CONFRONTING MULTICOLLINEARITY IN ECOLOGICAL MULTIPLE REGRESSION (In press:Ecology)

Michael H. Graham

Moss Landing Marine Laboratories

8272 Moss Landing Road

Moss Landing, California, 95039

Email: mgraham@mlml.calstate.edu

Fax: (831) 632-4403

Phone: (831) 771-4481

Key words: multiple regression; confounding factors; multicollinearity; sequential regression; principal components regression; structural equation modeling

Abstract

The natural complexity of ecological communities regularly lures ecologists to collect elaborate data sets in which confounding factors are often present. Although multiple regression is commonly used in such cases to test the individual effects of many explanatory variables on a continuous response, the inherent collinearity (multicollinearity) of confounded explanatory variables encumbers analyses and threatens their statistical and inferential interpretation. Using numerical simulations, I quantified the impact of multicollinearity on ecological multiple regression and found that even low levels of collinearity bias analyses ($r \geq 0.28$ or $r^2 \geq 0.08$), causing: (1) inaccurate model parameterization; (2) decreased statistical power; and (3) exclusion of significant predictor variables during model creation. Then, using real ecological data, I demonstrated the utility of various statistical techniques for enhancing the reliability and interpretation of ecological multiple regression in the presence of multicollinearity.

Introduction

Ecologists often use multiple regression to develop models that describe the regulation of particular aspects of organismal, population, and community ecology (dependent or response variables) by various environmental and biological factors (independent or explanatory variables) (James and McCulloch 1990). Multiple regression analyses, however, can be hindered by the complex nature of ecological data, in which targeted ecological responses are linked to many explanatory variables that are often correlated among each other (multicollinear). Multicollinear explanatory variables are difficult to analyze because their effects on the response can be due to either true synergistic relationships among the variables or spurious correlations. Ecologists often counter by designing experimental studies that break correlations among explanatory variables and analyzing the data with analyses of variance (ANOVA) that allow for the isolation of main effects

and higher-order interactions among individual explanatory variables (Scheffe 1959). In practice, however, ecological explanatory variables are often not under experimental control, in which case the explanatory variables of interest may be correlated. It is under these conditions that multiple regression is often used to analyze ecological data (James and McCulloch 1990).

The statistical and inferential problems of multicollinearity in multiple regression have been well established in the statistical literature (e. g. Cohen and Cohen 1983, Hocking 1996, Neter et al. 1996, Tabachnick and Fidell 1996, Draper and Smith 1998, Chatterjee et al. 2000), although problems specific to ecological data have rarely been discussed (James and McCulloch 1990, Phillipi 1993, Legendre and Legendre 1998, and see Mitchell-Olds and Shaw 1987 and Petraitis et al. 1996 for related discussions of fitness regression and path analysis, respectively). Yet, despite previous warnings by statisticians, only 32 of 294 (11%) papers published in Ecology, Ecological Monographs, Functional Ecology, Journal of Animal Ecology, and Journal of Ecology from 1993 to 1999 that used multiple regression for data analysis even discussed the potential presence of multicollinearity. Of these 32 papers, only 17 (53%) actually tested whether multicollinearity was present; of these 17 papers, 11 (65%) found significant multicollinearity, suggesting that ecological data are typically collinear. But how desperate is the problem for ecologists? The goal of this paper was two-fold: (1) to quantify through numerical simulation the statistical and inferential biases caused when multicollinearity is present in multiple regression analyses; and (2) to demonstrate the utility of various statistical techniques for enhancing the reliability and interpretation of ecological multiple regression in the presence of multicollinearity.

Theoretical Problems and Empirical Consequences

In multiple linear regression, data are fit to a linear model that predicts values of a response (Y) as the weighted sum of explanatory variables (X_i) and random error (ε): $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +$

49 $\cdots + \beta_i X_i + \varepsilon$, where β's are regression coefficients. The typical goal is to build a model using the
50 fewest variables to explain the greatest variability in the response, and to parameterize accurately
51 regression coefficients for those variables. If all explanatory variables are independent of each other,
52 each regression coefficient represents the total contribution of a given predictor to the response. If,
53 however, two or more variables are collinear to any extent, *partial* regression coefficients need to be
54 calculated to isolate the *unique* contribution of a particular explanatory variable (hereafter the
55 *predictor*) from that shared with other variables (hereafter *confounders*). This unique contribution is
56 the *extra sums of squares*. The distinction between unique and shared contributions is the crux of
57 multiple regression's statistical and inferential problems due to multicollinearity.

58 When data are standardized to a mean of zero and unit variance, the partial regression
59 coefficient for a predictor in the presence of a single confounder is defined as: $\beta = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2}$, where
60 r_{Y1} is the correlation between the response and predictor, r_{Y2} is that between the response and
61 confounder, and r_{12} and r^2_{12} are the correlation and coefficient of determination between the predictor
62 and confounder (Neter et al. 1996); β reduces to $\beta^* = r_{Y1}$ in the absence of multicollinearity (i. e. r_{12}
63 and $r^2_{12} = 0$). As such, partial regression coefficients decrease non-linearly with increasing
64 multicollinearity (as shown by Petraitis et al. 1996) and deviations from β^* will occur in the presence
65 of even the weakest multicollinearity (i. e. $\beta < \beta^*$ at all $r^2_{12} > 0$). The marginal statistics used to test
66 the significance of β (i. e. H_o:$\beta \neq 0$), which is typically used as a criterion to determine whether a
67 given predictor is to be included in a model, is defined as $t = \frac{\beta}{SE(\beta)}$ (or $t = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{MS_{residual}}}$). Here,
68 SE(β) is the standard error of the coefficient which increases linearly with increasing r^2_{12} (Neter et al.
69 1996). Power to detect an effect as significant will therefore also decrease non-linearly with
70 increasing multicollinearity.

If, during stepwise variable selection, a predictor is ultimately excluded from a model due to its low apparent significance, regression coefficients and marginal statistics of the other variables will change (Mitchell-Olds and Shaw 1987, Philippi 1993, Petraitis et al. 1996, Neter et al. 1996). The use of stepwise variable selection procedures that rely on calculation of marginal statistics may even exclude explanatory variables that are actually highly correlated with the response (i. e. decrease statistical power). Furthermore, although statistical significance and fit of a final model are not directly affected by multicollinearity (expected sums of squares and marginal statistics are not computed), interpretation of the model may be uncertain due to biased parameterization of partial regression coefficients for individual explanatory variables. Not only will the sum of r^2 for individual predictors generally differ from the R^2 of the final model, actual application of the final model to predict future values for the response can be grossly inaccurate, since none of the partial regression coefficients reflect shared contributions (Tabachnick and Fidell 1996).

These statistical difficulties in analyzing ecological data in the presence of multicollinearity were illustrated numerically by calculating marginal t-statistics (as described above) and *P*-values for a predictor in the presence of a single confounder (Figure 1). The purpose of the simulation was to estimate the level of multicollinearity that would result in the erroneous exclusion of significant predictors from a final model. In general: (1) apparent significance (*P* or apparent α) decreased rapidly with increasing multicollinearity; (2) weak predictors were more vulnerable to erroneous exclusion than strong ones; (3) predictors with high true significance became more vulnerable to erroneous exclusion as the correlation between the response and confounder (r_{Y2}) increased; and, (4) even if correlations between the response and confounders were relatively weak, low levels of multicollinearity (i. e. $r_{12} \geq 0.28$ or $r^2_{12} \geq 0.08$) resulted in significant predictors appearing insignificant.

