

# Regression dilution in energy management patterns

Lewis G. Halsey and Andrea Perna

Life Sciences, University of Roehampton, SW15 4JD, London, UK

Keywords: energy expenditure, heart rate, metabolic rate, regression bias

## Abstract:

Analysis of some experimental biology data involves linear regression and interpretation of the resulting slope value. Usually the x-axis measurements include noise. Noise in the x-variable can create regression dilution, and many biologists are not aware of the implications – regression dilution results in an underestimation of the true slope value. This is particularly problematic when the slope value is diagnostic. For example, energy management strategies of animals can be determined from the regression slope estimate of mean energy expenditure against resting energy expenditure. Typically, energy expenditure is represented by a proxy such as heart rate, which adds substantive measurement error.

With simulations and analysis of empirical data, we explore the possible effect of regression dilution on interpretations of energy management strategies. We conclude that unless  $r^2$  is very high, there is a good possibility that regression dilution will affect qualitative interpretation. We recommend some ways to contend with regression dilution, including the application of alternative available regression approaches under certain circumstances.

## Introduction

In experimental biology, the results of linear regressions are usually interpreted in terms of whether the relationship differs from the usual null hypothesis of 0, or by predicting values of y from x. Interpretation is less often based on the regression slope value. It is perhaps for this reason that many researchers are not aware of some of the problems arising from bias in linear regression slope estimates, which occur due to random measurement noise in the x axis. This bias in slope estimates is termed 'regression dilution' or 'attenuation bias', and results in an underestimate of the true slope value when the regression slope is calculated using the ordinary least squares (OLS) approach, which assumes that the x-axis values are error-free (Frost and Thompson, 2000; Smith, 2009). Regression dilution occurs because lower values of x tend to include a disproportionate number of values that are underestimates while higher values of x tend to include a disproportionate number of values that are overestimates (MacMahon et al., 1990) (for further explanation see Figure 1). The result is an increase in the x-value range, serving to spuriously attenuate the slope gradient towards 0. Measurement noise occurs due to any variation that causes the

observed values to be randomly different to the 'true' values (McArdle, 2003), such as inaccuracies during the recording of the x-value variable, sampling error and/or when the x-value variable is being used as a proxy. Although some comparative physiologists have highlighted the regression dilution problem (e.g. Green, 2001; Herrera, 1992; LaBarbera, 1989; McNerny and Purves, 2011; White, 2011; White and Kearney, 2014), there is value in revisiting this issue through application to an en vogue subfield of comparative physiology.

An area of comparative physiology for which analysis is based on interpretation of the gradient of linear regression slopes is energy management modelling. The amount of energy that animals can use to fuel their lives is finite and thus we expect animals to be strategic with their energy expenditure. One aspect of this energetics strategy is represented by patterns of energy management, which indicate the broad relationships between the energy an animal spends on 'background' processes such as cell growth and immune function against the energy it spends on 'auxiliary' processes such as locomotion (Halsey et al., 2019). The slope of the relationship between daily energy expenditure and background energy expenditure provides quick and easy insight into animals' energy management (Mathot and Dingemanse, 2015; Ricklefs et al., 1996). Slope estimates  $<1$  indicate the constraint pattern of energy management whereby an animal compensates during periods when auxiliary energy expenditure is high by decreasing background energy expenditure, and vice versa, thus constraining daily energy expenditure. A slope estimate of 1 is predicted by the independent pattern whereby variations in auxiliary energy expenditure do not correlate with variations in background energy expenditure (i.e. there is a lack of constraint of energy expenditure). Slope estimates  $>1$  indicate the performance pattern of energy management whereby greater auxiliary energy expenditure is associated with greater background energy expenditure. For a visual representation of this explanation, see Figure 2. Hence, this analytical process for categorising energy expenditure into one of three management strategies based on the relationship between daily and background energy expenditure is reliant on interpreting the gradient of the linear regression line. However, the x-value variable in analyses of energy management patterns from linear regression is prone to multiple sources of noise, particularly when a proxy for energy expenditure, such as heart rate, has been measured (Portugal et al., 2016).

These interpretations of slope estimates can strongly influence how we perceive animals respond to variations in their daily activity levels. For example, where the slope estimate  $<1$ , animals appear to be trading off the energy they expend on background metabolic costs with that which they expend on auxiliary costs, indicating a clear limit to their energy expenditure. In the case of humans, if they exhibit the constraint pattern then prescribed increases in exercise may be less effective at reducing weight than currently presumed; this is particularly pertinent for modern-living human populations in the midst of an obesity epidemic. It is therefore important that the slope estimates are accurate, yet the phenomenon of regression dilution may be causing inaccuracies and in turn encouraging a misinterpretation of the data.

