Web Ecology 8: 35–46.

Plotting partial correlation and regression in ecological studies

Jordi Moya-Laraño and Guadalupe Corcobado

Moya-Laraño, J. and Corcobado, G. 2008. Plotting partial correlation and regression in ecological studies. – Web Ecol. 8: 35–46.

Multiple regression, the General linear model (GLM) and the Generalized linear model (GLZ) are widely used in ecology. The widespread use of graphs that include fitted regression lines to document patterns in simple linear regression can be easily extended to these multivariate techniques in plots that show the partial relationship of the dependent variable with each independent variable. However, the latter procedure is not nearly as widely used in ecological studies. In fact, a brief review of the recent ecological literature showed that in ca 20% of the papers the results of multiple regression are displayed by plotting the dependent variable against the raw values of the independent variable. This latter procedure may be misleading because the value of the partial slope may change in magnitude and even in sign relative to the slope obtained in simple least-squares regression. Plots of partial relationships should be used in these situations. Using numerical simulations and real data we show how displaying plots of partial relationships may also be useful for: 1) visualizing the true scatter of points around the partial regression line, and 2) identifying influential observations and non-linear patterns more efficiently than using plots of residuals vs fitted values. With the aim to help in the assessment of data quality, we show how partial residual plots (residuals from overall model + predicted values from the explanatory variable vs the explanatory variable) should only be used in restricted situations, and how partial regression plots (residuals of Y on the remaining explanatory variables vs residuals of the target explanatory variable on the remaining explanatory variables) should be the ones displayed in publications because they accurately reflect the scatter of partial correlations. Similarly, these partial plots can be applied to visualize the effect of continuous variables in GLM and GLZ for normal distributions and identity link functions.

J. Moya-Laraño (jordi@eeza.csic.es) and G. Corcobado, Depto de Ecología Funcional y Evolutiva, Estación Experimental de Zonas Áridas, CSIC, General Segura, 1, Almería, ES–04001 Almería, Spain.

The use of multiple regression is particularly advisable when ecological questions involve several explanatory (independent) variables because usually the experimental manipulation of that many factors is logistically unfeasible (James and McCulloch 1990, Graham 2003). General linear models (GLM) and Generalized linear models (GLZ) with normally distributed errors, are extensions of multiple regression and ANOVA that include both continuous and categorical (e.g. treatment factors) explanatory (independent) variables (Littell et al. 1996, Agresti 2002, Quinn and Keough 2002). These latter methods are increasingly used in ecological studies. However, Ordinary least squares (OLS) multiple regression alone is still widely used in ecology (Philippi 1993) as demonstrated by our following survey of the literature. We searched for the use of multiple regression in the ecological literature. We consulted the Journal Citation Reports (JCR) in the ISI Web of Knowledge and chose the 20 journals with the highest impact factors within the ecology subject category. We used the following search engines: Ecological Society of America, Blackwell–Synergy, The Univ. of Chicago Press, Oxford Univ. Press and Allen Press. Some journals could

Accepted 12 May 2008

ISSN 1399-1183

not be used because either the search engines were not able to do a search within the full text in their articles (i.e. Wildlife Monographs, Ecosystems, Perspectives in Plant Ecology and Oecologia) or because they were review journals (i.e. Trends in Ecology and Evolution and Annual Review of Ecology, Evolution and Systematics). Thus, we skipped these journals and went to the journal immediately below in the JCR list until reaching a total of 20 journals. The included journals were: Ecology Letters, Ecological Monographs, Journal of Applied Ecology, Ecology, American Naturalist, Molecular Ecology, Journal of Ecology, Evolution, Conservation Biology, Global Change Biology, Ecological Applications, Global Ecology and Biogeography, Journal of Animal Ecology, Diversity and Distributions, Journal of Evolutionary Biology, Oikos, Functional Ecology, Behavioral Ecology, Journal of Biogeogrphy and Ecography). We used the year range 1997–2006 (from 14 Nov 2006) or later for those journals that first started after 1997. We also used the Web of Science within the ISI Web of Knowledge to know how many papers each journal had published in that year range. For each journal we did a search of the exact phrase 'multiple regression' within the full text of all articles, and recorded the number of entries. For each journal, we then selected at random five of the papers that included 'multiple regression' and we closely inspected them to verify if they in fact used multiple regression analysis or just mentioned the technique. Our survey and close inspection of 100 random papers was used to estimate the proportion of papers that use multiple regression in the ecological literature. We found 131 970 papers published since 1997 in the 20 ecological journals that were surveyed. We found 11 010 (8.3%) papers with the phrase 'multiple regression' in them. From the close inspection of 100 papers we found that only 74 actually used the technique, which gives an estimate of 6.2% of ecological papers using multiple regression. Two of the papers, however, had used general linear models. A list of the 100 random articles checked can be obtained from the authors upon request. In addition, the increasing easy access to computing time has made GLM, and especially GLZ, to become widely used tools to analyzing both the outcome of experiments as well as for predictive modelling. A survey in the Web of Science within the ISI Web of Knowledge under the Ecology subject category shows that the number of papers including 'Generalized linear models' or 'Generalised linear models' in their abstracts has steadily grown from 5 in 2001 to 43 in 2007. We assume that the number of papers published in Ecology in SCI journals has not increased so substantially as to affect this great difference. Unfortunately the data on how many papers were published in 2007 will be not available until the 2008 report from the Journal Citation Reports is released.

