sciendo DOI: 10.2478/bile-2018-0010

Biometrical Letters
Vol. 55 (2018), No.2, 139-146

Adapting Hellwig's method for selecting concomitant variables under a certain growth curve model

Mirosława Wesołowska-Janczarek, Monika Różańska-Boczula

Department of Applied Mathematics and Computer Science, University of Life Sciences in Lublin, Głęboka 28, 20-612 Lublin, Poland, e-mail: monika.boczula@up.lublin.pl

SUMMARY

This paper presents an application of Hellwig's method for selecting concomitant variables under a growth curve model, where the values of the concomitant variables change over time and are the same for all experimental units. The authors present a simple adaptation of the growth curve model to the multiple regression model for which Hellwig's method applies. The theoretical considerations are applied to the selection of significant concomitant variables for raspberry fruiting.

Key words: method of growth curves, concomitant variables, Hellwig's method.

1. Introduction

The analysis of data subject to the growth curve model with concomitant variables has been of interest to many statisticians (Fujikoshi, Rao 1991; Fus, Wesołowska-Janczarek 1998; Wang et al. 1999). These studies highlight the problem of determining which concomitant variables have a significant impact on the examined feature and which can be omitted. A special case occurs when the concomitant variables are features other than the tested Y feature and affect its values. For example, such a study may concern the yield of plants over time. From natural observations it appears that the yield may be affected by rainfall or air temperature, which are treated further as concomitant variables. A method for determining the influence of these variables on the yield of raspberries has been presented previously (Wesołowska-Janczarek, Fus 1996; Wesołowska-Janczarek

et al. 1997) for a model in which the influence of concomitant variables is assumed to be the same for all of the units, namely the studied plants.

Hellwig's (1987) method seems to be a simpler method of assessing the impact of concomitant variables for the model under consideration, in particular in the case of a small number of explanatory variables. This paper presents the adaptation of the growth curve model to the regression model, as well as an application of Hellwig's method to a sample set of experimental data.

2. Models and methods

The growth curve model with concomitant variables having the same values for all examined units has the following form (Wesołowska-Janczarek et al. 1997):

$$\underset{np}{\mathbf{Y}} = \underset{na}{\mathbf{A}}\,\underset{aq}{\mathbf{B}}\,\underset{qp}{\mathbf{T}} + \underset{1s}{\mathbf{1}_n}\,\underset{sp}{\boldsymbol{\gamma}'}\,\underset{np}{\mathbf{X}} + \underset{np}{\mathbf{E}} \tag{1}$$

where **Y** is the observation matrix $n = r \cdot a$ (r being the number of replications) of experimental units at p time points, **A** is the matrix of the experimental design dividing the units into a groups, **B** is the matrix of unknown coefficients in the growth curves, which are polynomials of degree q-1, **T** is the Vandermonde matrix which defines the internal observation structure, where $t_1, t_2, \ldots, t_p$ are time points at which the observed feature was measured,

$$\underset{qp}{\mathbf{T}} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ t_1 & t_2 & \ldots & t_p \\ \vdots & \vdots & \ldots & \vdots \\ t_1^{q-1} & t_2^{q-1} & \ldots & t_p^{q-1} \end{bmatrix}$$

1_n is a vector composed of n ones, γ is a vector of unknown regression coefficients determining the dependence of the examined feature on s concomitant variables, **X** is the matrix of s values of concomitant variables at p time points, and **E** is the matrix of random errors. The estimators of unknown matrix parameters obtained by the maximum likelihood method take the form

$$\hat{\mathbf{B}} = (\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'(\mathbf{Y} - \mathbf{1}_n\hat{\boldsymbol{\gamma}}'\mathbf{X})\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{T}'(\mathbf{T}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{T}')^{-1}$$

