

Poisson Regression

Data

2015 Family Income and Expenditure Survey (FIES) on households in the Philippines. Variables include

- + age: age of the head of household
- + numLT5: number in the household under 5 years old
- + total: total number of people other than head of household
- + roof: type of roof (stronger material can sometimes be used as a proxy for greater wealth)
- + location: where the house is located (Central Luzon, Davao Region, Ilocos Region, Metro Manila, or Visayas)

Data

Questions:

- + How is the age of head of household related to the number of people in the household?
- + Is the type of roof material related to the number of people in the household?

To answer these questions, our response variable is `total` (total number of people other than head). What kind of variable is this?

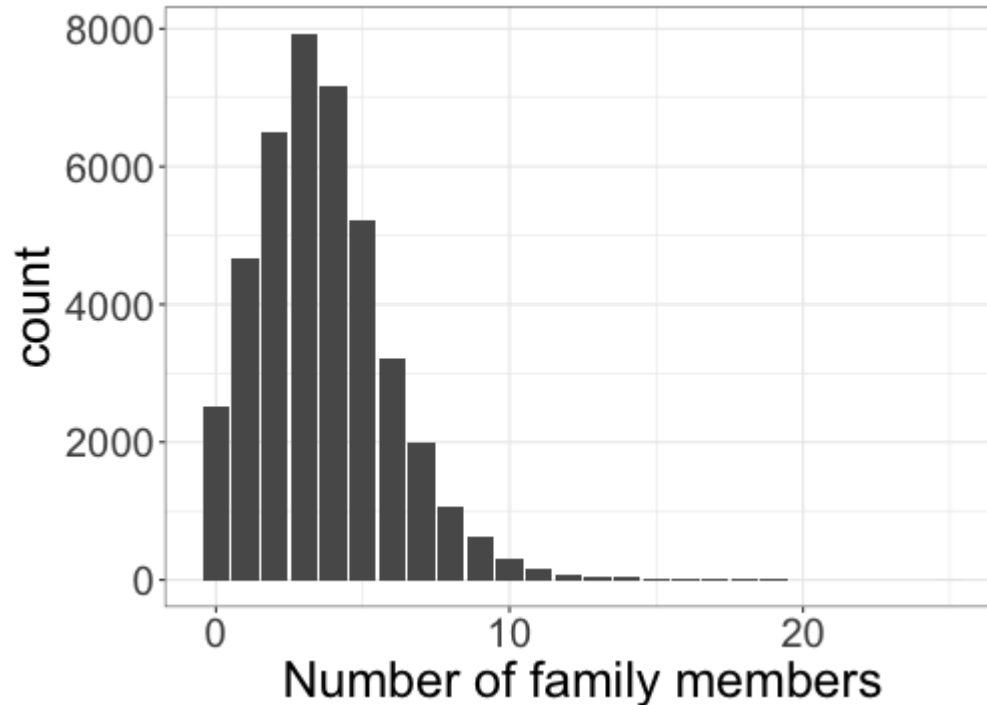
Building a parametric model

Step 1: Choose a reasonable distribution for Y

- + $Y_i = total_i$ is a count variable!
- + Unfortunately, we don't know any distributions for count data
- + Bernoulli and Categorical distributions are for categorical data
- + Normal distributions are for continuous data; count data is discrete

We need a new distribution!