The patterns of multiple regression and GLM are usually documented in publications using summary tables that include the intercept, the regression coefficients and/or the partial correlation coefficients along with tests of significance. This differs substantially from simple linear regression which is very often documented by displaying a scatter plot with the least-squares fitted line (Y versus X). When the distribution of errors in the model (i.e. the unexplained variance) conforms to a normal distribution, plots to document the partial relationship between each of the independent variables and the dependent variables (i.e. a plot of Y against X_1 while holding the remaining explanatory variables $X_{2...}X_k$ constant) are readily available (Larsen and McCleary 1972, Belsey et al. 1980, Velleman and Welsch 1981, Cook and Weisberg 1982, Neter et al. 1996, Montgomery et al. 2001), and can be very useful for several purposes (below). Although it is beyond the scope of the current paper, plots of partial relationships have also been devised for other types of distribution in generalized linear models (Hines and Carter 1993). These plots depict the true regression coefficient within the multiple regression model as the slope of a fitted line. Although, as we show here, not all kinds of these partial plots accurately reflect the right amount of partial correlation, through this paper we will refer to these plots in general as partial plots. This way to document the results of multiple regression has been used in a very minor proportion of ecological studies. In spite of its potential usefulness, a close inspection of the above 74 recent ecological papers that used multiple regression shows that only 1 of them (1.8%) included a partial plot to support the results (Findlay and Bourdages 2000). More importantly, 17.6% of the papers directly used a bi-variate plot of the dependent variable on the independent variable, just as it is usually done in simple linear regression. This latter procedure may be misleading because due to its inherent nature, the partial regression coefficient in multiple regression may change both in magnitude and in sign relative to the coefficient in a simple relationship (below).

The objectives of this paper are: 1) to introduce partial plots to ecologists who may be unfamiliar with them, 2) to highlight what partial plots are useful for, 3) to briefly review the use of partial plots in the ecological literature, 4) to emphasize in which circumstances the use of each kind of partial plot is more appropriate and 5) to simplify the different names given to these plots for future reference in ecological studies. The two main kinds of partial plots are partial regression plots and partial residual plots.

Partial regression plots

Given a multiple regression model such as $Y = b_0 + b_1X_1 + b_2X_2 + e$, where b_x is the intercept and b_1 and b_2 are the regression coefficients, and e denote the residual error, a partial regression plot involving the independent variable X_1 would be a plot of the residuals of the regression of Y on X_2 vs the residuals of the regression of X_1 on X_2. The slope of the least-squares fitted line in this plot matches the

regression coefficient (b_1). This kind of plot truly partials out the effect of X_2 in the relationship and as such, the Pearson correlation coefficient calculated from the point coordinates from this new plot is identical in magnitude to the partial correlation coefficient for the variable X_1 in the multiple regression model (Belsey et al. 1980, Cook and Weisberg 1982). Partial regression plots have also been called added variable plots, adjusted variable plots and partial regression leverage plots (Belsey et al. 1980, Cook and Weisberg 1982, Neter et al. 1996, Montgomery et al. 2001). The name 'added variable plots' denotes the fact that these plots can also be used to see how the overall fit of a model increases by adding a new independent variable (Neter et al. 1996, Montgomery et al. 2001). Partial regression plots are useful to show the slope and the true scatter of points around the partial line in an analogous way to bi-variate plots in simple linear regression and they also serve for graphic regression diagnostics in a more efficient way than the commonly used plots of regression residuals against predicted values (below). The same arguments stated here apply for GLM and GLZ. To obtain a partial regression plot from these other models, one just has to consider that some of the independent variables are categorical instead of continuous. Although the calculation of residuals is a little bit more tedious, most statistical packages (e.g. Statistica, SAS) have the option to calculate residuals from GLM and GLZ. If an identity link function and normal errors has been used for GLZ, then the residuals are identical than for GLM. We must stress, however, that here residuals are only used for graphical purposes, and that for statistical tests, full regression or GLM/Z models must always be used instead of residuals, because the standard errors (SE) of residuals are higher and therefore unnecessarily inflate the type II errors (Darlington and Smulders 2001, Garcia-Berthou 2001, Green 2001, Freckleton 2002, Moya-Laraño et al. in press).

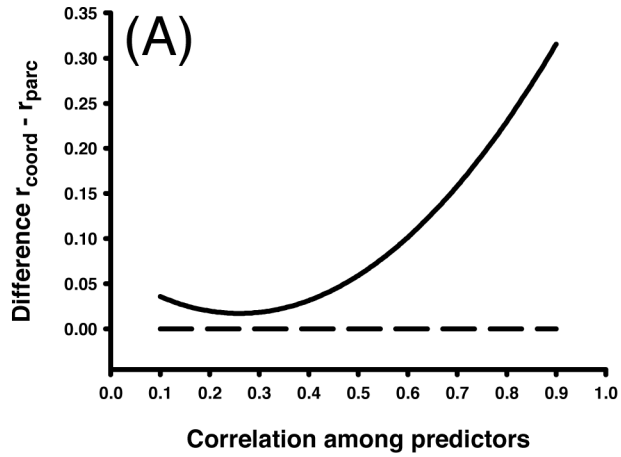
Partial residual plots

Given the above multiple regression model, a partial residual plot involving the independent variable X_1 would be a plot of the residuals (e) from the overall model ($Y = b_0 + b_1X_1 + b_2X_2 + e$) added to the partial predicted Y values from b_1 (i.e. b_1X_1) versus X_1. These plots were originally proposed as a better alternative to plots of residuals versus independent variables for regression diagnostics (Larsen and McCleary 1972). The slope of the least-squares fitted line in these plots also match the regression coefficient (b_1). However, although partial residual plots can be used for the same purposes as partial regression plots, they do not reflect the true partial relationship between the dependent and the independent variables in the regression model. This is because unlike partial regression plots, partial residual plots do not truly take into account the variance explained by the remaining explanatory variables (Cook and Weisberg 1982, p. 51). As a consequence, the scatter of the points around the line is underestimated relative to the scatter of the true partial correlation; the magnitude of the underestimation depends on the strength of the correlation between the independent variables (Cook and Weisberg 1982).

