2009

EVALUATING ECOLOGICAL STUDIES IN EPIDEMIOLOGY

Brenden T. Dufault

EVALUATING ECOLOGICAL STUDIES IN EPIDEMIOLOGY

(Spine title: Evaluating Ecological Studies in Epidemiology)

(Thesis format: Monograph)

By

Brenden T. Dufault

Graduate Program in Biostatistics and Epidemiology

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
University of Western Ontario
London, Ontario, Canada

ABSTRACT

The purpose of this research was twofold. The first objective was to assess the quality of modern cross-sectional ecological studies with a bibliometric review. The second objective was to investigate via simulation study the reliability of common ecological regression models for analysing count data.

The bibliometric review found that the quality and areas of application of the ecological literature is quite diverse. However, a large proportion of studies exhibited poor statistical practice and provided insufficient amounts of justification and information.

Linear, weighted linear, Poisson, and negative binomial regression were included in the simulation study based on their prevalence in the bibliometric review. The Poisson and negative binomial models had overly-liberal Type I error rates when faced with overdispersion or small samples respectively. Linear and weighted linear regression had highly robust Type I error rates. For all models, power decreased primarily as a function of overdispersion.

Key Words: ecological, cross-sectional, bibliometric review, simulation, Poisson, negative binomial, linear regression, epidemiology

ACKNOWLEDGMENTS

I must first and foremost gratefully acknowledge the generous help of my supervisor, Dr. Neil Klar, whose benevolent patience somehow always managed to exceed my ignorance. I am indebted to his perpetual insight, guidance, and mentorship. This thesis is the compilation of a long personal journey, and throughout each part of it his considerable expertise and, equally importantly, kind encouragement have been freely gifted to me. I could have had no better teacher these past two years.

I would also like to thank my advisor Dr. Yun-Hee Choi for donating so much of her time, and for providing me with help, support, and suggestions as I stumbled through this. I must further thank Dr. John Koval, who helped tremendously on several occasions when SAS confounded and foiled me. I am honoured to have been the benefactor of the knowledge and experience of the entire faculty and staff, who have continuously inspired and humbled me.

Finally, and with all my affection, I thank my family and friends. It will be a life's work to show you the depth of my love.

Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

Introduction

1.1 Introduction

The modern epidemiologist is well acquainted with a growing number of clinical and observational study designs, all of which have unique strengths and limitations that must be considered when undertaking etiologic research. One of the oldest and yet most troubling in terms of its limitations is the ecological design, which has been employed by statisticians and sociologists since the 19^{th} century (Dogan and Rokkan, 1969). The ecological design is the only design to be solely concerned with analyzing the exposures and health outcomes of groups of individuals, as opposed to individuals themselves. It seemingly presents an intuitive and convenient way of exploring social, political, environmental, and epidemiological hypotheses when aggregate data are available, and especially when individual data are not. As Wakefield (2008) points out, this appeal is particularly relevant in the modern era, where medical, governmental, and research institutions frequently amass large amounts of electronic information on the health risks and outcomes of various populations. For these reasons, ecological designs are frequently used by epidemiologists and social scientists.

In spite of the long history of the design and the allure of its simplicity, it is a method that has inspired a considerable amount of debate and discussion in the literature. Ever since Robinson (1950) popularized what is now referred to as the *ecological fallacy* in his seminal article published in the *American Sociological Review*, the research community has been aware

of and has been contending with this potential bias, which is unique to ecological analysis. The ecological fallacy, perhaps more aptly referred to as *cross level bias*, is committed when causal associations noted at the group level are incorrectly assumed to apply to the individuals within those groups. For example, Hemenway *et al.* (2001) conducted a study in which they investigated the association between prevalence of social trust and the prevalence of firearm ownership. They noted a positive association between levels of mistrust in a community and the prevalence of firearm ownership. Based on this analysis, it would be tempting to naively conclude that individuals less trusting of others are likelier to own a gun. While that may be true, it does not immediately follow from the association noted at the group level. In fact, it is entirely possible to observe a positive group-level association even though individuals who are more trusting are the owners of the firearms in the communities. That is, precisely the opposite relation may be observed at the individual level.

Ever since the ecological fallacy was famously described by Robinson (1950), much attention has been devoted to investigating the causes of the fallacy and to finding potential remedies for it. Essentially, the ecological fallacy is possible because the marginal nature of the data that exists at the aggregate level for each group is not uniquely determined by the internal frequencies on which the individual associations are based (Robinson, 1950). This means that analyses conducted at the group level are typically unable to capture or utilize information on the distributions of the exposures, effect modifiers, and confounders within each group (Wakefield, 2008). Although not immediately obvious, the face value of this concept is not especially difficult to grasp once it has been pointed out. It is perhaps surprising, then, that the causes and conditions for a lack of correspondence between group- and individual-level analyses are complex and often quite contrary to the reasons for bias at the individual level. Indeed, many causes of ecological bias could even be described as counter-intuitive. Conditions resulting in the ecological fallacy will be introduced in Section 1.2 and described in detail in Chapter 3.

1.2 Challenges of Ecological Studies

The efforts to reveal and understand the limitations of ecological studies primarily began with Robinson (1950), although the phenomenon was discussed by Thorndike (1939). Knapp (1977) reviews earlier statistical contributions to differences between analyses of individual and grouped observations dating back to the 19th century.

In addition to popularizing the term 'ecological fallacy', Robinson was also the first to attempt the construction of a formal association between individual-level and group-level correlations. Using ANOVA, he was able to show that the ecological correlation is a function of the overall individual-level correlation and the individual-level correlations calculated within each separate group, thus quantifying the potential discrepancy between the two levels of analysis.

Robinson (1950) also claimed that there is a positive relationship between the magnitude of the ecological correlation coefficient and the average size of the groups in the study. However, Knapp (1977) provided a more mathematically rigorous discussion and showed that this relationship does not hold in general. Nevertheless, it serves as an example of just how dubious ecological regression can be. Duncan *et al* (1961) extended Robinson's work using a similar approach, and were able to demonstrate the relationship between individual-level and ecological estimates of slope in a linear regression analysis. They show that the ecological regression coefficient is a weighted difference. In its case, however, the difference is between the average within-groups regression coefficient and the overall individual-level regression coefficient, the latter often being the etiologic relation of interest. It is important to note that even if the individual and grouped regression coefficients are in agreement, the correlation coefficients can still differ, a result that is surely not intuitive yet typical of ecological analysis. Such results make it clear that ecological analyses conducted with Pearson correlation coefficients alone are often on insufficient grounds for making strong etiologic claims.

More fundamental insights into the underlying reasons that give rise to the ecological fallacy when conducting linear regression are provided by Greenland and Morgenstern (1989). The

two main sources of bias unique to aggregate studies are confounding by group and effect modification by group, which are present in addition to the sources of confounding that can exist at the individual level.

Confounding by group may occur when there is an ecological correlation across groups between the disease rates in the unexposed and the average exposure level. Determining whether or not this is the case likely requires individual-level data, which is often unavailable to the researcher. Effect modification by group, when estimated using linear regression models (i.e. on an additive scale), can occur if the rate difference between the exposed and unexposed groups varies across groups at the individual level.

Either of these may come to pass if confounders or effect measure modifiers are unequally distributed across groups; if there is a contextual (ecological-level) effect of exposure that exists in addition to the individual-level exposure; or if the probability of the outcome for an individual depends on the occurrence of that outcome in other members of the same group, which may be the case for social phenomenon or infectious diseases.

Unfortunately, the complications extend beyond linear models. Richardson *et al* (1987) argue that while there is no mathematical bias resulting from aggregation when calculating relative risks based on simple linear risk models without confounders or effect modifiers, convex risk models such as the exponential model will have a multiplicative bias coefficient. Greenland (1992) highlights another issue: when dealing with ecologic log-linear models that have a main effect of interest and a set of confounders, their effects are implicitly assumed to be multiplicative with respect to the rate ratio. However, this is a biased estimate of the individual-level rate ratio, even if there is no unaccounted bias and the individual-level effects are also multiplicative. In addition, Wakefield (2008) points out that when a nonlinear risk model is aggregated over individuals, the resulting ecological risk model deals only with contextual effects and not individual effects, since the risk in this aggregate model depends solely on the proportion of exposed persons in the group. This is referred to as pure specification bias. Unless there is homogeneity of exposure within the group, this contextual effect can only correspond to the

individual association if average group exposure is the true causal factor.

The absence of individual-level data may also compromise attempts to adjust for confounding. That is, some conditions that allow for confounding in an ecological study do not present any opportunities for bias in an individual-level study. Specifically, a factor may cause ecological confounding even if it is unassociated with exposure in each group, so long as it is associated with the exposure at the aggregate level (Greenland and Morgenstern, 1989). This is in stark contrast to the individual-level condition for confounding, where a factor must be associated with the exposure in order to introduce any bias. Fortunately, this feature may also allow ecological studies to be safe from some individual-level bias: a within-group confounder may not produce ecological bias if it is unassociated with the exposure at the aggregate level.

Compounding the problem is the fact that control for confounders is more difficult in ecological analyses (Greenland and Robins, 1994). It is possible to perfectly measure and adjust for confounders, and yet have the ecological bias be increased, not attenuated. In a similar fashion, a lack of additivity at the individual level within groups may be a source of ecological bias (Richardson and Hemon, 1990). Unlike individual-level analyses, however, an ecological analysis cannot adjust for this by including a multiplicative interaction term in the model unless certain strict assumptions are met. Since verification of these assumptions requires knowledge about the joint distributions within groups, they are often untenable.

1.3 A Defense of Ecological Studies

Faced with challenges such as these, it might appear that ecological studies should be avoided. Indeed, this is what Robinson (1950) asserted when he discussed the ecological fallacy. Thankfully, this is not the case for several reasons. Chief among them is a philosophical shift in the concept of causation and the scope of epidemiology. While it was once believed that health outcomes could be explained solely by the causal processes at the level of the individual, it has

come to be understood that human health is more likely the result of coinciding factors that interact on multiple levels (Schwartz, 1994). March and Susser (2006) note that this multilevel conceptual approach to epidemiology can originally be traced as far back as Morris (1957), and is implicitly evident in the works of epidemiologists like John Snow. This view of causation was superseded by the individualistic risk-factor paradigm, which followed in the wake of medical advances achieved in the mid twentieth century. The undeniable impact of vaccines, antibiotics, and pharmaceuticals made ecological factors appear largely obsolete. The 1990s saw a return to the multilevel viewpoint as the reductionist approach proved to be inadequate in the face of modern infectious disease epidemics and for new branches of epidemiology. For example, the importance of contextual effects such as economic disparity, which exists solely at the group level, are central to some hypotheses of social epidemiology (Oakes and Kaufman, 2006).

Ecological studies need not be rough approximations to individual-level studies. Rather, they can instead be concerned with analyzing contextual effects at the group level, and since these processes are legitimate from a causal point of view, they are etiologic studies in their own right (Schwartz, 1994). Whether or not this group-level approach truly answers the question of interest is a matter of construct validity. It is the duty of the researcher to ensure that the level of measurement is suitable for the hypothesis at hand.

Ecological studies have the added benefit of being an inexpensive way to describe spatial variation in exposures and disease. This can be useful for directing policy and health interventions, often by quite literally mapping the associations between the health outcomes, social burdens, risks, and infrastructures in a given society. Szwarcwald *et al* (2000) adopted this very approach to determine clusters of mortality and to provide descriptive geographical relationships between health and socioeconomic conditions in Rio de Janeiro.

Another attractive feature of ecological studies is that they may allow the researcher to capture a sufficient amount of variation in the exposure variable if individuals in a given population are largely homogeneous (Morgenstern *et al*, 2008). This may be the case for social exposures that encompass large groups of people, such as laws or policies enacted at the

provincial level, or environmental exposures that do not vary greatly from region to region, such as the average annual hours of sunlight.

As previously mentioned, ecological analyses may be immune to individual-level confounders so long as they are unassociated with the exposure at the group level. This may be a crucially advantageous feature for certain etiologic questions, particularly those that might otherwise be susceptible to confounding by indication. Johnston (2000) tested the efficacy of this approach by comparing an ecological analysis of hospital treatment and mortality to an individual-level one, and found that the ecological approach reversed the direction of the association. Johnston concludes that the ecological analysis was able to counteract the effects of confounding by indication, and recommends the inclusion of group-level data in studies that may be vulnerable to this bias. Similarly, since ecological models rely on aggregate data that often comes from records and not from interviews or surveys, they may be relatively untouched by some issues that plague individual-level studies, such as recall bias or response bias (Schwartz, 1994).

Ecological studies present unique challenges. This design should be treated with utmost caution and used only when the researcher is fully informed of their limitations and either seeks to minimize the opportunities for bias or tempers the results with appropriate warnings. That being said, they are a class of study that remains useful and should be kept in the toolbelt of the epidemiologist. In addition to supplementing and motivating more common individual-level analysis, they provide a means of investigating many of the etiologic hypotheses that are continually becoming more relevant to the health concerns of contemporary researchers.

1.4 Types of Ecological Studies

The discussion has thus far referred to an over-arching ecological design without making reference to any possible subtleties or variations on it. However, ecological studies may be further classified by their type of design. Borrowing from the terminology of Koepsell and Weiss (2003) and Morgenstern (1995), we have the following main sub-designs:

- *Spatial ecological studies.* These compare the health outcomes and exposures of several groups for a given point in time or time period. Here, the designs do not statistically take into account any formal temporal considerations, and may be case-control, cross-sectional, etc. It is the simplest and most common class of ecological designs, with the cross-sectional sub-design being the simplest.

- *Longitudinal ecological studies.* These often employ time-series analysis to compare the change in outcome rates within a study group over periods of time. For example, Analitis *et al.* (2006) compared cardiovascular mortality rates in a pooled analysis of different cities by looking at the air pollution concentrations at different time lags.

- *Mixed ecological studies.* Also referred to as spatiotemporal designs, these combine temporal and spatial aspects into one model and investigate them simultaneously.

It can be easily seen that the ecological design encompasses all of the main approaches that are present in individual-level designs. In fact, when one considers that a set of ecological units are no different than a set of individuals as far as most study *designs* are considered, then it becomes apparent that the ecological class mirrors most of the individual-level designs.

1.5 Objectives

The objectives of this thesis are twofold:

(1) To describe and critically analyze the quality of ecological studies in the current epidemiological literature. This will be accomplished by conducting a bibliometric review of ecological studies published in select epidemiology journals. By so doing, we aim to provide an informed report of the strengths and shortcomings of the literature so that future researchers engaging in ecological analyses may improve upon their predecessors.

(2) To empirically investigate the statistical performances of the analytic methods most prevalent in the ecological studies that were identified in the bibliometric review. This will be achieved via simulation study, in which a computer code shall generate group-level data. By running a simulation study, we can iterate this process thousands of times and therefore make valid inferences regarding the average performances of these methods. Ultimately, we wish to provide researchers with a set of recommendations that they may consult when embarking on ecological studies of an etiologic nature.

However, this thesis shall not focus its attention on all types of ecological design. Since the cross-sectional design is the most widespread, and faced with the sheer volume and complexity of the analytic approaches available to the mixed and longitudinal designs, this thesis will only consider the studies and methods relevant to the cross-sectional design. This approach allows us to provide a more concentrated discussion.

1.6 Rationale

It is felt that the bibliometric review contained in this thesis shall fulfill a present need in the literature that has remained essentially unsatisfied. Insofar as we are aware, there is no published bibliometric review concerning the quality of recent epidemiological papers that rely on the ecological design. The only review comparable to the one we have proposed was conducted by Riva *et al.* (2007), in which the authors compile an analysis of all multilevel studies of small-area effects that were conducted between July 1998 and December 2005. Even though they considered topics of contextual (ecological) effects in addition to individual ones, and even though they presented a discussion regarding some of the more common conceptual and methodological issues surrounding multilevel studies and the ecological aspects thereof, the scope of their paper is considerably different from what this thesis proposes. Whereas the paper by Riva *et al.* focuses strictly on multilevel analyses and places more emphasis on the health outcomes studied rather than the methodological approaches, the bibliometric analysis contained herein will be tailored to ecological studies and will more critically analyze the statistical aspects performed and reported in each paper.

A thorough systematic review on this topic will be important for the epidemiological community. Since it is a diverse group comprised mainly of health researchers that lack vigorous training in statistics, it is occasionally necessary for the technical aspects of the literature to be scrutinized. For example, Weitzen *et al.* (2004) found that many medical research papers employing propensity score matching used the technique poorly and even more reported it inadequately. We will perform the same service for ecological studies.

The proposed simulation study will provide novel insight into the performance of key ecological regression models. While purely analytic work is more rigorous and has all the advantages of precise mathematics, it is often limited to asymptotic inferences for many modeling techniques. This prevents us from investigating small-sample behaviour and restricts us to generalities. A simulation, on the other hand, would allow us to simultaneously explore the effects of multiple parameters in a more realistic setting.

Previous bodies of work examining the ecological performances of statistical methods have had one of three general approaches:

- The first approach is to run the chosen statistical methods on real datasets and then compare their results to individual-level and multilevel associations, which are held as the gold standard. This was precisely the tactic adopted by Lancaster *et al.* (2006) when they analyzed such exotics as the stratified ecological model, the aggregated compound model, and the aggregated individual model. No inferences regarding the average performances of the studied methods are possible with this approach.

- The second general approach seen in the literature is to exploit computer simulation in order to evaluate new techniques that supplement ecological analysis with individual-level information or within-group distributional assumptions. For example, Guthrie and Sheppard (2001), as well as Wakefield and Haneuse (2008), used simulation studies to analyze novel methods that either aggregate individual-level data in such a way as to retain covariate information or combine the ecological information with case-control data.

- The final major approach compares advanced methods that have been adapted from the geostatistical literature, via real or simulated data. These methods are mainly suitable for the mapping of disease rates or the identification of spatial clustering. The papers by Green and Richardson (2002) and Kelsall and Wakefield (2002) typify this type of research, and are perhaps the papers most explicitly related to ecological studies. Their works investigate methods such as Gaussian random fields, spatially correlated Potts models, and spatial Markov models.

None of the approaches mentioned above focus on the common methods of analysis most relied on by epidemiologists while at the same time providing summary measures of performance gleaned from large-sample inferences. The simulation study we propose is oriented to serve the pragmatic needs of the research community and will present a unique contribution to the literature.

Since the ecological study is a design that has been well established in the literature, it is certainly worthwhile to evaluate how it has been applied and conducted, and to offer some empirically derived guidelines for proper statistical analysis. The objectives contained in this thesis seek to address these needs and to further investigate a research tool that has made significant contributions to human health, public policy, and social epidemiology.

Chapter 2

Bibliometric Review

2.1 Introduction

The objective of this bibliometric review is to evaluate how well modern ecological studies are designed, analysed, and reported. Sections 2.2.1 through 2.2.4 outline the methods and approaches by which we set out to achieve this, including: a rationale for the chosen journals and the selected time period; the inclusion and evaluation criteria for each study; and a rationale for the search method used to find publications. Results are presented in section 2.3, and are, followed by a discussion and summary of recommendations for future research in section 2.4.

2.2 Methods

2.2.1 Journals Studied

As Altman (1994) has noted, the technical quality and statistical legitimacy of medical research can often be lacking. Although he was referring mainly to publications found in clinical journals such as the British Medical Journal, his statement unfortunately holds a certain degree of truth for many research topics that pertain to human health, including epidemiology. Exactly how relevant is Altman's claim to the cross-sectional ecological studies undertaken by the epidemiological community? To answer this question, we concern our bibliometric review with a carefully chosen, specific set of journals. Although epidemiologists consult and publish in a wide variety of journals, it is not practical for one review to be simultaneously thorough

and comprehensive; the literature is simply too vast. Thus, we have selected *The American Journal of Epidemiology, Annals of Epidemiology, Epidemiology, International Journal of Epidemiology, Journal of Clinical Epidemiology,* and the *Journal of Epidemiology and Community Health* as the journals to be included in our study. We feel that these most explicitly and singularly serve the needs of epidemiologists and represent their works. Moreover, each of these journals has a significant impact factor and a high ranking within either epidemiology or public health. Thus, the significance of these journals as general, non-specialized forums for the research of epidemiologists makes them natural representatives of the discipline.

Any bibliometric review requires a second parameter in addition to its set of journals: a time period of interest. Unfortunately, there is no ecological equivalent to the Strength of Reporting of Observational Studies in Epidemiology (STROBE) guidelines whose date of publication we could use as a meaningful marker (von Elm *et al.*, 2007). We were therefore forced to base our choice on different grounds. The choice of our time period was motivated by the need to capture a significant amount of studies for the review, offset by the need to obtain a manageable amount of information. These considerations were coupled with the desire to make inferences regarding only the more recent publications. Thus, we have chosen to analyze all papers published between the years 2000 and 2008 inclusively, allowing us to focus on the works of the new millennium.

2.2.2 Search Method

All of the papers contained in this bibliometric review were found by searching through each issue of all selected journals for the time period January 1, 2000 through December 31, 2008. An alternative approach is to conduct computer searches of databases such as MEDLINE with the use of strategic keywords or MESH groups. While this may be an adequate means of obtaining many types of studies, we felt that the ecological design was so varied in its applications that some epidemiologists may use it unknowingly or without including the kinds of proper keywords that would be picked up by search engines or database categorization.

Our exhaustive search method allowed us to circumvent this potential problem and provide a definitive review containing all of the cross-sectional ecological papers in the selected journals.

2.2.3 Inclusion and Exclusion Criteria

Each potential study had to satisfy a number of inclusion criteria before it was admitted into the review. The research paper in question had to use a cross-sectional ecological design. This meant that ecological studies using temporal or spatiotemporal designs were excluded from the review, as were case-control, case-crossover, or cohort designs. For example, Lewis *et al* (2001) analyzed workplaces to determine risk characteristics for homicide. Workplaces were selected based on whether or not they had experienced the outcome of interest, namely homicide, which makes their study design case-control and not cross-sectional.

In addition, each study needed to present an original work with an etiologic research hypothesis. Consequently, descriptive studies of spatial trends were excluded from the analysis. Another consequence of this criterion was to exclude studies that introduced, discussed, or evaluated ecological methods with examples using real data, since the aim of such a study is not to establish causal associations. Likewise, ecological simulation studies were not included in our review. A study did not need to be entirely ecological in order to be included, so long as its ecological aspects met our inclusion criteria. That is, some ecological analyses were performed within papers that also employed multilevel techniques or conducted individual-level analyses, and we included these as well.

For certain papers, often those employing geospatial techniques, it was unclear whether they were testing etiologic hypotheses or if they were simply mapping the occurrences of diseases and risk factors for the purposes of public health policy. We reviewed these studies until a consensus could be reached. If a paper was deemed etiologic after such a series of reviews, it was included and then marked as once-tentative for the sake of our analysis.

We did not discriminate with regards to the length of the paper so long as it met the inclusion criteria mentioned above. Thus, papers that were deliberately brief and relegated to a

segment such as the "Short Report" section of the *Journal of Epidemiology and Community Health* were also included. These were noted as short reports for the bibliometric analyses.

2.2.4 Evaluation Criteria

The complexities and concerns of the ecological design are such that any good ecological paper will be especially cautious and premeditated. It should justify the choice of design and present its results in a careful, transparent fashion. The discussion should be written such that even readers unfamiliar with the ecological fallacy are mindful of its potential when making individual-level inferences. In other words, a proper ecological study needs to tread carefully and should not be undertaken lightly. When conducting this bibliometric review, it was therefore necessary to analyze the quality of the writing and reporting in addition to the statistical methodology. Thus, we designed each criterion to focus on one of the three major fundamentals that comprise any etiologic epidemiological publication. Specifically, we critiqued each paper according to a set of *a priori* criteria that collectively analyzed the following characteristics given below. Table 2.1 provides a summary of the criterion used in this bibliometric review.

1.*Aspects of Study Design*

Aspects of study design which were examined include: the sample size of each analysis, given by the number of ecological units; the level of aggregation at which the ecological units exist, recorded as, from smallest to largest: census tracts and neighbourhoods, municipalities, municipal areas or counties, states/provinces, or nations; whether the etiologic inference was at the group- or individual-level; and whether the ecological units were explicitly chosen (or constructed) to suit the etiologic hypothesis, or if they were seemingly motivated by convenience or necessity. By their nature, the two latter criteria had to occasionally be decided subjectively.

To see which areas of health and social epidemiology are presently the focuses of the cross-sectional ecological design, we also recorded the health outcomes according to consistent themes. We first grouped the outcomes into one of six broad classifications: mortality rates, incidence

rates, count data (number of events), prevalences, health services outcomes, or life expectancy. When possible, the outcomes in each classification were then further refined and given one of the following sub-classifications, which were motivated by the literature itself: cancer; cardiovascular disease; respiratory disease; other noninfectious diseases; infectious diseases; suicide or self harm; and criminal activity. Often, a given study had more than one outcome and thus more than one set of classifications. In this case, all were recorded.

2. *Statistical Methods and Practices*

The primary interest here was the analytic methodology employed by the authors. Therefore, we recorded any formal, statistical ecological analyses that were motivated by the etiologic hypothesis of the paper. To facilitate comparisons across studies, we tried to use consistent terminology whenever possible. For example, if it was appropriate, the regression techniques were often recorded as one of: Ordinary Least Squares (OLS) regression, Weighted OLS regression, Poisson regression, or Negative Binomial regression. Noting the statistical methods was intended to provide a measure of the analytical sophistication, diversity, and propriety of current ecological studies.

To more closely scrutinize the regression methods, we also included a measure of statistical validity. For each regression analysis, we recorded whether there were fewer than ten ecological units per covariate, more than ten ecological units per covariate, or more than twenty ecological units per covariate. Although ten is merely a rough guide, Vittinghoff and McCulloch (2007) have argued that fewer than ten observations per covariate may lead to invalid statistical inferences. Since the level of aggregation in any given study will be a limiting factor for sample size, and since the level of aggregation is expected to be highly variable across studies, we felt that this criterion would be especially relevant for the ecological design.

