國立交通大學

統計學研究所

碩士論文

有關近視眼研究的統計方法

Statistical Analysis for Studying the Progression of Myopia

研究生：傅驛爲

指導教授：王維菁 教授

中華民國一○三年六月

有關近視眼研究的統計方法

Statistical Analysis for Studying the Progression of Myopia

研究生：傅驛爲　　　　Student：Yi-Wei Fu

指導教授：王維菁 教授　　　　Advisor：Dr. Wei-jing Wang

國立交通大學

統計學研究所

碩士論文

A Thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
June 2014
Hsinchu, Taiwan, Republic of China

中華民國一○三年六月

有關近視眼研究的統計方法

研究生：傅驛爲　　　　　　　　　　指導教授：王維菁 教授

國立交通大學理學院

統計學研究所

摘要

在科技進步的現代社會，近視已是日趨嚴重的公衛問題，不僅盛行率增加，發生年齡亦下降，其中又以華人社會最為嚴重。同時眼科醫學亦著力於近視的預防與治療。在近視眼的研究中，統計分析是不可或缺的角色。本論文整理了有助於近視研究的統計方法，內容包含：最常見的橫斷面研究，可找出近視的風險因子；利用分析長期追蹤資料，探討近視度數演進的變化。此外眼科測度為“成對”但彼此相關的資料，存活分析可以用來探討眼睛發生重要“事件”所需的時間長度。我們亦回顧了台灣、新加坡及其他地區所發表有關於近視眼研究的醫學文獻，並整理這些論文使用的統計方法。最後我們以統計的角度，建議了未來研究可採用的方法。

關鍵詞：盛行率、橫斷面研究、長期追蹤研究、存活分析

Statistical Analysis for Studying the Progression of Myopia

Student：Yi-Wei Fu Advisor：Dr. Wei-jing Wang

Institute of Statistics

National Chiao Tung University

ABSTRACT

Nowadays myopia has become a global critical health problem, especially in Chinese community. Statistical analysis plays an important role in myopia research. In the thesis, we investigate related statistical methods, including cross-sectional studies for identifying possible risk factors for myopia and longitudinal studies for studying the evolution of myopia. Eye measurements are "paired" observations so that we review methods for analyzing such data. Furthermore survival analysis can also be applied to study critical event times such as the onset of myopia. We also review medical literature on myopia, most of which are recent empirical studies in Taiwan and Singapore. At the end of the thesis, we provide some suggestions for future research from the viewpoint of statistical analysis.

Keywords: cross-sectional study, longitudinal study, prevalence rate, survival analysis

誌　　謝

本論文可以順利的完成，首先我要感謝王維菁教授，在與老師作研究的過程中，老師教導了我如何從統計的角度來看待科學問題，進而培養正確的研究態度及方法，並適時地提醒我如何將龐大的資料作整合，接著以簡單明瞭的方式呈現出來是相當重要的能力，非常感謝老師這一年多來的教誨；也感謝所有統計所上所有的教授這兩年來的指導，讓我看到統計的不同面向並學習到許多統計工具；感謝郭姐及怡君姐，時常給予生活上或學業上的建議和叮嚀，並協助處理設備的問題；最後感謝研究室裡和我一起努力的同學們，除了在課業上的交流討論，也經常給予我精神上的鼓勵。

本論文的完成同時也代表著我將踏入人生的另一個階段，期許能將所學發揮在自己的領域，並保持自己在研究所生涯中培養的統計思維及熱情。希望能將完成此論文的滿足和快樂分享給所有曾經幫助和關心我的人。

傅驛爲 謹誌

2014年 6月

于交通大學統計學研究所

Contents

Tables

Figures

Chapter 1: Introduction

1.1 Motivation

Myopia affects many school-aged children nowadays. As technology advances in a very fast speed, fancy electronic products are becoming more and more popular for children. At the same time, the prevalence of myopia increases while its onset age decreases. Thus myopia prevention and control have become important health issues. In the thesis, we will examine statistical methods which can be used in myopia research and hope that our analysis can help clinicians to adopt suitable methods for analyzing their datasets.

The following descriptions about myopia are summarized from Wikipedia. *Myopia, commonly known as being "nearsighted" or "shortsighted", is a condition of the eye where the light that comes in does not directly focus on the retina but in front of it. As a result, the image that one sees when looking at a distant object to be out of focus, but in focus when looking at a close object. Youth* ***onset myopia*** *occurs in the early childhood or teenage, and the ocular power can keep varying until the age of 21. In some parts of Asia, myopia is very common. Singapore is believed to have the highest prevalence of myopia in the world; up to 80% of people there have myopia, but the accurate figure is unknown.*