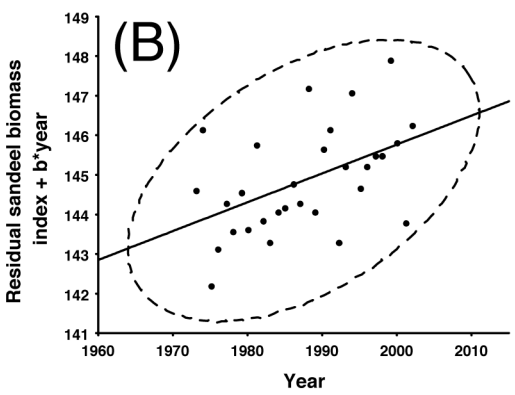
To illustrate this point we simulated regression models with 3 independent normally distributed and standardized ($\mu = 0$, $\sigma = 1$) variables ($Y = \beta_0 + \beta_1X_1 + \beta_2X_2 - \beta_3X_3 + \varepsilon$), in which for simplicity $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = 1$, and introduced different levels of correlations among the explanatory variables (from 0.1 to 0.9 in increments of 0.1; n = 9 models). To introduce the desired amount of correlation among the variables we used conventional procedures (Legendre 2000, Moya-Laraño and Wise 2007, Moya-Laraño et al. in press). We then took a sample of 60 data points for each regression model and calculated partial residual and partial regression plots. We also calculated Pearson correlation coefficients from the coordinates (r_{coord}) of both plots and compared these coefficients with the partial correlation coefficients obtained in the multiple regressions (r_{part}). Figure 1A shows the difference between both values (r_{coord}–r_{part}) for partial residual and partial regression plots for one of the independent variables. While the data displayed by partial regression plots perfectly matches the true correlation coefficients, the data displayed in partial residual plots increasingly overestimates partial correlation with an increase in the amount of correlation among the explanatory variables. We used the same simulated data to illustrate how displaying straight bi-variate plots of dependent variables against independent variables instead of partial plots may be misleading. With the highest correlation among independent variables (0.8 and 0.9), the relationship between Y and X_3, reverses from negative to positive if X_1 and X_2 are not included ($b > 0.67$, $r > 0.45$, $p < 0.0001$ in both cases). Thus, when reporting the results of multiple regression, and especially when the aim is to depict the true amount of partial correlation, partial regression plots should be the only ones included in publications. Partial residual plots give a better idea of how Y changes within the range of X. This is because unlike partial regression plots, partial residual plots maintain the scale for the original X variable in the plot. A good solution if the aim of the graph is to show how Y changes with the true values of X is to add the mean of the original (raw) variable to the residuals of Y and X in the partial regression plot. This will keep the partial correlation of the plot while re-scaling it to the original range of the raw variable.

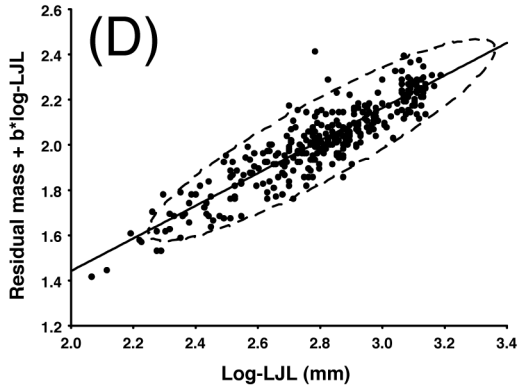
A brief survey of the ecological literature

We searched for the use of partial plots in the ecological literature using the same searching engines, journals and year ranges as above. We looked for the exact phrases: 'par-

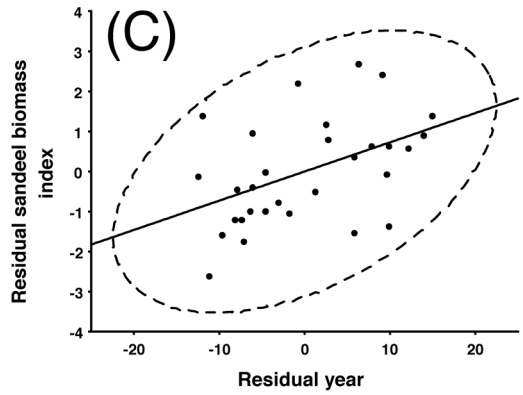


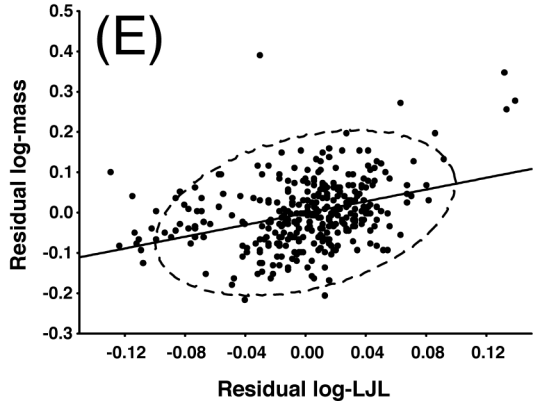
Partial residual plots





Partial regression plots





tial residual plots, partial regression plots, added variable plots, adjusted variable plots and partial regression leverage plots'. A list of the papers found can be obtained from the authors upon request. There was a clear trend towards an increase in the use of partial plots in recent years, since all the papers found were published after 1998. We found nine out of 131 970 papers published in major ecological journals included partial residual plots to support the main core of their results. We do not know, however, whether the authors just wanted to graphically display how Y changed with X or whether they wanted to show the true partial relationship. We also found four papers that correctly used partial regression plots but named them as partial residual plots. Several authors (nine papers) had also correctly used and named partial regression plots. Partial plots have also been used in more complex designs, such as GLMs that included both categorical and continuous variables (Anderson and Jetz 2005, Anderson et al. 2006). We re-calculated the models (either GLMs or multiple regressions) for two of the publications in which they used partial residual plots to display partial relationships (Anderson and Jetz 2005, Frederiksen et al. 2006) and compared the resulting correlation coefficients using the coordinates in the partial residual plots (r_{coord}) against the true partial regression coefficients (r_{part}). In both studies, due to the low correlation among predictors, the deviations of the correlation coefficients in partial residual plots from the true partial correlation coefficients were negligible for the plotted variables (Anderson and Jetz 2005, average difference $r_{coord}–r_{part}$ for three variables = 0.014; Frederiksen et al. 2006, average difference $r_{coord}–r_{part}$ for three variables = 0.012). Figure 1B shows the partial plots for one of the regression analyses as an example (Frederiksen et al. 2006). For more details about both studies and to see the remaining partial plots see Appendix A. Thus, the use of partial residual plots instead of partial regression plots by these authors did not contribute to substantially underestimate the scatter of points around the line.

