Original Research Article

DISPERSION OF COUNT DATA: A CASE STUDY OF POISSON DISTRIBUTION AND ITS LIMITATIONS.

ABSTRACT

Poisson distribution is one of the widely known distribution in the field of probability and statistics by statisticians. It has been widely applied in modeling of discrete observations including but not limited to the number of customers in a shop within a specified period, the number of accidents occurring within a specified time or the number of claims experienced by an insurance company within a specified period of time. Poisson regression model has been widely used in events where one response variable is influenced directly by other independent variables. One thing about Poisson model is that it is strict on the property of dispersion as it assumes that count data is equidispersed which is not the case in practice. By this assumption, the Poisson model states that the variance of the count data is equal to the mean which is not practically true. In most cases, the variance of real count data is always greater than the mean, a phenomenon described as over dispersion. This gives Poisson model a loss in its frequent use in modelling count observations. This paper seeks to study the concept of dispersion, how Poisson regression is applied and its possible limitations. A deep study of Poisson model is done, its properties up to the fourth moments outlined. A graphical representation of its probability density function is drawn from simulated data and its shape noted under different rates as it resumes symmetry as the rate increases. A histogram is also presented. An application to real data is done in R programing language and proof that Poisson regression is very poor on this analysis given. Finally, a counter distribution appropriate for taking care of over dispersion is analyzed and results compared. AIC is used to conclude that NB is better than Poisson regression model.

Keywords: Poisson, Dispersion, binomial regression, overdispersion, equidispersion, maximum likelihood estimation

1.0 INTRODUCTION

1.1 The Poisson Process

A Poisson distribution is a counting distribution which has had many uses over time including modeling risk in insurance. Because of the simplicity of its properties, the Poisson distribution

has been applied to many areas especially where events occur at a certain rate with randomness (occurrence not certain)[16].

A simple point process $\varphi = \{t_n : n \geq 1\}$ is a defined discrete sequence of strictly increasing points $0 < t_1 < t_2 < t_3 < \cdots < t_n$ with condition that $t_n \to \infty$ as $n \to \infty$. The Poisson process is therefore a counting process with discrete intervals.

1.2 Poisson Process as a Limit Process

1.2.1. Binomial Process

Consider a random variable $Z_n \sim Binomial(n, p(n))$ with all n and p positive. Let us also consider another parameter $\alpha > 0$ ie taking positive values only as a fixed real number and hence the $\lim_{n\sim\infty} np = \alpha$. Therefore the probability mass function of Z_n will converge to a Poisson Process (α) probability mass function, as $n \to \infty$. That means that, for any $\emptyset \in \{0,1,2,3,4 \ldots\}$ we say that

$$\lim_{n\to\infty} PZ_n(\emptyset) = \frac{e^{-\alpha}\alpha^{\emptyset}}{\emptyset!} \qquad 1$$

This therefore shows that the Poisson process is a limiting process to Binomial,[16].

If we consider tossing a coin and recording on a specified duration of time, say t, that is $(0,t]$ and we let N (t) be the number of events that either heads or tails show up with the relation $n \approx \frac{t}{\partial}$ slots of time within the specified interval, N (t) will give us the number of heads within n number of flips of the coin. As a result, we have a conclusion that $N(t) \sim Binomial(n, p)$ where we define $p = \alpha\partial$. Therefore:

$$np = n\alpha\partial$$
$$= \frac{t}{\partial} * \alpha\partial$$
$$= \alpha t \qquad 2$$

We can freely conclude that as $\partial \to 0$, the probability mass function of $N(t)$ will exhibit convergence to a Poisson Distribution with a revised rate of αt.

2.0 THE POISSON DISTRIBUTION

2.1 Construction

We consider a time interval t and a specified number of events occurring within the interval, say n independent and identically distributed events. This means that the occurrence of one event does not necessarily influence by any means the chances of occurrence of another event in the series. We also suppose that the probability of occurrence of one event within a small change in time δt is given as

$$P(1: \delta t) = \lambda \delta t \quad (3)$$

Where λ is a constant.

