# The Application of Hierarchical Linear Model in the Study of Tumor Progression

Shuang Li
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
690736804@qq.com

Pu Wang, Li Yingfang, Pei Yuchen
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
{wangpu, fangliying}@bjut.edu.cn
530795037@qq.com

***Abstract*—This paper presents how to analyze the tumor progression data using hierarchical linear model. In the recent 20 years, cancer has been a serious threat to the human health and life; therefore, how to analysis the longitudinal data of patients effectively has become an urgent problem to be solved. In this article, we describe the principle of the hierarchical linear model and its application in the longitudinal study. Take a group of followed-up data of lung cancer patients for example; the output is the estimation of various parameters. I discussed the meaning of the various parameters in the hierarchical linear model and find out the influence of different fixed individual characteristics and the time-varying factors on the tumor progression. At the same time, we made a reasonable analysis of the strengths and weaknesses of hierarchical linear model.**

***Keywords - hierarchical linear model; longitudinal data; cancer; time-varying.***

## I. INTRODUCTION

Longitudinal data, more specifically, can be considered as measurements on several variables for the same (groups of) individuals on a number of consecutive points in time. In this paper, we will be concerned with the analysis of individual growth as well as the average growth trend of a group of subjects [1]. For example, a longitudinal medical study offers a multifaceted way to analyze factors leading to a certain disease. This paper makes a further study for the physical tests and the physical symptoms of the subjects who are tested in frequent and irregular intervals for a long period of time.

Cross-sectional and simple time series approaches do not make full use of data available. Longitudinal medical data has both cross-sectional and time series characteristics. They are cross-sectional because they include many physical symptoms; we say that these data are differentiated across 'space' [2]. They are time series data because they represent many points in time. Longitudinal data methods are appropriate for these types of data.

There are two aspects concerning longitudinal data study; first, describing the individual development trend and the differences of the development trend between individuals, and second, explaining the development trend and the reason; the prediction variables can be unstable factors with time, and also can be fixed individual characteristics factors.

In the recent years, the techniques of longitudinal data analysis had made great progress; the hierarchical linear model had been widely used in longitudinal research. The term of the hierarchical linear model was firstly presented by Lindley and Smith in 1972 [3]; however, the traditional parameter estimation method (OLS) did not apply to the model because of the limitation of computing technology. In 1980s, iteratively reweighted generalized least squares and other methods had been used to estimate parameter, and there were some calculation software such as HLM (hierarchical linear model) [4]. In this article, we use HLM to analyze the medical followed-up data.

The remainder of this paper is organized as follows. First, we present the limitations of traditional statistical techniques and the advantages of hierarchical linear model in Section 2. In Section 3, we discuss the theory of hierarchical linear model. The application of hierarchical linear model in the study of tumor progression is discussed in Section 4. Concluding remarks are given in Section 5.

## II. THE ADVANTAGES OF THE HIERARCHICAL LINEAR MODEL

### *A. The assumptions of the homogeneity of variance and the independence of random error*

In the traditional statistical techniques, we always use variance analysis and multiple regression analysis to handle the longitudinal data. However, both of the two methods were used at the assumptions of the homogeneity of variance and the independence of random error, which were difficult to guarantee [5]. There were similarities between multiple tracking data from the same individual, and systematic errors might exist in the observations at the same time point, these issues made the assumption of random error independence hard to meet. Meanwhile, dependent variable occurred with a regular increase or decrease with time, which caused the increase or decrease of the variance. Those changes had a great effect on the homogeneity of variance. Therefore, traditional statistical techniques might lead to unreasonable or even wrong conclusions. Moreover, hierarchical linear model does not require the assumptions of the homogeneity of variance and the independence of random error, so it was more suitable for longitudinal studies.

