# Poisson Regression Model

Maniyar Sai Rauf Sai Rafik
Roll No. : 2026
Class : M.Sc.-II

Department of Statistics, School of Mathematical Science, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon.

17/03/2022

# Contents

# Introduction :

- The Poisson regression model and the Negative Binomial regression model are two popular techniques for developing regression models for counts.
- Regression models for forecasting counts: Well look at Poisson regression model. The Negative Binomial (NB) regression model is another commonly used model for count based data.
- Poisson regression models are generalized linear models with the logarithm as the link function.
- Poisson regression creates proportional hazard models, one class of survival analysis: see proportional hazard models for descriptions of Cox models.

# Why Poisson Regression ?

- In simple/multiple linear regression dependent /response variable is of quantitative/continuous type.
- In logistics regression response variable is of binary type.
- But when dependent variable is of count type ( ordinal) taking small value.
- In this scenario we use Poisson Regression .
- Eg. Number of accident occurred at Aakashwani Chowk on NH6 highway per day.

# What is count data in Poisson Regression ?

- Count data are observations that assume only non-negative integer values: 0, 1, 2, etc. Count data have a Poisson distribution if the frequencies of the values have the following features:
- Small-valued observations are quiet common
- Starting at some value, frequencies decrease very rapidly
- The average of observations is approximately equal to their variance .

## Assumptions :

- The response variable y is of count type, which follows Poisson distribution.
- Errors are independent of each others.
- Average of observations is approximately equal to their variance.
  i.e. Mean $\approx$ Variance

# Poisson Regression Model :

- We assume that the study variable $y$ a count variable and follows a Poisson distribution with parameter $\lambda > 0$ as

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

- Note that the mean and variance of a Poisson random variable are same and related as

$$E(y) = Var(y) = \lambda.$$

Based on a sample $y_1, y_2, \ldots, y_n$ can write

$$E(y_i) = \lambda$$

- And express the Poisson model as

$$y_i = E(y_i) + \epsilon_i, i = 1, 2, \ldots, n$$

where $\epsilon_i's$ are random error terms .

- We can define a link function g that relates to the mean of study variable to a linear predictor as

$$\begin{aligned} g(\lambda_i) &= \eta_i \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, i = 1, 2, \ldots, k \\ &= x_i' \beta \end{aligned}$$

and

$$\lambda_i = g^{-1}(\eta_i)$$

$$= g^{-1}(x_i'\beta).$$

The identity link function is

$$g(\lambda_i) = \lambda_i = (x_i'\beta).$$

- The **log-link function** is

$$g(\lambda_i) = \log(\lambda_i) = (x_i'\beta)$$

$$\implies \lambda_i = g^{-1}(x_i'\beta) = \exp(x_i'\beta).$$

- Poisson Regression model is given by

$$E(y_i) = \lambda_i = e^{x_i'\beta}, \quad i = 1, 2, \ldots, n$$

## Parameter Estimation :

- Maximum likelihood estimation of parameters :

  We use the method of maximum likelihood estimation to estimate the parameters of the Poisson regression model.The likelihood function is based on Poisson distribution with parameter $\lambda$ and then $\beta$'s are estimated through the link function.

- The likelihood function of $y_1, y_2, \ldots, y_n$ is

$$L(y, \lambda) = \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} \frac{exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

$$= \frac{\prod_{i=1}^{n} \lambda_i^{y_i} exp(-\sum_{i=1}^{n} \lambda_i)}{\prod_{i=1}^{n} y_i!}$$

Thus, $\log L(y, \lambda) = \sum_{i=1}^{n} y_i log(\lambda_i) - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} y_i!$

The parameter $\lambda_i$ can be related to $\beta$'s through the link function

$$\lambda_i = g^{-1}(x_i'\beta)$$

After choosing the proper link function, the log-likelihood function can be maximized using some numerical optimization techniques for a given set of data. Let $\hat{\beta}$ be the obtained maximum likelihood estimator of $\beta$.

## Parameter Estimation :

- Then the fitted Poisson regression model is

$$\widehat{y_i} = g^{-1}(x_i'\widehat{\beta})$$

- In case of log-link,

$$\widehat{y_i} = g^{-1}(x_i'\widehat{\beta}) = exp(x_i'\widehat{\beta})$$

## Testing Of Hypothesis :

- The test of hypothesis is the case of Poisson regression model is similar to the case of the logistic regression model. It is constructed as model deviance which is based on a large sample test using the likelihood ratio test statistic.
- The model deviance is defined as

$$y^*(\beta) = 2log(saturated\ model) - log(\hat{\beta})$$

where the saturated model is based on all the p parameters of the model, and it fits the data perfectly.