The no event probability is then obtained as follows,

$$P(0: \delta t) = 1 - \lambda \delta t \quad (4)$$

And the probability of no event in time interval $t + \delta t$ is calculated as follows:

$$\frac{d}{dt} P(0: t) = -\lambda P(0: t) \quad (5)$$

By integration we obtain:

$$P(0: t) = Qe^{-\lambda t} \quad (6)$$

We then generalize the result through procedure of integration and substitution to find that $Q = \lambda t$ and therefore

$$P(1: t) = \lambda t e^{-\lambda t} \quad (7)$$

according to [5].

Then the probability mass function of a random variable X following a Poisson distribution is:

If $X \sim Poisson\ (\zeta)$ where ζ is the parameter and range $\chi = \{0,1,2,3 \ldots\}$, is given by

$$P_X(t) = \begin{cases} \frac{\zeta^t e^{-\zeta}}{t!}, & if\ t \in \chi \\ 0, otherwise \end{cases} \quad (8)$$

2.2 Properties of Poisson distribution

The important properties under study here include the mean and variance of Poisson distribution.

The mean of a Poisson random variable X is given by:

$$E[X] = \sum_{i=1}^{n} xf(x) = \zeta \quad (9)$$

While the variance of the Poison Random variable X is given as:

$$V[X] = \sum_{i=1}^{n} x^2 f(x) = \zeta \quad (10)$$

Therefore, it is evident that the mean and variance of a Poisson random variable is equal. This limits its applicability to equidispersed data which is rare to find.

Skewness is given as

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\zeta}{\zeta^{\frac{3}{2}}} = \zeta^{\frac{-1}{2}} \quad (11)$$

The Kurtosis is given as:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{3(1 + 3\zeta)}{\zeta^2} - 3 \quad (12)$$

$$= \frac{\zeta + 3\zeta^2 - 3\zeta^2}{\zeta^2}$$

$$= = \zeta^{-1}$$

The characteristic function is given as:

$$\Theta(t) = e^{\zeta(e^{it}-1)} \quad (13)$$

Finally, the moment generating fuction is given as:

$$M(t) = e^{\zeta(e^{t}-1)} \quad (14)$$

As given by [11] and [14].

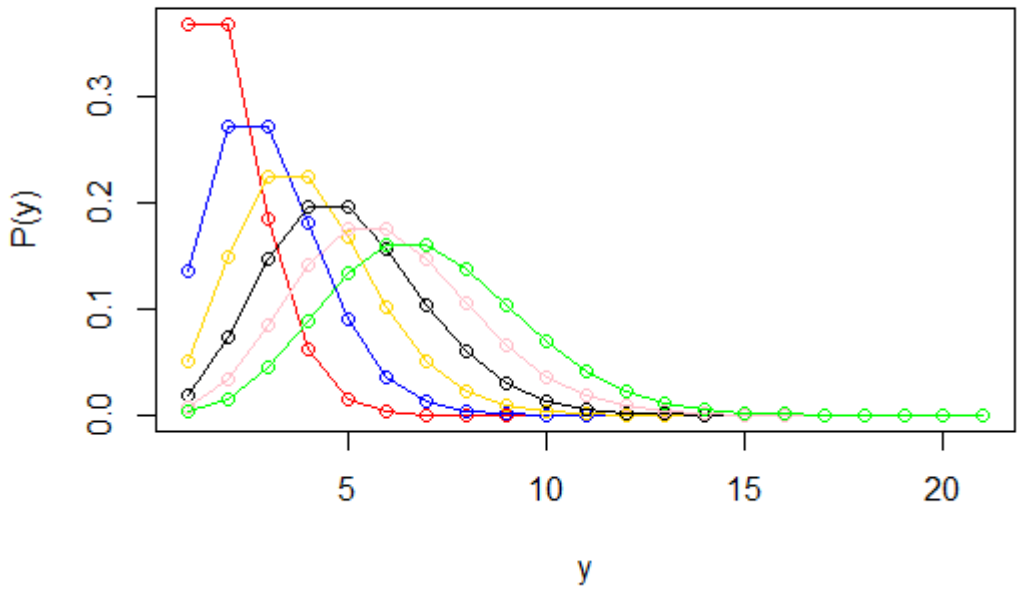

Fig. 1. The PDF of Poisson Distribution under different ζ from Red=1 down to Green=6. The graph becomes more symmetrical when the value of the parameter increases.

