EXTENDED KEY-FACTOR/KEY-STAGE ANALYSIS FOR LONGITUDINAL DATA

Kohji Yamamura

National Institute for Agro-Environmental Sciences, Tsukuba, Japan.

Key-factor/key-stage analysis was originally a descriptive approach to analyze life tables. However, this method can be extended to analyze longitudinal data in pharmaceutical experiments. By dividing the variance into components, the extended key-factor/key-stage analysis indicates which factor is influential, and through which stage the factor generates its influence in determining the outcome of treatments. Such knowledge helps us in constructing a class of nonlinear longitudinal models that can be interpretable than linear models. Example SAS programs and R programs are provided for the calculation.

Key words: descriptive approach; preliminary analysis; nonlinear model; repeated measurements

1. INTRODUCTION

The techniques for fitting models for longitudinal data (i.e., repeated measurement data or panel data) have been remarkably improved in recent years in several fields of study including medical science and economics (Wolfinger, 1997; Albert, 1999; Diggle et al., 2002; Wooldridge, 2002; Hsiao, 2003; Allison, 2005; Baltagi, 2008). In the analysis of a class of longitudinal data in which the effect of treatment is measured repeatedly over a period, the sophisticated approaches are divided into two categories of models: random coefficients models and covariance structure models (Verbeke and Molenberghs, 2000; Twisk, 2003; Hedeker and Gibbons, 2006). Several analyses use combined models having both random coefficients and covariance structure. Binary variables or count variables are analyzed as well as continuous variables by using the framework of a generalized linear mixed model (GLMM) or a generalized estimating equation (GEE) (Hardin and Hilbe, 2002; Molenberghs and Verbeke, 2005). When we explicitly include the random components of each subject (such as each individual), the models are called subject-specific models (SS models) or random-effects models. On the other hand, if we focus on the average response over all subjects, the models are called population-averaged models (PA models) or marginal models (Zeger et al., 1988; Hardin and Hilbe, 2002; Davis, 2003; Rabe-Hesketh and Skrondal, 2008). SS models are usually solved by using GLMM, while PA models are solved by using GEE. The development of several techniques such as adaptive quadrature methods and Bayesian methods using Markov chain Monte Carlo (MCMC) has increased the potential for estimating sophisticated mixed models (Gelman et al., 2003; Pinheiro and Chao, 2006). We can fit various linear models to the same data. Then, we can select the most appropriate model for prediction by using a criterion such as Akaike’s Information Criterion (AIC) (Akaike, 1973), QAIC, or QIC which is an extension of AIC (Pan, 2001; Burnham and Anderson, 2002).

Various models, including linear models and nonlinear models with covariance structures, should be compared by using an appropriate criterion such as AIC. However, we sometimes find difficulties in interpreting the linear models that were selected among a limited number of candidate models. Some kinds of descriptive approach will be useful as a preliminary tool in creating biological models that
should be compared (Der and Everitt, 2008). The use of summary statistics will be especially informative at the beginnings of analyses (Senn et al., 2000). Matthews et al. (1990) and Everitt (1995) recommended several summary statistics such as overall mean, area under curve (AUC), post-treatment means, and the final value of outcome. Yamamura (1999a) proposed key-factor/key-stage analysis to analyze life-table data by combining ANOVA and the conventional key-factor analysis used by Smith (1973) and Podoler and Rogers (1975). This method can be extended to analyze longitudinal data in pharmaceutical experiments. The extended key-factor/key-stage analysis indicates which factor is influential, and through which stage the factor generates its influence in determining the outcome of treatments. Then, we can construct a nonlinear mixed model that can be biologically more interpretable than linear models.

In this paper, we propose the extended key-factor/key-stage analysis as a preliminary tool to analyze longitudinal data. Example SAS programs and R programs for the analysis are shown in Appendices A and B placed at <http://cse.niaes.affrc.go.jp/yamamura/Key-factor_analysis_program.html>.

