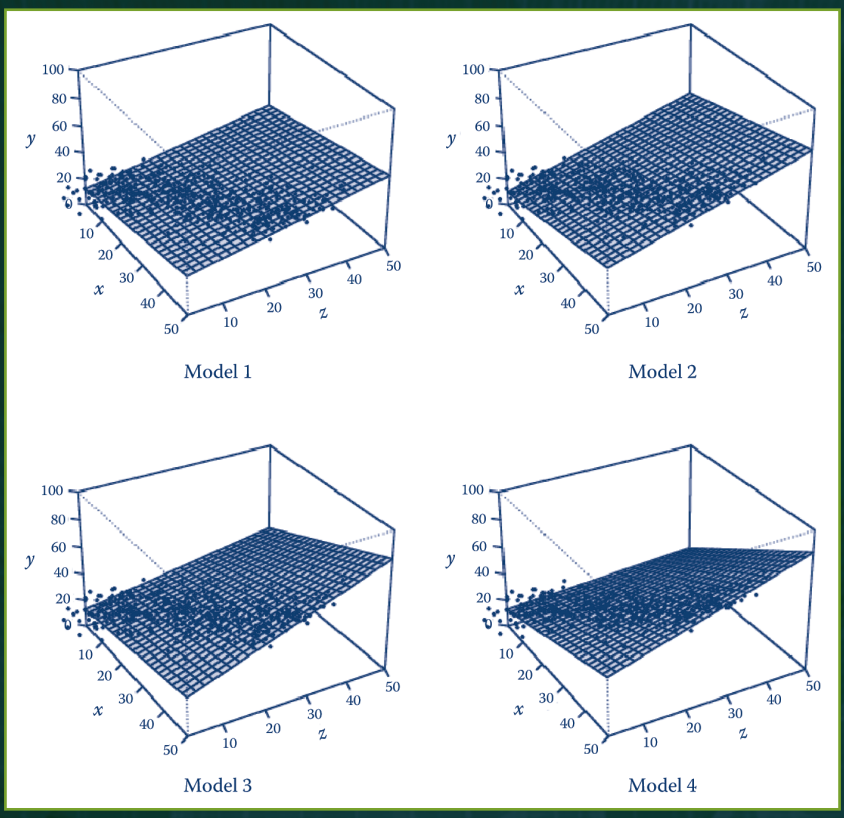
Statistical Thinking in Epidemiology



Yu-Kang Tu
Mark S. Gilthorpe

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Statistical Thinking in Epidemiology

Statistical Thinking in Epidemiology

Yu-Kang Tu
Mark S. Gilthorpe

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20110617

International Standard Book Number-13: 978-1-4200-9992-8 (eBook - PDF)

Contents

Preface

Many books have been written on epidemiological methods. The main difference between them and this book is that we emphasise statistical thinking more than applications of specific statistical methods in epidemiological research, because we believe it is vital to appreciate context and for this to happen, one has to stop and reflect, rather than plough in. This book is therefore not a textbook for the statistical methods commonly used by biostatisticians and epidemiologists; instead, we assume readers have a basic understanding of generalised linear models, such as multiple regression and logistic regression. We do, however, discuss some basic methods in great detail, such as Pearson's correlation and the analysis of covariance, and we show that sometimes it requires a lot of careful thinking to use these simple methods correctly.

We use a few real examples, some of which remain controversial in epidemiological research, to demonstrate our statistical thinking. However, we by no means feel that our thinking is the only approach to the problems we highlight, which are chosen because we believe they have an appeal to general readers. Postgraduate students in biostatistics and epidemiology may use this book as a supplementary reading to standard texts on statistical or epidemiological methods. Lecturers of postgraduate courses may use this book (we hope) as a good example for teaching statistical thinking in epidemiological research. Experienced researchers may find our book both intellectually entertaining and challenging, as some issues discussed are still controversial. We try to show how our statistical thinking of specific research questions develops and eventually leads us to a set of solutions, but our thinking is inevitably framed by our training, knowledge, and experience. For instance, we believe vector geometry is a very useful tool for intuitive understanding of the basic concepts and nuances of linear models, but we acknowledge that not all our readers will agree with us as some may not find thinking geometrically at all intuitive or helpful. Therefore, we welcome feedback from our readers to broaden our vision and improve our thinking, hopefully giving rise to better solutions to those problems discussed in this book.