3.0 DISPERSION OF COUNT DATA

3.1 Meaning of Dispersion

We experience the occurrence of count data in many fields in lives today. These data refer to how many complete number of times an event occurs within a certain fixed period of time. This can include the number of times passengers arrive in train station for travels within a day or the number of patients admitted in a hospital within a period of one hour. Also, the number of accidents occurring in a certain road within a period of one hour. These are examples of count data because they are discrete data, [8].

In most cases, there is always the need to investigate the relationship between the variables in count data. We can limit our arguments to count data arising from the number of accidents occurring within a given time interval. We ask ourselves a question, how does reckless driving and condition of the vehicle cause accidents? In this case where there is presence of co-variates, we model using Poisson regression as shown:

$$\lambda_t(\tau) = \lambda_0(\tau)e^{X\beta} \quad (15)$$

Where the β are covariates.

Poisson regression is a classical regression model that belongs to a class of generalized linear models, [10]. As a result the Poisson regression models the conditional mean from the data available, count data, through a set of covariates as shown above.

Remarks.

Although Poisson model is used in modelling count data, its used is limited to be applied to real data only because of its practical limitation of exhibiting equidispersion property,[15]. This means that its mean and variance are equal which may not be true in a practical situation. Real data do not exhibit equality in the mean and variance, [8].

3.2 Overdispersion

In a Poisson process, where a random variable Y follows a Poisson distribution, overdispersion occurs when there is inequality in the first two moments. When the variance(Y) > Mean (Y). According to a Poisson process, this is not true because the variance and the mean of a Poisson random variable is equal, [3]. Normally, it is always considered that the rate ζ is a constant over the period of time in a Poisson regression for covariates. In the event that the rate becomes a function of time, a non-stationary Poisson process is experienced, [6]. Overdispersion can as well be caused by heterogeneity and contagion, [3]

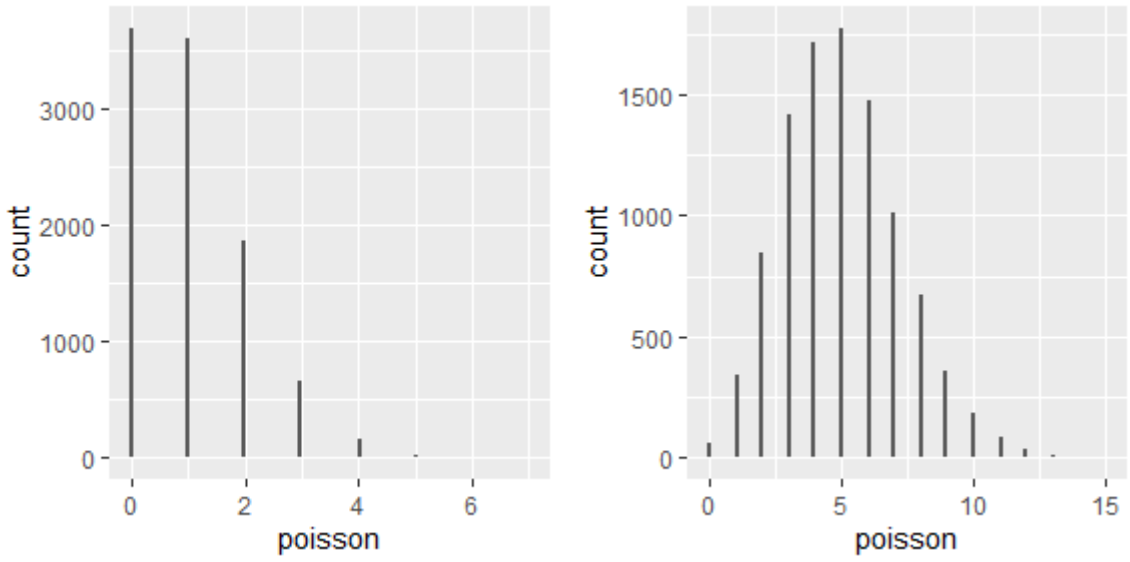

Fig. 2. (a) and (b) shows Poisson process count data with $\zeta = 1$ and $\zeta = 5$ in that order. As rate increases, the process becomes more symmetrical. (The two are generated simulated in R)

3.3. Underdispersion

In situations where the Poisson process random variable Y has mean smaller than the variance, we talk about underdispersion. Modelling such data with inappropriate models can give misleading results, [7].