We investigate the possibility of regression dilution affecting the slope estimates interpreted in the context of energy management patterns by:

- (1) running simulations of ecologically valid randomised samples of heart rate measurements to elucidate how different levels of measurement noise variance affect the slope estimate;
- (2) revisiting some of the data presented in (Halsey et al., 2019) and comparing the slope estimates of simple linear regressions fit to those data by different approaches, and then quantifying how the strength of the relationship appears to relate to the degree of regression dilution.

## Materials and Methods

Four approaches to linear regression are applied in the analyses of the present study: ordinary least squares, OLS; and three major axis approaches: major axis, MA; standard major axis, SMA; ranged major axis, RangedMA. While the OLS approach assumes that measurement noise only exists in the y-axis values, major axis approaches accept measurement noise in both axes, but each approach assumes different ratios in the magnitude of that noise between y and x (Legendre and Legendre, 1998; Quinn and Keough, 2002). This ratio is termed lambda, and thus in the current study lambda is calculated as the ratio of measurement noise in daily mean heart rate and measurement noise in daily minimum heart rate.

Through analysis of both simulated data and empirical data, we investigated the regression dilution caused by different values of lambda. The analyses were conducted in R v.3.4.0, and the various regression approaches were applied using the package `lmodel2()`.

### *Simulations*

To investigate the effects of different measurement noise ratios of daily mean heart rate and daily minimum heart rate (lambda), simulations were run involving 1000 iterations of datasets generated to represent ecologically valid ranges of heart rate values (beats / min). Each iteration was based on 100 values of daily mean heart rate, each associated with a value of daily minimum heart rate. Daily minimum heart rate values were randomly drawn from a distribution with mean 60 and standard deviation between 0 and 5 (see later). Daily mean heart rate is the summation of daily minimum heart rate and daily auxiliary heart rate (Halsey et al., 2019), thus 100 values of daily auxiliary heart rate were generated by drawing randomly from a distribution also with mean 60, and standard deviation 3. This process provided 100 values of true (i.e. without measurement noise) daily mean heart rate and daily minimum heart rate generated according to the independent energy expenditure pattern (no correlation between the two variables). Measurement noise was induced into the values of daily mean heart rate by randomly drawing values of noise from a normal distribution of mean 0 and standard deviation 3. Measurement noise was induced into the values of daily minimum heart rate also by randomly drawing from a normal distribution of mean 0, however the magnitude of the standard deviation was varied for each simulation in order to affect lambda.

Six simulations were run, the first with a standard deviation for the distribution of daily minimum heart rate of 0, and each subsequent simulation incorporating a unitary increase in that value, producing lambda values for each simulation of infinity ( $\infty$ ), 3, 1.5, 1, 0.75 and 0.6. For each iteration of each simulation, daily mean heart rate was regressed against daily minimum heart rate using four approaches to linear regression. By plotting each simulation separately and including the average slope estimates across all iterations, along with the average correct value across all iterations (very close to 1), it is possible to infer which approaches to the regression of daily mean heart rate against daily minimum heart rate are most accurate at various values of lambda.

### *Empirical data*

Empirical data were taken from the dataset presented in Halsey et al. (2019), which represents daily mean and minimum heart rate values for multiple individuals of each of 16 vertebrate species. To account for temporal autocorrelation in the data, for each species the dataset was reduced to every fifth data point. Certain species were then removed from the dataset due to typically small sample sizes per individual. For the remaining 11 species (represented by 12 datasets), a single individual was randomly selected (with the stipulation that the selected individual represented at least 20 data points, which is arguably important for SMA regression; Jolicoeur, 1990) and daily mean heart rate was linearly regressed against daily minimum heart rate using the four regression approaches stated earlier. This process resulted in a single  $r^2$  value and slope estimate calculated from each regression approach, per species. Finally, to investigate whether the correlations between daily mean heart rate and daily minimum heart rate with lower  $r^2$  values are subjected to greater regression dilution, the difference between the OLS slope estimate and each of the major axis fit slope estimates was regressed against  $r^2$ .