We obtained data from another study (Irschick et al. 2005) in which the authors correlated morphological traits with performance and habitat use in juvenile and adult male and female green anole lizards (*Anolis carolinensis*). Although the authors were not interested in knowing how different body parts correlated with body mass, here we analyze whether a morphological trait (i.e. lower jaw length, LJL) can positively explain mass independently of an known structural estimate of body size (i.e. snout–vent length, SVL). Thus, we included LJL and SVL as independent variables in a multiple regression explaining body mass. This is a good example of highly correlated data, since the correlation between SVL and LJL was very high ($r = 0.98$, $n = 337$, $p < 0.0001$) and thus they will be very useful to document the differences between partial residual and partial regression plots with highly correlated independent variables. We found that LJL significantly explained body mass independently of SVL ($r_{part} = 0.36$, $p < 0.0001$), indicating that larger mouth parts contribute to body mass independently of SVL. The amount of overestimation of the actual partial correlation coefficient in a partial residual plot was extremely high: $r_{coord}–r_{part} = 0.54$ (compare Fig. 1D, 1E).

Our brief literature review shows a clear trend towards an increase in the use of partial plots in recent years. However, we are still very far from these plots becoming a common practice for documenting partial regressions and correlations of core hypotheses in published papers, such as it is the case for simple linear regressions. In addition of their usefulness in displaying patterns in publications, partial plots can be very useful for regression diagnostics, such as checking for departures from normality, heterogeneity of variances and, more importantly, detecting the existence of outliers and influential data points.

Fig. 1. Two different kinds of partial plots reflect the true correlation coefficient very differently depending on the degree of correlation among predictors. (A) differences in estimates of partial correlations obtained from the coordinates in partial plots of X_3 from the simulated model: $Y = b_0 + b_1X_1 + b_2X_2 – b_3X_3 + \varepsilon$, with increasing amounts of correlation. Dashed line: partial regression plots; solid line: partial residual plots. The values on the x-axis are calculated as $r_{coord}–r_{part}$, where r_{coord} is obtained by calculating the Pearson correlation coefficient from the coordinates in the plot, and r_{part} is the true value of the partial correlation coefficient calculated from the multiple regression model. While the value of partial regression plots is identical to the true partial correlation independently of the amount of correlation between the independent variables (i.e. $r_{coord}–r_{part} = 0$ for all values), partial residual plots increase the overestimation of the true correlation coefficient with an increase in the degree of correlation among explanatory variables. (B) recalculated partial residual plot from Fig. 3d in Frederiksen et al. (2006). Residual sandeel biomass index (an estimate of sandeel larvae biomass, which is a very important food resource for several species of marine birds) was obtained from a regression model including diatom abundance and copepod biomass as independent variables. (C) partial regression plot for the data in Fig. 3d in Frederiksen et al. (2006). Both residual sandeel biomass index and residual year were controlled for diatom abundance and copepod biomass. The range of correlations among independent variables was 0.02–0.60. The remaining partial plots in Fig. 3 for this publication can be found in Appendix A. (D) partial residual plot and (E) partial regression plot of data from Irschick et al. (2005). LJL: lower jaw length. The controlling variable in the regression model was SVL (snout–vent length). Body mass in anole lizards is positively related to the relative size of the jaw independently of structural body size. Note how close the points are in the partial residual plot relative to the partial regression plot. To facilitate the visual assessment of the scatter of points around the line we show the 95% predicting ellipse (dashed line) in all partial plots.

Outliers and influential data points

In simple linear regression, outliers are data points that are far away from the cluster of points that originate the trend. Y-axis outliers are outliers that are far from the linear trend, whereas X-axis outliers are outliers that may fit well on the line but are very far away from the main data cluster (Neter et al. 1996). However, outliers may or may not be influential, in the sense that the overall parameter estimates and significance testing may be just barely affected by their presence. In both simple and multiple regression, there are a few statistics to detect outliers (e.g. hat matrix and studentized residuals) and to measure their influence on the overall model (e.g. Cook's distance, DFFITS and DFBETAS) (Belsey et al. 1980, Neter et al. 1996, Montgomery et al. 2001). However, in simple regression, the fastest way to detect outliers is by displaying a bi-variate scatterplot. Again, both partial plots can also be used to show outliers and influential observations in multiple regression (Larsen and McCleary 1972, Velleman and Welsch 1981, Hines and Carter 1993). Usually plots of residuals versus predicted values are displayed to detect this sort of problem as well as others, such as departures from normality and heteroscedascity of the residuals. We simulated an influential data point in one of the above simulated regression models (correlation among independent variables of 0.9) by adding a fourth (uncorrelated) normal (0,1) variable to the model that would take the arbitrary value of $X_4 = 3$ and Y = 3 for only one of the 60 cases. This newly created variable became wrongly significant in a multiple regression model ($b_4 = 0.30$, $r_{part} = 0.28$; $t_{55} = 2.1$, $p = 0.038$). When we plotted residuals vs predicted values, we observed a fairly homogeneous cloud of points, with no clear outlier showing up (Appendix B). However, using Cook's distance we found a very high value for that particular data point (0.84 versus 0.07 for the immediate highest distance in the data file), which indicates that this point is highly influential to the overall pattern (Neter et al. 1996). A partial regression plot for the variable of interest (Appendix B) is also very efficient for visually detecting this highly influential data point. Thus, had we used only the most classic approach of looking at the residuals versus the predicted values we would have wrongly concluded that there were no influential points, when in fact there was a very important one. Partial regression plots are thus also useful for detecting influential points; with the advantage over available statistics that they give information about which predictor variable is actually being affected by a highly influential point.

Non-linear trends

Because they underestimate the scatter of the data points, partial residual plots may be better than partial regression plots when checking for non-linearity in the relationship between the dependent and the independent variables (Hines and Carter 1993). Non-linearity may indicate the need to add an additional (e.g. quadratic) term in the model or the need to transform the data to make the relationship linear. To show this point, we simulated a regression model including a quadratic term ($Y = \beta_0 + \beta_1 X_1 - \beta_2 X_1^2 + \beta_3 X_2 - \beta_4 X_3 + \varepsilon$, with 0.9 of correlation among variables, $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$) and again randomly took 60 data points (Appendix C) and fitted the linear model $Y = b_0 + b_1 X_1 + b_3 X_2 + b_4 X_3 + e$. We then used partial residual plots and partial regression plots to check for non-linearity, which would indicate whether including a quadratic term in the fit would be appropriate. A partial residual plot (Fig. 2A) emphasizes a true non-linear pattern much more clearly than does a partial regression plot (Fig. 2B). Thus, a partial residual plot is useful for revealing the need to include a quadratic term in the regression model. Indeed, if the quadratic term is not included in the regression model, none of the predictors is significant ($p > 0.05$), but all four predictors (including the quadratic term) become highly significant after including the quadratic term in the model ($p < 0.01$). Usually, the conventional approach of plotting the residuals of the whole model against the fitted values will also be useful for detecting non-linear patterns. However, these latter plots do not provide information about which explanatory variable in the multiple regression shows a non-linear relationship with the dependent variable. Partial residual plots do provide such information and are therefore much more efficient.

