# Poisson Regression

# Contents

# Understanding Poisson Distribution

- Suppose we wish to study the number of accidents taking place on a busy highway in a year. Number of accidents is a count variable and the event 'accident' is considered as rare event
- There are several phenomena where a variable is a 'count' and is observed in a specific time period; such as,
  - Number of deaths caused by lightning in six months
  - Number of visits to a dentist per year
- **Such random variables do not follow normal distribution** and hence cannot be modeled using multiple linear regression
- The probability distribution best suited for such data is **Poisson distribution** and the regression model is **Poisson regression**

The distribution is named after French mathematician Siméon Denis Poisson

# Understanding Poisson Distribution

Poisson distribution is a limiting case of Binomial distribution where

- n (Number of trials) is very large ( $n \to \infty$ ) and
- p (Probability of success) is very small ( $p \to 0$ ) such that
- np is finite ( say $\lambda$ )

In other words, **chance of a success is very small and trial is repeated large number of times**

The Probability Mass Function is :

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

Poisson distribution is specified with a single parameter $\lambda$ .

For Poisson distribution Mean = Variance = $\lambda$

# Poisson Regression

**DEPENDENT** VARIABLE

Count

Often it is the count of the rare event. Counts are all positive integers

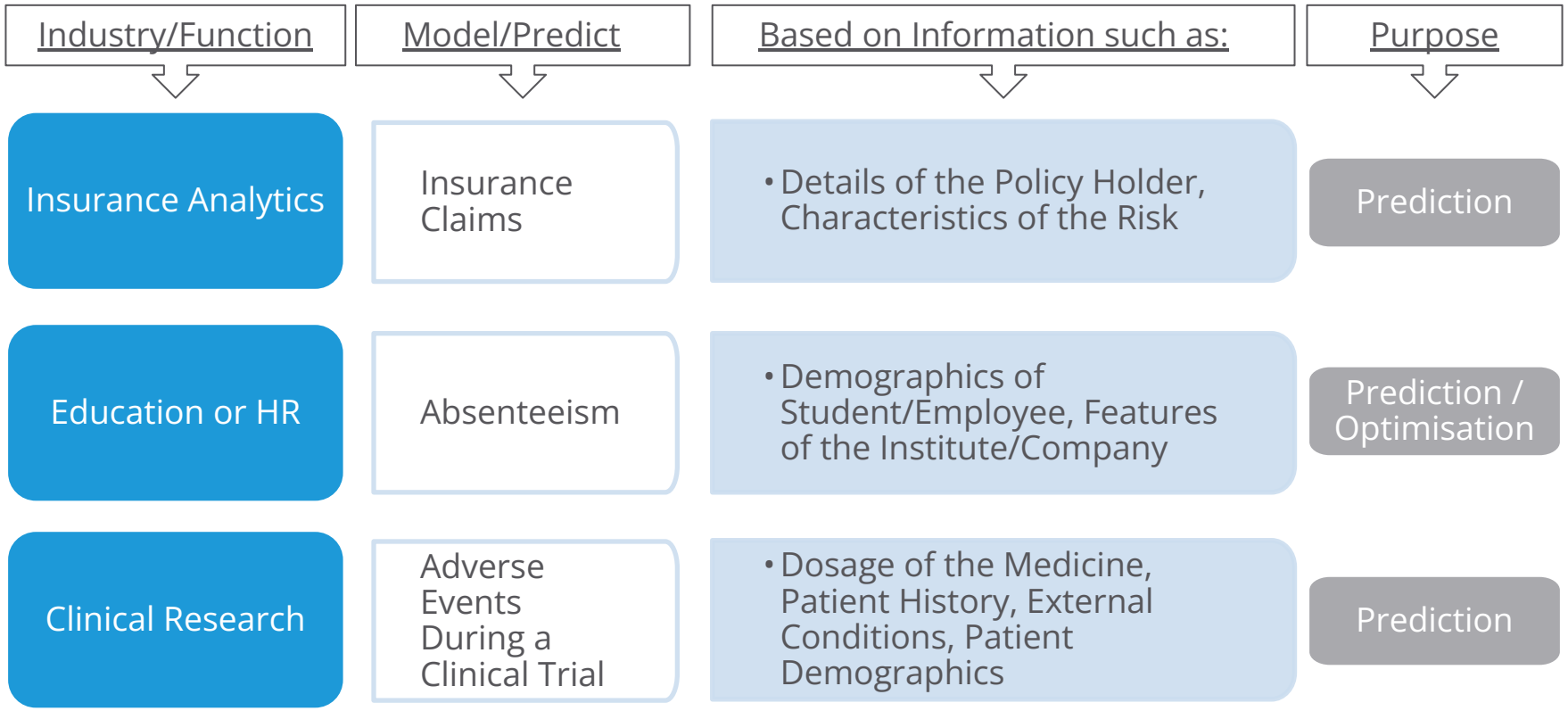
**INDEPENDENT** VARIABLE

Categorical or Continuous

Poisson regression is **most suitable in the case of rare events**

Poisson regression is a type of Generalised Linear Model, where the link function is a logarithm and the underlying distribution is Poisson

