Unpublished
143

ID 95211

MC
Please send me a copy of this.

~~Copy~~
~~KBL Jafri~~
~~Jerre Dubin~~
~~Mike Morris~~
~~Original to DB.~~

To: **D. Byerlee**
Director
Economics Program
CIMMYT

Copy: Mike Morris
Economics Program
CIMMYT

From: C. González
Consultant

Date: 12 March 1992

Nals'

Re: Revised version of report on "Methods of estimating Rate of Yield Gains for Plant Breeding Programs"

Attached is the new version of the report. I have added a new section FIXED EFFECTS, MIXED AND RANDOM MODELS and made a few changes on the previous version.

The changes are the following:

1. On page 5 "..., but this is not always so since the three values are mutually inconsistent." has been substituted with "..., but they cannot be regarded as a coherent summary since the three values are mutually inconsistent." The reason for this change is that although the values are inconsistent, they are still valid estimates (but not very useful).

2. On page 9 line 7 of the old version "precise" has been substituted with "accurate" (page 8, last line of the new one). Although both words are synonymous in English, in statistics they have very different meanings. The correct one in this case is "accurate".

3. On page 11, lines 8 and 9 of the old version ".. the first year" has been substituted with "... year 88" and "... the first variety" with "variety A" (page 10, lines 22 & 23 of the new one). The reason is that the coefficients estimate the differences from the levels of the factors for which no dummy variables were included and these are not necessarily the first ones.

4. On pages 15 and 16 of the old version, "South Region of Parana State" has been substituted with "region South of Parana State" which is the correct one.

5. On page 20 of the old version "Of course, ... can be used." has been substituted with what appears between lines 4 and 18, page 23 on the new version. I describe there the three stages procedure that should be followed in case data from individual trials be available.

6. Finally, I made drastic changes in the section CONCLUSIONS, most of them related to the exponential model. The method of estimating **g** is correct, especially if a weighted regression (by number of trials) is performed. What is not correct, from my point of view, is to adopt the exponential model without further consideration. I say that the exponential model is less realistic than the linear one, because the linear assumes that the absolute rate of gain is constant, whereas the exponential one assumes that the relative rate of gain is constant. This last assumption might be valid at the initial steps of a breeding program starting at very low yields, but not when high yield potentials has been attained. Perhaps for a very large number of years, the more realistic model would be the logistic.

   To make my point clear:

   $y_{k+1} = \exp[a + g(k+1)] = e^{a} \cdot e^{gk} \cdot e^{g}$, and

   $y_{k} = \exp[a + gk] = e^{a} \cdot e^{gk}$.

   Thus, the relative rate of gain is

   $y_{k+1}/y_{k} = e^{a} \cdot e^{gk} \cdot e^{g} \;/\; e^{a} \cdot e^{gk} = e^{g}$.

   That means that from one vintage to the following the yields increase exponentially, due to the release of new varieties. This is the assumption underlying the exponential model which for me is not very realistic.

MAR2012, 92

Methods of estimating rate of yield gains for plant breeding programs[1]

C.A. González [2]

Economics Miscellaneous Paper No. 92-1

Note:

[1] Paper prepared as part of a consultancy with CIMMYT's Economics Program - December 1991

[2] Present address: Apdo. Postal 105-383, 11581 México D.F., México

Methods of Estimating Rate of Yield Gains for Plant Breeding Programs

C.A. González

Abstract

In order to estimate rate of return to research investment in breeding, it is necessary to estimate the rate of gain in yield due to the introduction of new varieties. Traditionally such yield gains has been assessed by direct comparison of yields, restricted to differences with a standard or control. Problems of inefficiency arise when varieties x environments or vintages x environments tables are incomplete. By using information provided by both direct and indirect comparisons, least squares methods produce consistent and more efficient estimates. The paper compares both methods, and illustrates the use of different computational procedures to obtain the least squares estimates. It is found that approaches employed by statisticians and economists provide identical estimates of yield gains.

