Journal of Agricultural Science (2003), **140**, 383–393. © 2003 *Cambridge University Press*
DOI: 10.1017/S0021859603003332 *Printed in the United Kingdom*

Analysis of crop lodging using a latent variable model

D. J. ALLCROFT* AND C. A. GLASBEY

Biomathematics and Statistics Scotland, King's Buildings, Edinburgh, EH9 3JZ, Scotland

(*Revised MS received 29 May 2003*)

SUMMARY

A method for quantifying treatment comparisons for a situation in which there are too many zeros in the dataset for a conventional analysis of variance to be valid is presented. The method assumes the existence of a latent variable such that zero observations correspond to values below a threshold, and non-zero observations are transformed to fit the part of the distribution above the threshold. The method is known as Tobit analysis in econometrics. Parameters are estimated by maximum likelihood and standard errors obtained, all using standard numerical routines. Use of the method is demonstrated by analysis of a dataset of crop lodging, and it is anticipated to be widely applicable to other types of data for which high numbers of zeros prevent conventional analysis.

INTRODUCTION

There are many situations in which routine statistical analyses are performed, despite clear violation of the assumptions underlying the method of analysis. In such cases the results from the analysis will be invalid, however they are still often quoted, due to the absence of any established alternative method. One such example of this is addressed in the present paper, namely the use of analysis of variance (ANOVA) when there are many zeros present in a dataset. This will often cause the assumption of normality to be violated, a histogram of the raw data having a large spike at zero. The residual variation will be underestimated and so the value of F and hence significance levels of any treatment effects are likely to be overestimated.

The analysis of a model dataset of lodging (flattening of the crop) from a series of field trials evaluating variety performance is considered. Current practice in such trials is to estimate trial and variety effects using a conventional ANOVA, performed on untransformed data. However there is a preponderance of zeros and so, for the reasons described above, such analyses are clearly invalid. A possible alternative approach would be to use a two-stage process, firstly modelling the probability of zero lodging with a Bernoulli (0/1) variable and secondly using some continuous distribution to describe the % lodging when it occurs. Such modelling strategies were suggested by Aitchison (1955) and have been used by Pennington (1983) for plankton abundance estimation and Feuerverger (1979) for rainfall occurrence. Similar approaches for count data have been applied to defects in manufacturing by Lambert (1992) and to animal abundance estimation by Welsh *et al.* (1996). In the present paper an alternative, simpler methodology is developed, which allows both occurrence and amount of the measured quantity to be described by a single variable. This latent Gaussian variable is constructed such that zero observations correspond to the part of the distribution below some threshold, and non-zero observations are transformed to fit the distribution above the threshold. The method therefore effectively assumes that the zero observations are *censored*, i.e. an observed zero represents a response that simply did not score high enough on the given scale to produce a non-zero observation. Such models are commonly used in the field of econometrics, where they have been called Tobit models, first introduced by Tobin (1958) as limited dependent variable models. A good review of the various methodologies developed can be found in Amemiya (1984), and the main details are given in some standard econometrics textbooks, e.g. Johnston & DiNardo (1997). These models are relatively unknown in other disciplines, yet the idea is appealing for many physical and biological phenomena. Here, in some trials, most varieties display some lodging, but in other trials most do not, so the idea of a threshold, with different levels in different trials, seems entirely plausible. Similarly, with some varieties being more prone to lodging than others, a threshold that differs between varieties makes sense.

* To whom all correspondence should be addressed. Email: dave@bioss.ac.uk

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press

Table 1. *Dataset for the first season. Figures are per cent crop lodging*

	Trial						
Variety	1	2	3	4	5	6	7
1	0·0	0·0	0·0	0·3	7·7	0·0	0·4
2	0·0	0·0	0·0	0·0	1·7	0·0	0·0
3	66·7	1·3	0·7	1·0	6·7	0·0	0·0
4	0·0	0·0	0·0	0·0	0·0	0·0	0·0
5	0·0	0·0	0·0	0·0	2·7	0·0	0·0
6	0·0	0·0	0·7	0·3	10·0	0·0	0·0
7	0·0	0·0	0·0	0·0	5·0	0·0	0·0
8	3·3	0·0	0·0	1·7	28·3	0·3	0·0
9	0·0	0·0	0·0	0·0	37·7	0·0	0·0
10	0·0	0·0	0·0	0·0	1·0	0·0	0·0
11	0·0	0·0	0·0	0·0	4·3	0·0	0·0
12	0·0	0·0	0·0	0·0	1·0	0·0	0·0
13	0·0	0·0	0·0	0·3	30·0	0·0	0·0
14	0·0	0·0	0·0	1·3	1·7	0·0	0·0
15	13·3	0·0	0·0	2·7	13·3	0·3	0·0
16	10·0	0·0	0·0	0·0	8·3	0·3	0·2
17	3·3	0·0	0·0	0·0	4·3	0·0	0·2
18	0·0	0·0	0·0	0·7	0·0	0·0	0·0
19	26·7	0·0	0·0	0·7	2·3	0·0	0·0
20	0·0	0·0	0·0	0·0	2·0	0·0	0·0
21	0·0	0·0	0·0	0·0	2·7	0·3	0·0
22	0·0	0·0	0·0	0·3	3·3	0·0	0·0
23	26·7	0·0	0·3	0·0	16·7	0·3	0·2
24	0·0	0·0	0·0	2·0	15·0	0·0	0·0
25	0·0	0·0	0·0	1·0	11·7	0·3	0·0
26	0·0	0·0	0·0	0·0	0·0	0·0	0·2
27	0·0	0·0	0·0	0·0	13·3	0·0	0·0
28	0·0	0·0	0·3	0·0	5·0	4·0	0·2
29	0·0	0·0	0·0	1·0	11·0	0·0	0·0
30	3·3	3·0	0·0	2·0	11·0	0·0	0·2
31	0·0	0·3	0·3	0·0	9·3	0·3	0·0
32	30·0	1·3	0·0	0·3	8·3	0·0	0·0

The data and method are described in the next sections. Then, after *Results*, model diagnostics are considered; in addition, tests carried out to validate the method are also described.

DATA

Three similar datasets for successive seasons are considered, consisting of the percentage lodging of a cereal crop for between four and seven trials, involving between 27 and 34 varieties. For reasons of commercial confidentiality, further details cannot be given. Each season can be regarded as a complete block design, with variety the *treatment* effect and trial the *block* effect. In this paper, the authors concentrate on the dataset from the first season, involving seven trials and 32 varieties. The full dataset for this season is shown in Table 1. Sixty-six per cent of the values (147 out of 224) are recorded as zero.

A naive ANOVA on these data concludes there to be large variety and trial effects, but with so many observed zeros, assumptions have clearly been violated. To illustrate this, Fig. 1 shows plots of residuals, standardized to have unit variance, against fitted values obtained from ANOVA on untransformed data and square root transformed data. Neither plot is as it should be for a valid ANOVA, because for all the zero observations the residuals are -1 times the fitted values, hence all these residuals lie on a straight line. This also results in the variance decreasing as the size of the fitted value decreases. The square root transformation goes some way towards improving the constant variance assumption, but nevertheless the dominant feature of both plots is this line of residuals corresponding to the observed zeros and so whatever transformation is used, such a plot will never satisfy the requirements of appearing random with homogeneous variance.

METHOD

Transformation of data

The first step in this method is to find a transformation to a latent Gaussian variable, such that observed zeros correspond to the part of the distribution below a threshold, and non-zero observations (greater than zero) fit the part of the Gaussian distribution above the threshold. In order to investigate suitable transformations a large volume of data is desirable and so the combined data from all three seasons (full dataset not shown) are used for the exploratory work in this section, giving $n=495$ values, of which $n_0=284$ (57%) are zero. In the rest of the paper the data from the first season only are used.

To maintain a simple approach, only simple transformations of the data are considered, namely square root and log. Figure 2 illustrates firstly that using a square root transformation, a histogram of the transformed non-zero data values is a reasonable fit to the upper tail of a Gaussian distribution, and secondly that a plot of the normal scores of these non-zero values is fairly linear, further indicating that the transformation is a reasonable one. In order to produce these plots it was necessary to estimate both the threshold level and the variance of the Gaussian distribution. This is non-trivial, but not required for the actual fitting of the model as described in the next section, hence details are relegated to the Appendix.

Using untransformed or log transformed data, the corresponding graphs to Fig. 2 showed poor fits, the histograms having worse fits to the upper tail of the Gaussian distribution, and the normal probability plots displaying curvature, convex and concave respectively, reinforcing the choice here of the square root transformation.

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press

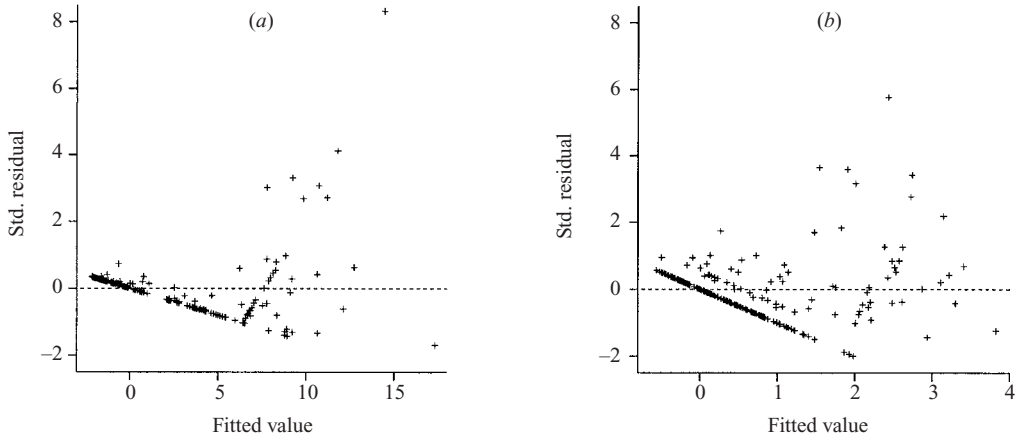


Fig. 1. Diagnostic plots resulting from conventional ANOVA – standardized residuals plotted against fitted values: (*a*) untransformed data; (*b*) square root transformed data.

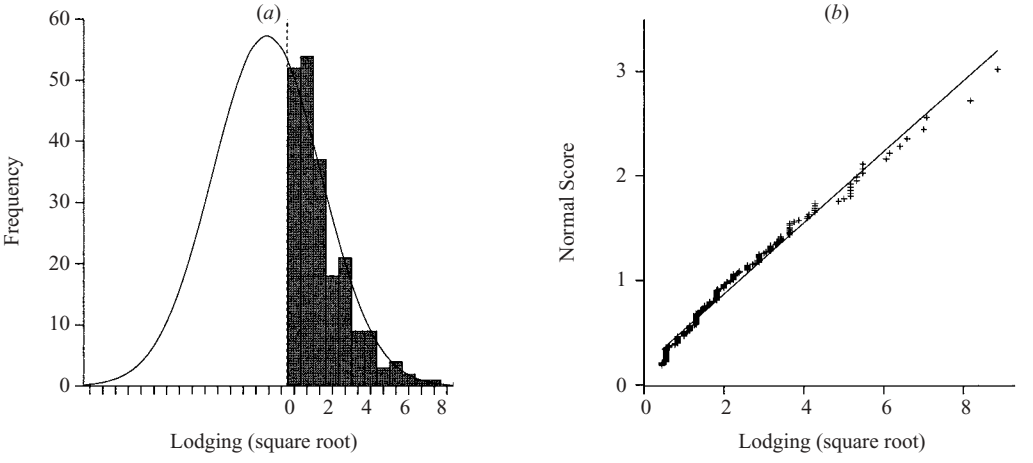


Fig. 2. (*a*) Histogram of square root transformed lodging, with superimposed thresholded Gaussian distribution. (*b*) Normal probability plot of upper tail to demonstrate suitability of square root transformation: (+) observed normal scores, (−) expected normal scores.

Note that in ANOVA the assumption being made is strictly that the *errors* are normally distributed, not the data themselves, though in practice, if the errors are normally distributed, the data will usually be approximately normal too, though this depends on the sizes of treatment and block effects. Therefore the ideas just described are useful in assessing the suitability of a transformation, but it is more important to examine the residuals after the model has been fitted.

Model estimation

The square root of the observed lodging for variety i in trial j will be denoted by y_{ij}, and z_{ij} will denote the corresponding value of the latent variable. Thus

$$y_{ij} = \begin{cases} z_{ij} & \text{if } z_{ij} \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

Consider the fixed effects model for the complete block design with V varieties and T trials:

$$z_{ij} = \mu + \nu_i + \tau_j + \varepsilon_{ij} \qquad (1)$$

Here μ is an overall mean, ν_i, $i = 1, \ldots, V$, are the variety effects, τ_j, $j = 1, \ldots, T$, are the trial effects, and the errors ε_{ij} are independent and identically distributed $N(0, \sigma^2)$. The last variety and trial effects are

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press

constrained such that the sum of each set of effects is zero, i.e.

$$\nu_V = -\sum_{i=1}^{V-1} \nu_i \quad \text{and} \quad \tau_T = -\sum_{j=1}^{T-1} \tau_j$$

The complete set of $(V+T)$ parameters will be written as $\theta = (\mu, \sigma^2, \nu_1, \ldots, \nu_{V-1}, \tau_1, \ldots, \tau_{T-1})$.

Equation 1 states the model in terms of the latent variable z, however it is also convenient to consider the corresponding expression in terms of the (square root transformed) lodging y:

$$y_{ij} = \mu + \nu_i + \tau_j + e_{ij} \tag{2}$$

The terms denoted by e_{ij} are such that $e_{ij} = \varepsilon_{ij}$ if $z_{ij} \geqslant 0$ and e_{ij} is an upper bound for ε_{ij} if $z_{ij} < 0$. The likelihood for the model can be written as the product of the $V \times T$ terms:

$$L = \prod_{y_{ij} > 0} \phi\left(\frac{e_{ij}}{\sigma}\right) \prod_{y_{ij} = 0} \Phi\left(\frac{e_{ij}}{\sigma}\right) \tag{3}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution,

$$\Phi(x) = \int_{X=-\infty}^{x} \phi(X)\, \mathrm{d}X \tag{4}$$

$\phi(\cdot)$ being the probability density function of the standard normal distribution,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right] \tag{5}$$

The likelihood quoted is equivalent to that for the Tobit model, given in e.g. Johnston & DiNardo (1997). So when lodging greater than zero is observed, a component for the associated error enters directly into the likelihood using the probability density function of the standard normal distribution, and when zero lodging is observed, the component in the likelihood reflects that we know that the error is less than a value determined by its trial and variety, and so the term involves the cumulative distribution function. This likelihood can be maximized using a suitable numerical routine to give maximum likelihood estimates for all parameters, $\hat{\theta}$. See *Validation of method, Numerical routines* for further details.

Note that for variety 4 there is no lodging in any of the seven trials. Such varieties are potentially a problem for the estimation procedure. A solution is to artificially set the lodging in one of the trials to a small positive number for any variety with completely zero lodging. It is most natural to do this for the trial with the highest level of lodging, here trial 5. This provides an anchor for that variety and allows its effect to be estimated, otherwise the maximum likelihood estimate for that variety effect would be negative infinity. Such an adjustment has negligible effect on the parameter estimates for other varieties. This was checked by re-running the analysis without the all-zero variety present. A small positive number of 0·01 was used, so one magnitude smaller than the smallest observed non-zero observation. Other small positive numbers were tried to confirm that the actual value used did not affect the resulting variety comparisons.

In addition to the estimated parameters $\hat{\theta}$, it might be useful to calculate and quote some alternative quantities. For example, the estimated mean for variety i on the latent scale is given by

$$\hat{\mu}_i = \hat{\mu} + \hat{\nu}_i$$

In the presentation of results from analyses involving transformations, it is usually desirable to back-transform estimates to the observed scale. For the present problem, natural quantities to present for each variety are the probability of observing zero lodging and the expected level of lodging conditional on it being non-zero. The probability of observing a zero value for variety i may be estimated by $\Phi(\hat{\xi}_i)$, where $\hat{\xi}_i = -\hat{\mu}_i/\hat{\sigma}$. The expected level of lodging, conditional on it being non-zero, is equivalent to the expectation of y^2 for a truncated normal distribution. This may be calculated using expressions given in Johnson *et al.* (1994), as

$$\left[1 + \frac{\phi(\hat{\xi}_i)}{1-\Phi(\hat{\xi}_i)}\hat{\xi}_i\right]\hat{\sigma}^2 + 2\frac{\phi(\hat{\xi}_i)}{1-\Phi(\hat{\xi}_i)}\hat{\mu}_i\hat{\sigma} + \hat{\mu}_i^2$$

Comparison of variety effects

The main objective of the analysis is to assess whether there are differences in levels of lodging between varieties. The effects obtained above allow a ranking of varieties, but standard errors are required if the significance of differences is to be assessed. Because maximum likelihood has been used to estimate parameters, the variance–covariance matrix of the parameter estimates can be obtained as the elements of the inverse of the Hessian matrix H, with elements given by

$$H_{ij} = -\frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j}$$

The matrix is easily computed using a numerical routine (see *Validation of method, Numerical routines*), and can then be inverted, again using a numerical routine, to obtain H^{-1}. The variance of the estimator $\hat{\theta}_i$ is then given by $(H^{-1})_{ii}$, and the covariance between any pair of estimators $\hat{\theta}_i$ and $\hat{\theta}_j$ is given by $(H^{-1})_{ij}$. In order to compare variety effects, the $(V-1)\times(V-1)$ submatrix C of H^{-1}, corresponding to $\hat{\nu}, \ldots, \hat{\nu}_{V-1}$, can be used. The diagonal elements of

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press

C then give the variances of the estimates of the variety effects

$$\mathrm{Var}(\hat{\nu}_i) = C_{ii} \quad \text{for } i = 1, \ldots, V-1$$

and the non-diagonal elements the covariances between pairs of variety effects

$$\mathrm{Cov}(\hat{\nu}_i, \hat{\nu}_j) = C_{ij} \quad \text{for } i, j = 1, \ldots, V-1$$

For the Vth variety, we have

$$\mathrm{Var}(\hat{\nu}_V) = \sum_{i=1}^{V-1} \sum_{j=1}^{V-1} C_{ij}$$

and

$$\mathrm{Cov}(\hat{\nu}_i, \hat{\nu}_V) = -\sum_{j=1}^{V-1} C_{ij}$$

Note that to quote standard errors of variety means (on the latent scale), information about the overall mean μ must be incorporated, i.e.

$$\mathrm{Var}(\hat{\mu}_i) = \mathrm{Var}(\hat{\mu}) + \mathrm{Var}(\hat{\nu}_i) + 2\,\mathrm{Cov}(\hat{\mu}, \hat{\nu}_i)$$

The standard errors of differences (S.E.D.) between pairs of varieties are determined by taking the square root of

$$\mathrm{Var}(\hat{\nu}_i - \hat{\nu}_j) = \mathrm{Var}(\hat{\nu}_i) + \mathrm{Var}(\hat{\nu}_j) - 2\,\mathrm{Cov}(\hat{\nu}_i, \hat{\nu}_j)$$

Pairwise comparisons can then be made via the usual t-tests, i.e.

$$\frac{|\hat{\nu}_i - \hat{\nu}_j|}{\sqrt{\mathrm{Var}(\hat{\nu}_i - \hat{\nu}_j)}}$$

By analogy with ANOVA, the degrees of freedom of this t-statistic should be $(V-1)(T-1)$. However, in the presence of censoring, many of these degrees of freedom are associated with censored observations and so this will be optimistic. In the absence of any obvious way to calculate the reduction in effective degrees of freedom, the commonly adopted practice of using 2·0 as the 95% level of the test statistic will be followed.

RESULTS

Table 2 shows the parameter estimates for the dataset from the first season. Estimated effects and standard errors, all on the latent scale, are given for all parameters. In addition, for each variety, the probability of observing zero lodging and the corresponding expected level of lodging given that it is non-zero are presented. As discussed below, if none of the data were censored, the estimated standard errors (E.S.E.) would be the same for all variety effects. However, comparison with the raw data in Table 1 shows that the varieties with the most censored values have the largest E.S.E.S. This is also easily seen for the trial effects – e.g. trials 2 and 3 have the least number of non-zero lodgings and, consequently, the trial effects for these trials have the largest E.S.E.S.

Table 2. *Parameter estimates for dataset from first season. All estimates and estimated standard errors (E.S.E.S) are on the latent scale*

	Estimate	E.S.E.	Prob. of observing zero	Expected value given non-zero
Grand mean	−0·963	0·220		
Residual variance	2·495	0·427		
Trial means:				
1	−0·05	0·376		
2	−2·51	0·562		
3	−2·39	0·532		
4	−0·61	0·358		
5	2·47	0·282		
6	−1·74	0·451		
7	−1·93	0·465		
Variety means:				
1	−0·53	0·739	0·63	1·92
2	−2·80	1·237	0·96	0·72
3	1·67	0·623	0·15	6·00
4	−3·67	1·406	0·99	0·53
5	−2·61	1·195	0·95	0·78
6	−0·34	0·722	0·59	2·11
7	−2·30	1·132	0·93	0·88
8	0·62	0·674	0·35	3·43
9	−0·84	0·883	0·70	1·65
10	−2·99	1·274	0·97	0·67
11	−2·38	1·149	0·93	0·85
12	−2·99	1·273	0·97	0·67
13	−0·50	0·796	0·63	1·95
14	−1·48	0·890	0·83	1·24
15	0·71	0·671	0·33	3·59
16	0·35	0·671	0·41	2·99
17	−0·49	0·743	0·62	1·96
18	−2·32	1·026	0·93	0·87
19	0·06	0·729	0·48	2·58
20	−2·74	1·223	0·96	0·74
21	−1·34	0·849	0·80	1·32
22	−1·50	0·892	0·83	1·23
23	1·23	0·626	0·22	4·73
24	−0·69	0·812	0·67	1·78
25	−0·34	0·733	0·58	2·11
26	−2·08	0·955	0·91	0·96
27	−1·69	1·017	0·86	1·13
28	0·19	0·655	0·45	2·75
29	−0·94	0·835	0·72	1·58
30	0·86	0·637	0·29	3·89
31	0·15	0·650	0·46	2·69
32	0·86	0·659	0·29	3·90

Table 3 displays the t-statistics for all pairwise comparisons that resulted in a test statistic greater than 2·0 and hence can be considered significant at the 5% level. For ease of viewing, the varieties

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press

Table 3. *Pairwise variety comparisons with* t*-statistic greater than 2·0. Varieties are presented in order of their estimated effects; the smallest effects correspond to rows, increasing down the table from the first row, and the largest effects correspond to columns, increasing to the right-hand side*

		Increasing variety effects →												Largest effects
		17	6	25	19	31	28	16	8	15	30	32	23	3
Smallest effects	4	2·02	2·12	2·12	2·38	2·47	2·50	2·60	2·77	2·83	2·94	2·93	3·19	3·48
	12				2·11	2·20	2·23	2·34	2·52	2·59	2·71	2·70	2·98	3·30
	10				2·10	2·20	2·23	2·34	2·52	2·59	2·71	2·70	2·98	3·30
	2				2·02	2·12	2·15	2·26	2·45	2·52	2·64	2·63	2·92	3·24
	20					2·10	2·12	2·23	2·43	2·49	2·62	2·61	2·90	3·23
	5					2·04	2·07	2·18	2·38	2·44	2·57	2·56	2·86	3·19
Increasing	11							2·08	2·28	2·35	2·48	2·47	2·77	3·12
variety effects	18					2·05	2·08	2·20	2·42	2·50	2·65	2·63	2·97	3·34
↓	7							2·04	2·24	2·31	2·45	2·44	2·74	3·09
	26							2·10	2·33	2·41	2·57	2·55	2·91	3·30
	27										2·14	2·12	2·45	2·83
	22										2·16	2·15	2·51	2·93
	14										2·14	2·13	2·49	2·91
	21										2·08	2·06	2·45	2·87
	29												2·09	2·52
	9													2·34
	24													2·32
	1													2·28
	13													2·17
	17													2·23
	6													2·12
	25													2·10

involved are listed in order of the size of their estimated effects.

MODEL CHECKING

As discussed in *Data*, for a conventional ANOVA, a plot of residuals versus fitted values allows the checking of both the normality of residuals and the homogeneity of variance assumptions. Figure 3(*a*) shows such a plot for the model fitted by the authors. If lodgings of zero are observed then the residuals are censored; the ▽ symbols on the graph represent not residuals themselves, but indicate that these residuals are censored at that value, i.e. may lie anywhere below this value. Clearly this plot is less informative than it would be for an uncensored dataset, nevertheless it seems plausible that had all the residuals been observed, they would be normally distributed with constant variance over the range of fitted values shown. The other plots in Fig. 3 allow further assessment of this. As the size of the fitted value decreases, there is more probability of the observation being censored, and so Fig. 3(*b*) displays a histogram of the observed and expected number of censored data points in terms of the size of the fitted value. The bins were created here such that each contains approximately the same proportion of the overall datapoints. The agreement seems reasonable. For Fig. 3(*c*) the cumulative distribution function of the standardized residuals is considered, estimating it by a modification of the Kaplan–Meier (product-limit) estimator, a common method for estimation of the survivor function for survival data (see e.g. Cox & Oakes 1984). In a typical survival dataset, the time of death t_k for patient k is either directly observed, or known to be greater than a time t_k, with the actual time of death unknown. This second case leads to a *censored* data point. If the time of death is known to be greater than t_k, the data point is *right-censored*. In contrast, the censored residuals in the present paper are *left-censored*, i.e. they are known to be less than or equal to a certain value. If the set of residuals from uncensored datapoints only are ordered, as $\{\varepsilon_{(k)}\}_{k=1,\ldots,m}$ and let $\{r_{(k)}\}_{k=1,\ldots,m}$ be the set of integers such that $r_{(k)}$ is the number of all residuals (both observed and censored) less than or equal to $\varepsilon_{(k)}$, then the product-limit estimator of the cumulative distribution function is defined as

$$\hat{F}(e) = \prod_{k:\varepsilon_{(k)} \geqslant e} \left(1 - \frac{1}{r_{(k)}}\right)$$

This results in the step-function shown in Fig. 3(*c*), which also shows the expected cumulative distribution function for a standard normal distribution. The fit is good.

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press

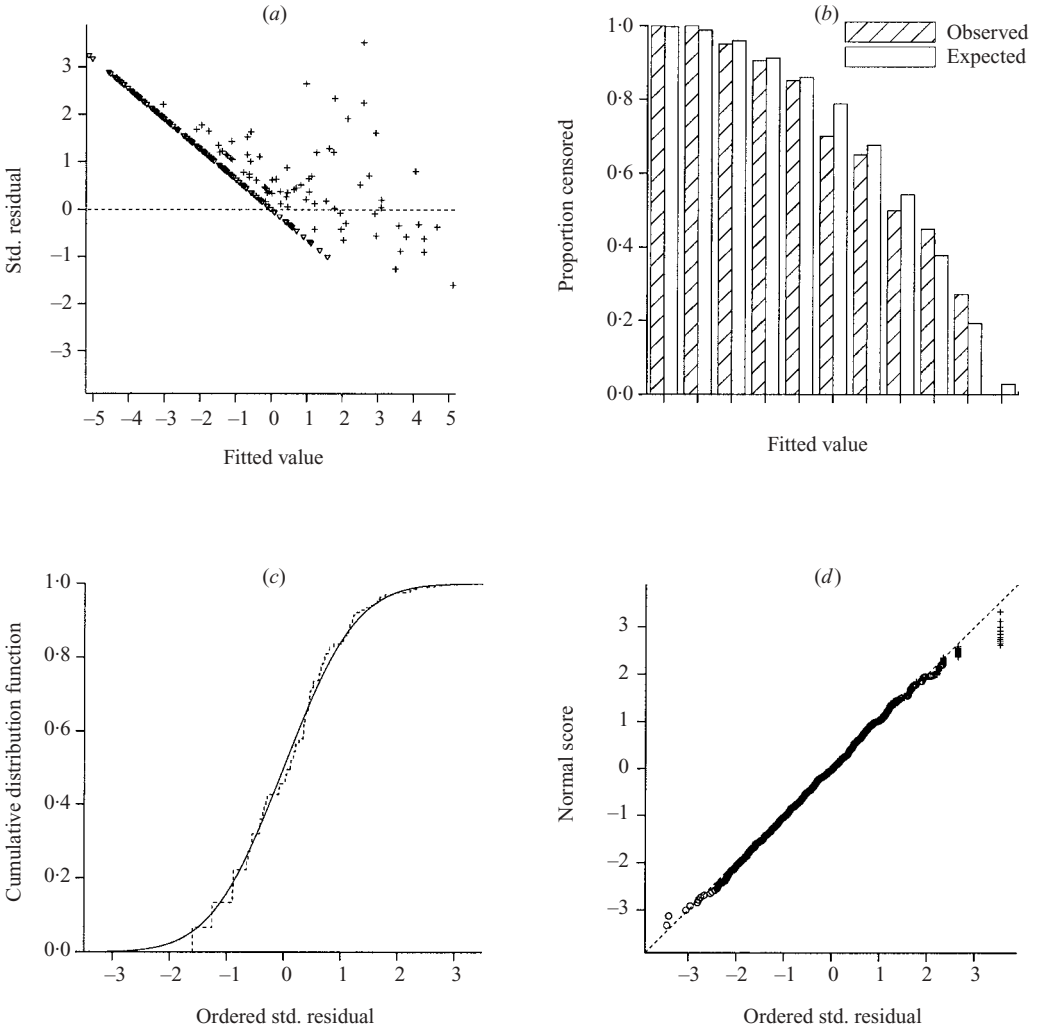


Fig. 3. Plots to assess goodness-of-fit of model fitted to square root transformed data. (*a*) Standardized residuals plotted against fitted values: (+) uncensored residual, (▽) residual censored at this value. (*b*) Proportion of censored values (observed and expected), grouping ordered fitted values into groups of 20. (*c*) Cumulative distribution function: (···) product-limit estimate, (—) expected function for a standard normal distribution. (*d*) Combined probability plot of residuals from 10 simulations of data: (+) uncensored residual, (○) simulated censored residual.

Finally, another common way to assess the normality of residuals is via a normal probability plot. The only way such a plot can be obtained in the presence of censored data is to simulate values for the censored residuals from the relevant part of the distribution, and use these as if they were from observed datapoints. Many simulations can be done and probability plots created for each realization. However, in the present paper many such plots have been combined into one, Fig. 3(*d*) showing the results obtained from 10 simulations. The linearity of the plot indicates that the residuals are indeed approximately normally distributed. All these plots therefore confirm the data to be broadly consistent with the fitted model.

It is also useful to compare such plots with those obtained using other transformations. Figure 4 therefore gives the corresponding plot to Fig. 3(*a*), with both untransformed data and log transformed data. It can be seen that in the untransformed case, the variation in residuals clearly increases as the fitted value increases. Conversely, for the log transformation, the variation decreases as the fitted value increases and there is some evidence of curvature. The square root transformation can therefore be seen to be the best of the three transformations considered.

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press

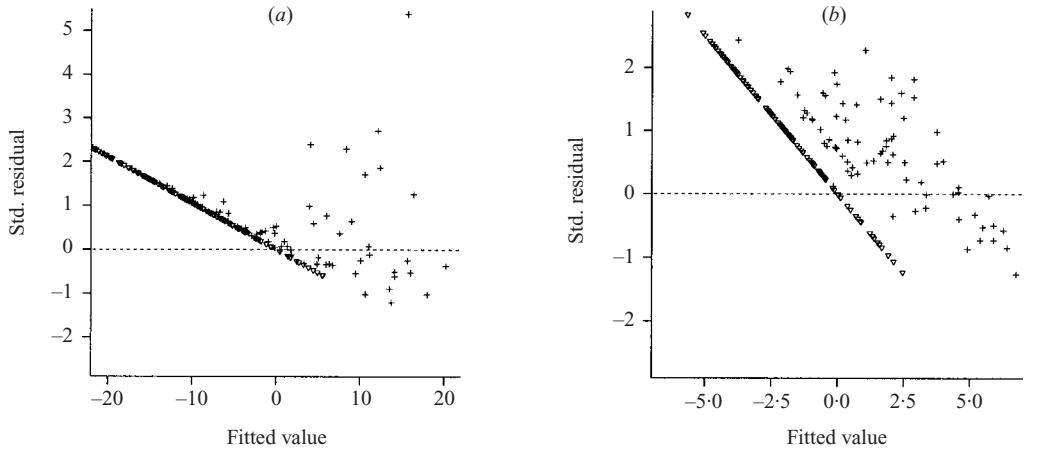


Fig. 4. Standardized residuals plotted against fitted values for other data transformations: (+) uncensored residual, (▽) residual censored at this value: (*a*) untransformed data; (*b*) log transformed data.

VALIDATION OF METHOD

Here some further work is described which was carried out to verify that the method is valid and gives the expected results.

Numerical routines

Both parameter estimates and standard errors have been obtained by off-the-shelf numerical routines, available in any programming language and many statistical packages. Specifically, the Fortran 90 programming language (Ellis *et al*. 1994) has been used, with IMSL numerical routines (IMSL 2002). A copy of the program used for the example in this paper is available from the authors on request. The likelihood was maximized using routine BCONF, a routine which minimizes a function of several variables subject to bounds on the variables (to ensure $\sigma^2 > 0$) using a quasi-Newton method and a finite-difference gradient; FDHES was used to approximate the Hessian using forward differences and function values only. For the problem in the present paper, both these procedures could have been performed analytically, through calculation of first and second derivatives of the log-likelihood function explicitly. Parameter estimates could then be obtained by solving the set of equations of the first derivatives all equated to zero, and the variance–covariance matrix found via the exact Hessian matrix obtained from the second derivatives (Stuart *et al*. 1999). These calculations were all performed and produced results very similar to those obtained numerically, the difference between parameter estimates being negligible and the elements of the Hessian all agreeing to at least 3 significant figures, hence the authors can be confident that the numerical routines are performing as intended. For the examples in the present paper, convergence was easily achieved in the likelihood maximization, from starting values of $\sigma^2 = 1$ and all other effects set to zero. If lack of convergence was encountered, an alternative set of starting values could be the estimates of variety and trial effects from a simple ANOVA on the (square root transformed) original data.

The model fitting could also be performed using procedures for fitting Tobit models that are available in some of the statistical and econometric packages. For example, LIMDEP (Greene 1986) has a TOBIT command and in the SAS package, Tobit models can be fitted using the LIFEREG procedure (SAS Institute 2000). The authors have verified that this latter package gives equivalent results to their own.

Uncensored data

For an uncensored dataset, the results obtained by the authors' program would be required to replicate results from using a standard ANOVA procedure. Here GenStat (GenStat 2000) has been used to verify this. Taking the parameter estimates obtained above, datasets were simulated of the same size as the observed. The uncensored version of this simulated dataset was then used directly to compare with the results obtained from GenStat ANOVA. Identical parameter estimates for the mean and all trial and variety effects were obtained. The only difference was in the estimation of the residual variation σ^2. This is, however, simply a consequence of the estimation method used. The method presented in the present paper is maximum likelihood and so the residual variation is estimated using a denominator of $n_V n_T$, whereas ANOVA adjusts this estimate for the degrees of freedom used to estimate the other parameters and so uses $(n_V - 1)(n_T - 1)$. This discrepancy would be

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press

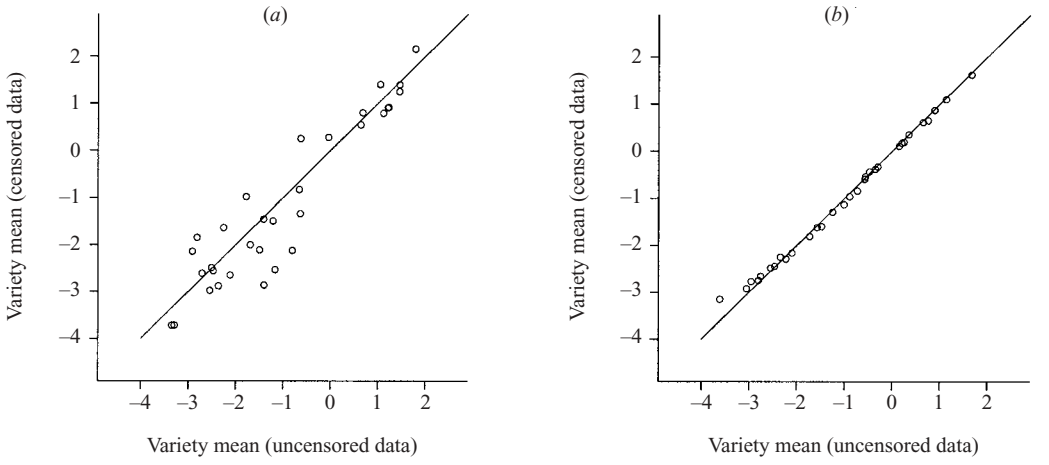


Fig. 5. Comparison of variety means estimated from censored and uncensored versions of the same simulated dataset: (*a*) for a single simulated dataset; (*b*) averaged over 100 simulated datasets.

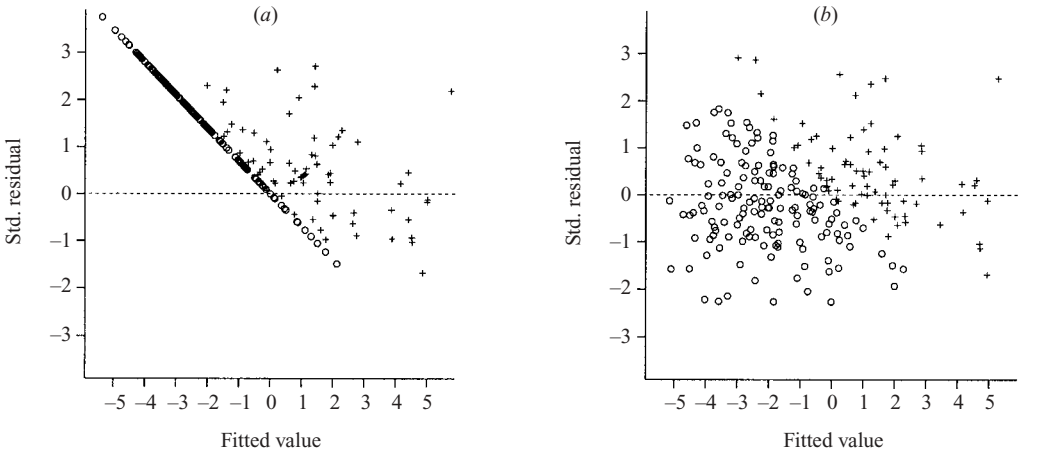


Fig. 6. Standardized residuals plotted against fitted values for a simulated dataset: (*a*) from analysis of censored version of dataset, (+) uncensored residual, (○) residual censored at this value; (*b*) from analysis of uncensored dataset, (+) uncensored residual, (○) residual simulated for censored data point.

rectified if a residual maximum likelihood technique were to be used, rather than full maximum likelihood; however in the presence of censored data, it is not obvious how this would be formulated. The result of using maximum likelihood is therefore that S.E.D.s are slightly more optimistic than those obtained by ANOVA. In the uncensored case, all varieties are equally replicated and so there is a single S.E.D. that applies to all pairwise comparisons. In contrast, as has already been observed, the estimated standard errors of variety effects increase when there is increased censoring.

Simulated datasets are also useful in verifying that the censoring of the data does not bias the parameter estimates. A simulated dataset can be censored in order to mimic the observed data and then analyses of uncensored and censored versions of the same dataset can be compared. Figure 5(*a*) allows comparison of estimated variety means for a single simulated dataset, demonstrating that the model can still produce reliable estimates when a dataset is heavily censored. (For this particular simulated dataset, 63% of values were censored.) There is more discrepancy for varieties with smaller means, as this is where most censoring occurs. Figure 5(*b*) shows the same picture averaged over 100 simulated datasets. The agreement confirms that any bias introduced as a consequence of estimating from the censored dataset is small.

Such simulated datasets may also be used to provide more insight into the residual plots. Figure 6

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press

shows residual plots obtained from a simulated dataset, (*a*) showing the residuals when the simulated data is censored before analysis, mimicking the observed data, and (*b*) showing the residuals when the uncensored dataset is analysed. Comparison of (*a*) and (*b*) shows that the censored residuals in (*a*) become approximately normally distributed in (*b*), as would be expected. Further, comparison of Fig. 6(*a*) with Fig. 3(*a*) confirms that the residual plot from the observed data is similar to that from a dataset simulated from the fitted model.

DISCUSSION

Using a dataset of crop lodging as an example, it has been demonstrated that data unsuitable for analysis by conventional ANOVA can be successfully transformed to a thresholded latent Gaussian variable and subsequently analysed by a valid full maximum likelihood method. The fact that the model can be fitted and standard errors obtained using numerical routines makes this method accessible to the general scientific community. In addition, the facilities for fitting Tobit models in some standard statistical and econometric packages should further encourage its use. The authors find it surprising that such a well-known technique within the econometrics discipline should be so little known outside of this area. It is anticipated that the method will be widely applicable to many physical and biological phenomena, including many in agriculture. Disease incidence is a good example, for which some varieties will display more resistance to disease than others, and there are many other analogous examples. The present authors have already used similar latent variable models for analysis of cow feeding behaviour (Allcroft & Glasbey 2002) and rainfall data (Allcroft & Glasbey 2003). The feeding data were binary and so the Gaussain variable was simply such that non-feeding periods corresponded to the part of the distribution below the threshold, and feeding to the part above. For the rainfall data, a quadratic power transformation was used, involving the estimation of four parameters. In that case 135 000 data values were available and so there is plenty of information to estimate these parameters. For the crop lodging data in the present paper, there is not as much data (495 values), and as the aim was to keep the model as simple as possible, only standard transformations were considered. In general, transformation parameters could be estimated simultaneously with the other parameters, by replacing y_{ij} with $f(x_{ij})$ in the model, Eqn 2, and in the subsequent likelihood, Eqn 3, where $f(\cdot)$ is the form of transformation desired and x_{ij} are data as observed.

Note that the pairwise *t*-tests between varieties are being carried out without the protection of an *F*-test for overall variety differences, as would be present in an ANOVA. However, if desired, this could be replaced by a likelihood ratio test (see e.g. Stuart *et al.* 1999) to test the overall significance of a term in the model. Typically though, *F*-tests are not required for variety trials, as most varieties will be similar and only unusual ones will be of interest (see e.g. Mead *et al.* 1993). Hence, even if the *F*-test is nonsignificant, pairwise *t*-tests can still be of interest.

Alternative diagnostics would be useful to assess the goodness-of-fit of these models. Zucchini (2002) suggests using plots akin to Fig. 3(*d*), but using *segments* instead of points to represent the censored data. However the data considered there were either discrete or interval-censored, whereas the example in the present paper has left-censored data and so segments would become half-lines here. Also, there are many more datapoints in the present work and so if all segments were plotted together the plot would be too cluttered. Ideas from the literature for regression using censored data have also been considered, e.g. Buckley & James (1979). In this context, Smith & Zhang (1995) suggest the use of *renovated scatterplots*, again representing censored points using line segments. However this whole approach is also unsuitable in the present work, as parameter estimates will only be unbiased if censoring is independent of the response; here this is certainly not the case. Nevertheless, despite the lack of availability of a sophisticated method for assessing model fit, all of the investigations carried out in the present paper suggest that the model is valid.

We thank the Scottish Executive Environment and Rural Affairs Department for financial support. We also thank Tony Hunter at BioSS and Jane Ryall at NIAB for highlighting this problem and Ann Cowling and a referee for bringing many useful references to our attention.

REFERENCES

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association* **50**, 901–908.

Allcroft, D. J. & Glasbey, C. A. (2002). A spectral estimator of ARMA parameters from thresholded data. *Statistics and Computing* **12**, 369–376.

Allcroft, D. J. & Glasbey, C. A. (2003). A latent Gaussian Markov random field model for spatio-temporal rainfall disaggregation. *Applied Statistics* **52**, 487–498.

Amemiya, T. (1984). Tobit models: a survey. *Journal of Econometrics* **24**, 3–61.

Buckley, J. & James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–436.

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press

Cox, D. R. & Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.
Davison, A. C. (1998). Normal scores. In *Encyclopedia of Biostatistics* (Eds P. Armitage & T. Colton), pp. 3067–3069. Chichester: John Wiley & Sons.
Ellis, T. M. R., Philips, I. R. & Lahey, T. M. (1994). *Fortran 90 Programming*. Harlow: Addison-Wesley.
Feuerverger, A. (1979). On some methods of analysis for weather experiments. *Biometrika* **66**, 655–658.
GenStat (2000). *GenStat for Windows. Release 4.2. Fifth Edition*. Oxford: VSN International Ltd.
Greene, W. H. (1986). *LIMDEP User's Manual*. New York: New York University.
IMSL (2002). *IMSL. Math/Library, Stat/Library*. Houston, Texas: Visual Numerics, Inc.
Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1, 2nd edition. New York: Wiley Interscience.
Johnston, J. & DiNardo, J. (1997). *Econometric Methods*, 4th edition. New York: McGraw-Hill.
Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.
Mead, R., Curnow, R. N. & Hasted, A. M. (1993). *Statistical Methods in Agriculture and Experimental Biology*, 2nd edition. London: Chapman & Hall.
Pennington, M. (1983). Efficient estimators of abundance, for fish and plankton surveys. *Biometrics* **39**, 281–286.
SAS Institute (2000). *SAS/STAT User's Guide, Version 8*. Cary, NC: SAS Publishing.
Smith, P. J. & Zhang, J. (1995). Renovated scatterplots for censored data. *Biometrika* **82**, 447–452.
Stuart, A., Ord, J. K. & Arnold, S. (1999). *Kendall's Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model*, 6th edition. London: Arnold.
Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26**, 24–36.
Welsh, A. H., Cunningham, R. B., Donnelly, C. F. & Lindenmayer, D. B. (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling* **88**, 297–308.
Zucchini, W. (2002). Generalized residuals and their applications. In *IBC2002. XXIst International Biometric Conference 2002. Proceedings – Manuscripts of Invited Paper Presentations*, pp. 403–423. Freiburg, Germany: International Biometric Society. (ISSN-1606-8653.)

APPENDIX

In order to produce the plots of Fig. 2, it is necessary to estimate the threshold level that divides the Gaussian distribution into the lower part corresponding to the zero observations and the upper part corresponding to the transformed non-zero observations. The variance of the distribution must also be estimated, thus allowing the scaling of the distribution to a standard normal.

The threshold y_T of the underlying standard Gaussian distribution is simply estimated as the value which divides the distribution such that the proportion of it below the threshold matches the proportion of zeros in the data, i.e.

$$y_T = \Phi^{-1}\left(\frac{n_0}{n}\right)$$

where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard normal distribution, given by Eqn 4.

To estimate the variance σ^2 we can minimize the sum of squares of observed and expected normal scores for the non-zero data values. Observed normal scores may be calculated from the ranked data, $y_{(i)}$, $i=1, \ldots, n$, as

$$\Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$$

(see e.g. Davison 1998) and the expected values of these normal scores are given by

$$y_T + \frac{1}{\tilde{\sigma}} y_{(i)}$$

Figure 2(*b*) has the order statistics $y_{(i)}$ on the x-axis and normal scores on the y-axis. The observed normal scores are plotted as points and the expected as the continuous line with slope $1/\tilde{\sigma}$.

It should be stressed that the estimation of y_T and σ^2 in this way is only required for production of the plots shown and not for the subsequent model estimation.

https://doi.org/10.1017/S0021859603003332 Published online by Cambridge University Press