Mixed-Effects Negative Binomial Regression with Interval Censoring: A Simulation Study and Application to Precipitation and All-Cause Mortality Rates among Black South Africans over 1997-2013

By

Christian Landon
Master of Public Health

Environmental Health

____

Matthew O. Gribble, Ph.D. D.A.B.T.

Committee Chair

Mixed-Effects Negative Binomial Regression with Interval Censoring: A Simulation Study and Application to Precipitation and All-Cause Mortality Rates among Black South Africans over 1997-2013

By

Christian Landon
B.S.
California State University, Fullerton
2016

Thesis Committee Chair: Matthew O. Gribble, Ph.D. D.A.B.T.

An abstract of
A thesis submitted to the faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health in Environmental Health
2019

Key Words: droughts; epidemiologic methods; errors, outcome measurement; multilevel analysis; vital statistics.

ABSTRACT

In research using epidemiological surveillance data, counts of health outcomes are often censored in order to protect privacy when the nonzero number of health outcomes occurring in specific times and places is small. Several common approaches to modeling such censored, hierarchically structured, over-dispersed count data neglect either the uncertainty in true counts from the censoring process, or the hierarchical structure of the spatiotemporally clustered data. Mixed-effects interval-censored negative binomial regression has potential to address these methodological issues directly. In this study, we conducted simulations to contrast the performance of mixed-effects interval-censored negative binomial regression against three other approaches, to illustrate how the extent of censoring, between-cluster variation, sample size, and strength of the true association can affect the findings from this method and comparison methods.

We assessed the bias in parameter estimates and standard errors, 95% confidence interval coverage, statistical power, and type I error rates of our model and the alternative approaches. The simulated data was generated under a hierarchical negative binomial process to which a mixed-effects negative binomial model was fit (Model 1). Then interval-censoring was imposed on the dataset and the interval-censored mixed-effects negative binomial regression was applied (Model 2). Next, we applied a condition on the dataset wherein the censored values were all deterministically imputed at a fixed value in the middle of the range of plausible counts. Under this condition that had some misclassification of the true counts, we applied mixed-effects negative binomial regression. Lastly, we then fitted fixed-effects negative binomial regression models that accounted for the interval-censoring, but neglected the hierarchy (Model 4). Building upon this, we applied the four modeling approaches to a real-world uncensored dataset of monthly mortality rates among black South Africans over 1997-2013, to examine the estimates of association of precipitation with mortality across Models 1-4, applying artificial censoring and deterministic imputation to mirror the simulations. Overall, in the simulated data, Models 1, 2, and 4 performed well in all measures. However, Model 3 performed increasingly poorly as the true effect size increased with the other parameters in the model. In the South Africa dataset, Models 1, 2, and 3 obtained similar estimates suggesting an inverse association of precipitation with mortality in black South Africans, while Model 4 gave a divergent finding. In conclusion, interval-censored mixed effects negative binomial regression should be considered as an analytical option when outcome data have both clustering and interval censoring.

Mixed-Effects Negative Binomial Regression with Interval Censoring: A Simulation Study and Application to Precipitation and All-Cause Mortality Rates among Black South Africans over 1997-2013

By

Christian Landon
B.S.
California State University, Fullerton
2016

Thesis Committee Chair: Matthew O. Gribble, Ph.D. D.A.B.T.

A thesis submitted to the faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health in Environmental Health
2019

Key Words: droughts; epidemiologic methods; errors, outcome measurement; multilevel analysis; vital statistics.

Acknowledgements

I would like to thank Dr. Matthew Gribble for taking time out of his day each week to teach me the concepts underlying this thesis. In the beginning I had yet to have any formal coursework on multi-level modeling and just the title of this thesis sounded daunting. Thank you to Dr. Robert Lyles of the Emory Rollins School of Public Health Biostatics Department for letting me expand upon your SAS code template and making this thesis possible. Thank you Dr. Noah Scovronick for providing the South Africa dataset and allowing this thesis to have real-world application. Last, but not least, thank you to my friends and family for your unconditional care and support. This incredible experience I had during my studies at Emory University would not have been possible without you.

TABLE OF CONTENTS

BACKGROUND

In public health research, models with count response variables are often used to describe patterns such as the number of deaths in a defined population, the number of days absent from school or work, the number of alcoholic drinks consumed per day, or the number of bacteria in dilution assays (Coxe, West, & Aiken, 2009). Poisson regression is a generalized linear model form commonly used to addresses research questions about counts, with the key assumption that the variance equals the mean and that observations are independent and identically distributed. However, in many areas of research, real-world count data is over-dispersed and may be more adequately described by negative binomial models that assume variance is proportionate, but not necessarily equal, to the mean (Smithson & Merkle, 2014). Other extensions to the basic Poison model are often necessitated by quirks of the data-generating processes being studied.

Observations in real-world studies are seldom independent; for example, in many surveillance studies, data on event counts are collected repeatedly from the same places over time (e.g., county-years of surveillance). Mixed-effect models account for this hierarchical clustering of observations by estimating an underlying distribution of cluster-specific parameters (e.g., random intercepts representing cluster-specific differences from the grand mean), and fitting the rest of the model (i.e., fixed effects) conditional on these cluster-specific parameters (Laird & Ware, 1982).

Censoring of the outcome variable is a frequent complication of applied research (Touloumi, Pocock, Babiker, & Darbyshire, 1999) (Wu, 1986). Censoring can arise in applied analyses of count data when events are rare, such as the number of deaths that occur in rural or very small populations, and data administrators (e.g., government agencies) censor the precise number of events in order to protect the privacy of persons who experienced those events (Bartell and Lewandowski, 2011), providing data analysts only with an interval within which the correct number of counts is contained. When the probability of censoring is related to the underlying exposure-outcome dose-response, this is an example of informative missingness and can introduce bias if unaddressed (Schluchter, 1992). Statistical methods exist to account for this censoring process directly (Terza, 1985) (Hilbe & Judson, 1998).

Mixed-effects Poisson and negative binomial regression models accounting for censoring of counts been used previously in a few applications (Quiroz, Wilson, & Roychoudhury, (2012) (Bartell & Lewandowski, 2011) (Lynch et al. (2018), but assessment of the performance of interval censored mixed-effects negative binomial models, vis-à-vis simplified alternatives, and versus mixed-effects negative binomial models fitted to uncensored complete data, remains to be explored.

The objective of this simulation study is to assess the performance of these methods, regarding bias in parameter estimates and standard errors, 95% confidence interval coverage, statistical power, and type I error rates. The contrasts between these approaches are then illustrated using a dataset of precipitation and monthly mortality rates among black South Africans over 1997-2013.

METHODS

SIMULATION

We developed a macro for SAS 9.4 software (SAS, Cary, NC) to conduct the simulations. The SAS IML (“interactive matrix language”) procedure is utilized within the macro to create the matrices, variables, and to set the parameter values for the simulations. In this study, we ran 1,000 simulations for each set of explored parameter values, including the baseline log-count (β_0), the variance of the random intercept reflecting differences in baseline log-counts (σ^2), the effect of the exposure on the outcome (β_1), the effect of the covariate (β_2), the negative binomial dispersion factor (α), the number of counties (N), and the number of study-years (K), resulting in 80,000 total simulations.

In order to examine how the models' performance was influenced by sample size, a five possible combinations of cluster-years were considered. Similarly to the Lynch et al. (2018) study, we will refer to the cluster-years as county-years throughout this paper given that the geographic unit of analysis is at the county level. The number of counties ranged from 10 to 500 and the number of years of observation per county ranged from 10 to 20 years. This resulted in a range of simulations with a minimum of 100 county-years and a maximum of 10,000 county-years. The combinations of county-years considered in this study are described in **Table 1**.

Data were simulated under four values of β_1: 0.01, 0.10, 0.20, and 0.50. We simultaneously changed the β_0 values: 2.25, 2.25, 2.25, and 0.00. The σ^2 parameter was also simultaneously changed and had values of 0.005, 0.05, 0.05, and 0.10. The parameter β_2 was held constant at 0, and α remained constant at a value of 0.25. This resulted in the following combinations of true parameter values for simulation across the different county-year levels.

$\beta 0 = 2.25$, $\beta 1 = 0.01$, $\beta 2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.005$.
$\beta 0 = 2.25$, $\beta 1 = 0.10$, $\beta 2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.05$.
$\beta 0 = 2.\ 25$, $\beta 1 = 0.20$, $\beta 2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.05$.
$\beta 0 = 0.00$, $\beta 1 = 0.50$, $\beta 2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.10$.

A uniformly distributed exposure variable ranging from 0-10 was generated as the primary independent variable. The outcome variable in the simulations was a negative binomial count variable. A uniformly distributed covariate ranging from 1-20 was also generated.

Thus, each simulation generated a complete negative binomial outcome dataset conditional on the exposure and other parameter values specified above.

Statistical Analysis of Simulated Data

We tested the performances of four models in this study that differed from each other in terms of how they modeled censored data, between-cluster variation, or both.

