Journal of Interdisciplinary History, XLVII:1 (Summer, 2016), 85–92.

Barry Edmonston

The Statistical Analysis of Longitudinal Data

Lives in Transition: Longitudinal Analysis from Historical Sources. Edited by Peter Baskerville and Kris Inwood (Montreal, McGill-Queen's University Press, 2015) 381 pp. $110.00 cloth $34.95 paper

Lives in Transition offers an innovative and useful discussion of data and methods for quantitative longitudinal historical research. After a helpful introduction, it includes three chapters about international migration, four about mobility in rural areas, two about mobility in urban areas, and three about ethnic groups during World War I. The chapters focus on four countries. New Zealand and Australia each receive a chapter; the United States receives two and Canada seven. Most of the analysis pertains to the second half of the 1800s, although two of the chapters concentrate on the first half of the 1800s and three deal with the early 1900s.

The book presents an excellent opportunity to discuss issues related to data collection and statistical analysis. This review essay first surveys various types of quantitative life-course data and how the chapters in this volume exemplify the collection, linkage, and analysis of such data before exploring the statistical analysis of quantitative longitudinal data.

QUANTITATIVE LIFE-COURSE DATA Quantitative life-course analysis has benefited greatly from the recent expansion of suitable data sources and the development of appropriate statistical methods. Most life-course analysis is based on four types of census, survey, or administrative data. One type, prospective data, derives from a particular group's responses to a regular series of surveys over time. In some cases, the prospective data include information from such administrative sources as employment or tax records. These data are expensive to collect and take a long time to accumulate for

Barry Edmonston is Research Professor, Department of Sociology, and Associate Director, Population Research Group, University of Victoria. He is the editor of the Special Issue "Life-course Perspectives on Immigration," *Canadian Studies in Population*, XL (2013), 1–102; with Eric Fong, *Canada's Population Situation* (Montreal, 2011).

 doi:10.1162/JINH_a_00942

analysis. This approach, however, has the advantage of collecting contemporary information as conditions change and events occur.

Another type, retrospective data, derives from surveys about events and conditions in the past. It is less expensive than the prospective type and provides immediate data for analysis, involving the duration between major events, such as the time between marriage and the first birth of a child. The disadvantages of retrospective data are unknown selection (for example, immigrants not always living or residing in their destination country) and recall bias (respondents not always remembering key events or their dates).

The third way for life-course information to be collected is by linking several censuses or surveys over time or linking these data with administrative records, primarily concerning prior or later immigration, military service, income, employment, birth, death, or marriage. This approach to collecting information about changes over time is relatively inexpensive, but the information available from census, survey, and administrative records is often limited.

The fourth data type is based on synthetic cohorts in censuses and surveys. This method usually relies on birth or immigration cohorts for analysis, rather than on individuals. Studies usually follow either a birth cohort (people born during, say, the Great Depression) or an immigrant cohort (a group of immigrants arriving during the same period) through time, using several censuses or surveys to compare its experiences in marriage or labor-force participation with that of other residents. It provides inexpensive data for long periods of time but confines analysis to cohort comparisons with potential selection biases.

Prospective and retrospective data are rare in historical research, unless researchers are familiar with a historical dataset that prospectively or retrospectively contains evidence for a group of individuals over time. There are, however, unique prospective datasets available for historical research. The Oakland Growth Study, under the leadership of Glen Elder, interviewed 167 adolescents born in 1920/1 in 1932 and four times thereafter to study the effects of the Great Depression on the American family. The last of the five interviews occurred in 1980/1, when the original cohort was nearing retirement. A list of other potential prospective or retrospective data sets, their data content, and availability for historical research would be useful to compile.

Most longitudinal data for historical research involves tracking individuals along several censuses or surveys, or linking those data to administrative sources with dates either before or after those of the census or survey data. Criminal records could be joined with later census data to investigate the relationship between different criminal statuses and subsequent employment, or death certificates could be linked to earlier census data to ascertain how occupations relate to causes of death.

The twelve chapters in this volume illustrate applications of several of these four types of data. Five of them make important use of the linkage between two or more censuses or surveys. Gordon Darroch's contribution links 1861 and 1871 microdata records for a study of agricultural occupational mobility in Ontario. Luiza Antonie, Baskerville, Inwood, and J. Andrew Ross analyze 1871 and 1881 census records to determine Canadian work patterns. Baskerville connects data for Perth County, Ontario, from the 1871 Canada census to either the 1881 Canada census or the 1880 U.S. census to trace migration paths. Although each of these five chapters includes a helpful discussion of the merits and limitations of linked census records, Sherry Olson's deserves special mention for its excellent advice regarding the fine points of how to link records and recover certain kinds of data, such as addresses. Evan Roberts connects a unique 1924/5 survey of 477 Chicago families to the 1920 and 1930 U.S. censuses, revealing the high degree of occupational and spatial mobility that existed even within a five-year period. His plans to compare the survey data with the forthcoming release of 1940 census data is keenly anticipated for the light that it could shed on family strategies for coping with the Great Depression.