To illustrate the usefulness of partial residual plots for documenting non-linear patterns in ecological data, we used a study in which quadratic terms in multiple regressions had been included (Hoffmann and Dodson 2005; < http://www.esapubs.org/archive/ecol/E086/014/default.htm>). In this latter study the authors documented patters of zooplankton species richness in both pristine and developed lakes. For developed lakes they found that the best model (based on AICc) was: zooplankton species richness = $A - Pr + D - D^2$; where A is log-transformed lake area, Pr is log-transformed primary productivity and D is arcsine-square-root transformed proportion of watershed developed. We intentionally fitted the model without the quadratic term (zooplankton species richness = A – Pr + D) and built both the partial residual and the partial regression plots for D. Partial residual plots clearly showed the need to fit a quadratic term (Fig. 2C), whereas partial regression plots failed to show a distinguishable curvilinear pattern (Fig. 2D). Thus, both simulated data and real ecological data showed the superiority of partial residual plots for visually detecting curvilinear patterns in multiple regression.

Conclusions

We have shown how partial plots can be used for the display of patterns in multiple regression as well as for regression

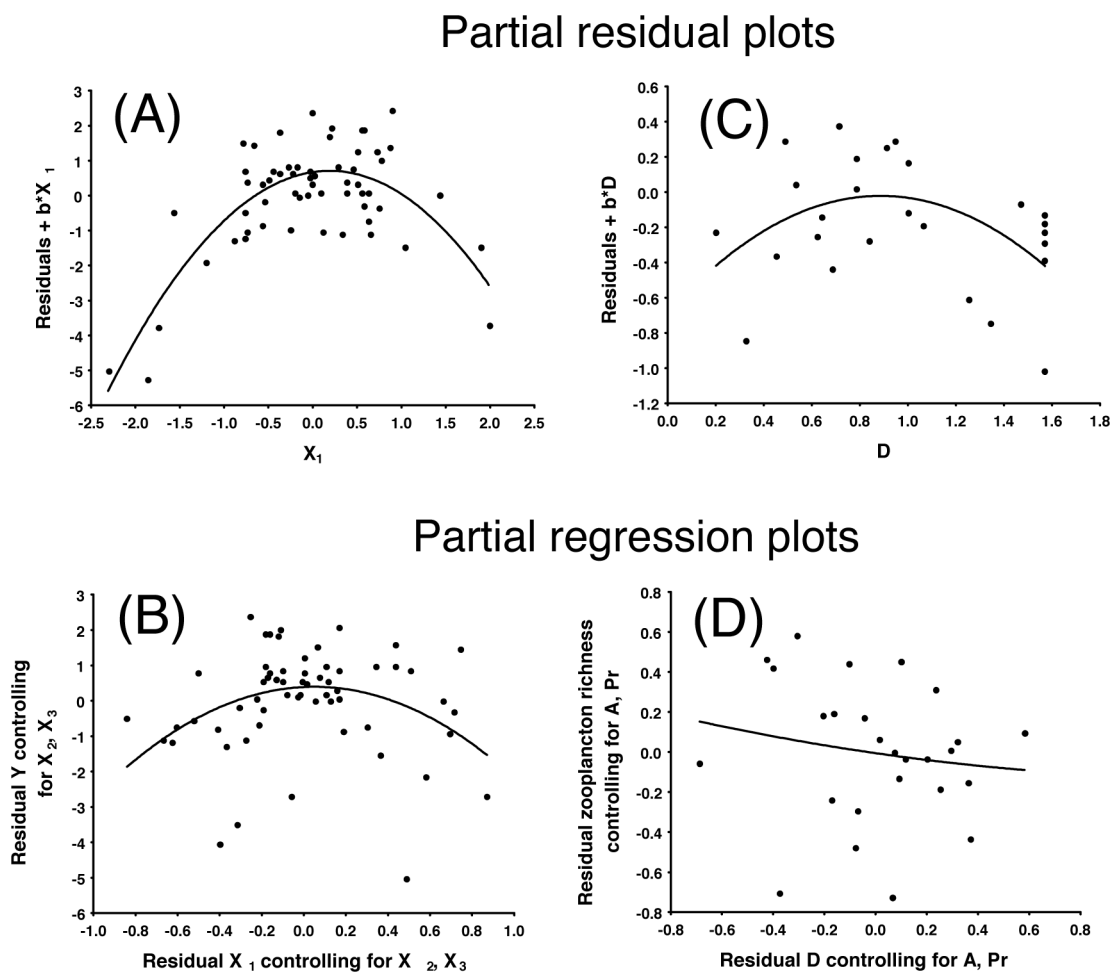


Fig. 2. Non-linear patterns in multiple regression are detected more efficiently in partial residual plots than in partial regression plots. (A) partial residual plot of X_1 in the simulated model $Y = b_0 + b_1X_1 - b_2X_1^2 + b_3X_2 - b_4X_3 + e$. (B) partial regression plot for the same variable. (C) partial residual plot for D in the model: zooplankton species richness = A – Pr + D in Hoffmann and Dodson (2005). A, log-transformed lake area; Pr, log-transformed primary productivity; D is arcsine-square-root transformed proportion of watershed developed. (D) partial regression plot for the same variable. Quadratic polynomials have been fitted to facilitate the assessment of the curvature. Note how partial residual plots show a true underlying curvilinear pattern more efficiently by making it more pronounced and how partial residual plots fail to show a pattern in a real-case scenario (plot D).

