# Poisson Regression

## Goodness-of-fit & overdispersion

Prof. Maria Tackett

01.31.22

Click for PDF of slides

# Announcements

- Reading: BMLR - Chapter 4 Poisson regression
- Mini-Project 01:
  - Draft **due Wed, Feb 02 at 12pm (noon)** in GitHub repo
  - Peer review in class Wednesday
  - Final write up and presentations **Wed, Feb 09 at 3:30pm**
- Thursday's class: Offsets and Zero-inflated Poisson model (ZIP)
- HW 02 **due Mon, Feb 07 at 11:59pm**
  - Released later today (will announce on GitHub Discussions)

# HW 01

- Exercise 2: The independence assumption is on the residuals (observations) not the predictors.
    - Multicollinearity is the correlation between predictors
- Read feedback carefully in Gradescope. Ask questions about feedback during office hours.

# Learning goals

- 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

# Recap

# 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)

# Poisson regression model

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

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

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

# Model 1: Household vs. Age

```
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 |

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

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 the 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.

# Model 3

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | -0.3843 | 0.1821 | -2.1107 | 0.0348 | -0.7444 | -0.0306 |
| age | 0.0704 | 0.0069 | 10.1900 | 0.0000 | 0.0569 | 0.0840 |
| age2 | -0.0007 | 0.0001 | -10.9437 | 0.0000 | -0.0008 | -0.0006 |
| locationDavaoRegion | -0.0194 | 0.0538 | -0.3605 | 0.7185 | -0.1250 | 0.0859 |
| locationIlocosRegion | 0.0610 | 0.0527 | 1.1580 | 0.2468 | -0.0423 | 0.1641 |
| locationMetroManila | 0.0545 | 0.0472 | 1.1542 | 0.2484 | -0.0378 | 0.1473 |
| locationVisayas | 0.1121 | 0.0417 | 2.6853 | 0.0072 | 0.0308 | 0.1945 |

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

# Goodness-of-fit

# 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

# Deviance residuals

The **deviance residual** indicates how much the observed data deviates from the fitted model

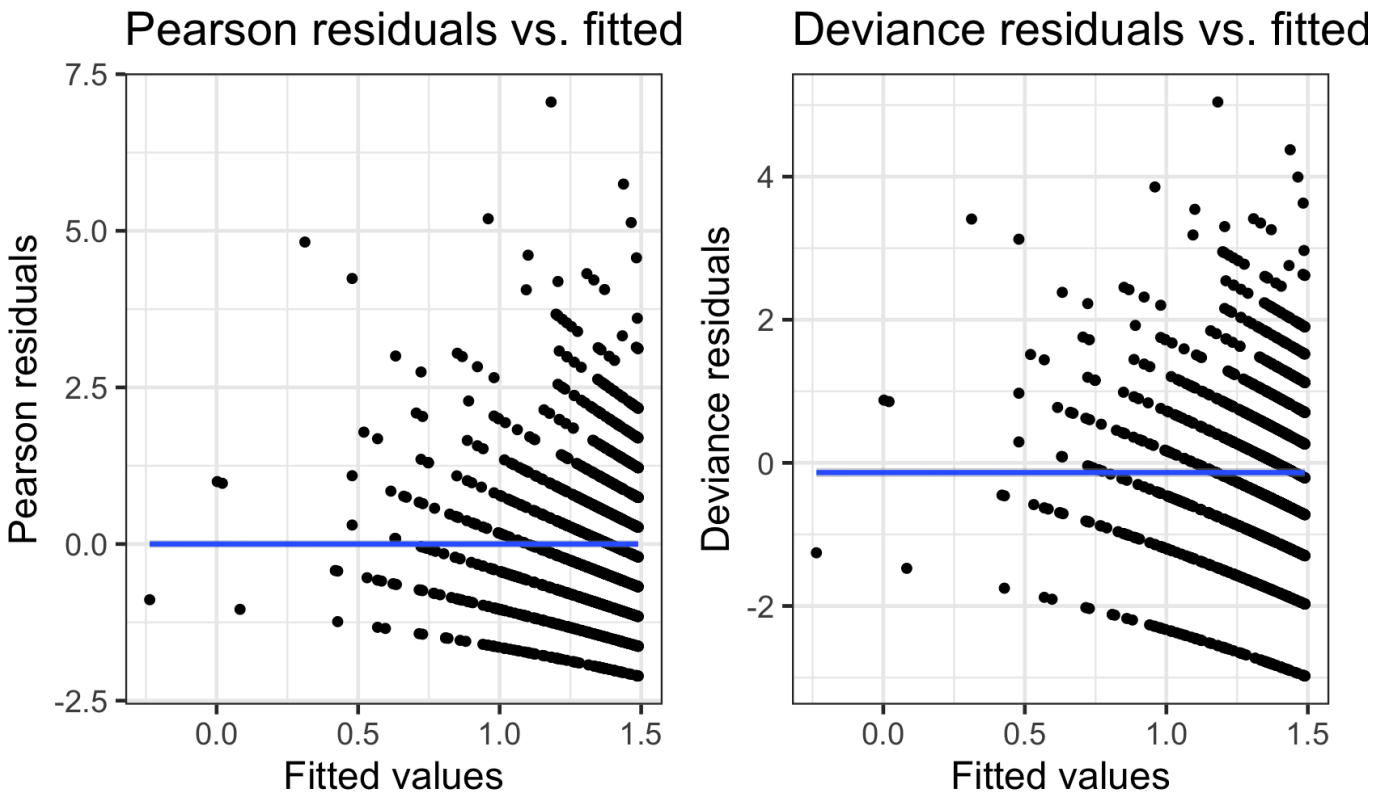
$$\text{deviance residual}_i = \text{sign}(y_i - \hat{\lambda}_i)\sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right]}$$

where

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

# Model 3: Residual plots

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



# Goodness-of-fit

- **Goal**: Use the (residual) deviance to assess how much the predicted values differ from the observed values. Recall
$(\text{deviance}) = \sum_{i=1}^{n} (\text{deviance residual})_i^2$

- If the model sufficiently fits the data, then

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

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?

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


Calculate the goodness-of-fit of Model 3 in R.

# Model 3: 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 $\approx 0$, so there is significant evidence of **lack-of-fit**.

# Lack-of-fit

There are a few potential reasons for 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
    - May need more flexibility in the model to handle **overdispersion**

# 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

❌ The p-values are artificially small

❌ This 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-poission

# Dispersion parameter

The **dispersion parameter** is represented by $\phi$

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

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 in the model

- 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)

# Model 3: Quasi-Poisson model