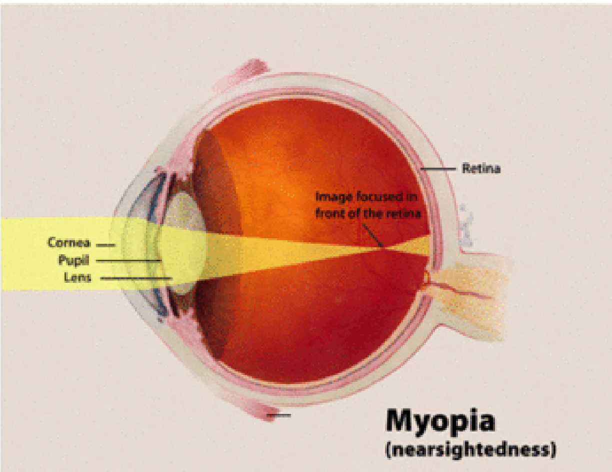

Figure 1: The phenomenon of Myopia

Data collected for myopia studies often involve complicated data structures and therefore provide an abundant area for statistical applications. Besides studies based on cross-sectional data, the development of myopia can be investigated via longitudinal follow-up. The problems of interests such as the onset age, the gap time between two clinical stages … etc. can be formulated under the framework of survival analysis. Ocular measurements are paired observations. To use the information of one eye or both eyes requires special attentions. In the thesis, we will first review the related methodology in separate domains. Then we will review existing medical literature on myopia most of which, however, use only elementary statistical methods. After presenting all the methods and showing some examples, we will give our suggestions.

1.2 Outline of the Thesis

Here is the outline of the thesis. In the second chapter, we introduce some medical background for myopia studies. In Chapter 3, we review statistical methods developed in separate areas including longitudinal data analysis and survival analysis and methods for analyzing paired data. In Chapter 4, we review some medical papers on studies of myopia. Chapter 5 contains our suggestions for the future study.

Chapter 2: Medical background on Myopia

2.1 Introduction of Myopia - Medical Background

Myopia is a type of refractive error. Refraction is the ability of crystalline lens that refractes light to focus on retina. If the focal point lies in front of the retina, it’s called myopia which is shown in Figure 1. The common medical definition of myopia is based on spherical equivalent to be at least -0.5D. According to spherical equivalent, children’s myopia can be classified as higher myopia (SE≦-3.0 diopters) and lower myopia (-3.0<SE≦-0.5 diopters). In myopia studies, researchers may also collect other biological measures, including cycloplegic autorefraction, keratometry, axial eye length, anterior chamber depth, crystalline lens thickness and vitreous chamber depth … etc. These measures provide useful information for assessing the conditions of eyes. Many studies focus on investigating the relationships between myopia and these measures.

It has been known that common risk factors of myopia include race, heredity and environment influences. In respect of race, French AM et al. (2013) discover that the chance of Asian children to have myopia (6.9% to 7.3%) is higher than Caucasian children (1.3% to 2.9%) in Australia. In addition, the incidence of myopia for 5 to 12 years old children is 14.1% in Shanghai and that for 5 to 16 years old children in Hong Kong is 14.4%. Jones et al. (2007) found that parents with myopia and how much time that children spend in sports and outdoors are the risk factors. Technology advancement has made myopia become an increasingly important health issue for children. According to the study of Lin et al. (2004), the mean ocular refraction for children in Taiwan began to progress to a myopic condition at the age of 11 in 1983 and increased to 8 in 2000

Once children had myopia, how to avoid myopia deterioration becomes the next issue.

Currently, myopia treatments including drugs, phototherapy and refractive surgery. Cycloplegic and ocular hypotensive agents are most commonly used in the drug treatment. Atropine is the most effective drug to avoid myopia deterioration. Phototherapy is the treatment that children wear glasses to correct refraction error. Refractive surgery is used to correct the refraction error and only suitable for adults above18 years old whose progression of myopia has ended.

Chapter 3: Literature Review on Related Statistical Methods

In this chapter, we review related statistical methods which can be applied to myopia studies.

3.1 Methods for Cross-sectional Studies

For cross-sectional studies, data are collected at one specific point in time. One goal of cross-sectional studies is to compare individuals with different characteristics at the same time. One example in myopia research is to investigate which factors, such as gender, race, nearwork and age … etc., can explain the condition of myopia for different children at the time of data collection.

Let (Y, Z) be the response and covariate vectors respectively. Many statistical methods have been developed to study how Z affects Y. If Y is a numerical variable with the range on the whole real line, the following linear model is often chosen to depict the influence of Z on Y:

$$Y = \beta^T Z + \varepsilon,$$

where ε is the error variable. The model states that for people with different values of Z, their expected values of Y, at the time of data collection, equals $\beta^T Z$. Generalized linear models provide a more flexible setting that allows the distribution of Y to be more flexible. Let $\mu = E(Y)$. The model assumes that after the transformation of $g(.)$, which is called the link function, the effect becomes linear with $g(\mu) = \beta^T Z$. Usually $g(.)$ is a mapping from the domain of μ to $(-\infty, \infty)$.

When Y is a binary variable, the logistic regression model is often chosen such that:

$$\log\left(\frac{\Pr(Y=1|Z)}{\Pr(Y=0|Z)}\right) = Z^T \beta,$$

Where $E(Y) = \Pr(Y=1) = \mu$ and $g(\mu) = \log(\frac{\mu}{1-\mu})$ is the log odds function. In other words,

$$\Pr(Y=1\,|\,Z)=\frac{\exp(Z^T\beta)}{1+\exp(Z^T\beta)}.$$

When Y is a categorical variable, taking K possible outcomes, say $0,\ldots,K-1$. The multinomial logistic regression model may be useful:

$$\Pr(Y=k)=\frac{\exp(Z^T\beta_k)}{1+\sum_{k=1}^{K-1}\exp(Z^T\beta_k)}$$

where $k=1,\ldots,K-1$ and $\Pr(Y=0)=\dfrac{1}{1+\sum_{k=1}^{K-1}\exp(Z^T\beta_k)}$.

