Generalized Linear Model - Poisson Regression

Claudio Mazzi
Department of Computer Science, University of Pisa
MeS Laboratory, Sant'Anna School for Advanced Studies, Pisa
claudio.mazzi@santannapisa.it

Applied Statistical Modelling
A.A. 2025-2026

Contents

1 Poisson Distribution

The Poisson distribution is a fundamental probability distribution used to model count data and events occurring within a fixed interval of time or space. It describes the likelihood of a given number of events happening when these events occur independently of each other and at a constant average rate. The probability mass function of the Poisson distribution is given by:

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad y = 0, 1, 2, 3, \ldots \tag{1}$$

where Y is the random variable representing the number of events, and $\lambda > 0$ is the expected number of events in the interval (known as the rate parameter). Key characteristics of the Poisson distribution include its mean and variance, both equal to λ. This equality implies a direct relationship between the distribution's central tendency and its variability, making it suitable for modeling processes where the occurrence rate is stable over time or space. However, when the variance exceeds the mean—a phenomenon known as overdispersion—alternative methods may be required, such as the negative binomial distribution.

The Poisson distribution is widely used in fields such as epidemiology, insurance, and traffic modeling, where it provides a natural framework for analyzing the frequency of rare events. Its simplicity and strong theoretical foundation make it an essential tool for generalized linear models (GLMs), particularly Poisson regression, which is designed to investigate relationships between predictors and count-based response variables.

In this analysis, we will deal with the Poisson Regression applied to different datasets, focusing on the implementation in R programming systems.

2 Introduction to Poisson Regression

When analyzing count data, several real-world questions naturally arise:

- Does daily air pollution influence the frequency of asthma-related Emergency Room visits?
- Does the number of daily customer complaints at a retail store depend on the time of year or promotional campaigns?
- Are the number of road accidents in urban areas influenced by traffic density and weather conditions?
- Has the number of invasive plant species recorded in protected ecosystems changed due to shifts in climate or conservation policies?

Each of these examples involves modeling a response variable, specifically a count, to one or more explanatory variables. A Poisson random variable is frequently used to model such count data. Its probability mass function is defined in Eq. 1, and it takes only non-negative values and has a theoretical upper bound of infinity. While the Poisson distribution itself is not memoryless, the time between successive events, modeled using an exponential distribution, exhibits this property. Moreover, the sum of independent Poisson random variables is also Poisson distributed, with a rate parameter equal to the sum of their rates. The distribution is positively skewed, with skewness decreasing as λ increases. Its kurtosis is $3 + 1/\lambda$, approaching that of a normal distribution as λ grows.

Finally, the Poisson distribution is closely related to other distributions. It can be derived as a limiting case of the binomial distribution when the number of trials is large, and the probability of success is small, with their product equal to λ. For large λ, the Poisson distribution can be approximated by a normal distribution with mean and variance equal to λ. These characteristics make it a foundational model for counting data, particularly when events occur independently and at a constant average rate.

As already said, the key parameter of interest in Poisson models, denoted as λ, represents the average rate of occurrences per unit of time or space. For example, in the context of road accidents, λ_i could represent the mean monthly number of fatalities in state i, potentially varying with the state's motorcycle density and climate.

While linear regression might seem like a natural analog for modeling λ, this approach is problematic. A model such as $\lambda_i = \beta_0 + \beta_1 x_i$ risks producing negative values for λ_i, which is mathematically infeasible given the non-negative nature of Poisson counts. Furthermore, the fundamental assumption of homoscedasticity in linear regression is violated because, for Poisson data, the variance increases proportionally with the mean $E(Y) = \mathrm{Var}(Y) = \lambda$.

To address these issues, Poisson regression models the natural logarithm of λ_i as a linear function of explanatory variables, i.e.,

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i.$$

This transformation ensures that $\lambda_i > 0$, and aligns with the variance structure inherent in Poisson data. Unlike linear regression, Poisson regression lacks an explicit error term because the Poisson distribution's structure inherently captures variability via the parameter λ, which governs both the mean and variance of the response variable.

2.1 Assumptions of Poisson Regression

Like linear least squares regression, Poisson regression relies on certain assumptions for valid inference:

- **Poisson Response:** The response variable must represent counts per unit of time or space and follow a Poisson distribution.
- **Independence:** Observations must be statistically independent.
- **Mean-Variance Equality:** The mean and variance of the Poisson-distributed response variable are assumed to be equal.

- **Log-Linearity:** The natural logarithm of the mean rate $\log(\lambda)$ is modeled as a linear function of the covariates.

These principles form the foundation for applying Poisson regression (PR) within the generalized linear model (GLM) framework, offering a robust method for analyzing count data.

Now, we take a look at the Poisson regression model in the context of different case studies inspired by the textbook *Beyond Multiple Linear Regression* of Paul Roback and Julie Legler, which is available at the link: https://bookdown.org/roback/bookdown-BeyondMLR/. Each case study is based on real data and real questions.

