Listy Biometryczne - Biometrical Letters
Vol. 37(2000), No. 1, 13-19

On the use of determination coefficient to describe goodness of fit of assessed growth curves

Mirosława Wesołowska-Janczarek

Instytut Zastosowań Matematyki, Akademia Rolnicza w Lublinie,
Akademicka 13, 20-934 Lublin

SUMMARY

In the paper a way of using the well-known determination coefficient for growth curves, assessed by the multivariate Potthoff-Roy's method, is proposed. Modified definitions of determination coefficients for each of the object levels and of the mean determination coefficient are given. The latter characterizes goodness of fit for all curves that are calculated by the given method. The problem is illustrated by two examples on real data.

KEY WORDS: determination coefficient, mean determination coefficient, goodness of fit, growth curves.

1. Introduction

The fit of polynomial regression of a studied feature as a time function for several groups of data simultaneously can be done by using the known growth curve method. A division into the groups can be conditioned by objects that were taken into consideration in the study. For each of the objects measurements are taken at the same p time points and on any number of experimental units $r_j (r_j > 1)$. In this method the following assumptions are made: vectors of observations for each of the experimental units are independent and have the same covariance matrices.

Under these assumptions a question arises, what is a measure of goodness of fit of the curves obtained by this method or whether functions obtained by the growth curve method are good approximation of the dynamic of the studied feature over the considered time period.

In the case of one-variable polynomial regression the determination coefficient R^2 is used for this purpose.

The application of this coefficient for the study of goodness of fit of individual functions obtained by Potthoff-Roy's method is proposed in this paper. This proposition is illustrated by two examples.

2. Determination coefficient R^2 for one-variate polynomial regression

In the one-variable polynomial regression method the function

$$y(t) = \beta_0 + \beta_1 t^1 + \beta_2 t^2 + \dots + \beta_{q-1} t^{q-1} \tag{1}$$

is fitted to the observed data $y_1, y_2, \dots, y_p$, which can be the values of explored feature at p successive time points $t_1, t_2, \dots, t_p$. Then $y_i = y(t_i)$ for $i = 1, 2, \dots, p$. The suitable model can be written in the matrix notation as

$$\mathbf{y} = \mathbf{T}'\boldsymbol{\beta} + \mathbf{e}, \tag{2}$$

where $\mathbf{y} = [y_1, y_2, \dots, y_p]'$ is the $p \times 1$ vector of observations, $\boldsymbol{\beta}$ is the $q \times 1$ vector of fixed unknown regression coefficients, the matrix $\mathbf{T}$ takes the form

$$\mathbf{T} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_p \\ t_1^2 & t_2^2 & \dots & t_p^2 \\ \dots & \dots & \dots & \dots \\ t_1^{q-1} & t_2^{q-1} & \dots & t_p^{q-1} \end{bmatrix} \tag{3}$$

and $\mathbf{e}$ is the $p \times 1$ vector of random errors. Moreover, the following assumptions are adopted: $\mathrm{E}(\mathbf{y}) = \mathbf{T}'\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{y}} = \sigma^2 \mathbf{I}_p$, where E is the expected value operator and $\boldsymbol{\Sigma}_{\mathbf{y}}$ is the covariance matrix of vector $\mathbf{y}$. It is very well known that $\widehat{\boldsymbol{\beta}} = (\mathbf{T}\mathbf{T}')^{-1}\mathbf{T}\mathbf{y}$ is the estimator of regression coefficients vector in (2).

Let $y_i^* = y^*(t_i)$ be the assessed value of a feature from equation (1) using $\widehat{\beta}_i$ instead of β_i ($i = 1, \dots, p$). In the matrix notation, we have $\mathbf{y}^* = \mathbf{T}'\widehat{\boldsymbol{\beta}}$ with

$$\mathbf{y}^* = [y_1^*, y_2^*, \dots, y_p^*]'. \tag{4}$$

The measure of goodness of fit of the function $y^*(t) = \widehat{\beta}_0 + \widehat{\beta}_1 t^1 + \dots + \widehat{\beta}_{q-1} t^{q-1}$ to observed points is the determination coefficient R^2, defined in the following way (see Seber, 1977, p. 111):

$$R^2 = \frac{\left[\sum_{i=1}^{p} (y_i - \overline{y})(y_i^* - \overline{y}^*)\right]^2}{\left[\sum_{i=1}^{p} (y_i - \overline{y})^2\right]\left[\sum_{i=1}^{p} (y_i^* - \overline{y}^*)^2\right]} = \frac{(nS_{yy^*})^2}{nS_y^2 \cdot nS_{y^*}^2}, \tag{5}$$

where $\overline{y} = \frac{1}{p}\sum_{i=1}^{p} y_i$, $\overline{y}^* = \frac{1}{p}\sum_{i=1}^{p} y_i^*$ and $\overline{y} = \overline{y}^*$.

