J Happiness Stud (2008) 9:197–205
DOI 10.1007/s10902-007-9044-7

RESEARCH PAPER

Does subjective well-being show a relationship between parents and their children?

Ferran Casas · Germà Coenders · Robert A. Cummins · Mònica González · Cristina Figuer · Sara Malo

Published online: 14 February 2007

Abstract The relationship between the subjective well-being of parents and their own 12–16-year-old children was explored in a Spanish sample of $N = 266$ families. A positive relationship was expected due to both a shared environment and the possibility of the genetic transmission of subjective well-being 'set-points'. A positive significant relationship was found for the summated scale of satisfaction domains forming the Personal Well-being Index, and for the specific domains of health and security for the future. However, no relationship was found for the other five domains that make up this Index or for satisfaction with life as a whole. We conclude while these results provide some evidence for the expected influence of a shared environment, they have failed to provide evidence for high heritability of set-points for subjective well-being.

Keywords Adolescents · Children · Well-being · Life satisfaction · Life domain satisfaction · Family

1 Introduction

In the socialization process that happens within families, there is no doubt that parents have an important influence on many aspects of their children's lives. Thus, it might be expected that the combination of such socialization together with shared genetic influences would cause children to resemble their parents in terms of their attitudes, beliefs, routines and values. Previous studies have tended to confirm this

F. Casas (✉) · G. Coenders · M. González · C. Figuer · S. Malo
Facultat de Ciències Econòmiques i Empresarials, Institut de Recerca sobre Qualitat de Vida (IRQV), Universitat de Girona, Campus Montilivi, Girona 17071, Spain
e-mail: ferran.casas@udg.es

R. A. Cummins
Australian Center on Quality of Life, Deakin University, Melbourne, VIC, Australia
e-mail: robert.cummins@deakin.edu.au

Appl. Spatial Analysis (2010) 3:45–65
DOI 10.1007/s12061-009-9030-z

Spatial Modelling of Car Ownership Data: A Case Study from the United Kingdom

Stephen Clark · Andrew O. Finley

Received: 4 September 2008 / Accepted: 8 July 2009 /
Published online: 23 July 2009

Abstract In this paper a model is formulated to estimate the strength of the relationship between household car ownership and income using cross-sectional data. Whilst reports of such studies are not uncommon in the transport literature, this study is different in that it takes explicit account of the spatial distribution of the data. By incorporating this spatial element in the model formulation, the residual errors in the model are uncorrelated and hence allows for the estimation of parameters that are, in a statistical sense, the best available. These spatial models are fitted to a large data set provided by the United Kingdom Office for National Statistics, covering the area of England and Wales. The recommended model form is a Hierarchical Bayesian spatial regression model with the parameters in the model estimated using the technique of Markov Chain Monte Carlo (MCMC). A common feature of all the spatial models is that the estimate of the elasticity of car ownership with respect to income is seen to be larger than that from a non-spatial model.

Keywords Car ownership · Bayesian inference · Markov Chain Monte Carlo · Spatial process

S. Clark (✉)
Institute for Transport Studies, University of Leeds, Leeds, LS2 9JT, UK
e-mail: s.d.clark@its.leeds.ac.uk

A. O. Finley
Department of Geography, Michigan State University, East Lansing, MI, USA
e-mail: finleya@msu.edu

Journal of Statistical Planning and Inference 142 (2012) 3242–3253

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Optimal predictive design augmentation for spatial generalised linear mixed models

Evangelos Evangelou [a,*], Zhengyuan Zhu [b,1]

[a] Department of Mathematical Sciences – University of Bath, Bath, UK
[b] Department of Statistics – Iowa State University, Ames, Iowa, USA

ARTICLE INFO

Article history:
Received 22 October 2011
Received in revised form
21 May 2012
Accepted 24 May 2012
Available online 1 June 2012

Keywords:
Generalised linear mixed models
Geostatistics
Predictive inference
Sampling design

ABSTRACT

A typical model for geostatistical data when the observations are counts is the spatial generalised linear mixed model. We present a criterion for optimal sampling design under this framework which aims to minimise the error in the prediction of the underlying spatial random effects. The proposed criterion is derived by performing an asymptotic expansion to the conditional prediction variance. We argue that the mean of the spatial process needs to be taken into account in the construction of the predictive design, which we demonstrate through a simulation study where we compare the proposed criterion against the widely used space-filling design. Furthermore, our results are applied to the Norway precipitation data and the rhizoctonia disease data.

1. Introduction

One of the most frequently used models for the analysis of geostatistical count data is the spatial generalised linear mixed model (SGLMM) (Diggle et al., 1998). Applications of SGLMM include Diggle et al. (1998) who looked into residual contamination from nuclear weapons testing and campylobacter infections in UK, Diggle et al. (2002) who studied the risk of malaria in Gambia, Zhang (2002) who analysed a root disease in an agricultural study, and Eidsvik et al. (2009) who examined precipitation data for the purpose of weather forecasting and for operating hydropower plants. This class of models assumes the existence of an unobserved Gaussian random field over the region of interest and that the observations, drawn at fixed locations, are conditionally independent given the value of the random field. The distribution of the random field may depend on unknown parameters and among the objectives is to use the sample to predict the value of the random field at every location in the region. The "plug-in" approach is a common method for prediction in these models from a frequentist point of view (Christensen, 2004; Evangelou et al., 2011) where in the first stage an estimate of the model parameters is obtained and in the second stage the predictive distribution of the random field is constructed conditional on the data and the parameter estimates. (Alternatively see Diggle et al., 1998; Christensen and Waagepetersen, 2002; Eidsvik et al., 2009 for a Bayesian solution.)

The objective of *spatial predictive design* (Zimmerman, 2006; Zhu and Stein, 2006) is to select the sampled locations within the region of interest in order to optimise, in some sense, the predictive capability of the sample. In summary, the strategy of optimal design comes down to developing some optimality criterion, such as the average prediction variance, and then searching over all possible sampling configurations for the optimal value of the criterion. We focus on the case

* Corresponding author. Tel.: +1 44 1225385673
E-mail address: ee224@bath.ac.uk (E. Evangelou).
[1] Partially supported by NSF DMS 0605434 grant.