3 Case Study: Household Size in the Philippines

International agencies use household size when determining the needs of populations, and household sizes determine the magnitude of the household needs. The Philippine Statistics Authority (PSA) spearheads the Family Income and Expenditure Survey (FIES) nationwide. The survey, which is undertaken every three years, is aimed at providing data on family income and expenditure, including levels of consumption by item of expenditure. Our data (available here https://github.com/EMbeDS-education/ComputingDataAnalysisModeling20242025/wiki/Datasets), from the 2015 FIES, is a subset of the 40,000 observations focused on five regions: Central Luzon, Ilocos, Davao, and Visayas.

In this exercise, we will address two research questions: (1) At what age are heads of households in the Philippines most likely to have the largest number of other members in their household? (2) Is this association similar for poorer households, as indicated by the presence of a roof made primarily of light or salvaged materials?

We begin by explicitly defining our response variable: Y = the number of household members excluding the head of the household. Next, we define the explanatory variables: the age of the head of the household, the type of roof (predominantly light/salvaged materials or predominantly strong materials), and the location (Central Luzon, Davao Region, Ilocos Region, and Visayas).

Since the response variable is a count, we employ a Poisson regression model, where the parameter of interest is λ, representing the average number of household members (excluding the head) per household. Our primary focus is on the relationship between household size and the age of the household head while controlling for location and income.

3.1 Data Preprocessing

Loaded the dataset, we have to perform several pre-processing passages since the original database presents many unuseful variables, and unpractical notation both for the columns name and for categorical values for *Roof* and *Region*.

Listing 1: Loading Libraries and Preprocessing the dataset

```
# Step 1: Install and Load Required Packages


library(dplyr)
library(tidyverse)
library(GGally)
library(ggplot2)
library(pROC)
library(stats)
library(nnet)
library(MASS)     # For ordinal logistic regression
library(pscl)     # For calculating Pseudo R^{2}
library(ggcorrplot)
library(smotefamily)

# POISSON REGRESSION - General / Over-dispersed / Zero-Saturated

# Step 2 - Household Size in the Philippines>
# Load the db
db_original <- read.csv("/Users/claudiomazzi/Documents/PhD/Course_AppSTAT/2_GLM/
    db/Fam_Income_Exp.csv")
```

```r
unique(db_original$Region)

# Pre processing the db
db_ph <- db_original %>%
  dplyr::select(Region,
                Age = Household.Head.Age,
                Roof = Type.of.Roof,
                NumLT5 =Members.with.age.less.than.5.year.old,
                Total = Total.Number.of.Family.members %>%
                mutate(Total = Total - 1) )

db_ph <- db_ph %>%
  mutate(Roof = case_when
                ( Roof == "Strong material(galvanized,iron,al,tile,
                            concrete,brick,stone,asbestos)" ~ "Pred. Strong
                               Material",
                  Roof == "Light material (cogon,nipa,anahaw)" ~ "Pred. Light
                      Material",
                  Roof == "Mixed but predominantly strong materials" ~ "Pred.
                      Strong Material",
                  Roof == "Mixed but predominantly light materials" ~ "Pred.
                      Light Material",
                  Roof == "Salvaged/makeshift materials" ~ "Pred. Light Material
                      ",
                  Roof == "Mixed but predominantly salvaged materials" ~ "Pred.
                      Light Material"
                  )
                )
db_ph <- db_ph %>%
  filter(Roof != "Not Applicable")

db_ph <- db_ph %>%
  filter(Region %in% c("VII - Central Visayas", "III - Central Luzon", "VI -
      Western Visayas",
                  "XI - Davao Region", "VIII - Eastern Visayas", "I - Ilocos
                      Region"))
db_ph <- db_ph %>%
  mutate(Region = case_when
          ( Region == "VII - Central Visayas" ~ "Visayas",
            Region == "III - Central Luzon" ~ "Central Luzon",
            Region == "VI - Western Visayas" ~ "Visayas",
            Region == "XI - Davao Region" ~ "Davao Region",
            Region == "VIII - Eastern Visayas" ~ "Visayas",
            Region == "I - Ilocos Region" ~ "Ilocos Region"
          )
  )

head(db_ph)
str(db_ph)
summary(db_ph)
dim(db_ph)
```

Figure 1 shows the header and the database structure after the preprocessing step. Note that the final database presents 1444 observations and 5 variables: *Region, Age, Roof, NumLT5* (the number in the household under 5 years of age), and *Total*, which stands for the total number of people in that specific household besides the head of household. As already said, this variable will be the response variable of our Poisson regression.