Exploring the response



- + Right skewed, unimodal distribution
- + We can use a *Poisson* distribution

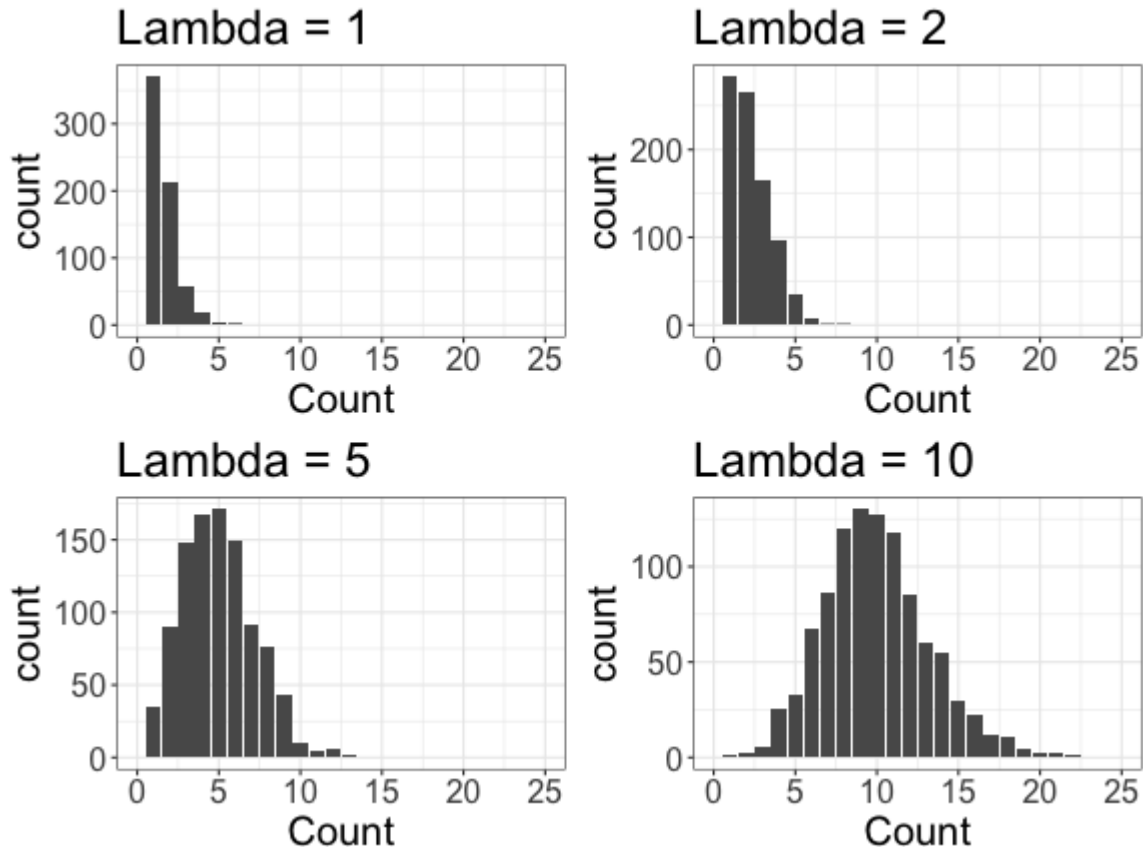
Poisson distribution

If $Y_i \sim \text{Poisson}(\lambda)$, then Y_i takes values $y = 0, 1, 2, \dots$ with probabilities

$$P(Y_i = y) = \frac{e^{-\lambda} \lambda^y}{y!}$$

Does this distribution look familiar?

Poisson distribution



How is λ related to the distribution?

Poisson distribution

If $Y_i \sim \text{Poisson}(\lambda)$, then Y_i takes values $y = 0, 1, 2, \dots$ with probabilities

$$P(Y_i = y) = \frac{e^{-\lambda} \lambda^y}{y!}$$

- + λ is the mean of the distribution
- + λ is also the variance! (the mean and variance are the same)
- + Our goal is to estimate λ , just like our goal was to estimate π in logistic regression