To illustrate the negative impact of these statistical biases on the reliability and interpretation

of ecological multiple regression, data were re-analyzed from a study of the effect of various environmental factors (wave orbital displacement, wave breaking depth, wind velocity, and average tidal height) on the shallow (upper) distributional limit of the subtidal kelp *Macrocystis pyrifera* (Graham 1997; Appendix 1). The overall severity of multicollinearity in these data was moderate, as wave orbital displacement, wave breaking depth, and wind velocity were strongly correlated among each other ($r \geq 0.6$; $VIF \geq 2$), but tidal height was only weakly correlated with the other variables ($r < 0.4$; $VIF = 1.17$). Although Neter et al. (1996) and Chatterjee et al. (2000) suggested that multicollinearity is only severe at VIFs > 10, it is clear from Figure 1 that VIFs as low as 2 can have significant impacts (see also Petraitis et al. 1996). When analyzed using separate linear regressions, all of the explanatory variables were significant or marginally significant predictors of the response (i. e. $P \leq 0.1$; Table 1). Backwards stepwise multiple regression, however, suggested that only wave orbital displacement and wind velocity were important (Table 1; forward selection yielded the same final model). Partial regression coefficients (β in standard and sequential regressions; Table 1) were often more than 1 SE lower than the non-partial regression coefficients (β in simple regressions; Table 1), reflecting the omission of variability in the response shared among predictors. Thus, although wave-breaking depth was initially identified as important (Table 1), this was due almost entirely to variability shared with wave orbital displacement. Many would argue that the removal of wave breaking depth was therefore necessary because it was a redundant variable, however, there was no evidence that wave-breaking depth wasn't the variable functionally responsible for the shared contribution. Clearly, for two highly collinear explanatory variables that have a strong shared contribution to the response, the decision as to which is the most important predictor, and should therefore be retained, is very ambiguous.

SOME OLD AND NOT-SO-OLD SOLUTIONS

If the entire goal of conducting a multiple regression analysis is to develop a model that best predicts variability in the response, and there is no interest in studying particular relationships between the response and explanatory variables, then the problems due to multicollinearity can be effectively ignored (i.e. "the proof is in the pudding" scenario). In most ecological studies, however, researchers are interested in examining the effects of particular explanatory variables, in which case various techniques are available for addressing the statistical pitfalls of multicollinearity. One approach is to avoid or stabilize the use of marginal statistics for variable selection. The easiest way to do this is to simply drop collinear variables from analysis (Philippi 1993, Legendre and Legendre 1998). Variable exclusion, however, ignores the unique contribution of the omitted variable and can result in a substantial loss of explanatory power (Carnes and Slade 1988, James and McCulloch 1990) as well as inferential problems in choosing which variables should remain (Mitchell-Olds and Shaw 1987). Another method is to avoid using marginal statistics during variable selection by predetermining model composition (*a priori* modeling). This circumvents the problem of choosing which collinear variables should be excluded. In the absence of a reasonable *a priori* model, marginal statistics can also be avoided by using an "all possible subsets" method of analysis (Furnival 1971). F-statistics and coefficients of determination are calculated for all possible combinations (subsets) of variables, and the subset with the greatest fit is identified as "best" using adjusted R^2 (or Akaike's Information Criteria, Mallow's C_p, PRESS, MSE, etc.; Neter et al. 1996). Since distinctions are not made between unique and shared contributions, all possible subsets analyses can help to identify reliably the final model that explains the most variability in the response, although the number of potential subsets can become analytically untreatable as the number of variables increases. An alternative to avoiding marginal statistics is to stabilize them using ridge regression, in which a constant is applied to the elements of the correlation matrix so that it is displaced from singularity, increasing the precision of the coefficients (Birkes and Dodge 1993). A problem with all of these

143 methods is that they still require the use of marginal statistics to estimate regression coefficients or
144 determine the relative importance of individual explanatory variables, and thus offer no refuge from
145 associated biases due to multicollinearity.

146 A more purposeful approach to solving the problems due to multicollinearity is to explore the
147 functional nature of the collinearities, rather than avoid them. This requires methods for identifying
148 and parameterizing the unique and shared contributions of explanatory variables to a response. Here
149 I used the kelp forest example data to illustrate how three such methods (residual/sequential
150 regression, principle components regression, and structural equation modeling) can improve the
151 reliability and interpretation of ecological multiple regression in the presence of multicollinearity.

152 *Residual and sequential regression* - When multicollinearity is limited to pairs of explanatory
153 variables, the easiest way to disentangle unique from shared contributions is simply to assume that
154 one variable is functionally more important than the other, assign the more important variable priority
155 over the shared contribution, and ignore the shared contribution when analyzing the less important
156 variable. This can be done by regressing the less important variable against the other, and replacing
157 the less important variable with the residuals from the regression (see for example, Graham 1997).
158 Priorities can be based on a researcher's own instincts and intuition, previously collected data, data
159 currently under analysis, or the results of prior experiments that estimated the relative importance of
160 one factor over another. Subsequent multiple regression analyses (residual regressions) will be
161 unbiased since the explanatory variables are no longer *statistically* collinear. As multicollinearity
162 among explanatory variables becomes more complicated, a modification of sequential regression (or
163 hierarchical regression) can be used. Here it is also assumed that some variables are functionally
164 more important than others, but *fixed* priorities are assigned to shared contributions for all variables
165 in the model (Tabachnick and Fidell 1996). Marginal statistics are computed for variables in order of
166 highest to lowest priority, with any given variable's marginal statistics ignoring variability already

167 explained by higher priority variables. As such, the rank (order) of marginal statistics remains
168 constant as variables are added or removed from the model, and the decision as to whether a
169 particular variable should remain in the model does not depend on the presence of other variables.
170 Furthermore, both unique and shared contributions are represented in the final parameterized model
171 by the regression coefficients and coefficients of determination. The major concern when using these
172 methods is whether the assigned priorities are relevant to the true functional importance of the
173 variables, and thus, it is vital that researchers are critical of the criteria used to assign priorities.

174 The final model from a sequential regression analysis of the example data is presented in
175 Table 1, where priorities were based on the unique contributions of each explanatory variable.
176 Regression coefficients and the rank of marginal statistics were constant for each variable selection
177 step (Appendix 2C) and confirmed that, by assigning fixed priorities, the decision as to whether a
178 particular variable should remain in the model does not depend on the presence of other variables and
179 model composition is not affected by the use of marginal statistics. It was concluded from this
180 analysis that the unique contribution of wave orbital displacement, plus its shared contribution with
181 winds, was the most important predictor of the response, but that the unique contribution of winds
182 was also important (Graham 1997). Note that, although the standard and sequential multiple
183 regressions yielded the same final models, with sequential regression analyses both unique and
184 shared contributions are represented by the regression coefficients and coefficients of determination,
185 and the individual r^2 values summed to R^2.