 ISBN: 978-1-61208-202-8 

### B. *Problems of missing values and unequal measurement interval*

Longitudinal studies needed to do repeatedly tracking observations of the same individual; it was prone to the loss of the sample. In the traditional statistical tools, there were two methods to handle the lost, one was removing the observed object which had missing values, another was fitting the missing values, and the former caused the waste of information, while the latter reduced the precision. While the HLM allowed the existence of missing values, it also made full use of the existing information. In addition, the traditional statistical technique required that all objects were observed at the same time interval. The HLM not only allowed the measurements of different time intervals, but also allowed the observed objects to have different observation schedules, this feature enabled researchers to have more convenience and flexibility [6].

### C. *The ability of dealing with the hypothesis*

HLM allowed researchers to put forward different assumptions in different levels, such as whether it had a significant increase or decrease? Whether the different individuals had the same change rate? Which factor could predict the difference of change rate between different individuals? It also could create multiple development models and select the assumption which is the most consistent with observation data through the square test.

## III. HIERARCHICAL LINEAR MODEL

Hierarchical linear model in different research fields had different terms, such as Multilevel Statistical Model, Mixed Model, Random Coefficient Model and so on, the diversity of the term reflected that this method had been widely used in different fields. No matter what the term was, the core content of the method was same, mainly to handled the nested structure data. For repeated tests, these data had nested structure that the measurement nested with the individual [7]. This method could both solve the individual development trend and the differences of the development trend between individuals, also it could directly deal with the unequal measuring time interval, and at the same time it could analysis the missing value of data reasonably.

We used the hierarchical linear model to analyze the longitudinal data; data were found in different hierarchies. At first, established a regression equation for the first level variables, in which the tracking results that came from different observation times were the first layer and the invariant individual characteristics or the dispose that had been accepted were the second layer data. Through the processing, it could explore the effect of different levels on the dependent variables. In the first floor of the data structure, the track observation result was considered as the dependent variable.

$$Yij = \beta_{0i} + \beta_{1i} X_{ij} + \varepsilon_{ij} \quad (1)$$

As in "(1)", subscript "0" means intercept, subscript "1" means slope, subscript "i" means the i-th observation object, Subscript "j" indicates the j-th observation time." $\beta_{0i}$ " is the intercept of the equation, it indicates the average of the i-th observed objects. " $\beta_{1i}$ "is the regression coefficient, it indicates the changing rate of the i-th observation object." $X_{ij}$ " means the values of the variable X when the i-th observed object is in the j-th observation time, " $\varepsilon_{ij}$ " means residual, the implication is that the measured value Y of the i-th object in the j-th observation time that cannot be explained by the independent variable X.

Equation (1) is similar with the general regression equation, the only difference is, intercept and slope are not constant [8]. Different observation object have different intercept and slope, they may subject to the affect of variable of the second layer. In the second layer of the data structures, the intercept and slope are used as the dependent variable in (1), and the individual characteristics or the dispose that have been accepted are considered as independent variables, then we create two regression equations for the second layer:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} W_{1i} + \mu_{0i} \quad (2)$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} W_{1i} + \mu_{1i} \quad (3)$$

In the above two equations, each parameter has two subscripts, if the first subscript is "0", this is the parameter that relates to the intercept of (1). If the first subscript is "0", this is the parameter that relates to the slope of (1). If the second subscript is "0", it means the intercept part of the second layer equation, if the second subscript is "1", it means the slope part of the second layer equation.

$\gamma_{00}$ is the intercept of (2), it can be understood as the average of the dependent variable Y when the independent variable W1 is 0.

$\gamma_{01}$ is the regression coefficients of the variables W1 in (2), it can be understood as the impact of the variable W1 to the initial value of the dependent variable Y.

$\gamma_{10}$ is the intercept of (3),it can be understood as the changing rate of the observed object when the variable W1 is 0.

$\gamma_{11}$ is the regression coefficients of the variable W1 in (3), it can be understood as the effect of the variable W1 on the changing rate.

$\mu_{1i}$ is the residual of (3), it can be understood as the changing rate of dependent Y that can't be explained by the variables W1 [9]. If the variance statistical test is significant, then the model needs to introduce new variables to explain the variation of the rate.