2. EXTENDED KEY-FACTOR/KEY-STAGE ANALYSIS

Let us consider a situation in which we observe subjects such as individuals at several different points of time after they have received different treatments. We refer to the treatment as a ‘factor’. We refer to the point of time as ‘stage’. Let $n$ be the total number of subjects, and $g$ be the number of stages. Let $y_j$ be the column vector of observation at the $j$th stage, $\bar{y}_j$ be the column vector of the mean; that is, $\bar{y}_j$ is a $n \times 1$ vector where all elements are the mean of $y_j$.

The difference between the first and final stage observations is expressed by,

$$y_e - y_1 = \sum_{j=1}^{g} (y_{j+1} - y_j),$$

(1)

where $(y_{j+1} - y_j)$ is a column vector of length $n$, expressing the change from the $j$th stage to the $(j + 1)$th stage. Let $k_j$ be the sum of the cross products of $(y_{j+1} - y_j)$ and $(y_e - y_j)$ around the mean:

$$k_j = [(y_{j+1} - y_j) - (\bar{y}_{j+1} - \bar{y}_j)][(y_e - y_j) - (\bar{y}_e - \bar{y}_j)].$$

Let $S(y_e - y_j)$ be the sum of squares of $(y_e - y_j)$ around the mean:

$$S(y_e - y_j) = [(y_e - y_j) - (\bar{y}_e - \bar{y}_j)][(y_e - y_j) - (\bar{y}_e - \bar{y}_j)].$$

Generally, we can divide the variance of the sum of two quantities, $A + B$, into the sum of covariance: $\text{var}(A + B) = \text{cov}(A + B, A + B) = \text{cov}(A, A + B) + \text{cov}(B, A + B)$. Thus, we can express $S(y_e - y_j)$ by the sum of $k_j$, by using Equation (1).

$$S(y_e - y_j) = [(y_e - y_j) - (\bar{y}_e - \bar{y}_j)][(y_e - y_j) - (\bar{y}_e - \bar{y}_j)]$$

$$= \sum_{j=1}^{g} [(y_{j+1} - y_j) - (\bar{y}_{j+1} - \bar{y}_j)][(y_e - y_j) - (\bar{y}_e - \bar{y}_j)]$$

$$= \left( \sum_{j=1}^{g} [(y_{j+1} - y_j) - (\bar{y}_{j+1} - \bar{y}_j)][(y_e - y_j) - (\bar{y}_e - \bar{y}_j)] \right)$$

$$= \sum_{j=1}^{g} k_j.$$  

(2)

Therefore, we can evaluate the influence of each stage on the variability of total change by comparing the value of $k_j$. This decomposition is the conventional key-factor analysis that was used by Smith (1973) and Podoler and Rogers (1975). This decomposition should be called ‘key-stage analysis’ instead of ‘key-factor analysis’, because it divides the variance into the components of each stage instead of components of each factor.
By using a similar argument, we can divide the sum of squares of $y_\varepsilon$ around the mean into several components. Let us define $y_\varepsilon = 0$. Then, Equation (1) yields the following expression of $y_\varepsilon$.

$$y_\varepsilon = \sum_{j=0}^{g-1} (y_{j+1} - y_j)$$  \hspace{1cm} (3)

Let $s_j$ be the sum of the cross products of $(y_{j+1} - y_j)$ and $y_\varepsilon$ around the mean:

$$s_j = \left[(y_{j+1} - y_j) - \bar{y}_j\right](y_\varepsilon - \bar{y}_\varepsilon) \text{ for } j = 1, 2, \ldots, g - 1.$$ Let $s_0$ be the sum of the cross products of $y_\varepsilon$ and $y_\varepsilon$ around the mean: $s_0 = (y_\varepsilon - \bar{y}_\varepsilon)^2$. Let $S(y_\varepsilon)$ be the sum of squares of $y_\varepsilon$ around the mean: $S(y_\varepsilon) = (y_\varepsilon - \bar{y}_\varepsilon)^2$. Then, we can express $S(y_\varepsilon)$ by the sum of $s_0$ and $s_j$ by using Equation (3).