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.05.008

Statistical Methods & Applications (2004) 13: 327–340
DOI: 10.1007/s10260-004-0082-6

SMA

Small area estimation for longitudinal surveys*

Maria Rosaria Ferrante, Silvia Pacei

Dipartimento di Statistica, Università di Bologna, Via Belle Arti n. 41, 40126 Bologna, Italy
(e-mail: {ferrante;pacei}@stat.unibo.it)

Abstract. Over the last few years many studies have been carried out in Italy to identify reliable small area labour force indicators. Considering the rotated sample design of the Italian Labour Force Survey, the aim of this work is to derive a small area estimator which "borrows strength" from individual temporal correlation, as well as from related areas. Two small area estimators are derived as extensions of an estimation strategies proposed by Fuller (1990) for partial overlap samples. A simulation study is carried out to evaluate the gain in efficiency provided by our solutions. Results obtained for different levels of autocorrelation between repeated measurements on the same outcome and different population settings show that these estimators are always more reliable than the traditional composite one, and in some circumstances they are extremely advantageous.

Key words: Small area estimators, rotation sampling, temporal correlation, local labour force indicators

1. Introduction

Statistical Agencies are often required to provide reliable estimates of parameters referring to the labour market in local areas. Nevertheless, the sample size of the official surveys within those areas is usually too small to obtain efficient results using direct survey estimators. Many studies dealing with the small area estimation problem have been discussed in the literature (Ghosh and Rao, 1994; Rao, 1999). In Italy, many research project have been carried out to identify a suitable estimation strategy for small area labour force indicators, based on the Italian Labour Force

* The present paper is financially supported by Murst-Cofin (2001) "L'utilizzo di informazioni di tipo amministrativo nella stima per piccole aree e per sottoinsiemi della popolazione" (National Coordinator Prof. Carlo Filippucci).

Biometrics 67, 1518–1531
December 2011
DOI: 10.1111/j.1541-0420.2011.01604.x

Variance Estimation for Systematic Designs in Spatial Surveys

R. M. Fewster

Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand
email: r.fewster@auckland.ac.nz

Summary. In spatial surveys for estimating the density of objects in a survey region, systematic designs will generally yield lower variance than random designs. However, estimating the systematic variance is well known to be a difficult problem. Existing methods tend to overestimate the variance, so although the variance is genuinely reduced, it is over-reported, and the gain from the more efficient design is lost. The current approaches to estimating a systematic variance for spatial surveys are to approximate the systematic design by a random design, or approximate it by a stratified design. Previous work has shown that approximation by a random design can perform very poorly, while approximation by a stratified design is an improvement but can still be severely biased in some situations. We develop a new estimator based on modeling the encounter process over space. The new "striplet" estimator has negligible bias and excellent precision in a wide range of simulation scenarios, including strip-sampling, distance-sampling, and quadrat-sampling surveys, and including populations that are highly trended or have strong aggregation of objects. We apply the new estimator to survey data for the spotted hyena *(Crocuta crocuta)* in the Serengeti National Park, Tanzania, and find that the reported coefficient of variation for estimated density is 20% using approximation by a random design, 17% using approximation by a stratified design, and 11% using the new striplet estimator. This large reduction in reported variance is verified by simulation.

Key words: Distance sampling; Encounter rate; Line transect sampling; Plot sampling; Poststratification; Quadrat sampling; Strip sampling; Systematic sampling; Variance estimation.

1. Introduction

Systematic survey designs are popular in spatial surveys such as strip sampling, quadrat sampling, and distance sampling from lines or points. The aim of these surveys is to estimate density of animals or plants (termed "objects") in a defined region. Systematic designs use a grid of equally spaced samplers—strips, lines, points, or quadrats—with a random start-point. They are easy to plan and implement in the field, and they generally yield lower variance than random designs in which samplers are placed randomly and independently in the survey region. This is because random designs include realizations where several samplers fall by chance into high-density or low-density parts of the region, whereas systematic designs ensure even coverage of the region for all realizations. In many situations, systematic designs are also more precise than stratified designs (Cochran, 1946).

The chief disadvantage of systematic designs is the difficulty of estimating the improved variance. A systematic sample is based on only one random start-point, so the samplers are not independent replicates. Wolter (1984, 1985) highlighted three common approaches to systematic variance estimation for sampling a finite population in social statistics:

1. Random estimation, ignoring the problem of nonindependent samplers and using estimators derived for random designs;
2. Poststratification, approximating the systematic design by a stratified design by grouping small sets of adjacent samplers into strata, and using stratified variance estimators;
3. Modeling the process producing the finite population, for example by proposing a model for the correlation in response between adjacent members of the population.

Similar ideas are used for spatial surveys. Most analyses ignore the problem (approach 1), but there is increasing recognition that this can be misleading. Millar and Olsen (1995), Simmonds and Fryer (1996), Kingsley (2000), and D'Orazio (2003) all used poststratification (approach 2), and Fewster et al. (2009) extended this scheme to provide estimators for strip or line-transect sampling where line lengths are not equal. However, the poststratification scheme is an approximation and does not yield unbiased estimates for the variance.

The aim of this article is to develop a new variance estimator for systematic spatial surveys. We create a model for the systematic variance, similar to approach 3 but exploiting the continuous nature of space. We show how the new variance estimator is applied to strip-sampling, line-transect distance-sampling, and quadrat or point-transect sampling surveys. We assess the estimator through a wide range of simulations, reproducing those in two recent studies in which correct variances were not always obtained (Fewster et al., 2009; Johnson, Laake, and Ver Hoef, 2010). We then apply the estimator to distance-sampling data for spotted hyenas in the Serengeti National Park, Tanzania (Durant et al., in press), and show that the new estimator can make a dramatic impact on the standard error and confidence interval width. This result is verified by further simulations. All computations are

Chapter 2
Individual Versus Ecological Analyses

2.1 Introduction

Analyses of disease maps frequently require the use of an ecological approach, partially because aggregates of cases allow such measures as rates to be computed. In addition, group averages of individual measures often are more readily available, tend to reduce impacts of measurement error, and help to preserve the confidentiality of individuals in each aggregation group. Given this context, the resulting problematic issue concerns drawing sound inferences about individuals from such grouped data. The general drawback to this type of inference is known as the ecological fallacy: most often a difference exists between an ecological regression and the regression based upon individuals underlying it (i.e., aggregate-level relationships do not necessarily hold at the individual level). Well-recognized impacts corrupting inference are aggregation bias (i.e., distortions of the information content of data attributable to loss of variability through observation aggregation), confounding variables (i.e., hidden or unknown variables lurking about in a study that cause distortions through their correlations with the response variable), and nonlinearity. One interesting exchange about this topic appears in the *Annals of the Association of American Geographers* (2000).

In this chapter, results of experiments with Syracuse, NY pediatric lead poisoning data demonstrate selected nonstandard spatial statistical analyses concerning individual versus ecological inference.