Estimating λ with maximum likelihood

$$P(Y_i = y) = \frac{e^{-\lambda} \lambda^y}{y!}$$

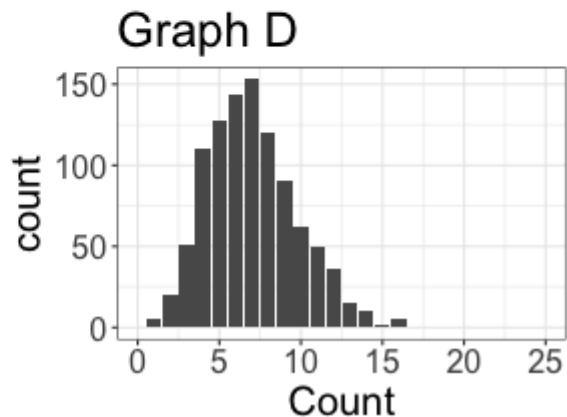
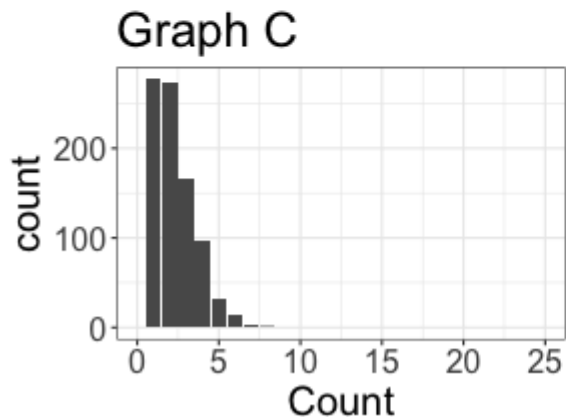
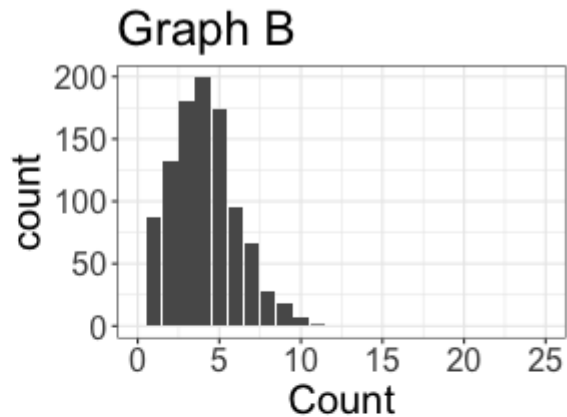
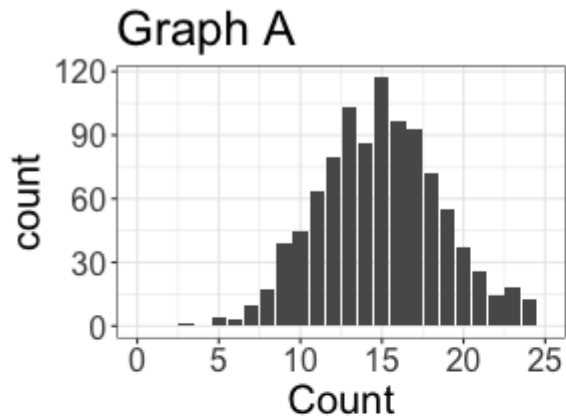
Suppose that $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Poisson}(\lambda)$.

What is the maximum likelihood estimate $\hat{\lambda}$?

Class activity

https://sta214-f22.github.io/class_activities/ca_lecture_17.html

Class activity



What do you think λ is for each graph?

Class activity

Y_i = number of dogs adopted from animal shelter

$$Y_i \sim \text{Poisson}(1.5)$$

What is the probability that at most two dogs are adopted?

Class activity

The Poisson distribution is for count data. Why is it ok for λ to not be a whole number?

Poisson regression

Y_i = number of people in household other than head

How is Y_i related to the age of the head of the household?

Step 1: Choose a reasonable distribution for Y

$$Y_i \sim \text{Poisson}(\lambda_i)$$

Step 2: Choose a model for any parameters

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{Age}_i$$

Why do you think we use $\log(\lambda_i)$ instead of just λ_i ?

Poisson regression

Y_i = number of people in household other than head

Step 1: Choose a reasonable distribution for Y

$$Y_i \sim \text{Poisson}(\lambda_i)$$

Step 2: Choose a model for any parameters

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{Age}_i$$

How is this different than linear regression with a log-transformed response?

Fitting the model

```
m1 <- glm(total ~ age, family = poisson,  
           data = fies)  
summary(m1)
```

```
...  
##               Estimate Std. Error z value Pr(>|z|)  
## (Intercept)  1.7135783  0.0085137  201.27  <2e-16 ***  
## age         -0.0035255  0.0001619  -21.78  <2e-16 ***  
...
```

$$\log(\hat{\lambda}_i) = 1.714 - 0.0035 \text{ Age}_i$$

How can we interpret the slope?

Fitting the model

$$\log(\hat{\lambda}_i) = 1.714 - 0.0035 \text{ Age}_i$$

For every additional year in age of the head of house, we expect the log of the *average* household size to decrease by 0.0035.

Can I interpret on the un-logged scale?

Assumptions

$$Y_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{Age}_i$$

What assumptions does our Poisson regression model make?

