Quantitative Methods in Population Health

Extensions of Ordinary Regression

Mari Palta

A JOHN WILEY & SONS, INC., PUBLICATION

Quantitative Methods in Population Health

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Peter Bloomfield, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Louise M. Ryan, David W. Scott, Adrian F. M. Smith, Jozef L. Teugels;*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall*

A complete list of the titles in this series appears at the end of this volume.

Quantitative Methods in Population Health

Extensions of Ordinary Regression

Mari Palta

A JOHN WILEY & SONS, INC., PUBLICATION

This book is printed on acid-free paper.♾

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Palta, Mari, 1948–
Quantitative methods in population health: extensions of ordinary regression / Mari Palta.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-45505-9 (cloth)
1. Medical statistics. 2. Regression analysis. 3. Population–Health aspects–Statistical methods. 4. Health surveys–Statistical methods. I. Title.

RA409.P34 2003
614.4′2′0727—dc21

2003050087

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

List of Figures

List of Tables

Contents

Preface

This text arose from my many years working with several long-term population-based observational studies. As I was asked to put together a third-semester statistics course for our new Ph.D. program in Population Health, I decided to assemble the information I had seen investigators and students need most often, and I also decided to answer as many questions as possible out of those I had typically been asked. The resulting mix of topics is guided by this experience. I have attempted to pull in and deal with the aberrations of observational data such as confounding and selection bias. Some traditional topics regarding small sample inference, analysis of variance, and experimental design are deemphasized, as I have found that they confuse rather than help population health researchers. I am using data sets from my own research and collaborations as examples to ensure that subject matter interpretations are meaningful, and that the reader becomes familiar with the "non-textbook" appearance of real data.

While keeping the material immediately applicable by providing detailed instructions for how to run and interpret procedures in SAS, I find it irresponsible to do so without creating some "common sense" about the methods and their assumptions. The beginning chapters lay the mathematical groundwork necessary for topics in later chapters. Whenever possible, I have made a point of inserting practical issues that are answered by specific mathematical derivations.

In addition, each topic starts with an explanation of the theoretical background necessary to allow reasonable judgment as to when the technique is applicable and to facilitate future learning of related methods and software. In the process the reader is exposed to some of the underpinnings of statistics that are often omitted from applied texts and courses. While the text is anchored in the terminology of the biostatistical tradition, I point to some important connections to techniques and terminology used in econometrics and psychometrics. Because of the historic emphasis of biostatistics on experiments and randomization, I have often found that econometric approaches provide further insight in the observational framework.

For progress in addressing current population health issues, it is necessary for researchers in epidemiology and health services to understand and apply regression analysis with weights to deal with unequal variance and correlated and longitudinal outcomes by mixed effects, generalized linear models, and generalized estimating equations. In addition, many data sets in these areas include survey weights. Increasingly, investigators are also called upon to examine the possible impact on their results of missing observations. I suggest straightforward methodology that can be implemented with standard software. The material is presented on a level that will make it accessible to epidemiologists and health services researchers, as well as to applied statisticians. This corresponds roughly to a third-semester applied statistics sequence for statistics non-majors. It is assumed that the reader is already well acquainted with ordinary and logistic regression analysis and has at least rudimentary knowledge of the SAS package.

The explanations are designed to assume as little background in mathematics and statistical theory as possible, except that some knowledge of calculus is necessary for certain parts, such as in understanding maximum likelihood and generalized linear models. The reader may wish to review the rules and uses of derivatives, which are not covered here. On the other hand, all relevant aspects of linear algebra and statistical theory are explained within the text. Important formulas are derived, but with an eye to avoiding excessive algebra.

SAS commands are provided for applying the methods. The SAS procedures emphasized are PROC REG, PROC MIXED, and PROC GENMOD, with occasional references to others. Useful data manipulation commands are introduced as needed to illustrate the techniques in the specific data sets, and the SAS ODS system is briefly introduced to accomplish viewing random effects from mixed models. However, basic commands used to read in data sets and annotate them are considered well known and are not always provided in the text.