We also recorded the use or non-use of ecological covariates in the analyses. Although more complex, confounding still occurs at the ecological level and under the right conditions may still be accounted for by the inclusion of covariates. Since covariate use is one of the foundational

premises of epidemiological analysis, we felt that our statistical critique should include this criterion. For the same reason, we also recorded whether or not the authors performed sample size calculations before beginning their study; an often-neglected consideration.

Spatial autocorrelation is a potential concern for many ecological analyses, since most of them use aggregated areal data. Whether researchers should adjust for spatial autocorrelation or include spatial dependence terms in their ecological models is presently debated. The concern is that the omission of spatial effects may cause a loss of information that could otherwise improve the estimation of model parameters. However, if the exposure has spatial structure, adjustment for spatial effects may 'adjust away' the causal association being studied (Wakefield, 2003). Although this is a controversial topic, we nonetheless recorded the number of studies accounting for spatial effects.

Our final statistical criterion determines whether or not appropriate regression covariates have been adjusted for age, gender, or both, if the outcome has been similarly adjusted. Rosenbaum and Rubin (1984) point out that area-based studies commonly regress age-adjusted outcomes on unadjusted predictors, and go on to state that this method often produces biased effect estimates for risk models that are linear at the individual level. They recommend that regression analyses with aggregate data incorporate either adjusted covariates, or crude covariates along with moments of the population distribution for the adjusted variable. We report how often this is done in practice.

3. *Reporting*

The sensitive nature of ecological studies requires that they be held to guidelines additional to those presently found in the STROBE statement. In this section, our objective was therefore to evaluate how well each paper presented and discussed the ecological aspects of its design and conclusions. We first searched the paper to see whether the authors mentioned one of the keywords 'ecologic', 'ecological', or 'aggregate' at any point in their article (excluding references). This criterion sought to test the proportion of authors that clearly inform the

reader of the study design. The next criterion analyzed whether or not the authors explicitly justified an ecological analysis, since it should be made evident to the reader that the design was not chosen lightly. A justification allows the reader to be informed as to the rationale, and also to decide for themselves whether or not the study design is truly warranted. It was not deemed essential for the justification to mention either of the above keywords, so long as it presented a clear explanation as to why an ecological analysis was either necessary or preferable.

Our last criterion recorded whether or not the authors sufficiently cautioned the interpretation of their results against undue individual-level inferences, preferably with the use of terms such as 'ecological fallacy' or 'cross-level bias'. The authors' statement did not need to be elaborate or include an in-depth explanation of the fallacy itself. It merely needed to point out that the results were not necessarily applicable at different levels of aggregation. We feel that it is the responsibility of every ecological researcher to address this concern, whether their stated inferences are fallible or not. It cannot be taken for granted that the reader is aware of the ecological fallacy, or that they have deduced the study design if it has not been stated by the authors.

Table 2.1: Summary of the Evaluation Criteria Used in the Bibliometric Review

Evaluation Criterion	Description
Study Design	
Sample Size	How many ecological units are used in the statistical analysis?
Level of Aggregation	At which level of aggregation do the ecological units exist? (e.g., census tracts or nations.)
Level of Inference	Are etiologic inferences made at the ecological level or the individual level?
Pre-specified Ecological Units	Are the ecological units selected or constructed *a priori*, or are they seemingly selected by convenience or necessity?
Health Classification of Primary Study Outcome(s)	To which one of our health classifications does the primary study outcome(s) belong? (e.g. cancer, infectious disease, cardiovascular diease, etc.)
Statistical Methods and Practices	
Analytic Methodology	What statistical methods are employed by the authors in the analysis of their data?
Statistical Validity of Regression	What is the ratio of the number of ecological units to the number of covariates in the regression analysis?
Use of Covariates	Do the authors use covariates in their ecological analysis?
Sample Size Calculations	Do the authors clearly state that they used sample size calculations to determine the necessary number of ecological units for their analysis?
Spatial Analysis	Do the authors investigate or adjust for spatial effects?
Covariate Adjustment for Age and Sex	If the outcome is sex- and/or age-adjusted, are the proper regression covariates similarly adjusted?
Reporting	
Statement of Study Design	Do the authors clearly and relevantly mention one of the terms 'ecological', 'aggregate', or some reasonable approximation in their article?
Justification of Study Design	Do the authors adequately justify a group-level analysis?
Discussion of Cross-level Bias	Are the results appropriately cautioned with some mention of the ecological fallacy or cross-level bias?

2.3 Results

Our search yielded a total of 117 cross-sectional ecological papers that met the inclusion criteria. Table 2.2 describes their distribution by journal and year of publication. From Table 2.2, it can be seen that the number of published cross-sectional ecological studies has remained relatively stable over the past nine years, and that journals more oriented to general or clinical epidemiology tend to publish far fewer of them than journals specializing in community health.

An additional twenty five papers were considered tentatively, but ultimately did not meet the inclusion criteria for one reason or another. Of the 117 papers that were included in the bibliometric review, sixteen of them had been subjected to repeated scrutiny after initially being deemed tentative. The once-tentative papers that were included in our review have been marked for the sake of our analyses, which will take their potential effects into account by calculating results with and without them as a sensitivity analysis.

Broadly speaking, the quality of the reviewed studies could best be described as bi-modal. It was found that many papers performed elementary analyses coupled with unclear, insufficient reporting. At the same time, however, we discovered that a significant number of studies employed analyses appropriately suited to the data and presented their results in a model fashion. In between these extremes, many papers would be adequate with respect to one or two of the three fundamentals only to perform unsatisfactorily in another. Overall, we feel that the ecological literature has demonstrated a clear and immediate need for improvement. To justify and elucidate this claim before we discuss implications and recommendations, the results of the bibliometric review are presented below in the same groups that comprise the three fundamentals of the evaluation criterion. Namely, we present in turn the results for the aspects of study design, the statistical methods and practices, and the quality of reporting.

Table 2.2: Number of Ecological Papers by Journal and Three-year Period

Journal	2000-2002	2003-2005	2006-2008	Total
American Journal of Epidemiology	4	3	5	12
Annals of Epidemiology	2	4	6	12
Epidemiology	5	0	3	8
International Journal of Epidemiology	12	9	6	27
Journal of Clinical Epidemiology	1	1	2	4
Journal of Epidemiology and Community Health	14	19	21	54
Total	38(32%)	36(32%)	43(36%)	117(100%)

2.3.1 Aspects of Study Design

We observed that cross-sectional ecological studies have been applied to a diverse set of etiologic hypotheses, and therefore have study outcomes touching on many areas of epidemiology. In spite of this, there were notable trends in the literature: mortality rates were found to be the most common class of outcome in our bibliometric review, comprising thirty-four percent of all outcomes, and numbering forty-one in total. Incidence rates were the next most prevalent, representing thirty-two percent of all outcomes. The remaining classifications were by far less populous among the analyses, with seventeen prevalence outcomes, seven count-data outcomes, seven life expectancy outcomes, and two health services outcomes. An additional seven analyses, with measures such as social inequality or community access to certain resources, were not classifiable as one of the above.

Certain types of the aforementioned sub-classifications appeared most often. The most common etiologic research area was cancer, which was the focus of twenty-two analyses (17% of total). Noninfectious diseases, all-cause mortality, and cardiovascular disease followed behind, representing fourteen percent, thirteen percent, and nine percent of the total analyses respectively. Infectious diseases, including HIV and AIDS, surprisingly received the attention

of a mere eleven analyses (9%). Respiratory disease outcomes comprised eight percent of all analyses; suicide and self harm comprised another seven percent; and criminal activity such as homicide accounted for five percent. Twenty-five analyses did not fall into one of the above broad sub-classifications, which perhaps serves to demonstrate the variety of cross-sectional ecological research.

The ecological units themselves seemed to be less diverse in their origins than the outcomes were in their applications. It appears that the vast majority of cross-sectional ecological research has its data dictated either by convenience or necessity. While we cannot be definitive about such a claim, we have been led to suspect that it is the case after observing that only eighteen studies out of the one-hundred seventeen (15%) had their ecological units explicitly motivated *a priori* by the etiologic hypothesis at hand.

With regards to the levels of aggregation of the ecological units, we noted that the literature had a pronounced reliance on, or preference for, small-area analysis. Units aggregated approximately at the level of census tracts or subsections of a city ("neighbourhoods") were by far the most numerous, being the observational units for forty-five studies (38% of total). Large-area studies occupying the other end of the extreme were the second most prevalent; international comparisons of the disease rates and risk factors of entire nations motivated twenty-four studies (21% of total).

All other levels of aggregation present in the literature occupied a middle ground between these two in terms of their size. These were, from smallest to largest: municipalities such as towns or cities; municipal regions and counties, which include not only major urban centers but also the contiguous areas around them; and states or provinces, which in this bibliometric review refer to the largest divisional units of nations. Approximately, these accounted for nine percent, sixteen percent, and nine percent of all studies respectively. Only four studies could not have their ecological units classified as one of the above, either because the units were disparate areas specially constructed by the authors for their study, or because the units were non-spatial, e.g. physician practices.

Insofar as we were able to tell, it appeared that a slight majority of studies intended their inferences to be applied at the individual level. Of the one-hundred seventeen in the review, forty-two (37%) were clearly oriented towards the individual. Conversely, twenty-seven studies (23%) either explicitly made inferences at the group level or otherwise tailored their discussions to group-level effects and outcomes. However, these results should be taken lightly, since a lack of clarity on the part of forty-eight studies (40%) prevented us from determining with any confidence their intended levels of inference.

The literature revealed that cross-sectional ecological studies employ sample sizes ranging from very small to extremely large, which was anticipated. The minimum sample size was 2 ecological units, and the maximum observed sample size was approximately 200,000 units. The median sample size was 62.5 ecological units. Due to the tremendous standard deviation of the sample sizes and their skewed distribution, the average value is not a meaningful measure to report.

We expected such a broad range chiefly due to the fact that ecological studies can often have their potential samples limited by their level of inference, or by the level of aggregation selected for their study. For example, Moulton and Benini (2003) were interested in community-level risk factors for the number of landmine victims in Chad and Thailand. The number of potential ecological units in their study is far fewer than if it had been possible (or desirable) to obtain units at smaller levels of aggregation, such as neighbourhoods. Perhaps one result of this is a high prevalence of ecological analyses predicated on sample sizes that would be considered small by the standards of individual-level studies: sixty-four papers (55%) relied on analyses with fewer than one-hundred observational units. Although these particular studies may have had adequate power to investigate their hypotheses, smaller sample sizes in general may be a statistical limiting factor that the ecological study design is forced to contend with.

Table 2.3: Results from Bibliometric Review for Study Design Criteria

Evaluation Criterion	Results (n=117)
Sample Size	
Median	62.5
Maximum	200,000
Minimum	2
Level of Aggregation [1]	
Census tract or subsection of city	45 (38%)
Municipality	11 (9%)
Municipal regions / counties	19 (16%)
Province or state	15 (13%)
National or nation-cluster	24 (21%)
Other	4 (3%)
Level of Inference [1]	
Individual	43 (37%)
Ecological	27 (23%)
Unclear	48 (40%)
Pre-specified Ecological Units	
Yes	18 (15%)
No	99 (85%)
Health Classification of Primary Outcome(s) [1]	
Violent crime	6 (5%)
Suicide and self-harm	8 (7%)
Respiratory disease	9 (8%)
Infectious disease	10 (9%)
Cardiovascular disease	10 (9%)
All-cause mortality	15 (13%)
Other non-infectious disease	16 (14%)
Cancer	20 (17%)
Other	25 (21%)

[1] Does not add up to 117

2.3.2 Quality of Statistical Methods and Practices

The first criterion of this section investigated analytic methodology. We discovered a great deal of diversity regarding the analytic methods employed in the cross-sectional ecological studies, with techniques ranging from the simplest possible to highly complex. In spite of this, there were also marked trends and preferences in the literature. There were several general types of analyses, especially regression techniques, which were by far the most prevalent. In particular, we noted that the following numbers of analyses were performed by each of the following: Ordinary Least Squares (OLS) regression (n=45); Poisson regression (n=19); Weighted OLS regression (n=18); the nonparametric methods of Chi-square tests for trend, Chi-square tests for homogeneity, Mann-Whitney Tests, nonparametric Spearman's rank correlation, and Wilcoxon's rank sum test (n=12); simple Pearson or Spearman correlations (n=11); logistic regression (n=7); negative binomial regression (n=6); and Bayesian analysis (n=6). An additional seventeen analyses did not fall into any of these categories, employing disparate techniques such as sophisticated geospatial methods, Monte Carlo approaches, spatial autoregressive models, and the simple model-free calculation of relative risks and standardized mortality ratios. There was also a wide range of approaches within any given analytic technique. For example, the Poisson regressions ranged from simple and univariate (Mezei *et al.* 2006), to hierarchical (Rezaeian *et al.* 2006), to complex Generalized Linear Mixed Model forms (Kleinschmidt *et al.* 2001). Thus, most techniques could be found in studies at both ends of the quality spectrum.

Note that the counts given above do not refer to the number of studies per se, but rather to the number of analyses. Since many studies performed more than one analysis and may have used more than one method, the distinction is important to bear in mind throughout the presentation of the results.

In spite of the preponderance of warnings in the literature about the limitations of OLS regression as an etiologic tool for ecological studies, it was nevertheless the single most prevalent analytic technique. In fact, it alone accounted for almost forty percent of the analyses. When

one includes weighted OLS regression in this figure, the number swells to fifty-three percent, implicating over half of the analyses. Even worse from a statistical perspective, eleven analyses (9% of the total) were conducted with nothing more than Pearson or Spearman correlation coefficients, thus neglecting to investigate even provisionally the effect measure estimates provided by regression coefficients. While the validity of simple regression must be decided on a case-by-case basis, the fact that over half of the analyses relied on at least one of its variations suggests that ecological researchers too often rely on unsophisticated techniques of dubious reliability.

In spite of this, there were a certain number of studies employing sophisticated techniques that appeared to be carefully considered and well suited for the etiologic hypothesis. These studies tended to use Poisson regression, negative binomial regression, Bayesian methods, and geospatial techniques while avoiding OLS regression and simple descriptive statistics.

The second criterion evaluated how many studies performed *a priori* sample-size estimates. The proper use of sample size formulae has been a neglected consideration in many epidemiological areas, and we observed that cross-sectional ecological studies are no exception; precisely one study claimed to have investigated the effects of sample size, rigorously or otherwise (Shakhtarin *et al.* 2003). The other studies may have provided an in-depth discussion of their sampling frame and techniques, as well as the population that was sampled, but would otherwise keep the reader uninformed as to how the number was arrived at. However, the use of *a priori* sample size calculations may be moot when one is confined by the availability of data or when the research is being used to motivate further studies, as may often be the case for ecological analyses.

We found the situation to be much improved with respect to the use of covariates in statistical analysis. Eighty-six studies (74% of the total) either stratified their analysis or employed covariates in addition to the main etiologic variable of interest. It remains surprising, however, that only three-quarters of ecological studies use covariate adjustment for their results when one considers that this procedure is a foundational principle of epidemiologic analysis, as

well as the fact that, under certain conditions, ecological studies may be able to employ covariate adjustment with the same efficacy as individual-level studies. The necessity of covariates must be decided on a case-by-case basis, however, and may depend on issues such as availability of data, consistency and quality of data across groups, etc.

Similarly, the reviewed studies also performed better with regards to the statistical validity of their regression. Although we were unable to determine the ratio of ecological units to predictor variables for five studies due to a lack of clear information, we calculated the ratios for the remaining one-hundred twelve studies. For nine of them this criterion did not apply, since no regression analyses were performed. Twenty-four regression analyses, approximately twenty-one percent of those remaining, had fewer than ten units per covariate, bringing their validity into question. It is interesting to note, though perhaps not surprising, that a disproportionately high number of these (66.7%) relied on either OLS or weighted OLS regression. Forty-eight analyses had more than ten units per covariate but fewer than twenty units per covariate, and forty regression analyses had more than twenty units per covariate. Naturally, those with more units per covariate tended to have far greater sample sizes. Sample size and regression techniques notwithstanding, however, there were no significant differences between those with fewer than ten ecological units per covariate and those with more than ten.

Almost one fifth of the reviewed studies included some investigation of spatial effects. Techniques such as spatially autocorrelated regression models, Moran's I statistic, kriging, and spatial scan statistics appeared to be the most common. The majority of these methods originate from the geostatistical literature. Some of them merely detect and describe spatial autocorrelation and clustering, while others seek to incorporate spatial processes into the effect measure estimation. Although the latter methods are controversial, the presence of spatial analyses nonetheless bodes well for the analytic complexity and sophistication of these particular ecological studies.

In terms of covariate adjustment, it was found that a small majority of papers do not adjust their covariates for age or sex when the study outcomes have been adjusted for these potential

confounders, as is recommended by Rosenbaum and Rubin (1984). Although most papers did not adjust their outcomes with respect to age or sex, of those that did only eleven studies properly adjusted their suitable (non-integral) covariates, while twenty-three did not. Thus, it would appear that over fifty percent of papers adjusting their outcomes in this manner are at significant risk of publishing conclusions based on biased effect measures.

Table 2.4: Results from Bibliometric Review for Statistical Methodology Criteria

Evaluation Criterion	**Results (n=117)**
Analytic Methodology [1]	
Number of analyses employing:	
OLS regression	45 (36%)
Poisson regression	19 (15%)
Weighted OLS regression	18 (14%)
Nonparametric methods	12 (10%)
Pearson or Spearman correlations	11 (9%)
Logistic regression	7 (6%)
Negative binomial regression	6 (5%)
Bayesian analysis	6 (5%)
Statistical Validity of Regression [1]	
Fewer than 10 units per covariate	24 (19%)
10 to 20 units per covariate	48 (38%)
More than 20 units per covariate	40 (32%)
Unclear	5 (4%)
Not applicable	9 (7%)
Use of Covariates	
Yes	86 (74%)
No	31 (26%)
Sample Size Calculations	
Yes	1
No	116
Spatial Analysis	
Yes	22 (19%)
No	95 (81%)
Covariate Adjustment for Age and Sex [1]	
Did not properly adjust covariates	23 (68%)

[1] Does not add up to 117

2.3.3 Quality of Reporting

From the bibliometric review it was discovered that the majority of authors clearly state their chosen study design. Eighty-one papers (69%) mentioned at least one of 'ecologic', 'ecological', or 'aggregate' at some point in the main body of their work, and did so in a clear and relevant fashion. The importance of doing so has been emphasized previously. In light of this, then, it is felt that the proportion of studies making unequivocal declarations regarding the ecological nature of their research is too low.

Overall, this level of quality extended to the other reporting criteria as well: Only forty-eight studies (41%) adequately and explicitly justified an ecological analysis, while the other 59% did not provide any rationale whatsoever as to why the design was either necessary or desirable. Most significantly, however, we found that the majority of studies do not sufficiently warn the reader about the possibility that their results may not be applicable to levels of aggregation different than those of the analysis. That is, most studies do not discuss in any terms the possibility that their results may be susceptible to cross-level bias. Of the one-hundred seventeen papers reviewed, fifty-two (44%) tempered their results appropriately, whereas sixty (51%) did not.

For five studies we felt that this criterion did not apply, either because their ecological investigations were performed alongside multilevel analyses or because the predictors were entirely group-level in nature, therefore precluding individual-level interpretations.

All of the above results underwent two sensitivity analyses. The first sought to explore any effects that the inclusion of 'short reports' might have had on the results. Since such reports are, by definition, limited in their length, it was decided that they would be likely to omit from their discussions some of the necessary information scrutinized in this review, and might therefore artificially reduce the overall quality of the literature. Six short reports were thus removed from the analysis, and the results were recalculated. While it is true that the short

Table 2.5: Results from Bibliometric Review for Quality of Reporting Criteria

Evaluation Criterion	**Results (n=117)**
Clearly State Study Design	
Yes	81 (69%)
No	36 (31%)
Justification of Study Design	
Yes	48 (41%)
No	69 (59%)
Discussion of Cross-level Bias	
Yes	51 (44%)
No	60 (51%)
Not applicable	6 (5%)

reports were likelier to neglect certain aspects of reporting, their omission had no appreciable effect on the results of the bibliometric review.

The second sensitivity analysis explored the effects of the included studies that had once been deemed tentative. There were sixteen such papers. As with the short reports, their removal did not appreciably affect the results or conclusions.

2.4 Discussion

The bibliometric review has revealed many weaknesses that appear to be prevalent for cross-sectional ecological studies published in journals of epidemiology. The combination of methodological errors, small sample sizes, dependence on extant aggregate data, and incomplete reporting has caused an overall reduction of quality and clarity that may limit otherwise important research. Our results demonstrate that all of the three aforementioned research fundamentals need to be significantly improved in order for the situation to be remedied.

In terms of the statistical aspects, it is evident that researchers must better heed many of

the important analytic results published in the biostatistical and epidemiologic literature. In particular, more attention needs to be paid to *a priori* sample size calculations, proper covariate adjustment, and the statistical validity of regression. It is also apparent that Ordinary Least Squares regression and simple approaches such as Pearson correlations have become the default techniques for many ecological analyses, which may hinder the potential of the research. In general, those seeking to publish with the use of this study design need to become better informed about the intricacies of statistical analysis at the ecological level and to adapt their methods as recommended.

Specifically, the discovery that, according to the results of Vittinghoff and McCulloch (2007), approximately twenty-percent of cross-sectional ecological analyses are likely not being performed within the bounds of statistical validity means that ecological researchers should become more aware of the impacts of their sample sizes and plan their analyses accordingly. We recommend that ecological researchers consult Vittinghoff and McCulloch (2007) as well as Rosenbaum and Rubin (1984) for guidance concerning the appropriate number of regression covariates and advice regarding any necessary covariate adjustment.

However, the area of greatest concern was not the quality of the statistical analysis. Our results have shown a severe paucity of important information when it comes to reporting the origins of the data, the intended level of inference, the nature of the design, and the study limitations. The sensitive nature of the ecological design makes it absolutely essential that all readers are sufficiently informed and thus able to decide for themselves whether or not the analysis and inferences are both valid and warranted. Too often, the authors seemingly assumed that their choice of study design was self-evident, did not require justification, and provided unambiguous results.

For the selection of ecological groups, the ideal is that every study chooses its observational units according to the specific needs of its research. This would optimize the inferences in terms of construct validity. However, as stated by Morgenstern (1995), ecological studies can at times be made possible or viable by extant aggregated data. While such data is surely a boon to

ecological research if used strategically, it can nevertheless be a harmful crutch if it is used blindly or without consideration as to the underlying inferences intended by the authors. Since less than one-sixth of the studies contained in this review make it explicit that their ecological units were not selected simply because they were the default option, there is reason to wonder why or how the data were chosen. The high prevalence of studies conducted at the census-tract/neighbourhood and national levels may be a result of the fact that easily available census data tends to exist at these scales of aggregation. In either case, the uncertainty surrounding the matter is due to a lack of clarity on the part of the literature.

Similarly, the aforementioned ambiguity regarding the intended levels of inference in many of the reviewed papers might at times make it difficult for the scientific community to properly apply the results. We found that many authors would merely state the scope of their analysis, but would not clarify the particular causal association of interest. Since many ecological analyses can have their causal associations interpreted in several ways, i.e. on several levels, it becomes necessary for all authors to clarify their discussions appropriately.

Also, authors need to exercise more care when cautioning the results of their investigations. This aspect was sorely neglected for many of the reviewed articles. The responsibility of presenting candid, clear interpretations of one's own research is perhaps greater for authors of ecological analyses than for those employing designs less prone to bias. It is necessary to provide a frank discussion that either makes reference to the perils of the ecological fallacy or explains why it cannot apply to the results. A commentary regarding not only the uses but also the limitations of the research is an indispensable part of any discussion, and is doubly so for ecological analyses.

In accordance with recommendations made by von Elm *et al.* (2007), we submit that there is presently a strong need for a set of guidelines that standardize the reporting of ecological studies, as well as their design and analysis. Many of the flaws identified in the literature by our review could be attenuated or prevented altogether if such a document were properly constructed and disseminated. Providing a set of guidelines that not only lists detailed criteria

but also explains their importance and the consequences of their neglect may serve to emphasize their need to the research community.

The lack of counsel in the STROBE document for the reporting of ecological analysis has proven to be a gap in the literature that should be filled sooner rather than later. From the results of our review, it is clear that the cross-sectional ecological design continues to be an approach relied upon by epidemiologists, and we therefore suggest that it is a worthwhile endeavour to augment the STROBE document with guidelines that standardizes its reporting. As stated by von Elm *et al.* (2007), recommendations on the reporting of research can improve reporting quality. While the STROBE document has surely been motivated by this sentiment, we feel that it must continue to improve in its role by working towards the inclusion of study designs that are presently excluded. The current standard of reporting for the cross-sectional ecological design makes it an ideal candidate.

In particular, the results of our review have pointed out that the following items, listed below in Table 2.6, need to supplement the STROBE document. For ease of reference, we present Table 2.6 with the same headings, titles, and numberings that are present in the STROBE document itself.

Table 2.6: Recommended Additions to the STROBE Document for the Reporting of Ecological Studies by Section of Publication and STROBE Item No.

STROBE Section	**Recommendations**
1.TITLE and ABSTRACT	
	Indicate the study's design with a commonly used term in the title or abstract *Ecological Study - use one of the following terms in your description: 'ecological', 'ecologic', or 'aggregate'*
INTRODUCTION	
3.*Objectives*	State specific objectives, including any prespecified hypotheses. *Ecological Study - declare to which level(s) of inference your investigation is intended to apply*
METHODS	
4.*Study Design*	Present key elements of the study design early on in the paper. *Ecological study - provide a rationale or justification for the design, taking your study objectives into account.*
6.*Participants*	*Ecological Study - Give the eligibility criteria for the ecological units, as well as their sources and any sampling methods used. Provide a rationale for their selection. If the units were aggregated for the study, explain how this was done and the criteria by which they were constructed.*
DISCUSSION	
19.*Limitations*	Discuss limitations of the study, taking into account sources of potential bias or imprecision. *Ecological Study - provide a discussion that either informs the reader of the potential for cross-level bias or explains why it is not a possibility for your results and inferences.*

Chapter 3

Discussion of Select Models and Biostatistical Results

3.1 Introduction

The preceding chapters have served as an introduction to ecological analysis. As we move forward, there are statistical details that need to be brought to light. The purpose of this chapter is to provide two things: a mathematical overview of the analytic models that will be investigated in the simulation study, and a discussion of the key biostatistical results introduced in Chapter 1.