1. INTRODUCTION

This paper is concerned with the use of series of variety trials to estimate the gain in yield from the introduction of new varieties. Analysis would be straightforward if all varieties, or vintages (sets of varieties released the same year) were in every year but this is rarely the case. There is a great deal of imbalance in the data since new varieties are continually being developed and added to the system, while others leave because they merit no further interest.

The method traditionally used consists in assessing the yield of each variety (vintage)[3] independently of other varieties (vintages) by direct comparison with standard controls. This method can give rise to inconsistencies. Even when it does not, it produces inefficient estimates (with high standard errors) as a result of not using all the available information.

Better methods, involving least squares adjustment of variety means, have been available for many years, but have been adopted very slowly because of lack of adequate computing facilities. Also many researchers using these various methods do not properly understand the differences between them. At the same time, there is much confusion between method of estimation and method of calculation.

[3] Vintage is used here to refer to the set of varieties released on a given year.

The present paper is intended to demonstrate the advantages of the least squares methods, and to clarify the confusion between estimation and calculation methods. For illustration purposes we will be using varieties x years tables, and then showing how the procedures can be easily extended to vintages x years tables.

2. SIMPLE AVERAGES, AND DIRECT AND INDIRECT COMPARISONS

As a trivial example, suppose that, in order to assess the yield gain due to the introduction of new spring wheat varieties, we have data from three trials conducted in different years: in 1988 with standard variety A and two new varieties B and C, in 1989 with A and B only, and in 1990 with A and C only (Table 1).

Table 1. Mean yield (t/ha) of three varieties tested in different years

		Variety		
Year	A	B	C	Average
1988	3.0	3.4	3.6	3.33
1989	3.8	4.8	-	4.30
1990	4.0	-	4.6	4.30
Average	3.60	4.10	4.10	3.98

Simple averaging gives means of 3.6, 4.1, 4.1 for A, B, C, suggesting a marked superiority of B and C over A, and no difference between B and C. However, in the single year when both were tested C was superior to B.

An alternative procedure, widely used for variety trials, and known as the method of direct comparisons, is to estimate the difference of two varieties only from trials in which both were included. This leads to the following estimates:

B is better than A by 0.7 {= [(3.4-3.0)+(4.8-3.8)]/2}
C is better than A by 0.6 {= 4.6-4.0}
C is better than B by 0.2 {= 3.6-3.4}

Under certain assumptions, these are legitimate estimates of variety yield differences, but they cannot be regarded as a coherent summary since the three values are mutually inconsistent. Consistency could be secured by restricting the direct comparisons to differences between a new variety and the control. This means that the comparison between two new varieties must be made through the controls. In the example here, B could be regarded as better than C by 0.1.

Summing up, this method consists in calculating, for each year or environment, the yield differences between each new variety and the control, and then calculating the mean yield differences over years, as shown on Table 2.

This method of estimation by direct comparisons between new varieties and a control, is the one used by Brennan (1986). Although the procedure produces unbiased estimates, these will often be unnecessarily imprecise. Another serious drawback of this method is that it does not allow for changes in the control, i.e. if we want to measure yield gain over a certain standard variety or vintage to be taken as a baseline, that variety or vintage must have been tested on every year of the period being considered.

Table 2. Yield performances (t/ha) between varieties

Year	B-A	C-A
1988	0.4	0.6
1989	1.0	-
1990	-	0.6
Mean difference	0.7	0.6

If the variance of each entry in Table 1 is σ^2, then the variance of each entry in Table 2 is $2\sigma^2$. The variance of the mean differences (B–A) and (C–A) are both σ^2, and the variance of the mean difference (C–B) is $2\sigma^2$.

3. ADJUSTED MEANS BY LEAST SQUARES

Variety means such as those given in the lower margin of Table 1 provide a sound basis for the comparison of varieties that are tested in all environments but they are not suitable for varieties, such as B, with incomplete results. On any assessment the mean 4.10 t/ha seriously underestimates variety B, because that variety was not tested in one of the years in which varieties gave higher yields. This difficulty can be overcome by adjusting the mean of variety B by adding an estimate of the year effect 1988-90 v. 1990.