The values of this coefficient are $0 \leq R^2 \leq 1$, and the closer the value of R^2 is to 1, the better the fit. The coefficient R^2 can be equal to 1 only if one observation corresponds to each value of t and cannot be equal to 1 if at any point t there is more then one value of y, regardless of goodness of fit of the curve (see Draper and Smith, 1998).

3. Determination coefficient as a measure of goodness of fit of growth curves

The growth curve model given by Potthoff and Roy (1964) can be treated as a generalization of the polynomial regression considered in the previous chapter. This generalization can take the following directions.

1^o. Not one curve but a few curves are fitted simultaneously for a given matrix of observations Y. This data are obtained from each of n independent experimental units devided into a groups at the same p time points.

2^o. In each of the groups there are r_j $(j = 1, ..., a)$ experimental units, $r_j > 1$. These groups can have the same number of units and $\sum_{j=1}^{a} r_j = n$.

3^o. Since the measurements are taken on each unit at p time points, the observations in each vector are correlated with the same covariance matrix for each p-vector of observations.

The following notation will be used. Let $\mathbf{Y} = [y_{ijk}]$ be the $n \times p$ matrix of observations $(i = 1, ..., p;\quad j = 1, ..., a;\quad k = 1, ..., n)$. The $n \times a$ matrix $\mathbf{A}$ consisting of zeros and ones is the design matrix that divides units into a groups, and matrix $\mathbf{T}$ of the form given in (3) defines a polynomial relation between the feature and time. Then, the growth curve model is of the form

$$\mathbf{Y} = \mathbf{ABT} + \mathbf{E}, \tag{6}$$

where $\mathbf{B}$ is the $a \times q$ matrix of fixed unknown coefficients of the assessed polynomial curves of the form (1) for each of the a groups and $\mathbf{E}$ is the matrix of random errors.

The maximum likelihood estimators of the coefficients of the growth curves and of the covariance matrix (see for example Srivastava and Khatri, 1979) are given by

$$\widehat{\mathbf{B}} = (\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'\mathbf{Y}\widehat{\mathbf{\Sigma}}^{-1}\mathbf{T}'(\mathbf{T}\widehat{\mathbf{\Sigma}}^{-1}\mathbf{T}')^{-1} \tag{7}$$

and

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{Y}'[\mathbf{I}_n - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}']\mathbf{Y}. \tag{8}$$

Then, the assessed values from the fitted curves can be obtained as

$$\mathbf{Y}^* = \mathbf{A}\widehat{\mathbf{B}}\mathbf{T}. \tag{9}$$

For the clarity of further considerations, the units will be divided into a independent groups of observations with the same number of units in each group, as in the balanced one-way classification. Then $r_j = r$. For example, we have a varieties with r plants for each variety and for each of the plants the observations are taken at p time points. The generalization to other classifications is simple.

In this case $\mathbf{A} = \mathbf{I}_a \otimes \mathbf{J}_r$, where $\mathbf{I}_a$ is the $a \times a$ identity matrix, $\mathbf{J}_r$ is a vector of r ones and $\otimes$ denotes the Kronecker product of two matrices, and $n = ar$.

To define a determination coefficient for each of the a groups, the following notation will be used. Let $\mathbf{Z} = [z_{ij}]$, $i = 1, ..., p$, $j = 1, ..., a$, be the $a \times p$ matrix of means z_{ij} of observations for the j-th group and the i-th time point, $z_{ij} = \frac{1}{r}\sum_{k=1}^{r} y_{ijk}$, and $\overline{z}_j = \frac{1}{p}\sum_{i=1}^{p} z_{ij}$ be the general mean for the j-th group, whereas Y_{ij}^* be the assessed value of a feature with a fitted curve for the j-th group at the i-th time point and $\overline{Y}_j^*$ be the mean of the values for the i-th group, which is $\overline{Y}_j^* = \frac{1}{p}\sum_{i=1}^{p} Y_{ij}^*$. Then the following definition can be formulated.

Definition 1. The determination coefficient of growth curve for the j-th group of observations is defined as

$$R_j^2 = \frac{\left[\sum\limits_{i=1}^{p}(z_{ij} - \overline{z}_j)(Y_{ij}^* - \overline{Y}_j^*)\right]^2}{\sum\limits_{i=1}^{p}(z_{ij} - \overline{z}_j)^2 \sum\limits_{i=1}^{p}(Y_{ij}^* - \overline{Y}_j^*)^2} = \frac{(nS_{zY^*})^2}{nS_z^2 \cdot nS_{Y^*}^2} \quad \text{for} \quad j = 1, ..., a. \tag{10}$$

The values of these coefficients are $0 \leq R_i^2 \leq 1$. Moreover, a general measure of goodness of fit curves obtained by Potthoff-Roy's method can be defined as a mean determination coefficient of the following form.