In this chapter, we primarily focus on ecological count data as the outcome of interest. As evidenced by the bibliometric review, count and rate responses in their various forms are the most common class of outcome for ecological analysis. Mortality rates, general incidence rates, prevalences, and raw counts comprised nearly 88% of the outcomes analysed by these studies. Thus, given the ability of generalized linear count-data models to handle rates as simple extensions of counts, we lose little by focusing our scope.

The statistical models themselves have been chosen for inclusion by the results of the bibliometric review. We selected the models that appeared most frequently and thus appear to be most relevant to cross-sectional ecological research. The analytic models included in this chapter are as follows: Ordinary Least Squares (OLS) regression, weighted LS regression, Poisson regression, and negative binomial regression.

OLS regression and the mathematical properties of its ecological regression coefficient will be presented in Section 3.2.1. Weighted OLS regression shall be discussed within an ecological context in Section 3.2.2. Poisson and negative binomial regression will be presented in Sections 3.3.1 and 3.3.2 respectively. Section 3.4.1 shall follow with a discussion of the ecological fallacy for linear regression models. The phenomena of pure specification bias, confounding by group, and effect modification by group will be discussed here. Mathematical descriptions of the covariate-adjustment insights provided by Rosenbaum and Rubin (1984) will also be presented in Section 3.4.1. Similarly, commentaries on the sources of the ecological fallacy and the results for covariate adjustment pertinent to generalized linear models will be given in Section 3.4.2.

Notation

Throughout this chapter, we will rely on a common set of notation to present the mathematics. For ease of reference, we summarize that notation here.

Borrowing from the work of Duncan *et al* (1961), assume that each ecological group has n_j individuals for a total of N individuals across groups, and that (X_{ij} , Y_{ij}) is an observation pair on the i^{th} individual in the j^{th} group, $j = 1...k$. Let X_{ij} denote the exposure to an effect of interest or a covariate, and let Y_{ij} denote the study outcome. Then define the following:

i) Expected number of events within group j: $E[Y_j]$, where $Y_j = \sum Y_{ij}$

ii) The within-group average exposure for group j: $\bar{X}_j = \sum_j X_{ij}/n_j$

iii) Grand mean over groups: $\bar{X} = \sum_j \sum_i X_{ij}/N$

iv) Total sum of squares of X: $S_{xxt} = \sum_j \sum_i (X_{ij} - \bar{X})^2$

v) Within-group sum of squares of X: $S_{xxw} = \sum_j \sum_i (X_{ij} - \bar{X}_j)^2$

vi) Between-group sum of squares of X: $S_{xxb} = \sum_j n_j (\bar{X}_j - \bar{X})^2$

vii) Total sum of products: $S_{xyt} = \sum_j \sum_i (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})$

viii) Within-group sum of products: $S_{xyw} = \sum_j \sum_i (X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j)$

ix) Between-group sum of products: $S_{xyb} = \sum_j n_j (\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y})$

An identical set of relationships exist for the outcome variable Y. Note that we use shall essentially use n_j interchangeably with t_j, which represents the person-time for group j. Population size is often taken as a proxy measure for person-time when length of exposure is considered to be equal between groups.

Furthermore, we have that:

The correlation ratios of the two variables are:

x) $C^2_{YA} = S_{yyb}/S_{yyt} = 1 - S_{yyw}/S_{yyt}$

xi) $C^2_{XA} = S_{xxb}/S_{xxt} = 1 - S_{xxw}/S_{xxt}$

Lastly, we have:

xii) Total regression coefficient: $\hat{\beta}_t = S_{xyt}/S_{xxt}$

xiii) Average within-group regression coefficient: $\hat{\beta}_w = S_{xyw}/S_{xxw}$

xiv) Between-group or ecological regression coefficient: $\hat{\beta}_b = S_{xyb}/S_{xxb}$

3.2 Linear Regression Models

3.2.1 OLS Regression

For the modeling of ecological rates with a single predictor, the OLS regression equation can be expressed as

$$\frac{Y_j}{n_j} = \beta_0 + \beta_1 X_j + \epsilon_j \qquad \epsilon_j \sim N(0, \sigma^2) \tag{3.1}$$

where the error terms ϵ_j are independent and identically distributed.

This simple statistical method was the single most prevalent form of analysis for the cross-sectional ecological studies contained in the bibliometric review. Unlike simple Pearson correlations, OLS regression provides parameter estimates that are not sensitive to the variability of the exposure within groups and allows for simultaneous covariate adjustment. Here, the outcome of interest is a continuous variable assumed to be approximately normally distributed but the predictors are not, and therefore this approach can be used for a wide range of applications. In cross-sectional ecological analysis, it is most commonly applied to measures of incidence and mortality rates across groups. For example, Zhang *et al.* (2000) fit models of international lung cancer mortality rates with dietary habits after adjusting for the prevalence of smoking and other potential confounders.

The relative ease of regressing ecological rates or group-prevalences on a set of exposures and confounders has likely been a driving force behind the popularity of ecological OLS regression. While it can certainly be an important tool in ecological analysis, it has potential limitations as well. As mentioned in Chapter 1, linear regression at the group level produces regression coefficients - and therefore effect measures - that are not simple functions of the group-level data. Analogous to the results for ecological correlation coefficients, ecological regression coefficients are functions of the average within-group regression coefficient and the overall (individual-level)

regression coefficient. This relationship is given below for the simple case of one predictor.

$$\hat{\beta}_t = \hat{\beta}_w + C_{XA}^2(\hat{\beta}_b - \hat{\beta}_w)$$

This is proven in Appendix B. As with correlation coefficients, this implies that the ecological regression coefficient cannot be expressed independently of non-ecological data.

3.2.2 Weighted Linear Regression

Weighted regression is an extension of ordinary linear regression. It is used when certain assumptions of OLS regression are violated. Weighted regression places different emphasis on each observation to correct for issues of heteroskedasticity, or, as is more common in ecological analysis, to obtain standard errors that account for differential population sizes or possible clustering effects. The model can be written as

$$\frac{Y_j}{n_j} = \beta_0 + \beta_1 X_j + \epsilon_j \qquad \epsilon_j \sim N(0, \sigma^2/n_j) \tag{3.2}$$

Note that the variance of the error terms for each observation are weighted by the corresponding group population. If the ecological groups do in fact differ in size, an unweighted regression may be inappropriate (Pocock *et al.*, 1981). The concern is that larger groups provide more reliable measures of prevalence, incidence rates, effect measures, etc, and should therefore receive more weight in the analysis.

One simple method for weighting the ecological observations is to assign as weights the population sizes or some function thereof. This was a common approach for the cross-sectional ecological papers in our bibliometric review. An alternative method is to use weights inversely

proportional to binomial sampling variance, but this can be too extreme. Pocock *et al.* (1981) claim that this approach tends to overweight larger areas, and they instead propose a weighting scheme based on a variance model that simultaneously incorporates sampling variation, unexplained variation, and variation accounted for by the predictors. The estimation of the variance components - and therefore the weights - relies on the combination of maximum likelihood and an iteratively reweighted least-squares approach. Interestingly, this is roughly the same estimation technique used by generalized linear models, which shall be discussed in section 3.3. The weighting method of Pocock *et al* (1981) is applicable to several forms of group-level mortality rates, such as standardized mortality ratios, but is likely to apply to group-level rates in general.

Insofar as we are aware, there are no large-sample inferences investigating the efficacy of these weighted regression approaches under varying ecological circumstances.

3.3 Generalized Linear Regression Models

The linear regression models of section 3.2 are all predicated on the assumption that the outcomes of interest are continuous and approximately Normal. Many forms of ecological data, however, cannot meet these assumptions. For example, the number of incident cancer cases in a given group and time period cannot be said to follow a continuous distribution of any sort; the data is confined to the set of nonnegative integers. Linear regression may then produce meaningless effect measures and incorrect standard errors (Gardner *et al*, 1995). In general, discrete outcomes belong to classes known as count data and binary data. Logistic regression is ideal for modeling binary data, whereas Poisson regression and negative binomial regression handle the modeling of counts. While they are different models with different assumptions and capabilities, they are nevertheless heavily interrelated. To best demonstrate their relationships with one another, we begin with Poisson regression.

3.3.1 Poisson Regression

Poisson regression models count data by assuming that the number of events from each ecological group follows a Poisson distribution. In ecological analysis, the number of occurrences of the outcome are counted by group. That is, the frequencies of the study outcome are known by group only. The expected number of events for a given group can be ecologically modeled as a function of the regional exposures and covariates. Poisson regression is not restricted to pure counts, however. If we know the amount of exposed person-time in a given group, we may transform the count into a rate and model that instead. If n_j represents the total amount of person time in group j, then entering the natural log of n_j into the Poisson equation provides an offset that accounts for any differential distribution of exposure time between groups. The inclusion of an offset term results in the following model:

$$E[Y_j] = exp(\beta_0 + log(n_j) + \beta_1 X_j) \tag{3.3}$$

Population incidence rates and mortality rates can then be ecologically modeled, therefore providing estimates of incidence rate ratios. This greatly increases the efficacy of Poisson regression. For example, Rezaeian *et al* (2006) adopted this approach in order to model suicide rates as function of regional deprivation. Recall that a similar offset for population size could be included instead of person-time. If exposure time is considered to be equal between groups, then entering the natural log of the group sizes would have the same effect as including the natural log of the person-time. These two approaches are relatively common. In either case, Poisson regression is ultimately nonlinear in form; a contrast to the models of section 3.2.

Regardless of whether Poisson regression is used at the individual or ecological level, it is implicitly assumed that the expected number of events is equal to the variance of the counts. This is called equidispersion, and is a direct result of the underlying Poisson formulation. This assumption is often violated by what is known as overdispersion, which occurs when the variance of the counts is significantly greater than the mean. Overdispersion is caused by a

positive within-group correlation or the absence of important explanatory covariates from the regression model(Hilbe, 2007). The effect of overdispersion is to provide false standard errors, and to therefore reduce the reliability of significance testing for parameter estimates. When this occurs, one is forced to adopt a model that does not make similar assumptions about the variance. There are several possible corrective models, all of which are extensions of Poisson regression itself.

One solution would be to adopt a method such as Generalized Estimating Equations (GEE) Poisson regression. GEE Poisson provides a regression equation and therefore effect measure estimates that are similar in form to those of ordinary Poisson regression. However, GEE Poisson allows for clustering within the data structure by estimating a nonparametric correlation matrix, and is therefore more capable of handling correlated count data. As a result, standard error estimates are likely to improve. A similar solution would be to adopt the variance estimation technique proposed by Moore and Tsiatis (1991), which directly builds from the GEE methods developed by Liang and Zeger (1986). Their method is also a less parametric version of ordinary Poisson regression since it does not specify the Poisson variance as being exactly equal to the mean. Instead, it defines the variance to be a general function of the mean with a random component that must be estimated from the data. This permits the variance component to accommodate overdispersion without introducing any further parametric assumptions. A more parametric but potentially more powerful approach is given by negative binomial regression, which we discuss in section 3.4.2.

It is worth noting that logistic and Poisson regression produce similar effect-measure estimates when outcomes are rare. This link between Poisson and logistic regression lies in the connection between counts, rates, and probabilities, as well as the fundamental connection between the Poisson and Binomial probability distributions. Under the Law of Rare Events, it can be shown that as the number of binomial trials approaches infinity while the number of successes remains constant the Poisson distribution with mean np comes to approximate the Binomial distribution (Kalbfleisch, 1985, page 125). In other words, rare events imply that the

number of binomial successes np, where n is the number of trials and p is the probability of success, is approximately equal to the Poisson mean μ. Since the underlying probability models converge in this special case, so do the regression models. Thus, we expect Poisson and logistic regression to be less and less distinguishable as outcomes become rarer. Rare outcomes are often the concern of ecological analysis, as demonstrated by the prevalence of such topics as regional cancer mortality or violent crime in the bibliometric review.

3.3.2 Negative Binomial Regression

Like Poisson regression, negative binomial regression is used to model count data. However, it is free from the equidispersion assumption and is therefore more capable of modeling correlated outcomes or data with excessive variance due to the inadequate explanatory power of the covariates or unmeasured intra-cluster correlation. Because of this, it is most frequently used to model Poisson data that is overdispersed, and is a direct extension of the Poisson model itself (Hilbe, 2007). As a result, the negative binomial regression model is the same as the Poisson model given by relation 3.3. However, the number of events Y_j now follows a negative binomial distribution.

Though the negative binomial distribution has multiple parameterizations, it is easiest to understand its connection to the Poisson model when it is derived as a gamma-Poisson mixed distribution. It can be shown that if a given observation follows the Poisson distribution with variance μ_j, and if the gamma distribution has variance μ_j^2/ϕ, then the negative binomial distribution has variance $\mu_j + \mu_j^2/\phi$. This is clearly a direct combination of the two probability distributions. The variance can also be rewritten as $\mu_j + \kappa\mu_j^2$, which allows intuitive inferences regarding the dispersion parameter κ and further illustrates that the negative binomial model is an extension of the Poisson. The value of κ is empirically estimated from the data and can be subjected to hypothesis testing. Note that in the absence of overdispersion $\kappa = 0$ and the negative binomial reduces to the Poisson.

There are many extensions of the negative binomial model, but the simplest and most general version has a logarithmic link between $E(\mu_j)$ and the covariates and predictors, and so is a nonlinear regression function as well. Note that the expected value of Poisson regression is the same as that of the negative binomial regression. Only the variances and underlying parameterizations differ. Note also that, like Poisson regression, negative binomial regression can accommodate incidence-rate outcomes with the inclusion of population or person-time offsets in the model.

Negative binomial regression is more restrictive than approaches such as GEE Poisson since it specifies the form of the variance. While this assumption may be untenable, it may increase efficiency of parameter estimation when the variance is approximately Gamma-Poisson in form (Gardner *et al.*, 1995).

3.4 Ecological Fallacy

3.4.1 Ecological Fallacy for Linear Models

Pure Specification Bias

The ecological regression, which is mainly limited to group means and prevalences, often does not follow the same functional form as the true individual-level risk model (Greenland, 1992). Pure specification bias, also known as model specification bias, arises because of this phenomenon. Specifically, pure specification bias arises because nonlinear risk models change their functional form when aggregated over individuals (Wakefield, 2008).

As stated by Greenland (1992), most of the ecological literature employs linear models for their analyses. This observation is in agreement with the results of our bibliometric review. He posits that the reliance on this type of risk model may be due to misapplications of the Aggregation Theorem, which is as follows:

Suppose we have K ecological groups. Let $\mathbf{X}$ denote the individual-level exposure and covariate vector. If the risk r follows a linear form at the individual level given by relationship (3.4)

$$r(\mathbf{x}) = \beta_0 + \beta\mathbf{X} \tag{3.4}$$

where $\beta\ \mathbf{X} = X_1\beta_1 + X_2\beta_2 + ... + X_m\beta_m$ is a vector of exposures and covariates.
Then, the ecological regression of the corresponding group averages will also be linear and have the same coefficients:

$$\bar{r}_j = \beta_0 + \beta\bar{\mathbf{X}}_j \tag{3.5}$$

Unfortunately, the Aggregation Theorem does not necessarily hold; the assumptions of additivity and linearity at the individual level may simply not be true. In this case the estimation of nonlinear risks with linear ecological regression models may result in a poor fit for obvious reasons, especially when estimating rate ratios (Greenland, 1992). However, nonlinear ecological regressions employing simple group means for the estimation of nonlinear individual-level risks may also produce substantially biased effect measures. This is because the Aggregation Theorem

does not extend to models that are, for example, exponential in form. Ignoring this fact often results in pure specification bias.

To illustrate, assume that event rates follow an exponential risk model at the individual level:

$$r(\mathbf{x}) = exp(\beta_0 + \beta\mathbf{X}) \tag{3.6}$$

This might suggest what has been referred to as the 'naive' analogous ecological model:

$$\bar{r}_j = exp(\beta_0^* + \beta^*\bar{\mathbf{X}}_j) \tag{3.7}$$

which regresses the average rate in group j on the group averages of the exposure and covariates. However, this model does not follow from the individual-level model of (3.6) except under special circumstances. In fact, fitting model (3.7) and interpreting β^* will most often result in biased effect-measure estimates (Morgenstern, 1982). Only when there is complete homogeneity of exposure in each group will model (3.7) disallow pure specification bias, since then the average value is equal to the individual value in every case (Wakefield, 2008).

Instead, the proper model for ecological estimation when the underlying risk follows the form of model (3.6) is given by

$$L_j = \beta_0 + \beta\bar{\mathbf{X}}_j \tag{3.8}$$

where L_j is the log-rate in ecological group j. This model will produce coefficients equal to those in model (3.6), thus guaranteeing accurate effect measures insofar as other possible biases have been accounted for. Unfortunately, L_j is an unlikely measure to exist in census data, and it rarely if ever is equal to the log of the group-average rates. Because of this, the ecological model most capable of estimating exponential risks is the one least likely to be used, and so pure specification bias may be a challenge for many ecological studies.

Confounding and Effect Modification by Group

Confounding by group and effect modification by group are the two unique sources of ecological bias that may occur when performing simple linear regression of ecological data. The phenomenon of confounding by group is well known, but in spite of this it has often been misunderstood; many researchers seem to think that it can occur only when individual-level confounders are differentially distributed between ecological groups (Greenland and Morgenstern, 1989). This is merely one source of confounding by group. In general, confounding by group arises when the background rates of disease, i.e. the rates in the unexposed populations, varies across groups. Such variation results from the differential distribution of extraneous risk factors, but these factors do not need to be confounders at the individual level within groups. More succinctly, it can be said that ecological bias results from confounding by group when there is a nonzero correlation between the background exposure rates and the average exposure levels within groups (Morgenstern, 1995).

Confounding by effect modification, on the other hand, occurs when the effect of the exposure is different across groups. This can occur when individual effect-modifiers are unequally distributed between groups, or when the average exposure has an effect that exerts itself independently of the individual-level exposure. The ecological regression, being confined to group prevalences and means, is unable to account for the within-group effect modification and as a result groups appear to be effect modifiers. Even in the absence of confounding by group, small amounts of effect modification by group can still result in largely biased ecological estimates (Greenland and Morgenstern, 1989).

It can be shown that linear ecological regression coefficients modeling rates can be decomposed into three terms: the average rate difference across groups, a bias term due to confounding by group, and a bias term due to effect modification by group (Greenland and Morgenstern, 1989). For the case of binary exposures, with r_{0j} being defined as the rate in the unexposed (background rate) and r_{1j} being the rate in the exposed in group j,

the crude rate in group j is given by

$$R_{.j} = p_j r_{1j} + (1 - p_j) r_{0j} = r_{0j} + p_j D_j = r_{0j}(p_j R_j + [1 - p_j])$$

where p_j is the proportion of the population exposed in group j, $D_j = r_{1j}$ - r_{0j} = the rate difference, and $R_j = r_{1j}/r_{0j}$ = the rate ratio. In the ecological linear regression of the crude rates on the exposure prevalences, the model is expressed as

$$R_{.j} = a + bp_j$$

From above, 1 + (b/a) is the ecological rate-ratio estimator and b is the ecological rate difference. Furthermore, from linear regression theory we may state that

$$b = \frac{cov(p_j, R_{.j})}{var(p_j)} = \frac{cov(p_j, r_{0j}) + cov(p_j, p_j D_j)}{var(p_j)}$$

where the covariances and variances are between groups, not within. Let $E(D_j)$ and $E(p_j)$ be the weighted averages of D_j and p_j over the groups. Then if we substitute

$$cov(p_j, p_j D_j) = E(D_j) + \frac{cov([p_j - E(p_j)]p_j, D_j)}{var(p_j)}$$

into the above, simplification and cancellation of terms reveals that

$$b = E(D_j) + \frac{cov(p_j, r_{0j})}{var(p_j)} + \frac{cov([p_j - E(p_j)]p_j, D_j)}{var(p_j)} \qquad (3.9)$$

The second term on the right-hand side of (3.9) can be seen as a term for confounding by group, since this source of bias will come to pass only if r_{0j} varies across groups, and therefore if $\text{cov}(p_j, r_{0j}) > 0$. The third term on the right-hand side can be seen as representing a bias component due to effect modification by group. This term is nonzero when D_j varies across groups, and therefore the covariance of $([p_j - E(p_j)]p_j, D_j)$ is positive. Thus, the ecological rate difference, when estimated via simple linear regression, is equal to the average individual-level rate difference $E(D_j)$ and two ecological bias terms.

Covariate Adjustment for Linear Models

A substantial number of ecological studies in the bibliometric review regressed age- or sex-adjusted outcomes on unadjusted predictors. The authors likely did this in an attempt to correct for the different distributions of age or sex between the ecological groups, and to therefore guard their estimates against confounding. However, Rosenbaum and Rubin (1984) have shown this approach to be erroneous and prone to bias. They base their argument on the supposition that the individual-level outcomes are modeled according to the following linear form:

$$E(Y_{aji}) = \alpha + \sum_m \beta_m a^m + \eta X_{aji} + \gamma^{\mathrm{T}} \mathbf{Z_j} + \delta^{\mathrm{T}} \mathbf{W_{aji}} \tag{3.10}$$

with Y_{aji} being the response of the i^{th} person with age a in group j. $\mathbf{Z_j}$ is a vector of characteristics of group j. $\mathbf{W_{aji}}$ is a vector of individual-level characteristics, and X_{aji} is a binary exposure variable. Note that this model includes m polynomial terms in age, though this may be replaced by other linear functions of age such as indicator variables.

One common method of attempting to adjust for age or sex in group-level analysis is to regress adjusted ecological outcomes on crude covariate averages. Another common approach is to regress age- or sex-specific group rates on crude covariate averages (e.g., separately regress the outcome rates of sexes or five-year age groups). Continuing with the notation of (3.10),

these two models are, respectively:

$$E(\tilde{Y}_{.j.}) = \alpha + \eta \bar{X}_{.j.} + \gamma^{\mathbf{T}} \mathbf{Z_j} + \delta^{\mathbf{T}} \bar{\mathbf{W}}_{\mathbf{.j.}} \tag{3.11}$$

$$E(\bar{Y}_{aj.}) = \alpha + \eta \bar{X}_{.j.} + \gamma^{T} \mathbf{Z_j} + \delta^{T} \bar{\mathbf{W}}_{\mathbf{.j.}} \tag{3.12}$$

The above models have been shown to be biased when attempting to estimate the individual-level parameters of model (3.10). Models (3.11) and (3.12) produce biased estimates because they neglect to completely control for the confounding effects of age. That is, they account for the association between age and the outcome of interest by direct adjustment, but they do not remove the association between age and the exposure, which is then inserted into the model. Our bibliometric review found model (3.11) to be a relatively common method of adjusting for age and sex.

According to Rosenbaum and Rubin (1984), there are three main (though oft-ignored) approaches for properly dealing with age distributions that differ between groups. The first of these, described by equation (3.13) below, is a weighted regression of the age-specific outcome rates on age and the corresponding age-specific predictors, averaged for each group.

$$E(\bar{Y}_{aj.}) = \alpha + \sum_{m} \beta_m a^m + \eta \bar{X}_{aj.} + \gamma^{\mathbf{T}} \mathbf{Z_j} + \delta^{\mathbf{T}} \bar{\mathbf{W}}_{\mathbf{aj.}} \tag{3.13}$$

Unbiased estimates for the individual-level parameters of model (3.10) may be obtained by the use of suitable weights for the group-average, age-specific outcome rates. For example, if the Y_{aji} are conditionally homoskedastic and uncorrelated, the appropriate weight for each observation is its group's population size. In general, however, model (3.13) will produce unbiased estimates for the parameters of model (3.10) with any set of positive weights (Rosenbaum and Rubin, 1984).

The second model regresses crude outcome rates on crude covariate averages and the moments of age for group j. That is, the outcome, exposure, and covariates are unadjusted for age. See model (3.14) below.

$$E(\bar{Y}_{.j.}) = \alpha + \sum_m \beta_m M_{jk} + \eta \bar{X}_{.j.} + \gamma^{\mathbf{T}} \mathbf{Z_j} + \delta^{\mathbf{T}} \bar{\mathbf{W}}_{\mathbf{j.}} \tag{3.14}$$

Note that M_{jk} is the k^{th} moment of age in group j, which may be estimated from frequency tabulations if they exist. This may be possible with census data. If the necessary moments of age are available for each group, then the parameters of model (3.10) can unbiasedly be estimated via weighted regression. As with model (3.13), the assumption of conditional homoskedasticity and independence entails that the appropriate weight for each observation is its corresponding group size.

The final model considered valid by Rosenbaum and Rubin (1984) regresses age-adjusted outcomes on age-adjusted predictors. This is an intuitive approach to solving the confounding problem and does not rely on rare data such as moments of age, as can be seen from relation (3.15). Here, both the outcome and the covariates are directly adjusted.

$$\begin{aligned} E(\tilde{Y}_{.j.}) &= \alpha + \sum_m \beta_m \sum_a f_a a^k + \eta \sum_a f_a \bar{X}_{aj.} + \gamma^{\mathbf{T}} \mathbf{Z_j} + \delta^{\mathbf{T}} \sum_{\mathbf{a}} f_a \bar{\mathbf{W}}_{\mathbf{aj.}} \\ &= \alpha' + \eta \tilde{X}_{.j.} + \gamma^{\mathbf{T}} \mathbf{Z_j} + \delta^{\mathbf{T}} \tilde{\mathbf{W}}_{\mathbf{j.}} \end{aligned} \tag{3.15}$$

Under the conditional homoskedastic and independence assumptions mentioned above, this model will produce unbiased estimates of the individual-level parameters if groups are weighted by the proper coefficient.