```
model3_q <- glm(total ~ age + age2 + 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 |
| age2 | -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**

| term | estimate | std.error |
|---|---|---|
| (Intercept) | -0.3843 | 0.1821 |
| age | 0.0704 | 0.0069 |
| age2 | -0.0007 | 0.0001 |
| locationDavaoRegion | -0.0194 | 0.0538 |
| locationIlocosRegion | 0.0610 | 0.0527 |
| locationMetroManila | 0.0545 | 0.0472 |
| locationVisayas | 0.1121 | 0.0417 |

**Quasi-Poisson**

| estimate | std.error |
|---|---|
| -0.3843 | 0.2166 |
| 0.0704 | 0.0082 |
| -0.0007 | 0.0001 |
| -0.0194 | 0.0640 |
| 0.0610 | 0.0626 |
| 0.0545 | 0.0561 |
| 0.1121 | 0.0497 |

# Quasi-Poisson: Inference for coefficients

| term | estimate | std.error |
|---|---|---|
| (Intercept) | -0.3843 | 0.2166 |
| age | 0.0704 | 0.0082 |
| age2 | -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**

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

# 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 $\lambda$

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

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

# Negative binomial regression model

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

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

# Negative binomial simulation exercise

Complete the Negative binomial regression exercise in R.

08:00

# Negative binomial regression in R

```
library(MASS)
model3_nb <- glm.nb(total ~ age + age2 + location, data = hh_data)
tidy(model3_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 |
| age2 | -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 |

# Acknowledgements

These slides are based on content in BMLR - Chapter 4 Poisson regression