Chapter 4 Part 1

04/27/2021

Packages

```
# Packages required for Chapter 3
library(gridExtra)
library(knitr)
library(kableExtra)
library(mosaic)
library(xtable)
library(pscl)
library(multcomp)
library(pander)
library(MASS)
library(tidyverse)
```

Poisson Regression - Learning Objectives

After finishing this chapter, you should be able to:

- Describe why simple linear regression is not ideal for Poisson data.
- Write out a Poisson regression model and identify the assumptions for inference.
- Write out the likelihood for a Poisson regression and describe how it could be used to estimate coefficients for a model.
- Interpret estimated coefficients from a Poisson regression and construct confidence intervals for them.
- Use deviances for Poisson regression models to compare and assess models.
- Use an offset to account for varying effort in data collection.
- Fit and use a zero-inflated Poisson (ZIP) model.

Introduction to Poisson Regression

Consider the following questions:

1. Are the number of motorcycle deaths in a given year related to a state's helmet laws?
2. Does the number of employers conducting on-campus interviews during a year differ for public and private colleges?
3. Does the daily number of asthma-related visits to an Emergency Room differ depending on air pollution indices?
4. Has the number of deformed fish in randomly selected Minnesota lakes been affected by changes in trace minerals in the water over the last decade?

Introduction to Poisson Regression

- Each example involves predicting a response using one or more explanatory variables, although these examples have response variables that are *counts* per some unit of time or space.
- A Poisson random variable is often used to model counts
- Since a Poisson random variable is a count, its minimum value is zero and, in theory, the maximum is unbounded.
- We'd like to model our main parameter λ, the average number of occurrences per unit of time or space, as a function of one or more covariates.

Example

- From Before: Are the number of motorcycle deaths in a given year related to a state's helmet laws?
- λ_i represents the average number of motorcycle deaths in a year for state i, and we hope to show that state-to-state variability in λ_i can be explained by state helmet laws.

Linear Least Squares Regression Model

- The parameter of interest is the average response, μ_i, for subject i, and μ_i is modeled as a line in the case of one explanatory variable.
- By analogy, it might seem reasonable to try to model the Poisson parameter λ_i as a linear function of an explanatory variable, but there are some problems with this approach.
  - A model like $\lambda_i = \beta_0 + \beta_1 x_i$ doesn't work well for Poisson data.
  - What are some issues we can imagine use such a model?
  - What about in terms of our linear regression assumptions?

Poisson vs Linear Regression

- A line is certain to yield negative values for certain x_i, but λ_i can only take on values from 0 to ∞.
- The equal variance assumption in linear regression inference is violated because as the mean rate for a Poisson variable increases, the variance also increases $E(Y) = \lambda$ and $SD(Y) = \lambda$.

Trying to Correct

- One way to avoid these problems is to model $\log(\lambda_i)$ instead of λ_i as a function of the covariates.
- The $\log(\lambda_i)$ takes on values from $-\infty$ to ∞. We can also take into account the increase in the variance with an increasing mean using this approach.

Poisson Regression

- Thus, we will consider the **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.
- For example, each state i can potentially have a different λ depending on its value of x_i, where x_i could represent presence or absence of a particular helmet law.
- Note that the Poisson regression model contains no separate error term like the ϵ we see in linear regression, because λ determines both the mean and the variance of a Poisson random variable.

