ANIMAL BEHAVIOUR, 2005, **70**, 967–971
doi:10.1016/j.anbehav.2005.01.016

Available online at www.sciencedirect.com

SCIENCE DIRECT®

COMMENTARY

The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies

LEIF ENGQVIST
Institute of Evolutionary Biology and Ecology, University of Bonn

(Received 24 November 2004; initial acceptance 31 December 2004; final acceptance 14 February 2005; published online 29 August 2005; MS. number: SC-1272)

In behavioural and evolutionary ecology, there are often large phenotypic differences between individuals in, for example, body size or large variation in abiotic conditions such as temperature, between measurements. This often inevitable source of variation may mask any effect of experimental treatment as it can have a large impact on the dependent variable of interest. In such cases, conventional statistical comparisons may have much lower power than desired. The inclusion of covariates in statistical analyses has proven a powerful method to control for such nonrandom differences between individual data points that cannot be controlled experimentally (Huitema 1980). To make correct conclusions, it is important to understand the basic assumptions underlying such a covariate analysis. In this paper I argue that this has evidently not been completely acknowledged in the scientific community. Sophisticated models relating responses to both one or more continuous covariates and one or more factors can be problematic. Factor is here used in the meaning of a categorical independent variable and its value divides individuals into discrete groups or categories, for instance experimental treatments. In the following, I use a simple one-factor ANCOVA design as an example, but the same general problem outlined here applies to all linear models with one or more covariates, including generalized linear models (GLIM) such as logistic regressions and even survival analysis.

The basic design of a one-factor fixed effect ANCOVA can be written as:

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \overline{X}) + \epsilon_{ij}$$

where Y_{ij} denotes the values for the dependent response variable of the *j*th subject in the *i*th category of the factor, μ is the mean intercept (the average value of the response parameter when the value of the covariate equals zero), α_i is the response to the *i*th category of the factor, X_{ij} the value for the covariate of the *j*th subject in the *i*th category of the factor, $\overline{X}$ is the mean value of the covariate for all individuals, β is the overall pooled regression coefficient (slope) within groups and ϵ the normally distributed error variance (cf. Huitema 1980). When performing an ANCOVA, we thus assume equivalent slopes among treatment groups (β). The test of homogeneity among slopes is therefore a key prerequisite to proceed to the ANCOVA itself. The easiest way to test this assumption is to include the interaction term between the covariate and the factor in the model. If the interaction term is nonsignificant, we can conclude that the slopes are homogeneous and then proceed to test whether the response differs between groups. This is formally done by testing for differences between treatment groups in the *Y* intercept for the regressions of the covariate on the response variable *Y*. This test is carried out by re-running the model, excluding the interaction term. Differences in the *Y* intercept (i.e. differences between groups when the value for the covariate equals zero) are generally of minor importance. However, since we assume homogeneity of slopes this difference can be inferred over the whole range of covariate values.

A significant interaction effect, on the other hand, indicates that the relation between the covariate and the response variable *Y* differs between groups. In such a case with heterogeneous slopes, the difference between groups will depend on the value of the covariate. To continue and perform an ANCOVA is therefore inappropriate in these cases. However, this first full model will of course also perform a significance test for the response to the factor. This test is not a test of the average response to the factor: it is solely a test for differences in the *Y* intercept. The effect of the factor is not fixed but conditional and will depend on the value of the covariate. Instead, after we have established that the slope coefficients are nonhomogeneous, we can continue in several different ways. We

Correspondence: L. Engqvist, Institut für Evolutionsbiologie und Ökologie, Rheinische Friedrich-Wilhelms-Universität Bonn, An der Immenburg 1, D-53121 Bonn, Germany (email: lengqvist@evolution.uni-bonn.de).

0003–3472/05/$30.00/0

can for instance (1) conclude that the response to the covariate is different between groups, which in a broader sense means that there are significant differences between groups (Cochran 1957), (2) perform a separate analysis for each group (which is rational only if the main interest is the response to the covariate and not the response to the factor), and (3) determine the regions of significance using the Johnson–Neyman procedure (Johnson & Neyman 1936; Huitema 1980).