2.2 Spatial Autocorrelation Effects

Frequently georeferenced data comprise geographic aggregates, with geographic variability constituting part of the focus of a study. Accordingly, analyses of disease maps are further complicated by the presence of spatial autocorrelation (SA) effects associated with georeferenced data, especially because less is known about impacts of these effects on binomial or Poisson random variables. Generally speaking, variance inflation is the principal impact of positive SA in linear statistical analyses.

D.A. Griffith, J.H.P. Paelinck, *Non-standard Spatial Statistics and Spatial Econometrics*, Advances in Geographic Information Science 1,
DOI 10.1007/978-3-642-16043-1_2, © Springer-Verlag Berlin Heidelberg 2011

Social Science Research 38 (2009) 693–716

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Racial diversity, minority concentration, and trust in Canadian urban neighborhoods

Feng Hou [a,*], Zheng Wu [b]

[a] Business and Labour Market Analysis Division, Analytical Studies Branch, Statistics Canada, Ottawa, Ont., Canada K1A 0T6
[b] Department of Sociology, University of Victoria, Victoria, BC, Canada V8W 3P5

ARTICLE INFO

Article history:
Available online 13 March 2009

Keywords:
Racial diversity
Minority concentration
Trust
Neighborhood effects

ABSTRACT

Using a sample of 42,329 respondents nested within 4254 Canadian urban neighborhoods, this study demonstrates the conceptual and empirical importance of making a distinction between neighborhood racial diversity and minority concentration, and examines how each is uniquely associated with trust. Our analysis shows that at a given level of racial minority concentration, Whites are more trusting when their minority neighbors are more evenly distributed across racial minority groups. Meanwhile, Whites are less trusting as the neighborhood share of racial minorities increases. Overall, the effect of racial minority concentration tends to prevail over that of racial diversity.

1. Introduction

As international migration has become a major means of alleviating the pressure of low fertility and labour shortages, many developed countries face mounting challenges to manage and adjust to rising racial/ethnic diversity. In recent years, across Western societies, there has been a growing preoccupation in public debates and policy initiatives regarding the possible barriers to social cohesion represented by increasing racial and ethnic diversity (Soroka et al., 2007). In the UK, for example, a number of government initiatives and policy documents in the early 2000s were centered on the assumption that limiting the social relevance of racial/ethnic diversity is a key condition for prosperity and strengthening the social fabric in British society (Cheong et al., 2007; Letki, 2008).

In the US, some studies have associated racial/ethnic diversity with a general decline in civic engagement, less efficient public policies, less provision of public goods, lower participation in social activities, and issues of trust across American cities (Alesina et al., 1999; Alesina and La Ferrara, 2000, 2002, 2005; Costa and Kahn, 2003a,b).[1] More specifically, recent empirical research in Canada, the US, UK, and Australia found that trust is negatively associated with racial diversity or minority concentration within neighborhoods (Leigh, 2006; Letki, 2008; Putnam, 2007; Soroka et al., 2006; Stolle et al., 2008). These results have been interpreted to suggest that racial/ethnic diversity reduces levels of trust, at least in the short term.

In this study, we question the interpretations of previous studies on the premise that they did not distinguish neighborhood racial diversity from racial minority concentration. These two constructs are conceptually distinct and should be treated as such. Racial diversity, as it is commonly measured in the trust literature, captures both the variety of racial groups and the spread of population distribution among racial groups within a neighborhood. Higher diversity is observed in places

* Corresponding author. Fax: +1 613 951 5403.
E-mail addresses: feng.hou@statcan.gc.ca (F. Hou), zhengwu@uvic.ca (Z. Wu).

[1] Some studies show that racial divisions seem to have more negative effects than diversity along ethnic ancestries (see Alesina and La Ferrara, 2005 for a review). However, there is evidence suggesting that diversity along culture dimensions such as language and lifestyle is linked to better amenities, higher productivity, and more innovations in American cities (Florida, 2002a,b; Ottaviano and Peri, 2006).

0049-089X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ssresearch.2009.03.002

ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 721−727

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

Missing covariate data in medical research: To impute is better than to ignore

Kristel J.M. Janssen[a,*], A. Rogier T. Donders[b], Frank E. Harrell Jr.[c], Yvonne Vergouwe[a], Qingxia Chen[c], Diederick E. Grobbee[a], Karel G.M. Moons[a]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands*
[b]*Department of Epidemiology, Biostatistics and Health Technology Assessment, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands*
[c]*Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, USA*

Accepted 14 December 2009

Abstract

Objective: We compared popular methods to handle missing data with multiple imputation (a more sophisticated method that preserves data).

Study Design and Setting: We used data of 804 patients with a suspicion of deep venous thrombosis (DVT). We studied three covariates to predict the presence of DVT: D-dimer level, difference in calf circumference, and history of leg trauma. We introduced missing values (missing at random) ranging from 10% to 90%. The risk of DVT was modeled with logistic regression for the three methods, that is, complete case analysis, exclusion of D-dimer level from the model, and multiple imputation.

Results: Multiple imputation showed less bias in the regression coefficients of the three variables and more accurate coverage of the corresponding 90% confidence intervals than complete case analysis and dropping D-dimer level from the analysis. Multiple imputation showed unbiased estimates of the area under the receiver operating characteristic curve (0.88) compared with complete case analysis (0.77) and when the variable with missing values was dropped (0.65).

Conclusion: As this study shows that simple methods to deal with missing data can lead to seriously misleading results, we advise to consider multiple imputation. The purpose of multiple imputation is not to create data, but to prevent the exclusion of observed data.

Keywords: Missing data; Complete case analysis; Multiple imputation; Bias; Coverage; DVT

1. Introduction

No matter how hard researchers try to prevent it, missing data occur frequently in medical research [1]. Commonly, researchers simply neglect all the data of patients with missing values because this is what standard software packages do when the data are analyzed (complete case analysis). Because this leads to a smaller dataset, it comes at least at the price of loss of power. Complete case analysis not necessarily leads to biased results. Under the condition that the missing values are missing completely at random (MCAR), meaning that the cause of missingness is pure coincidence, complete case analysis will not lead to biased results. As an alternative to complete case analysis, researchers tend to drop a variable from the analysis when it has missing values. However, both methods neglect valuable observed data.

Multiple imputation is a statistical technique that uses all observed data to fill in plausible values for the missing values [2−8]. This method receives increasing attention in the medical literature [9−16]. Nevertheless, many researchers seem unaware or uncertain about this approach to deal with missing values and still perform a complete case analysis or drop variables with missing values from the analysis [17]. The extent and sort of bias related to these approaches depend on the type of study. Diagnostic or prognostic studies often study the contribution of covariates (eg, patient characteristics and test results) in the prediction of a particular outcome by estimating the predictors' regression coefficients. For example, one may study the predictive effect of body mass index (BMI), age, gender, the intake of saturated fat, and other life style factors on the risk of cardiovascular diseases (CVD). Sometimes, these studies are aimed at developing a multivariable prediction model or risk score and estimate the ability of such a model to distinguish between patients at high and low risk of CVD. In etiologic studies, usually the effect of a specific

* Corresponding author. Tel.: +0031-8875-51752; fax: +0031-8875-55485.