3.2 Exploratory Data Analysis

Beyond the summary and structure of the database, shown in Figure 1, we perform several descriptive statistics to explore the database and make a priori hypothesis on distributions and the behavior of the selected response variable. From the following codes:

```
> head(db_ph)
   Region Age                 Roof NumLT5 Total
1 Visayas  58 Pred. Strong Material      0     2
2 Visayas  59 Pred. Strong Material      1     2
3 Visayas  73 Pred. Strong Material      0     0
4 Visayas  53 Pred. Strong Material      0     3
5 Visayas  49  Pred. Light Material      1     4
6 Visayas  37 Pred. Strong Material      0     0
> str(db_ph)
'data.frame':   2774 obs. of  5 variables:
 $ Region: chr  "Visayas" "Visayas" "Visayas" "Visayas" ...
 $ Age   : int  58 59 73 53 49 37 50 37 95 42 ...
 $ Roof  : chr  "Pred. Strong Material" "Pred. Strong Material" "Pred. Strong Material" "Pred. Strong Material" ...
 $ NumLT5: int  0 1 0 0 1 0 1 2 0 0 ...
 $ Total : num  2 2 0 3 4 0 5 6 2 2 ...
> summary(db_ph)
    Region               Age            Roof              NumLT5           Total
 Length:2774        Min.   :16.00   Length:2774        Min.   :0.0000   Min.   : 0.000
 Class :character   1st Qu.:41.00   Class :character   1st Qu.:0.0000   1st Qu.: 2.000
 Mode  :character   Median :50.00   Mode  :character   Median :0.0000   Median : 3.000
                    Mean   :51.62                      Mean   :0.4416   Mean   : 3.583
                    3rd Qu.:62.00                      3rd Qu.:1.0000   3rd Qu.: 5.000
                    Max.   :97.00                      Max.   :5.0000   Max.   :16.000
> dim(db_ph)
[1] 2774    5
```

Figure 1: Overview of the database on the Household in the Philippines

Listing 2: Descriptive Statiscisal

```r
# Descriptive statistics Total(response Variable Vs Predictor)
describe(db_ph$Total)

stats_by_roof <- aggregate(Total ~ Roof, data = db_ph, FUN = function(x)
  c(mean = mean(x, na.rm = TRUE), variance = var(x, na.rm = TRUE)))
stats_by_roof)

stats_by_Region <- aggregate(Total ~ Region, data = db_ph, FUN = function(x)
  c(mean = mean(x, na.rm = TRUE), variance = var(x, na.rm = TRUE)))
stats_by_Region
```

the average household size is 3.53, with values ranging from 0 to 16. Approximately 72.6% of these households have roofs made predominantly of light or salvaged materials. For households with roofs made from predominantly strong materials, the mean size is 3.52 (Variance = 5.67). In contrast, households with roofs made from predominantly light or salvaged materials have an average size of 3.57 (Variance = 4.97). Among the various regions, the Ilocos Region reports the largest average household size at 3.65, while the Davao Region has the smallest average household size at 3.42. Now, we explore the distributions of the response variable, to put the basis of the analysis, and its modeling through a Poisson distribution.

Figure 2 reveals 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 does not suggest that the number of people in a household is a normally distributed response. In figure 3 we show the distribution of the response variable concerning different age groups. It further shows that responses can be reasonably modeled with a Poisson distribution when grouped by a key explanatory variable: the age of the household head.

3.3 Checking Poisson Assumptions

For Poisson random variables, the variance of Y (i.e., the square of the standard deviation of Y) is equal to its mean, where Y represents the size of an individual household. As the mean increases, the variance also increases. Therefore, if the response variable is a count, and the mean and variance are approximately equal across groups of X, a Poisson regression model may be appropriate. To examine this, Table 4 displays age groups in 5-year increments, allowing us to compare the empirical means and variances of the number of individuals in a household for each age group. This serves as a method to assess one of the three Poisson Assumptions, which states that the mean equals the variance.

