Listy Biometryczne - Biometrical Letters
Vol. 38(2001), No. 1, 33-40

Weighted linear joint regression analysis

João Tiago Mexia[1], Dulce Gamito Pereira[2] and José Baeta[3]

[1]Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2825 Monte da Caparica, Portugal

[2]Departamento de Matemática, Universidade de Évora Colégio Luis António Verney, Rua Romão Ramalho 59, 7000 Évora, Portugal

[3]Estação Agronómica Nacional, 2780 Oeiras, Portugal

SUMMARY

The introduction of weights for the yields of the different cultivars in the different blocks enables us, while carrying out joint regression analysis (JRA), to consider the agricultural significance of the conditions pertaining to the blocks in a network. Moreover it is now possible to use incomplete blocks, thus lighting the local designs and enabling a better coverage of the region under study. Namely, this happens when α designs are used. To carry out the regressions adjustment for weighted linear JRA an algorithm is derived from the sum of sums of weighted residues. Only linear regressions are considered since the use of higher degree regressions does not significantly improve the adjustment.

KEY WORDS: joint regression analysis, incomplete blocks, randomized blocks, environmental indexes.

1. Introduction

Joint regression analysis (JRA) has been widely used in the joint analysis of design networks for cultivars comparison. In the classical version of the technique the designs in the network are randomized blocks. The J cultivars to be compared are present in every of the n blocks in the network. Following Gusmão (1985) the average yield of each block is used as a measure of it's productivity in the year under consideration. This productivity in the block is the environmental index for that year. The name of the technique was derived from the fact that it uses simultaneous linear regressions of the cultivars yields on the environmental indexes. This technique is straightforward, but has several limitations:

- All blocks are given the same relevance. Despite the regions covered by a design network having to be sufficiently homogeneous, see Gusmão et al. (1989), some designs and the corresponding blocks may be located in more representative agricultural situations than others.
- The technique can only be applied in the complete case in which all cultivars are present in all the blocks.
- The average of the yields in a block is taken as the true value of the corresponding environmental index, which is only approximately correct.

To overcome these limitations we introduce weights p_{ij}, $i = 1, ..., n$; $j = 1, ..., J$, for the pairs (block; cultivar). These weights will be null for absent cultivars and may take in consideration the relevance of the agricultural conditions. It may be convenient to give value one to the largest weights in order to standardize their values. Moreover, we will use least squares to estimate the environmental indexes and the regression coefficients. The technique could be extended to use higher order polynomial regressions, but (see Mexia et al., 1997; 1999) in many situations linear regressions give quite similar results.

As a parting remark we point out that JRA can now be applied when α designs are used, as is so often the case nowadays.

2. Adjustment

With $x = (x_1, ..., x_n)$ the vector of environment indexes, $\alpha = (\alpha_1, ..., \alpha_J)$ and $\beta = (\beta_1, ..., \beta_J)$ the vectors of regression coefficients, using least squares leads to minimization of

$$S(x, \alpha, \beta) = \sum_{j=1}^{J} \sum_{i=1}^{n} p_{ij}(Y_{ij} - \alpha_j - \beta_j x_i)^2,$$

where Y_{ij} is the yield in the i-th block of the j-th cultivar if it is present (if the cultivar is absent, $p_{ij} = 0$ and the value Y_{ij} is irrelevant).

A convenient algorithm to carry out this minimization is the ZIG-ZAG algorithm thus called since it alternatively minimizes $S(x, \alpha, \beta)$ with respect to the regressions coefficients and to the environmental indexes.

To apply this algorithm we start by adequately choosing an initial vector $x_0 = (x_{01}, ..., x_{0n})$. In the complete case we can use, as in classical JRA, the vector of block average yields. If, as in α designs, the blocks are grouped in super-blocks that contain every cultivar, for each block we can take the average yield of its super-block. In other

less structured cases we can try to use for each block its average yield. Once x_0^n is chosen, we minimize with respect to α and β. Since

$$S(x_0, \alpha, \beta) = \sum_{j=1}^{J} \left[\sum_{i=1}^{n} p_{ij}(Y_{ij} - \alpha_j - \beta_j x_{0i})^2 \right],$$

