COUNT DATA ANALYSIS USING POISSON REGRESSION AND HANDLING OF OVERDISPERSION

RAIHANA BINTI ZAINORDIN

UNIVERSITI TEKNOLOGI MALAYSIA

COUNT DATA ANALYSIS USING POISSON REGRESSION AND HANDLING OF OVERDISPERSION

RAIHANA BINTI ZAINORDIN

A dissertation submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

NOVEMBER 2009

To Mak and Abah

because I could not have made it without you both.

And to sunflower

because life should be meaningful.

ACKNOWLEDGEMENT

First and foremost, all praise be to Allah, the Almighty, the Benevolent for His blessings and guidance for giving me the inspiration to embark on this project and instilling in me the strength to see that this thesis becomes a reality.

I would like to say a big thank you to my amazing supervisor, P.M. Dr. Robiah binti Adnan for believing in me to finish up this dissertation. Her advice, encouragement and help mean a lot to me. So I can never thank her enough.

I would also like to thank my friend Ehsan who is never tired of helping and teaching me things that I do not know.

As always, I am forever grateful to my parents for all their love and sacrifice. To my sisters (Nordiana, Noraiman Amalina, Nurul Aina), thanks for giving me the reason to laugh and cry.

A heartfelt thanks goes to all my friends who are always willing to help and support me. And finally to my best friend, Ikhwan, thank you so much for being there.

ABSTRACT

Count data is very common in various fields such as in biomedical science, public health and marketing. Poisson regression is widely used to analyze count data. It is also appropriate for analyzing rate data. Poisson regression is a part of class of models in generalized linear models (GLM). It uses natural log as the link function and models the expected value of response variable. The natural log in the model ensures that the predicted values of response variable will never be negative. The response variable in Poisson regression is assumed to follow Poisson distribution. One requirement of the Poisson distribution is that the mean equals the variance. In real-life application, however, count data often exhibits overdispersion. Overdipersion occurs when the variance is significantly larger than the mean. When this happens, the data is said to be overdispersed. Overdispersion can cause underestimation of standard errors which consequently leads to wrong inference. Besides that, test of significance result may also be overstated. Overdispersion can be handled by using quasi-likelihood method as well as negative binomial regression. The simulation study has been done to see the performance of Poisson regression and negative binomial regression in analyzing data that has no overdispersion as well as data that has overdispersion. The results show that Poisson regression is most appropriate for data that has no overdispersion while negative binomial regression is most appropriate for data that has overdispersion.

ABSTRAK

Data bilangan adalah sangat lazim dalam pelbagai bidang, contohnya bidang sains bioperubatan, kesihatan awam dan bidang pemasaran. Regresi Poisson digunakan secara meluas untuk menganalisis data bilangan. Regresi Poisson juga sesuai untuk menganalisis data kadaran. Regresi Poisson merupakan sebahagian daripada model kelas model linear teritlak. Regresi ini menggunakan logaritma asli sebagai fungsi hubungan. Regresi ini memodelkan nilai jangkaan bagi pembolehubah maklum balas. Logaritma asli digunakan untuk memastikan supaya nilai ramalan bagi pembolehubah maklumbalas tidak akan berbentuk negatif. Pembolehubah maklumbalas dalam regresi Poisson dianggap mengikut taburan Poisson. Salah satu ciri taburan Poisson ialah nilai min pembolehubah adalah sama dengan nilai varians. Walaubagaimanapun, dalam aplikasi sebenar, data bilangan sering mempamerkan masalah lebih serakan. Masalah lebih serakan terjadi apabila nilai varians melebihi nilai min. Apabila ini terjadi, sesebuah data itu dikatakan terlebih serak. Masalah lebih serakan boleh menyebabkan kurang anggaran terhadap sisihan piawai yang kemudiannya memberi inferens yang salah. Selain daripada itu, keputusan ujian signifikan pula akan terlebih anggar. Masalah lebih serakan boleh diatasi dengan menggunakan kaedah kebolehjadian quasi dan juga regresi binomial negatif. Kajian simulasi telah dibuat untuk melihat keputusan regresi Poisson dan regresi binomial negatif dalam menganalisis data yang tidak mempunyai masalah lebih serakan dan juga data yang mempunyai masalah lebih serakan. Keputusan menunjukkan bahawa regresi Poisson adalah paling sesuai untuk data yang tidak mempunyai masalah lebih serakan manakala regresi binomial negatif adalah paling sesuai untuk data yang mempunyai masalah lebih serakan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Y	Response variable
X	Predictor variable
β	Regression coefficient
η	Link Function
$\mathbf{I}$	Information matrix
$\mathbf{U}$	Score fuction
$\mathbf{W}$	Weight matrix
E	Elasticity
X^2	Pearson chi-squares statistic
D	Deviance statistic
R	Pearson residual
Z	Wald statistic
ϕ	Dispersion parameter