Similarly, if we choose to include a nonsignificant interaction term in our final model, this would mean that we none the less assumed different slopes. A test for differences in Y intercepts between groups (treatment effect) would thus give us information not on the overall or average difference between groups, but only at one point in the covariate dimension, namely when the covariate equals zero.

The inclusion of a nonsignificant interaction term, however, not only limits the possibility of generalizing intercept differences; in most analyses it also eliminates the possibility of detecting even substantial differences in Y intercepts between groups. As the mean value of the covariate in most cases differs considerably from zero, random deviations of slopes within the confidence interval will be amplified by the extrapolation back to the Y intercept. This will amplify the standard errors of the intercepts (Fig. 1a). The effect will be smaller the steeper the slope and the larger the covariate's coefficient of variation. In contrast, analyses including a significant interaction term can also generate significant differences in intercept, because the nonrandom differences in slopes will be amplified, often generating a negative correlation between slope and intercept (Fig. 1b).

This obviously also means that analyses including covariate interaction terms, significant or not, are sensitive to different scaling of the covariate. For instance, if the covariate is temperature, an analysis including the interaction term will consequently present different F and P values for the effect of the factor depending on whether temperature was measured in degrees Celsius or Kelvin. This is, of course, not the case in an ANCOVA without an interaction term.

To conclude this brief review, which can be extracted from almost any statistical textbook (e.g. Huitema 1980; Sokal & Rohlf 1995; Pedhazur 1997; Goldberg & Scheiner

