Poisson Tutorial

Tina Yi

2023-11-16

1. **Overview:**

Generalized linear models (GLM) is a class of models in which the response variable Y is assumed to follow a distribution, which is assumed to be some (often nonlinear) function of $\mathbf{X}\beta$. The purpose of GLM is to build a linear relationship between the response and predictors through a link function.

Link Function connects the distribution of the outcome Y to $\mathbf{X}\beta$. It indicates how the expected value (mean) of the response relates to the linear combination of explanatory variables. It connects the mean of the distribution of the response variable to the predictors. The purpose of the link function is to allow GLMs to handle a variety of response variable distributions by transforming the predicted values to a suitable scale where a linear relationship with predictors can be assumed.

The General Structure of GLM is a linear relationship between the transformed expected response in terms of the link function and the explanatory variables: Link Function (Transformed expected value of the response) = $\mathbf{X}\beta$

The Poisson GLM is specifically designed for count data where the response variable represents the number of occurrences of an event within a fixed interval or space. The purpose of the Poisson GLM is to model discrete numeric random variables (count variables).

Two examples of research questions that could be answered using a Poisson GLM are:

- "What is the effect of distance to the nearest hospital on the number of ambulance calls received from different areas?" In this case, the count of ambulance calls could be modeled as a function of distance, with other covariates potentially included.
- "How does the presence or absence of a public park in a neighborhood affect the frequency of outdoor community events?" In this case, the count of events is the response variable, and the presence of a park is a binary predictor, possibly along with other socio-demographic factors.

2. **Probability Distribution**

The Poisson Distribution is assumed for the outcome in a Poisson regression model. This distribution is appropriate for modeling discrete numeric random variables (count variables), particularly for counts that represent the number of times an event occurs in a fixed interval of time or space. **The support** for the poisson distribution is the positive integers (includes 0). This means that the outcome variable can take on any non-negative integer value, reflecting the count nature of the data. **The parameter** for the poisson distribution is the lambda, the rate, the number of events on average in a given time frame, and the variance. It can take on any positive real number (> 0). **Parameter: rate=lambda=mean=variance (> 0)**

3. **The Model**

The general form of the Poisson GLM:

$$\log(\lambda_i) = \mathbf{X}\beta$$

We assume a Poisson distribution for the outcome:

$$y_i \mid x_i \sim \text{Poisson}\left(\lambda_i\right), i = 1, \ldots, n$$

We use a link function that ensures $\lambda_i > 0$ at any value of x_i:

$$\log\left(\lambda_i\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

The final Poisson model:

$$\log\left(E\left[y_i \mid x_i\right]\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, i = 1, \ldots, n$$

The link function is the natural logarithm In the context of a Poisson GLM. The log as the link function is appropriate for Poisson regression because it relates the linear predictors to the mean of the Poisson distribution, ensuring that the predicted mean count is always positive (>0).

Poisson Model Assumption:

- **Linearity in the Logarithmic Scale**: The relationship between the log of the expected counts (the mean of the Poisson distribution) and the linear predictors should be linear. This assumption can be checked by examining the deviance residuals from the model.
- **Mean=Variance**: A fundamental characteristic of the Poisson distribution is that the mean and variance are equal. This property can be evaluated through a dispersion test and by visually comparing the predicted counts with the observed counts in a graph. If the variance significantly exceeds the mean, the data exhibits overdispersion, suggesting that the Poisson model may not be the best fit.
- **Independence**: Each count or event should be independent of one another. This assumption is generally assured by the study design and is crucial for the model's estimates to be reliable.