Mari Palta
Madison, Wisconsin

Acknowledgments

I have been fortunate to collaborate with wonderful investigators in epidemiology and health services research. They have continually inspired me to look into new statistical issues and have provided me with the insights needed to link the statistics to the subject matter. My longest and closest collaborations have been with Mona Sadek-Badawi in research on the outcomes of very-low-birth-weight neonates, with Kit Allen in diabetes research, and with Terry Young in research in sleep disorders. We have together tackled many challenges that arise in long-term subject follow-up, such as the difficulty in locating subjects and the effects of some participants skipping samples or questionnaires. The research has also interested me, among many other things, in latent variables and in examining between and within individual effects. Many other individuals working on these projects, including Tammy LeCaire and Laurel Finn, have helped me obtain the data for the examples.

Many of the specifics addressed in this book arose from the research of Paul Peppard on the association of sleep apnea with blood pressure. These data form a core example that is traced through the various analytic techniques. I appreciate the many discussions and penetrating questions raised by Paul. Among health services researchers, Maureen Smith has perked my interest in econometric and other social science techniques and has lent me many books. Dennis Fryback shared his data and questions on health care cost analyses.

The book could not have been pulled together without the help of Lin Wang, a Ph.D. student in Statistics who critically reviewed the text, served as a sounding board for methodology, and helped with the formatting of the text. Other former statistics Ph.D. students whose research contributed to the methodology are T.-Y. Yao, Chin-Yu Lin, Wei-Hsiung Chao, Soomin Park, Lei Shen and Liang Li. Many students who have taken the class Quantitative Methods in Population Health, for which this text was initially drafted, have challenged me to explain statistics more clearly and have forced me to consider how different techniques relate to each other and why we do statistics the way we do.

Last but not least, I owe a lot to all the people who have supported me personally and who have pitched in when and wherever needed. My family has been wonderfully accommodating to my obsession with this book. I could not have completed this task without their love and support. In addition, everyone working on the Newborn Lung Project, including Aggie Albanese, Kathleen Madden, and Hana Said and the staff of the Diabetes Registry, make my immediate work environment fun and rewarding.

New methodological developments underlying certain parts of the book were supported by grant CA53786 from the National Cancer Institute and by grant P01 HL42242 from the National Heart Lung and Blood Institute.

M. P.

Acronyms

ANOVA	Analysis of variance
BLUE	Best linear unbiased estimator
GHb	Glycosylated hemoglobin
ML	Maximum likelihood
NICU	Neonatal intensive care unit
NHANES	National health and nutrition examination survey
REML	Restricted maximum likelihood
SAS	Statistical Analysis System®
SBP	Systolic blood pressure
se	standard error
VLBW	Very low birth weight

Introduction

Some Data Sets Used as Examples in This Text

In this book, we focus on extending ordinary regression analysis by considering situations where some of the usual assumptions are violated. As we discuss in more detail in future chapters, violations of assumptions are common in population health data sets. For example, when we want to model presence versus absence of a disease the outcome variable is binary. Because ordinary regression is designed for normally distributed outcomes, the presence of binary outcomes leads to the extension of ordinary regression to logistic regression. We discuss this extension and others that apply to non-normally distributed outcomes in later chapters. With normally distributed outcomes, we encounter situations with violation of equal variance and independence assumptions. For example, subjects may be followed longitudinally, which leads to correlated residuals. The fitting of models and inference in such cases is the topic of the earlier chapters. Throughout, we illustrate with the use of data sets that have accrued from population health research. The purpose of this introduction is to briefly describe these data sets. All analyses were run in SAS 8.2 [1] for Solaris. For graphics we occasionally used SAS 8.2 for Windows.

I.1 NEWBORN LUNG PROJECT

The Newborn Lung Project enrolled a cohort that included all very-low-birth-weight admissions to six neonatal intensive care units in Wisconsin and Iowa during 8/1/88–6/30/91. There were 1040 admissions during this time period, and some baseline data were collected on all of them. Neonatal nurses collected medical record information on factors such as birth weight, supplemental oxygen use the first 24 hours, and hospital of birth without identifiers. Parents were approached as soon as possible for informed consent for interview and medical record abstracting. Due to human subjects concerns, parents were not approached after the neonate had

died or if the neonate was in critical condition. A total of 810 infants survived the hospitalization, and the parents of 633 provided informed consent for abstracting. Recontract addresses were available only for the subgroup with informed consent. By age 5, six additional children had died. Among the 804 survivors, 438 were located, and a follow-up interview including health information and a functional assessment of 422 was performed. The parents of 345 children also gave informed consent for complete abstracting of medical records.