$$\hat{\boldsymbol{\gamma}}' = \left[\mathbf{1}_n' \mathbf{Y} - \mathbf{1}_n' \mathbf{A}(\mathbf{A}'\mathbf{A})^{-} \mathbf{A}' \mathbf{Y}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{T}'(\mathbf{T}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{T}')^{-1}\mathbf{T}\right] \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}'\mathbf{R},$$

where

$$\mathbf{R} = \left[n\mathbf{X}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}' - \mathbf{1}_n' \mathbf{A}\left(\mathbf{A}'\mathbf{A}\right)^{-} \mathbf{A}'\mathbf{1}_n \mathbf{X}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{T}'(\mathbf{T}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{T}')^{-1}\mathbf{T}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}'\right]^{-1}$$

$$\hat{\boldsymbol{\Sigma}} = \left(\mathbf{Y} - \mathbf{A}\hat{\mathbf{B}}\mathbf{T} - \mathbf{1}_n \hat{\boldsymbol{\gamma}}'\mathbf{X}\right)'\left(\mathbf{Y} - \mathbf{A}\hat{\mathbf{B}}\mathbf{T} - \mathbf{1}_n \hat{\boldsymbol{\gamma}}'\mathbf{X}\right).$$

The values of estimators are determined by the iterative method. It is assumed that

$$E(Y) = \mathbf{ABT} + 1_n \gamma' X \text{ and } \boldsymbol{\Sigma}_{\mathbf{Y}} = \underset{pp}{\boldsymbol{\Sigma}} \otimes \mathbf{I}_n ; \ \underset{pp}{\boldsymbol{\Sigma}} > 0 .$$

- The regression model considered by Hellwig has the form

$$\underset{p1}{\mathbf{y}} = \underset{ps}{\mathbf{X}}\underset{s1}{\boldsymbol{\beta}} + \underset{p1}{\boldsymbol{\varepsilon}}$$

where $\mathbf{y}$ is the vector of p examined feature observations (with dimensions $p \times 1$), $\mathbf{X}$ is the matrix of values of s features which in the growth curves are concomitant variables, $\boldsymbol{\beta}$ is the vector of s regression coefficients, and $\boldsymbol{\varepsilon}$ is the vector of random errors.

To transform the growth curve model (1), part of which refers to multiple regression, it is enough to sum the appropriate values of feature Y (at individual time points), that is, to perform a left-side multiplication of both sides of the model (1) by a row vector consisting only of ones, i.e. $\mathbf{1}_n'$. This gives

$$\underset{1p}{\mathbf{y}'} = \mathbf{1}_n' \mathbf{ABT} + \mathbf{1}_n' \mathbf{1}_n \boldsymbol{\gamma}' \mathbf{X} + \mathbf{1}_n' \mathbf{E} .$$

Finally, if $\mathbf{A}$ is a design matrix for a single classification, then the model of the vector of sums of observations has the form

$$\mathbf{y} = r\,\mathbf{T}'\mathbf{B}'\mathbf{1}_a + n\,\mathbf{X}'\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{2}$$

where the second component describes the relationship between the sums of Y values at p time points and the values of concomitant variables X_i (i =1, 2,…, s) which may be examined using Hellwig's method. The relationships between the Y feature and the subsequent X features considered in these models are linear,

hence in the matrix **X** only the empirical values of concomitant variables collected during the experiment are considered. An interesting issue is the special case of model (2), which additionally includes the products of variables X_i, i.e. the so-called interactions of independent variables. This issue, however, will be the subject of a future study.