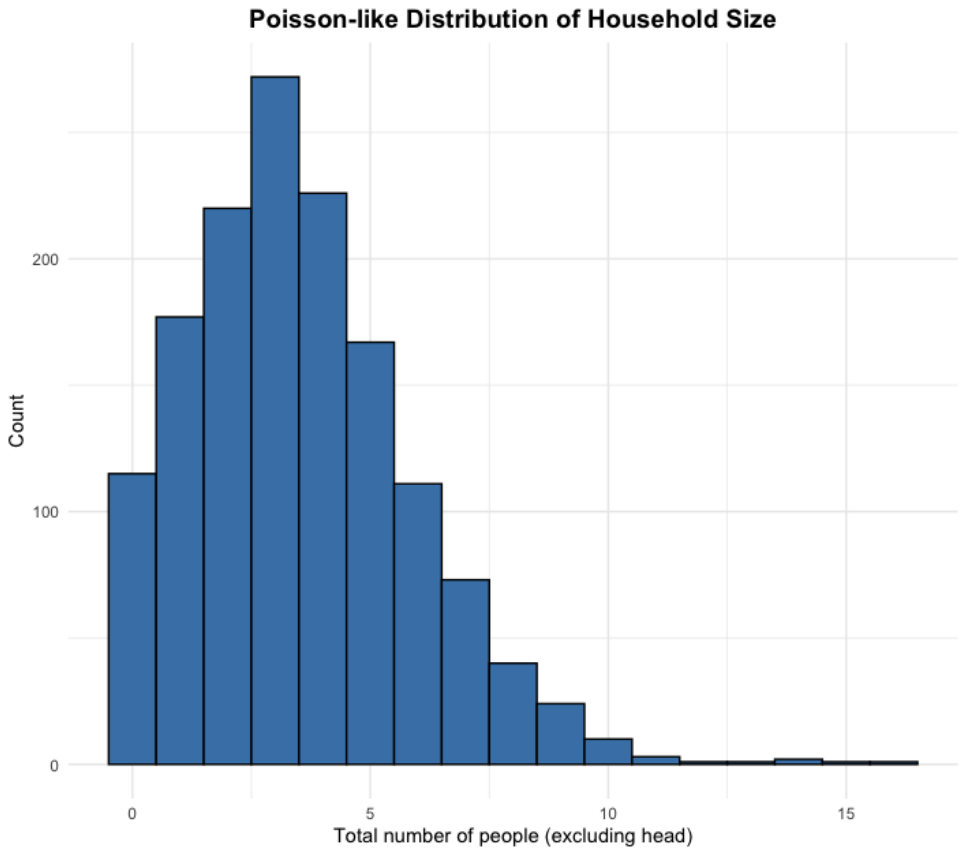

Figure 2: Distribution of household size in four Philippine regions.

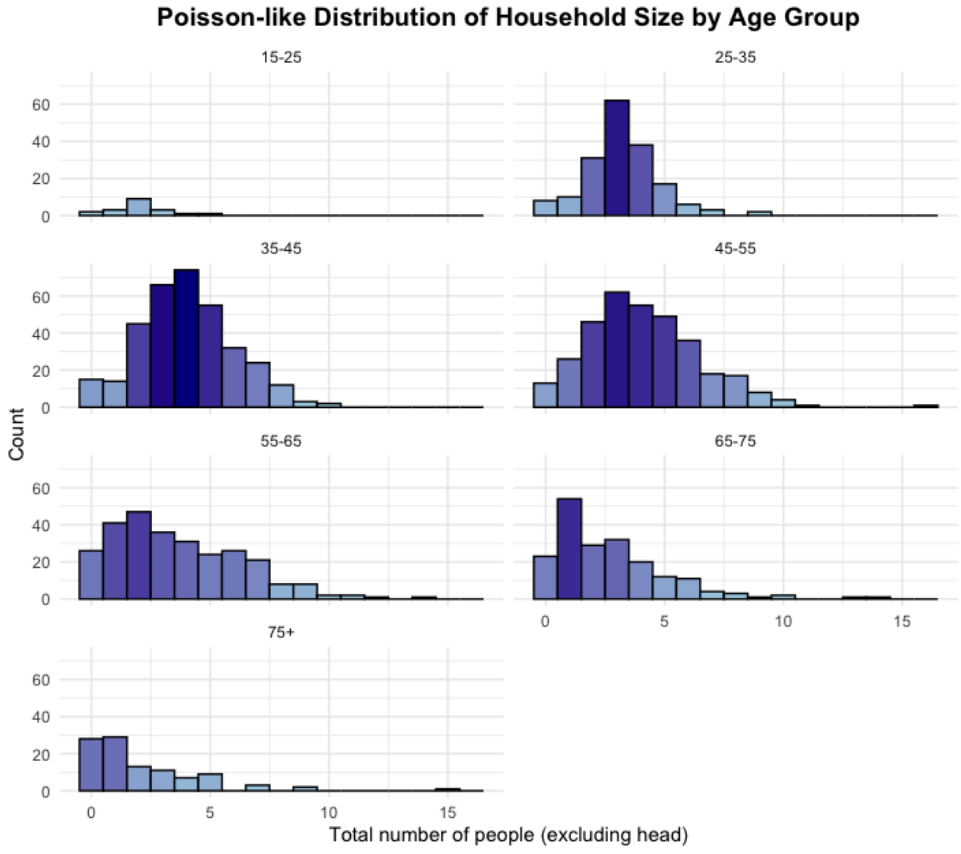

Figure 3: Distribution of household sizes by age group of the household head.

	Age_Group	Mean_Total	Variance_Total	Observations
	<fct>	<dbl>	<dbl>	<int>
1	15-25	2.05	1.50	19
2	25-35	3.22	2.35	177
3	35-45	4.02	3.96	342
4	45-55	4.12	5.51	336
5	55-65	3.63	7.14	274
6	65-75	2.70	5.71	193
7	75+	2.09	5.90	103

Figure 4: Compare mean and variance of household size within each age group.

If the assumption of mean equal to variance is violated, we often observe variances significantly larger than the means. In this analysis, as expected, variability increases with age. However, the variance appears to be smaller than the mean for lower age groups, while it exceeds the mean for higher ages. This suggests a potential violation of the mean = variance assumption, although the deviations are relatively modest.

The Poisson regression model also assumes that $\log(\lambda_i)$ is a linear function of age. Specifically,

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

To evaluate the linearity assumption for Poisson regression, we can plot $\log(\lambda_i)$ against age. Since λ_i is unknown, our best estimate is the observed mean household size for each age group (level of X). These observed means are referred to as empirical means. Taking the logarithm of the empirical means and plotting them against age provides a method to assess the linearity assumption.