Our students, collaborators, and colleagues have been very helpful in sharpening and improving our thinking, and the following is an incomplete list of those we would wish to acknowledge and thank for their help in this regard: our colleagues in the Division of Biostatistics and the Division of Epidemiology at the University of Leeds; Professor George Ellison at the London Metropolitan University; Professor David Gunnell, Professor Jonathan Sterne and Dr. Kate Tilling at the University of Bristol; Professor Vibeke Baelum at Aarhus University, Denmark; Dr. Samuel Manda at the

Biostatistics Unit, Medical Research Council, South Africa; and Professor Kuo-Liong Chien at the National Taiwan University, Taiwan. Dr. Chris Metcalfe at the University of Bristol and Dr. Jim Lewsey at Glasgow University kindly acted as external reviewers for our book and gave many useful comments and suggestions. Nevertheless, we take full responsibility for any weakness and errors in our thinking in this book. We would also like to thank Sarah Morris and Rob Calver at Chapman & Hall for their patience with this project.

Part of this book was written when the first author (YKT) was on study leave in Taiwan supported by an international joint project grant from the Royal Society and the National Science Council in Taiwan. In the last 5 years, the first author has been supported by a UK Research Council Fellowship jointly housed by the School of Medicine (Division of Biostatistics) and the Leeds Dental Institute, having enjoyed a large degree of academic freedom in research supported by both the second author (head of the Division of Biostatistics) and Dr. Margaret Kellett (dean of the Leeds Dental Institute). Their generous support is greatly appreciated.

Finally, we would like to thank our wives, Jorin and Amy, and our daughters, Emma and Zarana, for their unconditional love and support.

1

Introduction

1.1 Uses of Statistics in Medicine and Epidemiology

Correlation and regression analyses are among the most commonly used statistical methods in biomedical research. Correlation tests the linear relationship between two (usually continuous) variables, and linear regression tests the relationship between one outcome variable (also known as the dependent variable) and one or more explanatory variables (also known as independent variables or covariates).

As powerful personal computers and graphical user-interface (GUI) statistical software packages have become available and readily affordable in the last decade, complex statistical methods such as multivariable regression become more and more frequently used in medicine and epidemiology, despite involving complex numerical calculations. For instance, to obtain the regression coefficients for a multiple linear regression model with five covariates, one needs to invert a 5×5 matrix, which is very complex and might take days to do by hand. With the power of computers, it now only takes a few seconds to perform these calculations, and all the information relevant (or irrelevant) to the research questions can be shown on the computer screen.

Unfortunately, whilst biomedical researchers may be able to follow instructions in the manuals accompanying the statistical software packages, they do not always have sufficient knowledge to choose the appropriate statistical methods and correctly interpret their results. Many biostatisticians have noticed that the misuses of statistical methods in clinical research are common and the quality of statistical reports in clinical journals needs to be improved (Altman 1991b, 1998, 2002; Andersen 1990). Besides, the increasing use of more advanced and complex statistical methods does not necessarily provide greater insight into the research questions and thereby obtain more reliable knowledge. On the contrary, careless use of these methods can sometimes generate confusing or even misleading results.