186 *Principal components regression* - Alternatively, in principal components regression it is not
187 generally believed that multicollinearity can be understood best by a hierarchical assignment of
188 priorities, but that collinearities indicate the presence of underlying (latent) variables that are
189 responsible for the shared contributions (Tabachnick and Fidell 1996). A principal components
190 analysis is done on the explanatory variables which identifies vectors (i.e. the linear combinations of

variables) that account, successively, for the greatest variation in the observations of the explanatory variables; the principal components analysis is done in complete disregard of observed variability in the response. Scores of the orthogonal principal components are used as explanatory variables in a subsequent multiple regression analysis (Philippi 1993, Tabachnick and Fidell 1996, Legendre and Legendre 1998). Since principal components are orthogonal, partial regression coefficients and the rank of marginal statistics do not fluctuate as variables are added or removed and the results of principal components regression will be stable regardless of the severity of multicollinearity. Given that variable selection is unbiased in principal components regression, all principal components can and *should* be included during variable selection, avoiding the concerns of Mitchell-Olds and Shaw (1987) that explanatory power may be lost by limiting analyses to only those variables with high eigenvalues. The primary limitation of principal components regression lies in the biological interpretation of the principal components.

A principal components analysis was performed on the example data (Appendix 3). PC1 accounted for 64% ($\lambda = 2.57$) of the variability among the variables, with wave orbital displacement, wave breaking depth, and wind velocity loading heavily and positively on this PC (all loadings $\geq$ 0.86); average tidal height loaded moderately and negatively (loading = -0.54). PC1 thus represented high wave intensity, high wind velocity, and low tide height, or the occurrence of storms during low tides (see Graham [1997] for a detailed biological interpretation of these data). PC2 explained only 20% of the variability ($\lambda = 0.81$) and appeared to represent mostly tides (loading = 0.84; all others $\leq$ 0.26). PC3 explained less than 10% of the variability ($\lambda = 0.37$) and primarily represented wind activity (loading = 0.49; all others $\leq$ 0.19). PC4 explained approximately 6% of the variability ($\lambda =$ 0.26) and represented differences in the two estimates of wave intensity (OD and BD loaded –0.39 and 0.29 respectively; all others $\leq$ 0.13). The subsequent principal components regression confirmed the stability of regression coefficients and marginal statistics (Appendix 3C) and that individual r^2

values also summed to the total R^2 for the final model (Table 1). Not surprisingly, the PC that represented the occurrence of storms (PC1) explained the greatest amount of variation in the response. The importance of winds (PC3), however, was not emphasized in the principal components regression. Instead, PC4 was retained suggesting the importance of distinguishing between different aspects of wave intensity, despite the fact that PC4 explained only ~6% of the variability among explanatory variables. That the sequential and principal components regression analyses yielded different results when applied to identical data highlights the importance of determining whether latent variables are likely driving variability in the measured explanatory variables.

Structural equation modeling - Like residual/sequential regression and principal components regression, in structural equation modeling (SEM) it is generally assumed that the best functional multiple regression model is one that can account for both unique and shared contributions. Moreover, like *a priori* modeling, SEM does not simply explore data to search for relationships between the response and explanatory variables, but rather sets out to test and parameterize hypothesized relationships among the variables. As such, SEM can be used to develop accurate and meaningful final multiple regression models when collinearities among explanatory variables are thought to be present (Hayduk 1987, Loehlin 1987, Bollen 1989, Bentler 1995, Ullman 1996, Shipley 1999). Hypothetical causal links among variables (both unique and shared contributions) are specified and structural equations (models) are developed that represent each potential combination of links. Regression coefficients are then parameterized simultaneously for each link of each model (Bentler 1995, Ullman 1996) and the overall fit of the models are compared as with "all possible subsets" techniques (see above). In its generalized form, SEM directly incorporates latent variables into its models that can represent shared contributions (Ullman 1996; for ecological examples see Brown and Weis 1995, Bishop and Schemske 1998, Gough and Grace 1999), and thus avoids many of the problems identified by Petraitis et al. (1996) for path analysis. Still, the successful application

239 of SEM to ecological data is vulnerable to inferential errors made during model development and
240 selection (Ullman 1996; Shipley 1999): for example, alternate models may exist that differ greatly in
241 the form of their hypothetical causal links, yet may explain similar amounts of variability in the
242 response.

243 An SEM was developed for the example data, representing one potential relationship between
244 the four predictor variables (wave orbital displacement, wave breaking depth, average tidal height
245 and wind velocity) and the response (giant kelp shallow limit) (Figure 2; Appendix 4). It was
246 hypothesized that two latent variables were important in driving variability in the response. The first
247 structural equation specified that the latent variable *wave intensity* could be estimated by a linear
248 combination of wave orbital displacement and wave breaking depth. The second structural equation
249 specified that the latent variable *storm intensity* could be estimated by a linear combination of wind
250 velocity, average tidal height, and the latent variable wave intensity. The final structural equation
251 simply specified the linear relationship between the latent variable storm intensity and the response.
252 Again, the results of the parameterized SEM support the conclusions of the sequential and principal
253 components regressions, identifying the underlying importance of storm activity during low tides in
254 driving variability in giant kelp upper limits. Furthermore, by including latent variables into the
255 model, various unique and shared contributions among explanatory variables were explicitly
256 parameterized. However, although R^2 was almost identical among the various methods (i. e. 0.59-
257 0.60), the adjusted R^2 was in fact lower for the SEM (0.51) than the sequential (0.57) and principal
258 components (0.55) regressions, due to the greater number of SEM regression coefficients that needed
259 to be parameterized. Thus, although the incorporation of latent variables adds flexibility during
260 model development, SEM may not provide the greatest explanatory power for all data analyses.

261 *Post-analysis* – The application of one of the above techniques should not be considered the
262 final step in analysis of collinear data. First, each technique demands the standard set of parametric

263 assumptions: normality, constant variance and independence of error terms. As such, thorough
264 analysis of model residuals should always follow the application of multiple regression techniques.
265 Some techniques (e.g. principal components analysis) additionally require (1) non-singular matrices
266 of the correlation-covariance among explanatory variables, and (2) that the number of observations of
267 the response greatly exceeds the number of explanatory variables (Tabachnick and Fidell 1996).
268 Second, the generality of estimated regression coefficients should be validated against data that are
269 collected independently of those used during model parameterization. Such validation procedures
270 may also be useful for assessing whether a given multiple regression technique offers the greatest
271 explanatory power. Finally, structural equation modeling and residual, sequential, and principal
272 components regression all deal with shared vs. unique variance contributions differently, and
273 therefore provide diverse perspectives as to the nature of the underlying multicollinearity. As such,
274 ecologists will likely find it most useful to explore multicollinear data with a combination of
275 techniques.
276
277 ## CONCLUSION
278 This study has quantitatively shown that statistical and inferential problems created by
279 multicollinearity can be extremely severe under realistic ecological conditions. Although
280 straightforward techniques exist for diagnosing and remediating the effects of multicollinearity in
281 multiple regression, they are not commonly utilized in ecology. Still, most of these procedures only
282 help to stabilize the statistical analyses, making them less biased, less subjective, and more
283 repeatable, but only the statistical collinearity will have been removed from the data. The
284 explanatory variables are still, by nature and in nature, correlated, whether or not functionally. Aside
285 from designing manipulative experiments to break correlations among explanatory variables, no
286 technique exists that allows researchers to *infer* the different functional relationships between the

response and explanatory variables. Experiments, however, cannot be applied under all field situations and are especially difficult during the exploratory stage of data collection and model development. It is then that the determination of relative importance of individual explanatory variables via sampling, and thus a distinction between unique and shared variance contributions, becomes important. The suite of techniques described herein compliment each other and offer ecologists useful alternatives to standard multiple regression for identifying ecologically relevant patterns in collinear data. Each comes with its own set of benefits and limitations, yet together they allow ecologists to directly address the nature of shared variance contributions in ecological data.

