

# Inter-species regression analysis

Brigitte Pallmann<sup>a</sup>, A. D. Barbour<sup>a</sup>, D. J. Hosken<sup>b</sup> and P. I. Ward<sup>b</sup>

Universität Zürich

**Abstract.** When conducting inter-species regression analyses, the phylogenetic relationships between the individual species need to be taken into account. In this paper, a procedure for conducting such analyses is proposed, which only requires the use of a measure of relationship between pairs of species, rather than a complete phylogeny, and which at the same time assesses the importance to be attached to the relationships with regard to the conclusions reached. The procedure is applied to data from Minder (2002), relating testis size to mean hind tibia length, duct length and spermathecal area in 15 species of Scathophagidae.

**Keywords.** Inter-species regression, phylogenetic relationship, Ornstein-Uhlenbeck process, likelihood based inference.

## 1. Introduction

Comparative studies are a widely employed and powerful tool in evolutionary investigations. They have been used to elucidate macro-evolutionary patterns for many phenomena, including testis and sperm size evolution (e.g. Gage 1994; Hosken 1997), brain size evolution (e.g. Martin 1981; Pagel & Harvey 1989) and the scaling of metabolic rates (e.g. Thompson & Withers 1998; McNab 2002). Formerly, species values of characters of interest were regressed against putative predictor variables to elucidate possible relationships (e.g. Cummins & Woodall 1985). However, because of common descent, species do not represent independent data points, and hence species level analyses based on simple regression analyses have been criticised (Harvey & Pagel 1991). This is not to say that phylogeny has primacy of cause over other factors, merely that species level analyses

---

<sup>a</sup> Angewandte Mathematik, Winterthurerstrasse 190, CH-8057 ZÜRICH, Switzerland

<sup>b</sup> Zoologisches Museum, Winterthurerstrasse 190, CH-8057 ZÜRICH, Switzerland

may be misleading (Harvey 2000). For example, a simulation study by Martins & Garland (1991) investigated the across-species association between two variables, and found that the Type I error rate was 16%; when they employed phylogenetic control, the error in the regression was reduced to 5%. This paper, based on Pallmann (2001), presents a procedure for conducting regression analyses in the presence of phylogenetic relationships, which also assesses the importance of these relationships in the analysis. The procedure is based on a simple model of the way in which inter-species differences evolve.

In order to derive our procedure, we begin with a more detailed exposition of the underlying problem. In the classical regression model, the value  $y$  of the ‘dependent’ variable of interest is expressed linearly in terms of the values  $x^{(1)}, x^{(2)}, \dots$  of a number of explanatory ‘covariables’, up to an additive ‘error’  $e$ , which accounts for any variation in  $y$  not attributable to the covariables. Thus, for each of  $n$  observations indexed by  $i$ ,  $1 \leq i \leq n$ , we write

$$y_i = \beta^{(0)} + \beta^{(1)}x_i^{(1)} + \beta^{(2)}x_i^{(2)} + \dots + \beta^{(k)}x_i^{(k)} + e_i, \quad (1)$$

where the  $\beta^{(j)}$ ,  $0 \leq j \leq k$ , are the coefficients which relate the values of the covariables to that of  $y$ , and the  $e_i$  are needed because, in practice, it is usually impossible to find values  $\beta^{(0)}, \beta^{(1)}, \dots, \beta^{(k)}$  such that

$$y_i = \beta^{(0)} + \beta^{(1)}x_i^{(1)} + \beta^{(2)}x_i^{(2)} + \dots + \beta^{(k)}x_i^{(k)} \quad (2)$$

is *exactly* true for all  $i$ , if  $n \geq k + 2$ . The values of the parameters  $\beta^{(j)}$  are estimated and their significance tested with reference to some probabilistic model, which is assumed to govern the values  $e_i$  of the errors actually occurring; the simplest assumption is that the  $e_i$ ,  $1 \leq i \leq n$ , arise as realizations of *independent* random variables  $\varepsilon_i$ ,  $1 \leq i \leq n$ , which have zero mean and common unknown variance  $\sigma^2$ , and are normally distributed.

If the indices  $1 \leq i \leq n$  in fact represent  $n$  species, as in the setting introduced above, and if the measured values  $y_i$  and  $x_i^{(1)}, \dots, x_i^{(k)}$  are ‘typical’ values for the species, the assumption of independent errors may well be violated. This is because the variation in  $y$  unexplained by the  $x^{(j)}$  can be thought of in part as resulting from evolutionary drift in other, unobserved explanatory covariables, so that closely related species can be expected to exhibit rather similar values of  $e$ . Hence, when conducting regression analyses with such data, it seems important to take the phylogeny into account (Harvey & Pagel, 1991).