E-mail address: k.j.m.janssen@umcutrecht.nl (K.J.M. Janssen).

0895-4356/$ − see front matter
doi: 10.1016/j.jclinepi.2009.12.008

Journal of Adolescent Health 48 (2011) 52–58

ELSEVIER

JOURNAL OF ADOLESCENT HEALTH

www.jahonline.org

Original article

Environmental Influences on Young Adult Weight Gain: Evidence From a Natural Experiment

Kandice A. Kapinos, Ph.D.[a,*], and Olga Yakusheva, Ph.D.[b]

[a] *Institute for Social Research, University of Michigan, Ann Arbor, Michigan*
[b] *Department of Economics, Marquette University, Milwaukee, Wisconsin*

Article history: Received February 17, 2010; Accepted May 19, 2010
Keywords: Natural experiment; Adolescent obesity; Physical environment

A B S T R A C T

Objectives: This study investigated the importance of environmental influences in explaining weight gain and related behaviors among freshman college students.
Methods: We exploited a natural experiment that takes place on most college campuses in the United States - randomized dormitory assignments. We estimated the effects of living in dormitories with varying physical environment characteristics on weight gain and related behaviors (daily number of meals and snacks, weekly frequency of exercise) among randomly assigned freshman students.
Results: We found strong evidence linking weight and related behaviors to individual dormitories, as well as to specific characteristics of the dormitories. On average, students assigned to dormitories with on-site dining halls gained more weight and exhibited more behaviors consistent with weight gain during the freshman year as compared with students not assigned to such dormitories. Females in such dormitories weighed .85 kg ($p = .03$) more and exercised 1.43 ($p < .01$) times fewer; males consumed .22 ($p = .02$) more meals and .38 ($p = .01$) more snacks. For female students, closer proximity of the dormitory to a campus gym led to more frequent exercise (.54, $p = .03$), whereas living closer to central campus reduced exercise ($-.97$, $p = .01$).
Conclusions: Using a natural experiment to deal with the potential endogeneity of the living environment, this study found that the physical environment affects both students' weight changes and weight-related behaviors.

The increase in the prevalence of obesity in the United States in recent decades has resulted in considerable attention by public health and policy initiatives, the media, medical practitioners, and researchers alike. Numerous studies have investigated both the antecedents and consequences of being overweight or obese. The finding that body weight depends not only on biological factors, but also on environmental factors, implies that interventions that mitigate environmental influences are important in policies aimed at addressing this growing problem [1–3].

Obesity research focusing on the physical environment has investigated the role of the proximity, density, selection of healthy foods and eating facilities, and aspects of the built environment, such as "walkability," access to exercise facilities, parks, trails, urbanizaion, and crime [4–7]. Much of this work has found significant associations between characteristics of the physical environment and obesity. Living near supermarkets yields greater consumption of fruits and vegetables [4], whereas individuals who live in areas with higher concentrations of fast food restaurants tend to weigh more on average [8–10]. Individuals who have greater access to parks, gyms, or walking/jogging trails are more likely to engage in physical activity [6] and, not surprisingly, individuals who walk more and spend less time driving tend to have lower obesity rates [5]. A recent meta-analysis concluded that access to fast food and recreational facilities is consistently linked to weight-related behaviors and outcomes in adults [3].

However, all of this evidence relies on analyses that do not deal with the likely possibility that individuals choose to work

* Address correspondence to: Kandice A. Kapinos, Ph.D., Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48104.
E-mail address: kkapinos@umich.edu (K.A. Kapinos).

1054-139X/$ - see front matter © 2011 Society for Adolescent Health and Medicine. All rights reserved.
doi:10.1016/j.jadohealth.2010.05.021

TECHNICAL PAPER

ISSN:1047-3289 *J. Air & Waste Manage. Assoc.* **59**:1308–1316
DOI:10.3155/1047-3289.59.11.1308

An Optimal Spatial Configuration of Sample Sites for Air Pollution Monitoring

Naresh Kumar and Veronica Nixon
Department of Geography, University of Iowa, Iowa City, IA

Kaushik Sinha
Department of Computer Sciences, The Ohio State University, Columbus, OH

Xiaosen Jiang
Department of Computer Sciences, University of Iowa, Iowa City, IA

Sarah Ziegenhorn
Department of Geography, Macalester College, St. Paul, MN

Thomas Peters
Department of Occupational and Environmental Health, University of Iowa, Iowa City, IA

ABSTRACT

A novel sampling design is proposed that optimizes the spatial configuration of sampling sites and captures maximum intraurban variability in ambient air pollution with a minimum sample size. Unlike the classical sampling design, a deterministic approach is adopted and the redundancy in the site selection is minimized by controlling for spatial autocorrelation. The proposed design was tested and implemented in a medium-sized midwestern city. The analysis suggested that 32 sites were adequate to capture more than 95% of the total variance in airborne particulate 10 μm or less in aerodynamic diameter (PM_{10}). A list of 20 households was prepared around each of the 32 sites. Households were approached in order of their distance from these sites until one was recruited for intensive indoor and outdoor air pollution monitoring from spring through fall of 2008. Finally, 30 households located around the optimal sites participated in the study. One set of four photometric and gravimetric samplers was deployed for each indoor and outdoor environment. The average ambient PM_{10} concentration (monitored from April to September 2008) at the selected locations was lower but statistically insignificant as compared with the PM_{10} (computed using the data from mobile sampling in 2006) at the optimal sites.

IMPLICATIONS

The paper proposes an optimal spatial sampling design that captures the maximum variance in air pollution by deploying samplers at a minimum number of sites. This design is likely to have a major impact on air pollution monitoring strategies for capturing intraurban spatial variability for a specific air pollutant. This, in turn, will help epidemiologists to quantify intraurban exposure to ambient air pollution.