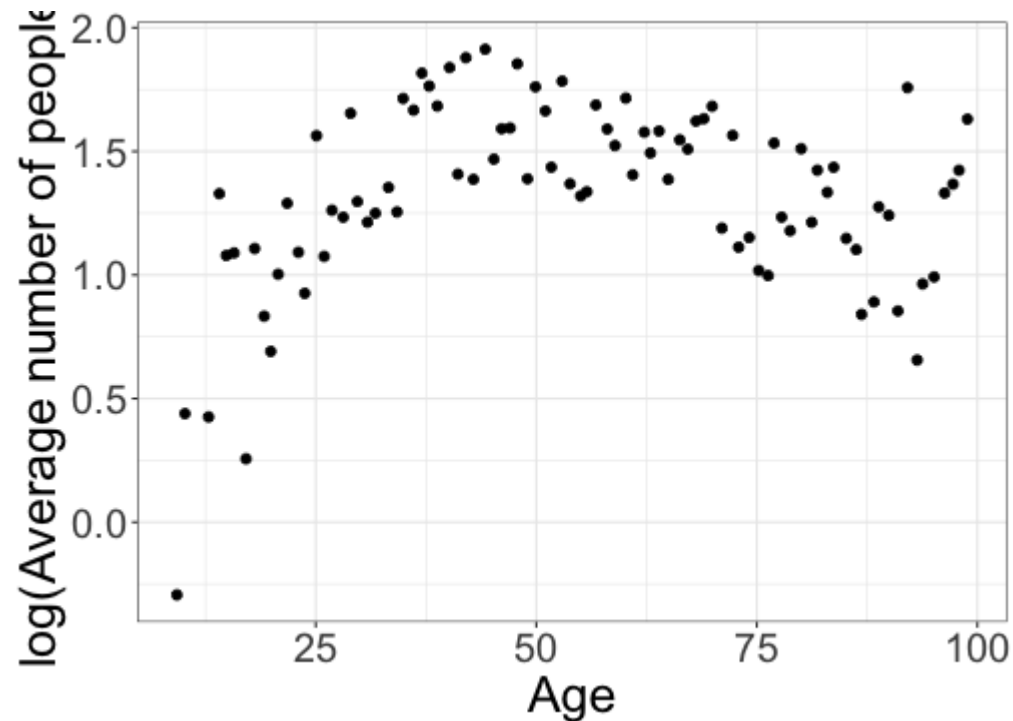
- + **Shape:** The shape of the regression model is correct
- + **Independence:** The observations are independent
- + **Poisson distribution:** A Poisson distribution is a good choice for Y_i

The shape assumption

Shape assumption: The shape of the regression model is correct

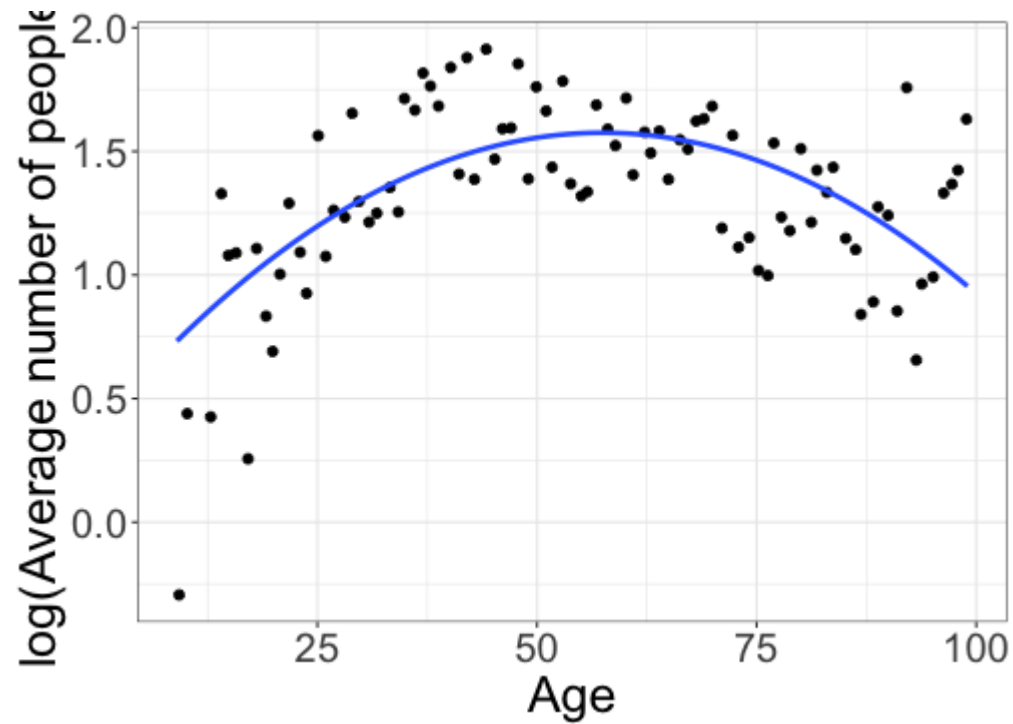
How can I assess this assumption?