$$S(y_\varepsilon) = (y_\varepsilon - \bar{y}_\varepsilon)^2$$

$$= \left[ \sum_{j=0}^{g-1} (y_{j+1} - y_j) - \sum_{j=0}^{g-1} \bar{y}_j \right] (y_\varepsilon - \bar{y}_\varepsilon)$$

$$= \left[ \sum_{j=0}^{g-1} (y_{j+1} - y_j) - (y_{j+1} - y_j) \right] (y_\varepsilon - \bar{y}_\varepsilon)$$

$$= \sum_{j=0}^{g-1} s_j.$$  \hspace{1cm} (4)

Let us next consider the influence of nominal factors. We first consider a balanced case where the same number of subjects is allocated for all combinations of levels of factors. Let $f$ be the number of factors, $m_i$ be the number of levels in the $i$th factor, and $r$ be the number of subjects (replicates) in each combination of the levels. Then, the total number of subjects ($n$) is given by $n = r \prod_{i=1}^{f} m_i$. We express $y_{ij} - \bar{y}_{ij}$ by a linear form of factors,

$$y_{ij} - \bar{y}_{ij} = Xb_j + e_j,$$  \hspace{1cm} (5)

where $X$ is the full-rank design matrix with elements such as 0 and 1. We are currently considering that the parameters of different factors are mutually orthogonal. $b_j$ is the column vector describing the coefficients at the $j$th stage, $e_j$ is the column vector of the residual component that has zero covariance between factors. Let $X_\varepsilon$ be the design matrix where the elements related to $h$th factor in $X$ are retained while other elements are replaced by 0. Then, Equation (5) is expressed by

$$y_{ij} - \bar{y}_{ij} = \sum_{h=1}^{f} X_\varepsilon b_j + e_j.$$  \hspace{1cm} (6)

We can then express the sum of squares of $y_\varepsilon$ around the mean by

$$S(y_\varepsilon) = (y_\varepsilon - \bar{y}_\varepsilon)^2$$

$$= \left[ \sum_{j=0}^{g-1} X_\varepsilon b_j + e_\varepsilon \right] \left[ \sum_{h=1}^{f} X_\varepsilon b_\varepsilon + e_\varepsilon \right]$$
\[
\sum_{k=1}^{f+1} (X_k b_k)(X_k b_k) + e_k^t e_k = \sum_{k=1}^{f+1} X_k b_k e_k + e_k^t e_k,
\]

because we are currently considering that \( X_k \) are mutually orthogonal; most of the cross-products equal zero. Thus, the sum of squares (SS) is divided into \( f \) factors and residual variance. This is the classical decomposition used in the conventional ANOVA; these \((f+1)\) quantities appear in the SS column of the ANOVA table in analyzing the influence of factors on the final stage \( y_g \). Similarly, the sum of the cross-products of \((y_{j+1} - y_j)\) and \( y_g \) around the mean, which is denoted by \( s_j \), is expressed by

\[
s_j = \sum_{j=1}^{f+1} (X_j (b_{j+1} - b_j))(X_j b_j) + (e_{j+1} - e_j)^t e_g, \quad j = 0, 1, \ldots, (g-1),
\]

where \( b_j = e_j = 0 \). By using Equations (8) and (4), we can divide \( S(y_g) \) into \((f+1) \times g\) components, \( r_{jh} \), that indicate the contribution of each factor through each stage:

\[
r_{jh} = \begin{cases} 
(X_j (b_{j+1} - b_j))(X_j b_j), & j = 0, 1, \ldots, (g-1), \quad h = 1, 2, \ldots, f, \\
(e_{j+1} - e_j)^t e_g, & j = 0, 1, \ldots, (g-1). 
\end{cases}
\]

Each component indicates the amount of contribution about how the \( h \)th factor and the error contribute via the \( j \)th stage in determining the quantity of final stage \( y_g \). We can list these components in an \((f+1) \times g\) table that can be called a ‘key-factor/key-stage table’. This table indicates which factor is influential, and through which stage the factor generates its influence. A graphical representation of the key-factor/key-stage table, which may be called ‘key-factor/key-stage graph’, enables us to interpret the results quickly.