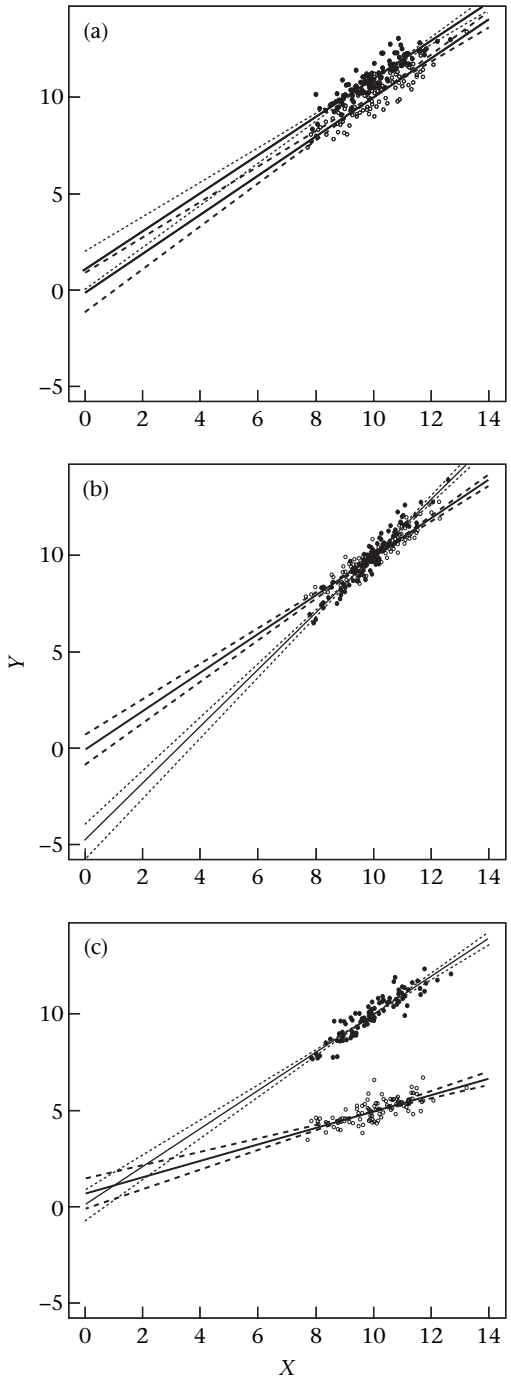

Figure 1. Illustrations of the problems with covariate analyses assuming different slopes. (a) A random sample from two populations ($n_1 = n_2 = 100$) with hypothetically equal slopes ($b_1 = b_2 = 1$) and a hypothetical difference in response (Y) of 1 unit. Hypothetical mean $\pm$ SD of covariate X was 10 ± 1 units and residual SD was 0.4 in this and in the following examples. Lines refer to the mean and 95% confidence limits to the regression estimates of each group. An ANCOVA on these data not assuming equal slopes failed to show significant differences between groups (group: $F_{1,196} = 2.85$, $P = 0.1$; slope: $F_{1,196} = 439.6$, $P < 0.0001$; interaction: $F_{1,196} = 0.105$, $P = 0.7$), whereas the difference is highly significant assuming (correctly) equal slopes (group: $F_{1,197} = 174.2$, $P < 0.0001$; slopes: $F_{1,197} = 829.4$, $P < 0.0001$). An analysis assuming homogeneous slopes has a power of $>99.9\%$ ($\alpha = 0.05$) to refute the null hypothesis of no difference between groups (all 1000 random samples). In contrast, an analysis including the interaction term has an approximate power of 0.056. (b) A random sample from two populations with hypothetically different slopes ($b_1 = 1$, $b_2 = 1.5$) and a hypothetical response difference of 0 units at the mean value of the covariate $X = 10$. An analysis including the interaction correctly revealed a significant difference in the Y intercept between groups (group: $F_{1,196} = 63.5$, $P < 0.0001$; slope: $F_{1,196} = 624.5$, $P < 0.0001$; interaction: $F_{1,196} = 66.5$, $P < 0.0001$). (c) A random sample from two populations with hypothetically different slopes ($b_1 = 0.5$, $b_2 = 1$) and a hypothetical response difference of 5 units at the mean value of the covariate $X = 10$. An analysis including the interaction correctly revealed no significant difference in the Y intercept between groups (group: $F_{1,196} = 0.974$, $P = 0.3$; slope: $F_{1,196} = 118.8$, $P < 0.0001$; interaction: $F_{1,196} = 97.1$, $P < 0.0001$).

2001; Quinn & Keough 2002): (1) nonsignificant covariate interaction terms must be removed before re-running the final analysis; (2) when there are significant interaction terms, which should not be removed, it is incorrect to interpret the response to the factor as an overall or average main effect.

None the less, these violations of the ANCOVA assumptions are recurrently seen in the literature. My aim in this paper was to find out if these are occasional mistakes or a more general problem. I chose to analyse research literature on behavioural and evolutionary ecology simply because this is the focus of my own research.

Methods

I scanned all papers in *Animal Behaviour*, *Behavioral Ecology* and *Journal of Evolutionary Biology* published between July 2003 and June 2004. I disregarded reviews, comments and theoretical models. Special emphasis was placed on the Material and Methods sections in search of a description of the statistical analyses used. I scrutinized the Results sections looking for words or test statistics indicating that a statistical model with at least one covariate and at least one factor was used. I also inspected all tables and figures. I classified two potential mistakes: (1) nonsignificant interaction terms between a covariate and a factor indicating homogeneity of slope were not removed before the final analysis; and (2) the results of models including significant interaction terms were inadequately interpreted with regard to the factor (usually the experimental treatment).

In papers reporting significant interaction terms, I therefore looked in more detail, searching for an interpretation of the test. Formulations such as '[…] overall, treatments differed significantly […]', '[…] when controlling for body size, treatment had a significant effect […]', '[…] treatment had an effect, furthermore slopes differed significantly […]', or with similar wording was classified as misinterpretations. If used in context with the interaction test only, statements such as 'treatment had an effect' were classified as acceptable, because different slopes between treatments can be viewed as a treatment effect (Cochran 1957): the response to the covariate is significantly different between treatments.

The inclusion of a nonsignificant interaction term in analyses was, if not explicitly stated in the text, generally deduced from the *df* of the residual deviance. In the simplest case, both the interaction term *df* and the factor *df* were given. Alternatively, I compared the factor *df* with the sample size. I made the conservative assumption that the analysis was correct if it was not possible to reconstruct which model was used in the final analysis, because either the factor *df* or the sample size was not stated.