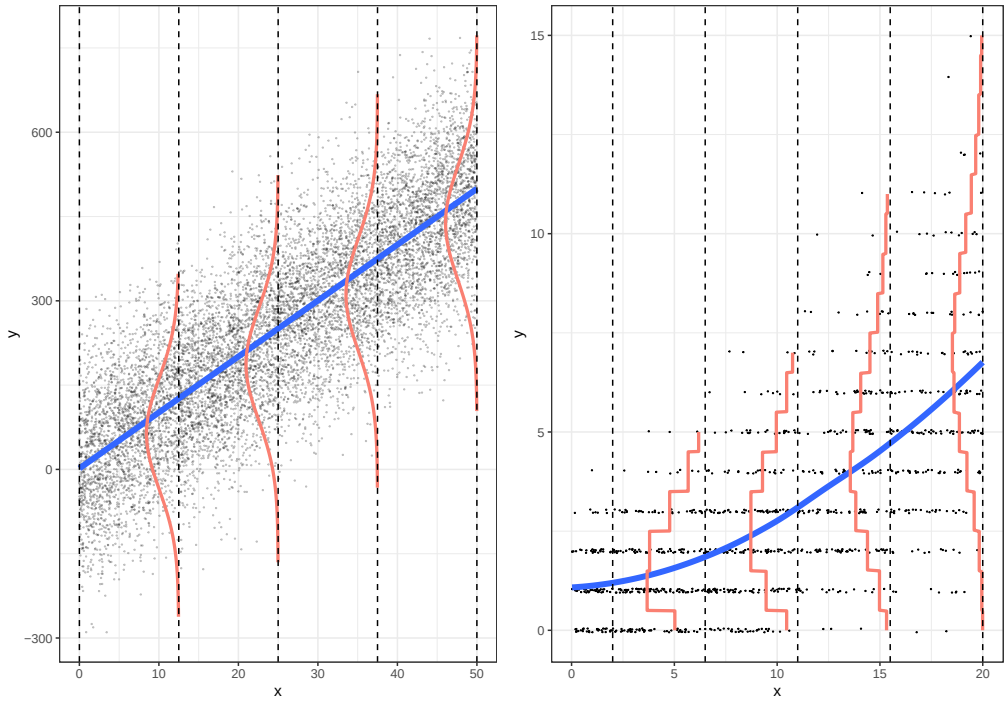
Poisson Regression Assumptions

Much like linear least squares regression (LLSR), using Poisson regression to make inferences requires model assumptions.

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, $\log(\lambda)$, must be a linear function of x.

A Graphical Look at Poisson Regression

- Code saved as LinRegVSPoisson.txt in our notes folder



- This illustrates a comparison of the LLSR model for inference to Poisson regression using a log function of λ.

Comparing Graphs

1. The graphic displaying the LLSR inferential model appears in the left panel. It shows that, for each level of X, the responses are approximately normal.

- The panel on the right side depicts what a Poisson regression model looks like. For each level of X, the responses follow a Poisson distribution (Assumption 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.

Comparing Graphs

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 (Assumption 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 (Assumption 4).

Case Studies Overview

- We take a look at the Poisson regression model in the context of three case studies.
- Each case study is based on real data and real questions.
- Modeling household size in the Philippines introduces the idea of regression with a Poisson response along with its assumptions. A quadratic term is added to a model to determine an optimal size per household, and methods of model comparison are introduced.
- The campus crime case study introduces two big ideas in Poisson regression modeling: offsets, to account for sampling effort, and overdispersion, when actual variability exceeds what is expected by the model.
- The weekend drinking example uses a modification of a Poisson model to account for more zeros than would be expected for a Poisson random variable. These three case studies also provide context for some of the familiar concepts related to modeling such as exploratory data analysis (EDA),

Case Study: Household Size in the Philippines

Read about this HERE

Household Size in the Philippines

- How many other people live with you in your home? -At what age are heads of households in the Philippines most likely to find the largest number of people in their household? Is this association similar for poorer households (measured by the presence of a roof made from predominantly light/salvaged materials)?
- Our data, from the 2015 FIES, is a subset of 1500 of the 40,000 observations [@PSA].
- Our data set focuses on five regions: Central Luzon, Metro Manila, Ilocos, Davao, and Visayas.

Map

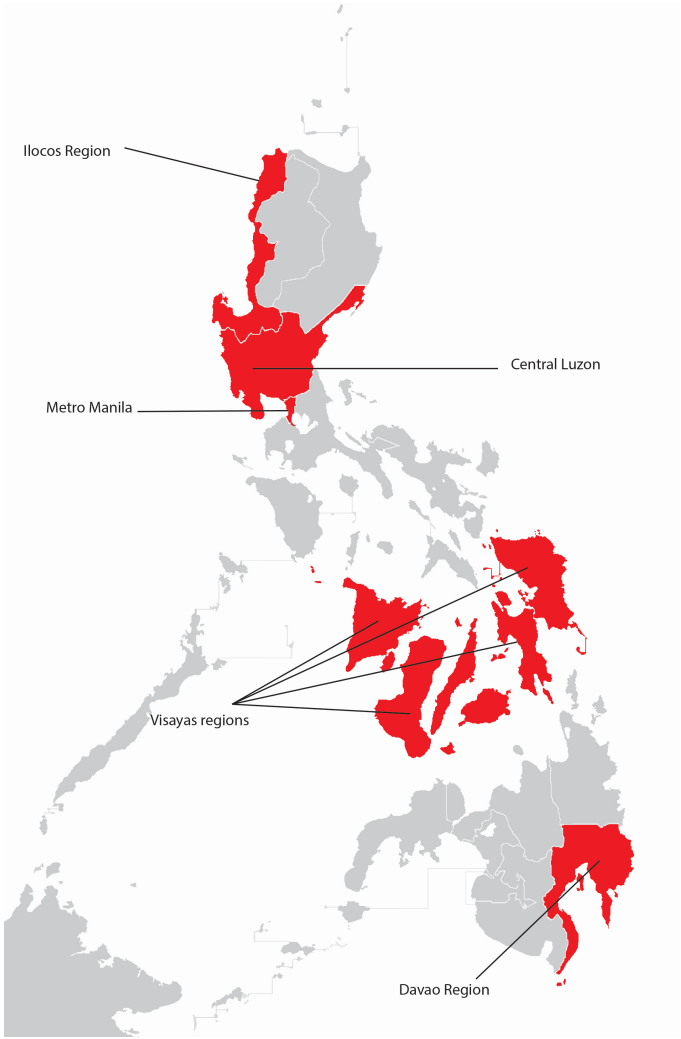


Figure 1: Regions of the Philippines.