INTRODUCTION

Several studies have documented the linkages between health outcomes and air pollution.[1–3] Most of these studies, including the National Morbidity, Mortality, and Air Pollution Study (NMMAP),[4,5] rely on air pollution data generated from centrally located monitoring stations that do not necessarily account for intra-city variability in exposure. However, the extent of intra-city variability in air pollution can exceed that of inter-city variability. A study conducted in Los Angeles that was based on American Cancer Society (ACS) data identified intra-city variability in air pollution exposure as a significant predictor of cancer risks.[6] Other studies, conducted in developed and developing countries, also suggest substantial intra-city variability in air pollution. For example, the interquartile range of airborne particulate 10 μm or less in aerodynamic diameter (PM_{10}) in the Delhi metropolitan area monitored at 113 spatially dispersed sites from July to December 2003 was 157 ± 18.1 μg/m^3 (95% confidence interval [CI]).[7] In another study that was conducted in Iowa City, IA, with an area of 112.5 km^2, the 3-week averages of particulate matter of aerodynamic diameter between 2.5 and 10 μm ($PM_{10-2.5}$) across 33 sites ranged from 9.3 to 20.1 μg/m^3.[8] These observations call for improved sampling strategies that can effectively account for intra-city/intraurban variability in air quality.

This article presents a deterministic spatial sampling design that adequately captures the intra-city variability in air pollution. The proposed design was adopted to draw a sample of households that represented population exposure to indoor and outdoor airborne particulates of different sizes in the area surrounding Iowa City and Coralville, IA (United States). The remainder of this article presents a theoretical framework of the proposed sampling design and its applications, followed by a detailed discussion.

Int J Public Health
DOI 10.1007/s00038-011-0276-2

ORIGINAL ARTICLE

Perceptions of the food environment are associated with fast-food (not fruit-and-vegetable) consumption: findings from multi-level models

Sean C. Lucan · Nandita Mitra

Received: 2 February 2011 / Revised: 30 June 2011 / Accepted: 4 July 2011

Abstract
Objectives Diets low in fruits and vegetables and/or high in fast foods are associated with obesity and chronic diseases. Such diets may relate to different aspects of neighborhood food environments. We sought to evaluate if people's perceptions of their neighborhood food environment are associated with reported fruit-and-vegetable and fast-food consumption.
Methods Cross-sectional analysis of a community health survey from Philadelphia, PA and four surrounding suburban counties ($n = 10{,}450$ individuals). We used mixed-effects multi-level Poisson models, nesting individuals within *neighborhoods*—i.e. census tracts ($n = 991$).
Results Negative perceptions of the food environment (perceived difficulty finding fruits and vegetables, having to travel outside of one's neighborhood to get to a supermarket, and perceived poor grocery quality) were each directly associated with fast-food consumption (incident rate ratios [IRRs] 1.31, 1.06, 1.20; $p < 0.001$, 0.04, < 0.001 respectively), but not significantly associated with fruit-and-vegetable consumption.
Conclusions Perceived difficulty finding or accessing produce and high-quality groceries may support the eating of more fast food. Neighborhoods where food-environment perceptions are worst might benefit from interventions to improve availability, accessibility, and quality of healthy foods, towards shifting consumption away from fast foods.

Keywords Fruits and vegetables · Fast food · Food environment · Multi-level models · Neighborhoods

S. C. Lucan (✉)
Department of Family and Social Medicine,
Albert Einstein College of Medicine/Montefiore Medical Center,
Bronx, NY, USA
e-mail: slucan@yahoo.com

N. Mitra
Department of Biostatistics, University of Pennsylvania,
Philadelphia, PA, USA

Introduction

Diet-related diseases are among the leading causes of death and disability in the developed world.(Michaud et al. 2001; Mokdad et al. 2004) Cardiovascular disease and cancer lead the list in the U.S.,(Mokdad et al. 2004) and are contributed by other diet-related conditions such as high blood pressure, diabetes, high cholesterol, and obesity (American Heart Association; National Cancer Institute 2009). Such conditions are all associated with dietary patterns high in fast foods (ready-to eat convenience items generally rich in unhealthy fats, sodium, and/or added sugars) and/or low in fruits and vegetables (Bazzano et al. 2003; Berkey et al. 2004; He et al. 2006; Key et al. 1999; Pereira et al. 2005; Rolls et al. 2004; Vainio and Weiderpass 2006).

While people's dietary patterns may depend in part on a host of individual factors,(Booth et al. 2001; Wetter et al. 2001) experts increasingly emphasize the importance of local environments in shaping individuals' dietary behaviors (Booth et al. 2001; Frieden 2010). Aspects of local environments that may be particularly important include the availability, accessibility, and quality of various foods. Food availability, food-store and restaurant accessibility, and overall grocery quality are characteristics of local food environments that may influence whether residents have predominantly healthy or unhealthy dietary patterns (Cheadle et al. 1991; Dibsdall et al. 2003; Franco et al. 2009;

Published online: 20 July 2011

Social Science Research 41 (2012) 977–990

Contents lists available at SciVerse ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Underage drinking, alcohol sales and collective efficacy: Informal control and opportunity in the study of alcohol use

David Maimon [a,*], Christopher R. Browning [b]

[a] Department of Criminology and Criminal Justice, University of Maryland, 2220 LeFrak Hall, College Park, MD 20742, United States
[b] Department of Sociology, Ohio State University, 214 Townshend Hall, Columbus, OH 43210, United States

ARTICLE INFO

Article history:
Received 16 April 2011
Revised 25 January 2012
Accepted 30 January 2012
Available online 9 February 2012

Keywords:
Neighborhoods
Underage drinking
Informal social control
Adolescents

ABSTRACT

Underage drinking among American youth is a growing public concern. However, while extensive research has identified individual level predictors of this phenomenon, few studies have theorized and tested the effect of structural social forces on children's and youths' alcohol consumption. In an attempt to address this gap, we study the effects of residential environments on children's and youths' underage drinking (while accounting for personality and familial processes). Integrating informal social control and opportunity explanations of deviance, we first suggest that while neighborhood collective efficacy prevents adolescents' underage drinking, individuals' access to local alcohol retail shops encourages such behavior. Focusing on the interactive effects of communal opportunities and controls, we then suggest that high presence of alcohol outlets and sales in the neighborhood is likely to increase youths' probability of alcohol consumption in the absence of communal mechanisms of informal social control. We test our theoretical model using the unprecedented data design available in the PHDCN. Results from a series of multilevel logit models with robust standard errors reveal partial support for our hypotheses; specifically, we find that alcohol sales in a given neighborhood increase adolescents' alcohol use. In addition, while the direct effect of collective efficacy is insignificantly related to children's and youths' alcohol consumption, our models suggest that it significantly attenuates the effect of local alcohol retailers and sales on underage drinking.