This book examines several common methodological and statistical problems in the use of correlation and regression in medical and epidemiological research: mathematical coupling, regression to the mean, collinearity, reversal paradox and statistical interaction. These problems are usually intertwined

with each other. For instance, in the analysis of the relation between change and initial value, mathematical coupling and regression to the mean are almost synonymous; mathematical coupling between explanatory variables usually gives rise to collinearity; and product interaction terms and their component variables are mathematically coupled and raise concerns over collinearity. Since the discussions of these five problems pervade all areas of medical and epidemiological research, it is necessary to study these problems in a framework that focuses on specific research themes. Therefore, this book aims to examine the problems of mathematical coupling and regression to the mean in the analysis of change in pre-test/post-test study designs. The evidence on the foetal origins of adult diseases hypothesis, which sometimes suffers problems of the reversal paradox caused by the adjustment of current body size as confounders, is used to illustrate the potential problems in selecting variables for statistical adjustment and testing statistical interaction.

The overall aim of this book is to explore these statistical problems, to critically evaluate the existing proposed solutions to these problems, and to develop new tools to overcome them. One specific feature of this book is that, wherever applicable, vector geometry is used as a mathematical tool, in novel ways, to illustrate and further develop the concepts behind these statistical methods. The advantage of a geometrical approach over an analytical (algebraic) approach is that geometry can be more readily visualised and may be more intuitive to non-statisticians. Some common uses and misuses of correlation and regression analyses can be more intuitively understood by using geometrical illustrations than using traditional algebraic formula. Moreover, geometry will not only enhance our understanding of statistical problems, but will also generate new insights into old problems, thereby paving the way to provide guidance as to how to avoid statistical errors or even to suggest solutions.

1.2 Structure and Objectives of This Book

This book begins in Chapter 2 with a concise introduction to the concepts underlying the uses of vector geometry in linear models. Although vector geometry has been shown to be a great tool for understanding statistical methods based on ordinary least squares techniques, discussions of vector geometry are, however, scattered in the statistical literature only, and none appear in the clinical literature. So the objective of Chapter 2 is to bring together the various elementary results in vector geometry and establish a set of basic geometric tools that will be utilised throughout this book.

In Chapter 3, we introduce the concepts of directed acyclic graphs (DAGs). DAGs have become popular in epidemiological research in recent years as it

provides a very useful tool for identifying potential confounders. It is very similar to path diagram for structural equation modelling, which is widely used in social sciences research for testing causal models.

Chapter 4 reviews a controversy in analysing the relation between change and initial value caused by mathematical coupling and regression to the mean. In the analyses of pre-test/post-test study design, it is difficult to distinguish between regression to the mean and mathematical coupling, and this difficulty has given rise to confusion amongst some statisticians surrounding how to solve the problem. The objectives of Chapter 4 are to demonstrate the inadequacy of current practice in clinical research, to correct a misconception amongst some statisticians and to review various statistical methods proposed to test the relation between change and initial value.

Chapter 5 explores the differences in statistical power and effect estimations between several statistical methods to analyse changes in the pre-test/post-test study design for both randomised and non-randomised studies. The objectives of Chapter 5 are first to illustrate, using vector geometry, the controversy about the adjustment of initial values, known as *Lord's paradox*, and to explain why and when this paradox arises. This is particularly pertinent to epidemiology since Lord's paradox arises where proper randomisation is not always feasible—as within most epidemiological research. Vector geometry and simulations are then used to show that the prevalent concept that analysis of covariance (ANCOVA) *always* has greater statistical power than other univariate methods is not strictly correct; only when sample sizes are large and/or the correlation between pre-test and post-test values are moderate or high does ANCOVA have greater statistical power than other univariate and multivariate methods.

The objective of Chapter 6 is to demonstrate that vector geometry can give new insights into the problems of collinearity and the detection of these problems in epidemiological research. Collinearity is especially important in observational studies, where the underlying premise of statistical adjustment in regression models is to control simultaneously for correlations amongst covariates and their correlations with the outcome. Vector geometry can elegantly show why principal component analysis, a commonly recommended solution to collinearity, is not always desirable.