The original purpose of the study was to establish severity scores for neonatal lung disease and to find risk factors associated with it. Later, we described functional and respiratory outcomes at ages 5 and 8 years and their predictors. The record abstracting led to longitudinal data on number of hospitalizations and clinic visits during every year of life.

Examples used in the text arise from data collected during the initial hospitalization and from the follow-up. For example, we analyze functional outcome at age 5 and hospitalizations during the first five years of life as outcomes in regression analysis. We briefly illustrate principal component analysis of socioeconomic indicators collected by this study. We will also use this data set to show how to use the available data on those not followed to examine selection bias.

Some references that present data included in this text and that provide further background on the Newborn Lung Project are listed at the end of this Introduction.

I.2 WISCONSIN DIABETES REGISTRY

The Wisconsin Diabetes Registry targeted all individuals <30 years of age diagnosed with Type I diabetes in a 28-county area in southern Wisconsin 5/1/87–4/30/92. The primary mode of recruitment was by physician, diabetes educator, and self-referral. Also, all hospitals and most multipractice clinics were telephoned every 3 months to ascertain any unreferred cases. A total of 733 cases were found. Out of these, 597 gave informed consent for participation. Participants underwent a baseline interview, were requested to submit blood samples every 4 months, were sent a questionnaire inquiring about hospitalizations and other events every 6 months, and underwent physical examinations at 4, 7, and 9 years of duration. The blood samples were used to determine glycosylated hemoglobin (GHb), an important indicator of glycemic control. The purpose of the study is to map the acute and chronic outcomes of Type 1 diabetes from diagnosis and to consider risk factors such as GHb from the earliest stages of the disease onwards.

Examples used in this text arise from the longitudinal glycosylated hemoglobin measurements performed on the blood samples. Figure I.1 obtained by PROC CHART shows the number of participants with GHb data for 1 year, 2 years, and so on, up to 14 years of follow-up, at the time the data sets for this text were compiled. In most analysis here, we will for simplicity average all the GHb measures in a given year. The commands use to produce the number of years of GHb measurements for each individual were:

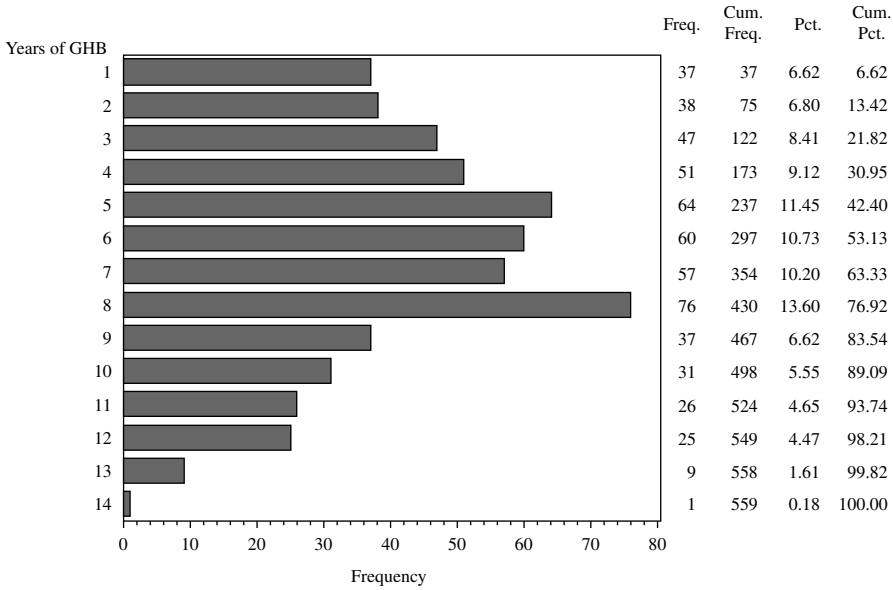

Fig. I.1 The number of participants with GHb

```
PROC SORT; BY ID;
PROC MEANS NOPRINT; BY ID; VAR GHB;
OUTPUT OUT=MM N=NG;
DATA B; SET MM;
LABEL NG='YEARS OF GHB';
PROC CHART; HBAR NG;
```

Some references pertaining to GHb measurements in the Wisconsin Diabetes Registry Study are listed at the end of this Introduction.