# Application Areas



# Statistical Model

$$\log(\lambda) = b_0 + b_1 x_1 + b_2 x_2 + \ldots\ldots + b_k x_k$$

Where, $\lambda$ is the conditional mean of Y given X

- Here ordinary least square method used in multiple linear regression is not appropriate as Y is discrete and RHS is continuous
- log(λ) is the link function used in Poisson Regression which establishes link between 'Y' and linear combination of X's
- Note that λ is greater than zero and its log will be negative if it lies between 0 and 1
- The regression coefficients are estimated using the method of maximum likelihood

# Case Study – Modeling Number of Complaints

## Background

- A company has recently launched a loyalty program under which they collected information about their customers. The next leg of the program aims to add 20,000 customers to their loyalty pool. The company wants to understand if the number of complaints by a customer can be modeled in order to set up a call centre with optimum strength.

## Objective

- To model number of complaints to prepare a road map for the call centre in the next leg of loyalty program

## Available Information

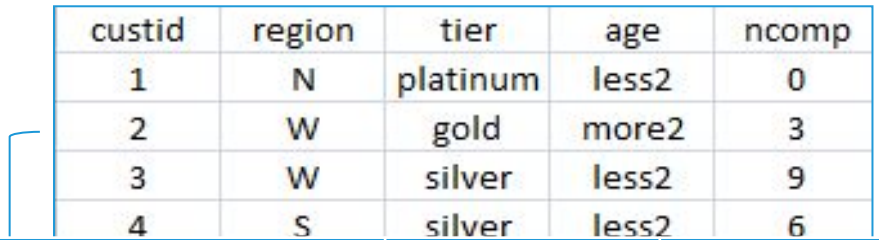
- Sample size is 113
- Information is available about **Region, Loyalty Tier, Complaints and Customer's Association with the Company**

# Data Snapshot

Complaints

Independent variables

Dependent variable



| custid | region | tier | age | ncomp |
|---|---|---|---|---|
| 1 | N | platinum | less2 | 0 |
| 2 | W | gold | more2 | 3 |
| 3 | W | silver | less2 | 9 |
| 4 | S | silver | less2 | 6 |

| Columns | Description | Type | Measurement | Possible values |
|---|---|---|---|---|
| custid | Customer ID | character | - | - |
| region | Region to which the customer belongs | categorical | E,W,N,S | 4 |
| tier | Loyalty program tier of the customer | categorical | platinum, gold, silver | 3 |
| age | Representing customer's association with the company | categorical | less2, more2 | 2 |
| ncomp | Number of complaints | Integer(count) | - | positive values |