Chapter 7 discusses the problem of the reversal paradox, which is perhaps better known in epidemiology as *Simpson's paradox* within categorical variable analysis. In regression analysis, or analysis of covariance, this phenomenon is known as Lord's paradox, and in general linear modelling, a widely discussed statistical paradox known as *suppression* or *enhancement* is another manifestation of the reversal paradox. The objectives of Chapter 7 are to use the foetal origins of adult diseases hypothesis as an example to show how and why the adjustment of current body size measures, such as current body

weight, can give rise to the reversal paradox in multiple linear regression. Vector geometry is used to illustrate the reversal paradox and question an alternative interpretation often proffered, namely, that catch-up growth has a greater impact than birth weight on health in later life. Computer simulations and evidence from empirical studies are then used to show that the negative relationships between birth weight and blood pressure are strengthened and positive relationships between birth weight and blood pressure are attenuated or even reversed after adjusting for one or more current body sizes.

Chapter 8 examines the role of statistical interaction within regression analyses by revisiting the four-model principle proposed in the foetal origins hypothesis literature. The objectives of this chapter are first to show that when three continuous variables, such as blood pressure, birth weight and current body weight, follow multivariate normality, the expected value of the partial regression coefficient for the product interaction term in the multiple regression model is zero. Computer simulations are used to show that categorising birth weight and/or current body weight, a common practice in epidemiological studies on the foetal origins hypothesis (and many epidemiological studies in general), gives rise to spurious interaction effects and can therefore potentially lead to seriously misleading conclusions.

Chapter 9 reviews the recent advances in identifying critical growth phases for health outcomes in later life. In recent years, some researchers' interests have shifted from birth size to the growth in body size during early childhood. In Chapter 9 we provide a concise introduction to those methods for identifying the critical growth phases and use a publicly available data set to illustrate and compare results. This chapter shows that whilst different statistical methods have been proposed to test the same hypothesis, the differences in those approaches seem to indicate that researchers may have different versions of the same hypothesis in mind. As statistical thinking is guided by the research hypothesis, differences in theorising and framing the research questions will inevitably lead to different thinkings.

After discussing the pros and cons of the proposed methods for identifying the critical growth phases within Chapter 9, we propose a new approach in Chapter 10 that uses partial least squares (PLS). PLS is widely used in chemometrics and bioinformatics, though it is rarely used in medical and epidemiological research. One apparent advantage of PLS is that it can deal with perfect collinearity that can arise due to the mathematical relationships amongst covariates. In Chapter 10 we use the same data set as in Chapter 9 to illustrate how PLS can be applied to lifecourse research.

Most of the data sets used in this book have been drawn from published literature that are easily obtainable. One data set was kindly made available by Dr. Lars Laurell.

1.3 Nomenclature in This Book

Italic capital letters (X) denote variables without measurement error.

Italic lower-case letters (x) denote observed variables measured with error.

Bold lower-case letters ($\mathbf{x}$) denote vectors.

Bold capital letters ($\mathbf{X}$) denote matrices.

Capital letters (X) with subscripts denote components of variables or vectors.

Subscripts of variables usually represent repeated observations on the same variables, though they may also represent levels within a multilevel framework.

1.4 Glossary

Due to our backgrounds and a number of extensive collaborations with dental researchers, we sometimes use examples from the dental literature to illustrate methodological problems. Consequently, non-dental readers might not be familiar with some of the terminology that frequently appears in some chapters. This section provides a non-technical explanation and illustration (Figure 1.1a and b) of the dental terminology.

Clinical attachment level (CAL): The distance between the cemento-enamel junction and the bottom of periodontal pocket. It is not common to measure CAL directly using a periodontal probe; instead, CAL is usually the sum of gingival recession (GR) and probing pocket depth (PPD).

Enamel matrix proteins derivatives (EMD): Extracts from developing teeth in young pigs; these mixtures of proteins are believed to be able to promote regeneration of lost periodontal tissue. A syringe is used to release EMD into periodontal infrabony defects.