Two chapters analyze data obtained by a linkage between census microdata and administrative records. John Cranfield and Inwood work with microdata from Canada's 1901 census and Canadian Expeditionary Force military records from 1914 to 1918 to examine differences in height between British and French soldiers. Based on similar linked data, Allegra Fryxell, Inwood, and Aaron van Tassel explore the participation of Australian Aboriginal soldiers in World War I, a topic that has received limited attention.

Four chapters delve into sets of administrative records to extract novel data with exceptional interest. Rebecca Kippen and Janet McCalman consult the criminal records of prisoners who arrived in Tasmania from 1826 to 1838 and follow up with administrative

records after their arrival, including useful data about a special group of working-class rioters transported to Tasmania during this period. Their efforts result in an interesting comparison group for understanding the selection bias of data about convicts and their outcomes in Tasmania. Hamish Maxwell-Stewart and Kippen deal with three administrative data sets—medical records for individual convicts transported to Australia, contextual data about the 289 convict vessels sailing from British or Irish ports, and various records collected after convicts' arrival. The upshot is intriguing evidence about the relationship between the condition of a vessel and male/female mortality, among other things. Rebecca Lenihan uses a genealogical register in New Zealand to study Scottish immigrants—a self-selected data set based on records already examined rather than a sample of all possible immigrant arrivals (single men who did not remain long and did not have descendants are likely missing from such genealogical registers). Linked longitudinal data does not apply to individuals only. Kenneth M. Sylvester and Susan Hautaniemi Leonard use agricultural censuses from Kansas to link data about twenty-five farming communities from 1875 to 1940 to test several ideas about the availability of land, labor mobility, and the evolution of family farms.

Kandace Bogaert, Jane van Koeverden, and D. Ann Herring follow a cohort of Polish-American military personnel in 1917—from their recruitment in the United States through their training at Camp Koscuiszko in Niagara-on-the-Lake, Ontario—to discover their mortality experience during the deadly 1918 influenza epidemic. Although the most common use of the term *cohort* is in relation to birth, such as a group of babies born during the 1930s, the demographic definition of *cohort* relates to a collection of people experiencing a common event during a common period of time, like these Polish-American soldiers.

These twelve chapters show how the analysis of quantitative longitudinal data can reveal new areas for social science and historical research. They provide helpful instruction regarding the creation of longitudinal data for the study of traditional topics as well as new areas of investigation.

STATISTICAL ANALYSIS OF LIFE-COURSE DATA Current life-course analysis relies on several multivariate statistical tools. Statisticians proposed new methods for the analysis of event changes observed in longitudinal data several decades ago; these initial models have

evolved into generalized linear mixed varieties (also known as multilevel or hierarchical models) that are suitable for analysis of continuous, binary, or counted data over time. Generalized linear mixed models, as implemented in several statistical packages, are described in applied statistics textbooks.[1] Other statistical methods, such as the double-cohort approach, have been developed in recent years for analysis of longitudinal immigration data.[2]

There are several specific issues in statistical analysis that the chapters in this volume bring to light. First, sample size should always be stated in tables and figures. In cases when analysis deals with a total sample, readers can correctly infer sample size, but in cases when analysis is limited to a selected group, sample size needs to be noted. Second, several chapters in this volume make appropriate use of logistic regression analysis, but they would have done well to cite several overall tests, including the chi-square fit and its statistical significance, as well the adjusted R-squared. Third, when giving rates, researchers should show the number of observations for the denominator so that readers can compare the sample size used for the different rates. Finally, figures that are perfectly comprehensible in color may need to be designed for presentation in black and white. The use of different types of shading or markers can help readers to identify different categories in the figures. Moreover, black-and-white figures can be improved by showing numbers or percentages if comparisons are not clear.

Several chapters in this volume use logistic regression models to indicate factors affecting a binary-outcome variable, such as labor-force participation. Because the binary logit model is nonlinear, it is a challenge to interpret the relationship between an explanatory variable and the outcome. One common approach—taken by some of the contributors to this volume—is to take the exponent of the logit regression coefficient, which indicates the expected change in the odds of the outcome. But, for most analysts, interpreting changes in the odds of an outcome is difficult because it depends on the base probability. Consider two examples: (1) If the odds of the outcome is 1/100, the corresponding probability is

1 See Sophia Rabe-Hesketh and Anders Skrondal, *Multilevel and Longitudinal Modeling Using Stata* (College Station, 2012; orig. pub. 2005); Judith D. Singer and John B. Willett, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence* (New York, 2003).