A Geostatistical Perspective

Most ecological studies rely on data that have been aggregated over spatial regions. Statistical geographers and spatial statisticians have shown that cross-level bias is related to the Modifiable Areal Unit Problem (MAUP) and the Change of Support Problem (COSP), two concepts rarely discussed in epidemiology.

The MAUP refers to the phenomenon by which the same data may lead to entirely different inferences depending on how they are aggregated. It may arise for two reasons. The first, referred to as an 'aggregation effect', is the tendency of larger scales of aggregation to be more and more dissimilar from their smaller constituent units. The second source of the MAUP is known as the 'zoning effect'. It occurs when effect measure modifiers or confounders are differentially distributed over regions, and so alternative formations of the areal units at the same or similar scales result in different conclusions (Crawford and Young, 2004). In other words, the inferences may be sensitive to how the region of analysis is partitioned into subunits. These two issues are likelier to come to pass when adjacent spatial units are less similar than distal units, i.e. when there is negative spatial autocorrelation. This is because aggregating dissimilar units obscures the underlying information.

The Change of Support Problem, on the other hand, refers to the inherent difficulty of observing data processes of one form and inferring to another. For example, we may observe meteorological point data at weather stations and from this construct smooth temperature surfaces over a region. By so doing, we have gone from one form of support (point) to a completely different one (surface). In ecological analysis, data exist at the aggregate level and are used to infer to the support of the individual. The one form of support is related to the other, but has different statistical and spatial properties. Moreover, they may be entirely different causal constructs. The MAUP and the ecological inference problem can be seen as special cases of the COSP (Crawford and Young, 2004).

3.4.2 Ecological Fallacy for Generalized Linear Models

Many of the mathematical results presented in section 3.4.1 have not been extended to nonlinear models or to generalized linear models. The wealth of formal insight for simple linear models appears to be lacking when it comes to more advanced analytic techniques. However, it is known that, as with ecological linear regression models, ecological generalized linear regression models are subject to the same sources of bias as individual-level analyses. They too are additionally susceptible to the sources of bias that are unique to ecological analysis.

It is worth noting that, in spite of this paucity, the results for pure specification bias mentioned in section 3.4.1 are particularly relevant for generalized linear models. This is due to the underlying connection between exponential functions and the logarithmic link functions used in the modeling of rates and probabilities. That is, since the most common form of generalized linear model is an exponential function, problems regarding the lack of correspondence between individual-level and group-level exponential regression coefficients are a threat. Most notably, Richardson *et al* (1987) have shown that exponential regression models develop a multiplicative bias term when aggregated over groups. More formally: if the mean level or prevalence of disease in group j is denoted by Y_j, then a convex risk model $f(x)$ has the following property:

$$Y_j = E[f(x)] \geq f(\bar{X}_j)$$

This is a result of Jensen's Inequality. For the specific case where the risk function is exponential and therefore of the form

$$f(x) = ce^{\gamma x}$$

It follows that

$$Y_j = E[ce^{\gamma x}] = ce^{(\gamma \bar{X}_j) b(\gamma, x)}$$

The bias coefficient $b(\gamma, x)$ will depend on the exponential regression coefficient, the mean $\bar{X}_j$, and the parameters of $f(x)$. The bias term will cause the relative risk to be biased unless it fortuitously cancels out from the numerator and denominator. This is rarely the case, and therefore the bias is likely to be nonzero for simple aggregate exponential models.

Covariate Adjustment for Nonlinear Models

Note that the results for age-adjustment discussed in section 3.3.1 were limited to linear models with continuous outcomes. If instead the model is of a generalized linear form with a binary outcome, Rosenbaum and Rubin caution that the logit model does not lead to the types of straightforward conclusions that they discuss. They go on to add that, in spite of this, the logit model is not immune to the problems that result from improper adjustment of covariates and outcomes. Following this line of reasoning, we should expect this cautionary statement to apply to the other generalized linear models discussed in this chapter. While the specific results may vary between models, it is likely that regressing adjusted rates on unadjusted predictors will potentially lead to bias irrespective of whether linear or generalized linear models are used.

3.5 Discussion

This chapter has presented some of the major mathematical complexities that are an inextricable challenge for any ecological analysis. We have formally elucidated the main points of the preceding chapters in order to justify the criterion used in the bibliometric review and to further motivate the simulation study. In particular, this chapter has illustrated some of the potential performance issues that may arise from: weighted ecological regression and the choice of weights; linear modeling versus generalized linear modeling for rate outcomes; the accuracy of effect-measure estimates and standard errors produced by the two interrelated generalized linear models when outcomes are rare; and the robustness of the various models to suboptimal ecological sample sizes.

Chapter 4

Simulation Study of Ecological Regression Models

4.1 Introduction

The purpose of this chapter is to present results of a simulation study comparing validity and power of several common statistical tests of association using ecological data, in which common ecological regression models were fit to group-level count data.

The objectives and selected tests of association are presented in Section 4.2. A discussion and justification for the parameter combinations investigated in the simulation are given in Section 4.3.1. Evaluation criteria for the tests are defined in Section 4.3.2. Simulation procedures are presented in Section 4.3.3. Results follow in Section 4.4, and are discussed in Section 4.5.

4.2 Study Objectives

The primary objective of the simulation study is to evaluate the validity and power of tests for association in an ecological context. Attention is limited to a continuous covariate with count-data outcomes from a Poisson or negative binomial distribution. Data for the study are generated to be typical of incidence rates seen in the ecological studies of the bibliometric review. The continuous exposure variable was generated to simulate the proportion of people exposed within each ecological group - as would be the case, for example, if obtained as the average of binary individual-level outcomes.

The test statistics we evaluate come from four common regression models, each of which shall be fit to the simulated data:

1. Ordinary Least Squares (OLS) regression
2. Weighted Least Squares regression, with weights proportional to population size
3. Poisson regression
4. Negative binomial regression

In this simulation, all test statistics shall test the null hypothesis of no association, H_0: $\beta_1 = 0$, versus the two-sided alternative H_a: $\beta_1 \neq 0$.

Secondary objectives include estimating the bias and precision of estimated regression coefficients. These analyses serve primarily to inform interpretation of the results concerning validity and power of the investigated test statistics.

4.2.1 Statistical Tests Investigated

The two linear models of interest mentioned previously both rely on the t-test for testing the significance of the regression coefficient. The two generalized linear models of interest, on the other hand, both use the Wald statistic and the Likelihood Ratio statistic. We briefly introduce these tests here.

Student's t-test follows the t-distribution and is of the general form

$$t = \frac{\hat{\beta}_1 - \beta_{H_0}}{\hat{\sigma}_1}$$

where $\hat{\sigma}_1$ is the estimated standard error of β_1 and β_{H_0} is the value under the null hypothesis. The test is similar for OLS and weighted LS models, except the weighted model incorporates the weights into the standard error and the parameter estimate. Thus, the two models will differ when groups have different values for the weighting variable.

The Wald test is similar in form to the t-test, and can be written as:

$$\chi^2_W = \frac{[\hat{\beta}_1 - \beta_{H_0}]^2}{\widehat{Var}(\hat{\beta}_1)}$$

Depending on how it is formulated, the Wald test is either asymptotically Normal or asymptotically chi-squared, with 1 degree of freedom. Both formulations are equivalent.

The Likelihood Ratio test is given by twice the difference between the maximized log likelihood under the alternative model and the maximized log likelihood under the null:

$$\chi^2_{LR} = 2[l(\hat{\beta}_a) - l(\hat{\beta}_n)]$$

where $\hat{\beta}_*$ is the maximum likelihood estimate of the regression parameter under the respective models. The Likelihood Ratio statistic asymptotically follows a chi-square distribution with 1 degree of freedom. Although the Poisson and negative binomial models employ the same tests, their performances may nonetheless differ since estimated standard errors and likelihood functions will not be identical.

4.3 Methods

4.3.1 Investigated Scenarios

We wish to tailor our simulation so that the scenarios of investigation reflect realistic circumstances often encountered by ecological researchers. To this end, we select the parameters and their specific values based on the characteristics of the literature noted in the bibliometric review. We list below the five parameters whose values shall determine the identity of each scenario:

- **Group Size:** number of individuals within an ecological group.
- **Ecological Sample Size:** the number of ecological units in a given scenario. It is equal to the number of counts that form the dependent variable in each regression analysis.
- **Overdispersion:** the degree to which the variance of the counts exceeds that expected under a Poisson assumption.
- **Exposure:** mean exposure prevalence of all groups, as well as the variability of exposure prevalences across groups.
- **Event Rate:** whether outcomes are rare or common, relative to group size.

Our simulation is particularly interested in the behaviour of the chosen regression models when faced with small numbers of ecological groups, overdispersion, and rare outcomes. A total of 240 scenarios were examined to evaluate test validity while a total of 48 scenarios were evaluated to test power. The remainder of Section 4.3.1 includes detailed descriptions of and justification for the selected parameter combinations.

A summary of the parameter values of interest is given in Table 4.1.

Group Sizes

Motivated by the literature, we choose three group sizes to reflect census tracts, counties, and municipalities. These levels of aggregation are commonly used and are therefore relevant.

According to Statistics Canada, census tracts have anywhere between 2,500 to 8,000 people. To represent this scale of aggregation, we take as our average group-size parameter value 2,500. To represent counties and municipalities, we will use the average group sizes of 20,000 and 250,000 respectively.

To model low variability between cluster sizes, we take the extreme case of equal group sizes and generate one of the above values for all ecological units. To model high variability, we include random variation by generating group sizes from a Uniform(a, b) distribution with a coefficient of variation of 0.20, where the coefficient of variation is defined as the ratio of the standard deviation to the mean. The Uniform mean is set to one of the above average values. This entails that variability increases with mean group size when variation is present, which is true in practice. See Table 4.1 for the mean, minimum, and maximum group sizes used in the simulation.

Ecological Sample Sizes

As noted in Chapter 2, studies using data at higher levels of aggregation often include fewer ecological units. For example, Treggiari and Weiss (2004) used data from nine geographic regions in the United States as part of their study comparing the incidence of mesothelioma to that of non-Hodgkins lymphoma. Sample sizes such as these represent an extremity whose effects are presently unknown. We therefore include the sample sizes of 10, 20, 30, 40, and 50 as the values for exploring small-sample circumstances. It is felt that beyond these values the models will begin to exhibit asymptotic properties, which we are not attempting to investigate.

Overdispersion

One objective of this simulation is to determine the effects of excess variance between ecological units. Study outcomes will be generated from both Poisson and over-dispersed Poisson data. Over-dispersed Poisson outcomes will be generated using the negative binomial distribution. The overdispersion parameter for the negative binomial distribution is denoted κ. When $\kappa = 0$ the negative binomial distribution reduces to the Poisson distribution. We set κ to 0, 0.05, 0.2 and 0.5.

In Chapter 5 we report results of an ecological analysis of international rates of homicide using data provided by Killias (1993). Killias (1993) was interested in the association between household gun ownership and homicide rates using data from 13 countries. Using a negative binomial regression model, we found that this association was highly statistically significant and that $\hat{\kappa} = 0.14$ (95% confidence interval (0.03, 0.25)). These results were used to inform our selection of values for κ in the simulation study.

Exposure

The nature of the exposure may entail that large amounts of the population are exposed, as is the case for common medical practices like prostate screening (Shaw *et al.*, 2004), or, conversely, that exposure is rare. Attention is limited to a binary individual-level exposure, which is modeled as a proportion at the ecological level. Ecological-level exposure is generated using a beta random variable. A beta random variable with parameters $\alpha = 2, \gamma = 18$ was used to generate data where exposure is likely to be uncommon (e.g. 10 percent exposed overall) and with little variability, while a Beta random variable with parameters $\alpha = 0.09, \gamma = 0.09$ generated data where exposure is common (e.g. 50 percent exposed) but highly variable. The latter will simulate scenarios in which most clusters are either entirely exposed or unexposed. The presence or absence of a group-level law would be an extreme example of this scenario. Plots of the two selected beta distributions are provided in Figure 4.1.

Event Rates

We limit our attention to rare outcomes. This is because linear regression models are at a known disadvantage when events are common, since a linear approximation of a nonlinear model fares best when the exponent is small in value. This can be seen from the Taylor expansion of an exponential regression function under that assumption:

$$e^{\beta X} \approx 1 + \beta X$$

Moreover, it was found that few studies transformed their outcomes to achieve linearity, and therefore an investigation of linear models attempting to circumvent the above problem with this technique would not seem relevant. Additionally, most of the literature tended to have primary outcomes for rare events such as cancer, infectious disease, self-harm, and violent crime. The majority of studies would therefore be best served by a simulation that gives its attention to exploring the effects of these circumstances. As such, we shall investigate event rates that occur for either 1% or 5% of the population on average.

Strength of Association

In our simulation, the regression coefficient denoted by β_1 is the parameter of interest. This is because β_1 measures the strength of the ecological association between exposure and incidence rate. To assess validity we set $\beta_1 = 0$, and to assess power we set $\beta_1 = 2.0$. An exponential regression coefficient of 2.0 is of realistic magnitude and prevents power from being too low or high for the parameter values investigated here.

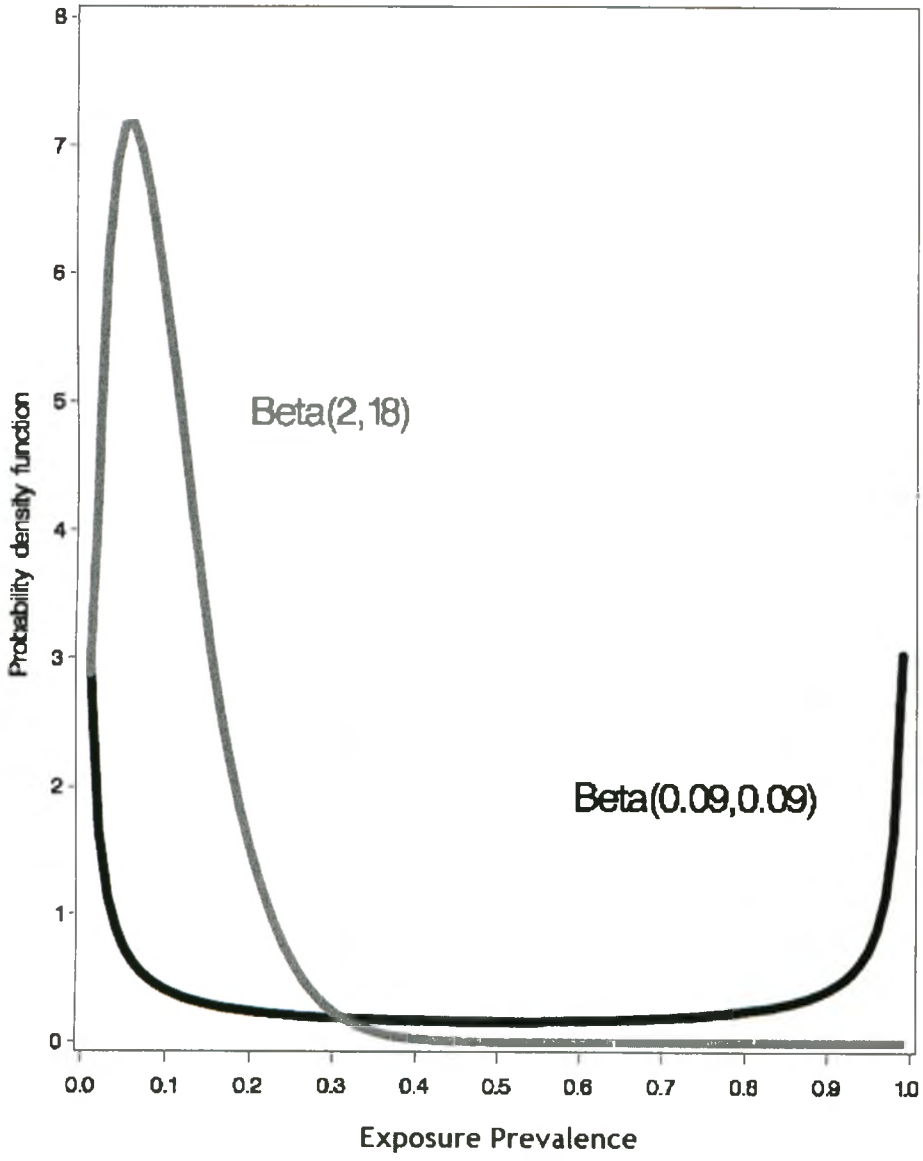

Figure 4.1: Plot of Beta Distributions Used to Randomly Generate Exposure

4.3.2 Evaluation Criteria

We have selected bias, power, and validity to be our evaluation criteria. Validity and power are calculated using empirical rejection rates for two-tailed tests assuming a 5% type I error rate using the six test statistics defined in Section 4.3.2. An exploration of bias may help understand any differences between the four models, and we therefore include it in the simulation as well. We define each criterion below:

1. **Power:** the ability of a given statistical test to correctly reject the null hypothesis of no association when a nonzero association truly exists. The inability to discover a statistically significant association is referred to as Type II error. Power is expressed as a value between 0 and 1, which in our case denotes the proportion of simulated runs in which the model did not commit a Type II error.

2. **Validity:** the converse of the above. A Type I error is committed when the null hypothesis of no association is incorrectly rejected and a nonexistent causal link is then believed to be real. In our simulation, 'validity' shall refer to the proportion of runs in which the model did not commit a Type I error.

3. **Bias:** a measure of any systematic difference that may exist between the average parameter estimate and the true value of that parameter. There are many measures of bias, but we have chosen to use Relative Bias as our means of assessment. Relative Bias indicates the difference between the average point estimate of the parameter and its true value, standardized to the scale of the true parameter. It is calculated as follows:

$$Relative\ Bias = \frac{\bar{\hat{\beta}} - \beta}{\beta}$$

Table 4.1: The Parameter Values Used in the Simulation Study

Parameter	**Values Investigated**		
Ecological Sample Size	10, 20, 30, 40, 50		
Group Size	*mean*	*(min, max)*	*coefficient of variation*
	2,500	(2500, 2500)	0.0
	2,500	(1634, 3366)	0.2
	20,000	(20000, 20000)	0.0
	20,000	(13072, 26928)	0.2
	250,000	(250000, 250000)	0.0
	250,000	(163397, 336602)	0.2
Exposure Rate	Low levels of exposure with little variability generated from a Beta(2,18) distribution, and highly variable exposures from a Beta(0.09, 0.09) distribution		
Event Rate	1% of cluster population size and 5% of cluster population size. This corresponds to the intercept values of β_0 = -3 and β_0 = -4.6 respectively		
Overdispersion	None ($\kappa = 0.0$), small ($\kappa = 0.05$), moderate (κ =0.20), and large ($\kappa = 0.50$)		
Strength of Association	No association ($\beta_1 = 0$), or moderate causal association ($\beta_1 = 2$)		

4.3.3 Simulation Procedures

This simulation has been designed to generate overdispersed count data according to the following model:

$$E[Y_i] = n_i e^{\beta_0 + \beta_1 X_i}$$

where Y_i refers to the number of events for ecological group i; n_i denotes the population size for group i; X_i is the proportion of the population experiencing the binary exposure; β_0 is a known constant; and β_1 is a known regression coefficient.

To achieve this, we use version 9.1 of the SAS Macro language. Each covariate in the above relation is created using the random-distribution generators provided by SAS. Exposure prevalence X_i is generated to follow a beta(α, γ) distribution, which is bound between [0,1]. Group size is generated according to a Uniform distribution, either with or without variation between groups. The known values β_0 and β_1 determine the event rates and are pre-specified. The resulting expected number of events $E[Y_i]$ is then passed as the mean to either a negative binomial or Poisson random generator. The overdispersion parameter κ is determined by the scenario being investigated. Thus, for each scenario, group-level count data is generated as a nonlinear function of random covariates, whose means and variances we control.

The counts are then passed to the four ecological regression models as the dependent variable. We set the linear regression model given by relation 3.1 to fit the count data solely as a function of exposure prevalence X_i. The weighted regression model given by relation 3.2 does the same, except that it weights each observation by group size. Both the Poisson and negative binomial regression employ models with exposure prevalence as the only predictor and group size as an offset variable, and so are of the form presented by relation 3.3. From each regression, we store the following information: the parameter estimate $\hat{\beta}_1$, whether a Type I error was committed, and whether a Type II error was committed. The stored data are then used to calculate large-sample estimates of validity, power, and relative bias.

For each scenario we randomly generated 1000 data sets. This number of data sets was

chosen to assure that a deviation of 0.014 ($= 1.96\sqrt{0.05 \times (1 - 0.05)/1000}$) from a nominal level of 0.05 is found to be statistically significant (see, e.g. Bradley (1978), Klar and Darlington (2004)). Consequently, Type I error rates exceeding 7% are printed in boldface as being overly liberal.

The values of the seeds, which initialize and determine the pseudo-random numbers used in the simulation, have been selected to maximize the likelihood of independence between the random number generators, runs, and scenarios. Independence is important for making valid inferences from simulated data (Burton *et al.*, 2006). We use the default seed in SAS, which is a function of the time and date. Although this does not allow the simulation results to be exactly replicated by other researchers, the alternative option of selecting fixed seeds is not recommended when using SAS Macro language.

For the two generalized linear regression models, we created additional variables that informed us if a given model did not achieve convergence. This was intended to act as a diagnostic for our simulation and to prevent us from unknowingly generalizing our results to parameter combinations that did not produce reliable estimates.

A factorial design was used when investigating validity. Therefore, all possible parameter combinations were simulated, with the following exceptions: the regression coefficient β_1 was always set to zero, and the exposure level for each group was arbitrarily generated from a beta(2,18) distribution. The exposure level is not important for validity, since the value of β_1 cancels its effects regardless.

Fewer scenarios were investigated for power. Attention was limited to sample sizes of 50, since the negative binomial model has unstable Type I error rates with fewer ecological units. Unlike validity, however, both exposure distributions were simulated. Only moderate event rates of 5% were generated. This is because the highly-variable exposures generated here increase the variability of event rates across groups, making the distinction between 1% and 5% less meaningful. All other parameter combinations were investigated, with $\beta_1 = 2$.

See Appendix C for the complete simulation code.

4.4 Results

The results of the simulation are tabulated and presented below. Tables 4.2 through 4.13 give Type I error rates for the linear models as well as the Wald and Likelihood Ratio test statistics of the generalized linear models. Convergence rates that are less than one-hundred percent for the simulations investigating Type I error are shown in Table 4.14. Results for power follow in Tables 4.15 and 4.16.

A notable discovery from these results is that the Wald statistic of the negative binomial model has large Type I error rates when sample sizes are small. Since its distributional properties are asymptotic, it would be expected that its validity improves with sample size. Although this was indeed the case, it was found that the Wald statistic did not consistently achieve nominal Type I error rates when the number of ecological units was less than 30 or 40. Even when the data originated from a negative binomial process, the Wald statistic exhibit inadequate small-sample behaviour.

The Likelihood Ratio test of the negative binomial model had better Type I error rates than its Wald counterpart. This result is in agreement with statistical theory, and also with previous empirical investigations. In spite of its superior performance, however, we found that it was unreliable for sample sizes smaller than 20 or 30.

Another important finding is the high validity of the two linear models. The results show that linear and weighted linear regression models are remarkably robust to the discrete nature of count data, overdispersion, cluster size, sample size, and event rates. The linear models maintained nominal Type I error rates for all of the parameter combinations we investigated. This may be a result of the fact that the negative binomial distribution begins to appear approximately Gaussian when the expected number of events and the number of trials are large. Thus, the underlying Normality assumption of linear regression was perhaps preserved, even though the data were log-linear and discrete. Figures 4.2 and 4.3 graph two of the negative binomial distributions used in this simulation. Figure 4.2 shows the distribution of counts when the overdispersion parameter $\kappa = 0.05$, the probability of success $p = 0.15$, and the expected

number of events is approximately 125. Note the strong resemblance to a Gaussian curve. Figure 4.3 shows the distribution of counts with the same number of expected events when $\kappa = 0.50$ and p = 0.016. Although Figure 4.3 is more skewed, it nonetheless bears a resemblance to a Gaussian curve as well.

It is important to note a potential limitation of the linear models. While they may have ideal Type I error rates and can therefore be used for hypothesis testing, their estimates of the strength of the association may be biased. This is because the true regression coefficient is in the exponent, and so is linear only on the natural-log scale. Our simulation found that when we set $\beta = 2$, both of the linear models consistently had relative biases between -0.93 and -0.98. Another limitation of linear models is that their intercepts and regression coefficients may lead to negative and therefore meaningless estimates of the rate ratio. However, these issues did not arise in our simulation under the null hypothesis, i.e, when the true value of the regression coefficient was zero. Regardless of the parameter combinations of the scenario, the relative bias of the linear models was often zero to the third or fourth decimal place, with the maximum relative bias being zero at two decimals.

The Poisson model proved to have overly liberal Type I error rates when the data were overdispersed. This is because overdispersion violates the inflexible variance structure of the Poisson model and causes the standard errors of the parameter estimates to be too small, and to therefore reject the null hypothesis too often. This is why its Type I error rates increased with the amount of overdispersion. Both the Wald and Likelihood Ratio test statistics proved to be highly susceptible. Less obvious was the finding that Type I error rates for the Poisson model were more sensitive to overdispersion when group sizes were larger.

It is interesting to note that, when exposure prevalence is low, the power of the four models are mainly dependent on the amount of overdispersion, with increasing overdispersion leading to lower power. This is likely due to the causal effect of exposure being obscured by the "noise" of the excess random variation. Although the impacts of group size and group variability were less notable, they produced an effect nonetheless. Specifically, for small amounts of overdispersion,

larger amounts of group variability reduce the power of the linear models. Also, we found that, for smaller levels of overdispersion, group size had a positive association with power.