Gingival recession (GR): The distance between the cemento-enamel junctions and the gingival margin. In a healthy condition, the level of gingival margin (i.e., gums) is assumed to be level with the cemento-enamel junction and, therefore, GR is zero. However, due to external trauma, periodontal diseases or aging, the gingival margins may recede; that is, move in the direction of the root apex, and GR

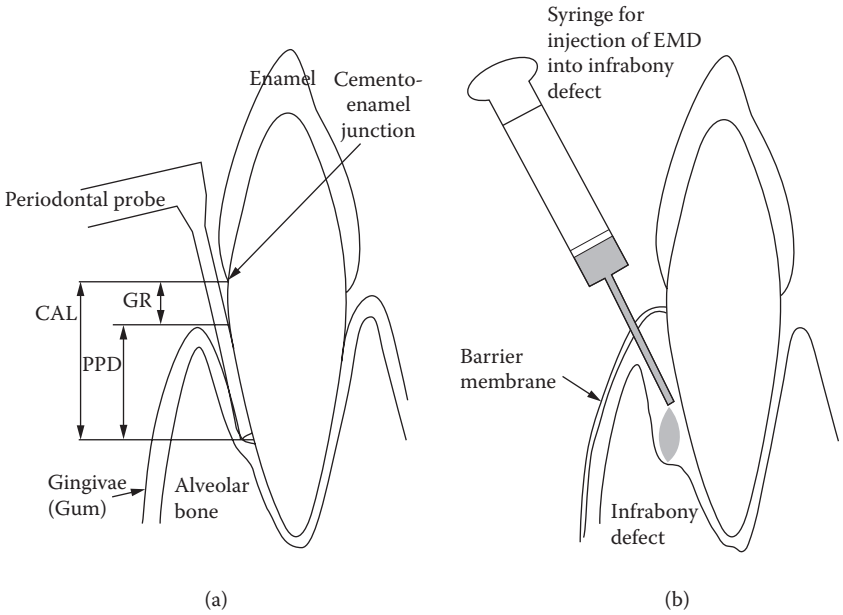


FIGURE 1.1
Illustration of the glossary. (a) Measurement of periodontal pocket depths; (b) application EMD in periodontal surgery.

is positive. In some circumstances, the gingivae (i.e., gums) might cover part of the crown of the tooth and the cemento-enamel junction becomes invisible, in which case GR is negative.

Guided tissue regeneration (GTR): A surgical procedure to regenerate lost periodontal tissue, especially in periodontal infrabony defects and furcation involvement in molars. A barrier membrane is placed around the tooth to cover the defect in order to allow progenitor cells from alveolar bone and/or periodontal membrane to repopulate on the root surface.

Probing (pocket) depth (PD): The distance between the gingival margin and the bottom of the periodontal pockets. Both PD and GR are usually measured with a periodontal probe with markings in millimetres. In this book, probing pocket depth, probing depth or pocket depth are all equivalent and used interchangeably.

2

Vector Geometry of Linear Models for Epidemiologists

2.1 Introduction

Vector geometry was first used in 1915 (Fisher 1915) by the great statistician Sir Ronald Fisher to address the problem of deriving the statistical distribution of the Pearson product–moment correlation coefficient. Fisher used his great geometric imagination to demonstrate that the correlation between two variables is the cosine function of the angle between two vectors in an n-dimensional space. However, as most statisticians do not have the same great geometrical insights as Fisher did, and most statisticians consider an algebraic approach more mathematically rigorous, the geometric approach has not thus far received as much attention as it perhaps deserves (Herr 1980). Nevertheless, some teachers of statistics have found a geometric approach to be more intuitive for people without a mathematical background and a very useful tool to develop the understanding of linear regression analyses. A few statistical textbooks (Saville and Wood 1991, 1996; Wickens 1995; Carroll et al. 1997) are written using mainly vector geometry, and some textbooks of statistics or econometrics on linear statistical models (Wonnacott and Wonnacott 1979, 1981; Fox 1997; Draper and Smith 1998) have chapters on vector geometry. However, vector geometry seems to have been used rarely to illustrate or explore specific methodological problems within biomedical research.