The smoothed curve shown in Figure 5 indicates a curvilinear relationship between age and the log of the mean household size. This suggests that including a quadratic term in the model could improve the fit. This observation aligns with the hypothesis that household size peaks at a specific age. It is important to note, however, that we are modeling the log of the true rate λ_i, not the log of the empirical means. Nonetheless, examining empirical means provides valuable insight into the relationship between $\log(\lambda)$ and x_i.

Moreover, We can extend Figure 5 by fitting separate curves for each region (see Figure 6). This allows us to see if the relationship between mean household size and age is consistent across regions. 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.

The independence assumption can be evaluated by considering the study design and how the data were collected. However, in this scenario, the information provided is insufficient to properly assess whether this assumption holds. If households were not selected randomly on an individual basis, but instead as groups from specific regions with distinct cultural practices regarding living arrangements, the independence assumption would not be satisfied.

4 Poisson Regression

In this section, we will begin by fitting a Poisson regression model where the logarithm of the expected response λ is modeled as a linear function of age. We will then extend the model to include a quadratic term for age, allowing us to assess whether this addition significantly improves the model's fit, based on patterns observed during the exploratory data analysis.

The R code for performing these analyses and presenting the regression results are the following:

Listing 3: Poisson Regression

```r
# Poisson Regression Model: Total ~ Age
model_poisson <- glm(Total ~ Age, family = poisson(link = "log"), data = db_ph)
summary(model_poisson)

# Adding quadratic term
db_ph <- db_ph %>%
  mutate(Age2 = Age^2)

```

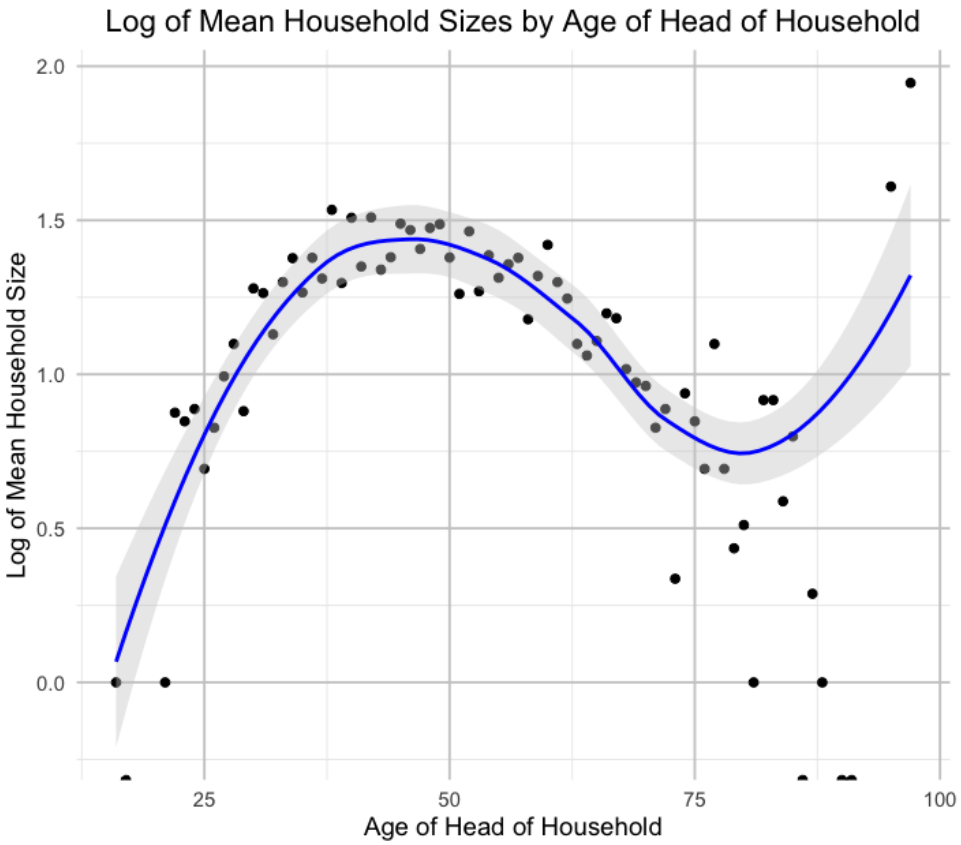

Figure 5: The log of the mean household sizes, besides the head of household, by age of the head of household, with loess smoother.

Figure 6: Empirical log of the mean household sizes vs. age of the head of household, with loess smoother by region.

```r
# Poisson Regression Model: Total ~ Age + Age^{2}
model_poisson_quad <- glm(Total ~ Age + Age2, family = poisson(link = "log"),
    data = db_ph)
summary(model_poisson_quad)
```

The results are shown in Figure 7-8.

```
> # Poisson Regression Model: Total ~ Age
> model_poisson <- glm(Total ~ Age, family = poisson(link = "log"), data = db_ph)
> summary(model_poisson)

Call:
glm(formula = Total ~ Age, family = poisson(link = "log"), data = db_ph)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.6077345  0.0499055  32.216  < 2e-16 ***
Age         -0.0068752  0.0009672  -7.108 1.17e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2435.7  on 1443  degrees of freedom
Residual deviance: 2384.5  on 1442  degrees of freedom
AIC: 6470.6

Number of Fisher Scoring iterations: 5
```

Figure 7: Summary of the Poisson Regression with $log(\lambda)$ linear in *Age*.