Model 1 a mixed-effects negative binomial regression of the following form fit to the complete simulated dataset:

$$\ln(\mu_{ij})=(\beta_0 + b_{0i}) + \beta_1 X_{ij} + \beta_2 Z$$

where x_{ij} is the exposure for county i in year j, Z is the covariate, and (b_{0i}) is a normally-distributed random intercept. Model 1 uses the following likelihood contribution for county *i* in year *j*:

$$\Pr\left(Y_{ij} = y_{ij} \middle| X_{ij} = x_{ij},\ Z = z\right) = \frac{\Gamma(\alpha^{-1}+y_{ij})}{\Gamma(\alpha^{-1})y_{ij}!}\left(\frac{\alpha\mu_{ij}}{1+\alpha\mu_{ij}}\right)^{y_{ij}}\left(\frac{1}{1+\alpha\mu_{ij}}\right)^{\frac{1}{\alpha}}$$

Model 2 has a similar linear predictive form and when the number of deaths (y_{ij}) is not censored (counts are 0 or $\geq$ 10), the same likelihood contribution is used. However, when the number of deaths is censored (deaths 1-9), the conditional likelihood contribution becomes:

$$\Pr\left(1 \le Y_{ij} \le 9 \middle| X_{ij} = x_{ij},\ Z = z\right)$$

based on the negative binomial model.

We use the SAS NLMIXED to specify this log-likelihood conditional on the random effects. Taking advantage of the recursive properties of the gamma function, contributions for the censored deaths are specified and the indicator for outcomes existing within the censoring window is defined within the procedure.

Model 3 has the same likelihood contribution as Models 1 and 2 when outcome counts are 0 or $\geq$ 10, but when the number of events is in the window [1,9], the outcome variable is forced to equal 5, and the likelihood contribution becomes:

$$\mathrm{If}\left(1 \leq \mathrm{Y_{ij}} \leq 9 \; then \; \mathrm{Y_{ij}} = 5 \;\middle|\mathrm{X_{ij}} = \mathrm{x_{ij}},\, \mathrm{Z} = \mathrm{z}\right)$$

This analytic approach is equivalent to deterministic imputation of censored counts by a count in the middle of the censoring window.

Model 4 accounts for censoring but does not account for the hierarchical structure of the data, and therefore its linear predictive form and conditional likelihood contribution are slightly different from the other models. When the number of deaths are in the censoring window [1-9], its likelihood contribution becomes:

$$\Pr(1 \leq \mathrm{Y} \leq 9 \;\; |\mathrm{X} = \mathrm{x},\, \mathrm{Z} = \mathrm{z})$$

A summary of all the model forms and likelihood contributions can be found in **Table 2.**

Simulation Summaries

Applying these 4 models to the 1,000 simulated datasets for each set of parameter values generated a set of 1,000 β_1 estimates and 1,000 β_1 standard errors of those estimates per approach for each data-generating process. The SAS MEANS procedure was applied to this distribution of β_1 estimates and distribution of β_1 standard errors to obtain the averages of those estimates for each modeling approach, applied to data from each data-generating process.

Confidence intervals (CIs) with intended 95% coverage for β_1 estimates were calculated under normality assumptions. Confidence interval coverage was calculated as the percentage of the 1,000 simulations wherein the confidence intervals generated within each dataset contained the true data-generating parameter value β_1. Statistical power was calculated as the percentage of the 1,000 simulations wherein the confidence intervals excluded zero. Both these conditions were tabulated using the SAS FREQ procedure.

The percentage bias of each β_1 parameter estimate was obtained by calculating the absolute value of the difference between the true value of β_1 and the estimated $\beta_{1,}$ and dividing by the true $\beta_{1;}$ then multiplying by 100 to express as a percentage. The SAS Compare procedure was used to perform this calculation.

Type I error rates were calculated as the percentage of the simulations, under a given set of parameter values where the true β_1 was 0, wherein the confidence intervals generated within each dataset excluded 0. The SAS FREQ procedure was used to tally the number of times per set of 1,000 simulations that this condition was met.

SOUTH AFRICAN DATA ANALYSIS

Mortality Data

Mortality data for black South Africans from each district of South Africa's civil registration system were obtained from Statistics South Africa (2013). The mortality file's completeness ranged from ~ 89% to ~ 94% throughout the study period of February 1st 1997 - December 1st 2013. This resulted in a total sample size of 10,607 district-months. Total population counts were used in this thesis as a surrogate for the black South African populations at risk of contributing to the black South African mortality outcomes. Population sizes were assumed to be constant within the years 1996-2000, 2001-2006, 2007-2010, and 2011-2013.

Standardized Precipitation Index

The primary exposure variable in the South African data set is based on the Standardized Precipitation Index (SPI). The SPI characterizes the meteorological drought conditions on a range of timescales. For this study, a 6-month timescale was used as a better indicator of climatological conditions (Keyantash, 2016). SPI is a common international indicator of can be compared across regions with markedly different climates. It quantifies observed precipitation as a standardized departure from a selected probability distribution function that models the raw precipitation data. The raw precipitation data are typically fitted to a gamma or a Pearson Type III distribution, and then transformed to a normal distribution (McKee et al., 1993). SPI values can be interpreted as the number of standard deviations by which the observed anomaly deviates from the long-term mean.

The dataset includes 6-month weighted averages of SPI values for every district in South Africa on a monthly timescale for the study period. The SPI ranges from -3 to 3, where values above 0 represent increasingly wet conditions and values below 0 represent increasingly dry conditions. This index was treated as a continuous variable for the main analysis, but coded as increasingly wet quartiles for a sensitivity analysis.

Model Specification and Statistical Analysis

Models analogous to those applied in the simulation study were fitted to the South African precipitation and mortality dataset (**Table 2**). In this application, SPI is the exposure variable, the covariate is the 17 year time period variable, and an added offset f_i reflects the ln(population) of each district at risk of developing mortality events.

In contrast to the simulation study, the true value for the effect of the exposure (SPI) on mortality is not known. As a result, the β1 estimate from Model 1, based on complete data, is used as the basis for comparison for the β1 estimates obtained from Models 2-4, which treat some of the outcomes as censored or imputed. The SAS procedure GLIMMIX was used to obtain plausible initial values for the negative binomial parameters used in the PARMS statement of the SAS NLMIXED procedure.

In addition to the main analysis modeling SPI as a continuous linear predictor, a sensitivity analysis was conducted to confirm that a linear dose-response was adequate, the SPI variable was transformed into quartiles to allow for a possibly non-linear dose-response. Because SAS NLMIXED does not have a 'class' statement available, SAS procedure GLIMMIX was used to obtain the β1 parameter estimates of the SPI quartiles in this sensitivity analysis. Model results were visualized using ggplot2 package (Wickham, 2016) in the R programming language (R Core team, 2019).

RESULTS

Simulation Study

All simulation results are displayed in appendix A. When β0 = 2.25, β1 = 0.01, β2 = 0.00, α = 0.25, and σ^2 = 0.005, all of the models had similar performance. Relative bias of the β1 parameter estimate was 1-2%. Model 3 had the smallest average standard error size. All four models had empirical confidence interval coverage that was permissive compared to the desired coverage level. Model 3 had the worst confidence interval coverage (93.16%) compared to 94.66% for Model 1, 94.52% for Model 2 and 94.60% for Model 4. Power varied from ~30% to ~99% depending on sample size.

When $\beta0 = 2.25$, $\beta1 = 0.10$, $\beta2 = 0.00$, $\alpha = 0.25$, and $\sigma^2 = 0.05$, differences between the models became more apparent. The mean percentage bias was small for Model 1 (0.06%), Model 2 (0.07%), and Model 4 (0.20%), whereas Model 3 had an average percentage bias of 3.22%. Model 1 had the smallest average standard error size, Model 2 had the second, Model 3 had the third, and Model 4 had the largest. Model 3 has very permissive confidence interval coverage at larger sample sizes: 81.30% for 5,000 county-years and 66.10% for 10,000 county-years. Statistical power reached 100% for all models at 500 county-years.

When $\beta0 = 2.\ 25$, $\beta1 = 0.20$, $\beta2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.05$, model 3 had the largest percent bias (2.13%) observed at 1,000 county-years. The same standard error size pattern that was seen at a true β1 value of 0.10 also occurred here. Confidence interval coverage for Models 1, 2, and 4 ranged from 92-95% for every county-year level, with Model 2 having best coverage. Model 3 reached coverage levels of approximately 90-94% until the simulations at 5,000 and 10,000 county-years, where its coverage dropped to 71.80% and 51.20%, respectively. Power was ~100% for all models at all sample sizes.

When $\beta0 = 0.00$, $\beta1 = 0.50$, $\beta2 = 0.00$, $\alpha = 0.25$, and $\sigma^2 = 0.10$, models 1 and 2 had biases below 0.13% at every county-year level, and Model 4 had a slightly higher amount (0.20 being the minimum and 0.26 being the maximum). In contrast, Model 3 had an average bias of 15.5% across all of the county-year levels. Model 3 had the smallest average standard error, followed by models 1, 2, and 4 respectively. Models 1, 2, and 4 had confidence intervals coverage of 93-95%. In contrast, Model 3 coverage peaked at 12.5% at 100 county-years, and for the rest of the county-year levels the coverage was ~0.00%. Power was ~100% for all of models at all sample sizes.

All models had similar type I error rates (4-5%). Model 3 had the highest rate of 6.10% at 100 county-years and model 1 had the lowest with 3.90% at 1,000 county-years.

Precipitation and Mortality Rates among Black South Africans

Descriptive statistics on South African mortality counts per district are summarized in Appendix B. The approximate population size of the districts ranged from 52,010 to 4,434,922 persons, with a mean of 876,296 over the study period.