ROC curves can also be used to evaluate the importance of risk factors on predicting the occurrence of the disease. Suppose Y indicates whether a subject has the disease or not and Z is a covariate taking continuous (numerical) value. Denote $\Delta(z)=I(Z\le z)$ and assume that the larger value of Z is associated with higher chance of having the disease. If we choose a cut-off point based on Z as a way to classify the result as being "positive" ($Z\ge z$) or "negative" ($Z<z$). Consider the following two-by-two table, where "TP", "FP", "FN" and "FN" represent "true positive", "false positive", "false negative" and "true negative" respectively.

		True condition	
		Y=1	Y=0
Diagnosis result	$\Delta(z)=1$	TP= $\Pr(Y=1,\Delta(z)=1)$	FP= $\Pr(Y=0,\Delta(z)=1)$
	$\Delta(z)=0$	FN= $\Pr(Y=1,\Delta(z)=0)$	TN= $\Pr(Y=0,\Delta(z)=0)$

Then define TPR and FRR as

$$\text{Sensitivity(z)}=TPR(z)=\frac{\Pr(Y=1,\Delta(z)=1)}{\Pr(Y=1)}$$

$$Specificity(z)=1-FPR(z)=\frac{\Pr(Y=0,\Delta(z)=0)}{\Pr(Y=0)}.$$

In other words, “sensitivity” is the probability that a man with the disease is correctly diagnosed; while “specificity” is the probability that a man without the disease is correctly diagnosed.

A useful test should have “high sensitivity and high specificity”, which is equivalent to “high TPR and low FPR”. An ROC curve plots $(FPR(z), TPR(z))$ in the (X, Y) axes based on many points of z. In many studies, the main purpose is to choose a cut-off point that makes the location of $(FPR(z), TPR(z))$ to be most near the upper left corner. This corresponds to choose a point $z*$ based on Z which gives the largest AUC, defined as the area under curve (AUC). Thus AUC becomes an important parameter with the value between 0 and 1 to evaluate the performance of a diagnostic test. Higher value of AUC means that the test has higher discrimination ability to diagnose this disease. The following figure shows the ROC curve for a diagnosis test and the 45 degree line is the curve by random guessing. The table below lists the range of AUC as a rule to evaluate the test.

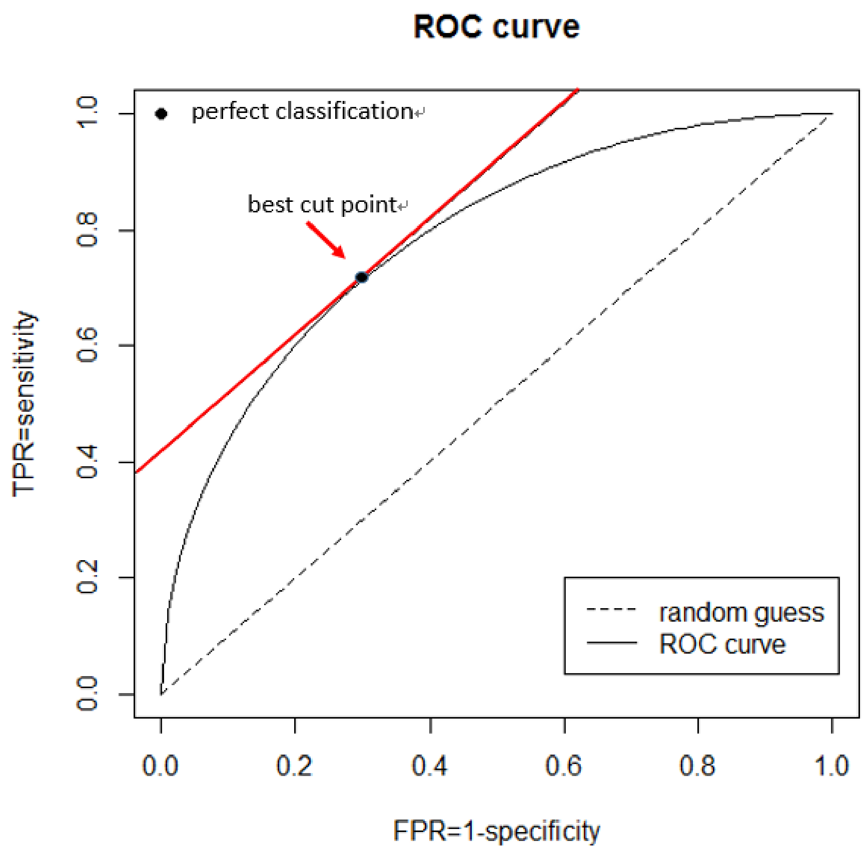

Figure 2: ROC curve and the best cut-off point

AUC=0.5	no discrimination
0.8>AUC≧0.7	acceptable discrimination
AUC≧0.8	excellent discrimination

Table 1: General criteria of AUC

3.2 Methods for Longitudinal Studies

For longitudinal studies, data are collected over a period of time. The main purpose is to study the progress of the response variable along the time and how the change of covariates affects the progression over the period. Thus the cost longitudinal studies is higher compared with cross-sectional studies. Longitudinal variables can be denoted as (Y_j, Z_j, t_j) which represent the response variable and the vector of covariates collected at time t_j respectively for $j = 1, ..., n$ where $t_1 < ... < t_n$.