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Count Data

An event count refers to the number of times an event occurs, for example the number of individuals arriving at a serving station (e.g.: bank teller, gas station, cash register, etc.) within a fixed interval, the number of failures of electronic components per unit of time, the number of homicides per year, or the number of patents applied for and received. In many fields such as in social, behavioral and biomedical sciences, as well as in public health, marketing, education, biological and agricultural sciences and industrial quality control, the response variable of interest is often measured as a nonnegative integer or count.

Significant early developments in count models, however, took place in actuarial science, biostatistics, and demography. In recent years these models have also been used extensively in economics, political science, and sociology. The special features of data in their respective fields of application have fueled developments that have enlarged the scope of these models. An important milestone in the development of count data regression model was the emergence of the generalized linear models, of which the Poisson regression is a special case.

In another case, an event may be thought of as the realization of a point process governed by some specified *rate of occurrence* of the event. The number of events may be characterized as the total number of such realizations over some unit of time. The dual of the event count is the *inter-arrival time*, defined as the length of the period between events. Count data regression is useful in studying the occurrence rate per unit of time.

The approach taken to the analysis of count data sometimes depends on how the counts are assumed to arise. Count data can arise from two common ways:

i) Counts arise from a direct observation of a point process.
ii) Counts arise from discretization of continuous latent data.

In the first case, examples are the number of telephone calls arriving at central telephone exchange, the number of monthly absences at workplace, the number of airline accidents, the number of hospital admissions, and so forth. The data may also consist of inter-arrival times for events.

In the second case, consider the following example. Credit rating of agencies may be stated as AAA, AAB, AA, A, BBB, B, and so forth, where AAA indicates the greatest credit. Suppose one codes these as $y = 0,1,..., m$. These are pseudocounts that can be analyzed using a count regression. But one may also regard this as an ordinal ranking that can be modeled using a suitable latent variable model such as ordered probit.

Typically, the characteristic of count data is that the counts occur over some fixed area or observation period and that the things that people count are often rare. Count data, even though numeric, can create some problems if it is analyzed using the regular linear regression because of the limited range of most of the values and because only nonnegative integer values can occur. Thus, count data can potentially result in a highly skewed distribution, one that cut off at zero. Therefore, it is often unreasonable to assume that the response variable and the resulting errors have a normal distribution, making linear regression a less appropriate option for analysis. A suitable way to deal with count data is to use Poisson distribution and log link function in the analysis. The regression model that uses these kinds of options is called Poisson regression or Poisson log-linear regression model.

Basically, the most popular methods to model count data are Poisson and negative binomial regression models. But Poisson regression is the more popular of the two and is applied to various fields.

1.2 Statement of the Problem

Count data often have variance exceeding the mean. In other words, count data usually shows greater variability in the response counts than one would expect if the response distribution truly were Poisson. This violates the Poisson regression assumption which strictly states that the mean is equal to the variance (equidispersion). The phenomenon where the variance is greater than the mean is called overdispersion. A statistical test of overdispersion is highly desirable after running a Poisson regression. Ignoring overdispersion in the analysis would lead to underestimation of standard errors and consequent of significance in hypothesis testing. The overdispersion must be accounted for by the analysis methods appropriate to the data. Poisson regression is not adequate for analyzing overdispersed data. Therefore, to overcome overdispersion, quasi-likelihood method will be used as well as negative binomial regression. Negative binomial regression is more adequate for overdispersed data. This is because negative binomial regression allows for overdispersion since its variance is naturally greater than its mean.

1.3 Objectives of the Study

The objectives of this study are:

i) To study the analysis of Poisson regression.
ii) To illustrate Poisson regression by analyzing count data manually and by using SAS 9.1.
iii) To demonstrate how to handle overdispersion in Poisson regression using quasi likelihood approach as well as negative binomial regression approach.
iv) To see the performance of Poisson regression and the performance of negative binomial regression in analyzing data that has no overdispersion as well as data that has overdispersion from simulation study.

1.4 Scope of the Study

This study will focus on the analysis of Poisson regression. This study will also focus on the overdispersion problem that exists when dealing with real life count data. Overdispersion happens when the variance is greater than the mean which violates the equidispersion property in Poisson distribution and thus need to be taken care of. In accordance to overdispersion problem, the performance of Poisson regression and negative binomial regression in analyzing data that has no overdispersion as well as data that has overdispersion will be examined from simulation study. The analyses in this study include manual analysis and analysis by using statistical package. Statistical package that is used in this study is SAS 9.1.