In models relating the numbers of deaths per district-month to precipitation, Models 1 estimated a rate ratio of 0.96 (95% CI 0.95, 0.97), Model 2 estimated a rate ratio of 0.96 (95% CI 0.95, 0.97), Model 3 estimated a rate ratio of 0.96 (95% CI 0.95, 0.97), and Model 4 produced a divergent estimated rate ratio of 0.95 (95% CI 0.93, 0.96). Although the divergent estimated rate ratio produced by Model 4 is only 1% different than the other models, at a national level this 1% difference in increased mortality rates corresponds to a difference of thousands of estimated deaths over time.

In a sensitivity analysis for possible non-linearity of the dose-response, there was a monotonic dose-response of increasing mortality with increasingly drought-like conditions. Comparing the second-wettest quartile of district-months to the wettest quartile of district-months resulted in a mortality rate ratio of 1.01, (95% CI 0.99, 1.04). Comparing the second-driest quartile of district-months to the wettest quartile of district-months resulted in a mortality rate ratio of 1.04 (95% CI 1.02, 1.06). Contrasting the driest quartile of district-months against the wettest quartile of district months resulted in a mortality rate ratio of 1.10 (95% CI 1.08, 1.12).

DISCUSSION

The simulation study indicates that all of the models perform relatively similarly when the model true parameters were set to $\beta0 = 2.25$, $\beta1 = 0.01$, $\beta2 = 0.00$, $\alpha = 0.25$, and $\sigma^2 = 0.005$ Where differences begin to come apparent is when the effect size of $\beta1$ and σ^2 were increased for the next two sets of simulations, where the true model parameter values were $\beta0 = 2.25$, $\beta1 = 0.10$, $\beta2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.05$ and $\beta0 = 2.\ 25$, $\beta1 = 0.20$, $\beta2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.05$ respectively. Here, model three begins to be increasingly influenced by differential measurement error of outcome by exposure. This influence became the most pronounced when measured at the true parameter values of $\beta0 = 0.00$, $\beta1 = 0.50$, $\beta2 = 0.00$, $\alpha = 0.25$, and $\sigma^2 = 0.10$.

Given that there is a true association between the exposure and outcomes in the simulation, as the exposure increases the number of events in a county should increase. As a result, due to this dose-response, counties at higher levels of exposure will have fewer events in the censoring interval. Thus, if the outcome is imputed with error (as seen in Model 3), the outcome measurement error would affect the unexposed more than the exposed. This leads to the large bias, non-existent confidence interval coverage, and falsely small standard error sizes we found when applying Model 3 to data simulated under the true values of $\beta0 = 0.00$, $\beta1 = 0.50$, $\beta2 = 0.00$, $\alpha = 0.25$, and $\sigma^2 = 0.10$. The other models were not influenced by this error likely due to the specified likelihood contribution used to address the censoring. Overall, a similar pattern emerged between Models 1, 2, and 4 as the true value of $\beta1$ increased with the other parameters. Model 2 performed very similarly to the uncensored model (model 1) and model 4 also performed similar, but had slightly decreased confidence interval coverage, and slightly increased degree of bias and standard error size.

In the South African dataset, Model 3 appeared to perform well, but Model 4 appeared to perform poorly in terms of β1 estimation. This likely occurred due to the larger between-district variance of the outcome observed in our South African data compared to the smaller between-county variance seen in the simulation study.

An additional finding from this dataset was observing a statistically significant association between SPI and mortality among black South Africans. It was suspected a U-shaped relationship could be found, but the SPI dose-response was found to be linear. This lack of a U-shaped relationship is likely contributed to the fact that the SPI does not take into consideration the intensity of precipitation (National Center for Atmospheric Research, 2019). However, it is intuitive that drier periods would be associated with increased mortality due to the effects on agriculture and soil composition.

Few studies have robustly examined mixed-effect negative binomial models with interval censoring. Similar models have been employed in research conducted by Bartell and Lewandowski (2011), Quiroz et al. (2012), and Lynch et al. (2018), but none compare interval censored+ mixed-effect negative binomial models performance versus imputation methods and censored regression that estimates fixed-effects only. This simulation study observed results that contrast the findings by Bartell and Lewandowski (2011) with regard to substitution (imputation) methods. They found the effects of substitution to be negligible when two conditions are met: (1) few observations are below the censoring cutoff and (2) the censoring cutoff is relatively small compared to most of the measurements. However, in our simulations where β0 = 2.25, β1 = 0.01, β2 = 0.00, α

= 0.25, and σ^2 = 0.005, the substitutions Model 3 implemented had negligible effects. In this case, approximately 50% of the data was within the censoring interval.

This simulation study had some notable limitations. Only four β1 effect sizes were explored. The research could have been improved upon if more effect sizes were explored, particularly negative β1 values. The results would have also provided stronger evidence if β_1 had been the only parameter value changed between simulations. However, at lower true values of β_1 this was not possible because a large enough effect size had to be produced in order for the SAS macro to generate observations where more than 9 deaths occurred. Subsequently, the effect size of β_0 had to be increased to a point where not only censored data was being generated. Additionally, a larger censoring interval than [1-9] could provide additional insight. However, coding logistics made it difficult to expand the interval to a much larger size (see appendix C). For future research, taking advantage of recursive properties or arrays within SAS procedure NLMIXED would simplify this issue and allow for examination of larger censoring intervals

Within the South Africa dataset, one apparent limitation of a methodological comparison is that the true association between SPI and mortality among black South Africans is not known, whereas the true exposure-outcome relationship in the simulation study is known. Therefore, Model 1 is used as an approximation for the true association. Monthly or yearly demographic information was not available at the time of this research. A large assumption was made when postulating that population size held constant over a 5-year period. This portion of the study could have been improved upon if more detailed demographic information was integrated and the population size of black South Africans was known. The effect size found was also fairly small compared to some of simulations, so it is possible that bias seen Model 3 did not manifest as it did in the simulation study due to this.

CONCLUSION

Overall, out of the censored models, Model 2, interval-censored mixed effects negative binomial regression, was the only model that performed well in both the conditions of the simulated and South African real-world dataset. Model 2 was not subject to large estimate bias, poor confidence interval coverage, or falsely decreased or inflated standard error sizes. This well-performing approach opens up further possibilities for more accurate research findings when examining environmental exposures and adverse human health outcomes.

References

Bartell, S. M., & Lewandowski, T. A. (2011). Administrative censoring in ecological analyses of autism and a Bayesian solution. *Journal of Environmental and Public Health*, *2011*, 1–5. https://doi.org/10.1155/2011/202783

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, *91*(2), 121–136. https://doi.org/10.1080/00223890802634175

Hilbe, J. (2011). Negative binomial regression. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511973420

Hilbe, J. M., & D. H. Judson. (1998). sg94: Right, left, and uncensored Poisson regression. *Stata Technical Bulletin 46*: 18–20. College Station, TX: Stata Press.

Keyantash, J., & National Center for Atmospheric Research Staff (Eds). (2016). The Climate Data Guide: Standardized Precipitation Index (SPI). Retrieved from: https://climatedataguide.ucar.edu/climate-data/standardized-precipitation-index-spi.

Laird, N., & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics, 38*(4), 963-974. doi:10.2307/2529876

Lynch, K., Lyles, R. H., Waller, L. A., Bell, J. E., & Gribble, M.O. (2018). Drought severity and all-cause mortality rates among adults in the United States: 1968-2014. Master's Thesis. Emory University Rollins school of public health. Atlanta, Georgia, United States.

McKee, T.B., Doesken, N. J. & Kliest, J. (1993). The relationship of drought frequency and duration to time scales. In Proceedings of the 8th Conference of Applied Climatology, 17-22 January, Anaheim, CA. American Meterological Society, Boston, MA. 179-18.

National Center for Atmospheric Research (n.d.). Standardized precipitation index (SPI). Retrieved from: https://climatedataguide.ucar.edu/climate-data/standardized-precipitation-index-spi

Osgood, D. W. (2000). Poisson-based regression analysis of aggregate crime rates. *Journal of Quantitative Criminology*, *16*(1), 21–43. Retrieved from: https://doi.org/10.1023/A:1007521427059

Quiroz, J., Wilson, J. R., & Roychoudhury, S. (2012). Statistical analysis of data from dilution assays with censored correlated counts. *Pharmaceutical Statistics 11*: 63-73. doi: 10.1002/pst.499

R Core Team (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: https://www.R-project.org/

SAS Institute Inc. (2008). SAS/STAT® 9.2 user's guide. Cary, NC: SAS Institute Inc.

Schluter, M. D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine*, *11*, 1861–1870.