Definition 2. The mean determination coefficient is the average of the determination coefficients obtained by the given growth curve method for studied objects; that is,

$$\overline{R}^2 = \frac{1}{a}\sum_{j=1}^{a} R_j^2.$$

The above coefficient can take values $0 \leq \overline{R}^2 \leq 1$, too.

The coefficients presented above can also be used for exploration of goodness of fit functions obtained by other growth curve methods, for example, by growth curve methods with time moving concomitant variables.

4. Examples

Two examples illustrating the application of the new measure to assess goodness of fit of a growth curve obtained by Potthoff-Roy's method will be given. For this purpose, real data will be used.

4.1. Sugar concentration in the roots of sugar beet

In this example, taken from a paper by Wesołowska-Janczarek (1996), the dynamic of sugar concentration in the roots of sugar beet for seven cultivars is compared. The following cultivars were considered: 1. Rejana, 2. Maria, 3. Danpol, 4. Freja, 5. Kawejana, 6. PN Mono-4 and 7. PN Mono-1. The root samples were taken at five dates: 20.08, 31.08, 10.09, 20.09 and 30.09 and the contents of sugar were measured for those samples. The dates were indicated as time points 1, 11, 21, 31 and 41.

The average values for each of the seven cultivars at successive time points are shown in matrix **Z** and the average values of concentration for each of the cultivars are shown in vector $\overline{\mathbf{Z}}$. They are as follows:

$$\mathbf{Z} = \begin{bmatrix} 38.62 & 62.77 & 65.25 & 82.22 & 127.52 \\ 47.62 & 61.47 & 82.70 & 94.81 & 95.79 \\ 35.52 & 46.87 & 54.52 & 65.95 & 70.41 \\ 55.72 & 82.28 & 89.22 & 109.58 & 109.26 \\ 42.80 & 62.95 & 72.61 & 85.00 & 83.10 \\ 36.52 & 46.04 & 62.47 & 77.36 & 77.27 \\ 43.97 & 68.25 & 81.00 & 71.62 & 104.65 \end{bmatrix}, \quad \overline{\mathbf{Z}} = \begin{bmatrix} 75.28 \\ 76.48 \\ 54.66 \\ 89.21 \\ 69.29 \\ 59.93 \\ 73.91 \end{bmatrix}.$$

Using Potthof-Roy's method, the functions of time were assessed. They are the third degree polynomials for each of the cultivars in the studied period. This functions are as follows:

$$f_1(t) = 34.212 + 4.474t - 0.232t^2 + 0.0043t^3,$$

$$f_2(t) = 46.901 + 0.755t + 0.079t^2 - 0.0017t^3,$$

$$f_3(t) = 34.377 + 1.011t + 0.002t^2 - 0.0001t^3,$$

$$f_4(t) = 52.476 + 2.798t - 0.052t^2 + 0.0004t^3,$$

$$f_5(t) = 40.493 + 2.105t - 0.003t^2 - 0.0001t^3,$$

$$f_6(t) = 36.404 - 0.018t + 0.098t^2 + 0.0018t^3,$$

$$f_7(t) = 39.349 + 5.434t - 0.254t^2 + 0.0039t^3.$$

Moreover, matrix $\mathbf{Y}^*$ and vector $\overline{\mathbf{Y}}^*$ contain estimated values of the trait from the above given functions at successive time points and their averages for all cultivars.

They are as follows:

$$\mathbf{Y}^* = \begin{bmatrix} 38.46 & 61.08 & 65.68 & 78.05 & 124.01 \\ 47.73 & 62.50 & 81.85 & 95.58 & 93.49 \\ 35.39 & 45.61 & 55.56 & 64.66 & 72.30 \\ 55.22 & 77.49 & 92.01 & 101.16 & 107.35 \\ 42.59 & 63.15 & 82.45 & 99.89 & 114.86 \\ 36.48 & 45.67 & 62.57 & 76.40 & 76.35 \\ 44.53 & 73.58 & 77.57 & 79.89 & 103.96 \end{bmatrix}, \quad \overline{\mathbf{Y}}^* = \begin{bmatrix} 73.46 \\ 76.23 \\ 54.70 \\ 86.65 \\ 80.59 \\ 59.49 \\ 75.91 \end{bmatrix}.$$