4. **Data Example:**

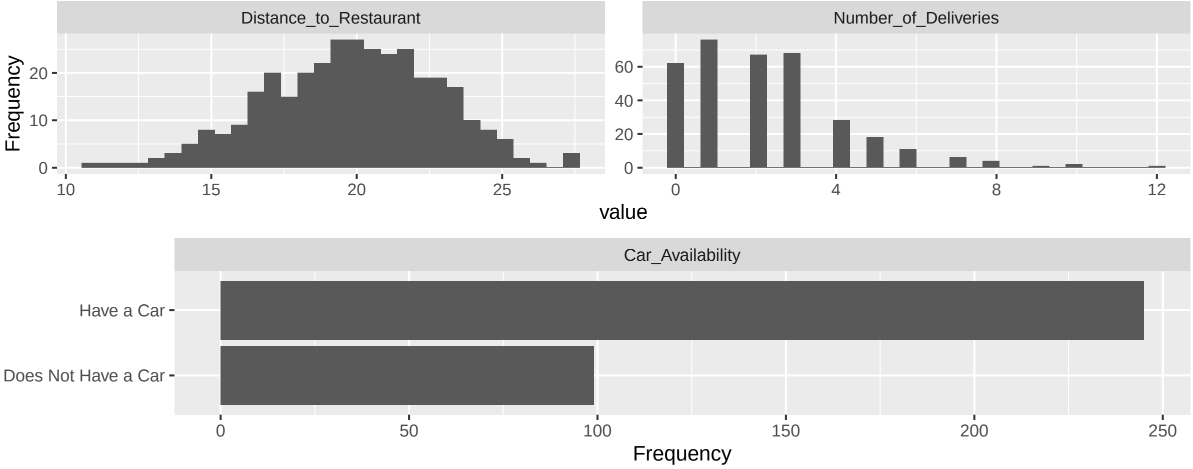
a. Introduction

Below is a simulated dataset generated using probability distribution functions in R, comprising three variables across 344 observations. The **research question** explores whether the distance to the nearest restaurant and the availability of a car affect the frequency of food delivery orders among urban dwellers. A **Poisson regression analysis** could be used to quantify how much of the variance in the number of deliveries can be explained by the distance to the closest restaurant and car availability.

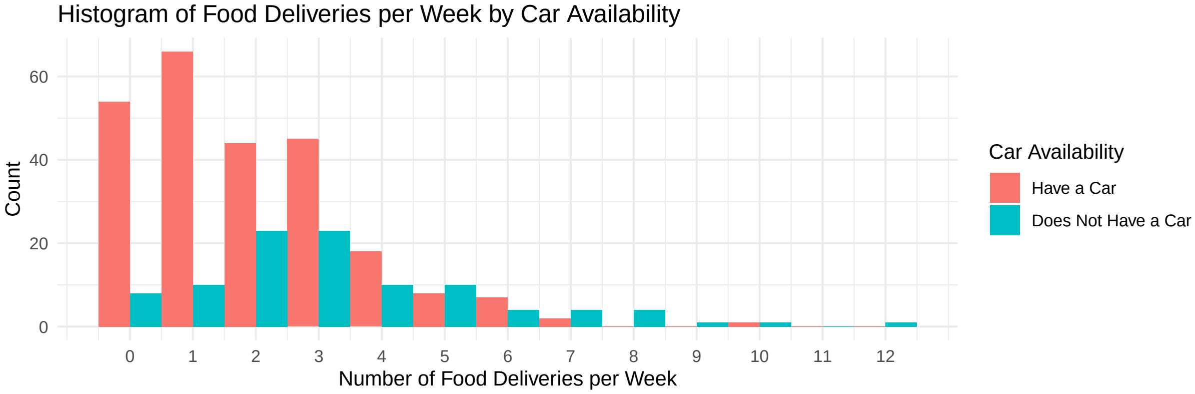
Table 1: Variable Descriptions

Variable	Type	Description
Y	Discrete Outcome	Number of Deliveries - The number of times an individual orders food delivery per week.
X1	Continuous Predictor	Distance to Restaurant - The distance from an individual's residence to the closest restaurant (measured in kilometers).
X2	Categorical Predictor	Car Availability - Whether the individual has a car or not, with 0 indicating 'Have a Car' and 1 indicating 'Does Not Have Car'.

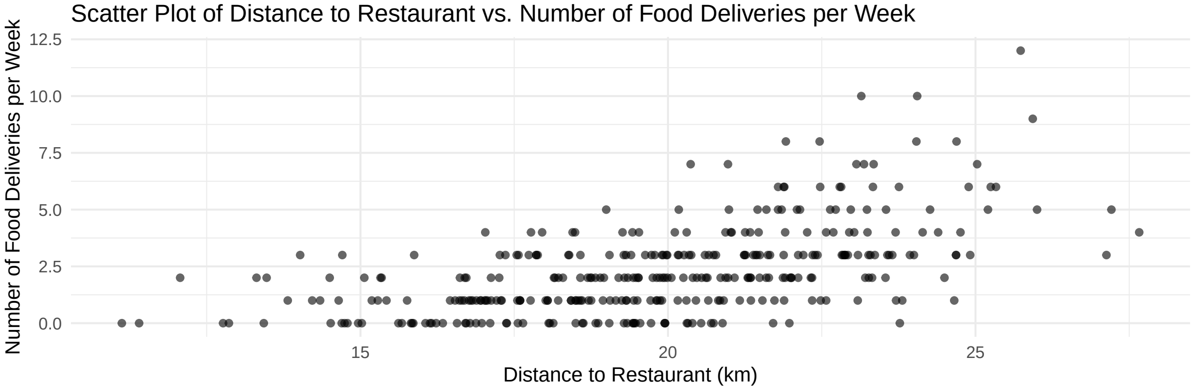
```
 Number_of_Deliveries Distance_to_Restaurant            Car_Availability
 Min.   : 0.000       Min.   :11.12         Have a Car         :245
 1st Qu.: 1.000       1st Qu.:17.86         Does Not Have a Car: 99
 Median : 2.000       Median :19.93
 Mean   : 2.317       Mean   :19.86
 3rd Qu.: 3.000       3rd Qu.:21.91
 Max.   :12.000       Max.   :27.66
```