ACKNOWLEDGMENTS

M. Edwards, S. Thrush, P. Wainwright, G. Leonard, L. Ferry-Graham, D. Strong, A. Ellison, and two anonymous reviewers provided useful comments on the manuscript and participated in various discussions of multicollinearity.

LITERATURE CITED

Bentler, P. M. 1995. EQS: Structural equations program manual. Multivariate Software Inc., Encino.

Birkes, D., and Y. Dodge. 1993. Alternative methods of regression. John Wiley & Sons, New York.

Bishop, J. G., and D. W. Schemske. 1998. Variation in flowering phenology and its consequences for lupines colonizing Mount St. Helens. Ecology **79**:534-546.

Bollen, K. A. 1989. Structural equations with latent variables. John Wiley & Sons, New York.

Brown, D. G., and A. E. Weis. 1995. Direct and indirect effects of prior grazing of goldenrod upon the performance of a leaf beetle. Ecology **76**:426-436.

Carnes, B. A., and N. A. Slade. 1988. The use of regression for detecting competition with multicollinear data. Ecology **69**:1266-1274.

311 Chatterjee, S., A. S. Hadi, and B. Price. 2000. Regression analysis by example. John Wiley & Sons,
312 New York.
313 Cohen, J., and P. Cohen. 1983. Applied multiple regression/correlation analysis for the behavioral
314 sciences. Lawrence Erlbaum, Hillsdale.
315 Draper, N. R., and H. Smith. 1998. Applied regression analysis. John Wiley & Sons, New York.
316 Furnival, G. M. 1971. All possible regressions with less computations. Technometrics **13**:403-408.
317 Gough, L., and J. B. Grace. 1999. Effects of environmental change on plant species density:
318 Comparing predictions with experiments. Ecology **80**:882-890.
319 Graham, M. H. 1997. Factors determining the upper limit of giant kelp, *Macrocystis pyrifera* Agardh,
320 along the Monterey Peninsula, central California, USA. Journal of Experimental Marine Biology
321 and Ecology **218**:127-149.
322 Hayduk, L. A. 1987. Structural equation modeling with LISREL: Essentials and advances. Johns
323 Hopkins University Press, Baltimore.
324 Hocking, R. R. 1996. Methods and applications of linear models: Regression and the analysis of
325 variance. John Wiley & Sons, New York.
326 James, F. C., and C. E. McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea
327 or Pandora's box? Annual Review of Ecology and Systematics **21**:129-166.
328 Legendre, P., and L. Legendre. 1998. Numerical ecology. Elsevier, Amsterdam.
329 Loehlin, J. C. 1987. Latent variable models. Lawrence Erlbaum, Hillsdale.
330 Mitchell-Olds, T., and R. G. Shaw. 1987. Regression analysis of natural selection: statistical
331 inference and biological interpretation. Evolution **41**:1149-1161.
332 Neter, J, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models.
333 Irwin, Chicago.
334 Petraitis, P. S., A. E. Dunham, and P. H. Niewiarowski. 1996. Inferring multiple causality: the

limitations of path analysis. Functional Ecology **10**:421-431.

Philippi, T. E. 1993. Multiple regression: Herbivory. Pages 183-210 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Chapman and Hall, New York.

Scheffe, H. 1959. The analysis of variance. John Wiley & Sons, New York.

Shipley, B. 1999. Testing causal explanations in organismal biology: causation, correlation and structural equation modeling. Oikos **86**:374-382.

Tabachnick, B. G., and L. S. Fidell. 1996. Using multivariate statistics. HarperCollins, New York.

Ullman, J. B. 1996. Structural equation modeling. Pages 709-812 in B. G. Tabachnick and L. S. Fidell, editors. Using multivariate statistics. HarperCollins College Publishers, New York.

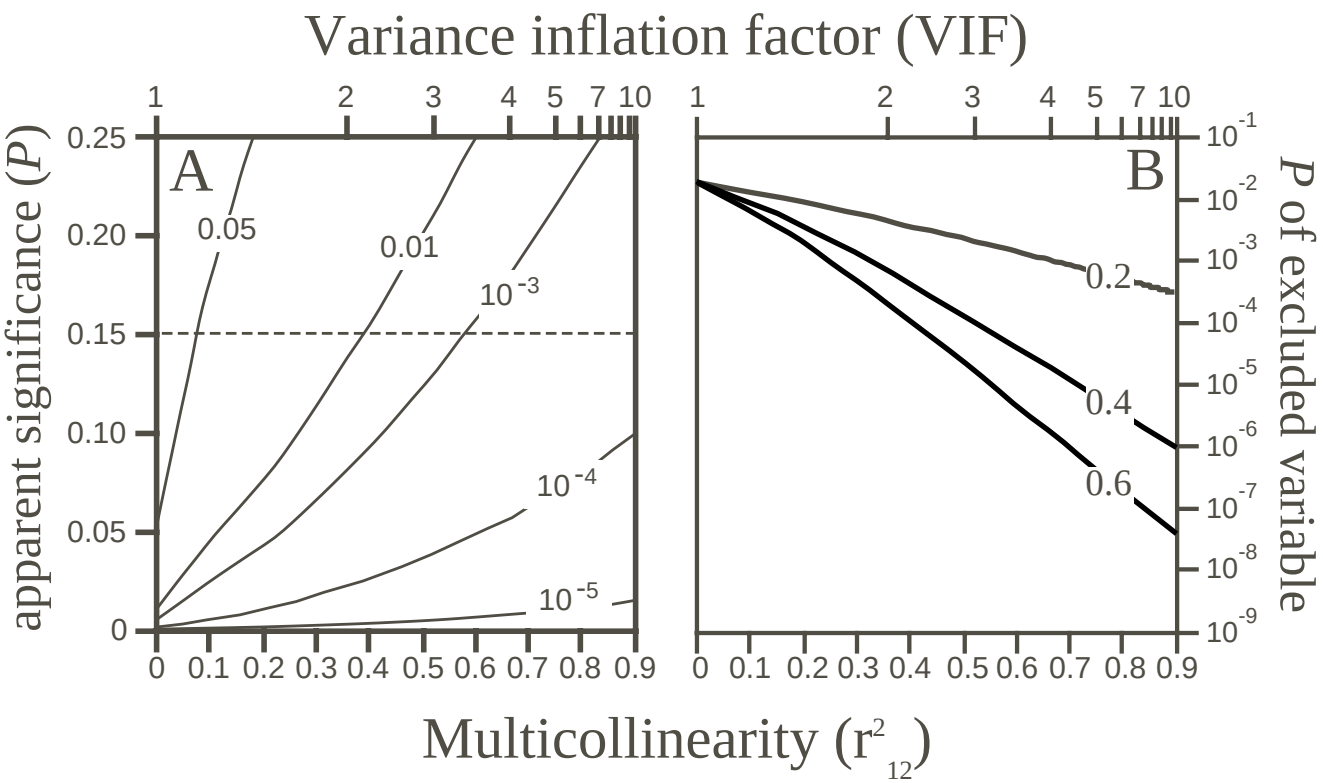
344 Table 1. Simple linear regressions, and final models from standard multiple regression, sequential
345 regression, and principal components regression (final models are after removal of insignificant
346 explanatory variables; *P*-values ≥ 0.15). Model intercepts were significant for all models (*P* <
347 0.0001) and are not given. r^2 in standard and principal components regressions represent total
348 contributions, whereas in the sequential regression, r^2 represents either unique or unique + shared
349 contributions as determined by assigned priorities.
350