4.0 Application to Real Life data and Remedies of Dispersion

Poisson distribution has been ineffective in the modelling of real count data because of errors associated with the final results as seen above. This demonstration has been done in R programming language with the following packages: stats, MASS, dplyr, ggplot2 and GlmSimulatoR. A generation of data following the distributions discussed is done using respective commands. Each graph is exported from the program for presentation.

4.1 Poisson Regression Model

Consider a sample of n observations of $y_1, y_2, y_3, y_4, \ldots$ in which the random variables follow a Poisson distribution, $y_i \sim P(\zeta_i)$ in which ζ_i is the mean in the i-th observation and is as well the variance which both depend on some variables, independent variables x. We then have a model

$$\zeta_i = x_i'\beta \qquad 16$$

Where the left hand side must be non-negative while the right hand side can take any real number. That makes it a disadvantage. Therefore to solve this problem, we introduce a link function log by taking the logarithm of the mean as follows:

$$\eta_i = \log(\zeta_i) \qquad 17$$

The logarithm transforms the mean of the Poisson regression model which becomes linear. Together with the link, the linear model obtained is therefore:

$$\log(\zeta_i) = x_i'\beta \qquad 18$$

This is the linear model that has β as coefficient and the x as the independent variables. An interpretation of this is that the coefficients represent the expected change in the $\log(\zeta_i)$ per unit change in x. the multiplicative model can be obtained by further solving equation 18 above by taking the exponents of both sides as follows:

$$\zeta_i = exp(x_i'\beta) \qquad 19$$

Therefore the coefficient exp (β) brings about a multiplicative effect which gives the model an advantage therefore solving our problem

We apply this regression analysis in R data package dataset on warpbreaks. This is a dataset containing three variables with breaks as the response variable and tension and wool as the independent variables. The data gives the number of warp breaks per loom which corresponds to a fixed length of yarn. A descriptive statistics is studied and glm fitted.

Estimation of Parameters.

For a model with a single covariate, the Maximum Likelihood estimation method of parameter estimation is used for both Poisson regression model and negative binomial regression model as follows:

The likelihood is obtained and maximized

We let:

$$\ell = \mathcal{L}(\beta) = \prod_{i=1}^{n} f(k)$$

$$\frac{\partial}{\partial \beta} ln\mathcal{L}(\beta) = 0 \qquad 20$$

We show that this derivative is maximum by taking the second derivative as follows:

$$\frac{\partial^2}{\partial \beta^2} ln\mathcal{L}(\beta) < 0 \qquad 21$$

Therefore the value of the covariate β is maximum.

The $\hat{\beta}$ is an unbiased estimator of β:

E[$\hat{\beta}$] = β and the variance is obtained as follows:

$$Var\left[\hat{\beta}\right] = \left(\frac{-\partial^2}{\partial \beta^2}\right)^{-1} \qquad 22$$

This is evaluated at $\beta = \hat{\beta}$

This is known as Cramer-Rao Lower Bound, [13]

The 95% confidence interval for the covariate β is thus obtained as follows:

$$\left[\hat{\beta} - 1.96\sqrt{Var\left[\hat{\beta}\right]}, \quad \hat{\beta} + 1.96\sqrt{Var\left[\hat{\beta}\right]}\right] \qquad 23$$

As outlined by [13].