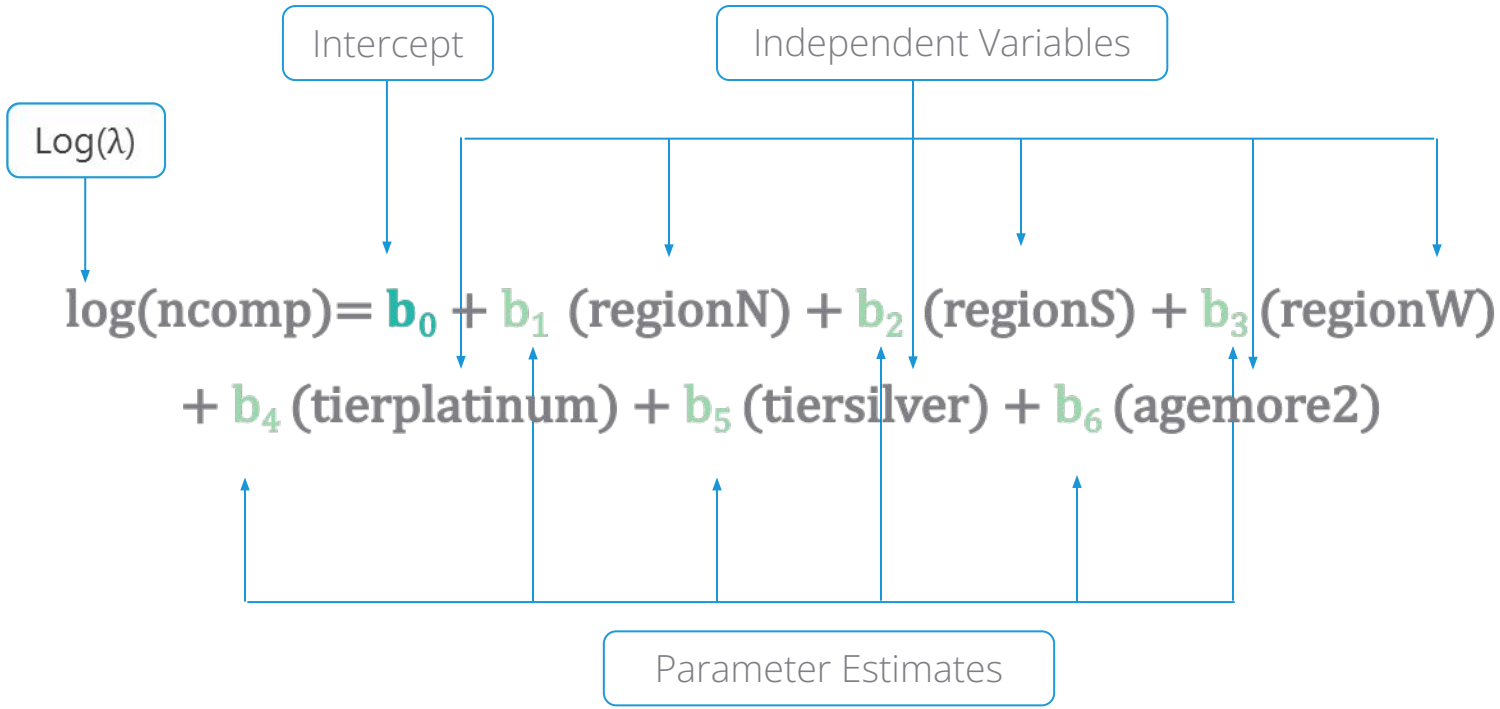
# Model for the Case Study

All independent variables (region, tier and age) in the case study are categorical

| Independent variables | Categories | Base category |
|---|---|---|
| region | East(E), West(W), North(N), South(S) | East(E) |
| tier | platinum, gold, silver | gold |
| age | less2, more2 | less2 |

# Model for the Case Study

The Poisson regression model is :



# Model Fitting in R

```
#Importing the Data
calldata<-read.csv("Complaints.csv",header=TRUE)

#Model Fitting
compmodel<-glm(formula=ncomp~region+tier+age,data=calldata,
               family='poisson')

summary(compmodel)
```

- **glm()** fits a generalised linear model.
- **family=poisson** ensures that a Poisson regression is used.

**summary()** yields model summary.

# Model Fitting in R

```
# Output
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.2919     0.1389   9.302  < 2e-16 ***
regionN       -0.1096     0.1420  -0.772 0.439968
regionS       -0.2286     0.1489  -1.535 0.124786
regionW       -0.4498     0.1613  -2.789 0.005290 **
tierplatinum  -0.6883     0.1754  -3.925 8.69e-05 ***
tiersilver     0.4410     0.1137   3.878 0.000105 ***
agemore2       0.1767     0.1077   1.641 0.100785
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 186.08  on 112  degrees of freedom
Residual deviance: 116.01  on 106  degrees of freedom
AIC: 456.86

Number of Fisher Scoring iterations: 5
```

**Interpretation:**

The **Estimate** column gives the estimates of coefficients of the independent variables in the model.