Getting Going

- Explicitly define our response, Y = number of household members other than the head of the household.
- Define the explanatory variables: age of the head of the household, type of roof (predominantly light/salvaged material or predominantly strong material), and location (Central Luzon, Davao Region, Ilocos Region, Metro Manila, or Visayas).
- Note that predominantly light/salvaged materials are a combination of light material, mixed but predominantly light material, and mixed but predominantly salvaged material, and salvaged matrial.
- Our response is a count, so we consider a Poisson regression where the parameter of interest is λ, the average number of people, other than the head, per household. We will primarily examine the relationship between household size and age of the head of household, controlling for location and income.

Data Organization

- Lets read our data into R calling it fHH1
- Lets look at the first few rows
- Each line of the data file refers to a household at the time of the survey:
  - `location` = where the house is located (Central Luzon, Davao Region, Ilocos Region, 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 in the household (either Predominantly Light/Salvaged Material, or Predominantly Strong Material, where stronger material can sometimes be used as a proxy for greater wealth)

Exploratory Data Analyses

- I would like you to be able to do investigate the following:
- mean and sd of the total
- proportions of the types of roof
- mean, sd, variance, and count of total by roof type
- mean, sd, variance, and count of total by location
- the distribution of total
- Create age groups and compare the distribution of total in each age group
- Using those same age groups, look for the mean, varainace, and frequency of total in each group (why?)

Comments

- For the rest of this case study, we will refer to the number of people in a household as the total number of people in that specific household *besides* the head of household.
- A fair amount of variability in the number in each house; responses range from 0 to 16 with many of the respondents reporting between 1 and 5 people in the house. Like many Poisson distributions, this graph is right skewed. It clearly does not suggest that the number of people in a household is a normally distributed response.
- Responses can be reasonably modeled with a Poisson distribution when grouped by a key explanatory variable: age of the household head. These last two plots together suggest that Assumption 1 (Poisson Response) is satisfactory in this case study.

Comments

- We display age groups by 5-year increments, to check to see if the empirical means and variances of the number in the house are approximately equal for each age group. This provides us one way in which to check the Poisson Assumption 3 (mean = variance).
- If there is a problem with this assumption, most often we see variances much larger than means.
  - Here, as expected, we see more variability as age increases. However, it appears that the variance is smaller than the mean for lower ages, while the variance is greater than the mean for higher ages.
  - Thus, there is some evidence of a violation of the mean=variance assumption (Assumption 3), although any violations are modest.

Checking the Linearity

- The Poisson regression model also implies that $\log(\lambda_i)$, not the mean household size λ_i, is a linear function of age; i.e., $log(\lambda_i) = \beta_0 + \beta_1 \text{age}_i. +$
- Therefore, to check the linearity assumption (Assumption 4) for Poisson regression, we would like to plot $\log(\lambda_i)$ by age.
- Unfortunately, λ_i is unknown. Our best guess of λ_i is the observed mean number in the household for each age (level of X).
- Because these means are computed for observed data, they are referred to as **empirical** means.

Checking the Linearity

- Taking the logs of the empirical means and plotting by age provides a way to assess the linearity assumption.
- The smoothed curve suggests that there is a curvilinear relationship between age and the log of the mean household size, implying that adding a quadratic term should be considered.
- This finding is consistent with the researchers' hypothesis that there is an age at which a maximum household size occurs.
- It is worth noting that we are not modeling the log of the empirical means, rather it is the log of the *true* rate that is modeled.
- Looking at empirical means, however, does provide an idea of the form of the relationship between $\log(\lambda)$ and x_i.

Checking Linearity in each Region

- This allows us to see if the relationship between mean household size and age is consistent across region.
- In this case, the relationships are pretty similar; if they weren't, we could consider adding an age-by-region interaction to our eventual Poisson regression model.
- Lets create this in R

Independence

- The independence assumption (Assumption 2) can be assessed using knowledge of the study design and the data collection process.
- In this case, we do not have enough information to assess the independence assumption with the information we are given.
- If each household was not selected individually in a random manner, but rather groups of households were selected from different regions with differing customs about living arrangements, the independence assumption would be violated.
- If this were the case, we could use a multilevel model like those discussed in later chapters with a village term.