Different adjusted means are obtained depending on the choice of estimate of year effect. The least squares adjustment (Patterson, 1978; Finney, 1980), uses the most accurate available estimate. Let y_{ij} be the yield of the i^{th} variety in year or environment j. Then the variety means are estimated by fitting the model

$$Y_{ij} = m + a_i + b_j, \qquad (i=1,2,3;\ j=1,2,3) \qquad (1)$$

where m is a general mean, a_i is the estimated effect of variety i and b_j the estimated effect of year or environment j. A constraint is imposed on the b_j so that $(m+a_i)$ is the mean over years or environments, for any variety that happens to have a complete set of results. Subject to this constraint, variety means are estimated by minimizing the sum of squares of residuals $y_{ij} - Y_{ij}$. In Appendix A we present the calculation, from first principles, of the least squares estimates of the variety means for the data in Table 1.

In Table 3 we present the simple averages and the least squares estimates of mean yield for all three varieties. For variety A, that has been tested in all three years, the least squares estimate of mean yield coincide with the simple average. For varieties B and C, least squares have adjusted yields upwards to compensate for the fact that they have not been tested in one of the high-yielding years.

Table 3. Least squares estimates of variety yields (t/ha)

	Variety			
Year	A	B	C	Average
1988	3.0	3.4	3.6	3.33
1989	3.8	4.8	-	4.30
1990	4.0	-	4.6	4.30
Average	3.60	4.10	4.10	3.98
Least squares estimate	3.60	4.28	4.28	4.05

In Table 4 we show the estimates of the variety mean yield differences and their variances, obtained by the method of restricted direct comparisons and by least squares.

Table 4. Variety yield differences by direct comparison and Least Squares estimates

Variety	Direct comparison		Least Squares	
Difference	Estimate	Variance	Estimate	Variance
B-A	0.7	σ^2	0.68	$0.93\sigma^2$
C-A	0.6	σ^2	0.68	$0.93\sigma^2$
C-B	-0.1	$2\sigma^2$	0.00	$1.33\sigma^2$

Both methods of estimation produce unbiased estimates. However, as can be seen from Table 4, the method of restricted direct comparisons sometimes leads to estimates with much larger variances than the least squares estimates. Among all linear unbiased estimates, the least squares estimates have minimum variance, i.e. they are the most accurate.

4. CALCULATION OF ADJUSTED MEAN YIELDS BY LEAST SQUARES

The classical approach to the analysis of experimental data is the analysis of variance, which is based on the pattern of calculations used before the advent of computers. The underlying concept is that the units are classified according to several factors, such as varieties and years in our model (1)

$$y_{ij} = m + a_i + b_j + \epsilon_{ij}.$$

The analysis of variance provides information about the sets of parameters (a_i), (b_j), and an estimate of σ^2, the variance of the ϵ_{ij}.

However, there are two computational problems in the factor-based approach. The first arises because of the lack of orthogonality between varieties and years, as a consequence of the unbalanced nature of the varieties x years tables. The second problem arises from the fact that the sets of parameters corresponding to the levels of each factor are not uniquely defined if there are two or more factors. Consider again our model above. It is not possible from any set of observations on combinations of varieties and years to deduce absolute values for either set of parameters without imposing an arbitrary restriction on the other set of parameters. This may be managed, for instance, by imposing one of the following arbitrary constraints: $b_i=0$ or $\Sigma b_i=0$.

The alternative to the analysis of variance approach is to view all models as essentially multiple linear regression models. This approach immediately overcomes the problem of the lack of orthogonality as a consequence of the unbalanced nature of the varieties x years tables. The constraint problem must be addressed directly by the elimination of one parameter from each set. Essentially if a_1 is omitted or set equal to zero then the remaining parameters in the set represent deviations from a_1.

The analysis of unbalanced data structures can be managed by a few general purpose computer packages, the most powerful being REML (Robinson, 1987), followed by Genstat and SAS. In the absence of any of these packages any unbalanced variety x environment or vintage x environment structure can be analysed by a multiple regression program, as will be illustrated in the following section. The multiple regression approach was the one followed by Godden (1988) and Byerlee (1990).