To simplified the problem, the second layer of the above regression equations only contains one variable W1, if there are multiple independent variables, accordingly, the slope part of (2) and (3) are needed, respectively, add $\gamma_{02} \cdot \gamma_{03} \cdot \gamma_{12} \cdot \gamma_{13}$ and so on.

 ISBN: 978-1-61208-202-8 

## IV. THE APPLICATION OF HIERARCHICAL LINEAR MODEL IN THE STUDY OF TUMOR PROGRESSION

### A. Data description

In order to study the condition trend of patients with the non-small cell lung cancer (NSCLC), treatment data of 52 patients, including six treatment data which interval is one month, were selected and two SPSS files were established. One file contains the patients' ID and the variable of "type" which is used as the number of measurement. The score of fact and symptom of each inspection are also included in it. Fact score is used as the output, the symptom and type are used as the predictor variables. Another file contains the id and sex as well as whether smoking in the initial state of the patients. Two SPSS files were read-in using HLM. The lengthways variation trend was analyzed respectively through two models as following.

### B. Unconditional growth model

The second layer of equations does not contain any independent variable in the unconditional mean model. The function of the model is to describe the variation trend of the total observed object, and to make decision of that weather to introduce the second layer of the explanatory variables. Then set the following two-layer equations in this model.

The first layer equation:

$$Fact = \beta_0 + \beta_1(\text{type}) + \beta_2(\text{symptom1}) + \varepsilon \quad (4)$$

The second layer equations:

$$\beta_0 = \gamma_{00} + \mu_0 \quad (5)$$

$$\beta_1 = \gamma_{10} + \mu_1 \quad (6)$$

For the variable type,0,1,2,3,4,5 are used to express it respectively, for the variable of symptom, In the Chinese traditional medicine, -1 indicates the asthenia syndrome, 0 indicates the compounding of the excess and deficiency syndromes, 1 indicates the old trauma sthenia. Use the software HLM to estimate the parameter, the result is showed in Table 1:

TABLE I. THE PARAMETER ESTIMATION1

| fixed effects | coefficient | Standard Error | T-ratio |
|---|---|---|---|
| $\gamma_{00}$ | 43.78 | 2.17 | 20.16 |
| $\gamma_{10}$ | -0.35 | 0.44 | -0.79 |
| $\gamma_{20}$ | 1.23 | 2.76 | 0.45 |
| **Random Effect** | **Variance Component** | **df** | **Chi-square** |
| $\mu_0$ | 118.15 | 20 | 51.77 |
| $\mu_1$ | 2.01 | 20 | 30.52 |
| $\mu_2$ | 153.54 | 20 | 44.17 |

The analysis indicates that in the first measurement, the average of the fact score of all patients is 42.64. Every month, the average of the fact score declined 0.35 points; at the same time, the fact score increase 1.23 points with the one level rising of the symptom.

In the following, we use the figures to explain the effects of the predictor variables (symptoms) on the development trends. For the patients whose symptoms are the compounding of the excess and deficiency (symptom1=0), the trend is 43.78-0.35*type, which is showed in figure 1.The figure2 shows that the patients whose symptoms is 0 of the previous three times, -1 of the fourth time, 0 of the fifth time, and -1 of the sixth time.