diagnostics. Bi-variate plots displaying dependent variables against each of the independent variables from a multiple regression should no longer be included in ecological publications because the slope may change in value and they may actually reflect opposite patterns compared to partial plots. These plots are only advisable when simple relationships are to be documented; i.e. when the relationship between Y and X is not influenced by other (measured) variables. Of the two available partial plots, only partial regression plots should be displayed in publications because they are the only ones that truly reflect the degree of partial correlation. Partial residual plots could be used if the only purpose is to show how Y changes within the range of observed X values. However, this can also be accomplished by adding the means of the raw variables to the residuals in partial regression plots. Partial regression plots are also more efficient than the conventional residuals versus predicted plots to detect influencing data points. However, partial residual plots are more efficient for detecting truly non-linear relationships between explanatory and dependent variables. The data necessary to plot partial correlations and regressions can easily be obtained by calculating the necessary models, obtaining the residuals from them and creating sub-sets of data. Some statistical packages even include the option to draw partial plots in their multiple regression modules (e.g. partial residual plots in STATISTICA 8.0 and partial regression plots in SAS 9.0 PROC REG (SAS Institute, Cary, USA in which they are called

partial regression residual plots) as well as in SPSS 15.0. However, these plots can be easily calculated and drawn with any statistical package that allows the calculation of residuals from GLM and GLZ models.

Because they reflect the true partial correlation coefficient, partial regression plots should be used more often in ecological studies when documenting main hypotheses from multiple regression and General and Generalized linear models involving normally distributed errors. We hope that this paper encourages the use of partial plots and consequently helps authors, reviewers and editors in the assessment of the quality of data in manuscripts.

Acknowledgements – We thank J. J. Soler and D. H. Wise for their helpful comments on an earlier version of this manuscript. We also thank K. Anderson, M. Frederiksen, D. J. Irschick and S. Dodson for sharing their data with us. We thank SAHFOS and ESA for permission to republish their data. This paper has been written under a Ramón y Cajal research contract from the Spanish Ministry of Science and Culture (MEC) to JML and a FPI scholarship (BES-2005-9234) to GC. This work has been partially funded by MEC grant CGL2004-03153 to JML and grant CGL2007-60520 to JML and GC.

References

Agresti, A. 2002. Categorical data analysis (2nd ed.). – Wiley.

Anderson, K. J. and Jetz, W. 2005. The broad-scale ecology of energy expenditure of endotherms. – Ecol. Lett. 8: 310–318.

Anderson, K. J. et al. 2006. Temperature-dependence of biomass accumulation rates during secondary succession. – Ecol. Lett. 9: 673–682.

Belsey, D. A. et al. 1980. Regression diagnostics. Identifying influential data and sources of collinearity. – Wiley.

Cook, R. D. and Weisberg, S. 1982. Residuals and influence in regression. Monographs on statistics and applied probability. – Chapman and Hall.

Darlington, R. B. and Smulders, T. V. 2001. Problems with residual analysis. – Anim. Behav. 62: 559–602.

Findlay, C. S. and Bourdages, J. 2000. Response time of wetland biodiversity to road construction on adjacent lands. – Conserv. Biol. 14: 86–94.

Freckleton, R. P. 2002. On the misuse of residuals in ecology: regression of residuals vs multiple regression. – J. Anim. Ecol. 71: 542–545.

Frederiksen, M. et al. 2006. From plankton to top predators: bottom-up control of a marine food web across four trophic levels. – J. Anim. Ecol. 75: 1259–1268.

Garcia-Berthou, E. 2001. On the misuse of residuals in ecology: testing regression residuals vs the analysis of covariance. – J. Anim. Ecol. 70: 708–711.

Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. – Ecology 84: 2809–2815.

Green, A. J. 2001. Mass/length residuals: measures of body condition or generators of spurious results? – Ecology 82: 1473–1483.

Hines R. J. O. and Carter E. M. 1993. Improved added variable and partial residual plots for the detection of influential observations in generalized linear models. – Appl. Stat. 42: 3–20.

Hoffmann, M. D. and Dodson, S. I. 2005. Land use, primary productivity, and lake area as descriptors of zooplankton diversity. – Ecology 86: 255–261.

Irschick, D. J. et al. 2005. Intraspecific correlations among morphology, performance and habitat use within a green anole lizard (*Anolis carolinensis*) population. – Biol. J. Linn. Soc. 85: 211–221.

James, F. C. and Mcculloch, C. E. 1990. Multivariate-analysis in ecology and systematics – panacea or Pandora box. – Annu. Rev. Ecol. Syst. 21: 129–166.

Larsen, W. A. and McCleary, S. J. 1972. Use of partial residual plots in regression-analysis. – Technometrics 14: 781–790.

Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. – J. Stat. Comput. Simul. 67: 37–73.

Littell, R. C. et al. 1996. SAS system for mixed models. – SAS Inst.

Montgomery, D. C. et al. 2001. Introduction to linear regression analysis (3rd ed.). – Wiley Interscience.

Moya-Laraño, J. et al. Analyzing body condition: mass, volume or density? – J. Anim. Ecol., in press.

Moya-Laraño, J. and Wise, D. H. 2007. Two simple strategies of analysis to increase the power of experiments with multiple response variables. – Basic Appl. Ecol. 8: 398–410.

Neter, J. et al. 1996. Applied linear statistical models. – McGraw–Hill.

Philippi, T. E. 1993. Multiple regression: herbivory. – In: Scheiner, S. M. and Gu revitch, J. (eds), Design and analysis of ecological experiments. Chapman and Hall, pp. 183–210.

Quinn, G. and Keough, M. 2002. Experimental design and data analysis for biologists. – Cambridge Univ. Press.

StatSoft, Inc. 2007. STATISTICA (data analysis software system), ver. 8.0. <http:www.statsoft.com>.

Velleman, P. F. and Welsch, R. E. 1981. Efficient computing of regression diagnostics. – Am. Stat. 35: 234–242.

Appendix A. Additional partial regression and partial residual plots from published ecological data.