If the time points are equally spaced which means that the influence of $(t_1, ..., t_n)$ can be ignored, one can consider the following model

$$Y_j = \beta_C Z_1 + \beta_L (Z_j - Z_1) + \varepsilon_j ,$$

where β_C describes the baseline effect while β_L measures the effect of covariate change on the same person's expected response value. It is equivalent to write

$$Y_j - Y_1 = \beta_L (Z_j - Z_1) + (\varepsilon_j - \varepsilon_1) .$$

With the data $\{(Y_{ij}, Z_{ij}, t_{ij})(i = 1, ..., m; j = 1, ..., n_i)\}$, there are two helpful scatterplots. Specifically we can plot $\{(Z_{i1}, Y_{i1})(i = 1, ..., m)\}$ in the (X,Y) axis which reveals the information of β_C and plot $\{(Z_{ij} - Z_{i1}, Y_{ij} - Y_{i1})\}$ which reflects the information of β_L .

Note that $(Y_{i1},...,Y_{in_i})$ are correlated. In the simplified situation with $n_i = n$, the dependence structure can be investigated via plotting r_{ij} versus r_{ik} $(k \neq j)$, where r_{ij} is the residual

$$r_{ij} = Y_{ij} - [\hat{\beta}_C Z_{i1} + \hat{\beta}_L (Z_{ij} - Z_{i1})]$$

for $j = 1,...,n$ with $(\hat{\beta}_C, \hat{\beta}_L)$ being the fitted values. There are $\binom{n-1}{2}$ such plots and we can pay attention to examine whether the association changes with respect to the change of $|k-j|$. The plots reveal information about the structure of the variance-covariance matrix of $(Y_{i1},...,Y_{in})$ $V = [\sigma_{jk}]: n \times n$, where $\sigma_{jk} = Cov(Y_{ij}, Y_{ik})$.

Note that β_L evaluates the "average" effect for a group of individuals with the same degree of covariate change on the change of response. If the focus is on a single individual, say the i-th observation at time t_j, we may consider

$$Y_{ij} = z_{ij}^T \beta + \gamma_i + \varepsilon_{ij}$$

where β describes the population average effect and γ_i is the person-specific effect which is usually assumed to follow a mean normal distribution with the variance explaining the magnitude of heterogeneity.

3.3 Survival Analysis

In survival analysis, the response variable of interest is the time, from a given starting point, to an event of interest. Denote the event time as T. The probabilistic behavior of T can be summarized by the survival function

$$S(t) = \Pr(T > t) = 1 - F(t) = \int_{u=t}^{\infty} f(u)du$$

where $f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$ is the density function, or the hazard function

$$\lambda(t) = \lim_{\Delta \to 0} \frac{\Pr(T \in [t, t+\Delta] \mid T \geq t)}{\Delta} = \frac{f(t)}{S(t^-)} = -\frac{d \log S(t)}{dt},$$

where $\log(.)$ here denotes the natural logarithm function. There is a one-to-one relationship between the two functions such that

$$S(t) = \exp(-\Lambda(t))$$

where $\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t)$ is the cumulative hazards function.

There are two popular regression models to describe how covariate Z affects the failure time T. The most popular choice is the Cox proportional hazards model which can be written as

$$\lambda_Z(t) = \lambda_0(t)\exp(Z^T\beta),$$

where $\lambda_0(t)$ is the baseline hazard function whose form is usually unspecified. Another common option is the AFT (accelerated failure time) model which can be written as

$$\log T = Z^T\beta + \varepsilon,$$

where ε is the error variable whose distribution is not specified. Note that the AFT model can also be written in terms of the hazard function

$$\lambda(t) = \lambda_0\{t\exp(-Z^T\beta)\}\exp(-Z^T\beta).$$

If survival data are obtained from a longitudinal study, some covariates may become time-dependent. The original form of the Cox model assumes that the influence of Z on the hazard is a constant. However in a longitudinal study, values of some covariates such as blood pressure or biochemical measures may change over time. To include time-dependent covariates, the model can be modified as

$$\lambda(t) = \lambda_0(t)\exp(Z^T(t)\beta).$$

The most special feature of survival analysis, compared with other areas of statistics, is the so-called "censoring" problem which happens when the occurrence of the event is not be accurately observed. There are several types of censoring. Here we introduce right censoring and interval censoring.

For right censoring, define C as the censoring time and one observes $X = \min(T, C)$ and $\delta = I(T \le C)$. Usually it is assumed that T and C are independent. Observed data can be written as $\{(X_i, \delta_i, Z_i)(i = 1, ..., n)\}$. The likelihood for right censored data can be written as

$$L(\theta) = \prod_{i=1}^{m} \{f_\theta(x_i)\Pr(C_i > x_i)\}^{\delta_i}\{S_\theta(x_i)g(x_i)\}^{1-\delta_i}$$

where $\theta: p \times 1$ is the parameter of interest and $g(.)$ is the density function of C. If C does not carry any information of θ, it becomes

$$L(\theta) \propto \prod_{i=1}^{n} f_\theta(x_i)^{\delta_i} S_\theta(x_i)^{1-\delta_i} = \prod_{i=1}^{n} \lambda_\theta(x_i)^{\delta_i} S_\theta(x_i) = \prod_{i=1}^{n} \lambda_\theta(x_i)^{\delta_i} \exp\{-\int_0^{x_i} \lambda_\theta(u)du\}.$$