5. ANALYSIS, BY MULTIPLE REGRESSION, OF THE UNBALANCED VARIETIES X ENVIRONMENTS DATA IN TABLE 1

The data in Table 1 can be analysed by multiple regression, by defining a dummy regression variable for each year except year 88 and for each variety except variety A. The regression coefficients then estimate the difference of each year from year 88, and of each variety from variety A. Thus, instead of fitting model (1) we will be fitting the following multiple regression model

$$Y = \text{constant} + \alpha_2 z_2 + \alpha_3 z_3 + \beta_2 w_2 + \beta_3 w_3, \qquad (2)$$

where z_2 is 1 for variety B and 0 otherwise, z_3 is 1 for variety C and 0 otherwise, w_2 is 1 for year 89 and 0 otherwise, and w_3 is 1 for year 90 and 0 otherwise. Thus α_2 estimates the difference of variety B from variety A, and α_3 the difference of variety C from variety A. Similarly β_2 estimates the difference of year 89 from year 88, and β_3 the difference of year 90 from year 88.

There are some packages, like MINITAB (1989), that, given the levels of a factor, with a simple command, generates one dummy variable for each level of the factor.

The estimates of the regression coefficients obtained by fitting model (2) to data in Table 1, using MINITAB are as follows (See Appendix B):

	Estimate	Stderror	t-ratio
Constant	2.8800	0.1697	16.97
α_2	0.6800	0.2117	3.21
α_3	0.6800	0.2117	3.21
β_2	1.0800	0.2117	5.10
β_3	1.0800	0.2117	5.10

and the estimate of σ^2 is 0.04800 with 2 degrees of freedom.

We have then that B-A = α_2 = 0.6800 and its variance is $(0.2117)^2$, and similarly for C-A = α_3. Omitting α_1 and β_1 from the model is equivalent to setting $\alpha_1 = \beta_1 = 0$. Now, from the estimates of the regression coefficients produced by the multiple regression program, the mean yield for variety i is calculated as:

$$\begin{aligned} &\text{constant} + \alpha_i + [\text{mean of the } \beta_j] = \\ &2.8800 + \alpha_i + (0 + 1.08 + 1.08)/3 = \\ &3.6000 + \alpha_i, \end{aligned}$$

thus giving

Variety	Mean effect
A	3.600
B	4.280
C	4.280

From the estimates of the variety mean effects it is possible to calculate the estimates of the differences between all possible pairs. However, the regression program provides only the standard error of the differences between the omitted variety and the rest. Thus, if standard errors of other differences were needed, it would be necessary to re-run the program omitting the dummy variable corresponding to one of the varieties involved in the differences whose standard errors were sought.

Clearly the use of multiple linear regression for analysis of varietal yield data with many varieties and years is tedious in terms of calculation of both adjusted mean yields, and especially for calculating standard errors of the differences. For this reason REML (Restricted Maximum Likelihood) program is preferred. REML uses an identical estimation procedure (i.e. least squares) but computes mean yields and standard errors within the program.

6. ANALYSIS, USING REML, OF THE UNBALANCED VARIETIES X ENVIRONMENTS DATA IN TABLE 1

REML, which is a program to estimate components of variance and is primarily intended for the analysis of unbalanced data, may be used with advantage for the efficient estimation of fixed effects in varieties x environments or vintages x environments incomplete tables. To obtain the fitted parameter values, REML internally uses the same multiple regression procedure described before. When there are redundant levels of fixed terms, REML set the value of the first level to zero, and the regression parameters for the remaining levels are in fact the differences from the first level.

The main advantage of using REML is that one does not need to bother about generating dummy variables, specifying which ones must be included in the multiple regression, and calculating the adjusted means and standard errors from the estimated effects. The user can specify the amount of output required, which may include variety and environment effects, with their standard errors, to adjusted variety mean effects and standard errors of differences between all pairs.

The mean effects of varieties, standard errors of differences between pairs, and estimate of σ^2 produced by REML for the data in Table 1 are as follows:

Mean effects of varieties

A	B	C
3.600	4.280	4.280

Standard errors of differences between pairs

	A	B
B	0.2117	
C	0.2117	0.2530

Estimate of σ^2 = 0.04800 with 2 degrees of freedom.