3. Hellwig's method

Hellwig's method is described as the optimal predicate selection method or information capacity indicator method. It is regarded as a formal method of selecting explanatory variables for statistical models, in particular econometric models. When using this method, first the vector $\mathbf{R}_0$ should be constructed, along with the matrix **R** of correlation coefficients of the form:

$$\mathbf{R}_0 = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_s \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1s} \\ r_{21} & 1 & \dots & r_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ r_{s1} & r_{s2} & \dots & 1 \end{bmatrix},$$

where r_i, $i = 1,...,s$ is the correlation coefficient of the explained variable *Y* and the potential concomitant variables $X_1, X_2,..., X_s$, and r_{ij}, $i, j = 1,...,s$ are the correlation coefficients between the variables $X_1, X_2,..., X_s$. Then, for each of *l* combinations of independent variables ($l = 2^s - 1$), an integral index

$$H_l = \sum_{i=1}^{k_l} h_{li}$$

is determined based on individual indicators

$$h_{li} = \frac{r_i^2}{1 + \sum_{i \neq j}^{k_l} |r_{ij}|},$$

where k_l is the number of potential explanatory variables in the *l*-th combination. The most desirable set of independent variables corresponds to the combination with the maximum integral index H_l. Grabiński et al. (1982) state that if the

maximum index H_l is close to unity, then the variables included in the *l*-th combination provide almost complete information about the endogenous variable Y. On the other hand, the difference $1 - H_l$ measures the influence exerted on the variable *Y* by predicates not covered by the study (and therefore unknown), among which there is also a random error. When this difference is close to zero, it can be assumed that the influence of variables not included in the study is completely random. Otherwise, it is a signal to search for new predicates that influence the analyzed phenomenon.

4. Selection of significant concomitant variables, using the example of raspberry fruiting

Here, the methods described above will be used to determine the effect of concomitant variables on the fruiting of raspberries over a set time interval corresponding to the harvesting period. The experiment was carried out in the years 1986–1991 at the Department of Pomology of the University of Life Sciences in Lublin, on the fields of the Experimental Farm in Felin. Results were collected for $a = 16$ raspberry varieties in 1988, which was the third year of fruiting. The fruiting period lasted 29 days, from June 27 to July 25. Raspberries were collected at $p = 12$ dates (time points). For each variety, fruit was collected from four plots $(r = 4)$ of 15 m^2. The studied varieties were Chilcotin, Matsqui, Nootka, Meeker, Pojedynek P-70, Barnaulska, Nowokitajewska, Rakieta, Iwanowska, Wisłucha, Wołżanka, Malling Promise, Malling Seedling, Veten, Norna and Canby. At each date when the raspberries were collected, the following were also determined: average daily air temperature (T), sum of daily rainfall (O) and insolation (U). Each of these three features was described by an average computed from the three days preceding the date of harvesting. These three variables were potential candidates for concomitant variables in the model determining the total raspberry yield over time {T, O, U}. It was assumed at the outset (in line with the model considered in this paper) that all plants, regardless of variety, react identically to these three co-existing factors. The following results were obtained using the methods described in the previous section:

- Using the model (1), the pattern of raspberry fruiting was described by a fourth-degree polynomial $(q = 5)$. It was found that variables such as average temperature, rainfall and insolation are significant for raspberry fruiting; this was confirmed by appropriate tests. The following regression coefficients were obtained: –0.00539 for temperature, 0.02976 for precipitation and –0.05124 for insolation. On the basis of these values, it was concluded that the strongest effect on yield was the negative effect of insolation; a slightly weaker, positive, effect came from precipitation; and a weaker negative effect came from temperature. Hellwig's method will indicate whether any of these variables can be considered as insignificant for raspberry yield, and will determine the impact of specific combinations of these variables.
- To apply Hellwig's method, the first part of model (2) was omitted and the calculations were performed only for the concomitant variables. The following were determined:

$$\mathbf{R}_0 = \begin{bmatrix} 0.21 \\ -0.15 \\ 0.41 \end{bmatrix} \begin{matrix} \text{T} \\ \text{O} \\ \text{U} \end{matrix}, \qquad \mathbf{R} = \begin{bmatrix} 1 & & \\ -0.33 & 1 & \\ 0.62 & -0.62 & 1 \\ T & O & U \end{bmatrix} \begin{matrix} \text{T} \\ \text{O} \\ \text{U} \end{matrix}.$$

Next, $l = 7$ combinations $K_1,...,K_7$ were created, for the analyzed concomitant variables T, O, U, and for each of them integral capacity indices H_l were determined based on the individual capacities h_{li}, as shown in Table 1.