I.3 WISCONSIN SLEEP COHORT STUDY

A survey questionnaire inquiring about sleep and sleep-related problems was sent to 6900 employees age 30–60 at four State of Wisconsin agencies. Completed surveys were received from 4927 respondents. A stratified random sample of the respondents was invited to spend the night in a completely equipped clinical sleep laboratory for overnight polysomnography and other tests. A total of 1370 individuals participated. Sleep studies were performed over an extended time period, resulting in some individuals being age 65 and older at the first visit. Blood pressure measurements were taken in the laboratory, and height and weight were measured. Subjects are reinvited for sleep studies every four years. The goals of the project are to identify risk factors and outcomes associated with sleep disorders.

Data used in this text are the longitudinal measures of systolic blood pressure and hypertension as associated with age, gender, and body mass index. Sample sizes in specific analyses vary slightly due to measurements sometimes being missing. For example, out of the 1370 total individuals, 5 had their first blood pressure measurement at the second visit. We also use data from a subproject on the cost of medical care for the cohort. A list of references from the study that involve these variables is provided below.

Suggested Reading

1. Palta M, Gabbert D, Weinstein MR, and Peters ME, Multivariate assessment of traditional risk factors for chronic lung disease in very low birth weight neonates, *Journal of Pediatrics,* **119:**285–292 (1991).
2. Palta M, Weinstein MR, McGuinness G, Gabbert D, Brady W, and Peters ME, A population study: Mortality and morbidity after availability of surfactant therapy, *Archives of Pediatrics and Adolescent Medicine*, **148:**1295–1301 (1994).
3. Palta M, Sadek M, Barnet JH, Evans M, Weinstein MR, McGuinness G, Peters ME, Gabbert D, Fryback D, and Farrell P, Evaluation of criteria for chronic lung disease in very low birth weight infants, *Journal of Pediatrics*, **132:**57–63 (1998).
4. Palta M, Sadek M, Evans M, Weinstein MR, and McGuinness G, Functional assessment of a multicenter VLBW cohort at age 5 years", *Archives of Pediatrics and Adolescent Medicine,* **154:**23–30 (2000).

Wisconsin Diabetes Registry

5. Allen C, Zaccaro D, Palta M, et al., Glycemic control in the first two years of insulin-dependent diabetes mellitus, *Diabetes Care,* **15**:980–987 (1992).
6. Palta M, Shen G, Allen C, Klein R, and D'Alessio D, Longitudinal glycosylated hemoglobin patterns from diagnosis in a population based cohort with IDDM, *American Journal of Epidemiology*, **114:**954–961 (1996).
7. Allen C, LeCaire T, Palta M, Daniels K, Meredith M, and D'Alessio D, Risk factors for frequent and severe hypoglycemia in Type I diabetes, *Diabetes Care,* **24:**1878–1881 (2001).

Wisconsin Sleep Cohort Study

8. Young T, Palta M, Dempsey J, Skatrud J, Weber S, and Badr S, Occurrence of sleep disordered breathing among middle-aged adults, *New England Journal of Medicine*, **328:**69–77 (1993).

9. Hla KM, Young TB, Bidwell T, Palta M, Skatrud J, and Dempsey J, Sleep apnea and hypertension—A population-based study, *Annals Internal Medicine*, **120:**382–388 (1994).
10. Young T, Finn L, Hla M, Morgan BJ, Palta M, Snoring as part of a dose–response relationship between sleep-disordered breathing and blood pressure, *Sleep*, **19** (supplement) (1996).
11. Peppard PE, Young T, Palta M, and Skatrud J, Association between sleep disordered breathing and hypertension, *New England Journal of Medicine*, **342:**1378–1384 (2000).
12. Young T, Peppard P, Palta M, Hla KM, Finn L, Morgan B, and Skatrud J, Population-based study of sleep disordered breathing as a risk factor for hypertension, *Archives of Internal Medicine*, **157:**1746–1752 (1997).