Here is a histogram showing the food deliveries per week by car availability. It shows that people who have a car tend to order fewer number of food deliveries per week, while people who do not have a car tend to order higher number of food deliveries per week. There are a lot of 0 counts and other lower counts of food deliveries per week on the graph for people have a car. It is a typical shape of poisson distribution, with lots of lower counts and decreases from there.



Here is a scatter plot showing the relationship between distance from home to the closest restaurant and number of food deliveries per week. As distance from home to the closest restaurant increases, the number of food deliveries per week also tends to increase.



b. Fit the Model

Observing the patterns in these graphs, it becomes apparent that a standard linear regression approach is not suitable, primarily because the count data contains large number of zero values, which make log transformation inapplicable — the logarithm of zero is undefined. Consequently, to accommodate the zero counts and the discrete nature of the data, we opt for Poisson regression, a more appropriate method for modeling count data outcomes. We assume a poisson distribution for the count outcome and model the mean of that distribution.

```
# Fit Poisson regression model
poismod <- glm(Number_of_Deliveries ~ Distance_to_Restaurant+Car_Availability,
               family = poisson(), data = poisson)

# Display summary of the model
summary(poismod)
```

```

Call:
glm(formula = Number_of_Deliveries ~ Distance_to_Restaurant +
    Car_Availability, family = poisson(), data = poisson)

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -2.98497    0.27586 -10.821  < 2e-16 ***
Distance_to_Restaurant               0.17639    0.01278  13.807  < 2e-16 ***
Car_AvailabilityDoes Not Have a Car  0.54754    0.07192   7.613 2.67e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 619.27  on 343  degrees of freedom
Residual deviance: 362.70  on 341  degrees of freedom
AIC: 1148.9

Number of Fisher Scoring iterations: 5
```

The results from the Poisson Regression model suggest that the Standard Errors are reasonably estimated, with no evidence of excessive inflation. The model indicates that both the distance to the nearest restaurant and the lack of car ownership are significantly associated with the number of weekly food deliveries. The degrees of freedom correspond to 344 observations, confirming that there were no missing values, as initially established during the exploratory data analysis. This ensures that the full dataset contributed to the model estimation.

```
library(MASS)
cbind (exp (coef (poismod)) , exp (confint (poismod) ) , summary (poismod) $coefficient [, 4])
```

```

                                                     2.5 %     97.5 %
(Intercept)                         0.05054097 0.02931505 0.08644191
Distance_to_Restaurant              1.19290683 1.16349686 1.22325083
Car_AvailabilityDoes Not Have a Car 1.72899709 1.50073326 1.98970956

(Intercept)                         2.745181e-27
Distance_to_Restaurant              2.320517e-43
Car_AvailabilityDoes Not Have a Car 2.671218e-14
```