Scovronick, N., Sera, F., Acquaotta, F., Garzena, D., Fratianni, S., Wright, C. Y., & Gasparrini, A. (2018). The association between ambient temperature and mortality in South Africa: A time-series analysis. *Environmental Research*, *161*, 229-235. Retrieved from: https://doi.org/10.1016/j.envres.2017.11.001

Smithson, M., & Merkle, E (2014). *Generalized linear models for categorical and continuous limited dependent variables*. Boca Raton, FL: CRC Press, Taylor & Francis Group

Statistics South Africa (2013). Mortality and causes of death in South Africa, 2013: Findings from death notification. Pretoria, South Africa: Stats SA

Terza, J. V. (1985). A Tobit-type estimator for the censored Poisson regression model. *Economics Letters, 18*(4), 361-365. doi:https://doi.org/10.1016/0165-1765(85)90053-9

Touloumi, G., Pocock, S., Babiker, A., & H Darbyshire, J. (1999). Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. *Statist*. *Med*. 18: 1215 -1233. doi: 10.1002/(SICI)1097-0258(19990530)18:10<1215::AID-SIM118>3.0.CO;2-6

UCLA Statistical Consulting Group. (2016). *Negative binomial regression | STATA annotated output*. Retrieved from: https://stats.idre.ucla.edu/stata/output/negative-binomial-regression/

Wickham, H. (2016) ggplot2: Elegant graphics for data analysis. New York, NY: Springer-Verlag

Wu, M. C. (1988). Sample size for comparison of changes in the presence of right censoring caused by death, withdrawal, and staggered entry. *Controlled Clinical Trials, 9*(1), 32-46. doi:https://doi.org/10.1016/0197-2456(88)90007-4

Tables

Table 1: Simulation Design and Sample Size

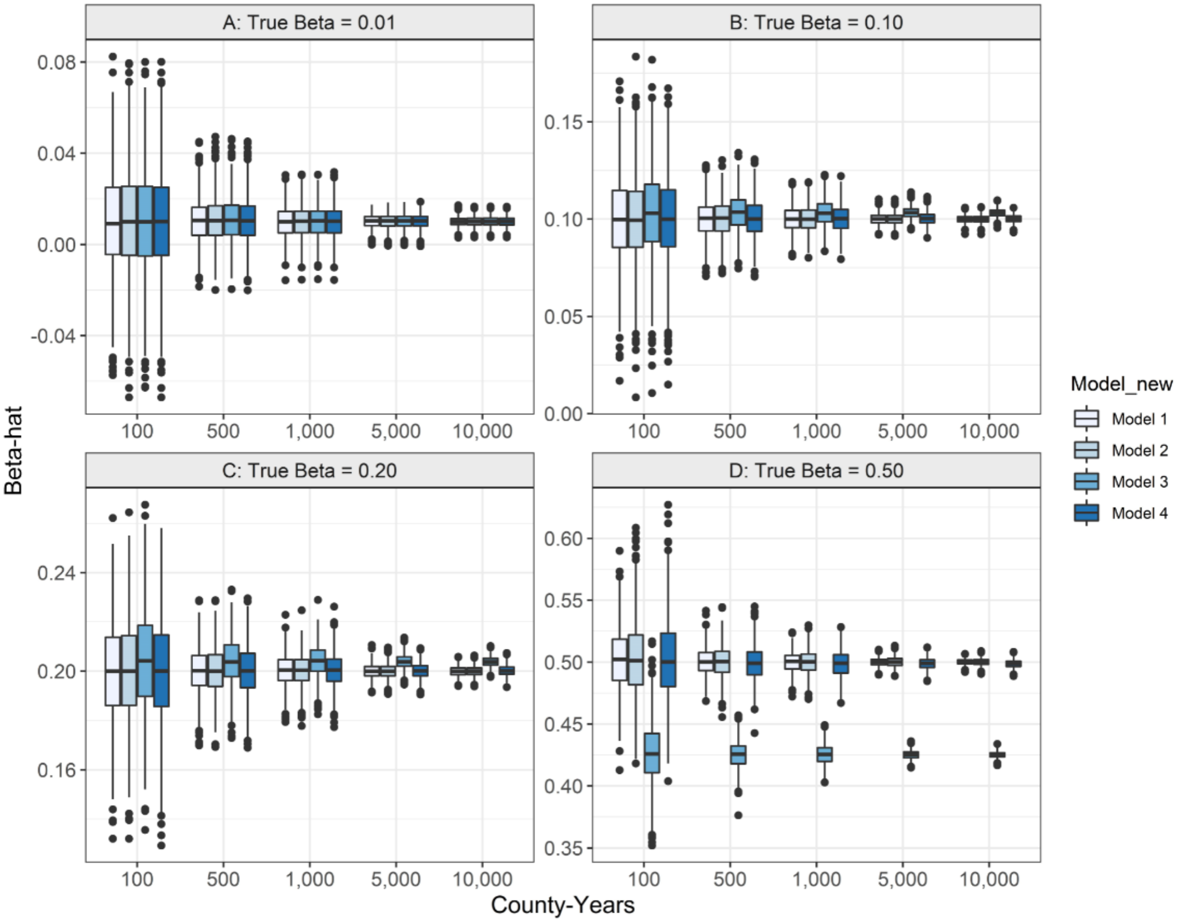
True Beta: 0.01				True Beta: 0.10			
K	N	K*N	Simulations	K	N	K*N	Simulations
10	10	100	4,000	10	10	100	4,000
50	10	500	4,000	50	10	500	4,000
50	20	1,000	4,000	50	20	1,000	4,000
250	20	5,000	4,000	250	20	5,000	4,000
500	20	10,000	4,000	500	20	10,000	4,000
True Beta: 0.20				True Beta: 0.50			
K	N	K*N	Simulations	K	N	K*N	Simulations
10	10	100	4,000	10	10	100	4,000
50	10	500	4,000	50	10	500	4,000
50	20	1,000	4,000	50	20	1,000	4,000
250	20	5,000	4,000	250	20	5,000	4,000
500	20	10,000	4,000	500	20	10,000	4,000

Where K represents the number of counties, N represents the number of years that counties were observed, and K*N represents the number of county-years (sample size).

Table 2. Model Descriptions and Forms

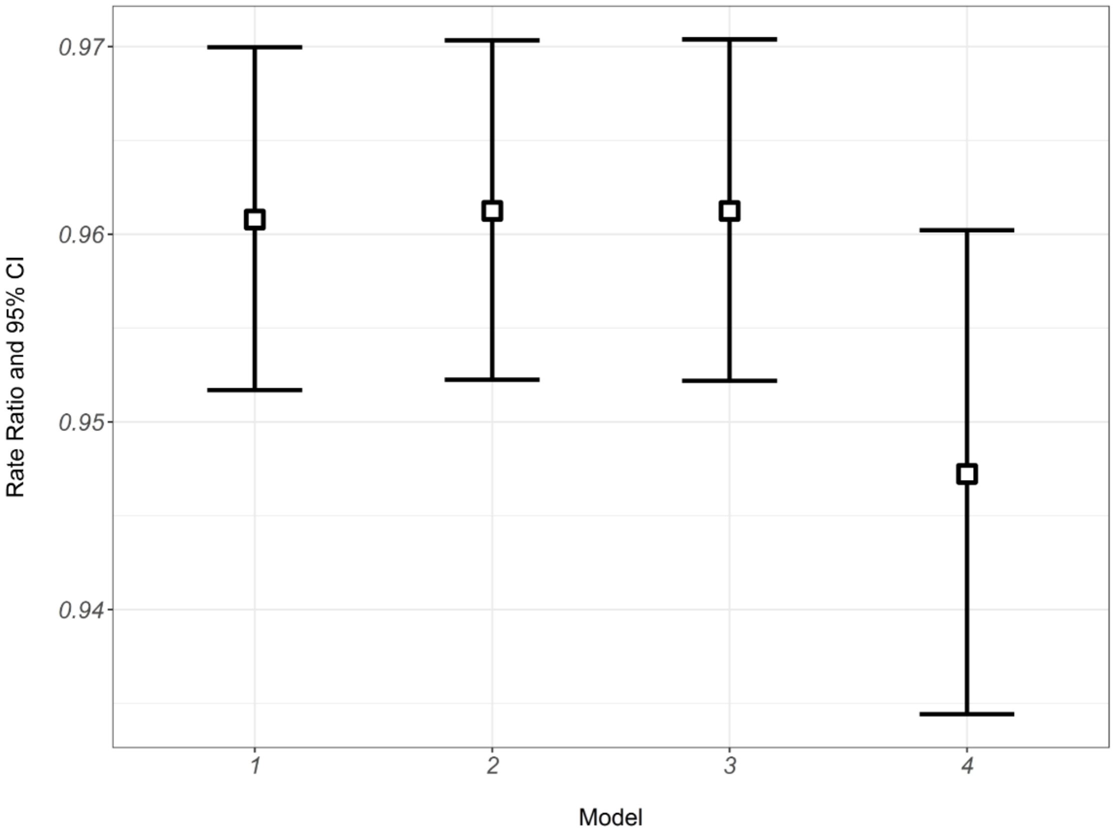
Model	Form	Description
1	$\ln(\mu_{ij})=(\beta_0+b_{0i})+\beta_1 X_{ij}+\beta_2 Z$ $\Pr(Y_{ij}=y_{ij}\vert X_{ij}=x_{ij},\, Z=z)=\frac{\Gamma(\alpha^{-1}+y_{ij})}{\Gamma(\alpha^{-1})y_{ij}!}\left(\frac{\alpha\mu_{ij}}{1+\alpha\mu_{ij}}\right)^{y_{ij}}\left(\frac{1}{1+\alpha\mu_{ij}}\right)^{\frac{1}{\alpha}}$	Mixed negative binomial model and likelihood contribution for county *i* in year *j*
2	$\ln(\mu_{ij})=(\beta_0+b_{0i})+\beta_1 X_{ij}+\beta_2 Z$ $\Pr(1\leq Y_{ij}\leq 9\vert X_{ij}=x_{ij},\, Z=z)$	Interval censored (on deaths 1-9) mixed-effects negative binomial model and conditional likelihood contribution.
3	$\ln(\mu_{ij})=(\beta_0+b_{0i})+\beta_1 X_{ij}+\beta_2 Z$ $\mathrm{If}\,(1\leq Y_{ij}\leq 9\ \mathit{then}\ Y_{ij}=5\ \vert X_{ij}=x_{ij},\, Z=z)$	Midpoint imputed mixed-effects negative binomial model and conditional likelihood contribution
4	$\ln(\mu)=(\beta_0)+\beta_1 X+\beta_2 Z$ $\Pr(1\leq Y\leq 9\ \ \vert X=x,\, Z=z)$	Interval censored (on deaths 1-9) fixed-effects negative binomial model and conditional likelihood contribution