Appendix B includes the relevant parts of the output from the MINITAB and REML programs, for analysing the data in Table 1.

7. FIXED EFFECTS, MIXED AND RANDOM MODELS

When running REML to produce the adjusted means of varieties as shown in the previous section, the variety and year effects were specified as fixed. The calculation of the adjusted means by the multiple regression approach or using REML with both effects fixed, is what Patterson (1982) and Patterson et al. (1989) call the FITCON procedure.

We will now use the example in Table 1 to compare four sets of variety means: unadjusted means, means given by the standard FITCON procedure for two-way tables, and two sets of REML means. In FITCON the effects of both factors, varieties and years, are regarded as fixed. REML1 specifies varieties as fixed and years as random, while REML2 treats both varieties and years as random.

Variety means

Variety	unadjusted	FITCON	REML1	REML2
A	3.60	3.60	4.28	3.64
B	4.10	4.28	4.27	4.24
C	4.10	4.28	4.27	4.24
min. pairwise s.e.	0.33*	0.21	0.21	0.19
max. pairwise s.e.	0.45*	0.25	0.25	0.22

* Calculated approx. using REML1 components of variance

Year means

	1988	1989	1990
FITCON	3.33	4.41	4.41
REML1	3.36	4.39	4.39
REML2	3.36	4.38	4.38

The first point to note is that the unadjusted mean for variety A, the only one appearing in every year, remains unchanged when either FITCON or REML1 is used. This would also apply to differences between varieties in pairs that are always grown in the same years. For other means and contrasts, however, simple averaging and FITCON give very different results. Inspection of the fitted year means provides an explanation. Neither variety B nor C were grown in one of the highestyielding years, so FITCON has adjusted the yields of these varieties upwards.

As judged by the the standard errors of differences, FITCON gives much more reliable results than simple averaging. This experience is common. Nevertheless, there is a possibility that adjustments might sometimes be made when they are not needed and this can result in decreased accuracy instead of the expected improvement.

REML1 avoids this problem. It makes almost the same adjustments as FITCON when differences between years are large, smaller adjustments than FITCON when the differences are smaller, and none at all if the differences are zero.

REML2 differs from REML1 only in that variety effects are treated as random instead of fixed. Yet the resulting means are very different, with yields of the best varieties decreased and yields of the poorest varieties increased, i.e. shrinkage has occurred. Varieties that are grown in only a few years are most affected.

The FITCON method of adjustment can be applied using either a multiple regression program or REML with a fixed effects model. For the REML1 and REML2 adjustments the REML package is needed. If only a multiple regression program is available, FITCON is the only possible adjustment to be made. If the REML package is at hand, REML1 should be the recommended adjustment, although the estimates obtained with REML1 will very rarely differ significantly from those obtained with FITCON.

In a varieties x years or vintages x years table, entries are means over a certain number **n** of trials. It is not uncommon that the number **n** of trials differs widely from one entry to another. As the variance of a single entry is inversely proportional to **n**, analysing the varieties x years table as though all entries were equally accurate would result in loss of efficiency and possibly misleading estimates of error. As will be explained in Section 9, if the values of **n** are known, this could be corrected by performing a weighted analysis, using the numbers of trials as weights. If the values of **n** are not available, there is no point in making a REML1 adjustment instead of a FITCON, as the improvement in the estimates will be negligible compared with the errors resulting from running an unweighted analysis.

8. ANALYSIS OF A VINTAGES X YEARS TABLE

Let us see how to extend the analysis of a varieties x years table to a vintages x years table. Results for twenty three spring wheat varieties tested between 1968 and 1989 in the region South of Parana State, Brazil, and released between 1969 and 1990, will be used for purpose of illustration. The varieties annual mean yields, in t/ha, are shown on Table 5.

We want to analyse the table of vintages x years, in order to obtain the vintages mean yields adjusted for years differences, and the yield gains over vintage 69 of all other vintages, as well as their standard errors.