Results

I investigated 457 empirical papers in *Animal Behaviour* (231), *Behavioral Ecology* (115) and *Journal of Evolutionary Biology* (111) published between July 2003 and June 2004. In total I found erroneous analyses in 28 of them (6.1%). Eighty papers (17.5% of all papers) used at least one analysis including a continuous covariate and a categorical factor. Thus, 35% of all covariate-type analyses were either misinterpreted or misapplied.

Thirteen articles comprised misinterpretations of the significance tests for the response to the factor in analyses with a significant covariate/factor interaction term, and in 21 articles I found analyses with an inadequate inclusion of a nonsignificant interaction term in the final model. In six articles both mistakes were made. Thus, in relation to total number of articles, these mistakes were approximately equally frequent but for different reasons. Whereas relatively few articles with covariate analyses had to handle significant interactions (28), these analyses were often misinterpreted (46.4%). On the other hand, most articles with covariate analyses involved nonsignificant interactions (67). Hence, in roughly one-third (31.3%) of these articles analyses were mistreated, as the interaction term was not removed in the final model.

To illustrate and underline that this is a nontrivial problem, I compared the different articles regarding the number reporting a significant response to the factor (e.g. treatment). Articles that neglected to remove a nonsignificant interaction term attained considerably fewer significant results than studies in which these terms were removed before the final analysis (see Fig. 2 for more details and cf. also Fig. 1a).

Covariate interaction terms can be mistreated in different kinds of statistical methods. I did not attempt to compare quantitatively between methods, because the sample size was rather small. However, I found mistakes in all forms of analyses, including the simplest form of ANCOVA and more complex GLM, repeated measures ANCOVA, logistic regressions and other GLIM and also survival analyses.

Discussion

Statistical analyses incorporating covariates can be a powerful tool when comparing groups with large within-group variance (Fisher 1932; Huitema 1980). Especially in ecological studies, where we often experience large phenotypic variance in traits potentially influencing the study variable, and time-consuming sampling often prevents us obtaining adequate sample sizes, this technique is often indispensable to eliminate irrelevant variance and achieve satisfactory statistical power. My aim in this short review was not to challenge this view. The use of covariate analyses, such as ANCOVA, GLM and GLIM, is increasing in behavioural and evolutionary ecology studies, and basically this is a positive development. However, it is worrying that so many analyses are flawed because of fairly simple mistakes. Of 80 studies using covariate analysis and published recently in three journals with large impact, at least 28 were inadequately analysed or interpreted. This is a substantial quantity and it is probably an underestimate. In seven additional studies, which I classified as correctly analysed, as there were no obvious inaccuracies, it was not possible to conclude from the article whether nonsignificant interactions were