At first, let's focus only on the simpler model, with only the linear term for predictor *Age*.

How can the coefficient estimates be interpreted in terms of this example? We consider how the estimated mean number in the house, λ, changes as the age of the household head increases by an additional year. But in place of looking at change in the mean number in the house, with a Poisson regression we consider the log of the mean number in the house and then convert back to original units. These results suggest that by exponentiating the coefficient β on age we obtain the multiplicative factor by which the mean count changes. In this case, the mean number in the house changes by a factor of $e^{-0.0068752} = 0.993$, or decreases by 0.7% (since $1 - 0.993 = 0.007$) with each additional year older. the household head is; or, we predict a 0.70% increase in mean household size for a 1-year decrease in the age of the household head (since $1/0.993 = 1.0070$).

Typically, the standard errors for the estimated coefficients are included in the Poisson regression output. Here the standard error for the estimated coefficient for age is 0.0009672. We can use the standard error to construct a confidence interval for Age. A 95% CI provides a range of plausible values for the age coefficient and can be constructed as:

$$\text{CI} = \hat{\beta} \pm Z^* SE(\hat{\beta}) \quad \Longrightarrow \quad \text{CI} = (-0.0106, -0.0049)$$

If there is no association between age and household size, there is no change in household size for each additional year, so λ_X is equal to λ_{X+1} and the ratio λ_{X+1}/λ_X is 1; in other words $\beta = 0$ and $e^{\beta} = 1$. Note that our interval for $e^{\hat{\beta}}$, $(-0.0106, -0.0049)$, does not include 1, so the model with age is preferred to a model without age; i.e., age is significantly associated with household size.

Another way in which to assess how useful age is in our model. Deviance measures how the observed data deviates from the model predictions. We want models that minimize deviance, so we calculate the drop-in-deviance when adding age to the model with no covariates (the null model). Both the deviances are shown in the summary of the model in Figure 7. So, the drop-in deviance is $2435.7 - 2384.5 = 51.2$

```
> model_poisson_quad <- glm(Total ~ Age + Age2, family = poisson(link = "log"), data = db_ph)
> summary(model_poisson_quad)

Call:
glm(formula = Total ~ Age + Age2, family = poisson(link = "log"),
    data = db_ph)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.683e-01  1.748e-01  -1.535    0.125
Age          7.104e-02  6.966e-03  10.199   <2e-16 ***
Age2        -7.506e-04  6.667e-05 -11.259   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2435.7  on 1443  degrees of freedom
Residual deviance: 2243.0  on 1441  degrees of freedom
AIC: 6331.1

Number of Fisher Scoring iterations: 5
```

Figure 8: Summary of the Poisson Regression with $log(\lambda)$ linear in quadratic *Age*.

with a difference of only 1 dof, so the addition of one extra term (age) reduced unexplained variability by 51.2. If the null model were true, we would expect the drop-in-deviance to follow a χ^2 distribution with 1 dof. Therefore, the p-value for comparing the null model to the model with age is found by determining the probability that the value for a χ^2 random variable with one degree of freedom exceeds 51.2, which is essentially 0. Once again, we can conclude that we have statistically significant evidence ($\chi^2_{dof=1} = 51.2, p = 8.34 \times 10^{-13} < 0.001$) that average household size decreases as the age of the head of household increases.

our exploratory data analysis in Section 4.4.2 suggests that a quadratic model might be more appropriate. A quadratic model would allow us to see if there exists an age where the number in the house is, on average, a maximum. The output for a quadratic model appears in Figure 8. As done before, we can assess the importance of the quadratic term computing the drop-in deviance test. But, more interesting is to assess the importance of adding the second-order term against the model without the quadratic model.

4.1 Comparing Models

To compare the two different models tested, we can provide an ANOVA test between models, as explained by the following code:

Listing 4: ANOVA test between linear and quadratic model

```
# Comparison
anova(model_poisson, model_poisson_quad, test = "Chisq")
```

As shown in Figure 9 the drop-in-deviance by adding the quadratic term to the linear model is $2384.6 - 2243.0 = 141.6$ which can be compared to a χ^2 distribution with one degree of freedom. The p-value is essentially 0, so we providing significant support for including the quadratic term. Similarly one can try to add different covariates, such as *Region*, and test the goodness of fit and the forecasting power of a more complex model.