-

## Testing Of Hypothesis :

- The statistic $y^*(\beta)$ has approximately $\chi^2_{(n-p)}$ distribution when n is large. The large value of $y^*(\beta)$ indicates that the model is not correctly fitted to the given data whereas small values of $y^*(\beta)$ indicating that model is well fitted to the given set of data in the sense that it is as good as the saturated model
- **Conclusion** :

  If $\lambda(\beta) \leq \chi^2_{(n-p,\alpha)} \implies$ The fitted model is adequate at $\alpha\%$ level of significance .

  and

  If $\lambda(\beta) > \chi^2_{(n-p,\alpha)} \implies$ The fitted model is not adequate at $\alpha\%$ level of significance .

# Some Statistical Packages used for Poisson Regression on different Softwares I

- Python : Poisson regression can be performed using the linear_model.**PoissonRegressor()** from sklearn module.
- MATLAB Statistics Toolbox: Poisson regression can be performed using the **glmfit** and **glmval** functions.
- Microsoft Excel: Excel is not capable of doing Poisson regression by default. One of the Excel Add-ins for Poisson regression is **XPost**.

# Some Statistical Packages used for Poisson Regression on different Softwares II

- R: The function for fitting a generalized linear model in R is **glm()**, and can be used for Poisson Regression.
- SAS: Poisson regression in SAS is done by using **GENMOD**.
- SPSS: In SPSS, Poisson regression is done by using the **GENLIN** command.
- STATA: STATA has a procedure for Poisson regression named **poisson**.

## Example :

- We have the in-built data set warpbreaks which describes the effect of wool type (A or B) and tension (low, medium or high) on the number of warp breaks per loom. Lets consider breaks as the response variable which is a count of number of breaks. The wool type and tension are taken as predictor variables.
- Input Data :

  > # Fit Poisson Regression Model by using RStudio
  > df = datasets::warpbreaks # In-build dataset
  > head(df)
  > tail(df)

# Example :

| | breaks | wool | tension | | breaks | wool | tension |
|---|---|---|---|---|---|---|---|
| 1 | 26 | A | L | 49 | 17 | B | H |
| 2 | 30 | A | L | 50 | 13 | B | H |
| 3 | 54 | A | L | 51 | 15 | B | H |
| 4 | 25 | A | L | 52 | 15 | B | H |
| 5 | 70 | A | L | 53 | 16 | B | H |
| 6 | 52 | A | L | 54 | 28 | B | H |

$>$ out $=$ glm(formula $=$ breaks $\sim$ wool $+$ tension, data $=$ df, family $=$ poisson)

$>$ summary(out)

Call:

glm(formula $=$ breaks $\sim$ wool $+$ tension, family $=$ poisson, data $=$ df)

## Example :

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -3.6871 | -1.6503 | -0.4269 | 1.1902 | 4.2616 |

Coefficients:

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 3.69196 | 0.04541 | 81.302 | <2e-16 *** |
| woolB | -0.20599 | 0.05157 | -3.994 | 6.49e-05 *** |
| tensionM | -0.32132 | 0.06027 | -5.332 | 9.73e-08 *** |
| tensionH | -0.51849 | 0.06396 | -8.107 | 5.21e-16 *** |

Null deviance: 297.37 on 53 degrees of freedom
Residual deviance: 210.39 on 50 degrees of freedom
AIC: 493.06
Number of Fisher Scoring iterations: 4

# Example

- In the summary we look for the p-value in the last column to be less than 0.05 to consider an impact of the predictor variable on the response variable. As seen the wooltype B having tension type M and H have impact on the count of breaks.

- The Null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean) whereas residual with the inclusion of independent variables.

- Above, we can see that the addition of 3 (53-50 =3) independent variables decreased the deviance to 210.39 from 297.37. Greater difference in values means a bad fit.

# References :


- Chapter 15 Poisson Regression Model, Shalabh, IIT Kanpur
- Dataset - Number of awards https://stats.idre.ucla.edu/stat/data/poisson_sim.csv
- warpbreaks dataset - https://cran.r-project.org/web/packages/greta/vignettes/example_models.html

*Thank You !*