The Wald and Likelihood Ratio statistics of the negative binomial model were far more comparable with respect to power. Although the Likelihood Ratio statistic seemed to have higher power overall, its superiority was marginal and did not manifest for small amounts of overdispersion. Effectively, the two methods demonstrated similar power. In fact, all of the regression models had very similar power. Negative binomial regression performed slightly better than OLS and Weighted LS regression only when there was variation between cluster sizes and overdispersion was small. However, at larger amounts of overdispersion the linear models tended to perform slightly better.

When exposure prevalence was highly variable, power was absolute for all models (results not shown). This is likely due to the fact that regression models have optimal power when the variability in the predictor is highest. This suggests that the models investigated here need not have power levels which are fatally susceptible to overdispersion. Note that the Poisson model has been excluded from the discussion of power since its large Type I error rates give it an unrealistic advantage.

Non-convergence was only an issue for the negative binomial model and only when the null hypothesis was true. Even then, it was relatively uncommon (see Table 4.14). Non-convergence appeared to be caused by negative estimates of the overdispersion parameter κ, which may have prevented the Hessian matrix from converging. Thus, non-convergence was likeliest when the data came from an equidispersed Poisson distribution. Small sample sizes were also a contributing factor. The lowest convergence rate was 93.1%. For every other model and parameter combination, convergence rates were 100%. When necessary, extra simulations were performed so that each scenario had 1000 runs.

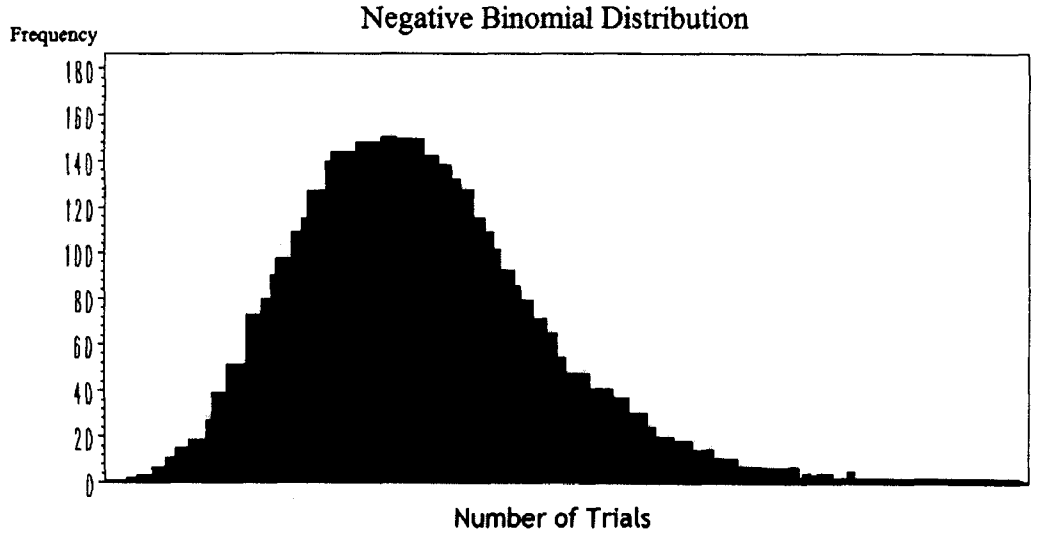

Figure 4.2: Negative Binomial Distribution when Overdispersion is Low

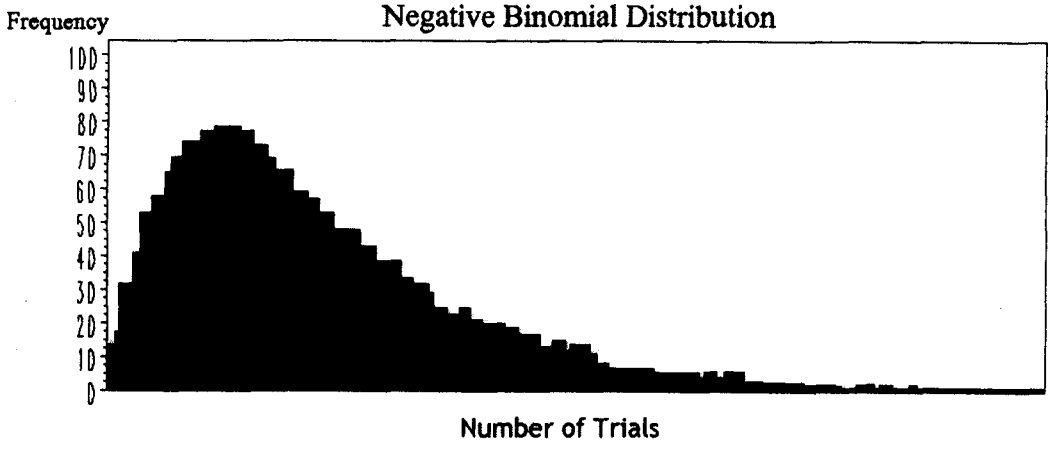

Figure 4.3: Negative Binomial Distribution when Overdispersion is High

Table 4.2: Type I error rates for testing $H_0 : \beta_1 = 0$ (nominal level=0.05)[1], with event rate of 5% and group size of 2500 (coefficient of variation = 0). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.054	0.050	0.060	0.040	0.046
	Weighted Linear	0.054	0.050	0.060	0.040	0.046
	Poisson Wald	0.050	0.051	0.055	0.042	0.047
	Poisson LR	0.049	0.053	0.055	0.042	0.047
	Negative Binomial Wald	**0.114**	**0.079**	**0.075**	0.058	0.061
	Negative Binomial LR	**0.085**	0.063	0.067	0.049	0.057
0.05	Simple Linear	0.044	0.037	0.050	0.052	0.063
	Weighted Linear	0.044	0.037	0.050	0.052	0.063
	Poisson Wald	**0.448**	**0.457**	**0.466**	**0.481**	**0.458**
	Poisson LR	**0.448**	**0.457**	**0.469**	**0.482**	**0.458**
	Negative Binomial Wald	**0.112**	0.066	0.064	0.060	0.068
	Negative Binomial LR	**0.080**	0.050	0.059	0.057	0.066
0.20	Simple Linear	0.051	0.049	0.050	0.058	0.054
	Weighted Linear	0.051	0.049	0.050	0.058	0.054
	Poisson Wald	**0.689**	**0.724**	**0.687**	**0.695**	**0.705**
	Poisson LR	**0.689**	**0.724**	**0.687**	**0.695**	**0.706**
	Negative Binomial Wald	**0.111**	**0.074**	**0.072**	**0.072**	0.066
	Negative Binomial LR	**0.082**	0.066	0.061	0.061	0.060
0.50	Simple Linear	0.048	0.054	0.041	0.053	0.051
	Weighted Linear	0.048	0.054	0.041	0.053	0.051
	Poisson Wald	**0.804**	**0.835**	**0.818**	**0.786**	**0.790**
	Poisson LR	**0.803**	**0.837**	**0.818**	**0.786**	**0.790**
	Negative Binomial Wald	**0.117**	**0.089**	**0.074**	0.068	0.054
	Negative Binomial LR	**0.083**	**0.073**	0.062	0.059	0.052

[1] Overly liberal Type I error rates are in bold

Table 4.3: Type I error rates for testing $H_0 : \beta_1 = 0$ (nominal level=0.05)[1], with event rate of 5% and group size of 2500 (coefficient of variation = 0.2). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.037	0.057	0.051	0.050	0.044
	Weighted Linear	0.044	0.054	0.057	0.050	0.051
	Poisson Wald	0.044	0.057	0.052	0.048	0.055
	Poisson LR	0.045	0.056	0.052	0.047	0.056
	Negative Binomial Wald	**0.136**	**0.097**	0.068	0.068	0.065
	Negative Binomial LR	**0.079**	**0.078**	0.060	0.062	0.062
0.05	Simple Linear	0.055	0.067	0.054	0.050	0.044
	Weighted Linear	0.058	0.068	0.064	0.056	0.045
	Poisson Wald	**0.464**	**0.463**	**0.469**	**0.479**	**0.455**
	Poisson LR	**0.465**	**0.465**	**0.471**	**0.480**	**0.457**
	Negative Binomial Wald	**0.134**	**0.080**	0.064	0.056	0.065
	Negative Binomial LR	**0.111**	0.064	0.054	0.055	0.058
0.20	Simple Linear	0.054	0.065	0.044	0.032	0.054
	Weighted Linear	0.044	0.070	0.054	0.041	0.063
	Poisson Wald	**0.695**	**0.727**	**0.670**	**0.677**	**0.702**
	Poisson LR	**0.695**	**0.727**	**0.670**	**0.679**	**0.702**
	Negative Binomial Wald	**0.098**	**0.086**	0.064	0.055	0.069
	Negative Binomial LR	**0.074**	0.069	0.056	0.049	0.061
0.50	Simple Linear	0.057	0.046	0.050	0.046	0.043
	Weighted Linear	0.060	0.053	0.061	0.060	0.055
	Poisson Wald	**0.786**	**0.794**	**0.788**	**0.790**	**0.780**
	Poisson LR	**0.786**	**0.793**	**0.798**	**0.791**	**0.781**
	Negative Binomial Wald	**0.113**	0.067	**0.077**	0.046	0.057
	Negative Binomial LR	**0.088**	0.053	0.066	0.043	0.049

[1] Overly liberal Type I error rates are in bold

Table 4.4: Type I error rates for testing $H_0 : \beta_1 = 0$ (nominal level=0.05)[1], with event rate of 5% and group size of 20,000 (coefficient of variation = 0). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.051	0.052	0.050	0.051	0.051
	Weighted Linear	0.051	0.052	0.050	0.051	0.051
	Poisson Wald	0.047	0.059	0.047	0.048	0.051
	Poisson LR	0.046	0.059	0.047	0.048	0.051
	Negative Binomial Wald	**0.114**	**0.084**	0.055	0.062	0.062
	Negative Binomial LR	**0.087**	0.068	0.052	0.055	0.056
0.05	Simple Linear	0.044	0.058	0.048	0.045	0.043
	Weighted Linear	0.044	0.058	0.048	0.045	0.043
	Poisson Wald	**0.775**	**0.767**	**0.781**	**0.752**	**0.765**
	Poisson LR	**0.775**	**0.767**	**0.781**	**0.753**	**0.764**
	Negative Binomial Wald	**0.114**	**0.081**	**0.072**	0.059	0.052
	Negative Binomial LR	**0.086**	0.069	0.065	0.055	0.051
0.20	Simple Linear	0.044	0.068	0.043	0.048	0.065
	Weighted Linear	0.044	0.066	0.043	0.048	0.065
	Poisson Wald	**0.877**	**0.891**	**0.898**	**0.902**	**0.902**
	Poisson LR	**0.878**	**0.891**	**0.898**	**0.902**	**0.902**
	Negative Binomial Wald	**0.120**	**0.099**	0.055	0.055	0.070
	Negative Binomial LR	**0.096**	**0.087**	0.047	0.049	0.066
0.50	Simple Linear	0.040	0.059	0.044	0.053	0.049
	Weighted Linear	0.040	0.059	0.044	0.053	0.049
	Poisson Wald	**0.944**	**0.924**	**0.919**	**0.924**	**0.921**
	Poisson LR	**0.944**	**0.924**	**0.919**	**0.924**	**0.921**
	Negative Binomial Wald	**0.117**	**0.084**	0.065	0.059	0.061
	Negative Binomial LR	**0.088**	0.069	0.055	0.055	0.056

[1] Overly liberal Type I error rates are in bold

Table 4.5: Type I error rates for testing $H_0 : \beta_1 = 0$ (nominal level=0.05)[1], with event rate of 5% and group size of 20,000 (coefficient of variation = 0.2). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.049	0.052	0.061	0.051	0.046
	Weighted Linear	0.051	0.052	0.065	0.056	0.048
	Poisson Wald	0.052	0.050	0.053	0.041	0.053
	Poisson LR	0.052	0.050	0.054	0.041	0.053
	Negative Binomial Wald	**0.111**	**0.076**	0.065	0.063	0.069
	Negative Binomial LR	**0.082**	0.062	0.057	0.053	0.065
0.05	Simple Linear	0.046	0.053	0.047	0.047	0.057
	Weighted Linear	0.059	0.058	0.053	0.060	0.055
	Poisson Wald	**0.773**	**0.781**	**0.775**	**0.788**	**0.788**
	Poisson LR	**0.773**	**0.781**	**0.774**	**0.789**	**0.787**
	Negative Binomial Wald	**0.126**	**0.092**	0.064	0.067	0.064
	Negative Binomial LR	**0.098**	0.081	0.061	0.062	0.059
0.20	Simple Linear	0.059	0.059	0.049	0.046	0.048
	Weighted Linear	0.057	0.063	0.053	0.055	0.047
	Poisson Wald	**0.892**	**0.902**	**0.892**	**0.889**	**0.889**
	Poisson LR	**0.892**	**0.902**	**0.892**	**0.889**	**0.889**
	Negative Binomial Wald	**0.125**	**0.086**	0.070	0.060	0.054
	Negative Binomial LR	**0.094**	0.062	0.067	0.054	0.051
0.50	Simple Linear	0.049	0.052	0.048	0.048	0.047
	Weighted Linear	0.055	0.055	0.057	0.056	0.058
	Poisson Wald	**0.928**	**0.922**	**0.915**	**0.920**	**0.925**
	Poisson LR	**0.929**	**0.922**	**0.915**	**0.920**	**0.925**
	Negative Binomial Wald	**0.111**	**0.076**	0.064	0.054	0.054
	Negative Binomial LR	**0.076**	0.053	0.058	0.049	0.044

[1] Overly liberal Type I error rates are in bold

Table 4.6: Type I error rates for testing H_0: $\beta_1 = 0$ (nominal level=0.05)[1], with event rate of 5% and group size of 250,000 (coefficient of variation = 0). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.046	0.057	0.044	0.057	0.052
	Weighted Linear	0.046	0.057	0.044	0.057	0.052
	Poisson Wald	0.041	0.053	0.046	0.055	0.046
	Poisson LR	0.041	0.053	0.046	0.055	0.045
	Negative Binomial Wald	**0.101**	**0.092**	0.064	**0.081**	0.059
	Negative Binomial LR	**0.075**	**0.075**	0.056	0.070	0.060
0.05	Simple Linear	0.051	0.040	0.058	0.064	0.051
	Weighted Linear	0.051	0.040	0.058	0.064	0.051
	Poisson Wald	**0.946**	**0.933**	**0.929**	**0.937**	**0.929**
	Poisson LR	**0.946**	**0.933**	**0.929**	**0.937**	**0.929**
	Negative Binomial Wald	**0.120**	0.066	**0.077**	**0.074**	0.061
	Negative Binomial LR	**0.086**	0.053	0.069	0.066	0.060
0.20	Simple Linear	0.052	0.044	0.053	0.043	0.062
	Weighted Linear	0.052	0.044	0.053	0.043	0.062
	Poisson Wald	**0.984**	**0.977**	**0.961**	**0.949**	**0.963**
	Poisson LR	**0.984**	**0.977**	**0.961**	**0.949**	**0.963**
	Negative Binomial Wald	**0.128**	**0.080**	0.066	0.049	0.070
	Negative Binomial LR	**0.099**	0.061	0.056	0.048	0.066
0.50	Simple Linear	0.058	0.061	0.049	0.046	0.060
	Weighted Linear	0.058	0.061	0.049	0.046	0.060
	Poisson Wald	**0.980**	**0.980**	**0.985**	**0.984**	**0.983**
	Poisson LR	**0.980**	**0.980**	**0.985**	**0.984**	**0.983**
	Negative Binomial Wald	**0.111**	**0.086**	0.068	0.064	0.070
	Negative Binomial LR	**0.088**	**0.080**	0.061	0.058	0.069

[1] Overly liberal Type I error rates are in bold

Table 4.7: Type I error rates for testing H_0: $\beta_1 = 0$ (nominal level=0.05)[1], with event rate of 5% and group size of 250,000 (coefficient of variation = 0.2). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.041	0.053	0.043	0.052	0.049
	Weighted Linear	0.038	0.055	0.051	0.046	0.042
	Poisson Wald	0.052	0.050	0.060	0.059	0.043
	Poisson LR	0.052	0.050	0.061	0.059	0.043
	Negative Binomial Wald	**0.086**	**0.064**	**0.077**	0.063	0.056
	Negative Binomial LR	**0.142**	**0.077**	**0.076**	**0.089**	0.068
0.05	Simple Linear	0.055	0.049	0.046	0.039	0.049
	Weighted Linear	0.055	0.052	0.055	0.044	0.054
	Poisson Wald	**0.945**	**0.926**	**0.926**	**0.920**	**0.935**
	Poisson LR	**0.945**	**0.926**	**0.926**	**0.920**	**0.935**
	Negative Binomial Wald	**0.112**	**0.072**	**0.071**	0.060	0.065
	Negative Binomial LR	**0.082**	0.059	0.060	0.050	0.057
0.20	Simple Linear	0.048	0.050	0.052	0.043	0.042
	Weighted Linear	0.050	0.061	0.065	0.053	0.056
	Poisson Wald	**0.966**	**0.967**	**0.963**	**0.969**	**0.966**
	Poisson LR	**0.966**	**0.967**	**0.963**	**0.969**	**0.966**
	Negative Binomial Wald	**0.106**	**0.079**	**0.074**	0.051	**0.072**
	Negative Binomial LR	**0.074**	0.063	0.065	0.047	0.067
0.50	Simple Linear	0.045	0.045	0.058	0.058	0.048
	Weighted Linear	0.053	0.055	0.056	0.063	0.057
	Poisson Wald	**0.983**	**0.973**	**0.978**	**0.976**	**0.979**
	Poisson LR	**0.983**	**0.973**	**0.978**	**0.976**	**0.979**
	Negative Binomial Wald	**0.113**	0.067	0.067	**0.073**	0.056
	Negative Binomial LR	**0.082**	0.058	0.054	0.064	0.053

[1] Overly liberal Type I error rates are in bold

Table 4.8: Type I error rates for testing H_0: $\beta_1 = 0$ (nominal level=0.05)[1], with event rate of 1% and group size of 2500 (coefficient of variation = 0). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.043	0.048	0.045	0.053	0.041
	Weighted Linear	0.043	0.048	0.045	0.053	0.041
	Poisson Wald	0.047	0.049	0.044	0.055	0.040
	Poisson LR	0.049	0.050	0.042	0.053	0.040
	Negative Binomial Wald	**0.106**	**0.085**	0.062	0.068	0.047
	Negative Binomial LR	**0.082**	0.068	0.053	0.064	0.044
0.05	Simple Linear	0.046	0.046	0.041	0.046	0.046
	Weighted Linear	0.046	0.046	0.041	0.046	0.046
	Poisson Wald	**0.186**	**0.207**	**0.206**	**0.171**	**0.189**
	Poisson LR	**0.185**	**0.210**	**0.207**	**0.168**	**0.188**
	Negative Binomial Wald	**0.098**	**0.080**	0.064	0.057	0.055
	Negative Binomial LR	**0.077**	0.068	0.059	0.052	0.051
0.20	Simple Linear	0.056	0.054	0.052	0.053	0.036
	Weighted Linear	0.056	0.054	0.052	0.053	0.036
	Poisson Wald	**0.450**	**0.418**	**0.441**	**0.459**	**0.400**
	Poisson LR	**0.454**	**0.420**	**0.443**	**0.459**	**0.401**
	Negative Binomial Wald	**0.103**	**0.087**	0.062	0.068	0.037
	Negative Binomial LR	**0.088**	**0.073**	0.054	0.062	0.038
0.50	Simple Linear	0.034	0.063	0.049	0.045	0.061
	Weighted Linear	0.034	0.063	0.049	0.045	0.061
	Poisson Wald	**0.566**	**0.584**	**0.569**	**0.585**	**0.610**
	Poisson LR	**0.566**	**0.585**	**0.569**	**0.588**	**0.609**
	Negative Binomial Wald	**0.090**	**0.107**	0.063	0.057	**0.073**
	Negative Binomial LR	0.060	**0.091**	0.056	0.052	0.070

[1] Overly liberal Type I error rates are in bold

Table 4.9: Type I error rates for testing $H_0 : \beta_1 = 0$ (nominal level=0.05)[1], with event rate of 1% and group size of 2500 (coefficient of variation = 0.2). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.065	0.044	0.043	0.049	0.044
	Weighted Linear	0.062	0.044	0.047	0.057	0.055
	Poisson Wald	0.049	0.046	0.047	0.056	0.054
	Poisson LR	0.049	0.046	0.045	0.058	0.053
	Negative Binomial Wald	**0.107**	0.069	0.060	**0.072**	0.063
	Negative Binomial LR	**0.082**	0.061	0.052	0.063	0.060
0.05	Simple Linear	0.058	0.049	0.052	0.057	0.059
	Weighted Linear	0.060	0.057	0.066	0.059	0.068
	Poisson Wald	**0.201**	**0.192**	**0.192**	**0.194**	**0.203**
	Poisson LR	**0.206**	**0.190**	**0.194**	**0.194**	**0.202**
	Negative Binomial Wald	**0.133**	**0.073**	**0.076**	0.061	0.069
	Negative Binomial LR	**0.091**	0.060	0.070	0.055	0.066
0.20	Simple Linear	0.047	0.050	0.040	0.066	0.051
	Weighted Linear	0.054	0.062	0.050	**0.077**	0.067
	Poisson Wald	**0.420**	**0.408**	**0.413**	**0.434**	**0.450**
	Poisson LR	**0.421**	**0.411**	**0.413**	**0.432**	**0.450**
	Negative Binomial Wald	**0.118**	**0.078**	0.058	**0.085**	0.052
	Negative Binomial LR	**0.088**	0.063	0.053	0.070	0.049
0.50	Simple Linear	0.053	0.041	0.052	0.048	0.034
	Weighted Linear	0.061	0.053	0.068	0.057	0.043
	Poisson Wald	**0.785**	**0.793**	**0.787**	**0.807**	**0.777**
	Poisson LR	**0.784**	**0.793**	**0.787**	**0.808**	**0.778**
	Negative Binomial Wald	**0.125**	**0.074**	**0.077**	0.060	0.041
	Negative Binomial LR	**0.092**	0.060	0.067	0.050	0.035

[1] Overly liberal Type I error rates are in bold

Table 4.10: Type I error rates for testing $H_0 : \beta_1 = 0$ (nominal level=0.05)[1], with event rate of 1% and group size of 20,000 (coefficient of variation = 0). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.062	0.051	0.052	0.048	0.042
	Weighted Linear	0.062	0.051	0.052	0.048	0.042
	Poisson Wald	0.064	0.050	0.051	0.038	0.037
	Poisson LR	0.065	0.050	0.051	0.037	0.040
	Negative Binomial Wald	**0.148**	**0.078**	0.069	0.058	0.049
	Negative Binomial LR	**0.109**	0.066	0.061	0.053	0.046
0.05	Simple Linear	0.060	0.056	0.043	0.043	0.050
	Weighted Linear	0.060	0.056	0.043	0.043	0.050
	Poisson Wald	**0.544**	**0.551**	**0.546**	**0.529**	**0.584**
	Poisson LR	**0.543**	**0.553**	**0.545**	**0.528**	**0.585**
	Negative Binomial Wald	**0.121**	**0.080**	0.055	0.063	0.061
	Negative Binomial LR	**0.090**	0.070	0.050	0.054	0.056
0.20	Simple Linear	0.046	0.063	0.041	0.037	0.044
	Weighted Linear	0.046	0.063	0.041	0.037	0.044
	Poisson Wald	**0.747**	**0.765**	**0.744**	**0.766**	**0.752**
	Poisson LR	**0.747**	**0.765**	**0.743**	**0.766**	**0.752**
	Negative Binomial Wald	**0.110**	**0.091**	0.053	0.051	0.060
	Negative Binomial LR	**0.086**	**0.078**	0.048	0.048	0.057
0.50	Simple Linear	0.045	0.038	0.056	0.050	0.052
	Weighted Linear	0.045	0.038	0.056	0.050	0.052
	Poisson Wald	**0.848**	**0.860**	**0.852**	**0.850**	**0.836**
	Poisson LR	**0.848**	**0.860**	**0.852**	**0.850**	**0.836**
	Negative Binomial Wald	**0.110**	0.065	**0.074**	0.067	**0.074**
	Negative Binomial LR	**0.080**	0.054	0.062	0.062	0.068

[1] Overly liberal Type I error rates are in bold

Table 4.11: Type I error rates for testing H_0: $\beta_1 = 0$ (nominal level=0.05)[1], with event rate of 1% and group size of 20,000 (coefficient of variation = 0.2). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.053	0.034	0.049	0.050	0.055
	Weighted Linear	0.050	0.040	0.045	0.049	0.051
	Poisson Wald	0.045	0.055	0.049	**0.073**	0.059
	Poisson LR	0.044	0.055	0.047	**0.072**	0.059
	Negative Binomial Wald	**0.110**	**0.076**	0.067	**0.092**	0.062
	Negative Binomial LR	**0.084**	0.066	0.059	**0.083**	0.057
0.05	Simple Linear	0.057	0.053	0.044	0.062	0.055
	Weighted Linear	0.058	0.062	0.050	0.069	0.066
	Poisson Wald	**0.564**	**0.554**	**0.558**	**0.566**	**0.563**
	Poisson LR	**0.563**	**0.552**	**0.558**	**0.566**	**0.564**
	Negative Binomial Wald	**0.123**	**0.075**	0.067	**0.073**	0.068
	Negative Binomial LR	**0.098**	0.065	0.061	0.070	0.066
0.20	Simple Linear	0.035	0.049	0.036	0.052	0.047
	Weighted Linear	0.041	0.059	0.050	0.060	0.068
	Poisson Wald	**0.734**	**0.754**	**0.765**	**0.757**	**0.742**
	Poisson LR	**0.735**	**0.754**	**0.766**	**0.755**	**0.743**
	Negative Binomial Wald	**0.114**	**0.081**	0.067	0.062	0.065
	Negative Binomial LR	**0.088**	0.069	0.061	0.059	0.063
0.50	Simple Linear	0.061	0.046	0.042	0.041	0.046
	Weighted Linear	0.065	0.060	0.052	0.049	0.059
	Poisson Wald	**0.833**	**0.840**	**0.857**	**0.844**	**0.845**
	Poisson LR	**0.832**	**0.840**	**0.856**	**0.844**	**0.845**
	Negative Binomial Wald	**0.132**	**0.078**	0.064	0.050	0.063
	Negative Binomial LR	**0.106**	0.066	0.050	0.046	0.056

[1] Overly liberal Type I error rates are in bold

Table 4.12: Type I error rates for testing H_0: $\beta_1 = 0$ (nominal level=0.05)[1], with event rate of 1% and group size of 250,000 (coefficient of variation = 0). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.059	0.052	0.053	0.057	0.044
	Weighted Linear	0.059	0.052	0.053	0.057	0.044
	Poisson Wald	0.051	0.041	0.058	0.052	0.050
	Poisson LR	0.051	0.042	0.058	0.052	0.050
	Negative Binomial Wald	**0.119**	**0.076**	**0.081**	0.067	0.055
	Negative Binomial LR	**0.097**	0.064	0.070	0.062	0.052
0.05	Simple Linear	0.042	0.042	0.053	0.058	0.055
	Weighted Linear	0.042	0.042	0.053	0.058	0.055
	Poisson Wald	**0.858**	**0.848**	**0.858**	**0.868**	**0.859**
	Poisson LR	**0.858**	**0.848**	**0.858**	**0.868**	**0.859**
	Negative Binomial Wald	**0.119**	**0.071**	0.068	**0.077**	0.061
	Negative Binomial LR	**0.094**	0.063	0.060	0.070	0.059
0.20	Simple Linear	0.049	0.057	0.061	0.058	0.053
	Weighted Linear	0.049	0.057	0.061	0.058	0.053
	Poisson Wald	**0.927**	**0.943**	**0.929**	**0.940**	**0.939**
	Poisson LR	**0.927**	**0.943**	**0.929**	**0.940**	**0.939**
	Negative Binomial Wald	**0.107**	**0.085**	**0.082**	**0.074**	0.070
	Negative Binomial LR	**0.083**	**0.072**	**0.074**	0.069	0.063
0.50	Simple Linear	0.052	0.045	0.052	0.045	0.050
	Weighted Linear	0.052	0.045	0.052	0.045	0.050
	Poisson Wald	**0.956**	**0.949**	**0.960**	**0.950**	**0.952**
	Poisson LR	**0.956**	**0.949**	**0.961**	**0.950**	**0.952**
	Negative Binomial Wald	**0.125**	**0.082**	0.070	0.063	0.062
	Negative Binomial LR	**0.098**	**0.072**	0.064	0.059	0.055

[1] Overly liberal Type I error rates are in bold

Table 4.13: Type I error rates for testing H_0: $\beta_1 = 0$ (nominal level=0.05)[1], with event rate of 1% and group size of 250,000 (coefficient of variation = 0.2). The overdispersion parameter is denoted by κ.