## Results

### *Simulations*

The outputs from the six simulations are presented in Figure 3, in both graphical and tabulated forms. When there is no noise in the measurements of daily minimum heart rate (the x-axis variable), lambda =  $\infty$  and the strength of the correlation (measured by the coefficient of determination;  $r^2$ ) is high, as would be expected. As the measurement noise in the x-axis variable is increased (and lambda decreases),  $r^2$  decreases. While all four regression approaches exhibit a decrease in slope estimate as lambda decreases, thus arguably all showing regression dilution, different regression approaches provide the most accurate slope estimate at different lambda values.

When lambda =  $\infty$ , the OLS slope estimate is almost identical to the correct slope of 1. The other regression approaches (MA, SMA and RangedMA) all return substantially greater slope estimates. The case is similar at lambda = 3, where the noise variance in the measures of daily minimum heart rate is one third the magnitude of the noise variance in the measures of daily mean heart rate. At lambda = 1.5, all major axis regression approaches

somewhat overestimate the slope while OLS somewhat underestimates it. At  $\lambda = 1$ , indicating the same magnitude of noise variance in both heart rate variables, OLS no longer provides the most accurate slope estimate, and SMA and RangedMA are both quite close to the true value of 1. In the last two simulations, where the noise variance in daily mean heart rate is larger than the noise variance in daily minimum heart rate ( $\lambda = 0.75$  and  $0.6$ ), all three major axis regression approaches provide at least reasonably accurate slope estimates while OLS returns a considerable underestimate. In all simulations, the MA approach provides a less accurate slope estimate than either SMA or RangedMA.

### *Empirical data*

The  $r^2$  value for the regression of each single individual representing each species, along with the simple linear regression slope estimate determined by each regression approach, is presented in Table 1.  $R^2$  was typically high ( $> 0.7$  for 8 of the 11 datasets), suggesting that the correlation between daily mean heart rate and daily minimum heart rate is often strong for this type of data. For every species, the slope estimate calculated from the OLS regression approach is lower than the slope estimate calculated for all of the major axis approaches. The difference between the OLS slope estimate and each of the major axis slope estimates covaries negatively with the  $r^2$  value of the relationship (Figure 4).

## Discussion

The energy management patterns exhibited by animals can be inferred from the slope estimates of linear regressions between daily mean heart rate and daily minimum heart rate. The present study examined how noise variance in heart rate measures could affect the accuracy of these regression slope estimates.

The simulations (Figure 3) show that when the noise variance in daily minimum heart rate is either non-existent or at least low compared to the noise variance in daily mean heart rate (thus  $\lambda$  is high), ordinary least squares (OLS) regression provides an accurate slope estimate; there is no appreciable regression dilution. This slope estimate is more accurate than the estimates returned from other regression approaches, which overestimate. However, once the noise variance in daily minimum heart rate is at least as large as that in daily mean heart rate (i.e.  $\lambda < 1$ ), the OLS slope estimate has attenuated considerably, thus becoming an inaccurate underestimate, while in contrast certain other regression approaches provide an accurate slope estimate. This is to be expected since while OLS regression assumes that the y-axis variable, but not the x-axis variable, is measured with noise (Quinn and Keough, 2002), the various major axes regression approaches (major axes, MA; standard major axes, SMA; ranged major axes, RangedMA) accept noise in both variables (Herrera, 1992).

There is of course noise in real measurements of daily minimum heart rate, thus  $\lambda \neq \infty$ .  $\lambda$  might be estimated at  $\sim 1$  since both daily mean and minimum heart rate are likely to incur the same forms of measurement noise: measurement technique imperfections, sampling variation and being used as a proxy for energy expenditure (though

see Smith, 2009). Moreover, possibly minimum heart rate has even greater noise variance than mean heart rate because while estimates of mean heart rate remain centred on the real value independently of the sample used for the estimation, estimates of minimum heart rate are affected by the duration of time over which minimum heart rate is calculated. The simulations indicate that if indeed  $\lambda$  is  $\sim 1$ , or is  $<1$ , OLS is not a viable regression analysis for interpreting energy management patterns.

Regression dilution will be greater when the coefficient of determination ( $r^2$ ) value is smaller, because measurement noise is here defined as any deviations from a perfect fit between the y- and x-variables (Smith, 2009). This phenomenon was confirmed by the simulations, and we also showed this in the empirical heart rate data sets (Figure 4); a higher  $r^2$  value for a species is associated with a higher slope estimate calculated using OLS regression. This suggests that  $\lambda$  is sufficiently low in some of these regressions that regression dilution is clearly apparent. Of course, we do not know the true value of each regression slope of empirical data. However, comparing the reduction in OLS slope estimate with the three other regression approaches (Fig. 3), it appears that when  $r^2 > 0.8$  the difference in slope estimate is minimal ( $< 0.1$ ), while  $r^2$  values of around 0.6 have a slope estimate difference of around 0.3, and substantially smaller  $r^2$  values have differences considerably larger.