CHAPTER ONE

Review of Ordinary Linear Regression and Its Assumptions

1.1 THE ORDINARY LINEAR REGRESSION EQUATION AND ITS ASSUMPTIONS

A linear regression equation can be alternatively specified as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{or} \\ \mu_{y|x} &= \beta_0 + \beta_1 x \quad \text{or} \\ E(y|x) &= \beta_0 + \beta_1 x \end{aligned} \tag{1.1}$$

to describe the quantitative relationship between a single predictor x and an outcome y. In the population health research projects described in the Introduction, y may be a measured GHb or the score of a very-low-birth-weight (VLBW) child on a test, or the systolic blood pressure (SBP) at a visit to the sleep clinic. In the first equation ϵ_i is a random regression error describing the deviation of a given value y_i from its mean. It can be viewed as capturing unmeasured influence on the outcome. In order to make both the first and the second equations of (1.1) correct, it is assumed that $E(\epsilon_i|x_i) = 0$. In other words, if the second equation is to describe the relationship of the mean y to x correctly, the random errors in the first equation must average to 0 for all x. This also implies that ϵ_i does not depend on x_i. The last two equations are just saying the same thing in different notation because the "expected value" $E(\cdot)$ of a variable is by definition the mean of that variable.

We assume that the reader is familiar with the "conditional on" notation implied by the "|". Conditioning on a variable means that the variable is (at that moment) considered a constant, so the parameters of the distribution of y may depend on x. In other words, when conditioning systolic blood pressure on a given age x, we are interested in the parameters of the distribution of blood pressure at that age. Estimation of the parameters of equations (1.1) usually proceeds by the method

Quantitative Methods in Population Health, by Mari Palta
ISBN 0-471-45505-9

of least squares. In dealing with the regression equation, forming estimators, and drawing inference, we commonly make a number of assumptions:

1.1.1 Straight-Line Relationship

Equation (1.1) implies that x and the mean of y are related in a straight-line fashion. This assumption can be alternatively stated as a constant difference in mean y between every pair of $x's$ that are separated by the same number of steps. For example, if y is systolic blood pressure from visit 1 in the Sleep Cohort Study and x is age, linearity implies that the difference in mean blood pressure between a 50-year-old and a 40-year-old is the same as that between a 40-year-old and a 30-year-old. Regardless of the level of x, $\mu_{y|x+1} - \mu_{y|x} = \beta_1$, so that the regression coefficient is the difference in mean with one step increase in x. Again, if y is systolic blood pressure and age x is recorded in years, β_1 is the increase in mean blood pressure every year. The linearity assumption is an inherent structural assumption, the validity of which is driven by the biological, sociological, and so on, mechanisms that relate y to x. When the linearity assumption holds, we are ahead statistically, because we need to estimate only two parameters β_0 and β_1 instead of a separate $\mu_{y|x}$ for every x.

Only in the situation that x is binary (e.g., designating two treatment groups) is the linearity assumption moot, or automatically satisfied. If y is systolic blood pressure and x is a 0–1 indicator of gender where 1 indicates male, then β_1 is the difference in mean blood pressure between males and females, and β_0 is the mean for females. In this situation, $\mu_{y|x}$ is simply a notation for representing the means of two groups (females and one-step difference involved). Since no assumptions are made on the mean structure, equations (1.1) estimate two parameters either way.

In other situations, the original x may just serve as a label for different groups, such as ethnic categories or treatments. The linearity assumption then makes little sense. However, we can expand (1.1) through the device of binary indicator variables, which bypass the linearity assumption, but again do not save us parameters as compared to estimating $\mu_{y|x}$ separately for each group. In the Wisconsin Sleep Cohort Study, we may wish to compare mean blood pressure between the four state agencies surveyed, by using three indicator variables. In SAS, indicator variables are created in many procedures by the CLASS statement [1].

In the simple cases presented in this chapter, we emphasize linearity of $\mu_{y|x}$ versus a single predictor. We can easily generalize equation (1.1) to more complicated cases by transforming y or x or by adding squared, cubic, and so on, terms in x. Note, however, that even when x or y is transformed or when polynomial terms are added, ordinary regression remains a linear expression of the regression parameters. This simplifies estimation. In Chapter 12, we will consider some situations when the regression equation for the mean is not linear in the parameters.

1.1.1.1 Example

OUTPUT PACKET I shows regression equations, plots of residuals versus predicted values, and mean plots for some variables from the data sets of interest. Later, we