As shown in Table 1, the maximum integral information capacity index corresponds to the third combination. By Hellwig's method, the conclusion would follow that insolation has the greatest impact on raspberry yield and is a variable that should undoubtedly be included in the model describing the dynamics of raspberry fruiting over time. It is notable that not only insolation alone, but also the combination of temperature and insolation has a fairly significant effect on the raspberry yield. A similar observation can be made for the combination of precipitation and insolation. It is apparent that the difference $1 - H_3 = 0.83$ is a large value. In our case, this measures the influence exerted on

Table 1. Values of indices of individual and integral capacity for all combinations of T, O, U

Combinations	h_{li}		H_l	
K_1 ={T}	h_T	0.044	H1	0.044
K_2 ={O}	h_O	0.022	H2	0.022
K_3 ={U}	h_U	0.166	H3	**0.166**
K_4 ={T, O}	h_T	0.033		
	h_O	0.017	H4	0.05
K_5 ={T, U}	h_T	0.027		
	h_U	0.102	H5	0.129
K_6 ={O, U}	h_O	0.014		
	h_U	0.103	H6	0.117
K_7 ={T, O, U}	h_T	0.023		
	h_O	0.012		
	h_U	0.074	H7	0.109

the total yield of raspberries by concomitant variables not included in the conducted experiment. This is certainly a signal to researchers to look for new concomitant variables that have a significant impact on the total yield of raspberries during the harvesting period.

5. Conclusions

The results obtained using the methods considered here are similar as regards the impact of concomitant variables, and indicate that insolation has the strongest effect on the yield of raspberries. Furthermore:

1. Regression coefficients determined by the growth curve method provide information about the direction of changes in the Y feature (negative or positive) and the subsequent concomitant variables X, while such information cannot be obtained using Hellwig's method.

2. The advantage of Hellwig's method is its simplicity and the fact that it provides additional information about the influence of combinations of concomitant variables on the studied phenomenon. It also signals the potential existence of predicates not included in the experiment but significantly affecting the analyzed feature.

3. In the present study, the conversion from the growth curve model to the regression model consisted in summing *n* observations of feature Y at each of the *p* time points. It is also possible to determine indices h_{li} and H_l based not on sums of observations, but on arithmetic means. This is because the Pearson correlation coefficients are independent of changes in the scale of the variables for which those coefficients are calculated, and therefore the values h_{li} and H_l calculated in Hellwig's method will be the same whether based on sums or on arithmetic means.

REFERENCES

Fujikoshi Y., Rao C.R. (1991): Selection of covariables in the growth curve model. Biometrika 78(4): 779-785.

Fus L., Wesołowska-Janczarek M. (1998): Comparison of two growth curve models with time moving concomitant variables. Colloquium Biometricum 28: 108-115.

Grabiński T., Wydymus S., Zeliaś A. (1982): Metody doboru zmiennych w modelach ekonometrycznych. PWN Warszawa.

Hellwig Z. (1987): Elementy rachunku prawdopodobieństwa i statystyki matematycznej. PWN Warszawa.

Wang S.-G., Liski E., Nummi T. (1999): Two-way selection of covariables in multivariate growth curve models. Linear Algebra and its Applications 289: 333-342.

Wesołowska-Janczarek M., Fus L. (1996): Parameters estimation in the growth curves model with time-changing concomitant variables. In Polish. Colloquium Biometricum 26: 263-277.

Wesołowska-Janczarek M., Fus L., Osypiuk Z. (1997): Zastosowanie metody krzywych wzrostu ze zmiennymi towarzyszącymi do badania owocowania malin. Colloquium Biometricum 27: 269-281.