When the density form of T is not specified, $S(t)$ can be estimated nonparametrically by the Kaplan-Meier estimator:

$$\hat{S}(t) = \prod_{u \le t} \left\{ 1 - \frac{\sum_{i=1}^{n} I(X_i = u, \delta_i = 1)}{\sum_{i=1}^{n} I(X_i \ge u)} \right\}.$$

In the presence of covariates, data can be written as $\{(X_i, \delta_i, Z_i)(i = 1, ..., n)\}$. Under the Cox model assumption $\lambda_Z(t) = \lambda_0(t)\exp(Z^T\beta)$ without specifying the form of $\lambda_0(t)$, the partial likelihood for β can be written as

$$L(\beta) = \prod_{i:\, \delta_i = 1} \frac{\exp(Z_i^T\beta)}{\sum_{j: X_j \ge X_i} \exp(Z_i^T\beta)}.$$

In the case of time-dependent covariates, the ideal data set structure be written as $\{(X_i, \delta_i, Z_i(s): 0 \le s \le X_i)(i = 1, ..., n)\}$ and the resulting score function for β can be written as

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left(Z_i(X_i) - \frac{\sum_{j: X_j \ge X_i} Z_j^T(X_i)\exp(Z_j^T(X_i)\beta)}{\sum_{j: X_j \ge X_i} \exp(Z_j^T(X_i)\beta)} \right).$$

However in practice it may not be possible to obtain the whole process of $Z_i(s)$ for $0 \le s \le X_i$. In some statistical software such as SAS or Splus, the nearest data from same individual may be used. Sometimes smoothing methods, such as kernel smoothing, may be adopted to impute the missing information.

Survival analysis can be adopted either for cross-sectional data or longitudinal data. In particular for longitudinal follow-up, observations are often collected at consecutive and distinct time points. It may happen that the exact event time may never be observed. Interval censoring occurs when we only have the information that T lies in an interval between two measurement times. Observed data can be written as $\{(L_i, R_i)(i=1,...,n)\}$ and we know that $T_i \in (L_i, R_i]$. The likelihood function can be written as

$$L(\theta) = \prod_{i=1}^{n} F_\theta(R_i) - F_\theta(L_i).$$

Unlike the Kaplan-Meier estimator suitable for right censored data, the nonparametric estimator for $S(t)$ under interval censoring has no explicit form (Turnbull,1974 and 1976). It has been shown that the nonparametric MLE can be obtained by solving the following self-consistency equation. The first step is to create grid intervals by rearranging $\{(L_i, R_i](i=1,...,n)\}$ in an ascending order and identifying the intervals such that a left-endpoint and a right-endpoint are adjacent to each other. Then denote the disjoint intervals as $\{(l_j, r_j]\ (j=1,...,m)\}$. The following figure depicts the construction of such intervals. Define the indicator $\delta_{ij} = \mathbf{I}\{(L_i, R_i) \subset (l_j, r_j]\}$ which shows whether the ith observed interval overlaps with the jth interval. Let p_j be the mass in $(l_j, r_j]$ which can be estimated by solving the following equation

$$p_j = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_{ij}p_j}{\delta_{i1}p_1 + ... + \delta_{im}p_m} = \frac{1}{n}\sum_{i=1}^{n} w_{ij} \quad (j = 1,...,m)$$

where w_{ij} measures the contribution of the ith observation for estimating the probability of interval $\Pr(T_i \subset (l_j, r_j])$ $(\sum_{j=1}^{m} w_{ij} = 1)$. In the figure, the weights for $(L_1, R_1]$ in estimating $\hat{P}_1, \hat{P}_2, \hat{P}_3$ are $P_1/(1\cdot P_1 + 1\cdot P_2 + 0\cdot P_3)$, $1\cdot P_2/(1\cdot P_1 + 1\cdot P_2 + 0\cdot P_3)$ and 0, respectively.

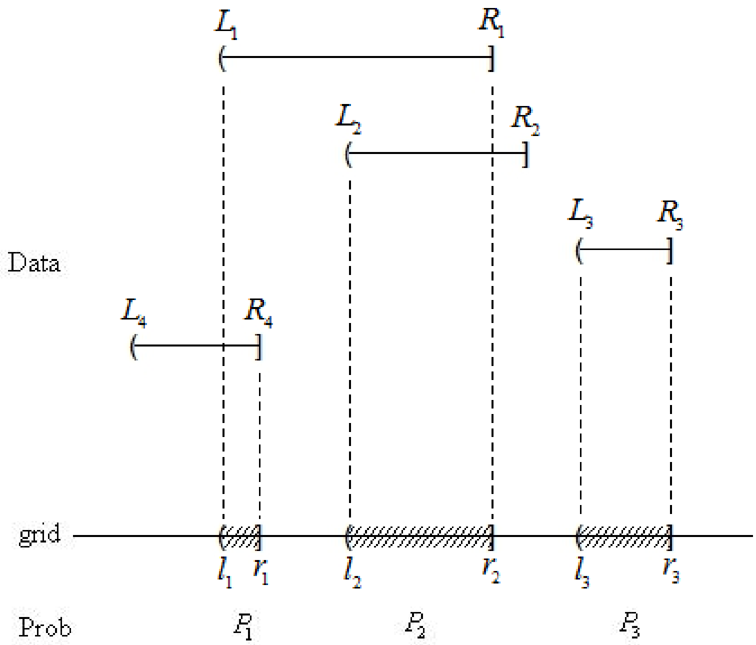

Figure 3: Construction of the mass interval censored data and the idea of self-consistency.