1. Introduction

Although underage drinking (i.e. alcohol consumption by individuals younger than 21 years old) has considerably declined during the last 15 years, alcohol use among adolescents and youth continues to be the leading drug problem in the US (Johnston et al., 2009). Specifically, while the 30-day prevalence of alcohol use has fallen by 40% among 8th graders, 30% among 10th graders and 16% among 12th graders between the years 1996 and 2008, the proportions of 8th, 10th and 12th graders who admitted to drinking an alcoholic beverage in the past month were still high in 2008 (16%, 29% and 43% respectively). Underage drinking involves profound consequences for young alcohol consumers. Several studies show, for instance, that alcohol consumption during adolescence has serious and significant consequences for the brain and the hormonal system (US Department of Health and Human Services, 2006), leading for instance to loss of memory (Nagel et al., 2005) and of verbal and non-verbal tasks (Brown et al., 2000). Further, it may lead to a lifetime of alcohol abuse and dependence (Grant and Dawson, 1997). Empirical studies also suggest that underage drinking is strongly tied to a

* Corresponding author.
E-mail address: dmaimon@umd.edu (D. Maimon).

0049-089X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ssresearch.2012.01.009

Adjusting Survey Weights When Altering Identifying Design Variables Via Synthetic Data

Robin Mitra and Jerome P. Reiter

Duke University, Durham, NC 27708, USA
{rm51, jerry}@stat.duke.edu
http://www.stat.duke.edu

Abstract. Statistical agencies alter values of identifiers to protect respondents' confidentiality. When these identifiers are survey design variables, leaving the original survey weights on the file can be a disclosure risk. Additionally, the original weights may not correspond to the altered values, which impacts the quality of design-based (weighted) inferences. In this paper, we discuss some strategies for altering survey weights when altering design variables. We do so in the context of simulating identifiers from probability distributions, i.e. partially synthetic data. Using simulation studies, we illustrate aspects of the quality of inferences based on the different strategies.

Keywords: Disclosure; Multiple imputation; Swapping; Synthetic data; Weights.

1 Introduction

Survey design variables often contain identifying information, for example race in a survey that over-samples minorities or establishment size in a probability proportional to size sample of businesses. To limit disclosure risks, statistical agencies may need to alter these variables before releasing the data to the public. It also may be necessary to alter the survey weights, which typically are deterministic functions of the design variables. Failure to do so can leave identifying information on the file, effectively defeating the purpose of the masking [1]. For example, an unaltered weight could reveal that a person was part of a minority group or could disclose the size of the establishment. Not altering weights also could affect the quality of data analysts' estimates, because the weights may not be appropriate for making the released sample representative of the population.

In this paper, we discuss some strategies for adjusting survey weights when altering design variables to limit disclosure risks. We do so in the context of simulating identifiers from probability distributions, i.e. partially synthetic data. Using simulation studies, we illustrate aspects of the data quality and confidentiality of the different strategies. We also examine the performance of the strategies when swapping identifiers.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 177–188, 2006.

Statistical Methodology 9 (2012) 285–304

Contents lists available at SciVerse ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Comparison of designs for generalized linear models under model misspecification

S. Mukhopadhyay [a,*], A.I. Khuri [b]

[a] Department of Mathematics, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India
[b] Department of Statistics, University of Florida, P.O. Box 118545, Gainesville, FL 32611-8545, USA

ARTICLE INFO

Article history:
Received 2 May 2011
Received in revised form
3 August 2011
Accepted 18 August 2011

Keywords:
Kriging
Linear predictor
Mean-squared error of prediction
Model bias
Response surface methodology

ABSTRACT

The purpose of this article is to demonstrate the use of the quantile dispersion graphs (QDGs) approach for comparing candidate designs for generalized linear models in the presence of model misspecification in the linear predictor. The proposed design criterion is based on the mean-squared error of prediction which incorporates the prediction variance and the bias caused by fitting the wrong model. The method of kriging is used to estimate the unknown function assumed to be the cause of model misspecification. The QDGs approach is also useful in assessing the robustness of a given design to values of the unknown parameters in the linear predictor. Three numerical examples are presented to illustrate the application of the proposed methodology.

1. Introduction

The true relationship between a given response and its control variables (modeled using the linear predictor in a generalized linear model (GLM)) is usually unknown, but is assumed to have a certain form. However, there is always the possibility that the fitted model is incorrect. Therefore, designs for GLMs should be compared on the basis of the mean-squared error of prediction which combines the prediction variance and the bias introduced by the misspecified model.

One of the main objectives of response surface methodology (RSM) is the choice of a design for fitting a model which describes the relationship between the response and its corresponding control variables. Several criteria, such as A-, D-, E-, and G-optimality are frequently used in linear models and GLMs for selecting the best design. However, most of these criteria are concerned with the reduction

* Corresponding author. Tel.: +91 2225767495.
E-mail address: siuli@math.iitb.ac.in (S. Mukhopadhyay).

1572-3127/$ – see front matter
doi:10.1016/j.stamet.2011.08.004

BMC Oral Health

RESEARCH ARTICLE Open Access

Contextual and individual assessment of dental pain period prevalence in adolescents: a multilevel approach

Marco A Peres[1*], Karen G Peres[1], Antônio C Frias[2], José Leopoldo F Antunes[3]

Abstract

Background: Despite evidence that health and disease occur in social contexts, the vast majority of studies addressing dental pain exclusively assessed information gathered at individual level.

Objectives: To assess the association between dental pain and contextual and individual characteristics in Brazilian adolescents. In addition, we aimed to test whether contextual Human Development Index is independently associated with dental pain after adjusting for individual level variables of socio-demographics and dental characteristics.

Methods: The study used data from an oral health survey carried out in São Paulo, Brazil, which included dental pain, dental exams, individual socioeconomic and demographic conditions, and Human Development Index at area level of 4,249 12-year-old and 1,566 15-year-old schoolchildren. The Poisson multilevel analysis was performed.

Results: Dental pain was found among 25.6% (95%CI = 24.5-26.7) of the adolescents and was 33% less prevalent among those living in more developed areas of the city than among those living in less developed areas. Girls, blacks, those whose parents earn low income and have low schooling, those studying at public schools, and those with dental treatment needs presented higher dental-pain prevalence than their counterparts. Area HDI remained associated with dental pain after adjusting for individual level variables of socio demographic and dental characteristics.