The reasons for doing so are quite simple. When the  $\varepsilon_i$  are in fact correlated, ordinary least squares (OLS) procedures still give parameter estimates which have the right

expectation and are asymptotically consistent. However, as is especially relevant when the number of species is fixed and perhaps not large, their precision is less than that of the best estimates possible for the actual correlation structure [Draper & Smith (1966, p. 80)]. Moreover, estimates of the precision of the OLS estimates, calculated in accordance with OLS assumptions, may seriously overestimate their actual precision [see, for example, Brockwell & Davis (1987, p. 212) for the consequences in the time series context]. In such cases, significance tests based on OLS are dangerous.

There is nonetheless still some debate about the efficacy of such phylogenetic control. This is primarily because the covariance of traits is explained by ecology and phylogeny, which typically overlap; hence, by controlling for phylogeny, variance due to ecology is also removed because of the overlap (McNab 2002). Despite such arguments, most investigators today use some form of phylogenetic control.

A number of methods have been proposed for incorporating phylogeny into the regression, including trait mapping (e.g. Markow 2002), nested analysis of variance (Stearns 1992), pairwise comparison (Møller & Birkhead 1992) and independent contrasts (Felsenstein 1985). These methods all involve the use of a pre-existing phylogeny. Here, we propose a simple and effective procedure, which can either be applied using a known phylogeny, or else just using an inter-species distance matrix from which a phylogeny could potentially be constructed. The procedure has the advantage that it not only respects the phylogeny, but also allows one to gauge its importance in the analysis.

## 2. Procedure

The basic idea is to return to the underlying assumption, that the  $e_i$ 's result from a process of evolution along the branches of the phylogenetic tree. The evolutionary model that we use, which can be thought of as a natural generalization of that of independent, normally distributed errors with common variance, supposes that the 'error' component evolves along each branch of the phylogenetic tree as an Ornstein–Uhlenbeck (O-U) process, as discussed in some detail in Felsenstein (1988). The O-U process looks locally in time like a Brownian motion, as would naturally be the result of many small random genetic changes; however, it also has a drift towards zero, which can be thought of as the result of selective pressure acting against departures from equality in (2), the strength of the drift being larger for larger departures. For our purposes, the main features of the process are that it is a Markov process, that its values are normally distributed, and that its autocorrelations

decay exponentially with elapsed time. Its distributions are entirely characterized by the variance  $\sigma^2$  of its equilibrium normal distribution and the exponential decay rate  $\lambda$  of correlations; we shall denote such a process by  $\text{OU}(\sigma^2, \lambda)$ .

Our model supposes that an  $\text{OU}(\sigma^2, \lambda)$  process starts in equilibrium at the root of the phylogenetic tree, and runs, with time corresponding to distance along the branch, until the first split. At this point, its value is taken as the initial value for two *independent*  $\text{OU}(\sigma^2, \lambda)$  processes, which then continue to run along the two branches until they split again; and so on. The species, the leaves of the trees, are assigned as values of  $e_i$  the values of the  $\text{OU}(\sigma^2, \lambda)$ -processes at the ends of the final  $n$  branches. This model of evolution along the tree results in values  $e_i$  realized from jointly normally distributed random variables  $\varepsilon_1, \dots, \varepsilon_n$  having equal variances  $\sigma^2$ , but now with correlations

$$C_{il}(\lambda) := \text{Corr}(\varepsilon_i, \varepsilon_l) = e^{-\lambda d_{il}}, \quad (3)$$

where  $d_{il}$  is the distance between species  $i$  and species  $l$  along the tree.