		Ecological Sample Size				
κ	**Model**	10	20	30	40	50
0	Simple Linear	0.050	0.027	0.045	0.052	0.050
	Weighted Linear	0.041	0.034	0.042	0.051	0.046
	Poisson Wald	0.054	0.042	0.047	0.044	0.042
	Poisson LR	0.053	0.041	0.047	0.044	0.043
	Negative Binomial Wald	**0.102**	**0.073**	0.067	0.065	0.064
	Negative Binomial LR	**0.092**	0.059	0.062	0.064	0.061
0.05	Simple Linear	0.043	0.057	0.056	0.048	0.050
	Weighted Linear	0.044	0.059	0.060	0.049	0.056
	Poisson Wald	**0.856**	**0.842**	**0.868**	**0.877**	**0.865**
	Poisson LR	**0.856**	**0.842**	**0.868**	**0.877**	**0.865**
	Negative Binomial Wald	**0.108**	**0.082**	0.070	0.070	0.061
	Negative Binomial LR	**0.077**	**0.072**	0.064	0.057	0.054
0.20	Simple Linear	0.055	0.053	0.047	0.057	0.064
	Weighted Linear	0.058	0.063	0.056	0.068	0.070
	Poisson Wald	**0.932**	**0.935**	**0.939**	**0.954**	**0.941**
	Poisson LR	**0.932**	**0.935**	**0.939**	**0.954**	**0.941**
	Negative Binomial Wald	**0.112**	**0.086**	0.058	**0.082**	0.063
	Negative Binomial LR	**0.089**	**0.074**	0.049	**0.072**	0.060
0.50	Simple Linear	0.051	0.050	0.053	0.047	0.055
	Weighted Linear	0.058	0.056	0.057	0.057	0.069
	Poisson Wald	**0.947**	**0.955**	**0.962**	**0.961**	**0.960**
	Poisson LR	**0.947**	**0.955**	**0.962**	**0.961**	**0.960**
	Negative Binomial Wald	**0.109**	**0.079**	**0.078**	0.062	0.063
	Negative Binomial LR	**0.079**	0.067	0.070	0.063	0.060

[1] Overly liberal Type I error rates are in bold

Table 4.14: Convergence rates of the negative binomial model for testing H_0: $\beta_1 = 0$. For all other models and scenarios, convergence rates were one-hundred percent.

Sample Size	**Group Size**	**(c.v.)**[†]	**Event Rate**	κ [‡]	**Convergence Rate (%)**
10	2500	0.00	5%	0.00	99.6
10	2500	0.20	5%	0.00	92.7
10	2500	0.00	1%	0.00	99.0
10	2500	0.00	1%	0.05	99.7
10	2500	0.20	1%	0.00	92.6
10	2500	0.20	1%	0.05	98.6
10	20,000	0.20	5%	0.00	94.1
10	20,000	0.20	1%	0.00	93.1
10	250,000	0.20	5%	0.00	91.3
10	250,000	0.00	1%	0.00	99.9
10	250,000	0.20	1%	0.00	94.7
20	2500	0.20	5%	0.00	99.7
20	2500	0.20	1%	0.00	99.8
20	20,000	0.20	5%	0.00	99.9
20	20,000	0.20	1%	0.00	99.9
20	250,000	0.20	5%	0.00	97.5
20	250,000	0.20	1%	0.00	99.5
30	250,000	0.20	5%	0.00	99.3
40	250,000	0.20	5%	0.00	99.9

[†] c.v. = coefficient of variation

[‡] κ = overdispersion parameter

Table 4.15: Power for testing H_0: $\beta_1 = 0$ when $\beta_1 = 2$, ecological sample size = 50, coefficient of variation = 0, event rates are 5%, and group-level exposure prevalence is low. Models with overly liberal Type I error rates have been omitted.

		Overdispersion Parameter			
Group Size	**Model**	0.00	0.05	0.20	0.50
2500	Simple Linear	1.00	0.945	0.532	0.268
	Weighted Linear	1.00	0.945	0.532	0.268
	Negative Binomial Wald	1.00	0.949	0.527	0.247
	Negative Binomial LR	1.00	0.947	0.526	0.252
	Poisson Wald	1.00	–	–	–
	Poisson LR	1.00	–	–	–
20,000	Simple Linear	1.00	0.952	0.542	0.272
	Weighted Linear	1.00	0.952	0.542	0.272
	Negative Binomial Wald	1.00	0.955	0.535	0.258
	Negative Binomial LR	1.00	0.953	0.539	0.264
	Poisson Wald	1.00	–	–	–
	Poisson LR	1.00	–	–	–
250,000	Simple Linear	1.00	0.963	0.536	0.260
	Weighted Linear	1.00	0.963	0.536	0.260
	Negative Binomial Wald	1.00	0.967	0.538	0.243
	Negative Binomial LR	1.00	0.964	0.536	0.253
	Poisson Wald	1.00	–	–	–
	Poisson LR	1.00	–	–	–

Table 4.16: Power for testing H_0: $\beta_1 = 0$ when $\beta_1 = 2$, ecological sample size = 50, coefficient of variation = 0.20, event rates are 5%, and group-level exposure prevalence is low. Models with overly liberal Type I error rates have been omitted.

		Overdispersion Parameter			
Group Size	**Model**	0.00	0.05	0.20	0.50
2500	Simple Linear	0.972	0.789	0.522	0.254
	Weighted Linear	0.973	0.881	0.511	0.254
	Negative Binomial Wald	1.00	0.937	0.533	0.261
	Negative Binomial LR	1.00	0.937	0.536	0.267
	Poisson Wald	1.00	–	–	–
	Poisson LR	1.00	–	–	–
20,000	Simple Linear	1.00	0.829	0.472	0.249
	Weighted Linear	1.00	0.836	0.482	0.263
	Negative Binomial Wald	1.00	0.963	0.512	0.239
	Negative Binomial LR	1.00	0.963	0.554	0.248
	Poisson Wald	1.00	–	–	–
	Poisson LR	1.00	–	–	–
250,000	Simple Linear	0.984	0.838	0.449	0.262
	Weighted Linear	0.988	0.835	0.444	0.266
	Negative Binomial Wald	1.00	0.979	0.531	0.249
	Negative Binomial LR	1.00	0.977	0.533	0.252
	Poisson Wald	1.00	–	–	–
	Poisson LR	1.00	–	–	–

4.5 Discussion

This simulation has revealed several key findings about the performance of linear and nonlinear regression models within the context of group-level count data. In particular, we have shown that the Wald estimator of the negative binomial regression model does not provide adequate Type I error rates with small sample sizes, even when the count data originates from a negative binomial process with known covariates and overdispersion values. Predictably, the Likelihood Ratio statistic provided better Type I error rates and so achieved nominal levels with fewer ecological units. Previous empirical investigations support this finding. For example, Lawless (1987) summarizes a simulation performed by Charmane Dean, who found that inferences regarding β using Likelihood Ratio statistics are satisfactory except in very small samples. This is because the Likelihood Ratio statistic quickly approaches the Chi-square distribution. This is of particular importance to researchers using SAS as their analytic tool, since SAS provides the Wald statistic by default but will only provide Likelihood Ratio statistics when ordered to do so. Thus, users unaware of the limitations of the Wald statistic will automatically and unknowingly receive the p-values that are likeliest to cause false-positive conclusions.

We have also shown that OLS and Weighted LS regression models have Type I error rates that are surprisingly robust to small ecological sample sizes, overdispersion, rare events, and group size. Unlike Poisson and negative binomial regression, these two models were consistently reliable when the data were overdispersed and sample sizes were less than 30. While this is certainly a desirable property for researchers with small ecological datasets violating assumptions of equidispersion, it is worth emphasizing that issues of cross-level bias remain in spite of the robust performance. As discussed in Chapter 3, the ecological fallacy is a known potentiality for linear models. Our simulation focused on aggregated data, and therefore Type I error rates refer to the group-level regression coefficient, not the individual-level one. This is true for the generalized linear models as well.

In light of these findings, it is recommended that researchers investigating group-level count data avoid the negative binomial model when faced with sample sizes less than 20, or perhaps 30. We also recommend that the Likelihood Ratio statistic be used in lieu of the Wald statistic when the data are believed to follow a Poisson or negative binomial process. In the case of very small samples, a combination of modeling approaches would perhaps be most prudent. Specifically, linear models could be used to test the null hypothesis, and since the generalized linear models are unbiased even with small samples, they could be used for effect measure estimation if there is evidence of association. We also recommend that the researcher be cautious when the data show evidence of overdispersion, since this can drastically reduce power for even the most robust model. The negative binomial model, which should have optimal power, was no exception.

Chapter 5

Example: An International Ecologic Study of Gun Ownership and Homicide Rates

5.1 Introduction

The purpose of this chapter is to provide an example of ecological regression using data from Koepsell and Weiss (2003) and Killias (1993). We shall use OLS, weighted LS, Poisson, and negative binomial regression to analyze the association between national gun ownership prevalence and national homicide rates. These data come from Killias (1993), who found a positive association between levels of gun ownership and rates of homicide and suicide (Spearman correlation coefficient $\rho_s = 0.658$). We will compare the results of the four regression methods to Killias' findings and to each other.

Killias (1993) used national-level data from 13 countries in Western Europe and North America. We present some of these data in Table 5.1. National-level data were chosen because it was felt that state-level data would not provide enough variability in the prevalence of gun ownership. We fit the four aforementioned regression models to the data to see whether they would support Killias' ecological correlation. For each model, we set the national homicide rate to be the outcome, with the percentage of households with firearms as the only predictor. Results are presented in Table 5.2. Fitted regression curves are shown in Figures 5.1 and 5.2.

Table 5.1: Levels of Gun-ownership and Homicide Rates per million persons in 13 Countries

Country	Overall Homicide Rate	% of Households With Guns	Population (millions)
Australia	19.5	19.6	16.7
Belgium	18.5	16.6	9.9
Canada	26.0	29.1	27.1
England & Wales	6.7	4.7	47.1
Finland	29.6	23.2	4.9
France	12.5	22.6	55.4
Netherlands	11.8	1.9	14.7
Norway	12.1	32.0	4.2
Scotland	16.3	4.7	5.1
Spain	13.7	13.1	40.5
Switzerland	11.7	27.2	6.2
United States	75.9	48.0	248.0
West Germany	12.1	8.9	60.7

5.2 Aspects of Study Data

It is interesting to note that the homicide, gun ownership, and population data all come from different sources and time periods. The percentage of households with firearms was estimated for each country via the International Crime Survey, which collected data from the year 1989. Homicide data between 1983 and 1986 came from the World Health Organization. Koepsell and Weiss (2003) obtain population estimates from the 1990 World Almanac. Obviously, the use of multiple sources and different dates may be problematic. The reliance on survey data such as these is a potential limitation for many ecological studies.

Also, the large population size, gun ownership level, and homicide rate of the United States as compared to the other countries are quite striking. These are by far the largest in each of their respective categories. This means that the United States may drive much of the association between homicide rates and gun ownership. England & Wales appear to occupy the opposite end of the spectrum, with homicide rates and gun ownership percentages in the single digits.

The average population size is 41.6 million, with a standard deviation of 65.2 million. Therefore, the estimated coefficient of variation is 1.57, which is much higher than the value used in the simulation. Based on the results of the simulation, however, we might expect the power of tests based on linear regression models to be reduced. If overdispersion is low, this may be at least partially offset by the considerable group sizes of these nations, which are orders of magnitude larger than the group sizes we investigated. The Wald tests and Likelihood Ratio tests should remain unaffected by the high variability in group size. Power for all models may be increased by the relatively large variability in exposure, which ranges from 1.9% to 48%. The net effect of these characteristics in terms of power is difficult to predict.

5.3 Results

From the results of Table 5.2, it can be seen that both the linear and generalized linear models found a highly significant positive association between national levels of gun ownership and homicide rates. At the ecological level at least, it would seem that these results support the findings of Killias. The weighted LS regression was better able to account for the disparate population sizes of the countries, and so explained more of the variability in homicide rates than the OLS model (Adjusted R^2: 0.91 vs 0.48 respectively). The negative binomial model detected overdispersion, and estimated its value to be 0.14. This is comparable to what was considered 'moderate' overdispersion in the simulation. As a result, we should expect the Poisson model to be highly susceptible to Type I error rates, and should take its conclusions with caution. Nevertheless, both of the generalized linear models should produce unbiased effect measure estimates, and we do find that their estimates are similar.

Table 5.2: Results of Four Regression Models Estimating the Effect of National Gun Ownership Prevalence on Homicide Rates for 13 Western Countries

Model	**Test**	$\hat{\beta}$ †	**95% Confidence Limits**	**P-value**	$\widehat{RR}$ ‡
Simple Linear	T-test	0.987	(0.365, 1.608)	0.005	72
Weighted Linear	T-test	1.678	(1.342, 2.017)	< 0.001	*Incalculable*
Poisson	Wald	0.052	(0.051, 0.053)	< 0.0001	185
Poisson	LR	0.052	(0.051, 0.053)	< 0.0001	185
Negative Binomial	Wald	0.036	(0.021, 0.051)	< 0.0001	36
Negative Binomial	LR	0.036	(0.020, 0.052)	< 0.0001	36

† $\hat{\beta}$ = estimated regression coefficient for proportion of households with firearms

‡ estimated rate ratio, comparing nations with firearms in every home to nations without firearms

5.4 Discussion

Based on the results of any of these models, we might tentatively conclude that the number of homicides in a country is associated with number of households with firearms. But what are the limitations of these data? We know from the bibliometric review that they are not atypical of cross-sectional ecological studies. They come from pre-existing tabulations and surveys; are at a common level of aggregation; the outcome is a rate and the exposure is binary; there are few covariates and no confounders or effect modifiers; and the sample size is small. The last of these characteristics is perhaps the most important. With only 13 countries in the analysis, we have little information on which to base our conclusions. From the simulation, we know that the negative binomial model has unreliable Type I error rates with this sample size. Since we appear to be faced with overdispersion, so too does the Poisson model. In fact, since the simulation found that the Poisson model performs worst when the data are overdispersed and group sizes are large, we should expect the group sizes of this data to render the Poisson model completely ineffectual. For hypothesis testing of the regression coefficient, we are therefore forced to turn to the linear models.

But which linear model? Either of them will reject the null hypothesis. However, they provide us with very different effect measures. We may calculate a rough rate ratio by using the fitted regression equations to calculate the predicted homicide rates when all households own firearms and when no households own firearms. With this method, the simple linear model estimates the rate ratio to be 71.5, meaning that the expected rate of homicides is approximately 72 times higher in a country where all homes own a firearm compared to a country that owns none. The weighted model, on the other hand, falls prey to one of the weaknesses of linear models; it provides a negative rate ratio, which is meaningless. Unfortunately, it is the weighted linear model that seems to explain more of the variability in the data, and so choosing the simple linear model may come at the expense of precision. On the other hand, the weighted model may be overly sensitive to the data from the US, which has an influentially large population. In all likelihood, the best approach would perhaps be to use a combination of the linear and generalized linear models, as suggested in Chapter 4.

Killias himself recognizes some of the limitations of the data. He remarks on the small sample size and the cross-sectional nature of the survey, which prevents any causal inferences. Unlike most of the papers in the bibliometric review, he justifies an ecological analysis by noting that the availability of firearms varies mostly with laws enacted at the state and national level, making individual-level study difficult. However, he does not acknowledge any of the ecological limitations of the data. The study is not stated to be ecological or aggregate, and the potential for the ecological fallacy is not mentioned. This is in spite of the fact that Killias concludes that the data suggest "the presence of a gun in the home increases the likelihood of homicide or suicide," a statement implying cross-level inference. The lack of clear reporting combined with the methodological simplicity and small sample size of the analysis means that this paper is typical of those in need of the recommendations suggested in Chapter 2.

5.5 Summary

This example demonstrates a few of the analytic complexities of ecological analysis when using small samples, some of which are known from the literature and some of which have been shown by the simulation. Highly variable, under-powered count data are not suitable for Poisson or negative binomial regression, and linear models can provide biased or even meaningless effect measures. These difficulties can complicate interpretation and make the proper modeling approach unclear. In this case, there is likely no single correct method. Instead, one should employ a battery of techniques. The ecological researcher should be aware of the capabilities of these models and the characteristics of their dataset before interpreting any results.

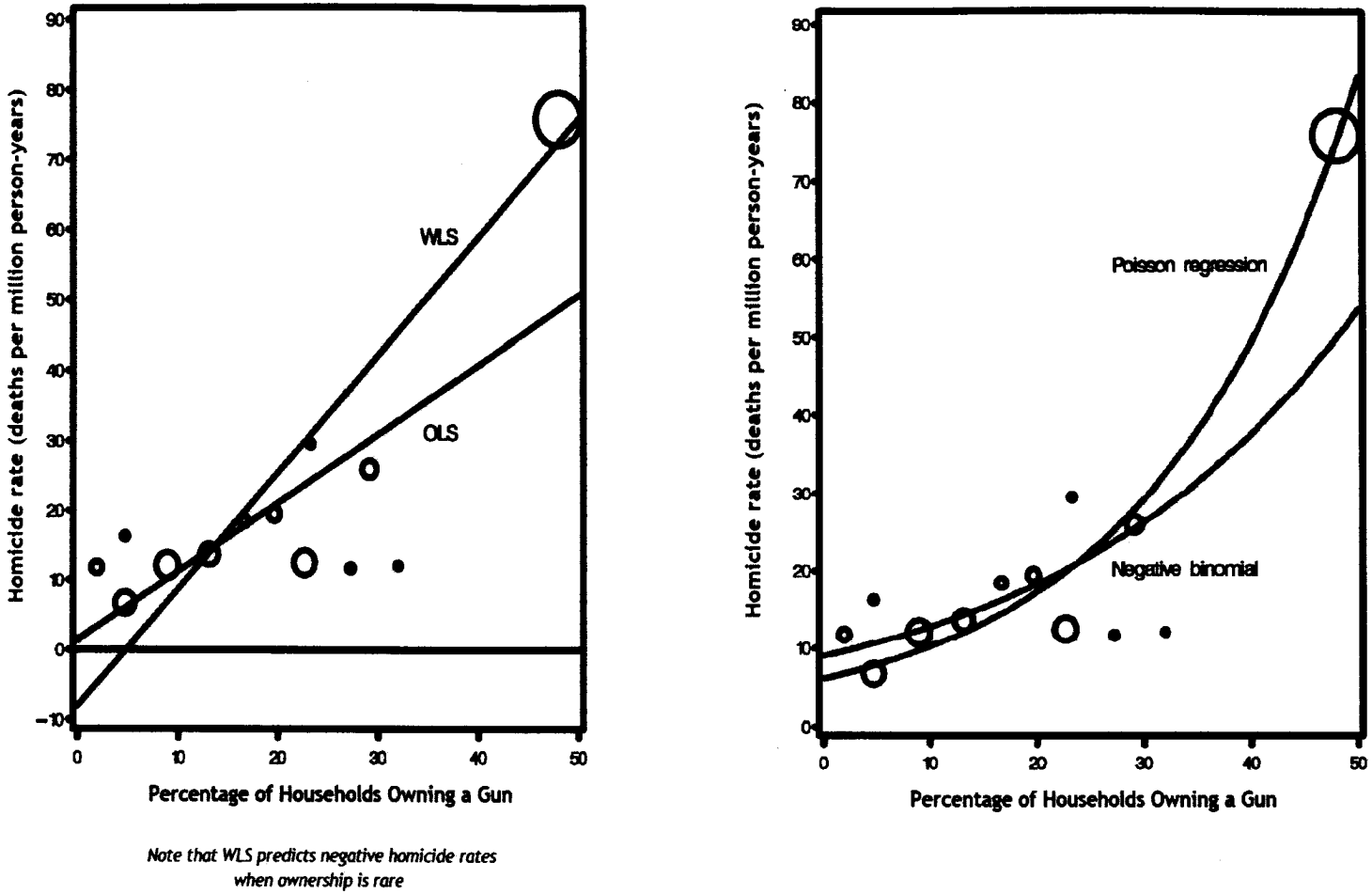

Figure 5.1: Linear and Generalized Linear Models Fit to National Homicide and Gun Ownership Data

Chapter 6

Discussion

6.1 Introduction

The primary purpose of this thesis was to critically evaluate the quality of ecological study in modern epidemiology and to assess the performance of four common regression methods within this context. Attention was restricted to cross-sectional ecological studies with a particular focus on event-rate data. This final chapter presents a brief discussion of the research contained in this thesis. Key findings are summarized in Section 6.2. Limitations of the research are described in section 6.3, while areas for future work are outlined in Section 6.4.

6.2 Key Findings

This research has found that the diversity of modern ecological studies in terms of quality, methodology, and areas of application is quite high. The reviewed papers collectively investigated social, environmental, economic, and medical data from all parts of the globe and used analytic tools ranging from the most rudimentary to the most complex. While this diversity implies the existence of high-quality papers, it also means that a great many studies methodologically unsound, unclear, and have little power. Unfortunately, we found that the latter seem to be the most prevalent. The largest problem is the poor quality of reporting and

the lack of clear justification. If epidemiologists are to continue using this study design, the recommendations presented in Table 2.6 should be heeded. Also, certain biostatistical results should be better incorporated into the analytic strategies used by ecological researchers. In particular, the covariate adjustment techniques discussed by Rosenbaum and Rubin (1984) as well as the validity issues for regression discussed by Vittinghoff and McCulloch (2007) must be addressed in future publications.

The simulation has demonstrated the fallibility of Poisson and negative binomial regression when faced with small sample sizes and overdispersion respectively. It has also discovered that Type I error rates of OLS and weighted LS models are robust to the parameter combinations used here. Implications for power seem to lie mostly with overdispersion, though extremity of group-level exposure prevalence appears to negate this. Thus, researchers investigating counts and rates should be wary of Type I error rates when using generalized linear models with small samples. As well, one must be equally concerned with power when the data are overdispersed and group-level exposure variability is not large.

6.3 Study Limitations

The bibliometric review is primarily limited in its mandate. We chose to scrutinize only cross-sectional studies for reasons mentioned in Chapter 1, and so all other forms of ecological design are beyond our scope. Comments regarding the prevalent characteristics and quality of the literature cannot be extended to temporal and spatiotemporal ecological analyses. In all likelihood, the overall quality and technical issues concerning these designs are significantly different than those facing the cross-sectional design. This is because the cross-sectional design is the simplest of the three, and will allow researchers with lower levels of expertise to participate. Another reason is that the temporal and spatiotemporal designs have unique statistical considerations, such as time-series autocorrelation.

The bibliometric review is also limited in the sense that only certain evaluation criteria have been used. While we attempted to cover the most important and interesting aspects, not all points could be addressed. For example, we did not investigate whether the cross-sectional nature of the data was mentioned or whether improper causal inferences were made from it. While this is not a hindrance, it does allow a few minor aspects of quality to remain unmeasured.