Conclusions: Girls, students whose parents have low schooling, those with low *per capita* income, those classified as having black skin color and those with dental treatment needs had higher dental pain prevalence than their counterparts. Students from areas with low Human Development Index had higher prevalence of dental pain than those from the more developed areas regardless of individual characteristics.

dental pain epidemiology, oral health, socioeconomic factors, multilevel analysis

Background

Dental pain is described as pain originating from innervated tissues of the tooth or immediately adjacent to it [1]. It is a subjective oral health indicator caused mainly by dental caries and should become uncommon when oral health improves [2]. Conditions such as erosion, trauma, and exfoliation of primary teeth can also cause dental pain [3]. In low-to-middle income countries, most caries remain untreated, and dental care may not be easily available and is not universally free in most of these countries [4]. Most international data on dental pain have reported period prevalence more than point prevalence, and range between around 10 and 30% depending on the case definition and assessment methods adopted [5]. Period prevalence refers to the number of persons known to have had pain at any time during a specified period, usually 6 months in dental pain studies, while point prevalence refers to the number of persons with pain at a specified point in time [6].

When children and adolescents are taken into account, dental pain may be of social concern because it

* Correspondence: mperes@ccs.ufsc.br
[1]Oral Epidemiology and Public Health Dentistry, Post-graduate Program in Public Health, Department of Public Health, Universidade Federal de University of Santa Catarina, Florianópolis, Brazil
Full list of author information is available at the end of the article

Research

Ecography 34: 1040–1048, 2011
doi: 10.1111/j.1600-0587.2011.06370.x

Subject Editor: Pedro Peres-Neto. Accepted 18 February 2011

Monitoring temporal trends in spatially structured populations: how should sampling effort be allocated between space and time?

Jonathan R. Rhodes and Niclas Jonzén

J. R. Rhodes (j.rhodes@uq.edu.au), The Univ. of Queensland, School of Geography, Planning and Environmental Management, Brisbane, QLD 4072, Australia. JRR also at: The Univ. of Queensland, The Ecology Centre, Brisbane, QLD 4072, Australia. – N. Jonzén, Dept of Theoretical Ecology, Ecology Building, Lund Univ., SE-223 62 Lund, Sweden.

Estimating temporal trends in spatially structured populations has a critical role to play in understanding regional changes in biological populations and developing management strategies. Designing effective monitoring programmes to estimate these trends requires important decisions to be made about how to allocate sampling effort among spatial replicates (i.e. number of sites) and temporal replicates (i.e. how often to survey) to minimise uncertainty in trend estimates. In particular, the optimal mix of spatial and temporal replicates is likely to depend upon the spatial and temporal correlations in population dynamics. Although there has been considerable interest in the ecological literature on understanding spatial and temporal correlations in species' population dynamics, little attention has been paid to its consequences for monitoring design. We address this issue using model-based survey design to identify the optimal allocation of sampling effort among spatial and temporal replicates for estimating population trends under different levels of spatial and temporal correlation. Based on linear trends, we show that how we should allocate sampling effort among spatial and temporal replicates depends crucially on the spatial and temporal correlations in population dynamics, environmental variation, observation error and the spatial variation in temporal trends. When spatial correlation is low and temporal correlation is high, the best option is likely to be to sample many sites infrequently, particularly when observation error and/or spatial variation in temporal trends are high. When spatial correlation is high and temporal correlation is low, the best option is likely to be to sample few sites frequently, particularly when observation error and/or spatial variation in temporal trends are low. When abundances are spatially independent, it is always preferable to maximise spatial replication. This provides important insights into how spatio-temporal monitoring programmes should be designed to estimate temporal trends in spatially structured populations.

The direction and magnitude of temporal trends in biological populations have critical implications for environmental management and policy (Gerber et al. 1999, IUCN 2001) and for understanding population dynamics (Clark and Bjørnstad 2004). Estimates of temporal trends can also be important for identifying ecosystem responses to climate change and other threatening processes (Rosenzweig et al. 2008). However, for broadly distributed species, monitoring regional population trends can be costly both in terms of time and financial resources (Pollock 2006, Nielsen et al. 2009). Consequently, identifying efficient sampling strategies to estimate temporal trends in these populations is of particular interest to ecologists and decision makers (Field et al. 2005, Rhodes et al. 2006). Quantifying temporal trends in spatially structured populations requires the monitoring of multiple, spatially distinct, sites over time. The most commonly used and effective design for achieving this is to monitor the same sites repeatedly over time (Urquhart and Kincaid 1999, Marsh and Trenham 2008). Other designs that rotate, or randomly select, sampling sites over time are also possible (McDonald 2003, de Gruijter et al. 2006), but are less commonly used (Marsh and Trenham 2008). In all cases, an important decision that must be made is how best to allocate sampling effort among spatial and temporal replicates. In fact, our ability to detect and quantify trends may depend critically on this decision (Carlson and Schmiegelow 2002). However, the development of general principles for informing this decision has received little attention in the ecological literature to date.

Complex interactions among spatial and temporal dynamics in spatially structured populations can complicate the interpretation and design of monitoring surveys (Bowers 1996, Brawn and Robinson 1996). For example, dispersal and habitat selection processes can introduce spatial relationships in population dynamics that affect our ability to detect temporal population trends (Jonzén et al. 2005). A particularly important issue for real populations is that the dynamics of distinct sub-populations are often spatially correlated due to processes such as dispersal and habitat selection, predation and correlated environmental noise (Moran 1953, Ydenberg 1987, Ripa 2000). Correlation is

Metrika (2005) 61: 137–156
DOI 10.1007/s001840400328

Metrika

Combining random sampling and census strategies - Justification of inclusion probabilities equal to 1

Horst Stenger[1] and Siegfried Gabler[2]

[1]Professor Dr. Horst Stenger L7, 3-5 University of Mannheim 68131 Mannheim, Germany (E-mail: stenger@rumms.uni-mannheim.de)
[2]PD Dr. Siegfried Gabler B2, 1 Centre for Survey Research and Methodology Postfach 12 21 55 68072 Mannheim, Germany (E-mail: gabler@zuma-mannheim.de)

Abstract. Very often values of a size variable are known for the elements of a population we want to sample. For example, the elements may be clusters, the size variable denoting the number of units in a cluster. Then, it is quite usual to base the selection of elements on inclusion probabilities which are proportionate to the size values. To estimate the total of all values of an unknown variable for the units in the population of interest (i.e. for the units contained in the clusters) we may use weights, e.g. inverse inclusion probabilities. We want to clarify these ideas by the minimax principle. Especially, we will show that the use of inclusion probabilities equal to 1 is recommendable for units with high values of the size measure.