The value of the decay rate  $\lambda$ , in combination with the values of the  $d_{il}$ , is seen from (3) to reflect the extent to which the inter-species correlations are important in the analysis. For instance, the correlations become weaker as  $\lambda$  increases, and  $\lambda = \infty$  corresponds to independent errors for each species. The behaviour of the model as  $\lambda \rightarrow 0$  is somewhat more subtle. For fixed  $\sigma^2$ , the extreme value  $\lambda = 0$  would correspond to having such strong dependence that all of the  $e_i$  were equal, which is unlikely in practice, since then (2) would be exactly satisfied (albeit with  $\beta^{(0)} + e_1$  in place of  $\beta^{(0)}$ ). For data which exhibit error in (1), whatever the choice of  $\beta^{(0)}, \beta^{(1)}, \dots, \beta^{(k)}$ , the best fit model with  $\lambda$  given but small has to be consistent with their intrinsic variability, and is thus not close to that with  $\sigma^2$  fixed and  $\lambda = 0$ ; rather, the errors in the fit are of more or less constant magnitude for all small values of  $\lambda$ . Now the locally Brownian behaviour of the  $\text{OU}(\sigma^2, \lambda)$  process has diffusion constant (infinitesimal variance)  $\tau^2 = \tau^2(\sigma^2, \lambda) = 2\lambda\sigma^2$  per unit branch length, and the limit as  $\lambda \rightarrow 0$  which preserves the variability in the data is thus that in which  $\sigma^2 \rightarrow \infty$  in such a way that  $\tau^2(\sigma^2, \lambda)$  tends to some finite limit  $\tau_0^2$ . The corresponding (driftless) process is then Brownian motion, with diffusion coefficient  $\tau_0^2$ . Thus, when fitting the  $\text{OU}(\sigma^2, \lambda)$  model to data, a best fit with large  $\lambda$  indicates more or less independent species, and one with  $\lambda$  very small indicates a Brownian-like model of evolution, a fact noted also by Blomberg *et al.* (2003) (with their parameter  $d$  corresponding to our  $e^{-\lambda}$ ).

Regression analysis based on this model is easy if the phylogenetic tree — specifically, all the tree-distances  $d_{il}$  between pairs of species — are known. Then, for any fixed  $\lambda$ ,

$0 < \lambda < \infty$ , the correlation matrix  $C(\lambda)$  can be calculated, and maximum likelihood for the linear model with normally distributed errors and known correlation matrix can be used to find estimates  $\hat{\beta}^{(0)}(\lambda), \hat{\beta}^{(1)}(\lambda), \dots, \hat{\beta}^{(k)}(\lambda)$  and  $\hat{\sigma}^2(\lambda)$  of the remaining model parameters, together with  $L(\lambda)$ , the maximum value of the log-likelihood for this value of  $\lambda$ . The value  $\hat{\lambda}$  to be used as an estimate of the true value of  $\lambda$  is now obtained by maximizing  $L(\lambda)$  iteratively with respect to  $\lambda$ , for instance using a golden section search; this yields the final parameter estimates

$$\hat{\lambda}, \hat{\beta}^{(0)}(\hat{\lambda}), \dots, \hat{\beta}^{(k)}(\hat{\lambda}), \hat{\sigma}^2(\hat{\lambda})$$

for the regression. The values  $e^{-\hat{\lambda}d_{il}}$  then indicate the strengths of the correlations between pairs of species, and their distribution contains information which may be useful for comparing studies.

For each given  $\lambda$ , the linear model theory also gives the standard deviations to be associated with the parameter estimates  $\hat{\beta}^{(j)}(\lambda)$ ,  $0 \leq j \leq k$ , and these can be used with  $\lambda = \hat{\lambda}$  as reasonable approximations to the standard deviations of the estimates  $\hat{\beta}^{(j)}(\hat{\lambda})$ , and hence for tests of hypotheses. However,  $\hat{\lambda}$  has been chosen from the data, and this source of variability is not included in such ‘plug-in’ approximations; bootstrapping data samples from the model obtained from the estimated parameters, and then using an identical estimation procedure, gives an alternative way of judging the actual precision obtained. If the value of  $\lambda$  is itself of interest, an approximate 95% confidence region based on large sample theory is given by the set of all  $\lambda$  such that

$$L(\hat{\lambda}) - L(\lambda) \leq 2$$

[c.f. Edwards (1972, p. 80)]. This region may include  $\lambda = \infty$ , in which case an analysis that neglects inter-species correlations should still be reliable. Again, bootstrapping data samples from the estimated model gives another measure of the variability in the estimates of  $\lambda$ .