Figure 1. Estimated Effects



Figures

Figure 3. Estimated Effects from South African Dataset
Interval Censored on Deaths 1-9



Appendix

Appendix A: Simulation Results
Results for Mean β1 Maximum Likelihood Estimates and Standard Errors

County-Years	True β_1	Model 1		Model 2		Model 3		Model 4	
		MLE	S.E.	MLE	S.E.	MLE	S.E.	MLE	S.E.
100	0.01	0.010	0.021	0.010	0.022	0.010	0.020	0.010	0.022
500	0.01	0.011	0.009	0.010	0.010	0.011	0.009	0.010	0.010
1,000	0.01	0.010	0.007	0.010	0.007	0.010	0.006	0.010	0.007
5,000	0.01	0.010	0.003	0.010	0.003	0.010	0.003	0.010	0.003
10,000	0.01	0.010	0.002	0.010	0.002	0.010	0.002	0.010	0.002
Mean	0.01	0.010	0.008	0.010	0.009	0.010	0.008	0.010	0.009

Mean results where the true values of β0 = 2.25, β1 = 0.01, β2 = 0.00, α = 0.25, σ² = 0.005.

		MLE	S.E.	MLE	S.E.	MLE	S.E.	MLE	S.E.
100	0.10	0.100	0.020	0.100	0.021	0.103	0.021	0.100	0.021
500	0.10	0.100	0.009	0.100	0.009	0.103	0.009	0.100	0.010
1,000	0.10	0.100	0.006	0.100	0.006	0.103	0.007	0.100	0.007
5,000	0.10	0.100	0.003	0.100	0.003	0.103	0.003	0.100	0.003
10,000	0.10	0.100	0.002	0.100	0.002	0.103	0.002	0.100	0.002
Mean	0.10	0.100	0.008	0.100	0.008	0.103	0.008	0.100	0.009

Mean results where the true values of β0 = 2.25, β1 = 0.10, β2 = 0.00, α = 0.25, σ² = 0.05.

		MLE	S.E.	MLE	S.E.	MLE	S.E.	MLE	S.E.
100	0.02	0.200	0.020	0.200	0.020	0.204	0.020	0.200	0.021
500	0.02	0.200	0.009	0.200	0.009	0.204	0.009	0.200	0.009
1,000	0.02	0.200	0.006	0.200	0.006	0.204	0.006	0.200	0.007
5,000	0.02	0.200	0.003	0.200	0.003	0.204	0.003	0.200	0.003
10,000	0.02	0.200	0.002	0.200	0.002	0.204	0.002	0.200	0.002
Mean	0.02	0.200	0.008	0.200	0.008	0.204	0.008	0.200	0.008

Mean results where the true values of β0 = 2.25, β1 = 0.20, β2 = 0.00, α = 0.25, σ² = 0.05.

		MLE	S.E.	MLE	S.E.	MLE	S.E.	MLE	S.E.
100	0.5	0.502	0.024	0.502	0.029	0.426	0.022	0.501	0.031
500	0.5	0.501	0.011	0.500	0.013	0.425	0.010	0.499	0.014
1,000	0.5	0.500	0.008	0.500	0.009	0.425	0.007	0.499	0.010
5,000	0.5	0.500	0.003	0.500	0.004	0.425	0.003	0.499	0.004
10,000	0.5	0.500	0.002	0.500	0.003	0.425	0.002	0.499	0.003
Mean	0.5	0.501	0.010	0.501	0.012	0.425	0.009	0.499	0.012

Mean results the true values of β0 = 0.00, β1 = 0.50, β2 = 0.00, α = 0.25, σ² = 0.10.

Results for β_1 mean Confidence Interval Performance and Statistical Power: True values of 0.01 and 0.10

County-Years	True β_1 [a]	Model 1		Model 2	
		CI Coverage	Power	CI Coverage	Power
100	0.01	94.10	9.20	93.70	9.13
500	0.01	94.50	20.50	94.20	18.80
1,000	0.01	95.30	33.70	94.90	31.00
5,000	0.01	94.70	92.30	95.50	91.00
10,000	0.01	94.70	99.70	94.30	99.50
Mean	0.01	94.66	51.08	94.52	49.89
		Model 3		Model 4	
		CI Coverage	Power	CI Coverage	Power
100	0.01	92.20	8.89	93.70	9.12
500	0.01	93.00	22.10	94.40	18.90
1,000	0.01	93.20	34.80	95.00	31.80
5,000	0.01	94.30	92.40	95.80	91.30
10,000	0.01	93.10	99.70	94.10	99.50
Mean	0.01	93.16	51.58	94.60	50.12

County-Years	True β_1 [b]	Model 1		Model 2	
		CI Coverage	Power	CI Coverage	Power
100	0.10	92.80	99.50	92.60	99.40
500	0.10	94.20	100.00	94.10	100.00
1,000	0.10	95.60	100.00	95.40	100.00
5,000	0.10	94.60	100.00	95.00	100.00
10,000	0.10	96.20	100.00	96.20	100.00
Mean	0.10	94.68	99.90	94.66	99.88
		Model 3		Model 4	
		CI Coverage	Power	CI Coverage	Power
100	0.10	91.90	99.40	92.90	99.00
500	0.10	93.80	100.00	94.50	100.00
1,000	0.10	92.00	100.00	94.40	100.00
5,000	0.10	81.30	100.00	95.30	100.00
10,000	0.10	66.10	100.00	95.30	100.00
Mean	0.10	85.02	99.88	94.48	99.80

[a] Mean results where the true values of $\beta 0 = 2.25$, $\beta 1 = 0.01$, $\beta 2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.005$.

[b] Mean results where the true values of $\beta 0 = 2.25$, $\beta 1 = 0.10$, $\beta 2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.05$.

Results for β_1 mean Confidence Interval Performance and Statistical Power: True values of 0.20 and 0.50

County-Years	True β_1 [c]	Model 1		Model 2	
		CI Coverage	Power	CI Coverage	Power
100	0.20	94.80	100.00	94.50	100.00
500	0.20	93.70	100.00	94.10	100.00
1,000	0.20	95.10	100.00	95.40	100.00
5,000	0.20	94.10	100.00	94.10	100.00
10,000	0.20	93.30	100.00	93.60	100.00
Mean	0.20	94.20	100.00	94.34	100.00
		Model 3		Model 4	
		CI Coverage	Power	CI Coverage	Power
100	0.20	94.10	100.00	94.70	100.00
500	0.20	90.90	100.00	92.80	100.00
1,000	0.20	90.03	100.00	94.30	100.00
5,000	0.20	71.80	100.00	93.60	100.00
10,000	0.20	51.20	100.00	93.40	100.00
Mean	0.20	79.61	100.00	93.76	100.00

County-Years	True β_1 [d]	Model 1		Model 2	
		CI Coverage	Power	CI Coverage	Power
100	0.50	94.20	100.00	93.50	100.00
500	0.50	94.70	100.00	94.40	100.00
1,000	0.50	95.00	100.00	95.70	100.00
5,000	0.50	95.40	100.00	95.60	100.00
10,000	0.50	95.40	100.00	94.30	100.00
Mean	0.50	94.94	100.00	94.70	100.00
		Model 3		Model 4	
		CI Coverage	Power	CI Coverage	Power
100	0.50	12.50	100.00	93.60	100.00
500	0.50	0.00	100.00	94.80	100.00
1,000	0.50	0.00	100.00	93.20	100.00
5,000	0.50	0.00	100.00	93.00	100.00
10,000	0.50	0.00	100.00	94.40	100.00
Mean	0.50	2.50	100.00	93.80	100.00

[c] Mean results where the true values of $\beta0 = 2.\ 25$, $\beta1 = 0.20$, $\beta2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.05$.

[d] Mean results where the true values of $\beta0 = 0.00$, $\beta1 = 0.50$, $\beta2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.10$.