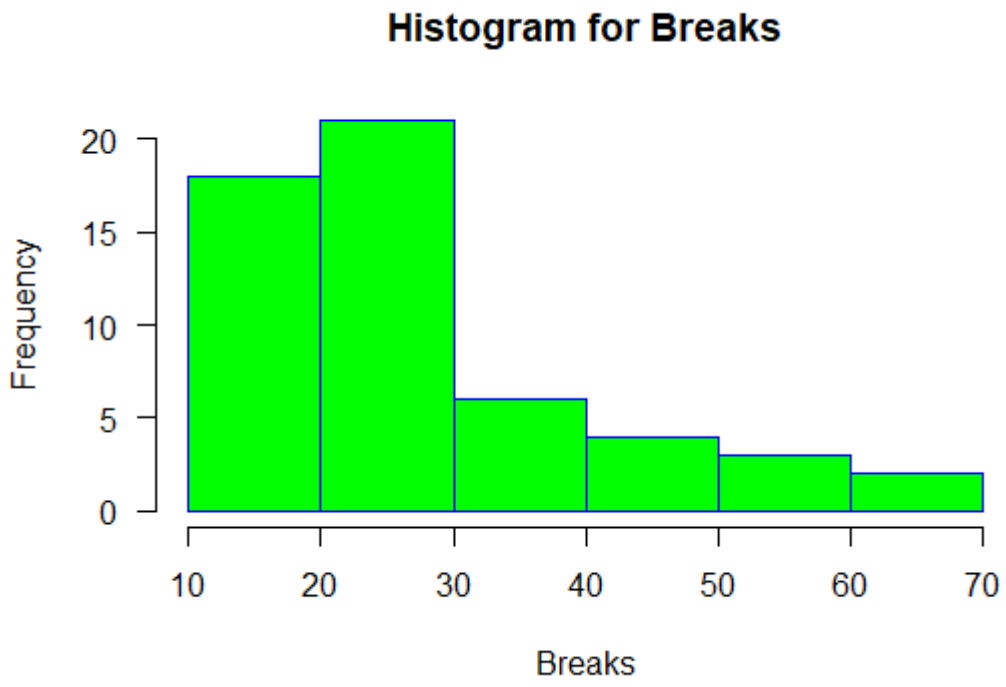

Fig. 3. A histogram of the number of Breaks.

Table 1. Poisson regression results still showing signs of data overdispersion. The ratios of deviance to degrees of freedom greater than 1.

```
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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    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
```

From the table, we see that obviously the model doesn't fit the data. The deviance and Pearson's Chi-Squared are both in 200's. To estimate the scale parameter by assuming that the variance is proportional and not equal to the mean, we obtain extra-Poisson variation as:

```
[1] 4.3 2.1
```

We see that the variance if far much larger than the mean.

4.2 Negative Binomial Regression Model

The reason why the negative binomial is used is because of the additional parameter. It is a generalization to Poisson regression model except for the presence of the extra parameter, θ and is used to model real data with overdispersion including the error/disturbance term, [1,2]. Therefore, the negative binomial is of the following form:

$$\log \zeta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \varrho\varepsilon_i \qquad 24$$

Where $\varrho\varepsilon_i$ is the error term or disturbance term, [4].

Table 2. Negative Binomial regression Results

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0144  -0.9319  -0.2240   0.5828   1.8220

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.6734     0.0979  37.520  < 2e-16 ***
woolB        -0.1862     0.1010  -1.844   0.0651 .
tensionM     -0.2992     0.1217  -2.458   0.0140 *
tensionH     -0.5114     0.1237  -4.133 3.58e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(9.9444) family taken to be 1)

    Null deviance: 75.464  on 53  degrees of freedom
Residual deviance: 53.723  on 50  degrees of freedom
AIC: 408.76