Method	Variable	□	SE	t	$F_{df=2,35}$	*P*	r^2
Simple	wave orbital displacement	0.194	0.030	-	43.58	< 0.001	0.55
	wave breaking depth	0.072	0.017	-	17.41	< 0.001	0.33
	wind velocity	0.018	0.003	-	29.15	< 0.001	0.45
	tidal height	-0.358	0.191	-	3.53	0.068	0.09
Standard	wave orbital displacement	0.139	0.038	3.62	-	0.001	0.55
	wind velocity	0.008	0.004	2.10	-	0.043	0.45
	Total	-	-	-	26.06	< 0.001	0.60
Sequential	wave orbit. displ. (1^{st} prior.)	0.194	0.028	6.91	-	< 0.001	0.55
	wind velocity (2^{nd} prior.)	0.008	0.004	2.09	-	0.043	0.05
	Total	-	-	-	26.06	< 0.001	0.60
Principal components	principal component 1	0.157	0.024	6.66	-	< 0.001	0.54
	principal component 4	-0.039	0.024	-1.69	-	0.100	0.03
	Total	-	-	-	23.62	< 0.001	0.57

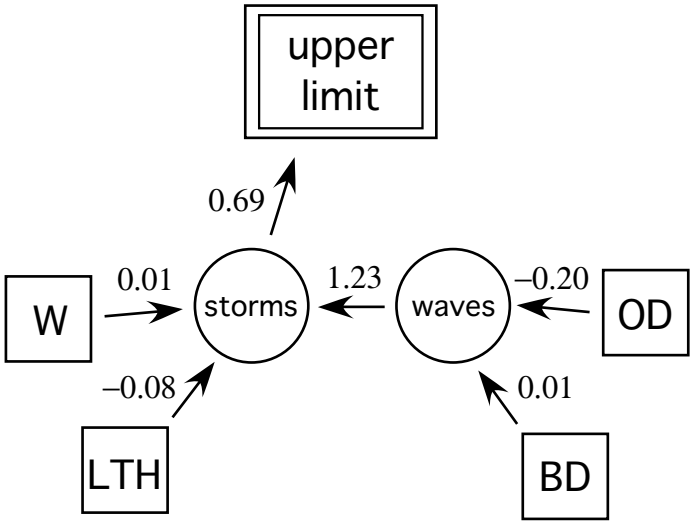
FIGURE LEGENDS

Figure 1. A) Effect of multicollinearity on predictor apparent significance (*P* or apparent □) in the presence of a single confounder. Multicollinearity was represented by r^2_{12} and variance inflation factors ($VIF = \frac{1}{1 - R^{2*}}$; R^{2*} is the R^2 when explanatory variable *i* is regressed on all other variables in model). r_{Y1} values were chosen to provide specific levels of "true" significance in the absence of multicollinearity (□; given on each line). $MS_{residual}$ equaled 0.1 for 35 df and was taken from real standardized data. Predictors with apparent *P*-values that increased to ≥ 0.15 were considered to be negatively affected by multicollinearity; all true *P*-values were ≤ 0.05 in this exercise. B) Effect of multicollinearity on the exclusion of a significant predictor. The y-axis is the true significance of predictors (expressed as *P*) that would have been excluded during variable selection. Data were obtained by setting t = 1.47 (corresponding to *P* = 0.15 for 35 df) and solving for "true" significance (r_{Y1}) as a function of various levels of multicollinearity (r^2_{12} or VIF) and confounder strength (r_{Y2}; given on each line).

Figure 2. A structural equation model representing the relationship between four measured explanatory variables (wave orbital displacement [OD], wave breaking depth [BD], average tidal height [LTH], and wind velocity [W]), two latent variables (storm intensity and wave activity), and the response variable (giant kelp shallow limit). Arrows depict the proposed links between each variable. Parameterized regression coefficients are associated with each link. The coefficients were parameterized using iterative normal-theory maximum likelihood available with EQS 6 for Windows. The latent variables were developed using the covariance matrix and varimax rotation, and initiated using adjusted principal components according to Bentler (1995).



Graham Figure 1



Graham Figure 2

Appendix 1. A) Original kelp forest example data and B) SYSTAT output for backwards stepwise multiple regression. *Response* = depth of giant kelp shallow limit (in meters); *OD* = wave orbital displacement (in meters); *BD* = wave breaking depth (in meters); *LTD* = average tidal height (in meters); *W* = wind velocity (in meters/s). Backwards stepwise multiple regression was run using MGLH:REGRESSION under the STATS menu in SYSTAT 5.2 for Macintosh. *Response* was the dependent variable; *CONSTANT, OD, BD, LTD*, and *W* were the dependent variables; Stepwise was set to *custom*, with a *backwards* step order, and *P = 0.15* to remove. Only standard SYSTAT output is presented.

A)

Response	OD	BD	LTD	W
3.241	2.0176	4.87	-0.59	-4.1
3.032	1.9553	4.78	-0.75	4.7
3.100	1.8131	3.14	-0.38	-4.9
3.156	2.5751	3.28	-0.16	-3.2
3.110	2.2589	3.28	0.01	5.6
3.127	2.5448	4.87	-0.19	4.1
3.456	2.6291	6.27	-0.14	5.3
3.244	3.1553	7.16	-0.43	23.5
3.565	3.4030	7.24	-0.34	13.3
3.116	2.8150	7.16	-0.33	-4.5
3.186	1.9012	4.78	-0.27	-4.7
3.210	2.1463	3.28	-0.20	-4.6
3.215	2.5851	3.47	-0.48	-4.0
3.368	2.0830	3.14	-0.26	-2.0
3.170	1.7192	4.78	-0.09	-2.8
3.625	3.5471	4.78	-0.17	3.5
3.445	3.6720	7.16	-0.34	13.2
3.680	4.7259	8.65	-0.54	20.6
3.618	3.6039	8.65	-0.70	14.1
3.824	4.1214	6.16	-0.46	8.3
3.345	3.4940	4.87	-0.31	6.8
3.168	3.4829	5.82	-0.34	-4.7
3.200	2.0793	3.47	-0.42	-3.7
3.038	2.0315	3.14	-0.37	-4.4
2.992	1.7356	3.14	-0.42	-4.7
3.001	1.4569	3.14	-0.25	-4.3
3.183	1.8559	3.28	-0.02	-4.4
3.277	2.6173	3.28	-0.29	4.6
3.231	2.8782	5.82	-0.31	7.6
3.517	2.7842	4.78	-0.32	17.8
3.125	3.3236	7.16	-0.10	-3.6
3.063	2.8799	3.47	-0.09	-3.7
3.155	1.9654	3.47	-0.07	-5.0