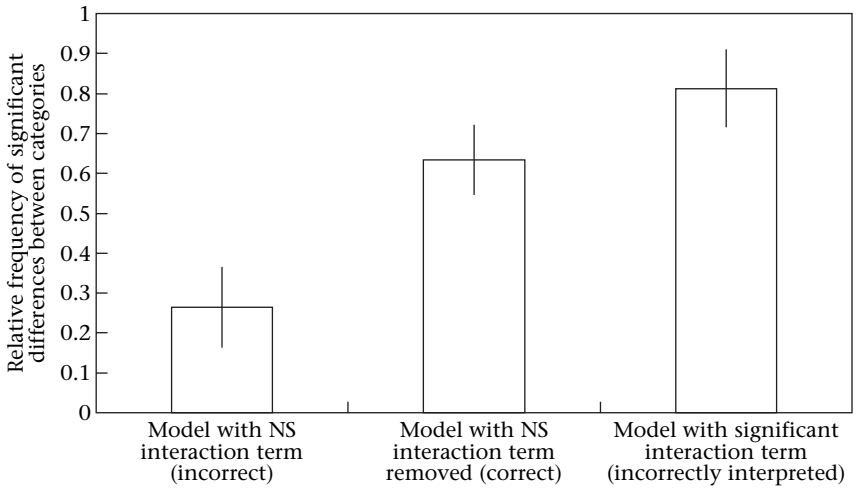

Figure 2. Comparison of different covariate models with respect to the relative frequency of significant differences between categorical groups ($\overline{X} \pm$SE). To avoid potential pseudoreplication, only the first analysis in an article was considered. Overall, groups differed significantly regarding frequency of establishing a statistically significant response to categorical factors (log likelihood ratio: $\chi^2_2 = 12.1$, $P = 0.002$; post hoc analysis: nonsignificant interaction term included versus correct analysis: $\chi^2_1 = 6.58$, $P = 0.01$; significant interaction term versus correct analysis $\chi^2_1 = 1.66$, $P = 0.22$). The lack of difference between the correct analyses and analyses with a significant interaction term does not mean that the conclusions in the latter are appropriate, because nonhomogeneity of slopes can both generate an apparent response (cf. Fig. 1b) and conceal apparent differences (cf. Fig. 1c).

removed or not. Furthermore, I focused only on one problematic aspect of covariate analyses. For instance, in another seven articles it was not stated whether homogeneity of slopes was tested at all. The significance level of the homogeneity test, which is not self-evidently 0.05, was also recurrently left out. A P value lower than, for instance, 0.1 still indicates that the slopes are likely to differ. One of the ramifications of assuming slopes are homogeneous when they are in fact not is a conservative effect on the ANCOVA F test on main effects (see Huitema 1980, pp. 102–103). Therefore, if a plot reveals that slopes look heterogeneous, it may be safer to continue with methods appropriate for such cases, which are discussed below. The frequent use of covariates to correct for initial differences (caused by, for instance, nonrandom assignment or pre-existing differences), and thus statistically to 'equate' comparison groups, is also inappropriate (see also Pedhazur & Schmelkin 1991; Quinn & Keough 2002). This clearly violates the assumption of no collinearity, that is, independence, between predictor variables.

The assumption of equivalent slopes among treatment groups is a key prerequisite to any analysis incorporating covariates to remove error variance. However, I have shown that many analyses unfortunately do not subsequently exclude the covariate interaction term, and thus actually assume different slopes between groups. From these analyses it is not possible to make a straightforward estimate of differences between groups. I found that about every fourth study carrying out a covariate analysis with equal slopes failed to remove the nonsignificant interaction term before the final analysis. This is a figure to be concerned about. The remedy, however, is simple: always remove the interaction terms from the model, when slopes for all groups can be assumed to be equal.

When slopes do differ, one cannot remove the interaction term. In these analyses it is incorrect to interpret the test for significant differences between groups as an 'average effect', as almost 50% of the studies in this review, facing this problem, erroneously did. The response, for example, treatment effect, will more exactly be dependent on the value of the covariate. This is comparable to the analogous situations with significant interactions in other models. In a two-factorial ANOVA, for instance, the response to one factor will depend on the value of the other. In such cases, when higher order interactions are significant, it is always inappropriate to draw conclusions on 'overall effects' from the test of lower order main effects. This has been pointed out by several authors over the years (e.g. Aiken & West 1991; Sokal & Rohlf 1995; Quinn & Keough 2002; Ruxton & Colegrave 2003), but seems to continue to be a problem in the behavioural sciences.

In a linear model, such as an ANCOVA, with heterogeneous slopes, the significance test for the treatment effect will test only for differences at the Y intercept, which is usually only of minor interest. Heterogeneous slopes thus present a problem in that it is not possible to test for significance using standard techniques. None the less, such a finding does imply that there are significant treatment effects (Cochran 1957). However, simply stating that the slopes are different is unsatisfactory, because in most cases the elevations and not the slopes are relevant. Most studies that were included in this survey made statements similar to: 'the effect increased with increasing body size' (if body size was used as covariate). However, it will often be of relevance to test post hoc for what values of the covariate there are significant group differences. Modifications of the Johnson–Neyman procedure (Johnson & Neyman 1936) suggested by Huitema (1980) and Wilcox (1987) provide solutions in this situation and are thus a generalization of an ANCOVA, relaxing the assumption of homogeneous slopes across treatments (see