Each term in Equation (9) is algebraically identical to the sum of elements of the \((j+1)\)th column of the corresponding SSCPM (sum of squares and cross products matrix) that appears in performing \( g \)-variate MANOVA in which we treat \( y_j \) and \((y_{j+1} - y_j)\) \((j = 1, 2, \ldots, g-1)\) as different variates. Hence, we can readily calculate these terms by performing MANOVA using some statistical software such as SAS (SAS Institute Inc., 2008a) or JMP (SAS Institute Inc., 2008b). Appendix A-1 and A-3 show example SAS programs to calculate the key-factor/key stage table and the key-factor/key-stage graph. Corresponding R programs are shown in Appendix B. These programs perform the key-factor/key-stage analysis by using Type I SSCPM (SAS Institute Inc., 2008a) when the data are not balanced or when the parameters of different factors are not mutually orthogonal intrinsically (for example, when several numerical factors are included in \( X \)). If the data contain observations that have missing value partially, we should use some kinds of imputation methods. Multiple imputation methods will be especially preferable in this case, because they enable the evaluation of the uncertainty that comes from the prediction of unknown missing values. Recent version of statistical software implements several kinds of multiple imputation methods (e.g., SPSS Inc., 2007; Arbuckle, 2008; SAS Institute Inc., 2008a). Appendix A-2 shows an example SAS program that uses proc MI, in which the multiple imputation is performed via MCMC (Markov chain Monte Carlo) by using the assumption that the variables follow a multivariate normal distribution. Appendix B shows the corresponding R program using the function MICE (van Buuren and Groothuis-Oudshoorn, 2011). Even if the data contain no observations that are partially missing, we should use the multiple imputation method instead of using Type I SSCPM if the data are highly unbalanced, that is, if the number of replications is highly different between levels of factors. Type I SSCPM
will be exactly appropriate only if the dominance between factors is clearly known; for example, if the factor B always works after factor A and if factor B does not influence the preceding effect of factor A.

We can use either of the two types of decomposition, Equations (2) and (4), for key-factor/key-stage analysis. If we are interested in the total change \((y_g - y_1)\) rather than the absolute quantity of \(y_g\), the decomposition about the difference (Equation 2) will be more useful. In this case, we divide \(S(y_g - y_1)\) into \((f + 1) \times (g - 1)\) components in a manner similar to that above. Each component is given by the sum of elements of the \(j\)th column of the corresponding SSCPM that appears in performing \((g - 1)\)-variate MANOVA in which we treat \((y_{j+1} - y_j)\) \((j = 1, 2, \ldots, g - 1)\) as different variates. Thus, the procedure of calculation is the same as that used in the decomposition of \(S(y_g)\), except for omitting \(y_1\) in this case. Then, we can construct another type of key-factor/key-stage table that consists of \((f + 1) \times (g - 1)\) components. This form of analysis is the original key-factor/key-stage analysis that was proposed by Yamamura (1999a).

3. EXAMPLE OF ANALYSES

3.1. Effect of drugs on the respiratory ability of asthma patients

A pharmaceutical company examined effects of three drugs, a standard drug, test drug, and placebo, on respiratory ability of asthma patients (Littell et al., 2000; Littell et al., 2002; Littell et al., 2006). The drugs were assigned to 24 patients each at random. A standard measure of respiratory ability called FEV1 was measured hourly for 8 hours following administration. FEV1 was also measured immediately prior to administration of the drugs. The change of mean FEV1 for each drug is shown in Figure 1.

Littell et al. (2006) started their analyses by identifying the covariance structure for the saturated mean model by using REML in SAS version 9.1 (SAS Institute Inc., 2004). Thus, they first analyzed ‘pure error’ in the terminology of Draper and Smith (1998). They handled the FEV1 before administration as a covariate. They compared five models having different covariance structures: CS covariance, unstructured covariance, AR(1) covariance, Toeplitz covariance, and AR(1) covariance + random intercept. Unstructured covariance yielded the smallest AIC, indicating that it is the best model. However, they adopted AR(1) covariance + random intercept, because it fitted sufficiently well. Then, they proceeded to the analysis of fixed effects and showed that the interaction between drug treatments and time was significant. They used two approaches for analyzing the interaction: one is based on comparisons among the drug × hour combination means and the other is based on comparing regression of the response. The first approach indicated significant differences among drug treatments for hour 1 to 5, but no statistically significant differences among drugs for hour 6, 7, and 8. Second approach indicated that there is no evidence of quadratic regression effects and that the linear regressions are different among the three drug treatments. Then, they finally fitted a 7-parameter model.

