Regression

In this session:

Dependent Variable	Regression Model
Continuous	Linear
Binary	Logistic
Multicategory (unordered) (nominal variable)	Multinomial Logit
Multicategory (ordered) (ordinal variable)	Cumulative Logit
Count variable	**Poisson Regression (Log-linear model)**

Poisson regression – Count data model - to predict count(+ integer) outcome

- ✓ Where it is applicable?
- ✓ Assumption of Poisson Regression
- ✓ Recap of Poisson distribution
- ✓ [redacted] Why Poisson Regression model is called log linear model?
- ✓ Process of developing Poisson Regression model.
- ✓ Extension of Poisson regression – negative binomial, over dispersed Poisson model, zero inflated Poisson model

Where is Poisson regression applicable?

Positive Integer outcome –

- The dependent variable is taking positive integer values like
  - ✓ Number of families visiting a restaurant in a given hour
  - ✓ Numbers of accident claims being made by car insurance holder in a year
  - ✓ Expected number of credit cards a customer may have?
  - ✓ Count of complete network failure in a year for a big size tech firm.
- Please note in all these cases, outcome can be
  - ✓ 0
  - ✓ 1
  - ✓ 2
  - ✓ .. Positive integer values only
- And most of the time dependent variable is smaller values like 0 , 1, 2 and there are few somewhat larger values
- So the average is small value.
- In such cases the Response variable Y has a Poisson distribution

Positive Integer outcome –

- If the average is large value, then for the response variable Y, the normal distribution is a good approximation for the Poisson distribution
- In Poisson regression - based on input variables, you tend to predict the count (i.e. dependent variable).
- In all such cases independent variable can be
  - ✓ Continuous
  - ✓ Continuous + dummy variable (categorical)
- And dependent variable is
  - ✓ Count variable

Assumption of Poisson regression

Assumption of Poisson Regression–

- The dependent variable Y is count
- Counts must be positive integers (i.e. 0,1,2,3...k).
- Counts must follow a Poisson distribution. Therefore, the mean = variance.
- Explanatory variables must be continuous, dichotomous or ordinal.
- Observation are independent of each other

- As we talked of Poisson distribution many a times, let's recap it quickly

In case of Poisson Distribution–

- The probability of dependent variable Y, taking value k (i.e. count = k) is given by

$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Where

- ✓ λ is the rate of event or average count of events per interval
- ✓ e=2.71828..
- ✓ k! = k*(k-1)*(k-2)* ……*1

An example of Poisson Distribution–

$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

- If defect rate for a manufacturer is 5%, then
- What is the probability that in a consignment of 100 TV
  - ✓ Q : There will be no defect at all?
  - ✓ A : here λ is 5
  - ✓ P(0) = $e^{-5} * \frac{5^0}{0!}$
  - ✓ Q : There will be one defect only?
  - ✓ A : P(1) = $e^{-5} * \frac{5^1}{1!}$

Process of Poisson Regression model development

Poisson Regression – summary and tips for interpretation of output

- If you have count as the dependent variable Y,
- Y'= Ln(Y) can be taken to develop a regression model (i.e. Poisson Regression)
- MLE is used to estimate the parameters of regression equation
- If Poisson regression gives Y'= Ln(Y) = a + bx
- Then please understand
  - ✓ If b = 0, then Y'= a i.e. Ln(Y) =a, hence expected / predicted Y = exp(a)
  - ✓ If b ≠ 0, then Y' i.e. Ln(Y) =a + bx , hence expected / predicted Y = exp(a + bx) = exp(a) * exp(bx) and
  - ✓ One unit of x increase, will change Y by exp(b) times (note multiplicative impact not the addition impact that happens in linear regression case)
  - ✓ If b > 0, then Y will become greater by exp(b) times
  - ✓ Like if b=2, then Y will become bigger with each unit increase of x, by exp(2) times
  - ✓ And similarly if b < 0 (like b=-2), then Y will decrease with each unit increase of x, by exp(-2) times

Practical challenges of Poisson Regression model development

- ✓ Some extensions of Poisson Regression
- ✓ Negative binomial
- ✓ Over dispersed Poisson Regression
- ✓ Zero inflated Poisson regression

Poisson Regression – practical challenges

- Basic assumption for Poisson regression : mean=variance

For std deviation of regression coefficient estimates

- When mean and variance are slightly different, the std error of parameter estimates should be obtained by robust procedure. This gives a narrower estimate of confidence interval for estimates.
- Variance > mean → Over dispersed data
- Dispersion parameter (simplistically) = variance / mean

Two alternative approach

- Smaller over dispersion – over dispersion Poisson Regression –
  - ✓ an extra parameter is included which estimates how much larger the variance is than the mean
  - ✓ This parameter estimate is then used to correct for the effects of the larger variance on the p-values
  - ✓ There are Test for Overdispersion by Cameron & Trivedi . If p value is > 0.05, one can apply Over dispersion Poisson regression

Poisson Regression – practical challenges

- Large over dispersion - Negative binomial –
  - ✓ It has one parameter more than the Poisson regression that adjusts the variance independently from the mean
  - ✓ The variation of this parameter can account for a variance of the data that is higher than the mean

Zero inflated Poisson Model

- Sometimes there are many, many more zeros than even a Poisson Model would indicate.
- This generally means there are two processes going on–there is some threshold that needs to be crossed before an event can occur.
- A Zero Inflated Poisson Model is a mixture model that simultaneously estimates
  - ✓ the probability of crossing the threshold (logistic model),
  - ✓ and once crossed, how many events (Poisson model)