1.5 Significance of the Study

This study will help the scientists to realize the use of Poisson regression in analyzing count data. Besides focusing on parameter estimation, this study will also help to highlight about the interpretation of coefficients. This study will also help to overcome overdispersion problem that occurs in Poisson regression which, if ignored, may cause underestimation of standard errors and which consequently gives misleading inference about the regression parameters. Clearly, this study is imperative and will give much benefit.

1.6 Outline of the Study

This dissertation consists of 6 chapters.

Chapter 1 gives rough idea about the study. It begins with the explanation on count data. This includes the characteristic of count data which is very important throughout the study. Chapter 1 also explains how the idea about the study came about. Furthermore, it also explains about the purpose of the study, the scope and the importance of the study.

Chapter 2 discusses the basic idea that is important in Poisson regression analysis. This chapter also discusses about common problems in Poisson regression as well as negative binomial regression other than previous studies done by previous researchers.

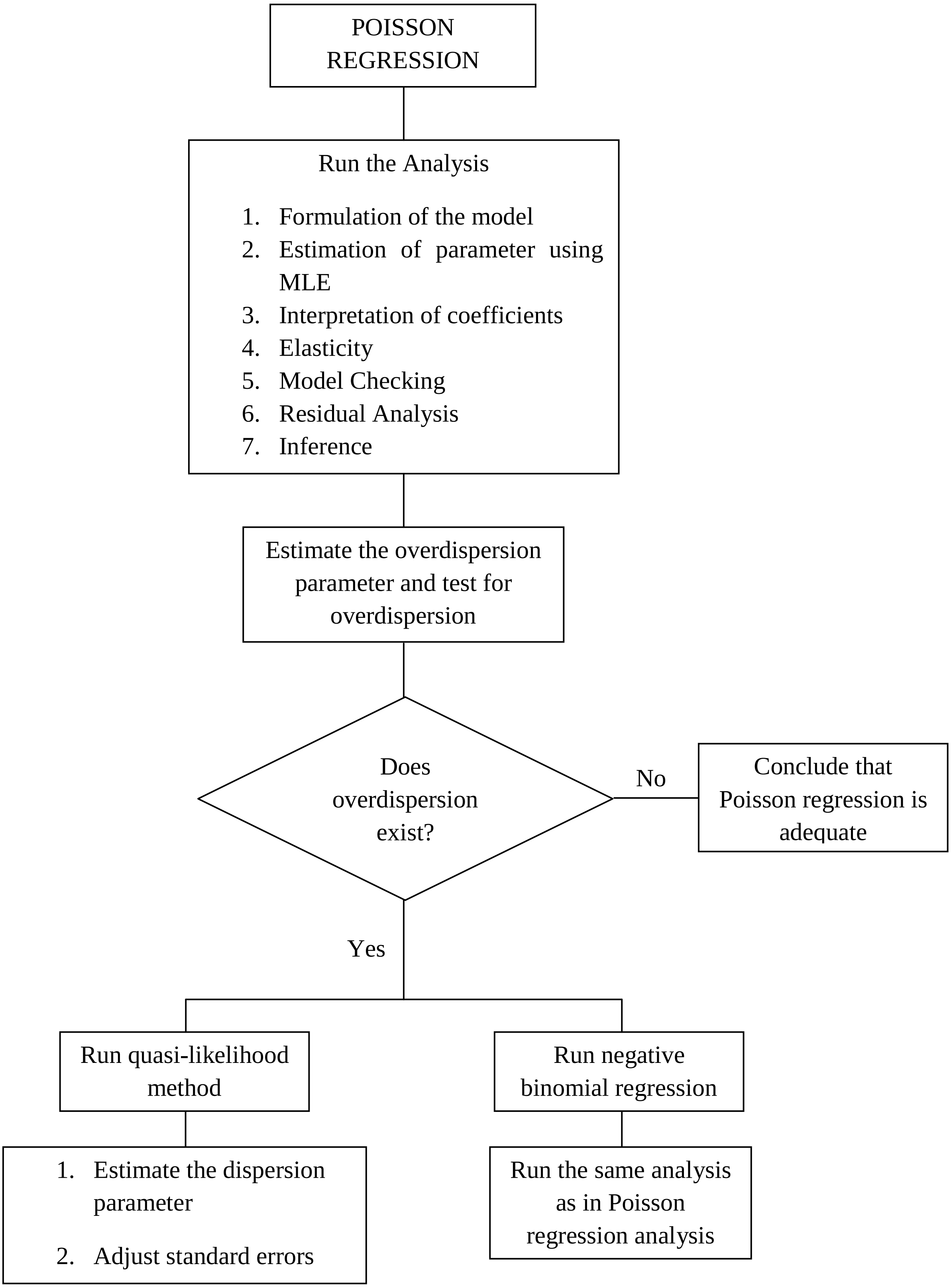
Poisson regression analysis can be found in Chapter 3. This chapter gives clear descriptions on formulation of Poisson regression model, manual computation of maximum likelihood estimates, and how to interpret coefficients in Poisson regression. It also includes other important analyses such as goodness of fit test, residual analysis and inference. Other than that, this chapter also discusses about the methods to handle overdispersion. To illustrate Poisson regression, an example is presented here. The analysis of this example is done manually.

