Applications of Mathematics

Karel Zvára

Analysis of variance as regression model with a reparametrization restriction

Applications of Mathematics, Vol. 37 (1992), No. 6, 453–458

Persistent URL: `http://dml.cz/dmlcz/104523`

ANALYSIS OF VARIANCE AS REGRESSION MODEL WITH A REPARAMETRIZATION RESTRICTION

KAREL ZVÁRA

(Received October 10, 1991)

Summary. Let us consider the linear model covering the one-way classification as a special case. In the paper the relationship between testing of some linear hypothesis and estimating of parameters in the linear model by common software packages is examined.

Keywords: ANOVA; Scheffé theorem; reparametrization

AMS classification: 62J10

1. MODEL

Let us consider a classical linear model

$$y_{ij} = \mu + \alpha_i + \mathbf{z}'_{ij}\boldsymbol{\beta} + e_{ij}, \tag{1}$$

where $e_{ij} \sim N(0,\sigma^2)$ for $j = 1,\dots,n_i$, $i = 1,\dots,k$, are independent random variables, $\mathbf{z}_{ij}$ are known real vectors satisfying

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \mathbf{z}_{ij} = 0, \tag{2}$$

$\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_k)'$ and $\boldsymbol{\beta} = (\beta_1,\dots,\beta_p)'$ are unknown parameters. Let us denote

$$n = \sum_{j=1}^{k} n_j.$$

The relation (1) can be rewritten in a matrix form as

$$\mathsf{y} = \mathsf{1}\mu + \mathsf{X}\boldsymbol{\alpha} + \mathsf{Z}\boldsymbol{\beta} + \mathsf{e}, \tag{3}$$

where $\mathsf{1} = (1, 1, \ldots, 1)'$, $\mathsf{Z} = (\mathbf{z}_{11}, \ldots, \mathbf{z}_{1n_1}, \ldots, \mathbf{z}_{kn_k})'$ and

$$\mathsf{X} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

The assumption (2) can be written as

$$\mathsf{1}'\mathsf{Z} = \mathsf{0}'. \tag{4}$$

Besides of it holds

$$\mathsf{X1} = \mathsf{1}. \tag{5}$$

Let us assume that

$$h(\mathsf{X}, \mathsf{Z}) = k + p, \tag{6}$$

therefore the matrix (X, Z) has linearly independent columns.

In case $\boldsymbol{\beta} = 0$ the relation (3) gives a common model of analysis of variance. Moreover, a Wishart's example cited in Rao (1973), p. 291, can be rewritten in the form (3), too. In this example growth rates of 30 pigs in dependence on initial weights, pen, sex and type of food are examined. Let effect α_i corresponds to i-th pen. The other effects are included in the vector $\boldsymbol{\beta}$. The requirement (4) is not restrictive. Instead of vectors $\mathbf{z}_{ij}$ we can consider vectors $\mathbf{z}_{ij} - \bar{\mathbf{z}}$ and the parameter μ we can replace by $\mu_0 = \mu - \bar{\mathbf{z}}'\boldsymbol{\beta}$.

Elements of vector $\boldsymbol{\alpha}$ are estimable if it is fulfilled the known reparametrization condition

$$\mathsf{n}'\boldsymbol{\alpha} = 0, \tag{7}$$

where the vector of weights $\mathsf{n} = (n_1, ..., n_k)'$ satisfies

$$\mathsf{1}'\mathsf{X} = \mathsf{n}'. \tag{8}$$

For computation of estimates of parameters μ, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ as well as for computation of the residual sum of squares we can use a common program for multiple linear regression, when we add a pseudo observation determined by the condition (7) to the used observations (e.g. Gentleman (1974), Zvára (1989)). Together with the primary observations the pseudo observation create the schema

$$\left(\begin{pmatrix} \mathsf{Z}, & 1, & \mathsf{X} \\ 0', & 0, & \varphi\mathsf{n}' \end{pmatrix}, \begin{pmatrix} \mathsf{y} \\ 0 \end{pmatrix}\right),$$

where $\varphi \neq 0$ is any real constant. The system of normal equations can be written as

$$\begin{pmatrix} \mathsf{Z}'\mathsf{Z}, & 0, & \mathsf{Z}'\mathsf{X} \\ 0' & n, & \mathsf{n}' \\ \mathsf{X}'\mathsf{Z}, & \mathsf{n}, & \mathsf{X} + \varphi^2\mathsf{n}\mathsf{n}' \end{pmatrix} \begin{pmatrix} \mathsf{b} \\ m \\ \mathsf{a} \end{pmatrix} = \begin{pmatrix} \mathsf{Z}'\mathsf{y} \\ 1'\mathsf{y} \\ \mathsf{X}'\mathsf{y} \end{pmatrix}. \tag{9}$$

According to Scheffé theorem (Scheffé (1959)), that makes base for described method of parameters m, a, b computation, the matrix of the system normal equations (9) is regular.