we can minimize separately the

$$\sum_{i=1}^{n} p_{ij}(Y_{ij} - \alpha_j - \beta_j x_{0i})^2, j = 1, ..., J,$$

thus obtaining the well known estimates

$$\begin{cases} \widetilde{\beta}_{1j} = \dfrac{\sum_{i=1}^{n} p_{ij} \sum_{i=1}^{n} p_{ij} x_{0i} Y_{ij} - \sum_{i=1}^{n} p_{ij} x_{0i} \sum_{i=1}^{n} p_{ij} Y_{ij}}{\sum_{i=1}^{n} p_{ij} \sum_{i=1}^{n} p_{ij} x_{0i}^2 - (\sum_{i=1}^{n} p_{ij} x_{0i})^2}, j = 1..., J \\ \widetilde{\alpha}_{1j} = \dfrac{\sum_{i=1}^{n} p_{ij} Y_{ij}}{\sum_{i=1}^{n} p_{ij}} - \widetilde{\beta}_{1j} \dfrac{\sum_{i=1}^{n} p_{ij} x_{0i}}{\sum_{i=1}^{n} p_{ij}}, j = 1, ..., J. \end{cases}$$

It would simplify these expressions if we could assume that $\sum_{i=1}^{n} p_{ij} = 1$, $j = 1, ..., J$, but this is not always possible since, for instance, different cultivars may be present at different locations. In the complete case or when blocks are grouped in complete super-blocks, one cultivar appearing once in each super-block, we can make this assumption.

Once $\widetilde{\alpha}_1 = (\widetilde{\alpha}_{11}, ..., \widetilde{\alpha}_{1J})$ and $\widetilde{\beta}_1 = (\widetilde{\beta}_{11}, ..., \widetilde{\beta}_{1J})$ are obtained, we minimize with respect to x. Since $S(x, \alpha_1, \beta_1) = \sum_{i=1}^{n} h_i(x_i \mid \alpha_1, \beta_1)$, with

$$h_i(x_i \mid \alpha_1, \beta_1) = \sum_{j=1}^{J} p_{ij}(Y_{ij} - \widetilde{\alpha}_{1j} - \widetilde{\beta}_{1j} x_i)^2, i = 1, ..., n,$$

we have only to minimize each of the $h_i(x \mid \widetilde{\alpha}_1, \widetilde{\beta}_1)$ with respect to x. It is straightforward to show that the minimums are

$$x_{1i} = \frac{\sum_{j=1}^{J} p_{ij} \widetilde{\beta}_{1j} Y_{ij} - \sum_{j=1}^{J} p_{ij} \widetilde{\alpha}_{1j} \widetilde{\beta}_{1j}}{\sum_{j=1}^{J} p_{ij} \widetilde{\beta}_{1j}^2}, i = 1, ..., n.$$

It is also straightforward to show that, with $x_1 = (x_{11}, ..., x_{1n})$,

$$\widetilde{S}_1 = S(x_1, \widetilde{\alpha}_1 \widetilde{\beta}_1) =$$

$$= \sum_{j=1}^{J} \left[\sum_{i=1}^{n} p_{ij} \left(Y_{ij} - \frac{\sum_{i=1}^{n} p_{ij} Y_{ij}}{\sum_{i=1}^{n} p_{ij}} \right)^2 - \widetilde{\beta}_{1j} \sum_{i=1}^{n} p_{ij} \left(Y_{ij} - \frac{\sum_{i=1}^{n} p_{ij} Y_{ij}}{\sum_{i=1}^{n} p_{ij}} \right) \left(x_i - \frac{\sum_{i=1}^{n} p_{ij} x_i}{\sum_{i=1}^{n} p_{ij}} \right) \right].$$

As it will be shown in the Appendix, we may rescale the estimated environmental indexes in order to maintain, from iteration to iteration, the minimum and maximum estimated environmental indexes. In this way the range of variation of the indexes is stabilized throughout the application of the algorithm. To do this rescaling we have only to replace the x_{1i} by the

$$\widetilde{x}_{1i} = m_0 + \frac{M_0 - m_0}{M_1 - m_1}(x_{1i} - m_1),$$

where the m_0 and m_1 (M_0 and M_1) are the minimums (maximums) of the $\{\widetilde{x}_{01}, ..., \widetilde{x}_{0n}\}$ and the $\{\widetilde{x}_{11}, ..., \widetilde{x}_{1n}\}$, respectively.

We can now initiate a second iteration using $\widetilde{x}_1 = (\widetilde{x}_{11}, ..., \widetilde{x}_{1n})$ as a provisional vector of estimated environmental indexes. As before we start by minimizing for α and β before minimizing for x. Then we obtain the sum $\widetilde{S}_2$ of sums of squares for residues and rescale the vector of estimated environmental indexes.

