Journal of Animal Ecology 2002
71, 542–545

FORUM

On the misuse of residuals in ecology: regression of residuals vs. multiple regression

ROBERT P. FRECKLETON
Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Summary

1. Residuals from linear regressions are used frequently in statistical analysis, often for the purpose of controlling for unwanted effects in multivariable datasets. This paper criticizes the practice, building upon recent critiques.
2. Regression of residuals is often used as an alternative to multiple regression, often with the aim of controlling for confounding variables. When correlations exist between independent variables, as is generally the case with ecological datasets, this procedure leads to biased parameter estimates. Standard multiple regression, by contrast, yields unbiased parameter estimates.
3. In multiple regression parameters are estimated controlling for the effects of the other variables in the model, and thus multiple regression achieves what residual regression claims to do.
4. Several measures of correlation exist that differ in the way that variance is partitioned among independent variables. These can be estimated multiply, or sequentially if reasons exist for estimating effects of variables in a hierarchical manner.

Key-words: general linear model, parameter bias, regression analysis.

Journal of Animal Ecology (2002) **71**, 542–545

Introduction

In analysing multivariable datasets it is common that in looking at the effect of some variable (x_1) on a dependent variable of interest (y), the effects of a third continuous variable (x_2) are to be controlled for, for instance because its effects may confound those of x_1. In such circumstances it has become common to perform a regression of y on x_2 and use the residuals from this regression in testing for the effects of x_1.

A recent article by García-Berthou (2001) pointed out that this is an inappropriate analysis in the case where x_1 is a categorical variable, and where the residuals from the regression of y on x_2 are subject to a t-test or an ANOVA to test for differences between the groups defined by x_1. As pointed out by García-Berthou (2001), the correct analysis is in fact an ANCOVA, or other general linear model (GLM) where the factorial and regression variables are included simultaneously. Although García-Berthou (2001) pointed out one analytical procedure in which residuals from regression are treated as data in subsequent analysis, the use of residuals as data is common in a range of analyses, particularly when the confounding variable, x_1, is continuous. This use of residuals as data is, for example, particularly common in controlling for the effects of body size in multivariable analyses.

In this paper it is argued that the practice of treating residuals from regression as if they are data is unjustified except in specialized circumstances. This is because in ecological data it is common to find that independent variables are correlated, and such correlations lead to biased parameter estimates or significance tests. This bias arises because, except in the case of fully balanced designs, the marginal (effect on y of changing x ignoring other covariates) and conditional (effect on y of changing x given other covariates) estimates of parameters are not the same. A similar point has been made by Darlington & Smulders (2001) in the context of analysing behavioural data. In their paper Darlington & Smulders concentrate the consequences of using residuals as data for hypothesis testing (i.e. rates of Type I and II errors). Below a different perspective is taken. It is argued that the estimation of effects in multiple regression is best viewed as consisting of two components:

Correspondence: Robert P. Freckleton, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. E-mail: robert.freckleton@zoology.oxford.ac.uk