The last and perhaps most obvious limitation of the review is that only six journals were included. These were selected to be representative of specialized epidemiology publications, and so inter-disciplinary journals such as the *New England Journal of Medicine* were omitted. These may hold their authors to different standards than the journals we used.

The principal limitation of the simulation study is that only two methods of count-data generation were employed: Poisson and negative binomial. As shown by Metcalfe and Thompson (2006), the results of a simulation can be quite sensitive to the methods of data generation. Thus, we cannot generalize our results or recommendations beyond datasets possessing the properties of these models. Researchers analysing count data with characteristics such as autoregression, time-dependent event rates, or variance structures different than those investigated here must not heed our results without caution.

Another limitation stems from the fact that only certain parameter combinations have been used. We attempted to investigate pragmatic scenarios and to therefore cover many plausible circumstances, but we could not analyse all of them. Scenarios with cluster sizes, cluster variation, regression coefficients, and population exposure rates different than those contained in this thesis are beyond the scope of our discussion. We have also limited our attention to the simple case of one predictor, and so issues such as adjustment for confounding and effect modification have not been included. Also, we have set the exposure to be binary at the individual level, thus leaving the effects of a continuous exposure uninvestigated. Again, these considerations potentially limit the generalizability of our findings.

6.4 Future Research

There are many possible extensions of the simulation performed in this thesis. Subsequent research could investigate more complex regression models that explore the effects of confounding or interaction; multivariate outcomes; spatial dependence; continuous exposures; and count data with non negative-binomial properties. In the latter case, new methods of analysis such as zero-inflated or zero-truncated negative binomial regression models could be introduced. Extensions of the Poisson model with additive gamma or multiplicative Gaussian random-effects could also be investigated, as these are capable of handling overdispersed count data and may be more comparable to the negative binomial model (Gupta *et al.*, 2004). The inclusion of spatial effects would be highly relevant, since this aspect of ecological study has become more prevalent in recent decades. An excellent discussion of ecological modeling with spatial clustering and overdispersion is provided by Clayton *et al.* (1993), and may serve as one motivation for future work.

Our simulation has been limited to group-level data. Other experiments may wish to assess the performance of linear and generalized linear models when individual-level relationships are known. This would require the generation and aggregation of individual outcomes and exposures, and so would be a more complicated task. However, such an analysis could address issues of cross-level bias. This would be an important insight into the modeling of ecological counts and rates. Issues such as pure specification bias, aggregation bias, spatial dependence, and contextual effects could potentially be explored with this approach. A discussion of the complexities facing cross-level inference with spatial and non-spatial regression techniques is provided by Wakefield (2007).

Finally, a bibliometric review of the temporal and spatiotemporal ecological designs would be a beneficial addition to the literature. This would allow a similar assessment of quality, and could identify further necessary additions to the STROBE document. This would also increase our understanding of current ecological practice in general, and may reveal common characteristics, strengths, and shortcomings between the three sub-designs.

References

Altman, D.G."The scandal of poor medical research", *British Medical Journal* 1994; 308: 283-284

Analitis A., Katsouyanni K., et al "Short-term effects of ambient particles on cardiovascular and respiratory mortality", *Epidemiology* 2006; 17 (12): 230-233

Bradley J.V. "Robustness?", *British Journal of Mathematical and Statistical Psychology* 1978; 31: 144-152

Burton A., Altman D.G., Royston P., Holder R.L. "The design of simulation studies in medical statistics", *Statistics in Medicine* 2006; 25: 4279-4292

Clayton D.G., Bernardinelli L., and Montomoli C. "Spatial Correlation in Ecological Analysis", *International Journal of Epidemiology* 1993; 22 (6): 1193-1202

Crawford C.A.G. and Young L.J. "A Spatial View of the Ecological Inference Problem", *Ecological Inference: New Methodological Strategies*, 2004. Ed. King G., Rosen O., and Tanner M.A. Cambridge University Press. ISBN: 0521835135

Dogan, M. and Rokkan S. *Quantitative Ecological Analysis in the Social Sciences*, 1969, MIT Press. ISBN: 0262040220.

Duncan O.D., Cuzzort R.P., and Duncan B.D. *Statistical Geography*, 1961, Glencoe, Illinois: The Free Press

von Elm E., Altman D.G., Egger M., Pocock S.J., Gotzsche P.C., and Vandenbroucke J.P. "The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) Statement", *Epidemiology* 2007; 18 (6): 800-804

Gardner W., Mulvey E.P., Shaw E.C. "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models", *Psychological Bulletin* 1995; 118(3): 392-404.

Green P.J. and Richardson S. "Hidden Markov models and disease mapping", *Journal of the American Statistical Association* 2002; 97 (460): 1055-1070

Greenland S. "Divergent biases in ecological and individual-level studies", *Statistics in Medicine* 1992; 11 (9): 1209-1223

Greenland S. and Morgenstern H. "Ecological bias, confounding, and effect measure modification", *International Journal of Epidemiology* 1989; 18 (1): 269-274

Greenland S. and Robins J. "Invited commentary: Ecologic studies biases, misconceptions, and counterexamples", *American Journal of Epidemiology* 1994; 139 (8): 747-760

Gupta P.L., Gupta R.C., and Ong S.H. "Modelling Count Data by Random Effect Poisson Model", *Sankhya: The Indian Journal of Statistics* 2004; 66 (3): 548-565

Guthrie K.A. and Sheppard L. "Overcoming biases and misconceptions in ecological studies", *Journal of the Royal Statistical Society Series A* 2001; 164: 141-154

Hemenway D., Kennedy B.P., Kawachi I., and Putnam R.D., "Firearm Prevalence and Social Capital", *Annals of Epidemiology* 2001; 11 (7): 484-490

Hilbe J., *Negative Binomial Regression*, 2007, Cambridge University Press.
ISBN: 0521857724

Johnston S.C. "Combining ecological and individual variables to reduce confounding by indication: case study - subarachnoid hemorrhage treatment", *Journal of Clinical Epidemiology* 2000; 53 (12): 1236-1241

Kalbfleisch, J.G. *Probability and Statistical Inference*, 1985, Springer-Verlag.
ISBN: 0387961445

Kelsall J. and Wakefield J. "Modeling spatial variation in disease risk: A geostatistical approach", *Journal of the American Statistical Association* 2002; 97 (459): 692-701

Killias M. "International correlations between gun ownership and rates of homicide and suicide", *Canadian Medical Association Journal* 1993; 148(10): 1721-1725

Klar N. and Darlington G. "Methods for modelling change in cluster randomization trials", *Statistics in Medicine* 2004; 23: 2341-2357

Kleinschmidt I., Sharp B.L., Clarke G.P.Y., Curtis B., and Fraser C. "Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in Kwazulu Natal, South Africa", *American Journal of Epidemiology* 2001; 153 (12): 1213-1221

Knapp T.R. "The unit-of-analysis problem in applications of simple correlation analysis to educational research", *Journal of Educational Statistics* 1977; 2 (3): 171-186

Koepsell T.D. andWeiss N.S. *Epidemiologic Methods: Studying the Occurrence of Illness*, 2003, Oxford University Press. ISBN-13 978-0-19-515078-0

Lancaster G.A., Green M., and Lane S. "Reducing bias in ecological studies: an evaluation of different methodologies", *Journal of the Royal Statistical Society Series A* 2006; 169 (4): 681-700

Lawless J.F. "Negative binomial and mixed Poisson regression", *Canadian Journal of Statistics* 1987; 15 (3): 209-225

Liang, K. Y. and Zeger, S. L. "Longitudinal data analysis using generalized linear models" *Biometrika* 1986; 73 (1): 13-22

Loomis D., Wolf S.H., Runyan C.W., et al. "Homicide on the job: community and workplace determinants", *American Journal of Epidemiology* 2001; 154 (5): 410-417

Lovasi G.S.,Quinn J.W., Neckerman K.M., Perzanowski M.S., and Rundle A. "Children living in areas with more street trees have lower prevalence of asthma", *Journal of Epidemiology and Community Health* 2008; 62 (7): 647-649

March D. and Susser E. "The eco- in eco-epidemiology", *International Journal of Epidemiology 2006*; 35 (6): 1379-1383.

Metcalfe C. and Thompson S.G. "The importance of varying the event generation process in simulation studies of statistical methods for recurrent events", *Statistics in Medicine* 2006; 25: 165-179

Mezei G., Borugian M.J., Spinelli J., Wilkins R., Abanto Z., and McBride M.L. "Socioeconomic status and childhood tumor and lymphoma incidence in Canada", *American Journal of Epidemiology* 2006; 164: 170-175

Moore, D. F. and Tsiatis, A. "Robust Estimation of the Variance in Moment Methods for Extra-binomial and Extra-Poisson Variation", *Biometrics* 1991; 47: 383401.

Morgenstern H. "Uses of ecologic analysis in epidemiologic research", *American Journal of Public Health* 1982; 72 (12): 1336-1344

Morgenstern H. "Ecological studies in epidemiology: concepts, principles, and methods", *Annual Review of Public Health* 1995; 16: 61-81

Morgenstern H., Rothman K.J., Greenland S., and Lash T.L. *Modern Epidemiology*, third edition, 2008, Lippincott Williams & Wilkins. ISBN 0781755646

Morris J.N. *Uses of Epidemiology.* Edinburgh: Livingstone, 1957

Moulton L.H. and Benini A.A. "Community level risk factors for number of landmine victims in Chad and Thailand", *Journal of Epidemiology and Community Health* 2003; 57: 956-959

Oakes J.M. and Kaufman J.S. *Methods in Social Epidemiology*, 2006, Jossey-Bass Publishers. ISBN: 0787979899

Pocock S.J., Cook D.G., and S.A. Beresford. "Regression of area mortality rates on explanatory variables: what weighting is appropriate?", *Applied Statistics* 1981; 30 (3): 286-295

Rezaeian M., Dunn G., St Leger S., and Appleby L. "Ecological association between suicide rates and indices of deprivation in the north west region of England: the importance of the size of the administrative unit", *Journal of Epidemiology and Community Health* 2006; 60: 956-961

Richardson S., Stucker I., and Hmon D. "Comparison of relative risks obtained in ecological and individual studies: some methodological considerations", *International Journal of Epidemiology* 1987; 16 (1): 111-120

Riva M., Guavin L., and Barnett T.A. "Toward the next generation of research into small area effects on health: a synthesis of multilevel investigations published since July 1998", *Journal of Epidemiology and Community Health* 2007; 61: 853-861

Robinson, W.S. "Ecological correlations and the behavior of individuals", *American Sociological Review* 1950; 15 (3): 351-357

Rosenbaum P.R. and Rubin D.B. "Difficulties with regression analyses of age-adjusted rates", *Biometrics* 1984; 40: 437-443

Schwartz, S. "The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences", *American Journal of Public Health* 1994; 84 (5): 819-823

Shakhtarin V.W., Tsyb A.F., Stepanenko V.F., *et al.* "Iodine deficiency, radiation dose, and the risk of thyroid cancer among children and adolescents in the Bryansk region of Russia following the Chernobyl power station accident", *International Journal of Epidemiology* 2003; 32: 584-591

Shaw P.A., Etzioni R., Zeliadt S.B., *et al.* "An ecologic study of prostate-specific antigen screening and prostate cancer mortality in nine geographic areas of the United States", *American Journal of Epidemiology* 2004; 160(11): 1059-1069

Szwarcwald C.L., Bastos F.I., Barcellos B., *et al.* "Health conditions and residential concentration of poverty: a study in Rio de Janeiro, Brazil", *Journal of Epidemiology and Community Health* 2000; 54 (7): 530-536

Thorndike E.L. "On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them", *American Journal of Psychology* 1939; 52: 122-124

Treggiari M.M. and Weiss N.S. "Occupation asbestos exposure and the incidence of non-Hodgkin's lymphoma of the gastrointestinal tract: an ecological study", *Annals of Epidemiology* 2004; 14 (3): 168-171

Vittinghoff E. and McCulloch C.E. "Relaxing the rule of ten events per variable in Logistic and Cox regression", *American Journal of Epidemiology* 2007; 165: 710-718

Wakefield J. "Disease mapping and spatial regression with count data", *Biostatistics* 2007; 8 (2): 158-183

Wakefield J. "Ecological studies revisited", *Annual Review of Public Health* 2008; 29: 75-90

Wakefield J. and Haneuse S. "Overcoming ecological bias using the two-phase study design", *American Journal of Epidemiology* 2008; 167 (8): 908-916

Weitzen S., Lapane K.L., Toledano A.Y., Hume A.L., and Mor, V. "Principles for modeling propensity scores in medical research: a systematic literature review, *Pharmacoepidemiology and Drug Safety* 2004; 13: 841-853

Zhang J., Temme E., and Kesteloot H. "Fish consumption is inversely associated with male lung cancer mortality in countries with high levels of cigarette smoking or animal fat consumption", *International Journal of Epidemiology* 2000; 29: 615-621

Appendix A: Studies Referenced in Bibliometric Review

American Journal of Epidemiology

Aamodt G., Bukholm G., Jahnsen J., *et al.* "The association between water supply and inflammatory bowel disease based on a 1990 - 1993 cohort study in southeastern Norway", 2008; 168 (9): 1065-1072

Bates S.J., Trostle J., Cevallos W.T., *et al.* "Relating diarrheal disease to social networks and the geographic configuration of communities in rural Ecuador", 2007; 166 (9): 1088-1095

Blanchard J.F., Bernstein C.N., Wajda A., *et al.* "Small-area variations and sociodemographic correlates for the incidence of Crohn's disease and ulcerative colitis", 2001; 154 (4): 328-335

Feltbower R.G., Manda S., Gilthorpe M.S., *et al.* "Detecting small-area similarities in the epidemiology of childhood acute lymphoblastic leukemia and diabetes mellitus, type 1: a Bayesian approach", 2005; 161 (12): 1168-1180

Gilpin E.A., Stillman F.A., Hartman A.M., *et al.* "Index for US state tobacco control initial outcomes", 2000; 152 (8): 727-738

Green C., Elliott L., Beaudoin C., *et al.* "A population-based ecologic study of inflammatory bowel disease: searching for etiologic clues", 2006; 164 (7): 615-623

Gutierrez E., Taucer F., De Groeve T., *et al.* "Analysis of worldwide earthquake mortality using multivariate demographic and seismic data", 2005; 161 (12): 1151-1158

Hodgson S., Nieuwenhuijsen M.J., Elliott P., *et al.* "Kidney disease mortality and environmental exposure to mercury", 2007; 165 (1): 72-77

Kleinschmidt I., Sharp B.L., Clarke G., *et al.* "Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in Kwazulu Natal, South Africa", 2001; 153 (12): 1213-1221

Mezei G., Borugian M., Spinelli J.J., *et al.* "Socioeconomic status and childhood solid tumor and lymphoma incidence in Canada", 2006; 164 (2): 170-175

Michelozzi P., Capon A., Kirchmayer U., *et al.* "Adult and childhood leukemia near a high-power radio station in Rome, Italy", 2002; 155 (12): 1096-1103

Shaw P.A., Etzioni R., Zeliadt S.B., *et al.* "An ecologic study of prostate-specific antigen screening and prostate cancer mortality in nine geographic areas of the US", 2004; 160 (11): 1059-1069

Annals of Epidemiology

Deuchert E. and Brody S. "Lack of autodisable syringe use and health care indicators are associated with high HIV prevalence: an international ecologic analysis", 2007; 17 (3): 199-207

Garland C.F., Garland F.C., and Gorham E.D.. "Epidemiologic evidence for different roles of ultraviolet A and B radiation in melanoma mortality rates", 2003; 13 (6): 395-404

Hall S.A., Kaufman J.S., Millikan R.C., *et al.* "Urbanization and breast cancer incidence in North Carolina, 1995-1999", 2005; 15 (10): 796-803

Hemenway D., Kennedy B.P., Kawachi I., *et al.* "Firearm prevalence and social capital", 2001; 11 (7): 484-490

Holtgrave D.R. and Crosby R. "Is social capital a protective factor against obesity and diabetes? Findings from an exploratory study", 2006; 16 (5): 406-408

Palma-Solis M., Vives-Cases C., and Alvarez-Dardet C. "Gender progress and government expenditure as determinants of femicide", 2008; 18 (4): 322-329

Peek-Asa C., Ramirez M.R., Shoaf K., *et al.* "GIS mapping of earthquake related deaths and hospital admissions from the 1994 Northridge, California, earthquake", 2000; 10 (1): 5-13

Rosenfeld R., Baumer E., and Messner S.F. "Social trust, firearm prevalence, and suicide", 2007; 17 (2): 119-125

Saadat M. and Zendeh-Boodi Z. "Correlation between incidence of self-inflicted burns and means of inbreeding coefficients, an ecologic study", 2006; 16 (9): 708-711

Treggiari M.M. and Weiss N.S. "Occupation asbestos exposure and the incidence of non-Hodgkin's lymphoma of the gastrointestinal tract: an ecological study", 2004; 14 (3): 168-171

Yeh Y.P., Chang H.J., Yang J., *et al.* "Incidence of tuberculosis in mountain areas and surrounding townships: dose-response relationship by goegraphic analysis", 2005; 15 (7): 526-532

Zhang J., Zhao Z., and Berkel H.J. "Animal fat consumption and pancreatic cancer incidence: evidence of interaction with cigarette smoking", 2005; 15 (7): 500-508

Epidemiology

Pikhart P., Bobak M., Kriz B., *et al.* "Outdoor air concentrations of nitrogen dioxide and sulfur dioxide and prevalence of wheezing in children", 2000; 11 (2): 153-160

Beard J.R., Earnest A., Morgan G., *et al.* "Socioeconomic disadvantage and acute coronary events: a spatiotemporal analysis", 2008; 19 (3): 485-492

Beaumont J.J., Sedman R.M., Reynolds S.D., *et al.* "Cancer mortality in a Chinese population exposed to hexavalent chromium in drinking water", 2008; 19 (1): 12-23

Graafmans W.C., Richardus J.H., Borsboom G., *et al.* "Birth weight and perinetal mortality: a comparison of "optimal" birth weight in seven western European countries", 2002; 13 (5): 569-574

Miller M., Azrael D., and Hemenway D. "Household firearm ownership and suicide rates in the United States", 2002; 13 (5): 517-524

Naess O., Piro F.N., Nafstad, P., *et al.* "Air pollution, social deprivation, and mortality: a multilevel cohort study", 2007; 18 (6): 686-694

Petridou E., Papadopoulos F.C., Frangakis C.E., *et al.* "A role of sunshine in the triggering of suicide", 2002; 13 (1): 106-109

Sieswerda L.E., Soskolne C.L., Newman S.C., *et al.* "Toward measuring the impact of ecological disintegrity on human health", 2001; 12 (1): 28-32

International Journal of Epidemiology

Anderson H.R., Gupta R., Kapetanakis V., *et al.* "International correlations between indicators of prevalence, hospital admissions and mortality for asthma in children", 2008; 37: 573-582

Antunes J.L.F., Waldman E.A., and Borrell C. "Is it possible to reduce AIDS deaths without reinforcing inequalities in health?", 2005; 34: 586-592

Aylin P., Morris S., Wakefield J., *et al* . "Temperature, housing, deprivation, and their relationship to excess winter mortality in Great Britain, 1986 - 1996", 2001; 30: 1100-1108

Bills C.B. and Li G. "Correlating homicide and suicide", 2005; 34: 837-845

Borroto R.J. and Martinez-Piedra R. "Geographic patterns of cholera in Mexico, 1991-1996", 2000; 29: 764-772

Clark M., Riben P., and Nowgesic E. "The association of housing density, isolation and tuberculosis in Canadian First Nations communities", 2002; 31: 940-945

Emch M., Ali M., Park J.K., *et al* "Relationship between neighbourhood-level killed oral cholera vaccine coverage and protective efficacy: evidence for herd immunity", 2006; 35: 1044-1050

Foliaki S., Kildegaard-Nielsen S., Bjorksten B., *et al.* "Antibiotic sales and the prevalence of symptoms of asthma, rhinitis, and eczema: the International Study of Asthma and Allergies in Childhood (ISAAC)", 2004; 33: 558-563

Fukuda Y., Nakamura K., and Takano T. "Cause-specific mortality differences across socioeconomic position of municipalities in Japan, 1973-1977 and 1993-1998: increased importance of injury and suicide in inequality for ages under 75", 2005; 34: 100-109

Houweling T., Kunst A.E., Looman C., *et al.* "Determinants of under-5 mortality among the poor and the rich: a cross-national analysis of 43 developing countries", 2005; 34: 1257-1265

Hurtig AK and San Sebastian M. "Geographical differences in cancer incidence in the Amazon basin of Ecuador in relation to residence near oil fields", 2002; 31: 1021-1027

Jackson L.E., Hilborn E.D., and Thomas J.C. "Towards landscape design guidelines for reducing Lyme disease risk", 2006; 35: 315-322

Jacobsen K.H. and Koopman J.S. "The effects of socioeconomic development on worldwide hepatitis A virus seroprevalence patterns", 2005; 34: 600-609

Jenum A.K., Stensvold I., and Thelle D.G. "Differences in cardiovascular disease mortality and major risk factors between districts in Oslo. An ecological analysis", 2001; 30: S59

Kerr-Pontes L., Montenegro A., Barreto M., *et al.* "Inequality and leprosy in northeast Brazil: an ecological study", 2004; 33: 262-269

Marshall S.W., Waller A.E., Dick R.W., *et al* "An ecologic study of protective equipment and injury in two contact sports", 2002; 31: 587-592

Moore S. "Peripherality, income inequality, and life expectancy: revisiting the income inequality hypothesis", 2006; 35: 623-632

Nagata C. "Ecologic study of the association between soy product intake and mortality from cancer and heart disease in Japan", 2000; 29: 832-836

Owen C.G., Whincup P.H., and Cook D.G. "Are early life factors responsible for international differences in adult blood pressure? An ecological study", 2005; 34: 649-654

Parslow R.C., McKinney P.A., Law G.R., *et al.* "Population mixing and childhood diabetes", 2001; 30: 533-538

Pearce J., Witten K., Hiscock R., *et al.* "Are socially disadvantaged neighbourhoods deprived of health-related community resources?", 2007; 36: 348-355

Salah A.B., Kamarianakis Y., Chlif S., *et al.* "Zoonotic cutaneous leishmaniasis in central Tunisia: spatio-temporal dynamics", 2007; 36: 991-1000

Shakhtarin V.V., Tsyb A.F., Stepanenko V.F., *et al.* "Iodine deficiency, radiation dose, and the risk of thyroid cancer among children and adolescents in the Bryansk region of Russia following the Chernobyl power station accident", 2003; 32: 584-591

Stewart A.W., Mitchell E.A., Pearce N., *et al.*. "The relationship of per capita gross national product to the prevalence of symptoms of asthma and other atopic diseases in children", 2001; 30: 173-179

Turrell G. and Mathers C. "Socioeconomic inequalities in all-cause and specific-cause mortality in Australia: 1985-1987 and 1995-1997", 2001; 30: 231-239

Zhang J., Temme E.H.M., and Kesteloot H. "Fish consumption is inversely associated with male lung cancer mortality in countries with high levels of cigarette smoking or animal fat consumption", 2000; 29: 615-621

Journal of Clinical Epidemiology

Amiri M., Kunst A.E., Janssen F., *et al.* "Cohort-specific trends in stroke mortality in seven European countries were related to infant mortality rates", 2006; 59 (12): 1295-1302

Dummer T.J.B., Dickinson H.O., Pearce M.S., *et al.* "Stillbirth risk with social class and deprivation: no evidence for increasing inequality", 2000; 53 (2): 147-155

Perez A., Dennis R.J., Rondon M.A., *et al.* "A Colombian survey found intensive care mortality ratios were better in private vs. public hospitals", 2006; 59 (1): 94-101

Schober E., Rami B., and Waldhoer T. "Small area variations in childhood diabetes mellitus in Austria: links to population density, 1989 to 1999", 2003; 56 (3): 269-273

Journal of Epidemiology and Community Health

Ayotte J.D., Baris D., Cantor K.P., *et al.* "Bladder cancer mortality and private well use in New England: an ecological study", 2006; 60 (2): 168-172

Baker M., Das D., Venugopal K., *et al.* "Tuberculosis associated with household crowding in a developed country", 2008; 62 (8): 715-721

Benach J., Yasui Y., and Borrell C. "Material deprivation and leading causes of death by gender: evidence from a nationwide small area study", 2001; 55 (4): 239-245

Bobak M., Murphy M., Rose R., *et al.* "Societal characteristics and health in the former communist countries of central and eastern Europe and the former Soviet Union: a multilevel analysis", 2007; 61 (11): 990-996

Boer J.W.D., Coutinho R.A., Yzerman E.P.F., *et al.* "Use of surface water in drinking water production associated with municipal Legionnaire's disease incidence", 2008; 62 (4): e1

Che D., Decludt B., Campese C., *et al.* "Sporadic cases of community acquired Legionnaires' disease: an ecologic study to identify new sources of contamination", 2003; 57 (6): 466-469

Connolly V., Unwin N., Sherriff P., *et al.* "Diabetes prevalence and socioeconomic status: a population based study showing increased prevalence of type 2 diabetes mellitus in deprived areas", 2000; 54 (3): 173-177

Corcoran P., Arensman E., and Perry I.J. "The area-level association between hospital-treated deliberate self-harm, deprivation, and social fragmentation in Ireland", 2007; 61 (12): 1050-1055

Doran T., Drever F., and Whitehead M. "Health underachievement and overachievement in English local authorities", 2006; 60 (8): 686-693

Dunn et al. "Income distribution, public services expenditures, and all cause mortality in US states", 2005; 59 (9): 768

Elliott M.R., Wang Y., Lowe R.A., *et al.* "Environmental justice: frequency and severity of US chemical industry accidents and the socioeconomic status of surrounding communities", 2004; 58 (1): 24-30