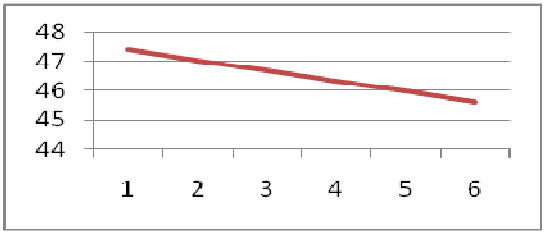


Figure 1. The variation trend1

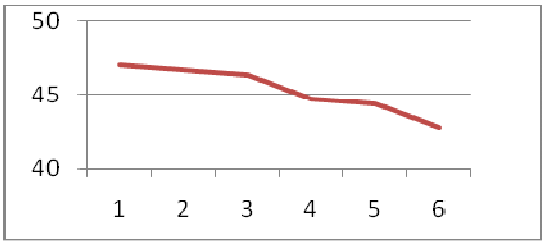


Figure 2. The variation trend2

It can be seen from the figures, if the time-varying predictor variables are included in the model, the model curve may vary based on the predictive value of different time points. The residual variation in (5) and (6) are significant that; indicating large differences in the fact score and rate of change. And thus the variables of second layer are needed to be introduced to get a better explanation.

### C. Combined model

In order to better explain the individual differences of the current score and the rate of change in (4), two second-level variables are introduced: gender (male is 0, female is 1) and whether smoking in the initial state (smoker is 1, nonsmoker is 0), independent variables are included in this two layers.

The first layer equation:

$$Fact = \beta_0 + \beta_1(\text{type}) + \beta_2(\text{symptom1}) + \gamma \quad (7)$$

The second layer equations:

$$\beta_0 = \gamma_{00} + \gamma_{01}(\text{sex1}) + \gamma_{02}(\text{smoke}) + \mu_0 \quad (8)$$

$$\beta_1 = \gamma_{10} + \gamma_{11}(\text{sex1}) + \gamma_{12}(\text{smoke}) + \mu_1 \quad (9)$$

 ISBN: 978-1-61208-202-8 

$$\beta_2 = \gamma_{20} + \gamma_{21}(\text{sex1}) + \gamma_{22}(\text{smoke}) + \mu_2 \quad (10)$$

TABLE II. THE PARAMETER ESTIMATION2

| fixed effects | coefficient | Standard Error | T-ratio |
|---|---|---|---|
| $\beta_0$ | | | |
| $\gamma_{00}$ | 44.79 | 5.60 | 8.00 |
| $\gamma_{01}$ | 4.06 | 5.68 | 0.71 |
| $\gamma_{02}$ | -5.74 | 5.66 | -1.01 |
| $\beta_1$ | | | |
| $\gamma_{10}$ | -0.82 | 1.06 | -0.77 |
| $\gamma_{11}$ | 0.12 | 1.12 | 0.11 |
| $\gamma_{12}$ | 0.83 | 1.11 | 0.74 |
| $\beta_2$ | | | |
| $\gamma_{20}$ | 3.57 | 7.15 | 0.50 |
| $\gamma_{21}$ | -5.44 | 7.29 | -0.75 |
| $\gamma_{22}$ | -0.48 | 7.26 | -0.07 |
| **Random Effect** | **Variance Component** | **df** | **Chi-square** |
| $\mu_0$ | 107.40 | 18 | 37.61 |
| $\mu_1$ | 1.87 | 18 | 28.50 |
| $\mu_2$ | 149.01 | 18 | 43.59 |

The estimation results of fixed parameters shows that, for the first level intercept $\gamma_{00}$ ,it means that the average of the fact sores is 44.79 when the type equals 0 and symptom equals 0; there is a significant difference between female and male in the initial state( $\gamma_{01}$ =4.06 , se=5.68,t=0.71), the female patients have a higher score than men have; smoking in the initial state have a negative predictive effect on the average of the score( $\gamma_{02}$ =-5.74,se=5.66,t=-1.01). Smoking patients, have a lower initial fact score. For the slope $\beta_1$ of the first level, the fact score has a significant descending trend with the increase of the check number ( $\gamma_{10}$ =-0.82, se=1.06,t=-0.77), over time, the gender has little effect on the rate of descending, the fact scores of smoking patients has a fast rate of descending than the nonsmokers'.

the slope $\beta_2$ of the first level, the fact score has a significant ascending trend with the increase of symptom, with the symptom changes, the gender has effect on the rate of ascending, female has a slower rate, ( $\gamma_{21}$ = -5.44, se=7.29, t=-0.75).Weather smoking in the initial time has a little effect on the ascending trend, the smoker has a slower rate ( $\gamma_{22}$ =-0.48 , se=7.26, t=-0.07).