Results for β_1 Mean Bias

County-Years	True β_1[a]	Model 1	Model 2	Model 3	Model 4
		Pct Diff	Pct Diff	Pct Diff	Pct Diff
100	0.01	1.25	0.21	0.84	0.14
500	0.01	5.27	4.58	5.30	4.55
1,000	0.01	0.84	0.85	0.27	0.90
5,000	0.01	1.94	1.69	2.24	1.66
10,000	0.01	0.02	0.04	0.50	0.04
Mean	0.01	1.86	1.47	1.83	1.46
County-Years	True β_1[b]	Model 1	Model 2	Model 3	Model 4
		Pct Diff	Pct Diff	Pct Diff	Pct Diff
100	0.10	0.04	0.13	3.08	0.30
500	0.10	0.17	0.16	3.40	0.13
1,000	0.10	0.01	0.00	3.22	0.15
5,000	0.10	0.00	0.01	3.24	0.27
10,000	0.10	0.06	0.06	3.16	0.17
Mean	0.10	0.06	0.07	3.22	0.20
County-Years	True β_1[c]	Model 1	Model 2	Model 3	Model 4
		Pct Diff	Pct Diff	Pct Diff	Pct Diff
100	0.20	0.03	0.11	2.10	0.70
500	0.20	0.03	0.06	2.04	0.13
1,000	0.20	0.16	0.17	2.13	0.24
5,000	0.20	0.02	0.01	1.95	0.08
10,000	0.20	0.02	0.02	1.93	0.1
Mean	0.20	0.05	0.07	2.03	0.25
County-Years	True β_1[d]	Model 1	Model 2	Model 3	Model 4
		Pct Diff	Pct Diff	Pct Diff	Pct Diff
100	0.50	0.12	0.01	17.65	0.24
500	0.50	0.13	0.10	14.98	0.20
1,000	0.50	0.01	0.02	14.93	0.26
5,000	0.50	0.04	0.03	14.95	0.25
10,000	0.50	0.04	0.03	14.97	0.26
Mean	0.50	0.07	0.04	15.50	0.24

Percent difference is calculated as the absolute difference between the $\beta1$estimate and the true $\beta1$ value divided by the true $\beta1$ value which is then multiplied by 100 to be converted into a percentage

[a] Mean results where the true values of $\beta0 = 2.25$, $\beta1 = 0.01$, $\beta2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.005$.

[b] Mean results where the true values of $\beta0 = 2.25$, $\beta1 = 0.10$, $\beta2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.05$.

[c] Mean results where the true values of $\beta0 = 2.25$, $\beta1 = 0.20$, $\beta2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.05$.

[d] Mean results where the true values of $\beta0 = 0.00$, $\beta1 = 0.50$, $\beta2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.10$.

Results for Type I Error Rate

County-Years	Model 1	Model 2	Model 3	Model 4
	Type I Error	Type I Error	Type I Error	Type I Error
100	5.00	5.10	6.10	5.20
500	5.40	4.90	5.40	4.80
1,000	4.80	5.10	5.70	4.90
5,000	3.90	3.60	4.00	3.10
10,000	4.30	4.10	4.70	4.00
Mean	4.68	4.56	5.18	4.40

Results from 1,000 simulations at each county-year level where the true values of $\beta0 = 2.25$, $\beta1 = 0.00$, $\beta2 = 0.00$, $\alpha = 0.25$, $\sigma^2 = 0.01$. A result was considered a type I error if the beta-1 estimate confidence interval did not contain zero when the value was truly zero. This frequency was then converted into a percentage.

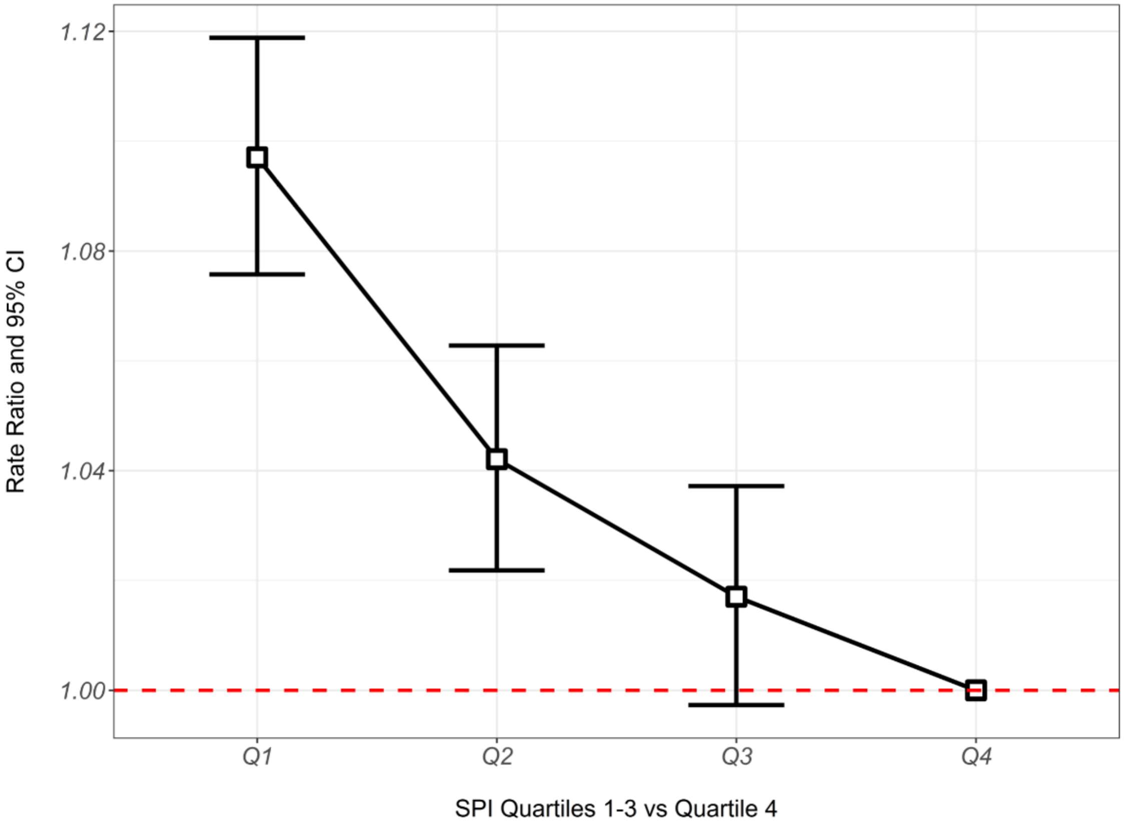
Appendix B: South Africa Tables and Figures

Black South African Mortality Statistics by District 1997-2013

Obs	District	N	Mean	Sum	Max	min	Median	Variance
1	**Totals**	10607	481.28	5104905	2548	0	426.0	168798.90
2	Alfred Nzo	203	271.96	55208	446	83	265.0	5990.85
3	Amajuba	204	354.12	72241	571	1	376.5	20301.38
4	Amathole	204	798.76	162947	1153	196	864.0	52986.39
5	Bojanala	204	752.97	153606	1116	112	812.0	49811.74
6	Buffalo City	204	579.69	118256	829	110	629.0	26493.77
7	Cacadu	204	127.83	26077	267	10	137.0	2085.76
8	Cape Winelands	204	69.87	14254	127	1	75.0	701.67
9	Capricorn	204	793.18	161809	1356	172	859.0	59038.16
10	Central Karoo	204	11.40	2325	31	0	11.0	38.43
11	Chris Hani	204	451.75	92156	709	79	486.5	18281.98
12	City of Cape Town	204	517.38	105545	787	38	552.0	26887.11
13	City of Johannesburg	204	1406.95	287018	2093	77	1510.5	223162.23
14	City of Tshwane	204	1052.10	214629	1674	108	1141.0	123350.85
15	Dr Kenneth Kaunda	204	490.94	100151	711	61	519.5	21419.19
16	Dr Ruth Segomotsi Mompati	204	314.07	64070	531	27	353.0	12959.03
17	Eden	204	65.08	13276	115	1	68.0	553.27
18	Ehlanzeni	204	1020.83	208249	1610	63	1101.0	155516.35
19	Ekurhuleni	204	1281.44	261413	1986	56	1362.5	209303.06
20	Fezile Dabi	204	343.48	70070	697	16	361.5	16469.17
21	Frances Baard	204	163.02	33257	275	7	173.0	3112.41
22	Gert Sibande	204	754.51	153920	1325	11	794.0	83503.61
23	Greater Sekhukhune	204	612.79	125010	934	102	660.5	31394.04
24	Joe Gqabi	204	201.39	41084	315	35	219.0	4843.20
25	John Taolo Gaetsewe	204	141.29	28823	255	12	152.5	2823.38
26	Lejweleputswa	204	543.87	110950	867	55	564.5	31766.26
27	Mangaung	204	556.18	113461	1079	50	597.5	43228.15
28	Mopani	204	554.59	113136	865	149	596.0	22125.46
29	Namakwa	204	6.87	1402	20	0	6.0	12.93
30	Nelson Mandela Bay	204	472.07	96302	835	55	494.5	44523.63

Obs	District	N	Mean	Sum	Max	min	Median	Variance
31	Ngaka Modiri Molema	204	500.42	102086	740	24	560.5	30851.06
32	Nkangala	204	675.45	137792	1048	45	723.5	58688.99
33	O.R.Tambo	204	555.70	113363	946	159	575.5	33676.68
34	Overberg	204	24.35	4967	55	2	23.0	120.17
35	Pixley ka Seme	204	50.18	10236	163	1	38.0	1232.32
36	Sedibeng	204	532.16	108560	891	3	585.5	39693.56
37	Sisonke	204	272.77	55645	463	29	285.0	10860.15
38	Siyanda	204	45.17	9215	132	0	46.5	369.70
39	Thabo Mofutsanyane	204	744.81	151941	1255	29	792.0	72250.88
40	Ugu	204	603.59	123133	895	4	678.5	54731.10
41	Vhembe	204	383.11	78154	588	95	398.0	11118.29
42	Waterberg	204	277.31	56572	462	18	309.5	11789.37
43	West Coast	204	47.72	9735	103	1	46.0	592.36
44	West Rand	204	486.65	99276	814	5	523.5	34681.04
45	Xhariep	204	139.29	28415	254	27	140.0	2031.83
46	Zululand	204	440.79	89921	750	27	507.5	32316.79
47	eThekwini	204	1692.04	345177	2548	270	1817.5	317555.36
48	iLembe	204	317.00	64667	531	7	349.5	15342.15
49	uMgungundlovu	204	763.92	155840	1180	41	822.5	85584.81
50	uMkhanyakude	204	313.47	63947	521	43	336.5	12571.25
51	uMzinyathi	204	333.10	67952	579	19	346.0	18144.11
52	uThukela	204	515.71	105204	879	25	541.5	33628.73
53	uThungulu	204	600.30	122462	938	25	623.0	41155.39

