Eur Respir J, 1993, 6, 325–327
Printed in UK - all rights reserved

Copyright© ERS Journals Ltd 1993
European Respiratory Journal
ISSN 0903 - 1936

EDITORIAL

Analysis of longitudinal data: choosing and interpreting regression models

J.H. Ware*

In this issue of the journal, Sherrill *et al.* [1] present results from a longitudinal study of pulmonary function levels of 1,524 men and women, aged ≥55 yrs at their initial examination. Participants were followed for up to 14 yrs to identify factors affecting the level and rate of decline of function. The authors report that respiratory symptoms and cigarette smoking were negatively associated with pulmonary function level and, in some cases, rate of decline of function, and quantify these effects. The paper is one of several [2–6] that have used new methods of longitudinal analysis to characterize individual patterns of pulmonary function growth during childhood and decline during adult life. This editorial discusses the advantages of longitudinal studies and the strategies for choosing and interpreting regression models, and comments on some aspects of the models used by Sherrill *et al.* [1].

Why are longitudinal studies important?

A data set is longitudinal if some study participants are observed on more than one occasion. Longitudinal designs are superior to cross-sectional designs in several ways. Firstly, only longitudinal data can provide information about individual rates of change over time. When cross-sectional data are used to estimate the effects of age, cigarette smoking, and other time-dependent variables, the estimates are based on comparisons between individuals, and can be confounded by other differences between age cohorts. Secondly, longitudinal designs provide the opportunity to measure participant characteristics prospectively, while cross-sectional studies require participants to recall a lifetime of smoking behaviour and respiratory health. Finally, longitudinal data give more efficient estimates of the effects of age and other variables that change over time than cross-sectional data, because subjects serve as their own controls. Although the advantages of repeated measurements are intuitively apparent, a formal demonstration of the increased efficiency is outside the scope of this editorial.

Choosing a longitudinal model

The longitudinal models used to analyse pulmonary function data are closely related to the models used in ordinary multiple linear regression. The form of the regression model is, in fact, identical to that used in ordinary multiple regression, but the methods used to estimate the regression coefficients must be modified, to account for the correlation between repeated measurements on the same subject. Consider the models used by Sherrill *et al.* [1]. If $\mathbf{y}_{ij}$ is the **j**th pulmonary function measurement for the **i**th subject, Sherrill and colleagues assume that $\mathbf{y}_{ij}$ depends linearly on age, height, symptom status and smoking status. For example, the regression model for forced expiratory volume in one second as percentage of forced vital capacity (FEV$_1$/FVC%) for men at any examination is given by the expression:

$$\begin{aligned} E(y_{ij}) = {} & 109.8 - 0.287\times \text{age} - 0.2081\times\text{height} - 1.412\times \\ & \text{cough} - 2.738\times\text{wheeze} - 3.43\times\text{dyspnoea} - \\ & 1.977\times\text{exsmoker} - 3.387\times\text{current smoker} + \\ & 0.136\times(\text{wheeze}\times\text{age}) - 0.138\times(\text{current} \\ & \text{smoker}\times \text{age}). \end{aligned}$$

Readers should interpret this function just as they would a regression model fitted to cross-sectional data. For example, FEV$_1$/FVC% is estimated to decline by 0.287% per year of age and to be 3.387% lower for current smokers than for nonsmokers. Moreover, the interaction term involving current smoking and age implies that the difference between current smokers and nonsmokers is estimated to increase by 0.138% per year of age.

A statistical issue arises in the estimation of the regression coefficients. Because the repeated observations on a single subject are correlated, ordinary multiple linear regression analysis will give inefficient estimates of these coefficients and very misleading standard errors. Thus, longitudinal analysis requires specification of two models, the linear regression model and the model for the covariances among the repeated measurements on the same subject. The regression coefficients are estimated by generalized least squares, rather than ordinary least squares, to account for the assumed covariance structure among the observations. Methods for linear regression analysis of longitudinal data [7–9] differ only in their approach to modelling these covariances. Laird and Ware [7] assumed that the correlation can be described by supposing that each individual has a growth curve and that the growth curve coefficients vary randomly among individuals. Jones and Boadi-Boateng [8] generalized this idea by assuming that the deviations from these individual

* Department of Biostatics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA.

growth curves are temporally auto-correlated. SHERRILL *et al.* [1] adopt the approach of Jones and Boadi-Boateng and assume that both individual growth curves are linear in age and that errors are auto-correlated. In a third influential paper, LIANG and ZEGER [9] allowed an arbitrary correlation structure but required subjects to be observed on a common set of occasions. Liang and Zeger proposed robust methods for estimation of standard errors of regression coefficients, that are valid even when the model assumed for the covariance structure is incorrect. From the users point of view, the similarities among these methods are more important than the differences. All yield estimates of linear regression models that account for intrasubject correlation.