4.2 Residual Plots

Residual plots may provide some insight into Poisson regression models, especially linearity and outliers, although the plots are not quite as useful here as they are for linear least squares regression. There are a

```
> # Comparison
> anova(model_poisson, model_poisson_quad, test = "Chisq")
Analysis of Deviance Table

Model 1: Total ~ Age
Model 2: Total ~ Age + Age2
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1442     2384.6
2      1441     2243.0  1   141.53 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: ANOVA test summary.

few options for computing residuals and predicted values. We have several options for creating residuals for Poisson regression models. In particular, deviance residuals have some useful properties that make them the best choice for Poisson regression. We define a deviance residual for an observation from a Poisson regression:

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

As its name implies, a deviance residual describes how the observed data deviates from the fitted model. Squaring and summing the deviances for all observations produces the residual deviance $= \sum(\text{deviance residual}_i^2)$. Relatively speaking, observations for good fitting models will have small deviances; that is, the predicted values will deviate little from the observed A careful inspection of the deviance formula reveals several places where the deviance compares Y to $\hat{\lambda}$: the sign of the deviance is based on the difference between Y and $\hat{\lambda}$, and under the radical sign we see the ratio $Y/\hat{\lambda}$ and the difference $Y - \hat{\lambda}$. When $Y = \hat{\lambda}$, that is, when the model fits perfectly, the difference will be 0 and the ratio will be 1 (so that its log will be 0).

Plot in Figure 10 shows the deviance residuals versus predicted responses for the first-order model. We note that it exhibits curvature, supporting the idea that the model may improved by adding a quadratic term. Other details related to residual plots can be found in various sources including McCullagh and Nelder (1989).

5 Advanced Applications of the `glm` Function in R for Poisson Regression

The `glm` function in R is a powerful tool for implementing generalized linear models (GLMs), including Poisson regression. In this section, we explore its extended capabilities, including how to add multiple covariates, introduce interaction terms, and handle common issues such as overdispersion and zero-inflated data.

5.1 Adding Covariates and Interaction Terms

Adding covariates to a Poisson regression model allows us to control for additional factors that may influence the response variable. Furthermore, we can include interaction terms to explore whether one covariate's effect depends on another's level. See the code below for a practical example

Listing 5: Adding covariates and interactions

```
# Adding an interaction term between age and gender
adv_model <- glm(Total ~ Age + Age2 + Roof + Region + Roof:Region, family =
    poisson(link = "log"), data = db_ph)

# Summarize the model
summary(adv_model)
```

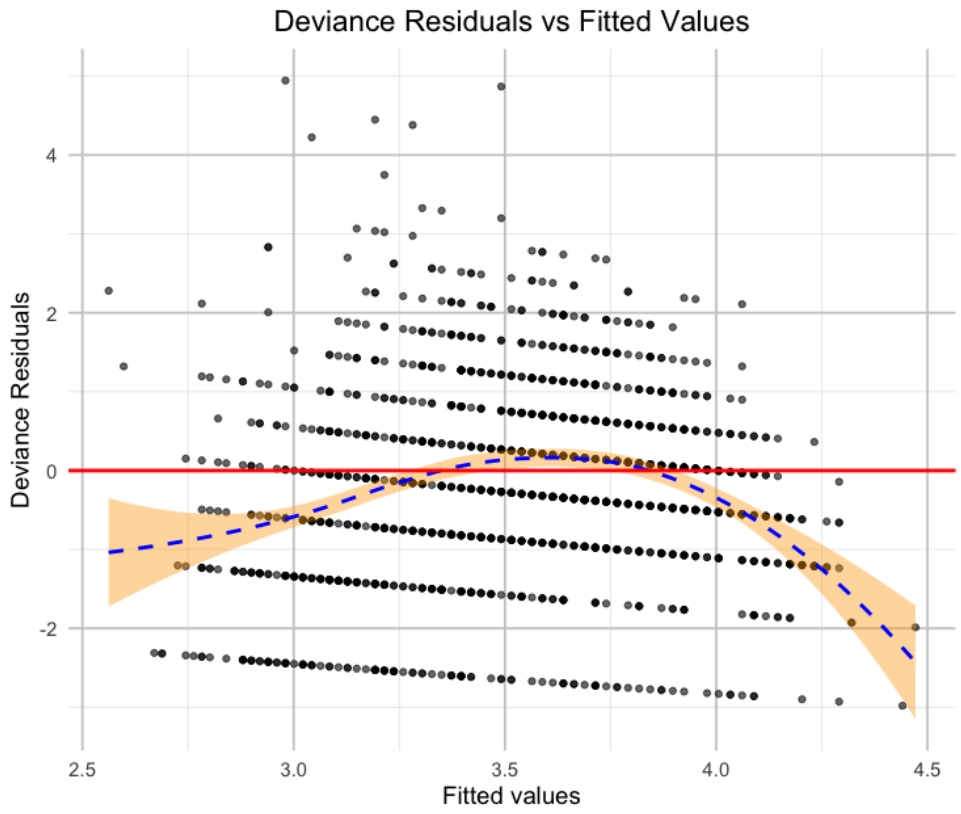

Figure 10: Residual plot for the Poisson model of household size by age of the household head.

5.2 Addressing Overdispersion

Overdispersion occurs when the variance of the response variable is greater than the mean, violating the assumptions of the Poisson distribution. One common approach to handle overdispersion is to fit a quasi-Poisson or a negative binomial model. The quasi-Poisson model adjusts the standard errors to account for overdispersion.

Listing 6: Quasi-Poisson distribution