Checking the shape assumption



What shape seems appropriate?

Second order polynomial



Poisson distribution assumption

$$Y_i \sim \text{Poisson}(\lambda_i)$$

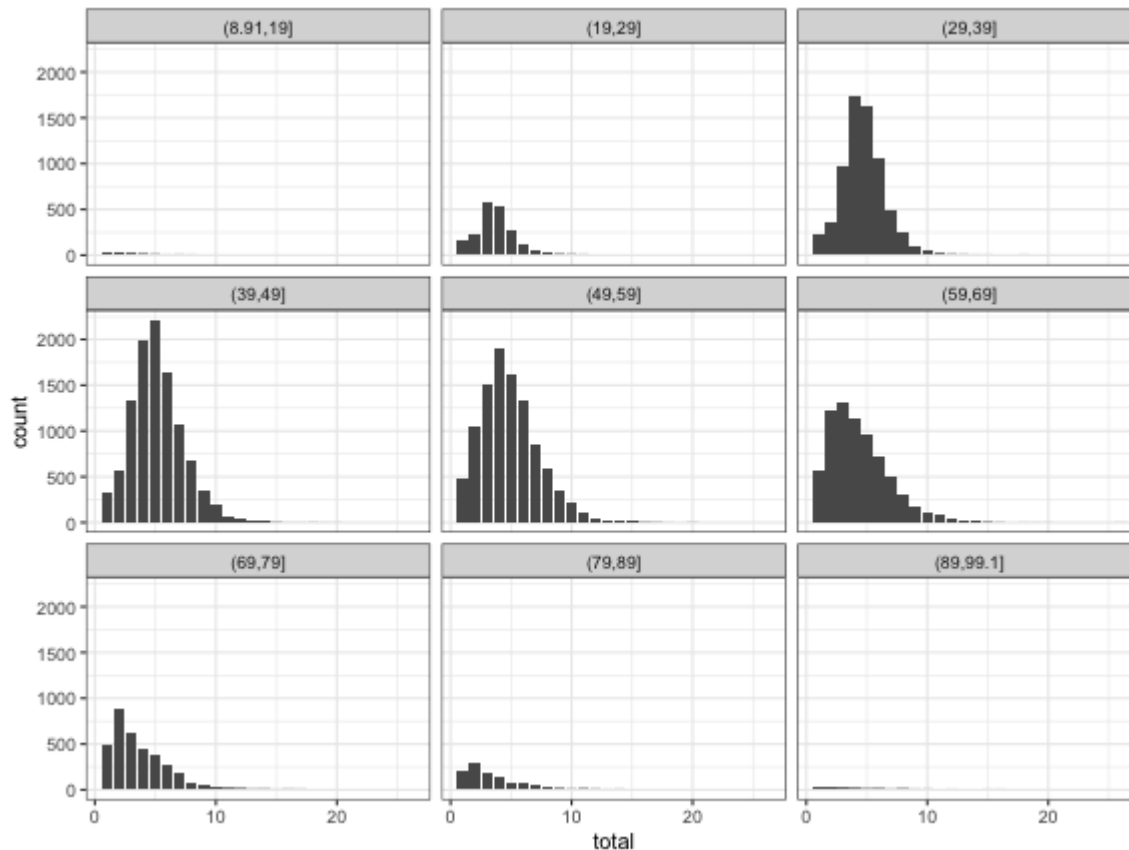
$$\log(\lambda_i) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Age}_i^2$$

Poisson distribution assumption: The Poisson distribution is a good choice for Y_i .

What are some characteristics of the Poisson distribution we could check?

Checking distribution shape

Look at the distribution of Y_i for different ages:



Checking mean and variance

```
mean(fies$total)
```

```
## [1] 4.635182
```

```
var(fies$total)
```

```
## [1] 5.199548
```

- + Rule of thumb: not too concerned if variance is less than twice the sample mean