How might regression dilution affect previous reports of energy management patterns based on analysis of the regression slope estimate of daily mean heart rate against daily minimum heart rate? Here we consider three published papers as brief case studies. Vezina et al. (2006) report a slope of 1.1 for captive, non-breeding zebra finches, with an  $r^2$  for the OLS regression of 0.35. This relatively low  $r^2$  value might suggest that the true slope value is somewhat higher than 1.1, which in turn could move interpretation of the energy management pattern exhibited by these birds from the independent pattern to the performance pattern. Careau (2017) reports that people training for a half-marathon exhibit an among-individuals slope of 2.60 ( $r^2 = 0.39$ ). Again, this  $r^2$  value is sufficiently low that we might be concerned the analysis includes a substantial degree of regression dilution. However, in this case the qualitative interpretation made of the slope estimate is perhaps unlikely to be affected because the among-individuals slope is already  $\gg 1$  (performance pattern).

Third, Halsey et al. (2019) argue that the species they analysed predominantly exhibit either the independent (slope = 1) or performance (slope  $> 1$ ) pattern at the across-individuals level; the evidence for this claim would be strengthened if regression dilution was not present since slope estimates would be higher. They also suggest that at the within-individual level there is a general tendency for species to exhibit an element of the compensation pattern (slope  $< 1$ ). For some species this interpretation is likely to be robust since the  $r^2$  values associated with the slope estimates are very high (e.g. red deer,  $r^2 = 0.96$ ; grey seals,  $r^2 = 0.83$ ). For other species, however, where the  $r^2$  values are relatively low, regression dilution might be falsely indicating that individual animals are exhibiting an element of energy compensation (e.g. Australasian gannets,  $r^2 = 0.40$ ; human beings,  $r^2 = 0.64$ ). Halsey et al. (2019) also make the claim that there is generally a 'left shunt' in slope

estimate from the across-individuals level to the within-individual level (Figure 3 in Halsey et al., 2019). This observation should be robust because the slope estimate confidence intervals are always larger at the across-individuals level (and hence measurement noise is greater), suggesting that the regression dilution is probably attenuating the size of this left shunt. Finally, the regression dilution in these analyses did not hide the insightful negative correlations found between slope value and mean heart rate per month (Figure 3 in Halsey et al., 2019), which suggest that species exhibit more energetic constraint during periods when daily energy expenditure is higher.

#### *How should we conduct energy management regressions?*

It is not the case that OLS should be substituted for an alternative approach simply because daily minimum heart rate includes noise variance. If the noise variance associated with daily minimum heart rate is fairly small compared to the daily mean heart rate noise variance (i.e.  $\lambda$  is large), our simulations confirm the advice of McArdle (1988) that OLS is appropriate (see also White, 2011). Smith (2009) argues it is usually the case in regression analyses of biological data that  $\lambda$  is large. In turn, he suggests that OLS is appropriate when the x-variable is thought to be affecting the y-variable, which is indeed the case in regressions of daily mean heart rate against daily minimum heart rate. However, at low  $r^2$  values OLS can underestimate the slope considerably; indeed, low slope estimates associated with a low  $r^2$  are suggestive of an inappropriate regression model (LaBarbera, 1989). Yet in this situation major axes methods can overestimate the slope (Figure 3; see also Kimura, 1992). Unfortunately, for datasets associated with energy expenditure such as heart rate or rate of oxygen consumption, the noise associated with measurement inaccuracies, with the use of these variables as proxies of energy expenditure and due to the fact that sampling is always imperfect, cannot easily be quantified. Therefore, it is difficult to ascertain whether  $\lambda$  is sufficiently large that OLS is a better approach than other methods (Smith, 2009). Based on the simulation results, and in agreement with McArdle (1988), one rule of thumb worth considering, however, is that if the major axis approach is taken then SMA or RangedMA may provide more accurate slope estimates than will MA.

Where the slope estimate is the focus of data interpretation and  $r^2$  is less than very high, researchers are advised to consider presenting their data using more than one regression approach. However, modelling approaches more complicated than single linear regression are usually based on OLS (Smith, 2009). In these situations, we would suggest that some simple linear regressions are also conducted, using a range of fitting approaches, to gain some insights into the potential impact of regression dilution on the slope estimates.