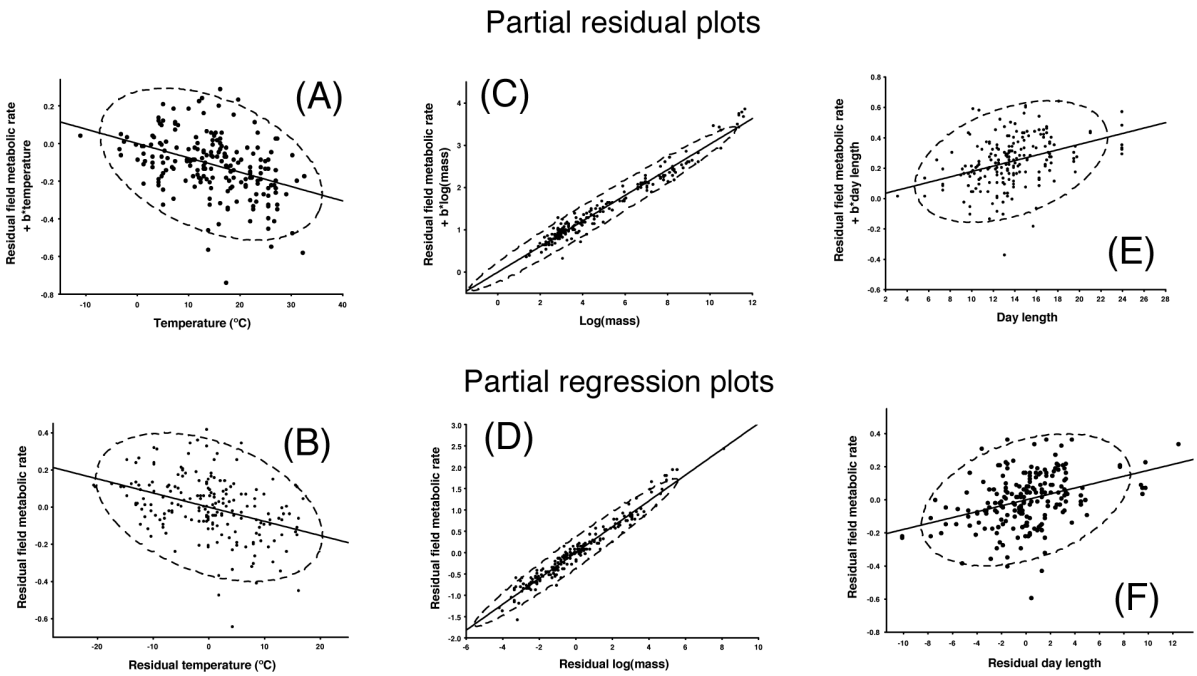


Fig. A1. Partial residual and partial regression plots from data in Fig. 4 of Anderson and Jetz (2005). The Y axis shows the residual field metabolic rate (FMR) for a mixture of birds and mammals. The graphs have been extracted from a GLM model that included: group (birds, mammals) and diet (nectarivores, carnivores, omnivores, herbivores) as categorical variables and two of the other three continuous variables in the graphs (i.e. temperature, log(mass) and/or day length). (A) partial residual plot of FMR vs temperature, (B) partial regression plot of FMR vs temperature, (C) partial residual plot of FMR vs log(mass), (D) partial regression plot of FMR vs log(mass), (E) partial residual plot of FMR vs day length and (F) partial regression plot of FMR vs day length. The study, which was conducted to explore the patterns explaining part of the residual variation in the allometry between body mass and energy expenditure, found that temperature and day length may be responsible for latitudinal variation in FMR. The average correlation among predictor variables was extremely low and non-significant (range 0–0.11), which explains why the difference in scatter around the lines for partial residual and partial regression plots is very similar. To facilitate the visual assessment of the scatter of points around the line, we show the 95% predicting ellipse (dashed line) in all partial plots.

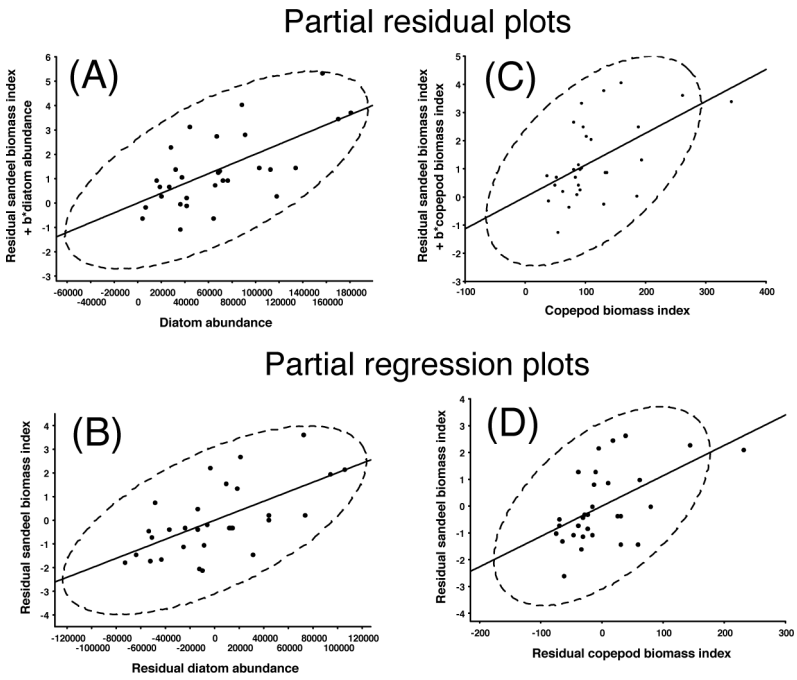


Fig. A2. Partial residual and partial regression plots from data in Fig. 3 of Frederiksen et al. (2006). The graphs were obtained from a multiple regression model that included year, diatom abundance and copepod abundance. The partial plots for year are in Fig. 1B in the printed part of this paper. (A) partial residual plot for diatom abundance, (B) partial regression plot for diatom abundance, (C) partial residual plot for copepod biomass and (D) partial regression plot for copepod biomass. Note the small difference between partial residual and partial regression plots due to the low correlation among predictors (see main text). To facilitate the visual assessment of the scatter of points around the line, we show the 95% predicting ellipse (dashed line) in all partial plots.

Appendix B. How a simulated influential point is detected in different kinds of plots

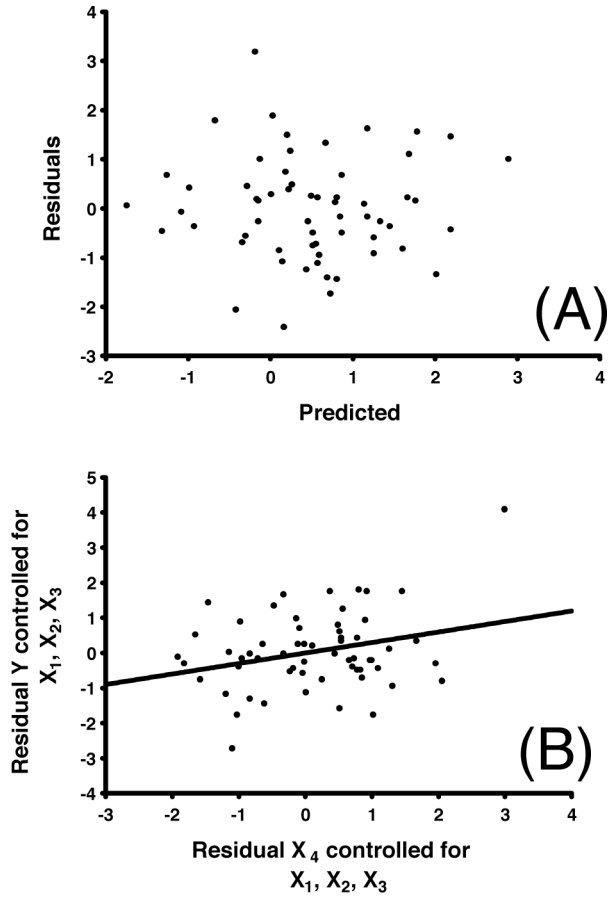


Fig. B1. Two graphic displays to detect influential data points in multiple regression using simulated data. (A) Plot of residuals vs fitted values from multiple regression. (B) the same data displayed in a partial regression plot shows a clear outlier. A very similar pattern would be shown in a partial residual plot (not shown). The independent variable X_4 is significant in a multiple regression model ($p = 0.038$). Eliminating the influential data point (top right in the partial regression plot) causes the significance of X_4 to disappear ($p = 0.386$). Note that in the plot of residuals vs fitted values the influential data point could easily have remained unnoticed.

Appendix C. Simulated data to build partial plots for assessing non-linear trends (Fig. 2)

X1	X1^2	X2	X3	Y
0.283065	0.080126	0.636109	–0.12527	–0.08815
–0.22635	0.051235	–0.1344	–0.22547	–0.12304
0.495625	0.245644	0.647152	–0.04904	–0.50296
0.567725	0.322312	–0.57743	0.275478	1.636631
0.388965	0.151294	0.603163	0.936586	0.097125
0.181334	0.032882	–0.36372	–0.41071	0.816304
0.878333	0.771469	1.213972	0.336596	0.725769
–0.54728	0.299516	–0.63687	–0.82273	–1.32291
–0.0692	0.004789	–0.25878	–0.60477	–1.03111
–0.36884	0.136041	–1.06838	–1.16906	–0.70071
0.199143	0.039658	0.256067	0.294812	1.513343
0.107103	0.011471	0.171459	0.706559	–1.07425
–0.88548	0.784077	–1.87662	–0.6095	–1.91789
1.894947	3.590824	1.010294	1.249495	–1.28096
0.369319	0.136396	0.371898	0.51486	0.154891
–0.5841	0.341177	–1.1929	–0.14883	–0.1015
–1.87366	3.510611	–1.72069	–1.92622	–7.05222
0.725868	0.526884	0.387154	0.860495	1.315733
–0.76669	0.587819	–1.0797	–1.18572	–1.83145
–0.6634	0.440097	–0.74757	–0.52848	0.572054
–0.77378	0.598735	–0.1301	–0.58318	–2.31848
–0.48897	0.239094	0.01406	–0.28229	–0.41084
0.093614	0.008764	–0.61308	–1.00952	–1.27222
0.507524	0.257581	0.19497	–0.09816	0.533644
–0.77892	0.606711	–0.85659	–1.25588	–0.73906
–0.2496	0.062302	0.113002	0.037685	–1.57909
–0.38102	0.145175	–0.72702	–0.72342	0.804862
–0.02441	0.000596	0.074932	0.072143	0.103545
–0.73112	0.53454	–0.78275	–0.90495	–0.77949
–0.16183	0.026188	0.076114	–0.19945	–0.84404
–0.20324	0.041307	–0.55276	–0.23553	–0.61027
–0.04675	0.002185	0.043822	–0.19652	–0.26864
0.547321	0.299561	0.874872	0.552828	–0.29631
0.44296	0.196213	0.344472	0.519469	0.506737
0.626119	0.392025	–0.06252	–0.45196	–0.92684
0.387133	0.149872	0.198559	0.768712	0.390229
–0.7856	0.617174	–1.00321	–0.59994	0.660695
1.041838	1.085427	0.584177	0.145434	–2.13508
–0.01112	0.000124	–0.37524	–0.07533	–0.27559
0.779431	0.607512	0.994223	1.116271	1.105686
–1.75231	3.070584	–2.09623	–1.94106	–5.52438
–0.19073	0.036377	–0.36479	–0.57089	–0.15332
–0.43983	0.193452	–0.55193	–0.07299	0.149565
0.56789	0.322499	0.741392	0.953965	–0.29297
–0.26918	0.072458	–1.07893	–0.53082	0.056734
0.749051	0.561078	0.75213	0.150372	–1.05145
–0.5812	0.337794	0.118218	0.383132	–1.14356
–2.30655	5.320167	–2.7767	–1.60255	–6.23952
0.655944	0.430263	0.142061	0.469282	–1.36045
–1.21357	1.472761	–0.937	–0.42573	–2.64126
1.985308	3.941446	1.574617	1.727165	–3.27408
1.418787	2.012956	0.944342	0.361842	–0.54419
0.558983	0.312462	0.205404	1.029157	2.085839
0.324766	0.105473	0.668624	0.777718	–1.22476
–0.00895	8.01E–05	–0.02049	0.643574	2.352469
–1.56674	2.454664	–2.18195	–1.69625	–1.9675
0.00507	2.57E-05	–0.04137	–0.15318	–0.15429
0.621033	0.385682	1.625419	0.948302	–0.93731
–0.74314	0.552257	–0.36084	–0.32917	–1.85287
0.893682	0.798667	1.216539	1.038355	2.427458