As we can see from Table 5, there are a few vintages that are represented by more than one variety in a particular year, thus giving rise to several observations for particular combinations of vintages and years. We handle this by including all the multiple observations for each combination of year and vintage, and performing a least squares analysis (FITCON), using either a multiple regression program, or a general purpose package like REML.

The relevant parts of the output from the multiple regression analysis using MINITAB, and from the analysis using the REML package are given in Appendix C.

The estimates of the regression coefficients corresponding to vintages 70 to 90, produced by MINITAB, are the adjusted yield gains of these vintages over vintage 69, with their corresponding standard errors. To calculate, from the regression estimates, the adjusted mean yield of a particular vintage, we apply the formula

Table 5. Mean yields (t/ha) of wheat cultivars in the region South of Parana State, Brazil - 1968/1989

Variety	Vintage	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978
1	69	1.63	1.13	1.73	1.85	0.58					1.73	2.31
2	69	1.44	1.27	1.71	1.80	0.59	0.74	1.45	0.83			
3	70	1.74	1.57	1.93	1.91	0.39						
4	71		1.60	1.99	1.91	0.40						
5	72			2.07	2.51	0.82	1.47	1.82	1.53	1.36	2.06	2.57
6	72			2.41	2.44	0.74	1.39	1.61	1.17	1.19	1.89	2.46
7	72			2.11	2.29	0.92	1.81	1.47	1.41	1.03		
8	74			2.17	2.22	0.85	1.69	1.45	1.23	1.13	1.87	
9	75			2.21	2.31	0.80	1.33	1.50	1.00	1.00		
10	76									1.51	1.60	2.63
11	78									1.44	1.89	2.67
12	78									1.78	2.19	2.45
13	80									1.51	1.73	2.41

Table 5 - (Continued)

Variety	Vintage	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
1	69	2.02	1.11	1.52	1.11							
2	69											
3	70											
4	71											
5	72	2.36	1.52	1.55	1.29	1.66	2.36	2.80	2.73	2.92	3.08	3.83
6	72	2.30	1.47	1.55	1.35							
7	72											
8	74											
9	75											
10	76	2.28	1.39	1.95	1.14							
11	78	2.30	1.72		1.49							
12	78	2.72	1.55	1.62	1.18							
13	80	2.33	1.34	1.59	1.65							
14	82		2.13	2.01	1.78	2.11	2.91					
15	82				1.86	1.85	3.20		3.19	2.50		
16	83			1.83	1.78	2.25	2.56	2.75	3.08	3.28	2.78	
17	84			2.02	1.87	2.48	2.95				3.02	3.84
18	85					1.69	2.69	2.94	2.81	2.89	3.08	
19	88								3.66	3.43	3.45	4.61
20	89										3.54	5.02
21	90										3.81	5.11
22	90										3.40	4.96
23	90										3.59	4.58

Constant + Vintage Coeff. + [Mean of years coefficcients]

= 1.5759 + Vintage Coeff. + [0.00 - 0.2257 + ... + 1.8539]/22 = 1.5759 + Vintage Coeff. + 0.08442

= 1.660 + Vintage Coefficient.

As an example for the first few vintages we obtain the following adjusted mean yields

Vintage 69: 1.660 + 0.000 = 1.660

Vintage 70: 1.660 + 0.0822 = 1.742

Vintage 71: 1.660 + 0.0867 = 1.747

Vintage 72: 1.660 + 0.39692 = 2.057

Vintage 74: 1.660 + 0.33606 = 1.996

Please note that these adjusted means calculated from the regression coefficients produced by a multiple regression program, coincide with the mean effects produced by REML when both vintages and years effects were specified as fixed.

Table 6 presents the relative yield gains over vintage 69 of the other vintages and the standard errors of these gains.