3.049	1.5116	3.28	-0.31	-5.0
3.082	1.7465	3.28	-0.34	-4.2
3.023	1.0967	3.14	-0.06	-4.4
3.042	2.9802	3.28	-0.28	4.1
3.515	3.0644	3.28	-0.52	8.5

B)

DEPENDENT VARIABLE: RESPONSE

MINIMUM TOLERANCE FOR ENTRY INTO MODEL = 0.010000

STEP #0; R= 0.775; RSQUARE= 0.601

IN

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
1 CONST						
2 OD	0.1433	0.046	0.55	0.38836	9.57	0.0040
3 BD	-0.0042	0.021	-0.03	0.42462	0.04	0.8402
4 LTD	-0.0596	0.142	-0.05	0.85087	0.17	0.6791
5 W	0.0079	0.004	0.30	0.47748	3.57	0.0677

OUT PART. CORR

---

none

STEP #1; R= 0.775; RSQUARE= 0.600

TERM REMOVED: BD

IN

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
1 CONST						
2 OD	0.1384	0.038	0.53	0.53399	13.0	0.0011
4 LTD	-0.0553	0.139	-0.05	0.87021	0.16	0.6938
5 W	0.0077	0.004	0.29	0.50605	3.69	0.0629

OUT PART. CORR

---

3 BD	-0.035	N/A	N/A	0.42462	0.04	0.8402

STEP #2; R= 0.773; RSQUARE= 0.598

TERM REMOVED: LTD

IN

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
1 CONST						
2 OD	0.1392	0.038	0.53	0.53564	13.0	0.0009
5 W	0.0081	0.004	0.31	0.53564	4.41	0.0431

OUT	PART. CORR					
3 BD	-0.025	N/A	N/A	0.43420	0.02	0.8861
4 LTD	-0.068	N/A	N/A	0.87021	0.16	0.6938

THE SUBSET MODEL INCLUDES THE FOLLOWING PREDICTORS:

CONSTANT
D
W

PARAMETERIZATION OF FINAL MODEL

DEP VAR: RESPONSE N: 38; MULTIPLE R: 0.773; SQUARED MULTIPLE R: 0.598; ADJUSTED SQUARED MULTIPLE R: 0.575;STANDARD ERROR OF ESTIMATE: 0.139

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
CONST	2.873	0.097	0.00	N/A	30.0	0.0000
OD	0.139	0.038	0.53	0.53564	3.62	0.0009
W	0.008	0.004	0.31	0.53564	2.10	0.0431

ANALYSIS OF VARIANCE

SOURCE	SS	DF	MS	F	P
REGRESSION	1.012	2	0.51	26.06	0.0000
RESIDUAL	0.679	35	0.02		

Appendix 2. A) Transformed (residual) kelp forest example data, B) residual transformation equations, and C) SYSTAT output for backwards stepwise sequential regression. *OD* = original wave orbital displacement (in meters); *W'* = wind velocity (in meters/s) after removing variability shared with *OD*; *LTD'* = average tidal height (in meters) after removing variability shared with *OD*, and *W'*; *BD'* = wave breaking depth (in meters) after removing variability shared with *OD*, *W'*, and *LTD'*. Transformed data were created by applying residual transformation equations (B) to original explanatory variables (in Appendix 1A). Backwards stepwise multiple regression was then run on the transformed data using MGLH:REGRESSION under the STATS menu in SYSTAT 5.2 for Macintosh. *Response* was the dependent variable (from Appendix 1A); *CONSTANT, OD, W', LTD'*, and *BD'* were the dependent variables; Stepwise was set to *custom*, with a *backwards* step order, and *P = 0.15* to remove. Only standard SYSTAT output is presented.

A)

OD	W'	LTD'	BD'
2.0176	-2.3069	-0.3345	0.8405
1.9553	6.9168	-0.4360	0.2463
1.8131	-1.7159	-0.1331	-0.4061
2.5751	-5.1990	0.1102	-0.9968
2.2589	5.7518	0.3348	-0.8777
2.5448	2.3071	0.1291	0.2540
2.6291	2.9337	0.1886	1.5515
3.1553	17.5546	0.0298	0.6901
3.4030	5.6697	0.0547	1.0574
2.8150	-8.1308	-0.0648	2.4984
1.9012	-2.1152	-0.0204	1.2344
2.1463	-3.6823	0.0541	-0.4803
2.5851	-6.067	-0.2150	-1.1031
2.0830	-0.6518	0.0107	-0.7307
1.7192	1.0228	0.1697	1.5349
3.5471	-5.1104	0.1606	-0.9351
3.6720	3.7400	0.0582	0.6741
4.7259	3.9715	-0.0754	0.4077
3.6039	5.1032	-0.2968	1.8369
4.1214	-4.2168	-0.0880	-0.7307
3.4940	-1.4492	0.0421	-1.0807
3.4829	-12.8737	-0.0658	0.3931
2.0793	-2.3266	-0.1608	-0.4776
2.0315	-2.7015	-0.1163	-0.6696
1.7356	-0.9888	-0.1729	-0.3671
1.4569	1.3069	-0.0046	0.1044
1.8559	-1.5071	0.2310	0.0242
2.6173	2.3140	0.0336	-1.5433
2.8782	3.5394	0.0380	0.5367

2.7842	14.3787	0.0955	-0.8858
3.3236	-10.6902	0.1792	2.1056
2.8799	-7.7722	0.1816	-1.0617
1.9654	-2.8519	0.1786	0.0667
1.5116	0.2348	-0.0684	0.1544
1.7465	-0.5629	-0.0894	-0.1825
1.0967	3.6569	0.1792	0.7131
2.9802	-0.6544	0.0459	-1.9251
3.0644	3.1728	-0.1630	-2.4704

B)

TRANSFORMATION EQUATION 1:

W regressed against *OD* USING SIMPLE LINEAR REGRESSION

DEP VAR: W N: 38 MULTIPLE R: 0.681; SQUARED MULTIPLE R: 0.464; ADJUSTED SQUARED MULTIPLE R: 0.449; STANDARD ERROR OF ESTIMATE: 6.032

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
CONST	-15.5166	3.297	0.00	N/A	-4.71	0.0000
OD	6.8019	1.218	0.68	1.0	5.59	0.0000

ANALYSIS OF VARIANCE

SOURCE	SS	DF	MS	F	P
REGRESSION	1135.46	1	1135.46	31.02	0.0000
RESIDUAL	1309.77	36	36.38		

W' = *W* +15.5166 + (-6.8019 * *OD*)

TRANSFORMATION EQUATION 2:

LTD regressed against *OD* and *W'* USING SIMPLE LINEAR REGRESSION

DEP VAR: LTD N: 38; MULTIPLE R: 0.360; SQUARED MULTIPLE R: 0.130; ADJUSTED SQUARED MULTIPLE R: 0.080; STANDARD ERROR OF ESTIMATE: 0.177

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
CONST	-14.6982	0.094	0.00	N/A	-1.57	0.1249
OD	-0.06152	0.035	-0.28	1.0	-1.78	0.0835
W'	-0.00676	0.008	-0.23	1.0	-1.43	0.1614

ANALYSIS OF VARIANCE

SOURCE	SS	DF	MS	F	P
REGRESSION	0.15	2	0.08	2.61	0.0877
RESIDUAL	1.02	35	0.03		

LTD' = *LTD* +14.6982 + (0.06152 * *OD*) +(0.00676 * *W'*)

TRANSFORMATION EQUATION 3:

BD regressed against *OD, W'*, AND *LTD'* USING SIMPLE LINEAR REGRESSION

DEP VAR: BD N: 38; MULTIPLE R: 0.759; SQUARED MULTIPLE R: 0.575; ADJUSTED SQUARED MULTIPLE R: 0.538; STANDARD ERROR OF ESTIMATE: 1.161

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
CONST	0.73579	0.634	0.00	N/A	1.16	0.2542
OD	1.52702	0.234	0.73	1.0	6.52	0.0000
W'	0.05388	0.032	0.19	1.0	1.68	0.1021
LTD'	-1.00787	1.147	-0.09	1.0	-0.88	0.3856

ANALYSIS OF VARIANCE

SOURCE	SS	DF	MS	F	P
REGRESSION	62.07	3	20.69	15.36	0.0000
RESIDUAL	45.81	34	1.35		

BD' = *BD* – 0.73579 + (-1.52702 * *OD*) +(-0.05388 * *W'*) +(1.00787 * *LTD'*)

C)

DEPENDENT VARIABLE: RESPONSE

MINIMUM TOLERANCE FOR ENTRY INTO MODEL = 0.010000

STEP #0; R= 0.775; RSQUARE= 0.601

IN

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
1 CONST						
2 OD	0.1942	0.029	7.40	1.0	45.0	0.0000
3 W'	0.0081	0.004	0.22	1.0	4.18	0.0489
4 LTD'	-0.0553	0.141	-0.04	1.0	0.15	0.6980
5 BD'	0.0043	0.021	-0.02	1.0	0.04	0.8403

OUT PART. CORR

---

none

STEP #1; R= 0.775; RSQUARE= 0.600

TERM REMOVED: BD'

IN

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
1 CONST						
2 OD	0.1942	0.0284	7.40	1.0	47.0	0.0000
3 W'	0.0081	0.0039	0.22	1.0	4.30	0.0457
4 LTD'	-0.0553	0.1393	-0.04	1.0	0.16	0.6938

OUT	PART. CORR					
5 BD'	-0.035	N/A	N/A	0.42462	0.04	0.8403

STEP #2; R= 0.773; RSQUARE= 0.598

TERM REMOVED: LTD'

IN

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
1 CONST						
2 OD	0.1941	0.0284	7.40	1.0	48.0	0.0000
3 W'	0.0081	0.0038	0.22	1.0	4.41	0.0431

OUT	PART. CORR					
4 LTD'	-0.068	N/A	N/A	1.0	0.16	0.6938
5 BD'	-0.035	N/A	N/A	1.0	0.04	0.8383

THE SUBSET MODEL INCLUDES THE FOLLOWING PREDICTORS:

CONSTANT
OD
W'

PARAMETERIZATION OF FINAL MODEL

DEP VAR: RESPONSE N: 38; MULTIPLE R: 0.773; SQUARED MULTIPLE R: 0.598; ADJUSTED SQUARED MULTIPLE R: 0.575; STANDARD ERROR OF ESTIMATE: 0.139

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P

CONST	2.748	0.076	0.00	N/A	36.0	0.0000
OD	0.194	0.028	0.74	1.0	6.91	0.0000
W’	0.008	0.004	0.22	1.0	2.09	0.0431

ANALYSIS OF VARIANCE

SOURCE	SS	DF	MS	F	P
REGRESSION	1.012	2	0.51	26.06	0.0000
RESIDUAL	0.679	35	0.02		

Appendix 3. A) Transformed (PC scores) kelp forest example data, B) principal components analysis (PCA), and C) SYSTAT output for backwards stepwise principal components regression. *PC1-4* are the saved principal components scores after the PCA (B) was done on the original explanatory variables (in Appendix 1A). The PCA was run using FACTOR:PRINCIPAL COMPONENTS under the STATS menu in SYSTAT 5.2 for Macintosh. The PCA utilized the correlation matrix with no factor rotations. Backwards stepwise multiple regression was then run on the transformed data using MGLH:REGRESSION under the STATS menu in SYSTAT 5.2 for Macintosh. *Response* was the dependent variable (from Appendix 1A); *CONSTANT, PC1, PC2, PC3,* and *PC4* were the dependent variables; Stepwise was set to *custom*, with a *backwards* step order, and *P = 0.15* to remove. Only standard SYSTAT output is presented.

A)

PC1	PC2	PC3	PC4
-0.1194	-1.9529	-1.0884	0.4859
0.3886	-2.7882	0.2971	0.9228
-0.8321	-1.0240	-0.0552	-0.1100
-0.6746	0.6035	0.0311	-1.0885
-0.6464	1.6024	1.6515	0.2333
-0.0340	0.7130	0.3070	0.4513
0.2707	1.2200	-0.2785	1.3548
1.7626	0.1170	1.8009	1.8021
1.3558	0.5953	0.0933	0.8514
0.3477	0.1418	-2.5273	0.7823
-0.5901	-0.1526	-0.8996	0.9476
-0.8651	0.1788	-0.0810	-0.4235
-0.2864	-1.2485	-0.4044	-1.3741
-0.7414	-0.1732	0.3983	-0.3001
-0.8016	0.8536	-0.4247	1.5977
0.3207	1.2078	-0.0484	-1.4851
1.4487	0.6912	0.0356	0.2927
2.7304	0.2294	-0.0694	-0.4223
2.1808	-1.2472	-0.8526	1.0804
1.3784	-0.0214	-0.4114	-1.6429
0.6181	0.4307	0.3622	-1.2696
0.3647	0.1887	-2.0314	-1.3717
-0.5576	-1.0948	-0.1569	-0.3522
-0.7316	-0.8714	-0.0388	-0.4732
-0.8090	-1.2850	-0.0229	0.0035
-1.1110	-0.3982	0.2358	0.7282
-1.1921	1.1163	0.1555	0.3211
-0.1816	-0.0200	1.1906	-0.8164
0.5818	0.3096	0.1576	0.5554
0.7670	0.2465	2.4016	0.6682

0.3259	1.6985	-2.4039	0.1415
-0.6122	1.1473	-0.2100	-1.4843
-1.0737	0.8814	-0.1131	0.1529
-1.0179	-0.7206	-0.0095	0.6115
-0.8507	-0.7902	0.0251	0.1936
-1.4915	0.5653	0.4526	1.5942
-0.0616	0.1748	1.0000	-1.5106
0.4394	-1.1248	1.5318	-1.6479

B)

PRINCIPAL COMPONENTS ANALYSIS OF *OD*, *BD*, *LTD*, AND *W*.