Interpreting the model

Once the simplicity of linear longitudinal models is appreciated, the reader can evaluate an investigator's model in familiar ways. For example, in the models described in their table 3, SHERRILL *et al.* [1] assume that pulmonary function measurements depend linearly on age and height, depend on symptom status as reported at the time of examination but not at previous examinations, and depend on smoking status (current or ex-smoking) with no measure of amount smoked. One aspect of this model is the choice of scale for the dependent variable. Analyses of adult pulmonary function measurements collected as part of the Six Cities Study [10] indicated that division of FEV_1 and FVC by the square of height controlled for the effects of height on pulmonary function level and also removed the variability induced by differences in height more effectively than inclusion of height or the square of height in the regression model. Although a model that assumes a linear dependence on height is clear and appealing, analysis of residuals and other methods for assessing goodness-of-fit can help determine which model provides better fit to the data. In the same data, the rate of loss of pulmonary function was found to accelerate with age [5], which would require inclusion of the square of age in the linear model.

An investigation of the residuals from the fitted model as a function of age would determine whether acceleration is occurring in the older adults studied by SHERRILL *et al.* [1]. It seems reasonable that respiratory symptoms, especially wheeze and dyspnoea, might have an acute effect on pulmonary function level that ceases when the symptom resolves. In our data, however, pulmonary function levels among current and ex-smokers depend linearly on the cumulative cigarette smoking measured in packyears [6, 10]. Choices such as these about the form of the regression model, are an important part of longitudinal analysis. The adequacy of any longitudinal model can be fully assessed only by investigating the goodness-of-fit of the model.

When is an analysis longitudinal?

Even when a data set contains repeated measurements and the authors use longitudinal methods, the estimates of regression coefficients can depend on cross-sectional as well as longitudinal information in the data. In the analyses of SHERRILL *et al.* [1], for example, the estimated regression coefficients summarise both differences between individual subjects at successive examinations and differences between subjects. The weights given to the two sources of information depend on several factors, including the relative sizes of the between- and within-subject variability. Statisticians call longitudinal models of this type "marginal" models because they provide an estimate of the expected pulmonary function level of a subject at a given age with given characteristics. Thus, the information is, in one sense, cross-sectional even though the design is longitudinal.

One aspect of this phenomenon merits special comment. SHERRILL *et al.* [1] use binary indicator variables for surveys to "correct for differences between surveys that could result from changes in equipment or techniques". Because the average change in age between examinations can be represented almost perfectly by these indicator variables, the average change in pulmonary function level between examinations will not contribute to the estimate of the effect of age on pulmonary function level. Thus, the coefficient for age is based primarily on cross-sectional data. The survey indicator variables should not, however, have appreciable effects on other regression coefficients.

In our data, cross-sectional and longitudinal models give very similar estimates for the effect of age on pulmonary function level, so the effects of using survey indicator variables should be small. Nevertheless, this effect illustrates the more general point that the use of longitudinal methods in the analysis of repeated measurements does not guarantee that the model describes how individuals change over time. With that issue in mind, we have developed methods of analysis that use only the individual changes between examinations to estimate longitudinal regression coefficients [5]. The coefficients obtained using these methods are purely longitudinal, in that they do not depend upon between-individual differences.

Conclusion

The statistical methods required to fit linear models to longitudinal data are now well-established. These methods are becoming widely available through new procedures offered in SAS, BMDP, and other statistical packages. As epidemiologists become more familiar with these models, we can anticipate wider use of longitudinal methods, more effective analysis of longitudinal studies, and an informed debate about the evidence that these studies provide about the growth and ageing of the lung. The important paper by SHERRILL *et al.* [1] illustrates the power of longitudinal designs and longitudinal analysis. The author hopes that this editorial will be of value to those who seek to understand and evaluate the results.

References

1. Sherrill DL, Lebowitz MD, Knudson RJ, Burrows B. – Longitudinal methods for describing the relationship between

pulmonary function, respiratory symptoms and smoking in elderly subjects: The Tucson Study. *Eur Respir J* 1993; 6: 342–348.

2. Berkey CS, Ware JH, Dockery DW, Ferris BG Jr, Speizer FE. – Indoor air pollution and pulmonary function growth in preadolescent children. *Am J Epidemiol* 1986; 123: 250–260.

3. Sherrill DL. Sears MR, Lebowitz MD. – The effects of airway hyperresponsiveness, wheezing and atopy on longitudinal pulmonary function in children: a six year follow-up study. *Pediatr Pulmonol* 1992; 13: 78–85.

4. Sherrill DL, Martinez FD, Lebowitz MD. – Longitudinal effects of passive smoking on pulmonary function in New Zealand children. *Am Rev Respir Dis* 1992; 145: 1136–1141.

5. Ware JH, Dockery DW, Louis RA, *et al.* – Longitudinal and cross-sectional estimates of pulmonary function decline in never-smoking adults. *Am J Epidemiol* 1990; 132: 685–700.

6. Xu X, Dockery DW, Ware JH, Speizer FE, Ferris BG Jr. – Effects of cigarette smoking on rate of loss of pulmonary function in adults: a longitudinal assessment. *Am Rev Respir Dis* (in press).

7. Laird NM, Ware JH. – Random effects models for longitudinal studies. *Biometrics* 1983; 38: 963–974.

8. Jones RH, Boadi-Boateng F. – Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics* 1991; 47: 161–175.

9. Liang KY, Zeger SL. – Longitudinal analysis using generalized estimating equations. *Biometrika* 1986; 73: 13–22.

10. Dockery DW, Speizer FE, Ferris BG Jr, *et al.* – Cumulative and reversible effects of lifetime smoking on simple tests of lung function in adults. *Am Rev Respir Dis* 1988; 137: 286–292.