Table 6. Relative yield gains over vintage 1969 (in %)

Vintage	Yield gain	Std Error
1970	4.95	5.97
1971	5.22	6.56
1972	23.9	3.49
1974	20.2	5.05
1975	15.0	5.31
1976	21.8	5.51
1978	29.3	4.68
1980	22.3	5.51
1982	45.1	5.49
1983	39.9	5.89
1984	47.2	6.35
1985	33.8	6.67
1988	71.7	7.70
1989	85.6	10.12
1990	83.3	7.66

9. IMPROVED ESTIMATES OF MEAN YIELDS BY WEIGHTED LEAST SQUARES

In the previous analysis, we treated all entries in Table 5, i.e. the variety means at each year, as if they were equally accurate, that is, as if they were all based on the same number of trials, which is very unlikely. It is possible to obtain more reliable estimates of the vintage mean yields if, knowing the number of trials for each combination of variety by year, we perform a weighted analysis, weighting by the number of trials of a given vintage in each year.

It is possible to perform a weighted FITCON analysis using either a multiple regression program or the general purpose package REML. Let us illustrate the procedure with fictitious data presented in Table 7, which give the number of trials or locations in each year.

Table 7. Mean yields of spring wheat cultivars (t/ha)

Variety	Vintage	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978
1	69	1.63 (1)	1.13 (2)	1.73 (2)	1.85 (3)	0.58 (4)					1.73 (4)	2.31 (3)
2	69	1.44 (1)	1.27 (2)	1.71 (2)	1.80 (2)	0.59 (3)	0.74 (3)	1.45 (2)	0.83 (2)			
3	71		1.60 (1)	1.99 (1)	1.91 (3)	0.40 (3)						
4	71			2.35 (1)	2.09 (2)	0.75 (2)	1.11 (3)	1.29 (4)	1.60 (4)	1.37 (3)	1.66 (3)	
5	72			2.07 (1)	2.51 (2)	0.82 (2)	1.47 (3)	1.82 (4)	1.53 (4)	1.36 (4)	2.06 (3)	
6	72			2.41 (1)	2.44 (1)	0.74 (2)	1.39 (3)	1.61 (4)	1.17 (4)	1.19 (4)	1.89 (4)	2.46 (4)
7	73			2.11 (1)	2.29 (2)	0.92 (2)	1.81 (2)	1.47 (2)	1.41 (3)	1.03 (2)		
8	73				2.22 (1)	0.85 (2)	1.69 (3)	1.45 (4)	1.23 (4)	1.13 (4)	1.87 (4)	

(Numbers in parenthesis indicate number of trials).

The relevant parts of the output from the weighted multiple regression analysis using MINITAB, and from the weighted FITCON analysis using the REML package are given in Appendix D. In the MINITAB program we need to give the values of a new variate called **Trials**, and then give the name of this variate as parameter of the subcommand WEIGHTS of the command REGRESSION. In the REML program, we have to define **trials** as a variate, read its values and give its name as parameter of the 'WEIGHT' directive. Interpretation of the output is identical to the unweighted case.

Of course, it is preferable that data for individual locations be available, in which case a joint varieties x locations x years or vintages x locations x years analysis could be performed. However, a possible drawback of this joint analysis is the sheer bulk of data that have to be handled simultaneously when numbers of varieties and locations are large. An alternative approach is to analyse the data in stages.

Three main stages can be identified. The first consists of analysing each individual trial. Data going forward to the second stage consist of a set of variety (vintage) means for each trial, together with standard errors. This second stage consists of the across locations analysis within each year. As a result of it an over-locations summary is prepared for each year. Then the results from all the years included are put together and the across years analysis performed.

10. CONCLUSIONS

The paper has illustrated the use of simple procedures for estimating adjusted mean yields from unbalanced data of varieties or vintages over years. This type of data is especially important in estimating yield gains in plant breeding programs, where new varieties are continuously being included in the programs and older varieties that were not successful or are no longer grown are dropped from the trials.

An important conclusion from the paper is that the method proposed by economists (Byerlee, 1990; Godden, 1988; Godden and Brennan, 1987) for analysing such data sets coincide with the one advocated by statisticians (Patterson, 1978; Finney, 1980). Both approaches give identical estimates of adjusted vintage or variety yields.

Economists have generally also included an additional step: calculating either the average annual yield gain or the average growth rate of yields due to release of new varieties (Byerlee, 1990; Godden, 1988). The general approach is to fit either a linear or an exponential model of yield growth over years. Although the exponential model has been reported to give a fit as good or better than the linear one (Byerlee, 1990), it seems to me that the exponential model is less realistic than the linear one. In any case it will be necessary to test the adequacy of the model.