\[
y_{it} = b_0y_{it0} + b_1t + u_{i0} + \epsilon_{it},
\]

(10)

where \(y_{i0}\) is FEV1 of the \(i\)th patient of the \(v\)th drug treatment at \(t\) hour \((t = 0, 1, 2, \ldots, 8)\). \(b_0\) is the coefficient for the FEV1 just before administration. \(b_1\) and \(b_2\) are intercept and slope for the \(v\)th drugs, respectively. \(u_{i0}\) is the random intercept having a normal distribution with a mean of zero. \(\epsilon_{it}\) represents residuals that follow a multivariate normal distribution with the covariance structure of AR(1). The fitted model is shown by the broken lines in Figure 1.

We have only one factor, drug treatment, in Equation (5) in performing the extended key-factor/key-stage analysis. Hence, the model for the key-factor/key-stage analysis is given by

\[
y_{iq} - \bar{y}_j = \beta_{ij} + \epsilon_{iq},
\]

(11)
where $y_{vij}$ is the FEV1 of the $i$th patient of the $v$th drug treatment at the $j$th stage ($j = 1, 2, \ldots, 9$); $\bar{y}_j$ is the mean of $y_{vij}$ at the $j$th stage; $\beta_{vj}$ is the influence of the $v$th drug treatment at the $j$th stage; $\varepsilon_{vij}$ is the remaining components that are orthogonal to $\beta_{vj}$. Unlike the residual of the model used in the mixed model approach, neither specific distribution nor specific covariance is assumed for $\varepsilon_{vij}$ in the extended key-factor/key-stage analysis. Notice that the quantities of $t$ and $j$ are different by 1; the observation at $t = 0$ (i.e., time = 0) corresponds to $j = 1$ (i.e., the first observation) in Equation (9).

The results of key-factor/key-stage analysis are summarized in Figure 2, which contains two lines that correspond to the two terms in the right-hand side of Equation (11). The SAS program for the extended key-factor/key-stage analysis is shown in Appendix A-1. The corresponding R program is shown in Appendix B. The error component at $t = 0$ in Figure 2 is given by $e'_1e'_1$; it corresponds to $j = 0$ in the second line of Equation (9). It is a cross-product of the residual components of the first observation and the last observation. This quantity (17.0) is much larger than the quantity of drug effect. It indicates that the influence of the initial variability among individuals before administration was much larger than the effect of drugs. The effect of drugs was thus very limited as compared to the amount of individual variability. The individual variability was nearly fixed at 2 hours after administration (0.0). The drug component of Figure 2 indicates that the effect of drugs emerges only just after the administration. The effect was 3.0 in the period from 0 to 1 hour, but it became consistently negative in the period from 1 to 7 hours. Thus, the effect of drugs consistently decreased in the period from 1 to 7 hours. However, the effect became slightly positive in the period from 7 to 8 hours (0.3). The effect of drugs is nearly zero immediately before administration (0.0). It indicates that the test drugs and placebo were allocated successfully at random to patients.

The extended key-factor/key-stage analysis immediately indicates the following 5-parameter nonlinear model.

$$
\begin{align*}
    y_{vit} &= \begin{cases} 
        b_v + u_v + e_{v0} & \text{for } t = 0 \\
        b_v + u_v + b_v(1 - b_v(t - 1)) + e_{v0} & \text{for } t \geq 1,
    \end{cases}
\end{align*}
$$

where $e_{v0}$ is the residual that follows a normal distribution at $t$ hour. Figure 2 indicated that the initial
variation among individuals is extremely large. Hence, the individual variability should be first incorpo-
rated by $b_0 + u_{vi}$ where $b_0$ is the intercept before administration, $u_{vi}$ is the effect of the $i$th individual in the
$v$th drug. Figure 2 also indicated that the influence of drugs occurs only just after the administration.
Hence, the influence of the $v$th drug between $t = 0$ and 1 should be incorporated by parameter $b_{1v}$. Figure
2 further indicated that the influence of drugs almost consistently decreased for $t > 1$. Hence, $b_{1v}$ should
be multiplied by a slope $(1 - b_2(t - 1))$.