Using the routine way of computation, we usually receive besides the mentioned estimates their standard errors and also t-statistics. In the paper we will show one possible interpretation of these t-statistics.

2. Variance matrix

Let us compute the variance matrix $\sigma^2\mathsf{G}$ of the estimate a. For that reason we need to find out the right lower submatrix of the inverse matrix of system (9). According to the well known formula (e.g. Rao (1973), p. 33)

$$\begin{pmatrix} \mathsf{A}, & \mathsf{B} \\ \mathsf{B}', & \mathsf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathsf{A}^{-1} + \mathsf{F}\mathsf{E}^{-1}\mathsf{F}', & -\mathsf{F}\mathsf{E}^{-1} \\ -\mathsf{E}^{-1}\mathsf{F}', & \mathsf{E}^{-1} \end{pmatrix}, \tag{10}$$

where

$$\mathsf{E} = \mathsf{D} - \mathsf{B}'\mathsf{A}^{-1}\mathsf{B}, \qquad \mathsf{F} = \mathsf{A}^{-1}\mathsf{B},$$

we will receive

$$\begin{aligned} \mathsf{G}^{-1} &= \mathsf{X}'\mathsf{X} + \varphi^2\mathsf{n}\mathsf{n}' - (\mathsf{X}'\mathsf{Z}, n) \begin{pmatrix} \mathsf{Z}'\mathsf{Z}, & 0 \\ 0', & n \end{pmatrix}^{-1} \begin{pmatrix} \mathsf{Z}'\mathsf{X} \\ \mathsf{n}' \end{pmatrix} \\ &= \mathsf{X}'\mathsf{X} - \mathsf{X}'\mathsf{Z}(\mathsf{Z}'\mathsf{Z})^{-1}\mathsf{Z}'\mathsf{X} + \left(\varphi^2 - \frac{1}{n}\right)\mathsf{n}\mathsf{n}' \\ &= \mathsf{C}^{-1} + \left(\varphi^2 - \frac{1}{n}\right)\mathsf{n}\mathsf{n}', \end{aligned}$$

denoting

$$\mathsf{C} = (\mathsf{X}'\mathsf{X} - \mathsf{X}'\mathsf{Z}(\mathsf{Z}'\mathsf{Z})^{-1}\mathsf{Z}'\mathsf{X})^{-1}.$$

Let us notice, that (using (5), (4) and (8) subsequently)

$$\begin{aligned} \mathsf{C}^{-1}\mathsf{1} &= \mathsf{X}'\mathsf{X}\mathsf{1} - \mathsf{X}'\mathsf{Z}(\mathsf{Z}'\mathsf{Z})^{-1}\mathsf{Z}'\mathsf{X}\mathsf{1} \\ &= \mathsf{X}'\mathsf{1} - \mathsf{X}'\mathsf{Z}(\mathsf{Z}'\mathsf{Z})^{-1}\mathsf{Z}'\mathsf{1} \\ &= \mathsf{n}, \end{aligned}$$

therefore

$$\mathsf{C}\mathsf{n} = \mathsf{1}. \tag{11}$$

Now we use another known formula (e.g. Rao (1973), p. 33)

$$(\mathsf{A} + \mathsf{u}\mathsf{v}')^{-1} = \mathsf{A}^{-1} - \frac{\mathsf{A}^{-1}\mathsf{u}\mathsf{v}'\mathsf{A}^{-1}}{1 + \mathsf{v}'\mathsf{A}^{-1}\mathsf{u}},$$

that is valid for a regular matrix A and vectors u, v satisfying $\mathsf{v}'\mathsf{A}^{-1}\mathsf{u} \neq -1$. Then using (11) we will get

$$\begin{aligned} \mathsf{G} &= \mathsf{C} - \left(1 + \left(\varphi^2 - \frac{1}{n}\right)\mathsf{n}'\mathsf{C}\mathsf{n}\right)^{-1}\left(\varphi^2 - \frac{1}{n}\right)\mathsf{C}\mathsf{n}\mathsf{n}'\mathsf{C} \\ &= \mathsf{C} - (1 + \varphi^2 n - 1)^{-1}\left(\varphi^2 - \frac{1}{n}\right)\mathsf{1}\mathsf{1}' \\ &= \mathsf{C} - \frac{1}{n}\mathsf{1}\mathsf{1}' + \frac{1}{n^2\varphi^2}\mathsf{1}\mathsf{1}'. \end{aligned} \tag{12}$$

Besides of the estimate a_r of the effect α_r a standard regression program computes standard error of this estimate

$$s\sqrt{c_{rr} - \frac{1}{n} + \frac{1}{n^2\varphi^2}} \tag{13}$$

and t-statistics

$$t_r = \frac{a_r}{s\sqrt{c_{rr} - \frac{1}{n} + \frac{1}{n^2\varphi^2}}} \tag{14}$$

where the symbol s^2 means the residual variance.

