Chapter 7

Summary of Results and Conclusion

This chapter provides a brief summary of results arrived and discussed in the preceding chapters and also provides the advantages and limitation of these results along with the scope for future study.

The principal objective of this thesis was to provide some contributions to the development of the methodology for modeling the count data. We initially focused on both theoretical and applications for analyzing the numerical (frequency) count data with excess number of zero counts. Subsequently, we extended the idea to the multipoint inflation and for categorical data. In this thesis, we proposed and extend few count models for analyzing both numerical as well as for categorical count data. Usually Poisson distribution is used for analyzing the numerical count data whenever the data has mean and variance are equal. The flexibility of the negative binomial distribution for modeling the count data; it is frequently applied over other standard discrete probability distributions. But, there are situations in the count data which contains excess number of certain counts especially zero count and it causes over dispersion. As a result, the standard count models such as Poisson and negative binomial distribution does provides better fit to over-dispersed count data. Therefore, the researchers started using mixed probability models for modeling over dispersed count data. Though these mixed probability distributions give relatively better fit compared to standard models, it does not provides good enough results due to the presence of particular count is much more higher than the other counts. Therefore, it opened many research problems to the researchers and become gateway for innovative ideas for modeling over dispersed count data. As a result, many inflated probability distributions came into existence for modeling over dispersed count data. This gave direction and motivation to us to do some contribution in research area of modeling count data particularly for over dispersed data.
We have extended zero inflated version of mixture of negative binomial distribution with one parameter Lindley distribution, two parameter Lindley distribution and Sushila distribution. We have derived the basic statistical measures for all three proposed zero inflated distributions. We have performed data analysis to compare the efficiency of these newly proposed zero inflated models with respect to log-likelihood, Chi-square and $p$-value and we got the results of ZINB-L distributions provides better fit compared to Poisson, NB and NB-L distributions. ZINB-TPL fits better than NB, ZINB, NB-L and NB-TPL. ZINB-S distribution provides better fit to over dispersed data than Poisson, NB, NB-S distributions. Further, we have performed the data analysis to study the efficiency of two point inflated negative binomial - Lindley distribution and we observed the result that it provides better fit than NB, NB-L and SPINB-L distributions.

It is universally accepted fact that the efficiency of the probability models depends on accurate estimation of its parameters through a function of sample observations (i.e., estimator). In case of zero inflated distributions, there exists an additional parameter and it is named as inflation parameter. Therefore, it is necessary to estimate the inflation parameter more accurately for zero inflated probability distributions. In this thesis, we have proposed two new estimators for estimation of inflation parameter of ZIP distribution called probability based inflation estimator (PBIE) and for inflation parameter of ZINB distribution called non-zero probability estimator (NZPE). Further we have compared the efficiency of PBIE with its corresponding MLE for different values of inflation rate using MSE and relative efficiency and arrived at the result that PBIE performs relatively better than its corresponding moment estimator and MLE.
We also discussed and considered some modeling techniques for count data which includes classical and Bayesian approach. In classical approach, we considered GLM for modeling the count data. Further, we made use of a non parametric method called ANN which is a data dependent and dynamic adoptive learning approach. While comparing the results of ANN, GLM and Bayesian methods for modeling count data, the results showed that ANN performs better in almost all different settings.

We considered zero inflated Poisson regression model, hurdle Poisson regression model and ANN for modeling over dispersed count data. Further, we compared suitability of these three models via a simulation study as well as actual data study for different inflation rates and different sample sizes. The accuracy of these three models is compared in terms of MSE, SE and bias. It is observed from both studies, we arrived at the conclusion that ANN is invariably more suitable model for modeling over dispersed count data.

For modeling zero inflated count data, the need for selection of a particular model based on its performance among the set of available models is important in this information era. Moreover, it is highly difficult and time consuming process to compare the performance of these models whenever the data is voluminous. Hence the need for selection of suitable model for given inputs based on past experience will be much useful. For this purpose, we have given an algorithm for model selection for modeling count data with the help of different classification techniques namely discriminant analysis, CART and random forest.

The material presented in this report has emphasized the need of modelling count data whenever one or more count is inflated and give many useful tips and tricks for modeling over dispersed count data.
Scope for Future Study

This thesis confined its study only with over dispersed count data mainly for zero inflated data which is over dispersed and right skewed but this study can be extended to other type of count data as well. There are many other count models available in the literature left unexplored about its mixture with other suitable distribution and its zero inflated version. These are some of other potential area of future research.