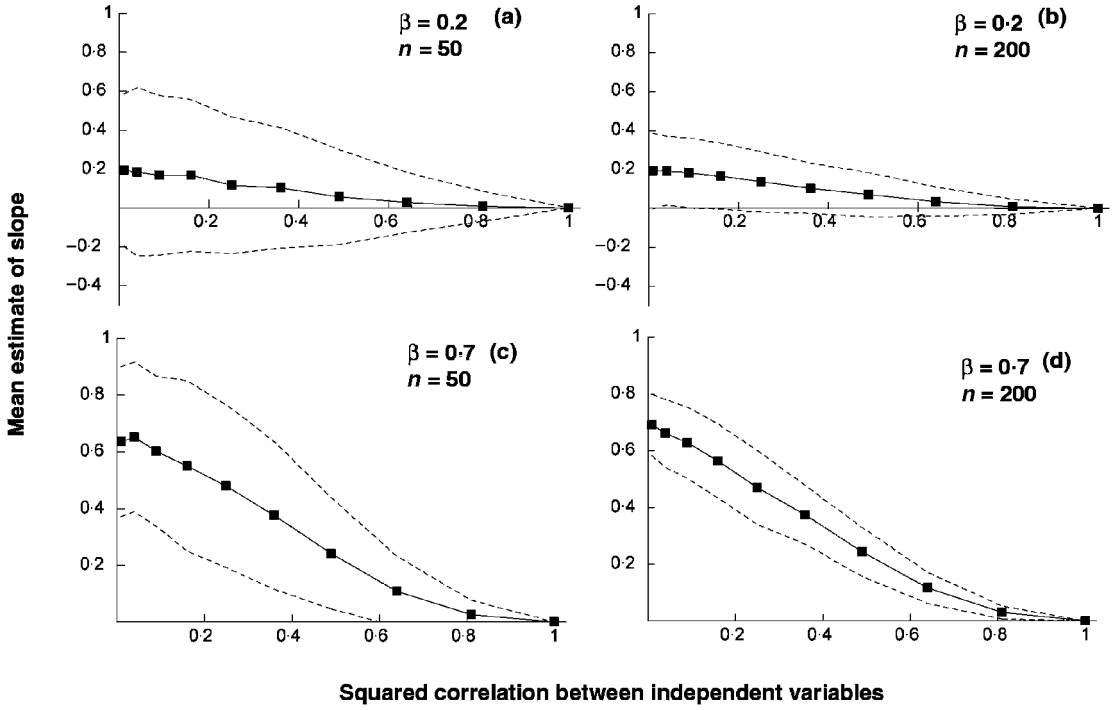

Fig. 1. Estimates of slope of simple linear relationships using residual regression (see main text for details of this). Data on a variable y were simulated according to the equation using $i = 1 \ldots n$ observations, where $n = 50$ (*a*, *c*) or $n = 200$ (*b*, *d*), and where β was set to either 0·2 (*a*, *b*) or 0·7 (*c*, *d*). ε was a normally distributed random error term. The degree of collinearity of the independent variables was varied by varying the squared correlation between x_1 and x_2, as shown. 1000 replicates were performed at each parameter combination. The dashed lines show the 95% confidence intervals for 1000 estimates of the slope relating y to x_2 using regression of the residuals from the relationship between y and x_1 on x_2.

(i) generating unbiased estimates of the parameters (i.e. slopes and intercept) for the data; and (ii) measuring how much variance is explained by each variable, and how much of this is independent of the other variables. It is highlighted that the second component of the analysis may be approached in several ways, depending on the question in hand, but that residual regression yields biased parameter estimates.

Residuals as data generate bias

In this section a simple simulation is used to demonstrate how the use of regression residuals as data in subsequent analyses leads to biased parameter estimates when correlation exists among independent variables. Data were simulated according to the following model:

$$y(i) = a + \beta_1 x_1(i) + \beta_2 x_2(i) + \varepsilon(i) \qquad \text{eqn 1}$$

Thus each of the $i = 1 \ldots n$ observations on variable y was generated from two independent variables x_1 and x_2, as well as a random error term ε. The relationship is a simple linear association, characterized by an intercept a (set to zero for simplicity) and slopes, β_1 and β_2, describing the effect of each of the independent variables. These slopes were assumed to be the same for both x_1 and x_2, hence their effects on y were identical. The error term was normally distributed with zero mean.

Figures 1 and 2 contrast two strategies for the analysis of the simulated data. In Fig. 1 the residual regression technique is employed, whereby regression of y on x_1 is performed first, then the residuals from this regression are regressed on x_2. In Fig. 2 standard least squares multiple regression (e.g. Sokal & Rohlf 1995) is employed, i.e. the effects of x_1 and x_2 are analysed simultaneously. Both Figs 1 and 2 show the estimate of the slope (β*) for the effect of x_2 on y when the actual value of β is 0·2 or 0·7 and for small ($n = 20$) and large ($n = 200$) datasets as the correlation between x_1 and x_2 is varied.

In both Figs 1 and 2 β*, the estimate of the true slope β, should be constant and unaffected by changing the correlation between the independent variables. It is clear, however, that the residual regression technique underestimates the effect of x_2 and that this bias can be large even from moderate correlations. Indeed, even when the r^2 for the association between x_1 and x_2 is only 0·2, the 95% confidence intervals do not include the true value of β, when β = 0·7, for the large ($n = 200$) dataset. By contrast the estimate of the true slope generated by least-squares multiple regression is unbiased and unaffected by the correlation between the independent variables. Only the sampling variance is affected, which becomes large when the correlation is very high, as is usual with multicollinearity (Tabachnick & Fidell 2000). Note that standard regression diagnostics such as variance inflation

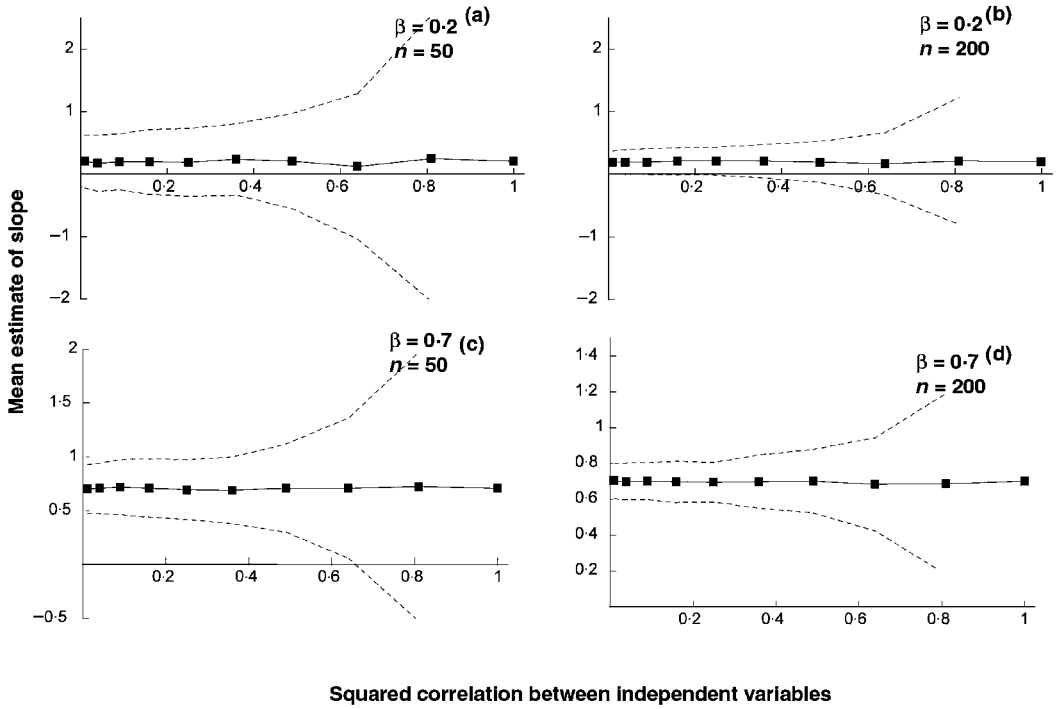

Fig. 2. As Fig. 1, but where the estimates of the effect of x_2 on y are derived through standard least-squares multiple regression.

factors (VIFs) would warn of an inflated variance resulting from high correlation between x_1 and x_2. By contrast the residual regression technique treats both x_1 and x_2 as independent and this inflation of the sampling variance would be missed.

The reason for the bias in Fig. 1 is that the effects of x_1 and x_2 are correlated and by removing the effect of x_1 only the effect that results from x_2 and is uncorrelated with x_1 remains. As the correlation between x_1 and x_2 increases, the unique component owing to x_2 is an increasingly small component of the total variance explained by x_1 and x_2 and the effect of x_2 is thus underestimated (and the effect of x_1 thus overestimated). Note that to estimate the true slope for the effect of x_2 using residual regression one would need to regress the residuals of the regression on y on x_1 on the residuals of the regression of x_2 on x_1 (e.g. see Baltagi 1999, pp. 72–74 for elaboration of this).

In summary, therefore, residual regression is a poor substitute for multiple regression since the parameters estimated from residual regression are highly biased, with this bias increasing with the correlations between the independent variables in the model.

Controlling for unwanted effects?

Note that in equation 1 no assumption was made about the order in which the effects of x_1 and x_2 occur. Thus the effects of x_1 or x_2 could occur in tandem or sequentially. What is important is the error structure of the model. In line with standard regression assumptions it is assumed that the variance of y is a simple additive function of the effects of the independent variables plus the error. Although, mechanistically, the effects of the x-variables may operate sequentially (e.g. the effect of x_1 may occur at one period in the life-cycle, those of x_2 later on) this does not affect the structure of the model. Given this structure the least-squares multiple regression provides the best linear unbiased estimates of the parameters of equation 1, e.g. see Grabill 1976 for a mathematical exposition of this point), whereas the residual regression provides biased estimates. It cannot therefore be argued that the residual regression controls for unwanted effects in estimating the parameters of equation 1. Conversely, if the idea that x_1 confounds the estimate of the effect of x_2 on y was incorrect, then residual regression technique would nevertheless yield a high estimate of the effect of x_1 on y, owing to the correlation between x_1 and x_2, and would thus underestimate the effect of x_2. In fact ordinary least-squares estimates the slope of the relationship between y and each x controlling for all other x variables in the model. Multiple regression thus actually achieves what residual regression claims to do.

Estimating variance components

When performing regression analysis using intercorrelated independent variables, the question will naturally arise, how much variation does each variable explain both in total and independently of each other? This is a separate issue from that of generating the best unbiased parameter estimates.

The variance in the dependent variable y may be thought of as having three components (e.g. see Fig. 5.3 in Tabachnick & Fidell 2000). v_1 and v_2 represent the variance explained by x_1 and x_2 independent of each other, respectively, while v_{12} represents the variance explained by both x_1 and x_2, i.e. the common variance

that these variables explain because of the correlation between them. v_r is the residual variance in y, i.e. that not explained by either x_1 or x_2. Three measures of association exist that vary in the way that these variances are partitioned. Squared correlation (r^2) measures the total explained by each variable relative to the total variance in y (e.g. for x_1, $r^2 = (v_1 + v_2)/(v_1 + v_2 + v_{12} + v_r)$). Semi-partial correlation (sr^2) measures the unique contribution of each variable (e.g. for x_1, $sr^2 = v_1/(v_1 + v_2 + v_{12} + v_r)$). Thirdly, partial correlation (pr^2) measures the contribution of each variable after all other variables have been accounted for (e.g. for x_1, $pr^2 = v_1/(v_2 + v_r)$), and the denominator is the total variance in y minus the effect of the other variable (v_2) and minus the variance in y common to both variables (v_{12}).

It is also important to note that variance can be estimated sequentally (as in Type III sums of squares) as well as adjusting for other terms in the model (Type I sums of squares) and correlations can be constructed based on a sequential partitioning of variance (Tabachnick & Fidell 2000). Thus for a given dataset the choice of coefficient will depend on the question being asked, the interrelationships between the independent variables as well as what, if anything, is known about the structure of the system. It is worth reiterating that in all cases the parameter estimates are the same, but would be biased in the case of residual regression.

Concluding remarks

Perhaps the only justification for treating residuals as data is in *post-hoc* diagnosis of fitted regression models. For instance, if a model is fitted to a series of observations on variables collected over time, the residuals from the regression could be regressed on the time of observation to check that the assumption that the residuals are independent of time is upheld. This would not preclude a correlation of the observed dependent variable with time, since one of the independent variables may correlate with time. Such analysis of residuals is simply a diagnostic check on model adequacy in the light of the assumptions and is not a rigorous testing or estimation procedure. Furthermore, the residual regression is unsuitable as method for model selection since degrees of freedom are usually not allocated appropriately (above, Darlington & Smulders 2001) and because the significance of variables will be extremely highly sensitive to the order in which they are entered.

For most applications the technique of residual regression is redundant and does not do what it claims to do. The claim for applying the technique is that the underlying sequence of effects of the independent variables is known (e.g. the effects of one variable take precedence over another). Even if this is the case, standard least squares regression should provide unbiased parameter estimates. Instead in such circumstances the problem should be viewed as one of working out how much variance each variable explains in isolation and in total. Moreover if situations exist in which either a hierarchical model is justified, or in which the structure of the relationship between the independent variables is known then techniques such as hierarchical regression and structural equation modelling exist to fit models under that account for such relationships (Shipley 2000). However, the usual application of regression analysis in ecology is to determine *whether* relationships between variables exits and how much variation these relationships explain. In such circumstances the standard least squares regression provides the best parameter estimates, and semipartial, partial and multiple correlation allow the variance explained by different variables to be clearly measured and dissociated.

Acknowledgements

I should like to thank Nick Dulvy, Phil Stephens and Andrew Watkinson for their comments on an earlier version of this MS and Emili García-Berthou and David Elstow for suggestions for improvement. This work was funded by the NERC (grant no. GR3/12939 to Paul Harvey and Mark Pagel).

References

Baltagi, B.H. (1999) *Econometrics*. Springer, Berlin.

Darlington, R.B. & Smulders, T.V. (2001) Problems with residual analysis. *Animal Behaviour*, **62**, 599–602.

García-Berthou, E. (2001) On the misuse of residuals in ecology: testing regression residuals vs. the analysis of covariance. *Journal of Animal Ecology*, **70**, 708–711.

Graybill, F.A. (1976) *Theory and Application of the Linear Model*. Duxbury Press, Boston, MA.

Shipley, B. (2000) *Cause and Correlation in Biology: a User's Manual to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press, Cambridge.

Sokal, R.R. & Rohlf, F.J. (1995) *Biometry*. W.H. Freeman, New York.

Tabachnick, B.G. & Fidell, L.S. (2000) *Using Multivariate Statistics*. Harper Collins, New York.

Received 11 September 2001; revision received 15 January 2002