Chapter 4 deals with the analysis of Poisson regression using SAS 9.1. A bigger data is used and more factors are considered. The data is a count data in the form of rate and it involves overdispersion. SAS codes are provided for convenience.

Chapter 5 presents the simulation study. Data is simulated using R 2.9.2 software and is analyzed by using SAS 9.1. The performance of Poisson regression and the performance of negative binomial regression in analyzing data that has no overdispersion as well as data that has overdispersion are presented in this chapter.

Lastly, the conclusions of the study are discussed in Chapter 6. This chapter summarizes the whole study. Some recommendations for further research are also made here.

1.7 Analysis Flow Chart



REFERENCES

Agresti, A. 1996. *An Intorduction to Categorical Data Analysis.* New York: John Wiley & Sons, Inc.

Atkins, D. C. and Gallop, R. J. 2007. Rethinking How Family Researchers Model Infrequent Outcomes: A Tutorial on Count Regression and Zero-Inflated Models. *Journal of Family Psychology.* Vol. 21, No. 4: 726 – 735.

Bailer, A. J., Reed, L. D., and Stayner, L. T. 1997. Modeling Fatal Injury Rates Using Poisson Regression: A Case Study of Workers in Agriculture, Forestry, and Fishing. *Journal of Safety Research.* 28(3): 177 – 186.

Cameron, A. C. and Trivedi, P. K. 1998. *Regression Analysis of Count Data.* Cambridge, UK: Cambridge University Press.

Carrivick, P. J. W., Lee, A. H., and Yau, K. K. W. 2003. Zero-Inflated Poisson Modeling to Evaluate Occupational Safety Interventions. *Safety Science*. 41: 53 – 63.

Chan, Y.H. 2005. Log-linear Models: Poisson Regression. *Singapore Med. J.* 46(8): 377 – 386.

Choi, Y., Ahn, H., and Chen, J. J. 2005. Regression Trees for Analysis of Count Data with Extra Poisson Variation. *Computational Statistics & Data Analysis*. 49: 893 – 915.

Dietz, E. and Bohning, D. 2000. On Estimation of the Poisson Parameter in Zero Modified Poisson Models. *Computational Statistics & Data Analysis.* 34: 441 – 459.

Dobson, A.J. 2002. *An Introduction to Generalized Linear Models*. 2nd Edition. New York: Chapman & Hall.

Dossou-Gbete, S. and Mizere, D. 2006. An Overview of Probability Models Statistical Modelling of Count Data. *Monografias del Seminario Matematico Garcia de Galdeano.* 33: 237 – 244.

Famoye, F. and Wang, W. Censored Generalized Poisson Regression Model. *Computational Statistics & Data Analysis.* 46: 547 – 560.

Fleiss, J.L., Levin, B. and Paik, M.C. 2003. *Statistical Methods for Rates and Proportions*. 3rd Edition. New York: John Wiley & Sons, Inc.

Frome, E.L. 1986. Regression Method for Binomial and Poisson Distributed Data. *The American Institute of Physics, New York.* 1 – 40.

Guangyong, Z. 2003. A modified Poisson Regression Approach to Prospective Studies with Binary Data. *American Journal of Epidemiology*. Vol. 159, No.7: 702 – 706.

Guo, J. Q. and Li, T. 2002. Poisson Regression Models with Errors-in-Variables: Implication and Treatment. *Journal of Statistical Planning and Inference*. 104: 391 – 401.

Heinzl, H. and Mittlbock, M. 2003. Pseudo R-squared Measures for Poisson Regression Models with Over- or Underdispersion. *Computational Statistics & Data Analysis*. 44: 253 – 271.