We obtain the following simple expression of Equation (12), by eliminating $b_0 + u_{vi}$ in Equation
(12), by substituting the equation for $t = 0$ into that for $t \geq 1$. This model contains only 4 parameters of
fixed effects.

$$y_{vi} = y_{o0} + b_{1v}(1 - b_2(t - 1)) + (e_{vi} - e_{00}) \quad \text{for } t \geq 1.$$  \hspace{1cm} (13)

We fitted Equations (10) and (13) using ML in NLINMIX macro of SAS version 9.2 assuming AR(1)
covariance + random intercept that was adopted by Littell et al. (2006). The SAS program for perform-
ing the estimation of nonlinear mixed model is shown in Appendix C-1. The AIC was 198.2 for Equation
(10) and 194.2 for Equation (13). The BIC was 221.0 for Equation (10) and 210.1 for Equation (13).
Thus, both AIC and BIC indicated that Equation (13) having only 4 parameters has higher power of
prediction by avoiding the overfitting, as compared to Equation (10) having 7 parameters. Furthermore,
the biological meaning of Equation (13), i.e., Equation (12), is clearer than that of Equation (10).

### 3.2. Influence of mother’s height on daughter’s height

Goldstein (1979) analyzed the influence of mothers’ heights on the growth curves of their
daughters. The heights of twenty girls were measured yearly from 6 to 10 years old. The mothers were
divided into three groups depending on their height: the short mother group (less than 155cm), medium
mother group (155-164 cm), and tall mother group (larger than 164 cm). The same data were analyzed
by Verbeke and Molenberghs (1997) and Fujikoshi, Kan, and Hijikata (2008). We use the corrected data
from Verbeke and Molenberghs (1997). The numbers of mothers in the three groups were 6, 7, and 7,
respectively. The change of mean height for each group is shown in Figure 3.

Verbeke and Molenberghs (1997) started their analyses by identifying the covariance structure
of residuals for the saturated mean model by using REML in SAS versions 6.11 and 6.12 (SAS Institute,
1997). They applied three kinds of covariance structures for residuals: unstructured, heterogeneous
Toeplitz, and heterogeneous AR(1). AIC indicated that the unstructured covariance is the best model.
However, they adopted heterogeneous AR(1), because it fitted sufficiently well. Then, they proceeded to analyze the mean structure by using ML with variance of heterogeneous AR(1). After several trials, they finally adopted the linear regression with a common slope for the short mother group and medium mother group, while they adopted the split-line regression for the tall mother group. The final model was expressed by a 6-parameter model

\[
\begin{align*}
\text{short mother group} & : y_{vit} = b_1 + b_2(t - 6) + e_{vit} \\
\text{medium mother group} & : y_{vit} = b_1 + b_2(t - 6) + e_{vit} \\
\text{tall mother group} & : y_{vit} = b_1 + b_2(t - 6) + b_3 \max\{t - 9, 0\} + e_{vit}
\end{align*}
\] (14)

where \(y_{vit}\) is the height of the \(i\)th girl in the \(v\)th mother group at age \(t\) \((t = 6, 7, 8, 9,\) and 10). \(b_1\) to \(b_6\) are fixed parameters. \(e_{vit}\) represents residuals that follow a multivariate normal distribution with the covariance structure of heterogeneous AR(1). The fitted model is shown by the broken lines in Figure 3.