This procedure must be repeated until the difference $\widetilde{S}_{m+1} - \widetilde{S}_m$ between successive sums of squares of residuals falls below a limit that is set before the calculations are started.

3. An application

Our first example concerns a network of eleven randomized block designs. These experiments were carried out in 1992 and 1993 by Estação Nacional de Melhoramento de Plantas (the Portuguese plant breeding board) who kindly allowed us to use the yield data. Each design had four blocks and all nine wheat cultivars (Celta, Helvio, TE9006, TE9007, TE9008, TE9110, TE9115, TE9204, Trovador) were present in each block. The yields in the different designs were, according to agricultural relevance, given the weights: 0.5; 0.6; 0.7; 0.8; 0.9; 1; 0.9; 0.8; 0.7; 0.7; 0.6; 0.5.

Applying the ZIG-ZAG algorithm we obtained the adjusted regression and determination coefficients shown in Table 1.

The adjusted regressions are jointly presented in Fig. 1. The minimum and maximum environmental indexes were 2212.59 and 8838.49.

Table 1. Regression and determination coefficients (complete case)

Cultivar	α	β	R^2
Celta	-341.03	1.21	0.90
Helvio	21.74	1.03	0.92
TE9006	-546.89	1.13	0.87
TE9007	-386.72	1.09	0.92
TE9008	460.71	0.93	0.89
TE9110	-252.41	0.92	0.80
TE9115	1375.00	0.52	0.50
TE9204	-19.35	1.09	0.90
Trovador	-311.04	1.07	0.89

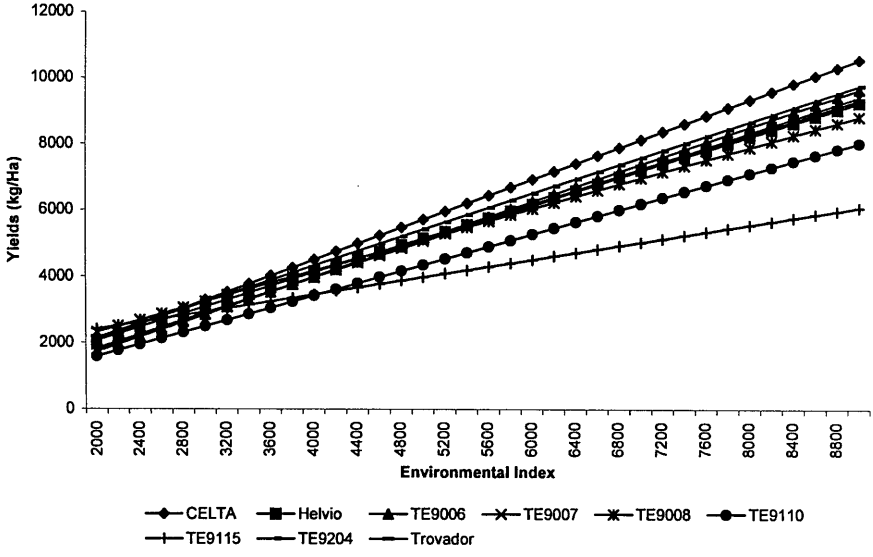

Figure 1. Joint linear regressions, the complete case

Next we consider a network of α designs where data was kindly made available by the Research Center for Cultivar Testing at Słupia Wielka. There were six field experiments, each with four super-blocks of five blocks. In each block four cultivars of winter rye were present.

As stated above we used as initial environmental indexes the super-block averages in order to apply the ZIG-ZAG algorithm. The results of the adjustments are given in Table 2. The adjusted regressions are presented in Fig. 2. The minimum and maximum environmental indexes were 3613.33 and 7273.33.

Table 2. Regression and determination coefficients (incomplete case)

Cultivar	α	β	R^2
URSUS	-2360.47	1.52	0.97
RAH 797	-2178.01	1.43	0.98
ESPRIT	-1397.61	1.32	0.94
RAH 897	-1496.16	1.30	0.97
RAPID	-1413.96	1.29	0.97
MARDER	-1344.10	1.29	0.93
RAH 496	-1232.25	1.28	0.95
RAH 596	-966.31	1.19	0.95
WID 196	-611.02	1.15	0.92
WARKO	-982.47	1.13	0.95
AMILO	-954.34	1.11	0.93
ADAR	-567.53	1.04	0.97
CHD 296	-284.94	1.00	0.88
SMH 1195	-458.55	1.00	0.90
RAH 697	390.54	1.00	0.87
ZDUNO	-245.97	0.99	0.95
CHD 396	-292.51	0.98	0.91
SMH 1295	-175.81	0.96	0.94
NAD 195	382.31	0.90	0.91
SMH 1094	660.28	0.79	0.88