The aim of this chapter is to provide an introduction to the various concepts of vector geometry within correlation and regression analyses. Vector geometry, as a mathematical tool, will be consistently and extensively explored in later chapters.

2.2 Basic Concepts of Vector Geometry in Statistics

The scatter plot, one of the most commonly used graphs to display the relationship between two continuous variables, can be viewed as a form of

geometry. In the scatter plot for two random variables, *X* and *Y*, each with *n* independent observations ($X_1 \ldots X_n$) and ($Y_1 \ldots Y_n$), there will be *n* points in *two*-dimensional space (i.e., on a plane). This is known as *variable space* geometry. The axes represent the variables *X* and *Y*, and the points are the observations made on each subject. In contrast, instead of using variables as axes, the same data can be displayed in what is termed *subject space* by using subjects as the axes; the variables *X* and *Y* are then two points in *n*-dimensional space. By drawing an arrow from the origin to each point, *X* and *Y* become two vectors with coordinates ($X_1 \ldots X_n$) and ($Y_1 \ldots Y_n$) in *n*-dimensional space (Wickens 1995; Fox 1997). Although it is impossible to visualise *n*-dimensional space, only a plane is required to visualise the relative relationship between the vectors representing *X* and *Y* in *n*-dimensional space by projecting onto a plane and effectively dropping the original axes. In general, the number of dimensions needed to draw the graph in subject space is no greater than the number of variables (Wickens 1995).

We use a numerical example to illustrate the difference in data presentation between variable-space (scatter plot) geometry and subject-space (vector) geometry. Suppose two men, Mr A and Mr B, have their body size measured: the heights of A and B are 180 cm and 170 cm, respectively, and their weights are 90 kg and 60 kg, respectively. When the data are presented in a scatter plot (i.e., variable-space), subjects A and B are two points in a two-dimensional plot, and the axes are *Height* and *Weight* (Figure 2.1a). When the same data are presented in vector geometry (subject space), *Height* and *Weight* become two vectors in a two-dimensional space, and subjects A and B are the axes (Figure 2.1b).

In Figure 2.1b, the two variables *Height* and *Weight* are presented in their raw data format. However, in general, it is very useful to represent the

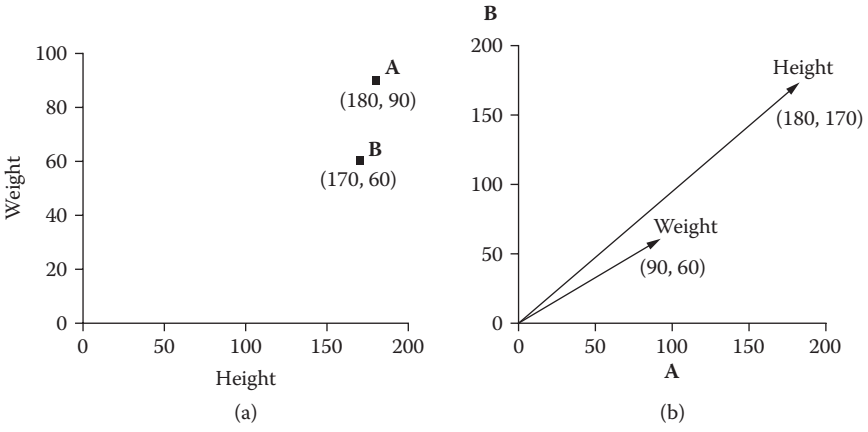


FIGURE 2.1
Illustrations of variable-space (a) and subject-space (b) geometry.