We have again only one factor, mother height, in Equation (5) in performing the extended key-factor/key-stage analysis. Hence, the model for the key-factor/key-stage analysis is given by

\[
y_{vij} - \bar{y}_j = \beta_{vj} + e_{vij},
\] (15)

where \(y_{vij}\) is the height of the \(i\)th girl in the \(v\)th mother group at the \(j\)th stage \((j = 1, 2, 3, 4,\) and 5); \(\bar{y}_j\) is the mean of \(y_{vij}\) at the \(j\)th stage; \(\beta_{vj}\) is the influence of the \(v\)th drug treatment at the \(j\)th stage; \(e_{vij}\) is the remaining components that are orthogonal to \(\beta_{vj}\). Unlike the residual of the model used in the mixed model approach, neither specific distribution nor specific covariance is assumed for \(e_{vij}\) in the extended key-factor/key-stage analysis.

The SAS program for the extended key-factor/key-stage analysis is shown in Appendix A-2. The corresponding R program is shown in Appendix B. The number of subjects is unbalanced between mother groups (6, 7, and 7). The program in Appendix A-2 uses the multiple imputation method instead of using Type I SSCPM to handle the unbalanced data. The results of key-factor/key-stage analysis are summarized in Figure 4, which contains two lines that correspond to the two terms in the right-hand side of Equation (15). The mother component in Figure 4 shows that the influence of mother height is much...
larger before the age of 6 than after the age of 6 (315 and 150, respectively). The error component before the age of 6, that is, the individual variability before the age of 6, is also fairly large (135). If we focus on the growth occurring during the observed period from the age of 6 to the age of 10, the influence of mother height is clearly positive between the ages of 6 and 9, while the influence is nearly zero in the period between the ages of 9 and 10 (3). Thus, the influence of mother height is amplified from the age of 6 to the age of 9 but not from the age of 9 to the age of 10.

By incorporating the knowledge obtained from the key-factor/key-stage analysis, we obtain the following 5-parameter nonlinear model.

\[
y_{it} = \begin{cases} 
  b_1 + b_3(t - 6) + e_{it} & \text{short mother group} \\
  b_1 + b_3(t - 6) + b_4(1 + b_5 \min(t - 6,3)) + e_{it} & \text{medium mother group} \\
  b_1 + b_3(t - 6) + b_4(1 + b_5 \min(t - 6,3)) + e_{it} & \text{tall mother group} 
\end{cases}
\]  

Figure 4

Key-factor/key-stage graph, showing the contribution of mother height and individual variability in determining the daughter height.

The first two terms in the right-hand side indicate the 'baseline' influence of age, that are common for all groups. Figure 4 indicated that the influence of mother height is very large before the age of 6. Hence the influence of mother height before the age of 6 should be first incorporated by using two parameters, \( b_3 \) and \( b_5 \). Figure 4 indicated that the influence of mother height was amplified until the age of 9. Hence, \( b_3 \) and \( b_5 \) should be amplified until the age of 9, by multiplying a slope \((1 + b_4 \min(t - 6,3))\). We fitted Equations (14) and (16) using ML in NLINMIX macro of SAS version 9.2 assuming heterogeneous AR(1) that was adopted by Verbeke and Molenberghs (1997). The SAS program for performing the estimation of nonlinear mixed model is shown in Appendix C-2. The AIC was 302.8 for Equation (14) and 299.8 for Equation (16). The BIC was 314.7 for Equation (14) and 310.8 for Equation (16). Thus, both AIC and BIC indicated that Equation (16) having 5 parameters was superior to Equation (14) having 6 parameters in predicting the height of daughters. Furthermore, the biological meaning of Equation (16) is clearer than that of Equation (14).

### 3.3. Influence of plant density on the abundance of insects

The previous examples contain one factor each: drugs and mother height, respectively. However, key-factor/key-stage analysis will be especially useful when multiple factors are included in the observation. For example, in pharmaceutical experiments, we can divide patients into various groups by using information such as genders, age, habit of smoking or drinking. We can incorporate the information as factors into key-factor/key-stage analysis. In usual ANOVA, the influence of covariates is separated to
yield the adjusted estimate of treatment. The extended key-factor/key-stage analysis uses completely the
same partition of variance (Equation 7) that is used in ANOVA. Hence, the key-factor/key-stage graph
indicates the adjusted influence of treatment separated from the influence of covariates. The graph also
indicates the adjusted influence of covariates separated from the influence of treatment and other covar-
iates. Then, if a factor is influential, we can quantitatively judge when the factor generates its influence.
We next show an example analysis of data that contain two factors.