In practice, the phylogenetic tree is rarely known precisely, complete with distances. However, the method proposed here can be expected to give useful results, even when the distances  $d_{il}$  are only approximately known; as long as the correlation structure is reasonably represented, gross errors in the conclusions arising from this source should be avoided. Thus, if any molecular or morphological data for the species are available, on the basis of which a tree can be reconstructed, this can be carried out, and the corresponding

tree distances used for the  $d_{il}$ . Alternatively, this relatively difficult step can be avoided by using the morphological and molecular data to define a measure of distance between pairs of species — in any case, often the starting point for a tree construction — and by then using these distances directly in place of the  $d_{il}$ . Here, there may be difficulty if the distances are not close enough to distances measured along a tree, because the resulting matrices  $C(\lambda)$  need not then be positive semi-definite for all values of  $\lambda$ , as has to be the case for correlation matrices. However, Bochner’s theorem [Defant & Floret (1993, p 316)] implies that  $p$ ’th powers of  $l_2$ -distances, for  $p \leq 2$ , never give rise to this problem, and that the same applies if a distance can be represented as a sum of such distances; thus, for instance, Hamming distance (number of mismatches) can be used for molecular data, and can be added to Euclidean distances between morphometrical characters.

### 3. Example

The procedure is illustrated by application to data in Minder (2002), with a regression of testis size  $y$  as a function of mean hind tibia length (HTL)  $x^{(1)}$ , spermathecal duct length  $x^{(2)}$  and spermathecal area  $x^{(3)}$  in 15 species of Scathophagidae (Diptera; true flies) [Table 1]. In the paper above, a corresponding analysis was made using the comparative analysis by independent contrasts program (CAIC) (Purvis and Rambaut 1994) to correct for the phylogeny, which was deduced from that of Bernasconi *et al.* (2000), itself derived from inter-specific differences in the sequence of 810 mDNA letters coding for the COI gene. Here, we look only at the 15 species considered by Minder (2002), and take the numbers of pairs of mDNA letters differing in species  $i$  and  $j$  as one measure of evolutionary distance  $d_{ij}^{(1)}$  [Table 2]. We also use a second matrix  $d^{(2)}$ , derived from  $d^{(1)}$  by setting

$$d_{ij}^{(2)} = -\log(1 - d_{ij}^{(1)}/105), \quad (4)$$

which is deduced from the assumptions that only about 140 of the sites are actually variable, and that the Jukes–Cantor model of evolution applies to them. Under these circumstances, the expected proportion of letters differing in a pair of species at evolutionary distance  $d$  would be  $q(d) = \frac{3}{4}(1 - e^{-\alpha d})$  if the mutation rate were  $\alpha$  per site per unit distance, giving

$$d_{ij}^{(1)} \approx 140q(d_{ij}) = 105(1 - e^{-\alpha d_{ij}}),$$

and we then standardize the scale of the distances  $d_{ij}$  to give  $d_{ij}^{(2)}$  by setting  $\alpha = 1$ . There is no particular reason to believe that these assumptions are precisely true, but they lead

to an estimate of evolutionary distance which is at least consistent with a genuine, if perhaps over-simplified, evolutionary model; the raw number of mismatches is not itself a natural measure of evolutionary distance. It turns out that there is in any case too little information in the data to justify an attempt at fitting a more sophisticated model for the distances.

The above procedure, with the correlation matrix  $C(\lambda)$  calculated for each given value of  $\lambda$  by substituting the evolutionary distance matrix  $d^{(2)}$  into (3), shows that  $x^{(2)}$  has no appreciable influence on  $y$ , and, leaving out this covariable, leads to a value  $\hat{\lambda} = 0.3676$  maximizing the log-likelihood at a value of 27.24, with parameter estimates

$$\hat{\beta}^{(1)} = 0.2121, \quad \hat{\beta}^{(3)} = 23.18 \quad \text{and} \quad \hat{\sigma} = 0.1798. \quad (5)$$

The value of  $\hat{\lambda}$  is such that the largest correlation (that between *Norellia striolata* and *N. spinimona*, only 9 differing pairs of mDNA letters) is estimated to be 0.821, a moderate correlation (between *Scathophaga sulla* and *S. furcata*, 41 pairs different) is estimated to be 0.336, and a smallish correlation (between *Clegiastra ciliata* and *Microprosopa pallidicauda*, 80 pairs different) is estimated to be 0.042. The standard deviations of  $\hat{\beta}^{(1)}$  and  $\hat{\beta}^{(3)}$ , as calculated from  $C(\hat{\lambda})$  and  $\hat{\sigma}$ , are 0.0947 and 7.84 respectively, and their estimated correlation is  $-0.275$  (cf. Younger (1985, Sections 11.5–11.8)).