2 See, for example, Edmonston and Sharon M. Lee, "Immigrants' Transition to Homeownership, 1991 to 2006," *Canadian Studies in Population*, XL (2013), 57–74.

about 0.01. If the effect of an explanatory variable doubles the odds, the odds increase to 2/100, and the probability doubles to about 0.02. In this instance, the interpretation is fairly clear because a doubling of the odds ratio increases the outcome probability by about two. (2) If the odds of the outcome are 1/1, the corresponding probability is 0.50. If the odds double to 2/1, however, the corresponding probability increases to 0.67. In this instance, there is no clear intuitive interpretation for the relationship of a change in the odds and the change in the corresponding probability.

The key point is that a constant change in the odds ratio does *not* correspond to a constant change in the outcome probability. The change in the outcome probability is affected by the base probability. In most analyses, when the base probability is not close to zero, it is preferable to interpret binary logit coefficients in terms of changes in the outcome probabilities. In medical research, the outcome variable is often close to zero; hence, health researchers routinely cite the odds ratio for interpreting logistic regression coefficients. If, for example, the coefficient for cigarette smoking is 2.1, and the probability for a non-smoker to contract lung cancer is 0.0001, then cigarette smokers are 8 times (the exponent of 2.1 is about 8) more likely to get lung cancer, with a probability of 0.0008. In this case, the interpretation of the odds ratio is clear. However, the interpretation is not apparent when the outcome probability is not close to zero, which is the case for many outcomes in historical and social-science research.

A variety of approaches for interpreting the relationship between explanatory and outcome variables for binary logit models is possible when the outcome probability is not close to zero.[3] A useful one is to compute predicted values or probabilities of the outcome for specified values of the explanatory variables. Such predicted probabilities for the outcome are also called predictive margins, or adjusted predictions, in the statistical literature. Predicted probabilities have a straightforward interpretation because they indicate the outcome probability for a specific value of an explanatory variable, holding constant all other explanatory variables.

Consider a logit regression model that predicts labor-force participation (a binary variable coded 0 if not in the labor force

3 J. Scott Long and Jeremy Freese, *Regression Models for Categorical Dependent Variables Using Stata, Second Edition* (College Station, 2006; orig. pub. 2001).

and 1 if in the labor force) based on sex and education. The predicted probability for females is the average probability for labor-force participation if everyone in the data is treated as female while all other variables are held constant with their estimated regression coefficients. The predicted probabilities for females and males, in this example, give the expected probability of labor-force participation by sex, holding education constant.

To make this example specific, consider the following estimated logit regression equation: $Y = 0.5 + 0.1$ Sex $+ 0.2$ Education, where Y is a binary outcome variable (0 for not in the labor force and 1 for in the labor force), Sex is a dummy variable (0 for female and 1 for male), and Education is a dummy variable (0 for less than high school and 1 for high school). Further suppose a data set of 100 adults—50 females and 50 males. If the constant term is 0.5, the effect of sex (being male) is 0.1, and the effect of education (being a high-school graduate) is 0.2, then males are 1.11 times (1.11 is the exponentiation of 0.1) more likely to be in the labor force—holding education constant—compared to females. But, in this case, there is no clear intuitive interpretation for an odds ratio of 1.11. How can the use of predicted probabilities assist the interpretation?

We calculate the predicted probability of labor-force participation as follows. For females, we assume that each of the 100 persons is female and has the observed regression coefficient for education. Although we assume that each person has 0.0 for sex because we assume that everyone is female, we note each person's actual education and take into account the estimated effect (0.0 for less than high school and 0.2 for high school). The sum of the values for all 100 persons is calculated and divided by the observed sample size, which is 100 in this case, to yield a predicted probability. In this hypothetical case with an equal number of males and females and an equal distribution of education groups for each sex, the predicted probability would be 0.65 for female labor-force participation, holding education constant. The similar calculation for males produces a predicted probability of 0.70 for male labor-force participation, holding education constant. These two values—65 percent for females and 70 percent for males—offer a useful interpretation of the logit regression coefficients for sex in terms of predicted probabilities, holding all other explanatory variables constant.

In this simple example, the coefficient of 0.1 for sex means that 65 percent of females are predicted to be in the labor force compared to 70 percent of males, holding their observed education constant. In other words, males are predicted to have labor-force participation rates that are 5 percentage points higher than females. Usually, there are more explanatory variables and more categories in some variables. Furthermore, logit regression equations often include continuous variables, which are evaluated using the observed variable value multiplied by the estimated regression coefficient. Fortunately, the calculation of predicted probabilities for binary logit models is easy to achieve. With Stata software, for example, the margin command performs this calculation after estimating a logistic regression model.[4]

Lives in Transition provides a valuable sampling of empirical research by historians using longitudinal data. It improves upon the current understanding of a variety of historical issues by focusing on different aspects and stages of individual life courses and identifying questions for further study. It demonstrates the important empirical challenges and presents a variety of substantive topics for longitudinal analysis. Overall, the chapters show that analysis of longitudinal data illuminates historical studies in new ways, providing insights about the factors affecting changes in individual lives.

4 Stata, *Stata Base Reference Manual: Release 14* (College Station, Texas, 2014).