Key words: Asymptotically minimax strategies, RHC-strategy, stratification

AMS Classification 2000: Primary 62D05. Secondary 62C20

1 Introduction

Suppose a population consists of N clusters of sizes $x_1, \ldots, x_N$, i.e. cluster i comprises x_i units, and it is desired to estimate the total number of units belonging to a specified class. Let y_i be the number of units of cluster i which are members of the class of interest. Hence, $0 \leq y_i \leq x_i$ and we are looking for

$$y = \sum_{1}^{N} y_i$$

or for y/x with $x = \Sigma x_i$. Then the Hansen-Hurwitz strategy (1943) may be used with higher selection probabilities for the larger clusters or we may give large clusters higher inclusion probabilities and use the Horvitz-Thompson

CHILD & ADOLESCENT PSYCHIATRY & MENTAL HEALTH

RESEARCH Open Access

Modelling the relationship between obesity and mental health in children and adolescents: findings from the Health Survey for England 2007

Paul A Tiffin[1*], Bronia Arnott[2], Helen J Moore[1] and Carolyn D Summerbell[1]

Abstract

A number of studies have reported significant associations between obesity and poor psychological wellbeing in children but findings have been inconsistent. Methods: This study utilised data from 3,898 children aged 5-16 years obtained from the Health Survey for England 2007. Information was available on Body Mass Index (BMI), parental ratings of child emotional and behavioural health (Strengths and Difficulties Questionnaire), self-reported physical activity levels and sociodemographic variables. A multilevel modelling approach was used to allow for the clustering of children within households. Results: Curvilinear relationships between both internalising (emotional) and externalising (behavioural) symptoms and adjusted BMI were observed. After adjusting for potential confounders the relationships between obesity and psychological adjustment (reported externalising and internalising symptoms) remained statistically significant. Being overweight, rather than obese, had no impact on overall reported mental health. 17% of children with obesity were above the suggested screening threshold for emotional problems, compared to 9% of non-obese children. Allowing for clustering and potential confounding variables children classified as obese had an odds ratio (OR) of 2.13 (95% CI 1.39 to 3.26) for being above the screening threshold for an emotional disorder compared to non-obese young people. No cross-level interactions between household income and the relationships between obesity and internalising or externalising symptoms were observed. Conclusions: In this large, representative, UK-based community sample a curvilinear association with emotional wellbeing was observed for adjusted BMI suggesting the possibility of a threshold effect. Further research could focus on exploring causal relationships and developing targeted interventions.

Keywords: Obesity, Children, Adolescents, Mental Health, Statistical Modelling

Background

Childhood obesity is a serious health problem in the Western world with evidence of continued high rates [1,2]. Moreover, excess adiposity in children tracks throughout adulthood [3] and is linked to serious physical health risks [4]. Thus, a continued paediatric obesity epidemic will be associated with increased long-term health and social care costs and decreased productivity at a time of global economic downturn [5]. Rates of mental health problems in young people are also high, and increasing, with around one in ten children aged 5-16 years having a diagnosable condition [6,7]. Like obesity, mental ill health has been identified as a major cause of persistent disability with attendant economic implications [8].

Obesity has been shown to be associated with poor mental health in studies of working-age adults [9,10] with most research focussed on depression. A meta-analysis pooling the results of 17 cross-sectional studies concluded that the association between obesity and depression was highly statistically significant and possibly varied by gender [11]. There are many plausible reasons why excess adiposity may be associated with poor psychological adjustment. These include: the impact of obesity on self-esteem and social confidence; the direct effect of hormonal and metabolic changes on brain function [12,13]; the result of changes in dietary behaviour and physical activity levels

* Correspondence: p.a.tiffin@durham.ac.uk
[1]School of Medicine and Health, Wolfson Research Institute, Durham University Queen's Campus, University Boulevard, Stockton-on-Tees, TS17 6BH, UK
Full list of author information is available at the end of the article

Economics of Education Review 30 (2011) 528–545

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Education and labor market risk: Understanding the role of data cleaning☆

Alexander Whalley*

School of Social Sciences, Humanities and Arts, University of California - Merced & NBER, United States

ARTICLE INFO

Article history:
Received 18 May 2009
Received in revised form
14 December 2010
Accepted 23 December 2010

JEL classification:
J2
I6

Keywords:
Human capital
Labor market risk

ABSTRACT

This paper examines whether conclusions about the relationship between education and labor market risk depend on the use of commonly applied procedures to clean data of extreme values. The analysis uses fifteen years of data from the Panel Study of Income Dynamics to demonstrate that conclusions about the relationship between education and labor market risk are sensitive to how extreme values of labor income are treated. The untrimmed estimates imply that college graduates experience 75% less transitory labor market risk than high school dropouts. However, applying commonly used trimming procedures results in estimates of a one standard deviation transitory labor market shock for high school dropouts being reduced by between $2700 and $4500, or 14% and 24% of annual earnings. The results demonstrate that seemingly innocuous sample selection procedures can have substantive implications.

1. Introduction

Human capital is one of the central assets most people hold. A vast body of economic research has investigated the question of how education affects the level of income (see Card, 2001 for a review). Recently, however, economists have also studied the effect of education on the riskiness of labor income, as evidence that markets for sharing idiosyncratic labor income risk are incomplete has accumulated (i.e. Cochrane, 1991). When markets are incomplete the relevant rate of return to education is risk-adjusted, rather than based on expected income alone.

Crucial for the estimation of the risk-adjusted rate of return to education is the responsiveness of labor income risk to education. While much progress has been made, the findings remain mixed. Hubbard, Skinner, and Zeldes (1994b), Carroll and Samwick (1997), and Guvenen (2009) all find little statistically significant relationship between education and labor income volatility. In contrast, Meghir and Pistaferri (2004) find a statistically significant relationship between education and labor income volatility.[1]

One important way the labor income volatility studies differ is in terms of the data cleaning procedures applied to minimize the impact of measurement error. As seemingly innocuous data cleaning procedures can have substantive effects (Bollinger & Chandra, 2005) and different studies of labor income volatility adopt different data cleaning procedures, this paper considers whether commonly applied data cleaning procedures can explain the mixed findings.

☆ I thank Thomas DeLeire, James Feigenbaum, Shawn Kantor, Beomsoo Kim, Michael Wall, and Katie Winder for helpful comments and conversations. Financial support from the Washington Economic Club Dissertation Fellowship is gratefully acknowledged. All errors are my own.

* Tel.: +1 209 228 4027; fax: +1 209 228 4007.
E-mail address: awhalley@ucmerced.edu

[1] In contrast, recent applications of mean-variance models of asset pricing to human capital indicate that education is more valuable than the expected rate of return alone would indicate. Palacios-Huerta (2003) shows that risk adjusted rates of return to education are greater than the effect of education on expected income alone. Hartog and Vijverberg (2007) show that broadly based school curriculums do indeed reduce risk. In addition, earnings losses from job displacement have been shown to decline with education (see Stevens, 1997; Farber, 1997, 2005).

0272-7757/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.econedurev.2010.12.013