Now we briefly compare survival analysis with logistic regression. Suppose T refers to the onset age of an event and A is the current age. Denote $Y = I(T \le A)$. By fitting a logistic model, we have

$$\log\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right) = \log\left(\frac{\Pr(T \le A)}{\Pr(T > A)}\right) = Z^T\beta,$$

where $\Pr(T \le A) = \int_u F(a) f_a(u)du$, $\Pr(T > A) = \int_u S(a) f_a(u)du$ and $f_a(.)$ is the density function of A. We see that the age distribution for observations in the sample will have a confounding effect on the analysis. Also logistic regression analysis for event-type data in practical applications often ignores the fact that T is subject to censoring and makes the results more mis-leading.

3.4 Methods for analyzing paired data

Myopia data usually contain paired observations. Here we review different methods for analyzing paired data. Let (Y_1, Y_2) be some measures obtained from left and right eyes respectively. If the main purpose is to make comparison based on the means, define $D_i = Y_{i1} - Y_{i2}$, $\bar{D} = \sum_{i=1}^{n} D_i / n$ and S_D is the standard deviation of D_i $(i = 1, \ldots, n)$. We may consider the paired T statistic given by

$$t = \frac{\bar{D}}{s_D / \sqrt{n}}$$

which can be used to test whether the mean of $E(D) = 0$. Other nonparametric tests are available such the sign test or signed rank test.

Sometimes the purpose is to find the association between the two variables rather than making comparison. To describe the association between paired variables (X, Y), which may be (Y_1, Y_2) in the above case, we can consider Pearson's correlation coefficient:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

which can be estimated by

$$r_{X,Y} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}.$$

Kendall tau correlation is a rank correlation coefficient and is more robust than Pearson's correlation. Before defining Kendall's tau, we first introduce the concept of concordance. Let (X_i, Y_i) and (X_j, Y_j) be two independent replications from (X, Y). They are "concordant" if

$(X_i - X_j)(Y_i - Y_j) > 0$; while they are "discordant" if $(X_i - X_j)(Y_i - Y_j) < 0$. If $(X_i - X_j)(Y_i - Y_j) = 0$, the pair is neither concordant nor discordant. After calculating the number of concordant pairs, n_c , and the number of discordant pairs, n_d ,the Kendall tau correlation coefficient can be written as

$$\tau = (N_c - N_d) / \binom{n}{2}$$

where $N_c = \sum_i \sum_{j>i} I[(X_i - X_j)(Y_i - Y_j) > 0]$ is the number of concordance pairs and

$N_d = \sum_i \sum_{j>i} I[(X_i - X_j)(Y_i - Y_j) < 0]$ is the number of discordance pairs.

Chapter 4: Statistical Applications in Myopia Research

In this chapter, we will review how statistical methods are applied to study myopia in Taiwan and Singapore. In particular, Singapore has the highest prevalence rate in the world.

4.1 Myopia research in Taiwan

In Taiwan, myopia has also been a serious health problem. The research team of National Taiwan University conducted five nationwide surveys for 7 to 18 years-old schoolchildren in 1983, 1986, 1990 and 2000 (Lin et al., 2001 and 2004). It was found that the prevalence of myopia for 7 years-old children increased from 5.8% in 1983 to 21% in 2000; while the prevalence for 12 years-old children increased from 36.7% in 1983 and to 61% in 2000 rapidly. The onset age of myopia had decreased during the period of 17 years. Specifically the average onset age of myopia decreased from 11 years-old in 1983 to 8 years-old in 2000. They also compared the differences in gender and the living environments. For example for children of the same age, girls had higher prevalence rate and more severe myopia status than boys. School children living in cities had higher prevalence rate and more severe myopia than those living in countries. Shih et al. (2010) further studied the progression of myopia based on the longitudinal study and found that the growth patterns of myopia were also different for children living in cities and countryside. For example the average growth rates of myopia for 10-15 years old children who lived in cities were 0.43D/year and 0.50D/year for boys and girls, respectively; while the rates for children living in countryside were 0.24D/year and 0.31D/year for boys and girls, respectively. This implies that environmental factors play some roles in the development of myopia.

Another research team in Chung Shan Medical University investigated possible factor that might influence the status of myopia for elementary school students in Taiwan. Cheng et al. (2013) selected three elementary schools in Tamsui, Taichung and Tainan, located in the northern, middle and southern parts of Taiwan, respectively. They found that the condition of myopia is affected by the level of nearwork, outdoor physical activities and the use of spectacles … etc.

The above studies focused more on epidemiological issues. Alternatively some researchers investigated the relationship between myopia and possible biometrical measures. For example Pei-Yao Chang et al. (2010) studied the relationship between myopia and the axial length and corneal hysteresis.

4.2 Myopia Research in Singapore

In Singapore, about 85% of people have myopia. Chew et al. (1988) found that people with higher education level were more likely to get myopia and the condition was more severe. SCORM (the Singapore Cohort study Of the Risk factors for Myopia) is a cohort study which has collected the information of schoolchildren from several schools since 1999. The projected was conducted by researchers in National University of Singapore. They found that breast feeding, birth weight, parental smoking and outdoor activities are related to myopia. Genetic investigation by studying monozygotic twins was also pursued. The influences of race, culture and education level has also be examined. The article "Historical Overview of Myopia Research in Singapore" (2008) provides a useful summary about the research development, important findings and key persons.