Figure 2. Linearity Assesment of Estimated Model Effects
from South Africa Dataset



Appendix C: SAS Code Examples

Simulation Data Generation and Sample Models

```
libname Drought 'h:\bob\Gribble';

options ps=66 ls=90 nodate nonumber nonotes;

title1 'PROGRAM: Sim Pgm Template 09_13_18.sas';
title2 'Simulating data from Negative Binomial model';

ods listing;

%let nsim=1000;


%macro iternb;
   %do q=1 %to &nsim;


proc iml worksize=70 symsize=250;


k=50;       ** Number of counties **;
n=20;         ** # observations (years)per county **;
ntot=n*k;

sig1sq=.05;

**Set parameters for NB regression simulation**;
**NOTE: 1/alpha has to be an integer to generate data, but not in
analysis of data **;

  bet0=2.25;  bet1= 0.2;  bet2=0;
  alpha=.25;  r=1/alpha;


t={1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20};   ** Let years
go from 1 to 20 **;
tmat=j(n,k,0);
w=j(n,k,0);


  do j1=1 to n;
    w[j1,]=1:k; *** matrix with ID #s for exporting to proc nlmixed **;
  end;

  do i=1 to k;
  tmat[,i]=t`; *** matrix with obs. times for exporting to proc nlmixed
**;
  end;

 wvec=shape(w`,ntot,1);        ***STRING OUT W and T INTO VECTORS***;
 tvec=shape(tmat`,ntot,1);
```

```
START DATAGEN;

gamm0is=j(k,1,0);
ymat=j(n,k,0);
indexmat=j(n,k,0);
linpred=j(n,k,0);
mumat=j(n,k,0);
pmat=j(n,k,0);
RVx=j(r,1,0);

t={1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20};   ** Let years
go from 1 to 20 **;
tprime=t`;

   do i=1 to k;

      gamm0is[i,]=0 + sqrt(sig1sq)*RANNOR(0);


   do j=1 to n;
      indexmat[j,i]=10*RANUNI(0); **Generate drought index for each
county/year as uniform(0,10);

        linpred[j,i]=(bet0 + gamm0is[i,]) + bet1*indexmat[j,i] +
bet2*tprime[j,];
        mumat[j,i]=exp(linpred[j,i]);

        pmat[j,i]=1/(1+mumat[j,i]/r);

**Generating each NB outcome**;

do randindx=1 to r;
     u=RANUNI(0);
     RVx[randindx,]=floor(log(u)/log(1-pmat[j,i]));
    end;

  **print RVx;

  ymat[j,i]=sum(RVx);

   end;
   end;

yvec=shape(ymat`,ntot,1);   ***STRING OUT Y INTO A VECTOR***;
indexvec=shape(indexmat`,ntot,1); ***STRING OUT Drought Indices INTO A
VECTOR***;


FINISH DATAGEN;

run datagen;

datmat=wvec||tvec||yvec||indexvec;

create dat from datmat;
append from datmat;
```

```
truevals=bet0||bet1||bet2||alpha||sig1sq;

create truevals from truevals;
append from truevals;


QUIT;

data dat; set dat;
  rename COL1=id;
  rename COL2=year;
  rename COL3=Ydeaths;
  rename COL4=DroughtIndx;
run;

data dat; set dat;
  Ylt10=0;
  if Ydeaths < 10 then Ylt10=1;

  Yobserved=0;
  if Ydeaths=0 | Ydeaths ge 10 then Yobserved=1;
run;
```

```
                         ****** Model 1 ******;

proc nlmixed data=dat;
        parms bet0=2.25, bet1=.2, bet2=0, alpha=.2, sigsq1=.05;

        bounds sigsq1 >= 0;

        alphainv=1/alpha;
        linp=(bet0+g0i) + bet1*DroughtIndx + bet2*year;
        mu=exp(linp);
        p=1/(1+mu*alpha);

        loglike=lgamma(alphainv+Ydeaths) - lgamma(alphainv) -
        lgamma(1+Ydeaths)
                + Ydeaths*log(1-p) + alphainv*log(p);

        model Ydeaths ~ general(loglike);
        random g0i ~ normal(0, sigsq1) subject=id;
  run;
```

```
                         ****** Model 2 ******;

proc nlmixed data=dat;
       parms bet0=2.25, bet1=.2, bet2=0, alpha=.2, sigsq1=.05;
       bounds sigsq1 >= 0;
       alphainv=1/alpha;
       linp=(bet0+g0i) + bet1*DroughtIndx + bet2*year;
       mu=exp(linp);
       p=1/(1+mu*alpha);
       prYeq0=p**alphainv;
       prYeq1=alphainv*(p**alphainv)*(1-p);
       prYeq2=((alphainv+1)*alphainv/fact(2))*(p**alphainv)*(1-p)**2;

      prYeq3=((alphainv+2)*(alphainv+1)*alphainv/fact(3))*(p**alphainv)
      *(1-  p)**3;

      prYeq4=((alphainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(4))*
      (p**alphainv)*(1-p)**4;

      prYeq5=((alphainv+4)*(alphainv+3)*(alphainv+2)*(alphainv+1)*alpha
      inv/fact(5))*(p**alphainv)*(1-p)**5;

      prYeq6=((alphainv+5)*(alphainv+4)*(alphainv+3)*(alphainv+2)*(alph
      ainv+1)*alphainv/fact(6))*(p**alphainv)*(1-p)**6;

      prYeq7=((alphainv+6)*(alphainv+5)*(alphainv+4)*(alphainv+3)*(alph
      ainv+2)*(alphainv+1)*alphainv/fact(7))*(p**alphainv)*(1-p)**7;

      prYeq8=((alphainv+7)*(alphainv+6)*(alphainv+5)*(alphainv+4)*(alph
      ainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(8))*(p**alphainv)
      *(1-p)**8;

      prYeq9=((alphainv+8)*(alphainv+7)*(alphainv+6)*(alphainv+5)*(alph
      ainv+4)*(alphainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(9))*
      (p**alphainv)*(1-p)**9;

      CDFterm=prYeq0 + prYeq1 + prYeq2 + prYeq3 + prYeq4 + prYeq5 +
      prYeq6 + prYeq7 + prYeq8 + prYeq9;

   ** log-likelihood function when Y values are detectable ***;

      if Yobserved=1 then do;
      loglike=lgamma(alphainv+Ydeaths) - lgamma(alphainv) -
      lgamma(1+Ydeaths)
          + Ydeaths*log(1-p) + alphainv*log(p);
  end;

 ** log-likelihood function when Y values are interval censored on
[1,9] ***;

      else if Yobserved=0 then do;
      loglike=log(CDFterm - prYeq0);
  end;

      model Ydeaths ~ general(loglike);
      random g0i ~ normal(0, sigsq1) subject=id;
  run;
```

```
                        ****** Model 3 ******;


proc nlmixed data=dat;
      parms bet0=2.25, bet1=.2, bet2=0, alpha=.2, sigsq1=.05;
      bounds sigsq1 >= 0;
      alphainv=1/alpha;
      linp=(bet0+g0i) + bet1*DroughtIndx + bet2*year;
      mu=exp(linp);
      p=1/(1+mu*alpha);

      if Ydeaths lt 10 and Ydeaths gt 0 then Ydeaths = 5;


      loglike=lgamma(alphainv+Ydeaths) - lgamma(alphainv) -
      lgamma(1+Ydeaths)
          + Ydeaths*log(1-p) + alphainv*log(p);

      model Ydeaths ~ general(loglike);
      random g0i ~ normal(0, sigsq1) subject=id;
  run;
```

```
                         ****** Model 4 ******;

proc nlmixed data=dat;
       parms bet0=2.25, bet1=.2, bet2=0, alpha=.2, sigsq1=.05;
       bounds sigsq1 >= 0;
       alphainv=1/alpha;
       linp=(bet0) + bet1*DroughtIndx + bet2*year;
       mu=exp(linp);
       p=1/(1+mu*alpha);

       prYeq0=p**alphainv;
       prYeq1=alphainv*(p**alphainv)*(1-p);
       prYeq2=((alphainv+1)*alphainv/fact(2))*(p**alphainv)*(1-p)**2;

      prYeq3=((alphainv+2)*(alphainv+1)*alphainv/fact(3))*(p**alphainv)
      *(1-p)**3;

      prYeq4=((alphainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(4))*
      (p**alphainv)*(1-p)**4;

      prYeq5=((alphainv+4)*(alphainv+3)*(alphainv+2)*(alphainv+1)*alpha
      inv/fact(5))*(p**alphainv)*(1-p)**5;

      prYeq6=((alphainv+5)*(alphainv+4)*(alphainv+3)*(alphainv+2)*(alph
      ainv+1)*alphainv/fact(6))*(p**alphainv)*(1-p)**6;

      prYeq7=((alphainv+6)*(alphainv+5)*(alphainv+4)*(alphainv+3)*(alph
      ainv+2)*(alphainv+1)*alphainv/fact(7))*(p**alphainv)*(1-p)**7;

      prYeq8=((alphainv+7)*(alphainv+6)*(alphainv+5)*(alphainv+4)*(alph
      ainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(8))*(p**alphainv)
      *(1-p)**8;

      prYeq9=((alphainv+8)*(alphainv+7)*(alphainv+6)*(alphainv+5)*(alph
      ainv+4)*(alphainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(9))*
      (p**alphainv)*(1-p)**9;

      CDFterm=prYeq0 + prYeq1 + prYeq2 + prYeq3 + prYeq4 + prYeq5 +
      prYeq6 + prYeq7 + prYeq8 + prYeq9;

  ** log-likelihood function when Y values are detectable ***;

      if Yobserved=1 then do;
      loglike=lgamma(alphainv+Ydeaths) - lgamma(alphainv) -
      lgamma(1+Ydeaths)
          + Ydeaths*log(1-p) + alphainv*log(p);
  end;

 ** log-likelihood function when Y values are interval censored on
[1,9] ***;

      else if Yobserved=0 then do;
      loglike=log(CDFterm - prYeq0);
  end;

      model Ydeaths ~ general(loglike);run;
```