3. Significance of α_r

Considering (5) the linear model (3) we can express using parameters

$$\boldsymbol{\delta} = \boldsymbol{\alpha} + \mu \mathbf{1}$$

and $\boldsymbol{\beta}$ in the following form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\delta} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e}.$$

Under the assumption (6) we have the full rank model, therefore both vectors of parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ are estimable. For the estimate $\mathbf{d}$ of the parameter $\boldsymbol{\delta}$ it holds (e.g. considering (9))

$$\begin{aligned}\operatorname{var} \mathbf{d} &= \sigma^2(\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} \\ &= \sigma^2 \mathbf{C}.\end{aligned}$$

The requirement $\alpha_r = 0$ for a given r ($r = 1, \dots, k$) is not a testable linear hypothesis in the model (3). However, the situation is different for the linear hypothesis

$$H : \delta_r - \frac{1}{n}\sum_{i=1}^{k} n_i \delta_i = 0, \tag{15}$$

that can be expressed after a small adjustment also as

$$H : \alpha_r - \frac{1}{n}\sum_{i=1}^{k} n_i \alpha_i = 0.$$

With the condition (7) it si the same as the hypothesis

$$\alpha_r = 0.$$

The linear hypothesis (15) we can write as

$$\mathbf{h}'\boldsymbol{\delta} = 0, \tag{16}$$

where

$$\begin{aligned}\mathbf{h} &= \left(\mathbf{I} - \frac{1}{n}\mathbf{n}\mathbf{1}'\right)\mathbf{j}_r, \\ \mathbf{j}_r &= (0, \dots, 1, \dots, 0)'.\end{aligned}$$

We can compute (using (11))

$$
\begin{aligned}
\operatorname{var} \mathsf{h}'\mathsf{d} &= \sigma^2 \mathsf{h}'\mathsf{C}\mathsf{h} \\
&= \sigma^2 \mathsf{j}_r' \Big(\mathsf{I} - \frac{1}{n}\mathsf{1}\mathsf{n}'\Big)\mathsf{C}\Big(\mathsf{I} - \frac{1}{n}\mathsf{n}\mathsf{1}'\Big)\mathsf{j}_r \\
&= \sigma^2 \mathsf{j}_r' \Big(\mathsf{C} - \frac{1}{n}\mathsf{1}\mathsf{n}'\mathsf{C} - \frac{1}{n}\mathsf{C}\mathsf{n}\mathsf{1}' + \frac{1}{n_2}\mathsf{1}\mathsf{n}'\mathsf{C}\mathsf{n}\mathsf{1}'\Big)\mathsf{j}_r \\
&= \sigma^2 \mathsf{j}_r' \Big(\mathsf{C} - \frac{1}{n}\mathsf{1}\mathsf{1}'\Big)\mathsf{j}_r \\
&= \sigma^2 \Big(c_{rr} - \frac{1}{n}\Big).
\end{aligned}
$$

Therefore the test statistic for testing hypothesis (15) equals the ratio

$$
t_r = \frac{a_r}{s\sqrt{c_{rr} - \frac{1}{n}}}. \tag{17}
$$

We can see that for a sufficiently large value of the expression $n^2\varphi^2$, the statistics (14) a (17) in fact coincide. The hypothesis (15), that the r-th class in one-way analysis of variance with covariances "does not differ from the mean," we can test by a program for the linear regression using the statistics (14).

Let us go back to the Wishart's example. We ask whether growth rates in the pen III are significantly different from average growth rate. For $\varphi = 0.01$, $\varphi = 1$ and $\varphi = 100$ we get according to (14) $t_3 = -0.26$, $t_3 = -2.39$ and $t_3 = -2.40$, respectively. The last value can be considered as exact, therefore the growth rates in pen III can be considered as statistically significantly different from average growth rate, even if the separate influence of the pen factor on the growth rate on the same level is not significant $\big(F_{4,19} = 2.35 < F_{4,19}(0.95) = 2.90\big)$.

References

[1] *Gentleman, W. M.*: Basic procedures for large, sparse or weighted linear least squares problems, Applied Statistics *23* (1974), 448-454.

[2] *Rao, C. R.*: Linear statistical inference and its applications, Wiley, New York, 1973.

[3] *Scheffé, H.*: The analysis of variance, Wiley, New York, 1959.

[4] *Zvára, K.*: Regression analysis, Academia, Prague, 1989. (In Czech.)

Author's address: Dr. *Karel Zvára*, CSc., Matematicko-fyzikální fakulta UK, Sokolovská 83, 186 00 Praha 8.