Yamamura and Yano (1999) examined the influence of plant density and block (i.e., spatial po-
sition) on the abundance of insects: larvae of beet semi-looper Autographa nigrisigna (Walker) (Noctu-
idae). Cabbage plants were planted with two levels of densities: sparse plots (0.25 plants/m²) and dense
plots (4 plants/m²). The experimental plots (each 10×10 m) were replicated with 4 blocks separated by
weeded ground. They estimated the logarithmic abundance of insects (per plant) for 1st, 2nd, 3rd, and
4th stage larvae. In this case, we have two factors, plant density and block, in Equation (5) in performing
the extended key-factor/key-stage analysis. The model for the key-factor/key-stage analysis is given by

\[ y_{vj} = \beta_v j + \gamma w j + \varepsilon_{vwj}, \]  

where \( y_{vj} \) is the logarithmic abundance of insects at the \( v \)th plant density treatment of the \( w \)th block at
the \( j \)th stage; \( \beta_v j \) is the mean of \( y_{vj} \) at the \( j \)th stage; \( \gamma w j \) is the influence of the \( w \)th plant density treatment
at the \( j \)th stage; \( \varepsilon_{wj} \) is the influence of the \( w \)th block at the \( j \)th stage; \( \varepsilon_{wv} \) is the remaining components that
are orthogonal to \( \beta_v j \) and \( \gamma w j \). The key-factor/key-stage graph is shown by Figure 5. The SAS program and
R program for the extended key-factor/key-stage analysis are given in Appendix A-3 and Appendix B,
respectively. Figure 5 contains three lines that correspond to the three terms in the right-hand side of
Equation (17). The graph indicates that the influence of plant density changed in the course of the growth
of insects, from showing a negative influence at the stage 1 (\(-0.066\)) to a positive influence in the latter
stages (0.007, 0.018, and 0.048). Such a negative influence of plant density on the initial abundance of
insects per plant is called ‘resource diffusion effect’ (Yamamura, 1999b, 2002). The block influenced in
all stages; the influence was generally larger than that of plant density. The error component from the
stage 1 to stage 2 is negative (\(-0.114\)). Such ‘negative variability’ indicates that the variability in stage
1 was adjusted between stage1 and stage 2, and that the variability in stage 2 became smaller than the
variability in stage 1. The error component from the stage 3 to stage 4 was very large (0.230). It indicates
that the abundance of the 4th stage larvae was considerably influenced by the unknown variability that
occurred from the stage 3 to stage 4.

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-

Figure 5 Key-factor/key-stage graph, showing the contribution of plant density, block, and residual variabi-
4. DISCUSSION

We proposed the extended key-factor/key-stage analysis as a handy tool for the preliminary analysis. The analysis indicates a class of nonlinear longitudinal models that can be interpretable than linear models; we showed it by using two well-known longitudinal data. We can perform the extended key-factor/key-stage analysis by using SSCPM that is calculated in MANOVA by using statistical software such as SAS and R. We provided an example SAS macro program and an example R function in Appendix A and B, respectively. Numerical factors are allowed as well as nominal factors and interactions in these programs. A multiple imputation method can be used when the data contains missing values or an unbalanced number of subjects.

Key-factor/key-stage analysis further clarifies the stage where the individual variability is amplified or damped. Thus, the key-factor/key-stage analysis will be also useful in studying the source of individual variability. The error component in Figures 2 and 4 indicated that the amount of increase of individual variability becomes smaller with time in these examples. Hence, we did not add variance-amplifying terms in these cases. If the error component in key-factor/key-stage graph showed increasing trends, however, we should add variance-amplifying terms such as random slopes.

AKNOWLEDGMENTS

We would like to thank two anonymous reviewers for their comments that greatly improved the manuscript.

REFERENCES