Both linear and exponential models are represented by equations (3) and (4) respectively:

$$y_{ijk} = a + gk + \epsilon_{ij} \tag{3}$$

$$\ln(y_{ijk}) = a + gk + \epsilon_{ij} \tag{4}$$

where

y_{ijk} is the mean yield of variety i in year j, and k is the year of release of that variety,

$\ln(y_{ijk})$ is the natural logarithm of y_{ijk},

in model (3), g is the average annual yield increase in absolute terms (i.e. g measures t/ha/year yield gains),

while in model (4) g is the average yield increase in relative terms (i.e. 100g measures the percent per year yield increase due to the release of new varieties), and exp(g) is the average annual proportion of yield gain (i.e. 100exp(g) measures the percent yield gain on a given year over the previous year, due to the release of new varieties).

In both models ϵ_{ij} is an error term with a Normal distribution with mean 0 and variance σ^2/n_{ij} where n_{ij} is the number of trials in which variety i was tested in year j.

Whichever model is used, it is more appropriate to estimate g by performing a weighted regression with the numbers n_{ij} of trials (locations) as weights.

Even so, caution must be used in interpreting results. In the first place, genetic gains were estimated for released varieties based (usually) on experimental station yield trials. These results will only be relevant (a) if farmers plant a mosaic of varieties similar to the released varieties (in fact, farmers are more likely to emphasize higher yielding varieties--see Byerlee and Moya) and (b) if relative yield gains of new varieties under farmers´ management are similar to those achieved on station. In the first case, actual gains in farmers´ fields can be emphasized by a Varietal Improvement Index (Brennan, 1986), if data are available on varieties planted by farmers over years.

Finally, an important underlying assumption for the estimation of vintage mean yields is that, over and above varieties x years interaction, yields of a given variety remain fairly constant over time. In fact most varietal yield trials include long term checks whose yields decline over time, as a result of the breakdown of disease resistance. Consequently, mean yields of vintages corresponding to these long term checks are usually underestimated. Methods to avoid underestimating vintage mean yields when check varieties decline steadily over time need to be developed.

References

Brennan, J.P. (1986). Impact of Wheat Varieties from CIMMYT on Australian Wheat Production. Agricultural Economics Bulletin 5. Department of Agriculture New South Wales.

Byerlee, D. (1990). Technical Change and Returns to Wheat Breeding Research in Pakistan's Punjab in the Post-Green Revolution Period. PARC/CIMMYT Paper 90-7.

Byerlee, D. and Moya, P. (1991). Impacts of International Wheat Breeding Research in the Developing World, 1965-1990. Draft paper.

Finney, D.J. (1980). The estimation of parameters by least squares from unbalanced experiments. J. Agric. Sci. **95**, 181-189.

Godden, D. (1988). Technical change embodied in new varieties of English winter wheat and spring barley. Research and Development in Agriculture, **5**(2), 117-122.

Godden, D. and Brennan, J. (1987). Technological Change Embodied in Southern N.S.W. and British Wheat Varieties. Paper presented at the 31st Annual Conference of the Australian Agricultural Economics Society, University of Adelaide, 9-12, 1987.

MINITAB (1989). MINITAB Reference Manual, Release 7. Minitab Inc., State College, Pennsylvania.

Patterson, H.D. (1978). Routine least squares estimation of variety means in incomplete tables. J. natn. Inst. agric. Bot., **14**, 401-412.

Patterson, H.D. (1982). FITCON and the analysis of incomplete varieties x trials tables. Utilitas Mathematica, **21A**, 267-289.

Patterson, H.D., Thompson, R., Hunter, E.A. and Williamns, E.R. (1989). Analysis of non-orthogonal data using REML. Edited by R.A. Kempton. Scottish Agricultural Statistics Service, Edinburgh.

Robinson, D.L. (1987). Program REML. Scottish Agricultural Statistics Service, Edinburgh.