The maxima of the log-likelihood for the submodels obtained by omitting either  $x^{(1)}$  or  $x^{(3)}$  are both more than 2 smaller than that obtained above, indicating that, at the 5% level, neither submodel should be preferred to the model estimated above. This means that hind tibia length and spermathecal area are associated with testis size across the Scathophagida after phylogenetic control. The same result was found by Minder (2002) using the CAIC approach, confirming the validity of our procedure. However, the decision to retain the hind tibia length  $x^{(1)}$  at this level is marginal. The log-likelihood for the model neglecting inter-species correlations is 25.72, less than 2 smaller than the maximum log-likelihood obtained above, so that this model, with  $\lambda = \infty$ , is within the 95% confidence interval for  $\lambda$ ; on this assumption, the parameter estimates are

$$\tilde{\beta}^{(1)} = 0.2065, \quad \tilde{\beta}^{(3)} = 19.82 \quad \text{and} \quad \tilde{\sigma} = 0.1799, \quad (6)$$

with standard deviations 0.129 and 7.92 for  $\tilde{\beta}^{(1)}$  and  $\tilde{\beta}^{(3)}$  respectively. However, because the natural assumption is that evolutionary correlation is to be expected, it seems nonetheless preferable to accept the model with  $\hat{\lambda} = 0.3676$ .

In order to judge the validity of the procedure, and to obtain an alternative assessment of the variability in the estimates, the model (5) and the correlation structure  $C(\hat{\lambda})$  for the errors were held fixed, and data from this model distribution were simulated 1'000 times. The estimation procedure was then applied individually to each of the 1'000 resulting sets of data. These led to mean values

$$\bar{\beta}^{(1)} = 0.2137, \quad \bar{\beta}^{(3)} = 23.33 \quad \text{and} \quad \bar{\sigma} = 0.1513, \quad (7)$$

for the parameter estimates, with empirical standard deviations of 0.112 and 8.33 and an empirical correlation of  $-0.260$  for the 1'000 estimates of  $\beta^{(1)}$  and  $\beta^{(3)}$ , and an empirical standard deviation of 0.0388 for the 1'000 estimates of  $\bar{\sigma}$ . The mean values in (7) are very much in accord with the parameters of the model (5), though the estimated standard deviation  $\hat{\sigma}$  of the error in (5) is almost a standard deviation higher than the empirical average  $\bar{\sigma}$ . The empirical standard deviations of the estimates  $\hat{\beta}^{(1)}$  and  $\hat{\beta}^{(3)}$  are a little higher than the values calculated from  $C(\hat{\lambda})$  and  $\hat{\sigma}$ , as is to be expected, because the variability in the estimation of  $\lambda$  is also incorporated in the simulations.

The estimates of the correlation structure in the simulated data were by no means as stable. The empirical distribution of the estimates of the large correlation had a broad peak around 0.8 and a second, smaller peak at 0, with a range of values between represented at lower density. The moderate correlation was more or less uniformly distributed between 0 and 0.8, except for a noticeable peak at 0; the small correlation, but for an occasional large value, mostly took values between 0 and 0.05. This variability in the estimates reflects the rather flat log-likelihood for values of  $\lambda > \hat{\lambda}$  in the model (5), with  $L(\infty)$  only 1.5 smaller than the maximum  $L(\hat{\lambda})$ , and is consistent with around 20% of the simulations estimating the model without inter-species correlations as being best. This is not particularly surprising, in view of the small number (15) and large variability ( $\hat{\sigma} = 0.1798$ , as compared with estimated effects  $\hat{\beta}^{(1)} \times \text{SD}(x^{(1)}) = 0.0822$  and  $\hat{\beta}^{(3)} \times \text{SD}(x^{(3)}) = 0.0503$ ) of the observations; indeed, the very closely related species *N. striolata* and *N. spinimona* even had very different values of  $x^{(1)}$ . However, there was no apparent correlation between the estimates of  $\lambda$  and those of the model parameters  $\beta^{(1)}$  and  $\beta^{(3)}$ .

The same analyses can also be conducted with correlations based on the evolutionary distance matrix  $d^{(1)}$ . The results are broadly the same, though the maximum log-likelihood achieved is somewhat smaller (26.75), indicating a slightly worse fit. This difference is actually enough to result in the preferred choice of model including only  $x^{(3)}$ , the spermathecal area, if the test is made at the 5% level; once again, the decision is marginal.