also Quinn & Keough 2002). By means of this technique it is possible to identify regions of significance, or rather nonsignificance, throughout the range of the covariate. However, no study in this survey made use of this procedure. In many analyses that I encountered, I do not doubt that there actually were significant 'overall' treatment effects. However, with heterogeneous slopes it is not possible to conclude this from the initial ANCOVA without additional analyses. In these cases, using the Johnson–Neyman procedure would have made it possible to make statements such as 'there were significant differences between treatments over the whole range of the covariate and treatment effects increased with increasing value of the covariate' or possibly 'treatment effects increased with increasing value of the covariate, and groups were significantly different for values of the covariate above'. These statements give a much more detailed and substantial description than for instance 'the effect increased with increasing value of the covariate'. The use of the Johnson–Neyman technique is hampered by its absence as a standard feature in most commercially available statistical packages (but see Hunka & Leighton 1997; D'Alonzo 2004). However, the calculations for the simple, but very common, one covariate-one category with two groups-case are not so demanding and can be obtained in for instance Huitema (1980), White (2003) and D'Alonzo (2004).

In this short survey I wanted to draw attention to some mistakes related to interaction terms in linear models. I have focused on interactions between covariates and categorical factors in ANCOVA-type analyses, as these seem unsettlingly common. The same problem of course applies to first-order effects in multiple regression when higher order interactions are included (see Aiken & West 1991).

Klaus Reinhold encouraged me to perform this study. He, Jutta Schneider and two anonymous referees also gave helpful criticism on the manuscript. I also thank Deutsche Forschungsgemeinschaft for financial support (LE 469-1/1).

References

Aiken, L. S. & West, S. G. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park: Sage.

Cochran, W. G. 1957. Analysis of covariance: its nature and uses. *Biometrics*, **13**, 261–281.

D'Alonzo, K. T. 2004. The Johnson–Neyman Procedure as an alternative to ANCOVA. *Western Journal of Nursing Research*, **26**, 804–812.

Fisher, R. A. 1932. *Statistical Methods for Research Workers*. 4th edn. Edinburgh: Oliver & Boyd.

Goldberg, D. E. & Scheiner, S. M. 2001. ANOVA and ANCOVA: field competition experiments. In: *Design and Analysis of Ecological Experiments* (Ed. by S. M. Scheiner & J. Gurevitch), pp. 46–67. Oxford: Oxford University Press.

Huitema, B. E. 1980. *The Analysis of Covariance and Alternatives*. New York: J. Wiley.

Hunka, S. & Leighton, J. 1997. Defining Johnson–Neyman regions of significance in the three-covariate ANCOVA using mathematica. *Journal of Educational and Behavioral Statistics*, **22**, 361–387.

Johnson, P. O. & Neyman, J. 1936. Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, **1**, 57–93.

Pedhazur, E. J. 1997. *Multiple Regression in Behavioral Research: Explanation and Prediction*. 3rd edn. Fort Worth, Texas: Harcourt Brace.

Pedhazur, E. J. & Schmelkin, L. P. 1991. *Measurement, Design, and Analysis: an Integrated Approach*. Hillsdale, New Jersey: Erlbaum.

Quinn, G. P. & Keough, M. J. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.

Ruxton, G. D. & Colegrave, N. 2003. *Experimental Design for the Life Sciences*. Oxford: Oxford University Press.

Sokal, R. R. & Rohlf, F. J. 1995. *Biometry*. 3rd edn. New York: W.H. Freeman.

White, C. R. 2003. Allometric analysis beyond heterogenous regression slopes: use of the Johnson–Neyman technique in comparative biology. *Physiological and Biochemical Zoology*, **76**, 135–140.

Wilcox, R. R. 1987. Pairwise comparisons of J independent regression lines over a finite interval, simultaneous pairwise comparison of their parameters, and the Johnson–Neyman procedure. *British Journal of Mathematical and Statistical Psychology*, **40**, 80–93.