a **negative binomial regression model**

- This has more flexibility than the quasi-Poisson model, because there is a new parameter in addition to λ

Let Y be a **negative binomial random variable**,
$Y \sim NegBinom(r, p)$, then

$$P(Y = y_i) = \binom{y_i + r - 1}{r - 1}(1-p)^{y_i} p^r \qquad y_i = 0, 1, 2, \ldots, \infty$$

$$E(Y) = \frac{r(1-p)}{p} \qquad SD(Y) = \sqrt{\frac{r(1-p)}{p^2}}$$

Negative binomial regression model

- **Main idea**: Generate a λ for each observation (household) and generate a count using the Poisson random variable with parameter λ

$$\text{If} \quad Y|\lambda \sim \mathit{Poisson}(\lambda)$$
$$\text{and } \lambda \sim \mathit{Gamma}\left(r, \frac{1-p}{p}\right)$$
$$\text{then } Y \sim \mathit{NegBinom}(r, p)$$

Negative binomial regression model

- **Main idea**: Generate a λ for each observation (household) and generate a count using the Poisson random variable with parameter λ
  - Makes the counts more dispersed than with a single parameter

$$
\begin{aligned}
&\text{If} \quad Y|\lambda \sim \mathit{Poisson}(\lambda) \\
&\text{and } \lambda \sim \mathit{Gamma}\left(r, \frac{1-p}{p}\right) \\
&\text{then } Y \sim \mathit{NegBinom}(r, p)
\end{aligned}
$$

Negative binomial regression model

- **Main idea**: Generate a λ for each observation (household) and generate a count using the Poisson random variable with parameter λ
  - Makes the counts more dispersed than with a single parameter
- Think of it as a Poisson model such that λ is also random

$$\text{If} \quad Y|\lambda \sim \mathit{Poisson}(\lambda)$$
$$\text{and } \lambda \sim \mathit{Gamma}\left(r, \frac{1-p}{p}\right)$$
$$\text{then } Y \sim \mathit{NegBinom}(r, p)$$

Negative binomial regression in R

Use the `glm.nb` function in the **MASS** R package.

The **MASS** package has a `select` function that conflicts with the `select` function in **dplyr**. You can avoid this by (1) always loading **tidyverse** after **MASS**, or (2) use `MASS::glm.nb` instead of loading the package.

Negative binomial regression in R

```
hh_age_loc_nb <- MASS::glm.nb(total ~ age + I(age^2) +
                                 location, data = hh_data)
tidy(hh_age_loc_nb) |>
  kable(digits = 4)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.3753	0.2076	-1.8081	0.0706
age	0.0699	0.0079	8.8981	0.0000
I(age^2)	-0.0007	0.0001	-9.5756	0.0000
locationDavaoRegion	-0.0219	0.0625	-0.3501	0.7262
locationIlocosRegion	0.0577	0.0615	0.9391	0.3477
locationMetroManila	0.0562	0.0551	1.0213	0.3071
locationVisayas	0.1104	0.0487	2.2654	0.0235

Negative binomial vs. Quasi-Poisson

	Quasi-Poisson		Negative binomial	
term	estimate	std.error	estimate	std.error
(Intercept)	-0.3843	0.2166	-0.3753	0.2076
age	0.0704	0.0082	0.0699	0.0079
I(age^2)	-0.0007	0.0001	-0.0007	0.0001
locationDavaoRegion	-0.0194	0.0640	-0.0219	0.0625
locationIlocosRegion	0.0610	0.0626	0.0577	0.0615
locationMetroManila	0.0545	0.0561	0.0562	0.0551
locationVisayas	0.1121	0.0497	0.1104	0.0487

Exercise

Suppose

$$
\begin{aligned}
Y|\lambda &\sim \text{Poisson}(\lambda) && (1)\\
\lambda &\sim \text{Gamma}\left(r, \frac{p}{1-p}\right). && (2)\\
& && (3)
\end{aligned}
$$

It follows that

$$Y \sim \text{NegBinom}(r, p).$$

Exercise

We are given that:

$$Y \mid \lambda \sim \text{Poisson}(\lambda),$$

which means that the conditional probability mass function (PMF) of Y, given λ, is

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

Additionally, we are given that:

$$\lambda \sim \text{Gamma}\left(r, \frac{p}{1-p}\right).$$

Thus,

$$p(\lambda) = \frac{1}{\Gamma(r)(\frac{p}{1-p})^r} \lambda^{r-1} e^{-\lambda(\frac{1-p}{p})}.$$

Solution

The marginal distribution of Y is found by integrating out λ from the joint distribution of Y and λ. That is:

$$P(Y = y) = \int_0^\infty P(Y = y, \lambda) d\lambda \tag{4}$$

$$P(Y = y) = \int_0^\infty P(Y = y \mid \lambda) f(\lambda) d\lambda. \tag{5}$$

This follows from the fact that

$$P(A, B) = P(A \mid B) P(B).$$

Solution

It follows that

$$p(y) = \int_0^\infty P(Y = y \mid \lambda) f(\lambda) d\lambda \tag{6}$$

$$= \int_0^\infty \frac{\lambda^y e^{-\lambda}}{y!} \times \frac{1}{\Gamma(r)(\frac{p}{1-p})^r} \lambda^{r-1} e^{-\lambda(\frac{1-p}{p})} d\lambda \tag{7}$$

$$= \frac{p(1-p)^{-r}}{\Gamma(r) y!} \int_0^\infty \lambda^{y+r-1} e^{-\lambda(\frac{1}{p})} d\lambda. \tag{8}$$

Solution

Observe that

$$\int_0^\infty \lambda^{y+r-1} e^{-\lambda(\frac{1}{p})} d\lambda$$

is the kernel (part without the normalizing constants) of a Gamma distribution with parameters $a = 1/p$ and $b = y + r$.

This implies that

$$\int_0^\infty \lambda^{y+r-1} e^{-\lambda(\frac{1}{p})} d\lambda = \frac{\Gamma(y+r)}{(1/p)^{y+r}} = \Gamma(y+r) p^{y+r}.$$

Solution

Fact: $\Gamma(c) = (c-1)!$ when c is an integer.

Using the Gamma kernel fact, it follows that

$$
\begin{aligned}
p(y) &= \frac{p(1-p)^{-r}}{\Gamma(r)y!} \int_0^\infty \lambda^{y+r-1} e^{-\lambda(\frac{1}{p})} d\lambda && (9) \\
&= \frac{p(1-p)^{-r}}{\Gamma(r)y!} \times \Gamma(y+r) p^{y+r} && (10) \\
&= \frac{\Gamma(y+r)}{\Gamma(r)y!} p^y (1-p)^r && (11) \\
&= \frac{(r+y-1)!}{(r-1)!y!} p^y (1-p)^r, && (12)
\end{aligned}
$$

which is a Negative Binomial distribution(r, p), where r is the number of successes until the experiment is stopped and p is the success probability.

Additional resources

You may find this post helpful, which outlines different parameterizations of the Gamma distribution.

https://timothy-barry.github.io/posts/2020-06-16-gamma-poisson-nb/

If you have taken STA 360 (or will take it), this derivation is known as calculating the marginal distribution.

References

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