```
# Fit a quasi-Poisson model
quasi_model <- glm(Total ~ Age + Region, family = quasipoisson(link = "log"),
    data = db_ph)

# Summarize the quasi-Poisson model
summary(quasi_model)
```

The negative binomial model provides an alternative approach, estimating an additional dispersion parameter.

Listing 7: Negative Binomial

```
# Load the MASS package for negative binomial regression
library(MASS)

# Fit a negative binomial model
nb_model <- glm.nb(Total ~ Age + Roof, data = db_ph)

# Summarize the negative binomial model
summary(nb_model)
```

5.3 Handling Zero-Inflated Data

Zero-inflated data occur when there are more zero counts than expected under a standard Poisson model. In such cases, we can use zero-inflated Poisson (ZIP) models, which combine a Poisson distribution with

a logistic model for predicting excess zeros. The `pscl` package in R allows us to fit ZIP models using the `zeroinfl()` function.

Listing 8: Zero-Inflated

```r
# Load the pscl package
library(pscl)

# Fit a zero-inflated Poisson model
zip_model <- zeroinfl(Total ~ Age + Roof | gender, data = db_ph, dist = "poisson")

# Summarize the ZIP model
summary(zip_model)
```

In the previous example the first part of the formula (`response ~ age + gender`) models the Poisson count process, while, the second part (`| gender`) models the logistic process for the excess zeros.

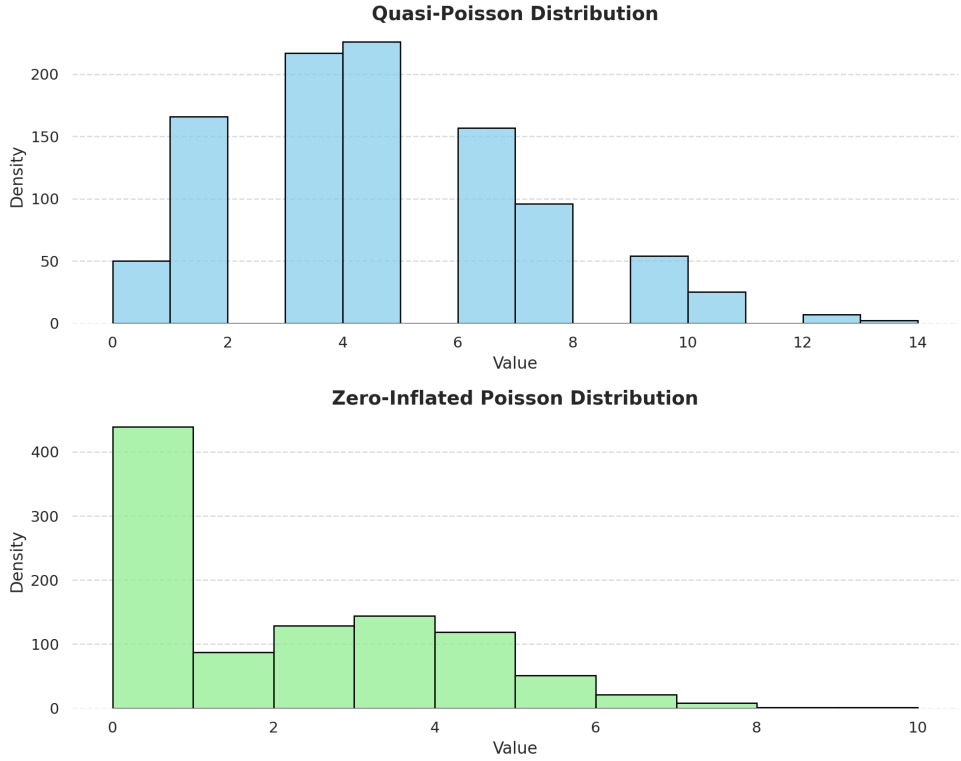

Figure 11: Distribution of values for Quasi-Poisson and Zero-Inflated Poisson models. The Quasi-Poisson distribution accounts for overdispersion, where the variance exceeds the mean, as shown by the broader spread of values. The Zero-Inflated Poisson distribution incorporates an excess probability of zeros, evident in the higher frequency of zero occurrences compared to the standard Poisson distribution.

Model selection is crucial in Poisson regression analysis. It is recommended to use goodness-of-fit tests, residual diagnostics, and information criteria such as AIC or BIC to compare models and select the most appropriate one. When interpreting coefficients, particular care should be taken in models that include interaction terms or zero-inflation components, as these can complicate the interpretation of the effects of individual covariates. Additionally, diagnosing overdispersion and zero inflation is an important step. Overdispersion can be checked by evaluating the ratio of residual deviance to degrees of freedom, while zero inflation can be detected by inspecting the number of zero counts in the data relative to the expected counts under a standard Poisson model.

By understanding and leveraging the flexibility of the `glm` function and related methods, we can build robust models that account for complexities in the data, including overdispersion and zero inflation, ensuring accurate and interpretable results.