Exeter D.J. and Boyle P.J. "Does young adult suicide cluster geographically in Scotland?", 2007; 61 (8): 731-736

Exeter D.J., Feng Z., and Flowerdew R. "Shrinking areas and mortality: an artefact of deprivation effects?", 2005; 59 (11): 924-926

Fabio A., Li W., Strotmeyer S, *et al.* "Racial segregation and county level intentional injury in Pennsylvania: analysis of hospital discharge data for 1997-1999", 2004; 58 (4): 346-351

Gutierrez-Fisac J.L., Gispert R., and Sola J. "Factors explaining the geographical differences in disability free life expectancy in Spain", 2000; 54 (6): 451-455

Hales S., Black W., Skelly C., *et al.* "Social deprivation and the public health risks of community drinking water supplies in New Zealand", 2003; 57 (8): 581-583

Herttua K., Mkel P., Martikainen P., and Siren R. "The impact of a large reduction in the price of alcohol on area differences in interpersonal violence: a natural experiment based on aggregate data", 2008; 62 (11): 995-1001

van Hooijdonk C., Droomers M., Deerenberg I.M., *et al.* "Higher mortality in urban neighbourhoods in The Netherlands: who is at risk?", 2008; 62 (6): 499-505

Jerrett M., Eyles J., Dufournaud C., *et al.* "Environmental influences on healthcare expenditures: an exploratory analysis from Ontario, Canada", 2003; 57 (5): 334-338

Kelleher et al. "Indicators of deprivation, voting patterns, and health status at area level in the Republic of Ireland", 2002; 56 (1): 36

Knox E.G. "Atmospheric pollutants and mortalities in English local authority areas", 2008; 62 : 442-447

Kopp M.S., Skrabski A., Kawachi I., *et al.* "Low socioeconomic status of the opposite sex is a risk factor for middle aged mortality", 2005; 59 : 675-687

Kousa A., Moltchanova E., Viik-Kajander M., *et al.* "Geochemistry of ground water and the incidence of acute myocardial infarction in Finland", 2004; 58 (2): 136-139

Laaser U., Breekencamp J., Ulrich A., and Hoffmann B. "Can a decline in the population means of cardiovascular risk factors reduce the number of people at risk?", 2001; 55 (3): 179-184

Laurent O., Filleul L., Havard S., Deguen S., Declercq C., and Bard D. "Asthma attacks and deprivation: gradients in use of mobile emergency medical services", 2008; 62 (11): 1014-1016

Lawlor D.A., Maxwell R., and Wheeler B.W. "Rurality, deprivation, and excess winter mortality: an ecological study", 2002; 56 (5): 373-374

Lobmayer P. and Wilkinson R.G. "Inequality, residential segregation by incomes and mortality in US cities", 2002; 56 (3): 183-187

Lovasi G.S., Quinn J.W., Neckerman K.M., *et al.* "Children living in areas with more street trees have lower prevalence of asthma", 2008; 62 (7): 647-649

Maheswaran R., Payne N., Meechan D., *et al.* "Socioeconomic deprivation, travel distance, and renal replacement treatment in the Trent region, United Kingdom 2000: an ecological study", 2003, 57 (7): 523-524

Manjer J., Berglund G., Bondesson L., *et al.* "Intra-urban differences in breast cancer mortality: a study from the city of Malmo in Sweden", 2000; 54 (4): 279-285

McLean G., Sutton M., and Guthrie B. "Deprivation and quality of primary care services: evidence for persistence of the inverse care law from the UK Quality and Outcomes Framework", 2006; 60 (11): 917-922

Merlo J., Ostergren P-O, Hagberg O., *et al.* "Diastolic blood pressure and area of residence: multilevel vs ecological analysis of social inequality", 2001; 55 (11): 791-798

Middleton N., Sterne J.A.C., and Gunnell D. "The geography of despair among 15-44-year-old men in England and Wales: putting suicide on the map", 2006; 60 (12): 1040-1047

Mitchell R. and Popham F. "Greenspace, urbanity, and health: relationships in England", 2007; 61 : 681-683

Mohr S.B., Garland C.F., Gorham E.D., *et al.* "Could ultraviolet B irradience and vitamin D be associated with lower incidence rates of lung cancer?", 2008; 62 (1): 69-74

Moulton L.H. and Benini A.A. "Community-level risk factors for numbers of landmine victims in Chad and Thailand", 2003; 57 (12): 956-959

Ocana-Riola R., Saurina C., Fernandez-Ajuria A., *et al.* "Area deprivation and mortality in the provincial capital cities of Andalusia and Catalonia (Spain)", 2008; 62 (2): 147-152

Pickett K.E., Kelly S., and Brunner E. "Wider income gaps, wider waistbands? An ecological study of obesity and income inequality", 2005; 59 : 670-674

Rezaeian M., Dunn G., St Leger S., *et al* "Ecological associations between suicide rates and indices of deprivation in the north west regions of England: the importance of the size of the administrative unit", 2006; 60 (11): 956-961

Rosano A., Botto L.D., Botting B., *et al.* "Infant mortality and congenital anomalies from 1990 to 1994: an international perspective", 2000; 54 (9): 660-666

Ruger P. and Kim H-J. "Global health inequalities: an international comparison", 2006; 60 (11): 928-936

Sanmartin C., Ross N.A., Tremblay S., *et al* "Labour market income inequality and mortality in North American metropolitan areas", 2003; 57 (10): 792-797

Setel P., Whiting D., Hemed Y., *et al.* "Educational status is related to mortality at the community level in the three areas of Tanzania, 1992-1998", 2000; 54 (12): 936-937

Shi L., Macinko J., Starfield B., *et al.* "Primary care, infant mortality, and low birth weight in the states of the USA", 2004; 58 (5): 374-380

Skrabski A., Kopp M., and Kawachi I. "Social capital in a changing society: cross sectional associations with middle aged female and male mortality rates", 2003; 57 (2): 114-119

Skrabski A., Kopp M., and Kawachi I. "Social capital and collective efficacy in Hungary: cross sectional associations with middle aged female and male mortality rates", 2004; 58 (4): 340-345

Smits J., Westert G.P., and van den Bos G. "Socioeconomic status of very small areas and stroke incidence in the Netherlands", 2002; 56 (8): 637-640

Stanistreet D., Bambra C., and Scott-Samuel A. "Is patriarchy the source of men's higher mortality?", 2005; 59 (10): 873-876

StonehamM., Goldacre M., Seagroatt V., *et al.* "Olive oil, diet and colorectal cancer: an ecological study and a hypothesis", 2000; 54 (10): 756-760

Tunstall H., Mitchell R., Gibbs J., *et al.* "Is economic adversity always a killer? Disadvantaged areas with relatively low mortality rates", 2007; 61 (4): 337-343

de Vogli R., Mistry R., Gnesotto R, *et al.* "Has the relation between income inequality and life expectancy disappeared? Evidence from Italy and top industrialized countries", 2005; 59 (2): 158-162

Wood R., Sutton M., Clark D., *et al.* "Measuring inequalities in health: the case for healthy life expectancy", 2006; 60 (12): 1089-1092

Woods L.M., Rachet B., Riga M., *et al.* "Geographical variation in the life expectancy of birth in England and Wales is largely explained by deprivation", 2005; 59 (2): 115-120

Young F.W. "An explanation of the persistent doctor-mortality association", 2001; 55 (2): 80-84

Appendix B: Ecological Regression Coefficient

This Appendix proves the relationship between the ecological regression coefficient $\hat{\beta}_b$, the average within-group regression coefficient $\hat{\beta}_w$, and the total regression coefficient $\hat{\beta}_t$ that was described in Section 3.2.1. To do so, we use the notation defined in Section 3.1.

Since $S_{xxt} = S_{xxw} + S_{xxb}$ and since the total sum of products can be expressed as the sum of S_{xyw} and S_{xyb}, this relationship can be demonstrated fairly easily:

$$\begin{aligned}
\hat{\beta}_w + C_{XA}^2(\hat{\beta}_b - \hat{\beta}_w) &= \frac{S_{xyw}}{S_{xxw}} + (\frac{S_{xxb}}{S_{xxt}})(\frac{S_{xyb}}{S_{xxw}} - \frac{S_{xyw}}{S_{xxw}}) \\
&= \frac{S_{xyw}}{S_{xxw}} + \frac{S_{xxb}S_{xyb}}{S_{xxt}S_{xxb}} - \frac{S_{xxb}S_{xyw}}{S_{xxt}S_{xxw}} \\
&= \frac{S_{xyw}(S_{xxt} - S_{xxb})}{S_{xxt}S_{xxw}} + \frac{S_{xxb}S_{xyb}}{S_{xxT}S_{xxb}} \\
&= \frac{S_{xyw}S_{xxw}}{S_{xxt}S_{xxw}} + \frac{S_{xxb}S_{xyb}}{S_{xxt}S_{xxb}} \\
&= \frac{S_{xyw}}{S_{xxt}} + \frac{S_{xyb}}{S_{xxt}} \\
&= \frac{S_{xyw} + S_{xyb}}{S_{xxt}} \\
&= \hat{\beta}_t
\end{aligned}$$

Appendix C: Simulation Code

```
title 'SIMULATIONS OF ECOLOGICAL REGRESSION';
title2 'OLS, Weighted OLS, Poisson, and NB';
options ls=80 ps=80;
options mprint;

%macro ecological(num_runs, overdispersion, k, Beta_0, Beta_1,
Lower_pop, Upper_pop, sample_size, alpha1, alpha2);

%global PoissonCnvrgCount NBCnvrgCount Lowerpop Upperpop Over_dispersion
        Beta0 Beta1 LinearTypeIRate LinearTypeIIRate WeightedTypeIRate
        WeightedTypeIIRate PoissonTypeIRate PoissonTypeIIRate
        PoissonLR_TypeIRate PoissonLR_TypeIIRate
  NBTypeIRate NBTypeIIRate NBLR_TypeI_Rate NBLR_TypeII_Rate
  AvgDispersionEstimate SampleSize Number_runs
  AverageLinearStdErr AverageWeightedStdErr AveragePoissonStdErr
  AverageNBStdErr AverageLinearBeta_Est AverageWeightedBeta_Est
  AveragePoissonBeta_Est AverageNBBeta_Est Linear_RelBias
  Weighted_RelBias Poisson_RelBias NB_RelBias;

%let SampleSize = &sample_size;
%let Number_runs = &num_runs;
%let Beta0 = &Beta_0; %let Beta1 = &Beta_1;
%let Over_dispersion = &overdispersion; %let Lowerpop = &Lower_pop;
%let Upperpop = &Upper_pop;
%let LinearRegAlpha=0; %let LinearRegBeta=0; %let LinearStdErr = 0;
%let LinearBeta_Est = 0;
%let WeightedRegAlpha = 0; %let WeightedRegBeta = 0;
%let WeightedStdErr = 0; %let WeightedBeta_Est = 0;
%let PoissonRegAlpha = 0; %let PoissonRegBeta = 0;
%let PoissonStdErr = 0; %let PoissonCnvrgCount = 0;
%let PoissonBeta_est = 0; %let PoissonLR_ChiSq = 0; %let NB_LR_ChiSq = 0;
%let NB_LR_RegAlpha = 0; %let NB_LR_RegBeta = 0;
%let PoissonLR_RegBeta = 0; %let PoissonLR_RegAlpha = 0;
%let NBRegAlpha = 0; %let NBRegBeta = 0; %let NBStdErr = 0;
%let NBCnvrgCount = 0; %let NBBeta_Est = 0; %let Dispersion_Est = 0;
%let SumLinearStdErr = 0; %let SumWeightedStdErr = 0;
%let SumPoissonStdErr = 0; %let SumNBStdErr = 0;
```

```
%let SumLinearBeta_Est = 0; %let SumWeightedBeta_Est = 0;
%let SumPoissonBeta_Est = 0; %let SumNBBeta_Est = 0;
%let SumDispersion = 0;


%do i=1 %to &num_runs;

%if &overdispersion=1 %then %do;
  data NegBinomial;
   do i=1 To &sample_size;
    prop_exposed = rand('beta', &alpha1, &alpha2);
    pop_size = &Lower_pop + (&Upper_pop - &Lower_pop)*ranuni(0);
    Lpop_size=log(pop_size);
    lambda = exp(log(pop_size) + &Beta_0 + &Beta_1*prop_exposed);
    p=&k/(lambda + &k);
    Y = rand('negbinomial',p,&k);
    Event_rate = Y/pop_size;
    output;
    keep pop_size prop_exposed Event_rate Y Lpop_size;
   end;
%end;

%if &overdispersion=0 %then %do;
  data NegBinomial;
   do i=1 To &sample_size;
    prop_exposed = rand('beta', &alpha1, &alpha2);
    pop_size = &Lower_pop + (&Upper_pop - &Lower_pop)*ranuni(0);
    Lpop_size=log(pop_size);
    lambda = exp(log(pop_size) + &Beta_0 + &Beta_1*prop_exposed);
    Y = rand('poisson',lambda);
    Event_rate = Y/pop_size;
    output;
    keep pop_size prop_exposed Event_rate Y Lpop_size;
   end;
%end;


/* Running Linear Regression */
proc reg data=NegBinomial;
 model Event_rate = prop_exposed;
 ods output ParameterEstimates = par_est; /* getting info */
```

```
run;
data LinearReg;
 set par_est;
 if Variable = "prop_exposed" then do;
  call symputx('LinearBeta_Est', Estimate);
  call symputx('LinearProbt', put(probt, 7.5));
  call symputx('LinearStdErr',StdErr);
 end;
run;

/* Calculating type I Error for linear regression */
%if &Beta_1 = 0 AND &LinearProbt < 0.05 %then %do;
  %let LinearRegAlpha = %eval(&LinearRegAlpha + 1);
%end;
/* Calculating type II Error for linear regression */
%if &Beta_1 ne 0 AND &LinearProbt > 0.05 %then %do;
  %let LinearRegBeta = %eval(&LinearRegBeta + 1);
%end;
/* Adding up standard errors. */
%let SumLinearStdErr = %sysevalf(&SumLinearStdErr + &LinearStdErr);
/* Adding up estimates of Beta_1 */
%let SumLinearBeta_Est = %sysevalf(&SumLinearBeta_Est + &LinearBeta_Est);


/* Running Weighted Linear regression */
proc reg data=NegBinomial;
 model Event_rate = prop_exposed;
 weight pop_size;
 ods output ParameterEstimates = Weighted_par_est;
run;
data WeightedReg;
 set Weighted_par_est;
 if Variable = "prop_exposed" then do;
  call symputx('WeightedBeta_Est', Estimate);
  call symputx('WeightedProbt', put(probt, 7.5));
  call symputx('WeightedStdErr', StdErr);
 end;
run;

/*Calculating type I error for weighted regression */
%if &Beta_1 = 0 AND &WeightedProbt < 0.05 %then %do;
```

```
  %let WeightedRegAlpha = %eval(&WeightedRegAlpha+1);
%end;
/*Calculating type II error for weighted regression */
%if &Beta_1 ne 0 AND &WeightedProbt > 0.05 %then %do;
  %let WeightedRegBeta = %eval(&WeightedRegBeta+1);
%end;
/*Adding up standard errors. */
%let SumWeightedStdErr =
                        %sysevalf(&SumWeightedStdErr + &WeightedStdErr);
/*Adding up Beta_1 estimates */
%let SumWeightedBeta_Est =
                        %sysevalf(&SumWeightedBeta_Est + &WeightedBeta_Est);


/* Poisson Regression */
proc genmod data=NegBinomial;
 model Y = prop_exposed / offset=Lpop_size dist=poisson link=log type3;
 ods output ParameterEstimates = Poisson_par_est;
 ods output Type3 = Poisson_LR;
 ods output ConvergenceStatus = Poisson_convergence;
run;
data PoissonReg;
 set Poisson_par_est;
 if parameter = "prop_exposed" then do;
  call symputx('PoissonBeta_Est' Estimate);
  call symputx('PoissonChiSq', put(ProbChiSq, 7.5));
  call symputx('PoissonStdErr', StdErr);
 end;
run;
data PoissonLR;
 set Poisson_LR;
 call symputx('PoissonLR_ChiSq', put(ProbChiSq, 7.5));
run;
data PoissonConvrg;
 set Poisson_convergence;
 call symputx('PoissonStatus', Status);
run;

/* Calculating type I error for Poisson regression */
%if &Beta_1 = 0 AND &PoissonChiSq < 0.05 %then %do;
  %let PoissonRegAlpha = %eval(&PoissonRegAlpha+1);
```

```
%end;
/* Calculating type II error for Poisson regression */
%if &Beta_1 ne 0 AND &PoissonChiSq > 0.05 %then %do;
  %let PoissonRegBeta = %eval(&PoissonRegBeta+1);
%end;
/* Calculating type I error for Poisson LR test */
%if &Beta_1 = 0 AND &PoissonLR_ChiSq < 0.05 %then %do;
  %let PoissonLR_RegAlpha = %eval(&PoissonLR_RegAlpha+1);
%end;
/* Calculating type II error for Poisson LR test */
%if &Beta_1 ne 0 AND &PoissonLR_ChiSq > 0.05 %then %do;
  %let PoissonLR_RegBeta = %eval(&PoissonLR_RegBeta+1);
%end;
/* Adding up standard errors */
%let SumPoissonStdErr =
                    %sysevalf(&SumPoissonStdErr + &PoissonStdErr);
/* Adding up estimated values of Beta_1 */
%let SumPoissonBeta_Est =
                    %sysevalf(&SumPoissonBeta_Est + &PoissonBeta_Est);
/* Counting number of times Poisson regression did not converge */
%if &PoissonStatus ne 0 %then %do;
  %let PoissonCnvrgCount = %sysevalf(&PoissonCnvrgCount + 1);
%end;


/* Negative Binomial Regression */
proc genmod data=NegBinomial;
 model Y = prop_exposed / offset=Lpop_size dist=Negbin link=Log type3;
 ods output ParameterEstimates = NB_par_est;
 ods output Type3 = NBLikelihood;
 ods output ConvergenceStatus = NB_convergence;
run;
data NBReg;
 set NB_par_est;
 if parameter = "prop_exposed" then do;
  call symputx('NBBeta_Est', Estimate);
  call symputx('NBChiSq', put(ProbChiSq, 7.5));
  call symputx('NBStdErr', StdErr);
 end;
 if parameter = "Dispersion" then do;
  call symputx('Dispersion_Est', Estimate);
```

```
 end;
run;
data NBLR;
 set NBLikelihood;
 call symputx('NB_LR_ChiSq', put(ProbChiSq, 7.5));
run;
data NBConvrg;
 set NB_convergence;
 call symputx('NBStatus', Status);
run;


/* Discarding non-convergent negative binomial data: */
%if &NBStatus=0 %then %do;

/* Calculating type I error for negative binomial regression */
%if &Beta_1 = 0 AND &NBChiSq < 0.05 %then %do;
  %let NBRegAlpha = %eval(&NBRegAlpha + 1);
%end;
/* Calculating type II error for negative binomial regression */
%if &Beta_1 ne 0 AND &NBChiSq > 0.05 %then %do;
  %let NBRegBeta = %eval(&NBRegBeta + 1);
%end;
/* Calculating type I error for negative binomial LR test */
%if &Beta_1 = 0 AND &NB_LR_ChiSq < 0.05 %then %do;
  %let NB_LR_RegAlpha = %eval(&NB_LR_RegAlpha+1);
%end;
/* Calculating type II error for negative binomial LR test */
%if &Beta_1 ne 0 AND &NB_LR_ChiSq > 0.05 %then %do;
  %let NB_LR_RegBeta = %eval(&NB_LR_RegBeta+1);
%end;
/* Adding up standard errors */
%let SumNBStdErr = %sysevalf(&SumNBStdErr + &NBStdErr);
/* Adding up estimated values of Beta_1 */
%let SumNBBeta_Est = %sysevalf(&SumNBBeta_Est + &NBBeta_Est);

%end;

/* Adding up dispersion parameter estimates */
%let SumDispersion = %sysevalf(&SumDispersion + &Dispersion_Est);
/* Counting number of times NB regression did not converge */
```

```
%if &NBStatus ne 0 %then %do;
  %let NBCnvrgCount = %sysevalf(&NBCnvrgCount + 1);
%end;

proc datasets;
delete NegBinomial LinearReg WeightedReg PoissonReg NBReg;

%end; /* end of &num_runs loop */

/* Calculating summary statistics here */
%let LinearTypeIRate = %sysevalf(&LinearRegAlpha/&num_runs);
%let LinearTypeIIRate = %sysevalf(&LinearRegBeta/&num_runs);
%let AverageLinearBeta_Est = %sysevalf(&SumLinearBeta_Est/&num_runs);
%let AverageLinearStdErr = %sysevalf(&SumLinearStdErr/&num_runs);
%let WeightedTypeIRate = %sysevalf(&WeightedRegAlpha/&num_runs);
%let WeightedTypeIIRate = %sysevalf(&WeightedRegBeta/&num_runs);
%let AverageWeightedBeta_Est = %sysevalf(&SumWeightedBeta_est/&num_runs);
%let AverageWeightedStdErr = %sysevalf(&SumWeightedStdErr/&num_runs);
%let PoissonTypeIRate = %sysevalf(&PoissonRegAlpha/&num_runs);
%let PoissonTypeIIRate = %sysevalf(&PoissonRegBeta/&num_runs);
%let PoissonLR_TypeIRate = %sysevalf(&PoissonLR_RegAlpha/&num_runs);
%let PoissonLR_TypeIIRate = %sysevalf(&PoissonLR_RegBeta/&num_runs);
%let AveragePoissonBeta_Est = %sysevalf(&SumPoissonBeta_Est/&num_runs);
%let AveragePoissonStdErr = %sysevalf(&SumPoissonStdErr/&num_runs);
%let NBTypeIRate = %sysevalf(&NBRegAlpha/(&num_runs - &NBCnvrgCount));
%let NBTypeIIRate = %sysevalf(&NBRegBeta/(&num_runs - &NBCnvrgCount));
%let NBLR_TypeI_Rate = %sysevalf(&NB_LR_RegAlpha/(&num_runs - &NBCnvrgCount));
%let NBLR_TypeII_Rate = %sysevalf(&NB_LR_RegBeta/(&num_runs - &NBCnvrgCount));
%let AverageNBBeta_Est = %sysevalf(&SumNBBeta_Est/(&num_runs - &NBCnvrgCount));
%let AverageNBStdErr = %sysevalf(&SumNBStdErr/(&num_runs - &NBCnvrgCount));
%let AvgDispersionEstimate = %sysevalf(&SumDispersion/&num_runs);

%if &Beta_1 ne 0 %then %do;
  %let Linear_RelBias =
              %sysevalf((&AverageLinearBeta_Est - &Beta_1)/&Beta_1);
  %let Weighted_RelBias =
              %sysevalf((&AverageWeightedBeta_Est - &Beta_1)/&Beta_1);
  %let Poisson_RelBias =
              %sysevalf((&AveragePoissonBeta_Est - &Beta_1)/&Beta_1);
  %let NB_RelBias =
              %sysevalf((&AverageNBBeta_Est - &Beta_1)/&Beta_1);
```

```
%end;

%if &Beta_1 = 0 %then %do;
 %let Linear_RelBias = &AverageLinearBeta_Est;
 %let Weighted_RelBias = &AverageWeightedBeta_Est;
 %let Poisson_RelBias = &AveragePoissonBeta_Est;
 %let NB_RelBias = &AverageNBBeta_Est;
%end;

%mend ecological; /* end of macro */



/* now printing simulation results: */

proc printto print='/dev/null' log='/dev/null';
%ecological(1000, 0, 20, -3, 0, 2500, 2500, 10, 2, 18);

proc printto print='simulation_output.lst1' log='/dev/null';
data simulation_results;

 Overdispersion = &Over_dispersion;
 Cluster_lower_limit = &Lowerpop;
 Cluster_upper_limit = &Upperpop;
 Beta0 = &Beta0;
 Beta1 = &Beta1;
 Number_of_runs = &Number_runs;
 SampleSize = &SampleSize;

 Linear_TypeI_Rate = &LinearTypeIRate;
 Linear_TypeII_Rate = &LinearTypeIIRate;
 AvgLinear_StdErr = &AverageLinearStdErr;
 AvgLinear_Est = &AverageLinearBeta_Est;
 Linear_Relative_Bias = &Linear_RelBias;

 Weighted_Type_I_Rate = &WeightedTypeIRate;
 Weighted_Type_II_Rate = &WeightedTypeIIRate;
 AvgWeighted_Est = &AverageWeightedBeta_Est;
 AvgWeighted_StdErr = &AverageWeightedStdErr;
 Weighted_Relative_Bias = &Weighted_RelBias;
```

```
 Poisson_Type_I_Rate = &PoissonTypeIRate;
 Poisson_Type_II_Rate = &PoissonTypeIIRate;
 PoissonLR_TypeI_Rate = &PoissonLR_TypeIRate;
 PoissonLR_TypeII_Rate = &PoissonLR_TypeIIRate;
 AvgPoisson_Est = &AveragePoissonBeta_Est;
 AvgPoisson_StdErr = &AveragePoissonStdErr;
 Poisson_Relative_Bias = &Poisson_RelBias;

 NB_Type_I_Rate = &NBTypeIRate;
 NB_Type_II_Rate = &NBTypeIIRate;
 NBLR_TypeI_Rate = &NBLR_TypeI_Rate;
 NBLR_TypeII_Rate = &NBLR_TypeII_Rate;
 AvgNB_Est = &AverageNBBeta_Est;
 AvgNB_StdErr = &AverageNBStdErr;
 NB_Relative_Bias = &NB_RelBias;

 Avg_Dispersion = &AvgDispersionEstimate;
 NumTimesPoisson_no_cnvrg = &PoissonCnvrgCount;
 NumTimesNB_no_cnvrg = &NBCnvrgCount;

proc print data = simulation_results;
run; run; quit;
```