## 4. Discussion

The method that we propose for conducting inter-species regression analyses is developed from the engagingly simple idea of using likelihood-based methods in conjunction with the Ornstein-Uhlenbeck model of evolution. The result is a procedure which can be carried out without knowing the complete phylogeny — a measure of the evolutionary distance between pairs of species suffices — and which, at the same time, assesses the importance of the phylogeny for the analysis. It is thus rather surprising that it has not been introduced earlier.

The procedure involving branch length modification in Blomberg *et al.* (2003) seems to come closest. Their model is one without covariates,  $y_i = \beta^{(0)} + e_i$ , and their main interest is therefore in the strength of the correlations between the  $e_i$ ; we are principally interested in estimating and testing regression coefficients, and properly accounting for the correlations is a means towards this end. It should be noted that the analysis given in Blomberg *et al.* (2003, Appendix 2) is incorrect, in that they set the variance of the character at the root of the phylogenetic tree to be 0, but do not then take into account the subsequent drift from this (unknown) value towards its mean, as a result of the OU-process drifting towards 0. Their formulae then also have the drawback of involving not only the inter-species distances, but also the root to tip distances, requiring a more precise knowledge of the underlying phylogeny than is the case for our analysis.

Our example illustrates that the method works much as expected when applied to a ‘typical’ biological data set. Although the detailed correlation structure was not reliably estimable, because there were few data points and a low signal to noise ratio, this still did not prevent the regression coefficients and their precisions being successfully estimated. Indeed, the procedure performed very well in a number of respects. It estimated the regression parameters and the variability of these estimates satisfactorily; it highlighted the extent to which the highly variable data did not support accurate estimation of the underlying correlation structure; and it indicated that, with these data, assuming that phylogeny had little impact on trait associations and treating species as independent data points would have had little impact on the final conclusions.

## Acknowledgement

ADB was supported in part by Schweizerischer Nationalfonds Projekt Nr. 20-67909.02, and by the Institute for Mathematical Sciences of the National University of Singapore.

DJH was supported in part by Schweizerischer Nationalfonds Projekt Nr. 31-56902.99.

## References

- Bernasconi, M. V., Pawlowski, J., C. Valsangiacomo, C., Piffaretti, J.-C., Ward, P. I., 2000. Phylogeny of Scathophagidae (Diptera, Calyptratae) based on mitochondrial DNA sequences. *Mol. Phyl. Evol.* 16, 308-15.
- Blomberg, S. P. B., Garland, T. Jr., Ives, A. R., 2003. Testing for phylogenetic signal in comparative data: behavioural traits are more labile. *Evolution* 57, 717-745.
- Brockwell, P. J., Davis, R. A., 1987. *Time series: theory and methods*. Springer, New York.
- Cummins, J. M., Woodall, P. F., 1985. On mammalian sperm dimensions. *J. Reprod. Fert.* 75, 153-175.
- Defant, A., Floret, K., 1993. *Tensor norms and operator ideals*. North Holland Mathematical Studies 176, North Holland, Amsterdam.
- Draper, N. R., Smith, H., 1966. *Applied regression analysis*. Wiley, New York.
- Edwards, A. W. F., 1972. *Likelihood*. Cambridge University Press.
- Felsenstein, J., 1985. Phylogenies and the comparative method. *Amer. Nat.* 125, 1-15.
- Felsenstein, J., 1988. Phylogenies and quantitative characters. *Ann. Rev. Ecol. Syst.* 19, 445-471.
- Gage, M. J. G., 1994. Associations between body size, mating pattern, testis size and sperm lengths across butterflies. *Proc. Roy. Soc. B* 258, 247-254.
- Harvey, P. H., 2000. Why and how phylogenetic relationships should be incorporated into studies of scaling, in: Brown, J. H., West, G. B., (Eds.), *Scaling in Biology*, Oxford University Press, Oxford, pp. 253-265.
- Hosken, D. J., 1997. Sperm competition in bats. *Proc. Roy. Soc. B* 264, 385-392.
- Markow, T. A., 2002. Perspective: Female remating, operational sex ratio, and the arena of sexual selection in *Drosophila* species. *Evolution* 56, 1725-1734.