We now summary some papers of Saw et al., who is the leader of SCORM. Then we will give some comments from the viewpoint of statisticians.

4.3 Studying Risk Factors on Myopia of Different Severity

In the paper by Saw et al. The main purpose is to investigate the relationship of nearwork activities and myopia for elementary school-age children. The cross-sectional study conducted in 1999 collected 1005 children, aged 7 to 9 years, from two schools in Singapore. One school is ranked among the top 20 schools in Singapore, and the other is ranked among the bottom 20 schools. For each child, ophthalmological measurements including the axial eye length, anterior chamber depth, crystalline lens thickness and vitreous chamber depth were taken from two eyes. Besides the lab results, additional questionnaires were given to children's parents to record the information about children's nearwork activities such as reading, time on computer or video games and other possible risk factors, such as parental myopia, socioeconomic status, and light exposure history.

Now we summarize the statistical methods. First, the correlation between the refractive errors for the left and right eyes was found to be 0.94. As result they decided to use the right eyes data in analysis. The severity of myopia is classified into three levels: : "higher myopia" (SE≦-3.0 diopters), "lower myopia" (-3.0<SE≦-0.5 diopters) and "nonmyopes" (SE>-0.5 diopters), where "SE" is the abbreviation of "spherical equivalent". To investigate the relationship between the level of myopia and possible risk factors, they used ANCOVA which combines ANOVA and regression. Specifically ANCOVA evaluates whether population means of a dependent variable (DV) are equal across levels of a categorical variable, while statistically controlling for the effects of other variables which are of less interest.

We present their results. The prevalence rates of higher myopia and lower myopia were 8.1% and 24.3%. Chinese students had higher prevalence (37.0%) than non-Chinese (19.9%). For children with higher myopia, there are 10% Chinese students with higher myopia and 2.9% non-Chinese students in both schools. Compared with the children with lower myopia or no myopia, children with higher myopia were more likely to have higher cylinder power, longer axial lengths, deeper anterior chambers, longer vitreous chambers, steeper corneas, and a higher ratio of axial length (AL) to corneal radius (CR). In respect of family background, there exist positive associations between higher myopia prevalence rates and larger housing type, higher family income, more advanced father's and mother's education ($P < 0.001$, for each). To assess the effect of nearwork activities, higher myopia is strongly related to the number of books reading per week. In summary, severe myopia is related to the above mentioned biological measurements. Children growing up in a richer or more educational environment are more likely to have myopia.

Multiple logistic regression analysis was also conducted to study the effect of "reading more than two books per week" on the incidence of high myopia. It was found that the crude odds ratio of higher myopia for reading more than two books per week was 3.15 (95% CI, 1.96-5.04), whereas the odds ratio adjusted for other risk factors was 3.05 (95% CI, 1.80-5.18). Note that Chinese parents in Singapore usually encouraged reading and had higher income.

4.4 A Longitudinal Study for Predicting Myopia

Here we introduce how a longitudinal study can be used to predict the occurrence of myopia based on the paper of Jones et al. (2007). The study recruited 514 children in the third grade (aged from 8 to 9 years) who did not have myopic in the right eye and then the subjects were followed up until the eighth grade. The major interest was to investigate whether there was any difference between those who developed myopia and those who did not developed myopia within the five years given that their eye conditions were in similar at the baseline. Besides the family information, at each follow-up, children's information about the time on activities or biometrical measures to assess the visual condition was collected.

Several T tests were performed to compare the two groups at the eighth grade. Furthermore, a number of simple logistic regression analysis was conducted. For example, let $Y=1$ indicate that the student developed myopia within the study period and $Y=0$ otherwise and Z can be chosen from the following variables, namely cycloplegic sphere, corneal power, axial length, hours of sports, hours of reading, hours of TV, hours of studying, hours of computer, diopter hours, father myopia, mother myopia and number of myopia parents. Logistic regression model assumes that

$$\log\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right)=\beta_0+\beta_1 Z\,.$$

We summarize their findings. The odds ratio was 2.17 with only one parent having myopia; while the ratio was 5.4 with both parents having myopia. Note that parents' myopia conditions may also affect their parenting behavior. To examine this issue, the number of myopic parents is treated as a categorical variable of tree levels (0,1,2) and the other categorical variable is constructed based on the number of hours of sports and outdoor activity per week. Applying the Chi-square test, it was found that the number of physical activities per week and the number of myopic parents are correlated. Some covariates are further evaluated by AUC based on the ROC curve. The results show that the number of myopic parents, the time on sports and outdoor activity per week and the time on reading per week were significantly associated with children's future myopia. Then, the authors used these three variables to construct the multivariate logistic model. The reading time was not significant in this model. The final model included the following explanatory variables: cycloplegic sphere, corneal power, axial length, the number of hours of sports and outdoor activity per week, the number of myopic parents and the interaction of sport and myopic parents.