c. Coefficient Estimates Interpretation

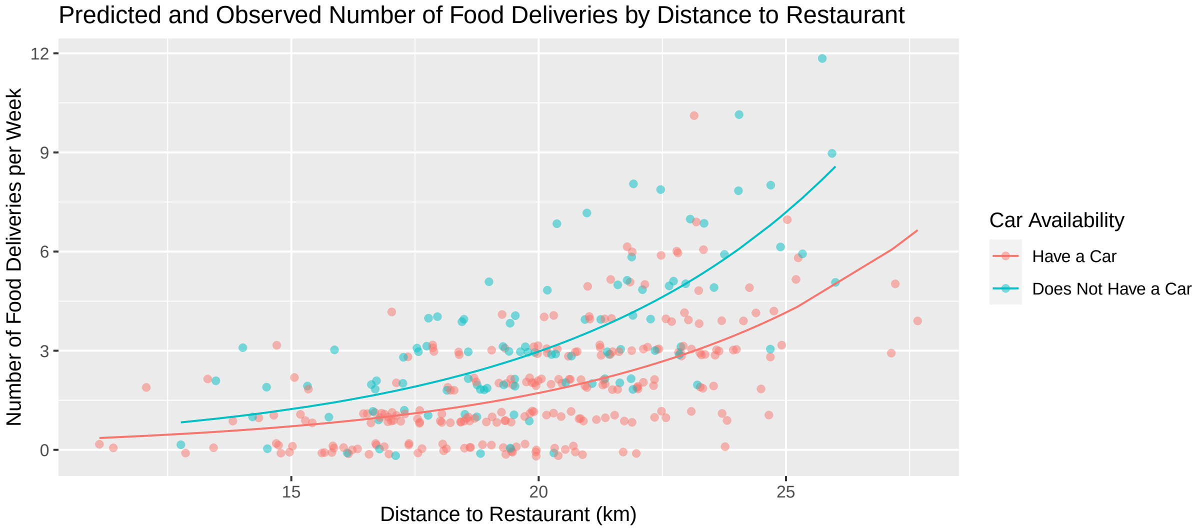
- Controlling the car availability, the expected number of food deliveries per week increases 19% per one unit (kilometer) increase in the distance from home to the closest restaurant. We are 95% confident that the true percent increase is between 16% and 22%. Distance to restaurant is statistically significantly associated with the number of food deliveries per week. (p<0.001)
- Controlling distance to restaurant, the expected number of food deliveries per week increases 72% for those do not have a car compared to those have a car. We are 95% confident that the true percent increase is between 5% and 98%. Do not have a car is statistically significantly associated with the number of food deliveries per week.(p<0.001)

d. Model Result

This plot shows the predicted and observed number of food deliveries per week by distance from home to the closest restaurant grouped by car availability. In this plot, the points show the observed number of food deliveries per week, and the line shows the predicted number of food deliveries per week. We can see that there's a lot of 0 counts for food deliveries per week for those who have a car. Our regression model shows that there is a significant difference between the number of food deliveries per week for those have a car and those do not have a car, which matches what we see on the graph. Those without a car tend to place more food delivery orders than those with a car. As distance from home to the closest restaurant increases, number of food deliveries per week also increases.

```r
library(tidyverse)
# Get predictions
poisson$preds <- predict(poismod, type = "response")

# Now create the plot
ggplot(poisson, aes(x = Distance_to_Restaurant, y = preds, colour = Car_Availability)) +
  geom_point(aes(y = Number_of_Deliveries),
             alpha = 0.5,
             position = position_jitter(h = 0.2)) +
  geom_line(aes(y = preds, group = Car_Availability)) +
  labs(x = "Distance to Restaurant (km)",
       y = "Number of Food Deliveries per Week",
       colour = "Car Availability") +
  ggtitle("Predicted and Observed Number of Food Deliveries by Distance to Restaurant")
```



e. Model Assessment

The assumption that is unique to poisson model is that **mean and variance are equal**. As the mean number of counts grows, so does the variance. If the assumption does not hold, the data exhibits overdispersion, suggesting that the Poisson model may not be the best fit, then we should consider other modeling options such as negative binomial regression. There are two ways to test this assumption:

- **Mehod 1: Use Dispersion Test to Assess Overdispersion**

H0: mean and variance are equal.

H1: mean and variance are not equal (true dispersion is greater than 1).

The following dispersion test for our poisson model shows P-value=0.8142. It fails to reject H0. Therefore, there aren't sufficient evidence that this poisson model is not good. The mean=variance assumption holds, so we can go with the poisson model.

```
library(AER)
dispersiontest(poismod)
```

```

    Overdispersion test

data:  poismod
z = -0.89364, p-value = 0.8142
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  0.936963
```

- **Method 2: Visualizing Prediction and Observe Graph**

However, it is not always great to solely assess the mean equals variance assumption based on the 0.5 cutoff using the overdispersion test. We should also visualize predictions and look at the distribution of outcomes to assess whether the mean equals variance assumption holds, whether the poisson model is appropriate.

From the above graph showing the predicted and observed number of food deliveries per week by distance from home to the closest restaurant grouped by car availability, we can observe that as the mean number of food deliveries per week goes up, so does the variance. When mean is close to 0, there is little variance. As mean increases a bit, variance also seems to spread a bit more. When we keeps moving along, as mean grows, the spread of the data (variance) also grows. Therefore, the mean equals variance assumption holds, indicating the poisson model is appropriate.

When assessing two poisson models, we can use deviance and AIC since they're based on the likelihood. Lower deviance and AIC indicates a better fitted poisson model.