South African Data SAS Code

```
                    ****** Model 1 ******;

proc nlmixed data=format;
      parms bet0=3.2076, bet1= -0.04197, bet2=.000151, alpha=.1950,
      sigsq1=.1.3586;
      bounds sigsq1 >= 0;
      alphainv=1/alpha;
      linp=(bet0+g0i) + bet1*exposure + bet2*date + lnpop;
      mu=exp(linp);
      p=1/(1+mu*alpha);

      loglike=lgamma(alphainv+deaths) - lgamma(alphainv) -
      lgamma(1+deaths)+ deaths*log(1-p) + alphainv*log(p);

      model deaths ~ general(loglike);
      random g0i ~ normal(0, sigsq1) subject=district;

      ods output parameterestimates=ests1;
      title1 'NB regression with random effects, using general LL
      facility';
run;

data ests1;
      set ests1;
      model = '1';
run;

proc print data = ests1;
run;
```

```
                          ****** Model 2 ******;

  proc nlmixed data=format;
      parms bet0=3.2076, bet1= -0.04197, bet2=.000151, alpha=.1950,
      sigsq1=.1.3586;
      bounds sigsq1 >= 0;

      alphainv=1/alpha;
      linp=(bet0+g0i) + bet1*exposure + bet2*date + lnpop;
      mu=exp(linp);
      p=1/(1+mu*alpha);

      prYeq0=p**alphainv;
       prYeq1=alphainv*(p**alphainv)*(1-p);
       prYeq2=((alphainv+1)*alphainv/fact(2))*(p**alphainv)*(1-p)**2;

      prYeq3=((alphainv+2)*(alphainv+1)*alphainv/fact(3))*(p**alphainv)
      *(1-p)**3;

      prYeq4=((alphainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(4))*
      (p**alphainv)*(1-p)**4;

      prYeq5=((alphainv+4)*(alphainv+3)*(alphainv+2)*(alphainv+1)*alpha
      inv/fact(5))*(p**alphainv)*(1-p)**5;

      prYeq6=((alphainv+5)*(alphainv+4)*(alphainv+3)*(alphainv+2)*(alph
      ainv+1)*alphainv/fact(6))*(p**alphainv)*(1-p)**6;

      prYeq7=((alphainv+6)*(alphainv+5)*(alphainv+4)*(alphainv+3)*(alph
      ainv+2)*(alphainv+1)*alphainv/fact(7))*(p**alphainv)*(1-p)**7;

      prYeq8=((alphainv+7)*(alphainv+6)*(alphainv+5)*(alphainv+4)*(alph
      ainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(8))*(p**alphainv)
      *(1-p)**8;

      prYeq9=((alphainv+8)*(alphainv+7)*(alphainv+6)*(alphainv+5)*(alph
      ainv+4)*(alphainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(9))*
      (p**alphainv)*(1-p)**9;

       CDFterm=prYeq0 + prYeq1 + prYeq2 + prYeq3 + prYeq4 + prYeq5 +
      prYeq6 + prYeq7 + prYeq8 + prYeq9;

   ** log-likelihood function when Y values are detectable ***;

      if Yobserved=1 then do;
      loglike=lgamma(alphainv+deaths) - lgamma(alphainv) -
      lgamma(1+deaths)
          + deaths*log(1-p) + alphainv*log(p);
  end;

   ** log-likelihood function when Y values are interval censored on
[1,9] ***;

      else if Yobserved=0 then do;
      loglike=log(CDFterm - prYeq0);
  end;
  model deaths ~ general(loglike);
  random g0i ~ normal(0, sigsq1) subject=district;

  ods output parameterestimates=ests2;
title1 'NB regression with random effects, Interval censored [1-9]’;
run;

data ests2;
set ests2;
model = '2';
run;

proc print data = ests2;
run;
```

```
                         ****** Model 3 ******;

proc nlmixed data=format;
      parms bet0=3.2076, bet1= -0.04197, bet2=.000151, alpha=.1950,
      sigsq1=.1.3586;

      bounds sigsq1 >= 0;

      alphainv=1/alpha;
      linp=(bet0+g0i) + bet1*exposure + bet2*date + lnpop;
      mu=exp(linp);
       p=1/(1+mu*alpha);

      if deaths lt 10 and deaths gt 0 then deaths = 5;

      loglike=lgamma(alphainv+deaths) - lgamma(alphainv) -
      lgamma(1+deaths) + deaths*log(1-p) + alphainv*log(p);

      model deaths ~ general(loglike);
      random g0i ~ normal(0, sigsq1) subject=district;

      ods output parameterestimates=ests3;
      title1 'NB regression W/ Midpoint Imputation';
run;

data ests3;
set ests3;
model = '3';
run;

proc print data = ests3;
run;
```

```
                         ****** Model 4 ******;


  proc nlmixed data=format;
      parms bet0=3.2076, bet1= -0.04197, bet2=.000151, alpha=.1950,
      sigsq1=.1.3586;
      bounds sigsq1 >= 0;
      alphainv=1/alpha;
      linp=(bet0) + bet1*exposure + bet2*date + lnpop;
      mu=exp(linp);
      p=1/(1+mu*alpha);

      prYeq0=p**alphainv;
      prYeq1=alphainv*(p**alphainv)*(1-p);
      prYeq2=((alphainv+1)*alphainv/fact(2))*(p**alphainv)*(1-p)**2;

      prYeq3=((alphainv+2)*(alphainv+1)*alphainv/fact(3))*(p**alphainv)
      *(1-p)**3;

      prYeq4=((alphainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(4))*
      (p**alphainv)*(1-p)**4;

      prYeq5=((alphainv+4)*(alphainv+3)*(alphainv+2)*(alphainv+1)*alpha
      inv/fact(5))*(p**alphainv)*(1-p)**5;

      prYeq6=((alphainv+5)*(alphainv+4)*(alphainv+3)*(alphainv+2)*(alph
      ainv+1)*alphainv/fact(6))*(p**alphainv)*(1-p)**6;

      prYeq7=((alphainv+6)*(alphainv+5)*(alphainv+4)*(alphainv+3)*(alph
      ainv+2)*(alphainv+1)*alphainv/fact(7))*(p**alphainv)*(1-p)**7;

      prYeq8=((alphainv+7)*(alphainv+6)*(alphainv+5)*(alphainv+4)*(alph
      ainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(8))*(p**alphainv)
      *(1-p)**8;

      prYeq9=((alphainv+8)*(alphainv+7)*(alphainv+6)*(alphainv+5)*(alph
      ainv+4)*(alphainv+3)*(alphainv+2)*(alphainv+1)*alphainv/fact(9))*
      (p**alphainv)*(1-p)**9;

       CDFterm=prYeq0 + prYeq1 + prYeq2 + prYeq3 + prYeq4 + prYeq5 +
      prYeq6 + prYeq7 + prYeq8 + prYeq9;

   ** log-likelihood function when Y values are detectable ***;

       if Yobserved=1 then do;
       loglike=lgamma(alphainv+deaths) - lgamma(alphainv) -
      lgamma(1+deaths)   + deaths*log(1-p) + alphainv*log(p);
  end;

   ** log-likelihood function when Y values are interval censored on
[1,9] ***;

      else if Yobserved=0 then do;
      loglike=log(CDFterm - prYeq0);
  end;

      model deaths ~ general(loglike);
      ods output parameterestimates=ests4;
      title1 'NB regression with fixed effects';
run;

data ests4;
set ests4;
model = '4';
run;

proc print data = ests4;
run;
```

```
            ****** Examine the Exposure in Quartiles ******;

proc glimmix data = Quartiles;
      class district Quarter (ref = '4');
      model deaths = Quarter date / dist=negbin link=log solution
      offset=lnpop;
      random intercept / sub= district;
      estimate 'Rate ratio of Q1 vs Q4' Quarter 1 0 0 -1 / exp cl;
      estimate 'Rate ratio of Q2 vs Q4' Quarter 0 1 0 -1 / exp cl;
      estimate 'Rate ratio of Q3 vs Q4' Quarter 0 0 1 -1 / exp cl;
run;
```