4.5 Myopia and other eye diseases

Tong et al. (2006) conducted a longitudinal study to assess the relationships between the severity of myopia and other conditions of eyes (prevalence of anisometropia, changes in the inter-eye difference in spherical equivalent and the change of axial length). The study recruited 1979 children aged 7 to 9 years from 3 Singaporean schools. Several biometrical measurements on eyes were taken per year and then continued for 4 years. Based on their baseline conditions, children were classified as "both eyes myopic", "both eyes hyperopic", "one eye myopic and the other eye emmetropic" and "one eye hyperopic and the other eye emmetropic".

To make comparison for different groups, Mann-Whitney test and Kruskal-Wallis test were considered. The result showed the prevalence rate of anisometropia was not associated with gender at any visit and was associated with age. The prevalence rate of anisometropia in those with "at least one myopic eye" was significantly different from "nonmyopes".

Chapter 5: Our Suggestions on Statistical Analysis

We find that although statistical methods played an important role in myopia research, some advanced but useful statistical methods have not been adopted. For example myopia is a progressive status and practitioners may be interested in some particular stages. We may define $(T_1, T_2, \ldots, T_k)$ as the times to reach these stages. If the events have an ordered structure, we have $T_1 < T_2 < \cdots < T_k$ but their order can also be arbitrary. searchers can apply techniques of multivariate survival analysis to study the association structure of the event times or study how covariates affect the joint behavior.

We also noticed that most literature utilized on a half of data based on paired observations. Appropriate use of all data which needs to handle the association of pairs will increase the efficiency. For example for two variables (Y_{1i}, Y_{2i}) on the same person i, we may assume

$$Y_{ji} = \beta^T Z_{ji} + \alpha_i + \varepsilon_{ji} \ (j = 1,2; i = 1,\ldots,m)$$

where β stands for the fixed-effect parameter and $\alpha_i \overset{iid}{\sim} N(0, \sigma_\alpha)$ is the random effect specific to the ith person. The magnitude of σ_α reflects the degree of association for the two measurements. Alternatively GEE methods can also be used by employing the working correlation matrix.

The onset ages of myopia of two eyes can be denoted as (T_1, T_2) and their association can be described by the odds ratio function originally defined by Oakes (1989) such that

$$\theta(t_1,t_2)=\frac{S(t_1,t_2)\dfrac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2}}{\dfrac{\partial S(t_1,t_2)}{\partial t_1}\dfrac{\partial S(t_1,t_2)}{\partial t_2}}$$

where $S(t_1,t_2)$ is the joint survival function of (T_1,T_2). Statistical inference of $\theta(t_1,t_2)$ based on bivariate censored data has been discussed by many statisticians. One can refer to the paper of Wang and Wells (2000) which summarizes related references.

References

〔1〕 Au Eong, Kah Guan, Primary Eye Care In Singapore: Looking Back, Looking Forward, Sight & Eye International Pte Limited, 2008.

〔2〕 Cheng CY, Huang W, Su KC, Peng ML, Sun HY, Cheng HM, "Myopization factors affecting urban elementary school students in Taiwan", Optom Vis Sci, 900, pp. 400-6, 2013 Apr.

〔3〕 Jones LA, Sinnott LT, Mutti DO, Mitchell GL, Moeschberger ML, Zadnik K, "Parental History of Myopia, Sports and Outdoor Activities, and Future Myopia", Invest Ophthalmol Vis Sci, 48, pp.3524-32, 2007 Aug.

〔4〕 Lin LL, Shih YF, Hsiao CK, Chen CJ, Lee LA, Hung PT, "Epidemiologic study of the prevalence and severity of myopia among schoolchildren in Taiwan in 2000", J Formos Med Assoc, 100, pp. 684-91, 2001 Oct.

〔5〕 Lin LL, Shih YF, Hsiao CK, Chen CJ, "Prevalence of myopia in Taiwanese schoolchildren: 1983 to 2000", Ann Acad Med Singapore, 33, pp. 27-33, 2004 Jan.

〔6〕 Peter Diggle, Patrick Heagerty, Kung-Yee Liang, Scott Zeger, Analysis of Longitudinal Data, Oxford University Press,2013

〔7〕 Saw SM, Chua WH, Hong CY, Wu HM, Chan WY, Chia KS, Stone RA, Tan D. "Nearwork in Early-Onset Myopia", Invest Ophthalmol Vis Sci, 43, pp.332-9, 2002 Feb.

〔8〕 Shih YF1, Chiang TH, Hsiao CK, Chen CJ, Hung PT, Lin LL, "Comparing myopic progression of urban and rural Taiwanese schoolchildren", Jpn J Ophthalmol, 54, pp. 446-51, 2010 Sep

〔9〕 Tong L, Chan YH, Gazzard G, Tan D, Saw SM, "Longitudinal Study of Anisometropia in Singaporean School Children", Invest Ophthalmol Vis Sci, 47, pp. 3247-52, 2006 Aug.

〔10〕 Weijing Wang, Martin T. Wells, "Model Selection and Semi-parametric Inference for Bivariate Censored Data" , Journal of American Statistical Association, 95, pp. 62-72, 2000 Jan

附 錄：眼科醫學名詞之中譯

anterior chamber depth	前房深度
autorefractor	自動驗光機
axial length	眼軸長
corneal	角膜
corneal curvature radius	角膜曲率半徑
crystalline lens thickness	水晶體厚度
cycloplegia	睫狀肌麻痺劑
ocular power	視力
retina	視網膜
spherical equivalent	屈光度
vitreous chamber depth	玻璃腔深度