Using data from matrices $\mathbf{Z}$, $\overline{\mathbf{Z}}$, $\mathbf{Y}^*$ and $\overline{\mathbf{Y}}^*$, the determination coefficients for successive cultivars were calculated from (10). The values of those coefficients are:

$$\begin{aligned} R_1^2 &= 0.9980, \quad R_2^2 = 0.9966, \quad R_3^2 = 0.9919, \quad R_4^2 = 0.9657, \\ R_5^2 &= 0.9191, \quad R_6^2 = 0.9996 \quad \text{and} \quad R_7^2 = 0.9539. \end{aligned}$$

Moreover, the mean determination coefficient is $R^2 = 0.9750$.

The values of those coefficients prove that the fit of this curves is good for each of the cultivars, and we can confirm that time is a main factor deciding about the sugar concentration in the sugar beet roots in the studied period. Moreover, the fit of the curves by the use of this method is good, as the high value of $\overline{R}$ suggests.

4.2. Dynamic of raspberries fruitbearing

This data were collected by the Department of Orcharding, Agriculture University in Lublin, in 1987, in the second year of fruitbearing of raspberries. The fruitbearing period lasted 36 days in that year. Fruits were collected at 14 time points from 16 cultivars used in the experiment.

On the basis of this data, the fourth degree polynomial functions were assessed by Potthoff-Roy's method for each of the cultivars and determination coefficients for each cultivar were obtained by the method given in Section 2. The values of these coefficients are:

$$\begin{array}{llll} R_1^2 = 0.3480 & R_2^2 = 0.1890 & R_3^2 = 0.3561 & R_4^2 = 0.1634 \\ R_5^2 = 0.0087 & R_6^2 = 0.4691 & R_7^2 = 0.0489 & R_8^2 = 0.0520 \\ R_9^2 = 0.1625 & R_{10}^2 = 0.3197 & R_{11}^2 = 0.6092 & R_{12}^2 = 0.4106 \\ R_{13}^2 = 0.0025 & R_{14}^2 = 0.0450 & R_{15}^2 = 0.0046 & R_{16}^2 = 0.4269 \end{array}$$

and the value of the mean determination coefficient is $\overline{R}^2 = 0.2260$. In this case, a very strong difference between the values of the coefficients can be noticed, from 0.0025 in the case of cultivar number 13 to 0.6092 for cultivar number 11. It means that the course of fruitbearing of different cultivars of this plant is varied to a great degree and there is a strong influence of other factors, apart from time, on this feature. The cultivar number 11 is the most stable and the successive most stable ones are 6, 16 and 12. On the basis of the value of $\overline{R}^2$ it is easy to see that, in general, the dynamic of raspberries fruitbearing is poorly described by this functions and it is necessary to seek other factors determining the course of fruitbearing of this plant.

5. Conclusions

The examples confirm that the determination coefficients defined in the paper provide a good measure of the degree of the fit of polynomials obtained by means of growth curve method to the actual course of real processes. The low values of the obtained coefficients may, therefore, suggest that either additional concomitant variables should be taken into consideration or an entirely different method, such as the random coefficient growth curve method should be chosen.

References

Draper, N.R., Smith, H. (1998). *Applied Regression Analysis.* 3rd ed., J. Wiley, New York.

Potthoff, R.F., Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51**, 313-326.

Seber, G.A.F. (1977). *Linear Regression Analysis.* J. Wiley, New York.

Srivastava, M.S., Khatri, C.G. (1979). *An Introduction to Multivariate Statistics.* Elsevier North Holland, New York.

Wesołowska-Janczarek, M. (1996). Zastosowanie krzywych wzrostu w rolnictwie. *Fragmenta Agronomica* XIII, **3**(57), 6-53.

Received 20 October 1999; revised 10 March 2000

O zastosowaniu współczynnika determinacji do określenia dobroci dopasowania oszacowanych krzywych wzrostu

Streszczenie

Zaproponowano sposób zastosowania znanego współczynnika determinacji, używanego w przypadku wielomianowej regresji jednej zmiennej, dla funkcji dopasowanych wielozmienną metodą krzywych wzrostu Potthoffa-Roy'a. Podano zmodyfikowaną definicję tego współczynnika dla grup obserwacji oraz średniego współczynnika determinacji charakteryzującego dobroć dopasowania łącznie wszystkich krzywych uzyskanych podaną metodą. Problem zilustrowano dwoma oryginalnymi przykładami wykorzystującymi dane pozyskane w badaniach nad roślinami.

Słowa kluczowe: współczynnik determinacji, średni współczynnik determinacji, dobroć dopasowania, krzywe wzrostu.