#### Acknowledgements

We thank Jon Green for discussions on elements of a draft version of this article.

#### Funding information

No funding was received for this study.

Table 1.  $r^2$  and slope estimates calculated from various simple linear regression fits, for daily mean heart rate against daily minimum heart rate of single individuals from a range of vertebrate species.

Species	$r^2$	OLS	MA	SMA	RMA	Mean MA	n
Barnacle goose	0.77	1.02	1.18	1.16	1.16	1.17	73
Greylag goose	0.72	0.74	0.85	0.87	0.86	0.86	79
Great cormorant	0.23	0.69	2.06	1.45	1.37	1.63	21
Australasian gannet	0.47	1.09	1.9	1.58	1.54	1.67	40
Little penguin	0.64	0.97	1.28	1.22	1.24	1.25	39
Macaroni penguin	0.77	1.09	1.28	1.24	1.25	1.26	56
Eider duck	0.26	0.48	0.88	0.93	1.08	0.96	42
Przewalski horse	0.95	1.04	1.07	1.07	1.07	1.07	41
Alpine Ibex	0.8	1.46	1.71	1.63	1.65	1.66	31
Red deer (dataset 1)	0.93	0.97	1.01	1.01	1.01	1.01	95
Red deer (dataset 2)	0.82	1.19	1.35	1.31	1.33	1.33	122
Roe deer	0.71	1.32	1.7	1.57	1.49	1.59	25
<b>Mean</b>	0.67	1	1.36	1.25	1.25	1.29	55.3
<b>1 S.E.</b>	0.02	0.02	0.03	0.02	0.02	0.02	2

OLS =ordinary least squares, MA = major axis, SMA = standard major axis, RangedMA = ranged major axis. Mean MA is the mean of the MA, SMA and RangedMA values.



## References

Frost, C. and Thompson, S. G. (2000). Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163, 173-189.

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

Halsey, L. G., Green, J. A., Twiss, S. D., Arnold, W., Burthe, S. J., Butler, P. J., Cooke, S. J., Grémillet, D., Ruf, T. and Hicks, O. (2019). Flexibility, variability and constraint in energy management patterns across vertebrate taxa revealed by long-term heart rate measurements. *Functional Ecology*.

Herrera, C. M. (1992). Interspecific variation in fruit shape: allometry, phylogeny, and adaptation to dispersal agents. *Ecology* 73, 1832-1841.

Jolicoeur, P. (1990). Bivariate allometry: interval estimation of the slopes of the ordinary and standardized normal major axes and structural relationship. *Journal of Theoretical Biology* 144, 275-285.

Kimura, D. K. (1992). Symmetry and scale dependence in functional relationship regression. *Systematic Biology* 41, 233-241.

LaBarbera, M. (1989). Analyzing body size as a factor in ecology and evolution. *Annual review of ecology and systematics* 20, 97-117.

Legendre, P. and Legendre, L. (1998). Numerical ecology: second English edition. *Developments in environmental modelling* 20.

MacMahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J. B. p., et al. and stroke, a. (1990). Part 1. Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* 335, 765-774.

Mathot, K. and Dingemanse, N. (2015). Energetics and behavior: unrequited needs and new directions. *TREE* 30, 199-206.

McArdle, B. (1988). The structural relationship: regression in biology. *Canadian Journal of Zoology* 66, 2329-2339.

McArdle, B. H. (2003). Lines, models, and errors: regression in the field. *Limnology and Oceanography* 48, 1363-1366.

McInerny, G. J. and Purves, D. W. (2011). Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution* 2, 248-257.

Portugal, S. J., Green, J. A., Halsey, L. G., Arnold, W., Careau, V., Dann, P., Frappell, P. B., Grémillet, D., Handrich, Y., Martin, G. R. et al. (2016). Associations between Resting, Activity, and Daily Metabolic Rate in Free-Living Endotherms: No Universal Rule in Birds and Mammals. *Physiological and Biochemical Zoology* 89, 251-261.

Quinn, G. P. and Keough, M. J. (2002). Experimental design and data analysis for biologists: Cambridge University Press.

Ricklefs, R. E., Kornazewski, M. and Daan, S. (1996). The relationship between basal metabolic rate and daily energy expenditure in birds and mammals. *The American Naturalist* 147, 1047-1071.

Smith, R. J. (2009). Use and misuse of the reduced major axis for line-fitting. *American Journal of Physical Anthropology* 140, 476-486.

White, C. R. (2011). