REGRESSION WITH CORRELATED ERRORS

C.A. GLASBEY

SYSTEMATIC RESIDUALS

When data exhibit systematic departures from a fitted regression line (see for example Figs 1 and 2), either the regression function is inappropriate, or the errors are correlated, or both. In most cases it is assumed that the function is deficient, and it is changed. But there are situations where the assumption of independent errors is not wholly plausible. For example, some sources of error will persist over several observations when repeated measurements are made on a single experimental unit.

Systematic departures may be modelled either by another regression function, or by correlated errors. To illustrate, consider

$$y_i = a + bx_i + c \sin x_i + e_i \qquad i=1, \ldots, n,$$

where a, b, $x_1, \ldots, x_n$ are constants, c is normally distributed with mean 0, variance τ^2, and $e_1, \ldots, e_n$ are independently normally distributed with means 0, variances σ^2. Two models are equally valid for the y's, either they are independently normally distributed with means $a+bx_i + c\sin x_i$, variances σ^2, or they are correlated, with means $a+bx_i$, variances $\tau^2 \sin^2 x_i + \sigma^2$ and covariances $\tau^2 \sin x_i \sin x_j$ between y_i and y_j.

In general the modelling objective determines the choice: for a simple summary it may be preferable for the regression function to explain all systematic variability, whereas a correlated stochastic component may be of more assistance in understanding the data generating mechanism. A succinct summary of data is often achieved by using the regression function to describe the long-term trends and the correlations the short-term fluctuations. This will be the assumed case from now on.

CORRELATED ERRORS

In the presence of correlated errors, ordinary least squares regression parameter estimators remain unbiased, but they may be inefficient, and the conventional estimators of the variances of

these estimators are usually biased. For example, consider

$$y_i = a + b\,(i - 5.5) + e_i \qquad i=1, \ldots, 10,$$

where $e_1, \ldots, e_{10}$ are normally distributed with means 0, variances 1, and the correlation between e_i and e_j is $r^{|i-j|}$. The least squares estimators of a and b are

$$\hat{a} = 0.1\,(y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_{10})$$

$$\hat{b} = 0.055\,(y_{10} - y_1) + 0.042\,(y_9 - y_2) + 0.030\,(y_8 - y_3) + 0.018\,(y_7 - y_4) + 0.006\,(y_6 - y_5)$$

If r = 0.9 then

$$\text{var}\,(\hat{a}) = 0.73 \qquad \text{var}\,(\hat{b}) = 0.017$$

The generalised least squares estimators, that is those with minimum variance, are

$$\tilde{a} = 0.36\,(y_1 + y_{10}) + 0.04\,(y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9)$$

$$\tilde{b} = 0.107\,(y_{10} - y_1) + 0.003\,(y_9 - y_2) + 0.002\,(y_8 - y_3) + 0.001\,(y_7 - y_4) + 0.0004\,(y_6 - y_5)$$

and

$$\text{var}\,(\tilde{a}) = 0.68 \qquad \text{var}(\tilde{b}) = 0.015$$

Therefore $\hat{a}$ is 93% efficient and $\hat{b}$ is 88% efficient. However the least squares estimators of the variances, based on the assumptions of independent e's, have expectations

$$E\,(\hat{\text{var}}\,(\hat{a})) = 0.016 \qquad E\,(\hat{\text{var}}\,(\hat{b})) = 0.0020$$

and are therefore very biased. Results for other values of r are given below

r	1000 x var ($\hat{a}$)	1000 x var ($\tilde{a}$)	efficiency %	1000 x E (v$\hat{a}$r ($\hat{a}$))	bias %
-0.9	9	6	68	121	1318
-0.6	30	27	91	116	292
-0.3	57	56	98	110	91
0.0	100	100	100	100	0
0.3	173	171	99	85	-51
0.6	325	308	95	59	-82
0.9	728	679	93	16	-98

r	10000 x var ($\hat{b}$)	10000 x var ($\tilde{b}$)	efficiency %	10000 x E (v$\hat{a}$r ($\hat{b}$))	bias %
-0.9	31	9	28	146	370
-0.6	48	38	81	141	196
-0.3	78	75	96	133	70
0.0	121	121	100	121	0
0.3	182	176	97	103	-44
0.6	250	226	90	71	-72
0.9	172	151	88	20	-88

These results show that least squares estimators are not too inefficient provided that r is positive, but that estimated variances can be severely biased, typically downwards for positive r.

The simplest solution, in general, is to discard the biased standard errors. This approach is most useful when no estimate of precision is required, for example when data are available from independent units and within-unit variability is of little importance (Rowell and Walters, 1976). Alternatively, if it can be assumed that the errors arose from a particular stochastic model, any parameters can be estimated jointly with the regression ones by maximising the likelihood. Models may be either empirical or mechanistic. The mechanistic approach requires knowledge of the processes by which the data were generated, whereas the empirical method is purely data-based. Seber and Wild (1989) discuss and review a wide range of possible models.

EMPIRICAL MODELS

To adopt an empirical approach, gross assumptions have to be made about the correlation structure. Otherwise there are

far too many possible models for one to be identified from a single set of observations. For serially-structured data it may be reasonable to assume that the error process is stationary, in which case the correlations between errors depend solely on the time separation between them. Further, it may be possible to assume that the errors are a sample record from an autoregressive-moving average process of low order (Gallant and Goebel, 1976).

However, maximum likelihood estimators are not as efficient as generalised least squares would suggest, because covariance parameters have to be estimated. For example, in the previously considered first-order autoregressive case, the maximum likelihood estimator of r can be approximated by

$$\tilde{r} = 10 \Sigma \hat{e}_i \hat{e}_{i+1} / 9 \Sigma \hat{e}_i^2,$$

and an approximately unbiased estimate of the error variances is given by

$$\tilde{\sigma}^2 = \Sigma \hat{e}_i^2/7,$$

where $\hat{e}_1, \ldots, \hat{e}_{10}$ are the regression residuals. Commencing from the least squares estimate $\hat{\beta}$, r and σ^2 are estimated, and used to re-estimate β by generalised least squares. Then r and σ^2 are re-estimated, and so on until convergence. From 1000 simulations with r = 0.9,

$$\text{var } (\tilde{a}) = 0.73,$$

which is no less than the variance of the least squares estimator. Further, standard errors are biased:

$$E \text{ (v}\tilde{a}\text{r } (\tilde{a})) = 0.08.$$

This is mainly a consequence of r being underestimated. The average value from the simulations was 0.2.

Residual maximum likelihood estimation (Cooper and Thompson, 1977) is more complicated, but does reduce the bias in $\tilde{r}$, the average result being 0.6. Also, r was sometimes estimated to be 1, which implied that a could not be estimated with any accuracy at all. However, the slope parameter was still estimable, and

resulted in

$$\text{var } (\tilde{b}) = 0.016, \qquad E\ (\tilde{\text{var}}\ (\tilde{b})) = 0.012.$$

Little improvement in efficiency has been gained over least squares, but standard errors are of the right magnitude. An alternative, is to simply modify the least squares standard errors, based on

$$\text{var } (\hat{b}) = \Sigma\ \Sigma\ w_i\ w_j\ \text{cov } (y_i, y_j),$$

where $\hat{b} = \Sigma\ w_i\ y_i$. Therefore

$$\hat{\text{var}}\ (\hat{b}) = \Sigma\ \Sigma\ w_i\ w_j\ \tilde{\sigma}^2\ \tilde{r}^{|i-j|}.$$

From the simulations

$$E\ (\hat{\text{var}}\ (\hat{b})) = 0.013.$$

If the assumed error model is incorrect, then least squares can be more efficient than supposed maximum likelihood estimation. For example, if the above case is modified so that y_1 and y_{10} are independent of the remaining eight data values, but the first-order autoregressive model is still assumed to hold, then

$$\text{var } (\hat{b}) = 0.012 \qquad \text{var } (\tilde{b}) = 0.021$$

MECHANISTIC MODELS

The mechanisms by which errors are correlated are sometimes known precisely; for instance, when data are obtained by applying algebraic operations such as summing, averaging or differencing to independent observations. On other occasions the covariances are known to within a few parameters. Many non-linear regression equations have some justification in terms of underlying deterministic models such as differential equations (Sandland and McGilchrist, 1979) or compartment systems (Matis and Wehrly, 1979). By making these models stochastic, regression functions and error processes can be generated with shared parameters. If a stochastic model is appropriate, then, in fitting it to data,

the most efficient parameter estimators are obtained, because the regression curve is fitted efficiently, and also because extra information on the parameter values is recovered from the error covariances.

Glasbey (1988) used two data sets to illustrate the methods and problems encountered. Drug-induced currents in ion-channels, as in Fig 1, were satisfactorily represented by a stochastic compartment system. First-order linear stochastic difference equations were used to model milk yield of cows, as for example in Fig 2. In this case, the proposed model did not fit adequately, and the results it produced were potentially very misleading.

In general, for mechanistic error models it is recommended that (i) an independent observation error is included, (ii) the goodness-of-fit is tested by refitting the model with separate parameters in the regression and error components, and (iii) a stochastic model is not used unless it has a sound scientific basis. It should not be assumed that a particular stochastic model is appropriate simply because its deterministic counterpart fits well. Sources of error are usually many and varied and it is safer to model them separately from the regression model.

CONCLUSIONS

Maximum likelihood estimators which are based on the wrong error variance model may be even less efficient, and the estimated standard errors even more biased, than the ordinary least squares ones (Engle, 1974). Moreover, regression parameters can have a different interpretation when errors are modelled by a correlated process. The conjunction of regression model and error model describe a data set, so a change in the error model forces a compensatory change in the regression model. For example, the fitted regression will make larger systematic departures from the data if errors are assumed to be highly correlated than if they are assumed to be independent. Therefore, although the estimation of regression parameters with almost any choice of variance matrix has become computationally easy, it is beset with statistical difficulties.

Fig 1

Current through end-plate membrane of muscle fibre after a voltage jump, and least squares fit of exponential curve.

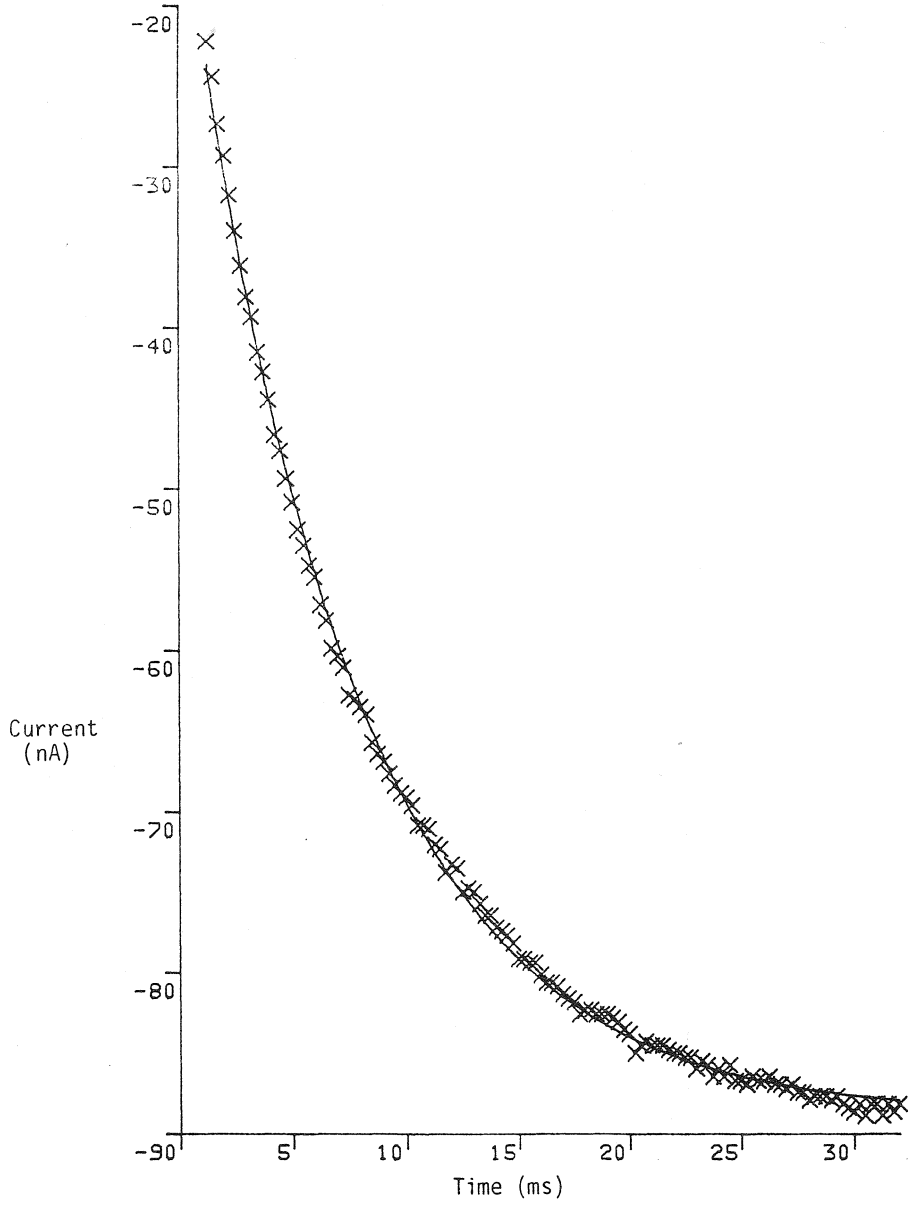
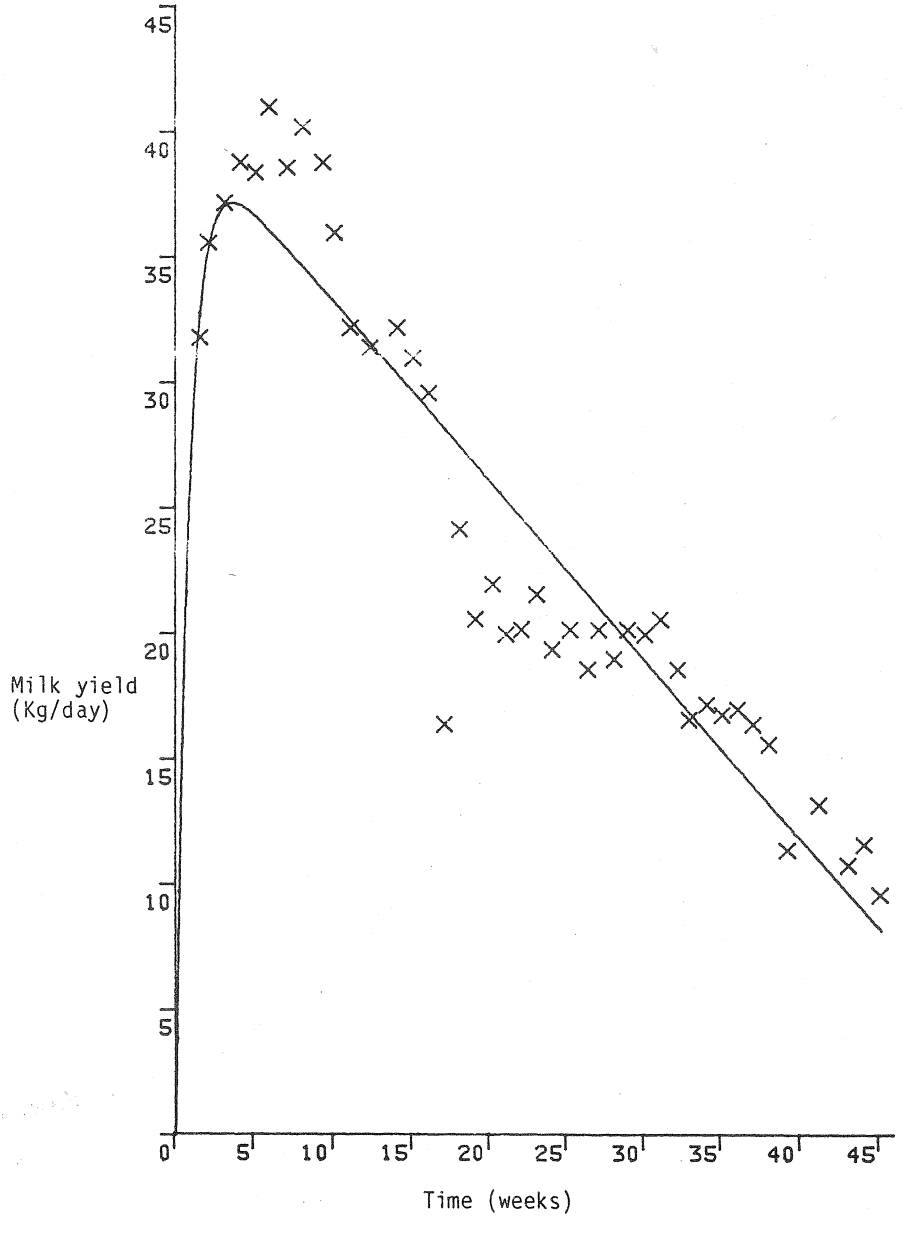


Fig 2

Daily milk yields of a cow at weekly intervals, and least squares fit of a lactation curve (exponential plus linear trend).



REFERENCES

Cooper, D.M. & Thompson, R. (1977) A note on the estimation of the autoregressive-moving average process, Biometrika, 64, 625-628.

Engle, R.F. (1974) Specification of the disturbance for efficient estimation, Econometrica, 42, 135-146.

Gallant, A.R. & Goebel, J.J. (1976) Nonlinear regression with autocorrelated errors, Journal of the American Statistical Association, 71, 961-967.

Glasbey, CA (1988) Examples of regression with serially correlated errors. The Statistician, 37, 277-291.

Matis, J.H. & Wehrly, T.E. (1979) Stochastic models of compartmental systems, Biometrics, 35, 199-220.

Rowell, J.G. & Walters, D.E. (1976) Analysing data with repeated observations on each experimental unit, Journal of Agricultural Science, Cambridge, 87, 423-432.

Sandland, R.L. & McGilchrist, C.A. (1979) Stochastic growth curve analysis, Biometrics, 35, 255-271.

Seber, G.A.F. & Wild, C.J. (1989) Nonlinear Regression, New York, Wiley, chapers 6-8.

Scottish Agricultural Statistics Service
JCMB, The King's Buildings
Edinburgh EH9 3JZ
Scotland