- Martin, R. D., 1981. Relative brain size and basal metabolic rate in terrestrial vertebrates. *Nature* 293, 57-60.
- Martins, E. P., Garland, T. Jr., 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution* 45, 534-557.
- McNab, B. K., 2002. *The physiological ecology of vertebrates: a view from energetics*. Comstock Publishing Associates, Cornell University Press, Ithaca, New York.
- Minder, A., 2002. Co-evolution of sperm, testis and the female tract morphology in Scathophagidae. MSc Thesis, University of Zurich.
- Møller, A. P., Birkhead, T. R., 1992. A pairwise comparative method as illustrated by copulation frequency in birds. *Amer. Nat.* 139, 644-656.
- Pagel, M. D., Harvey, P. H. 1989. Taxonomic differences in the scaling of brain size on body size among mammals. *Science* 244, 1589-1593.
- Pallmann, B., 2001. Parameterschätzung für die komparative Methode in der Evolutionsbiologie. Diploma Thesis, Institute of Mathematics, University of Zurich.
- Purvis, A., Rambaut, A., 1994. *Comparative analysis by independent contrasts (CAIC)*, Version 2. Oxford University.
- Stearns, S. C. 1992. *The evolution of life histories*. Oxford University Press, Oxford.
- Thompson, G. G., Withers, P. C., 1998. Standard and maximal metabolic rates of goannas (Squamata: Varanidae). *Physiol. Zool.* 70, 307-323.
- Younger, M. S., 1985. *A first course in linear regression*, 2nd Edn. PWS Publishers, Boston.

**Table 1:** Testis size, mean HTL, duct length and spermathecal size for 15 species of Scathophagidae: data from Minder (2002).

Species	Testis size (mm <sup>2</sup> ) $y$	mean HTL (mm) $x^{(1)}$	Duct Length (mm) $x^{(2)}$	Spermathecal Area (mm <sup>2</sup> ) $x^{(3)}$
<i>Cordilura albipes</i>	.169	2.410	.534	.00490
<i>Cleigastra apicalis</i>	.078	2.080	.412	.00769
<i>Cordilura ciliata</i>	.435	3.290	.604	.01743
<i>Cordilura pubera</i>	.332	2.775	.727	.01611
<i>Microprosopa pallidicauda</i>	.477	2.125	.531	.00795
<i>Norellia liturata</i>	.382	2.095	.962	.02497
<i>Norellia spinimana</i>	.547	2.295	1.086	.02048
<i>Norellia striolata</i>	.855	3.110	1.384	.02397
<i>Phrosia albilabris</i>	.319	2.380	.519	.01485
<i>Scathophaga cineraria</i>	.486	2.750	.561	.01046
<i>Scathophaga furcata</i>	.965	2.430	.541	.02195
<i>Spaziphora hydromyzina</i>	.134	2.000	.235	.01049
<i>Scathophaga stercoraria</i>	.544	2.815	.672	.01044
<i>Scathophaga suilla</i>	.461	2.380	.386	.01002
<i>Scathophaga taeniopa</i>	.699	2.695	.479	.01347

**Table 2:** Numbers of differing pairs between the sequences of 810 mDNA letters coding for the COI gene in 15 species of Scathophagidae: original sequences from Genbank.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	Species
* 70	64	69	72	73	87	83	62	67	75	82	63	65	70	a:	<i>Cordilura albipes</i>
* 92	93	64	69	76	75	81	60	71	74	55	54	61	b:	<i>Cleigastra apicalis</i>	
* 63	80	77	88	88	71	73	68	89	71	72	77	c:	<i>Cordilura ciliata</i>		
* 81	83	99	96	75	84	88	96	81	81	85	d:	<i>Cordilura pubera</i>			
* 66	79	73	82	63	74	55	52	59	63	e:	<i>Microprosopa pallidicauda</i>				
* 67	65	77	59	70	70	58	57	63	f:	<i>Norellia liturata</i>					
* 9	95	67	80	74	69	72	79	g:	<i>Norellia spinimana</i>						
* 94	64	77	71	64	68	75	h:	<i>Norellia striolata</i>							
* 69	78	83	68	70	72	i:	<i>Phrosia albilabris</i>								
* 36	62	35	20	30	j:	<i>Scathophaga cineraria</i>									
* 73	43	41	45	k:	<i>Scathophaga furcata</i>										
* 56	56	62	l:	<i>Spaziphora hydromyzina</i>											
* 29	38	m:	<i>Scathophaga stercoraria</i>												
* 22	n:	<i>Scathophaga suilla</i>													
* o:	<i>Scathophaga taeniopa</i>														