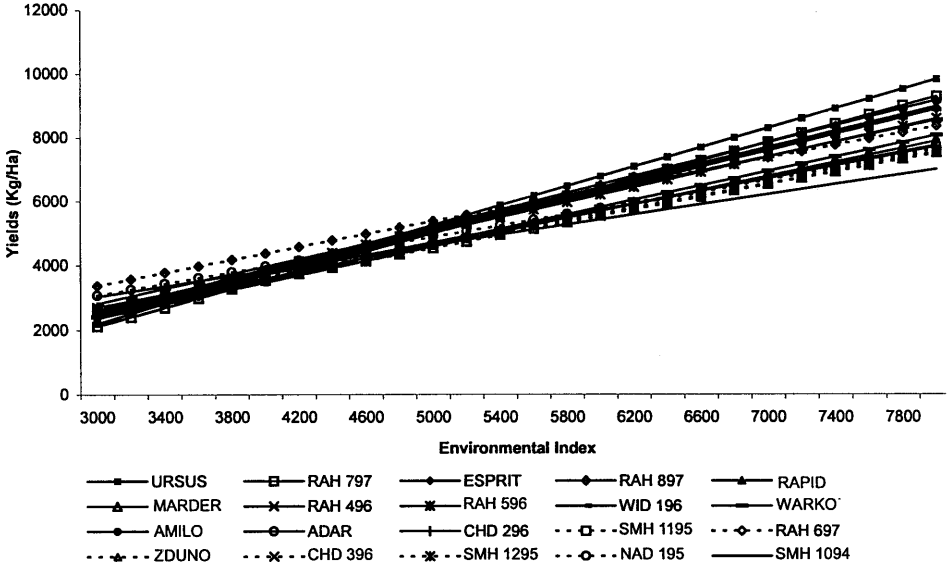

Figure 2. Joint linear regressions, the incomplete case

Thus the ZIG-ZAG algorithm is equally easy to apply in both the complete and incomplete case with α designs and the final results are of the same type. This is interesting since the use of α designs is now widespread.

4. Appendix

We now consider briefly certain results useful to establish the existence of minimums and the possibility of rescaling. We start with

PROPOSITION 1. *Writing* $(x_1, \alpha_1, \beta_1)\tau(x_2, \alpha_2, \beta_2)$ *when, with* $c \neq 0$, $x_2 = cx_1$, $\alpha_2 = \alpha_1$ *and* $\beta_2 = c^{-1}\beta_1$, *we establish an equivalence relation in* $R^n \times R^J \times R^J$. *If* $(x_1, \alpha_1, \beta_1)\tau(x_2, \alpha_2, \beta_2)$, *then* $S(x_2, \alpha_2, \beta_2) = S(x_1, \alpha_1, \beta_1)$. *The goal function* S *takes all its values for* $x \neq 0_n$, *where* 0_n *is the vector of* n *zeros, on* $\mathbf{V} = \{(x, \alpha, \beta); \| x \| = 1\}$ *and has a set of absolute minimums saturated for* τ.

Proof. It is straightforward to show that τ is an equivalence relation. Besides this, if $(x_1, \alpha_1, \beta_1)\tau(x_2, \alpha_2, \beta_2)$, $Y_{ij} - \alpha_{2j} - \beta_{2j}x_{2i} = Y_{ij} - \alpha_{1j} - \beta_{1j}x_{1i}$, $i = 1, ..., n$, $j = 1, ..., J$, and so $S(x_2, \alpha_2, \beta_2) = S(x_1, \alpha_1, \beta_1)$. Moreover, taking $c = \| x_1 \|^{-1}$, whenever $x_1 \neq 0_n$, we get $(x_2, \alpha_2, \beta_2) \in \mathbf{V}$ such that $(x_1, \alpha_1, \beta_1)\tau(x_2, \alpha_2, \beta_2)$ and so, according to the second part of the thesis, all the values of $S(x, \alpha, \beta)$, for $x \neq 0_n$, are taken $\mathbf{V}$. If $\| x \| = 1$ we have $S(x, \alpha, \beta) \geq S(x, \widetilde{\alpha}(x), \widetilde{\beta}(x))$.