Ismail, N. and Jemain, A.A. 2005. Generalized Poisson Regression: An Alternative For Risk Classification. *Jurnal Teknologi*. 43(C): 39 – 54.

Jahn-Eimermacher, A. 2008. Comparison of the Andersen – Gill Model with Poisson and Negative Binomial Regression on Recurrent Event Data. *Computational Statistics & Data Analysis*. 52: 4959 – 4997.

Jovanovic, B. D. and Hosmer, D. W. 1997. A Simulation of the Performance of C_p in Model Selection for Logistic and Poisson Regression. *Computational Statistics & Data Analysis*. 23: 373 – 379.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E., and Nizam, A. 1998. *Applied Regression Analysis and Other Multivariable Methods*. 3rd Edition. USA: Duxbury Press.

Kohler, M. and Krzyzak, A. 2007. Asymptotic Confidence Intervals for Poisson Regression. *Journal of Multivariate Analysis.* 98: 1072 – 1094.

Kokonendji, C.C., Demetrio, C,G,B, and Dossou-Gbete, S. 2004. Overdispersion and Poisson-Tweedie Exponential Dispersion Models. *Monografias del Seminario Matematico Garcia de Galdeano.* 31: 365 – 374.

Kutner, M.H., Nachtsheim, C.J., and Neter, J. 2004. *Applied Linear Regression Models.* Fourth Edition. New York: McGraw-Hill.

Lee, J, Nam, D., and Park, D. 2005. Analyzing the Relationship Between Grade Crossing Elements and Accidents. *Journal of Eastern Asia Society for Transportation Study*. Vol. 6: 3658 – 3668.

Lee, Y., Nelder, J.A., Pawitan, Y. 2006.*Generalized Linear Models with Random Effects, Unified Analysis via H-likelihood.* USA: Chapman & Hall.

Liu, J. and Dey, D. K. 2007. Hierarchical overdispersed Poisson Model with Macrolevel Autocorrelation. *Statistical Methodology.* 4: 354 – 370.

Lord, D., Washington, S. P., and Ivan, J. N. 2005. Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis and Prevention*. 37: 35 – 46.

Luceno, A. 1995. A Family of Partially Correlated Poisson Models for Overdispersion. *Computational Statistics & Data Analysis*. 20: 511 – 520.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models.* 2nd Edition. London: Chapman & Hall.

Norliza binti Adnan (2006). *Comparing Three Methods of Handling Multicollinearity Using Simulation Approach.* Master of Science (Mathematics). Universiti Teknologi Malaysia, Skudai.

Osgood, D.W. 2000. Poisson-Based Regression Analysis of Aggregate Crime Rates. *Journal of Quantitative Criminology*. Vol. 16, No. 1: 21 – 43.

Pedan, A. Analysis of Count Data Using the SAS System. *Statistic, Data Analysis, and Data Mining*. Paper 247-26: 1 – 6.

Pradhan, N. C. and Leung, P.S. 2006. A Poisson and Negative Binomial Regression Model of Sea Turtle Interactions in Hawaii's Longline Fishery. *Fisheries Research*. 78: 309 – 322.

Spinelli, J. J., Lockhart, R. A., and Stephens, M. A. 2002. Tests for the Response Distribution in a Poisson Regression Model. *Journal of Statistical Planning and Inference*. 108: 137 – 154.

Strien, A.V., Pannekoek, P., Hegemeijer, W. and Verstrael, T. 2000. A Loglinear Poisson Regression Method to Analyse Bird Monitoring Data. Proceeding of the International Conference and 13th Meeting of the European Bird Census Council. 33 – 39.

Tang, H, Hu, M, and Shi, Q. 2003. Accident Injury Analysis for Two-Lane Rural Highways. *Journal of Eastern Asia Society for Transportation Study*. Vol. 5: 2340 – 2443.

Tsou, T. S. 2006. Robust Poisson Regression. *Journal of Statistical Planning and Inference*. 136: 3173 – 3186.

Vacchino, M. N. 1999. Poisson Regression in Mapping Cancer Mortality. Environmental *Research Section A*. 81: 1 – 17.

Wang, K., Lee, A. H., and Yau, K. K. W. and Carrivick, P. J. W. 2003. A Bivariate Zero-Inflated Poisson Regression Model to Analyze Occupational Injuries. *Accident Analysis and Prevention*. 35: 625 – 629.

Wang, Y., Smith, E. P., and Ye, K. 2006. Sequential Designs for a Poisson Regression Model. *Journal of Statistical Planning and Inference.* 136: 3187 – 3202.