# Individual Testing in R

```
# Identifying significant variables
summary(compmodel)
# Output
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.2919     0.1389   9.302  < 2e-16 ***
regionN       -0.1096     0.1420  -0.772 0.439968
regionS       -0.2286     0.1489  -1.535 0.124786
regionW       -0.4498     0.1613  -2.789 0.005290 **
tierplatinum  -0.6883     0.1754  -3.925 8.69e-05 ***
tiersilver     0.4410     0.1137   3.878 0.000105 ***
agemore2       0.1767     0.1077   1.641 0.100785
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

## Interpretation:

- The p-values for **regionW, tierplatinum, tiersilver are <0.05**
- **regionW** is significant, with a negative coefficient : likelihood of complaints coming from West region is lower by -0.4498 compared to complaints from East region
- **tierplatinum** and **tiersilver** are significant, with a negative and positive coefficients respectively : compared to the base tier Gold, customers from Silver category tend to complain more whereas complaints from Platinum customers are the least

# Goodness of Fit

| | |
|---|---|
| **Objective** | To test the **null hypothesis** that **the model is a good fit** |

Null Hypothesis ($H_0$): Model is a good fit

Alternate Hypothesis ($H_1$): There is significant lack of fit

| | |
|---|---|
| **Test Statistic** | Pearson's chi − sq, $\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} \sim \chi^2_{(n-k)}$ <br> or <br> Deviance: $G^2 = 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i) \sim \chi^2_{(n-k)}$ |
| **Decision Criteria** | Reject the null hypothesis if p-value < 0.05 |

# Goodness of fit in R

```
#Goodness of Fit
s1<-summary(compmodel)
res_deviance<-s1$deviance
df<-s1$df.residual
pvalue<-1-pchisq(res_deviance,df)
pvalue
```

- creating an object s1 to store summary of **compmodel**
- storing residual deviance in **res_deviance**
- storing the corresponding degrees of freedom in **df**
- **pchisq()** calculates Chi-square value by using (**Residual**, **degrees of freedom**) as the arguments

```
# Output
[1] 0.2380154
```

**Interpretation**:
p-value > 0.05, Do not reject $H_0$.
The model can be considered to be a good fit .

# Predictions in R

```
#Predictions
calldata$ncomppred<-round(predict(compmodel,calldata,type='response'))
head(calldata)
```

- **predict()** requires model object, data and type.
- **type='response'** gives the predicted probabilities

```
# Output
```

| | custid | region | tier | age | ncomp | ncomppred |
|---|---|---|---|---|---|---|
| 1 | 1 | N | platinum | less2 | 0 | 2 |
| 2 | 2 | W | gold | more2 | 3 | 3 |
| 3 | 3 | W | silver | less2 | 9 | 4 |
| 4 | 4 | S | silver | less2 | 6 | 5 |
| 5 | 5 | E | silver | less2 | 7 | 6 |
| 6 | 6 | N | silver | less2 | 5 | 5 |

**Interpretation**:
The last two columns are observed and predicted vales of "ncomp"

# Introduction to Zero-Inflated Poisson Regression

- One common cause of over-dispersion is excess zeros, which in turn are generated by an additional data generating process. In this situation, zero-inflated poisson regression should be considered
- Zero-inflated models attempt to account for excess zeros. In other words, two kinds of zeros are thought to exist in the data, "true zeros" and "excess zeros". Zero-inflated models estimate two equations simultaneously, one for the count model and one for the excess zeros.

```
# Zero-Inflated Poisson Regression
install.packages("pscl")
library(pscl)

zip_model<-zeroinfl(formula= dependent variable~ independent
variables| variable causing zero inflation, data= data)

summary(zip_model)
```

# Offset Variable in Poisson Regression

- Poisson regression can be used to analyze **not only the count data but also the rate data**
- **Rates are simply counts divided by a measure like total count or time**. For ex. Insurance claim rate is measured as number of claims divided by the total number of the policy-holders (say N)
- The **log transformed regression variable** with the **constant coefficient of 1** for each observation is known as **'Offset'**
- The Poisson model is fit to the counts & **uses the log of the denominator** (usually N) **as an offset variable**
- Model in R using offset variable ( C is number of claims, CAR is vehicle type ,AGE is vehicle age and N is number of policies)

- claimmodel<-glm(formula=C~offset(log(N))+CAR+AGE,data=claimdata, family=poisson)

# Quick Recap

In this session, we learnt the basics of Poisson regression.

| | |
|---|---|
| Poisson Regression | • Is used to model count of a rare event |
| Model Building | • glm function is used to perform Poisson Regression<br>• Family="poisson" is used inside glm function |
| Parameter estimation | • Parameters are estimated by maximum likelihood estimation method |
| Check Variable Significance | • Hypothesis Testing |

# Quick Recap

| | |
|---|---|
| Measure Goodness of Fit and Predictions | • Check Residual Deviance and Degrees of Freedom as a measure of goodness of fit<br>• Deviance follows a chi-squared distribution, with degrees of freedom equal to the difference in the number of parameters<br>• Predict the count using the estimated model parameters |
| Zero-Inflated Poisson Regression | • Is used when there are excess zeros in the count which are generated by an additional data generating process |