Now, according to Weierstrass theorem, $\overline{S}(x) = S(x, \widetilde{\alpha}(x), \widetilde{\beta}(x))$ has at least an absolute minimum in $\oplus_1 = \{x; \| x \| = 1\}$. These absolute minimums of $\overline{S}(x)$ are the absolute minimums of $S(x, \alpha, \beta)$, for $x \neq 0_n$. Since $S(0_n, \alpha, \beta) = S(x, \alpha, 0) \geq S(x, \widetilde{\alpha}(x), \widetilde{\beta}(x))$, the absolute minimums of $S(x, \alpha, \beta)$ will have $x \neq 0_n$.

Lastly, since $S(x, \alpha, \beta)$ takes the same value for all points in a τ equivalence class, the set of its absolute minimums is saturated for τ. □

When we rescale the vector of estimated environmental indexes at the end of each iteration we are applying this proposition. Moreover, it guarantees the existence of least squares estimators.

During the u-th iteration of ZIG-ZAG algorithm we minimize the functions $S_1(\alpha, \beta \mid \widetilde{x}_{u-1}) = S(\widetilde{x}_{u-1}, \alpha, \beta)$ and $S_2(x \mid \widetilde{\alpha}_u, \widetilde{\beta}_u) = S(x, \widetilde{\alpha}_u, \widetilde{\beta}_u)$. It is interesting to observe that

$$\begin{cases} S_1(\alpha, \beta \mid \widetilde{x}_{u-1}) = \sum_{j=1}^{J} \left[\sum_{i=1}^{n} p_{ij}(Y_{ij} - \alpha_j - \beta_j \widetilde{x}_{(u-1)i})^2 \right] \\ S_2(x \mid \widetilde{\alpha}_u, \widetilde{\beta}_u) = \sum_{i=1}^{n} \left[\sum_{j=1}^{J} p_{ij}(Y_{ij} - \widetilde{\alpha}_{uj} - \widetilde{\beta}_{uj} x_i)^2 \right] \end{cases}$$

and that the functions $\sum_{i=1}^{n} p_{ij}(Y_{ij} - \alpha_j - \beta_j \widetilde{x}_{(u-1)i})^2$, $j = 1, ..., J$, and $\sum_{j=1}^{J} p_{ij}(Y_{ij} -$

$\widetilde{\alpha}_{uj} - \widetilde{\beta}_{uj}x_i)^2$, $i = 1, ..., n$, are convex in (α_j, β_j), $j = 1, ..., J$, and in x_i, $i = 1, ..., n$, respectively. Thus $S_1(\alpha, \beta \mid \widetilde{x}_{u-1})$ and $S_2(x \mid \widetilde{\alpha}_u, \widetilde{\beta}_u)$ will be convex in (α, β) and in x, respectively, since when we add convex functions of distinct variables we obtain convex functions. Then, see Bazaraa et al. (1992, p. 113), $S_1(\alpha, \beta \mid \widetilde{x}_{u-1})$ and $S_2(x \mid \widetilde{\alpha}_u, \widetilde{\beta}_u)$ will have unique minimums.

References

Bazaraa M.S., Sherali H.D. and Shetty C.M. (1992). *Nonlinear Programming. Theory and Algorithms.* 2nd ed., John Wiley & Sons, New York.

Gusmão L. (1985). An adequate design for regression analysis of yield trials. *Theor. Appl. Genet.* **72**, 98-104.

Gusmão L., Mexia J.T. and Gomes M.L. (1989). Mapping of equipotential zones for cultivar yield pattern evolution. *Plant Breeding* **103**, 293-298.

Mexia J.T., Amaro A.P., Gusmão L. and Baeta J. (1997). Upper contour of a joint regression analysis. *Journal of Genetics and Breeding* **51**, 253-255.

Mexia J.T., Pereira D. and Baeta J. (1999). L_2 environmental indexes. *Biometrical Letters* **36**, 137-143.

Received 12 November 2000; revised 20 August 2001

Łączna analiza ważonej regresji liniowej

Streszczenie

Wprowadzenie wag dla plonów różnych odmian w różnych blokach pozwala na uwzględnienie w analizie regresji agrotechnicznej istotności różnych bloków w serii doświadczeń. Pozwala też na wykorzystanie bloków niekompletnych (np. układów typu α), co wpływa na zmniejszenie doświadczeń i lepszą ich reprezentatywność. Dla przeprowadzenia ważonej regresji liniowej opracowano algorytm oparty na sumach ważonych reszt. Rozważa się tylko regresję liniową, gdyż regresja wyższego rzędu nie wprowadza istotnej poprawy dopasowania.

Key words: łączna analiza regresji, bloki niekompletne, bloki losowane, indeksy środowiskowe.