Number of Fisher Scoring iterations: 1


              Theta:  9.94
          Std. Err.:  2.56

 2 x log-likelihood:  -398.764
```

5.0 Interpretations of the Results.

5.1 The AIC

The Akaike Information criteria for the two models are as follows.

Table 3. Akaike Information Criteria for the models

S. No.	Model	AIC
1.	Poisson	493.06
2.	Negative Binomial	408.76

The AIC helps us to know which model among the two describes the pattern of the data well. Normally, the one with the lowest value is picked. As far as the AIC is concerned, the Negative Binomial regression model is better than Poisson regression model.

5.2. The Deviance

The deviance is divided into two, the null and the residual deviance. The null deviance tells us how much we can rely only on the intercept as a predictor of the response variable. The residual deviance shows how much we can rely on the intercept together with all the variables as predictors of the response variables. The higher the difference between the two the better.

Fisher scoring algorithm shows the number of iterations before the model stops.

6.0. Conclusion

As seen in this paper, the Poisson model assumes that the variance and the mean of count data is equal, meaning that the data is equidispersed. It has been proven through application and comparison that when Poisson distribution is used in modelling count data, very incorrect inferences is made with wrong conclusion as it seems to underestimates parameter. Akaike Information Criterion has also been used to provide this proof by comparing it with the binomial. Therefore when we prove this, we don't claim perfection. We therefore hope that other better models like NB model and its mixtures be used in modeling count data.

6.1. Recommendations

We therefore recommend that the study be expanded to provide more regression models that best captures the count data with minimum possible error. These distributions can be obtained through mixtures of other heavy tailed and flexible models that can provide better fit for the data.

REFERENCES

[1]. Agresti A, (1990). Categorical data analysis. New York: John Wiley & Sons.

[2]. Allison P, (1999). Logistic regression using the SAS system: theory and application. Cary (NC): SAS Institute.

[3]. Barron, D. N. (1992). The analysis of count data: Overdispersion and autocorrelation. *Sociological methodology*, 179-220.

[4]. Byers, A. L., Allore, H., Gill, T. M., & Peduzzi, P. N. (2003). Application of negative binomial modeling for discrete outcomes: a case study in aging research. *Journal of clinical epidemiology*, *56*(6), 559-564.

[5]. Cowan, G. (2009). Derivation of the Poisson distribution. *Royal Holloway, University of London. Dec*.

[6]. Cox, D. R., & Isham, V. (1980). *Point processes* (Vol. 12). CRC Press.

[7]. Harris, T., Yang, Z., & Hardin, J. W. (2012). Modeling underdispersed count data with generalized Poisson regression. *The Stata Journal*, *12*(4), 736-747.

[8]. Klakattawi, H. S., Vinciotti, V., & Yu, K. (2018). A simple and adaptive dispersion regression model for count data. *Entropy*, *20*(2), 142.

[9]. Landefeld CS, Palmer RM, Kresevic DM, et al, (1995). A randomized trial of care in a hospital medical unit especially designed to improve the functional outcomes of acutely ill older patients. N Engl J Med; 332:1338–44.

[10]. Nelder, J.A.; Wedderburn, R.W. Generalized linear models. J. R. Stat. Soc. Ser. A 1972, 135, 370–384.

[11]. Papoulis, A., & Probability, R. V. (1984). Stochastic process. *New York*.

[12]. Tippett, L. H. C. (1950) Technological Applications of Statistics. Wiley. Page 106

[13]. Van den Bos, A. (1994). A Cramér-Rao lower bound for complex parameters. *IEEE Transactions on Signal Processing [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on], 42 (10)*.

[14]. Weisstein, E. W. (2003). Poisson distribution. *https://mathworld. wolfram. com/*.

[15]. Rasch, G. (1960). ON GENERAL LAWS AND THE MEANING OF MEASUREMENT IN. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, June 20-July 30, 1960* (Vol. 2, p. 321). Univ of California Press.

[16]. Pishro-Nik, H. (2016). Introduction to probability, statistics, and random processes.

APPENDIX

The R codes for both regression analysis for the two models are enclosed here

For the Poisson and Negative Binomial regression:

```
library(datasets)
data <- warpbreaks
data
ls.str(warpbreaks)
hist(data$breaks, main = "Histogram for Breaks", xlab = "Breaks", border = "Blue",
     col = "Green", las=1, breaks = 5)

poisson <- glm(breaks ~ wool + tension, data, family = poisson(link = "log"))
summary(poisson)
m1 <- glm.nb(breaks ~ wool + tension, data, data = warpbreaks)
qchisq(0.95, df.residual(poisson))
deviance(poisson)
pr <- residuals(poisson, "pearson")
pr
sum(pr^2)
phi <- sum(pr^2)/df.residual(poisson)
round(c(phi, sqrt(phi)),1)
library(foreign)
library(brglm2)
library(brglm)
library(MASS)
m1<- glm.nb(breaks ~ wool + tension, data = warpbreaks)
sum(m1)
summary(m1)
```

For the Poisson probability mass function with increasing values of \zeta (ζ)

```
colors<- c("Red","Blue","Gold","Black","Pink","Green")
poisson.dist <- list()
a <- c(1,2,3,4,5,6)
for (i in 1:6) {poisson.dist[[i]] <- c(dpois(0:20, i))}


plot(unlist(poisson.dist[1]), type = "o", xlab="y", ylab = "P(y)",
     col = colors[i])
for (i in 1:6) {
  lines(unlist(poisson.dist[i]), type = "o", col = colors[i])
```