LATENT ROOTS (EIGENVALUES)

PC1	PC 2	PC3	PC4
2.56537	0.80552	0.37082	0.25829

COMPONENT LOADINGS

VARIABLE	PC1	PC 2	PC3	PC4
OD	0.87771	0.26037	-0.09692	-0.39044
BD	0.87347	0.16099	-0.35373	0.29328
LTD	-0.54211	0.83786	0.04084	0.04939
W	0.85917	0.09901	0.484399	0.13188

VARIANCE EXPLAINED BY COMPONENTS

PC1	PC 2	PC3	PC4
2.56537	0.80552	0.37082	0.25829

PERCENT OF TOTAL VARIANCE EXPLAINED

PC1	PC 2	PC3	PC4
64.1342	20.1380	9.2706	6.4572

C)

DEPENDENT VARIABLE: RESPONSE

MINIMUM TOLERANCE FOR ENTRY INTO MODEL = 0.010000

STEP #0; R= 0.775; RSQUARE= 0.601

IN

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
1 CONST						
2 PC1	0.1571	0.0235	0.73	1.0	45.0	0.0000
3 PC2	0.0267	0.0235	0.12	1.0	1.29	0.2651
4 PC3	0.0219	0.0235	0.10	1.0	0.87	0.3564
5 PC4	-0.0398	0.0235	-0.19	1.0	2.86	0.1003

OUT	PART. CORR
none	

STEP #1; R= 0.768; RSQUARE= 0.590

TERM REMOVED: PC3

IN

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
1 CONST						
2 PC1	0.1571	0.0235	0.73	1.0	45.0	0.0000
3 PC2	0.0267	0.0235	0.12	1.0	1.29	0.2640
5 PC4	-0.0398	0.0235	-0.19	1.0	2.87	0.0994

OUT	PART. CORR					
4 PC3	0.161	N/A	N/A	1.0	0.87	0.3564

STEP #2; R= 0.758; RSQUARE= 0.574

TERM REMOVED: PC2

IN

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
1 CONST						
2 PC1	0.1571	0.0235	0.73	1.0	44.0	0.0000
5 PC4	-0.0398	0.0236	-0.19	1.0	2.85	0.1005

OUT	PART. CORR					
3 PC2	0.19	N/A	N/A	1.0	1.29	0.2640
4 PC3	0.158	N/A	N/A	1.0	0.87	0.3582

THE SUBSET MODEL INCLUDES THE FOLLOWING PREDICTORS:

CONSTANT

PC1
PC4

PARAMETERIZATION OF FINAL MODEL

DEP VAR: RESPONSE N: 38; MULTIPLE R: 0.758 SQUARED MULTIPLE R: 0.574;
ADJUSTED SQUARED MULTIPLE R: 0.550 STANDARD ERROR OF ESTIMATE: 0.143

VARIABLE	COEFF	SE	STD COEF	TOLER	F	P
CONST	3.2498	0.0233	0.00	N/A	140.0	0.0000
PC1	0.1571	0.024	0.74	1.0	6.65	0.0000
PC4	-0.0398	0.024	-0.19	1.0	-1.69	0.1005

ANALYSIS OF VARIANCE

SOURCE	SS	DF	MS	F	P
REGRESSION	0.97	2	0.49	23.6	0.0000
RESIDUAL	0.72	1	35	0.02	

Appendix 4. A) Protocol for creating EQS structural equation models, B) EQS program file, 3) EQS output (regression coefficients) for structural equation model on the original explanatory variables (in Appendix 1A). Only truncated EQS output is presented.

A)

1. CREATE PROGRAM FILE BY USING *TITLE/SPECIFICATIONS* UNDER *BUILD EQS* MENU.

a. USE DEFAULT SETTINGS WITH: NORMAL THEORY ESTIMATORS SET TO *ML* (MAXIMUM LIKELIHOOD), VARIABLES SET TO 5, AND CASES SET TO 38.

2. CREATE EQUATIONS USING *EQUATIONS* UNDER *BUILD EQS* MENU.

a. SET *NUMBER OF VARIABLES* TO 5; SET *NUMBER OF FACTORS* TO 2

b. BUILD THE FOLLOWING *CREATE EQUATIONS* BOX:

	F1	F2	Response	OD	BD	LTD	W	E or D
Response	*							1
OD								1
BD								1
LTD								1
W								1
F1		*				*	*	1
F2				*	*			1

c. USE DEFAULT *CREATE VARIANCE/COVARIANCE* BOX

3. RUN EQS

B)

RESULTING EQS PROGRAM file

```
1       /TITLE
2       EQS model created by EQS 6 for Windows -- c:\eqs6\kelp.ess
3       /SPECIFICATIONS
4       DATA='c:\eqs6\kelp.ess';
5       VARIABLES=5; CASES=38; GROUPS=1;
6       METHODS=ML;
7       MATRIX=RAW;
8       ANALYSIS=COVARIANCE;
9       /LABELS
10      V1=Response; V2=OD; V3=BD; V4=LTD; V5=W;
```

```
11      /EQUATIONS
12      V1 =  + *F1  + 1E1;
13      F1 =  + *F2  + *V4 + *V5 + 1D1;
14      F2 =  + *V2 + *V3 + 1D2;
15      /VARIANCES
16      V2 = *;
17      V3 = *;
18      V4 = *;
19      V5 = *;
20      E1 = *;
21      D1 = *;
22      D2 = *;
23      /COVARIANCES
24      /PRINT
25      FIT=ALL;
26      TABLE=EQUATION;
27      /TECHNICAL
28      ITERATION= 100;
29      EITERATION= 100;
30      AITERATION= 100;
31      HITERATION= 100;
32      /END
```

C)

MAXIMUM LIKELIHOOD SOLUTION (NORMAL DISTRIBUTION THEORY)

EQS STANDARDIZED SOLUTION:

RESPONSE =V1 = 0.860*F1 + 0.510 E1
F1 = 0.917*F2 - 0.065*LTD + 0.393*W + 0.000 D1
F2 = -0.777*OD + 0.049*BD + 0.627 D2

EQS ONLY OUTPUTS STANDARDIZED SOLUTIONS. IN ORDER TO COMPARE AMONG STANDARD, SEQUENTIAL, AND PC REGRESSION RESULTS, EQS STANDARDIZED REGRESSION COEFFICIENTS WERE TRANSFORMED TO UN-STANDARDIZED

VALUES .

AN UN-STANDARDIZED REGRESSION COEFFICIENT (β) EQUALED THE STANDARDIZED REGRESSION COEFFICIENT (β') MINUS THE AVERAGE VALUE OF THE VARIABLE (AVE[X]) DIVIDED BY THE STANDARD DEVIATION OF THE VARIABLE (SD[X]):

$$\beta = \left[\frac{\beta' - AVE(X)}{SD(X)} \right]$$

YIELDING THE FOLLOWING EQS UN-STANDARDIZED SOLUTION

RESPONSE =V1 = 0.69*F1
F1 = 1.23*F2 - 0.08*LTD + 0.01*W
F2 = -0.20*OD + 0.01*BD