The parameter estimation results of the random effects show that, comparing with the unconditional growth model, taking into account of the influence by the gender and smoking, the variance of the intercept and the slope have decreased, indicating that the two new independent variables explained a part of the intercept and slope variance, but from the random effects of the test results, the remaining unexplained variance is till significant; therefore, it is also needed to introduce a new variable to explain individual difference [10].Meanwhile, when increase the number of the parameters, the running time has no significant increase.

## V. CONCLUSION AND FUTURE WORK

According to the analysis of tumor progression data by the hierarchical linear model, the model not only describes the individual dependent trend and the differences between individuals, but also gives the explanation. The model is specifically used with the medical data; it can study the influence of various factors on the fact score, and can study the influence by the fixed individual characteristics or the unstable time-varying factors. In the future work, importing more variables to analyze the impacts on the output and getting more favorable medical conclusions will be taken into consider.

Hierarchical linear model can handle the relationship between the tested data and the time variable data, and it can make a valid estimation of the parameters of the non-equilibrium measurements. So by using the hierarchical linear model to deal with the repeated measurements data, it needn't require the individuals to have the same number of observation times, and does not require the same time interval between the different individual tests. However the model has its drawbacks, firstly, compare with traditional estimation methods, its more complicated [11]. Secondly, using the hierarchical linear model to analyze data requires more than three times of tracking data, most of the data is hard to meet the requirement. Finally, the outcome variable in level 1 must be equivalent in each test. If all tests use the same test tool, the equivalence of the outcome variables can basically be assured, which means the measurement results are comparable.

## REFERENCES

[1] Laaksonen, M., Simell, B., Salakoski, T., and Simell, O. Integrated data management and analysis environment for medical longitudinal research with machine learning based prediction models. 2009 WRI World Congress on Computer Science and Information Engineering, CSIE, 31 March-2 April 2009, pp. 552-556.

[2] Frees, E W., and Miller, T W. Sales forecasting using longitudinal data models. International Journal of Forecasting, 2004.

[3] Fearn, T. Towards a Bayesian package[C]. Wien, Austria: Physica-Verlag, 1978, pp. 473-479.

[4] Van, D L R., Vrijburg, K., and De, L J. Review of two different approaches for the analysis of growth data using

 ISBN: 978-1-61208-202-8 

longitudinal mixed linear models: comparing hierarchical linear regression (ML3, HLM) and repeated measures designs with structured covariance matrices (BMDP5V)[J]. Computational Statistics and Data Analysis. 1996, pp. 583-605.

[5] Liu, H Y., and Meng, Q M . A Review on Longitudinal Data Analysis Method and It's Development. Advances in Psychological Science 2003, pp. 586-592.

[6] Liu, H Y. How to Abstract Developmental Variations:Latent Growth Mixed Model[J]. Advances in Psychological Science. 2007, pp. 539-544.

[7] Ren, X P., and Pan, C X. On the Study of Forecast of the Foodstuff Output Based on the Input Factors—The Longitudinal Data From 1978 -2005 of Henan Province. China Machine Press.

[8] Singer, J D., and Willett, J B. Applied Longitudi nal Data Anal ysis: Modeling Change and Event Occurrence. New York: Oxford University Press, 2003.

[9] Gai, X S., and Zhang, X H. The Application Of Multilevel Model In Longitudinal Research. Psychological Science 2005 , pp. 429-431.

[10] Molenberghs, G., and Verbeke, G M. Models for discrete longtitudinal data , new York:springer, 2005.

[11] Liu, L., Ma, J Z.,and Johnson, B A. A multi-level two-part random effects model, with application to an alcohol-dependence study B-4968-2009. Statistics in medicine. 2008, pp. 3528–3539.

 ISBN: 978-1-61208-202-8 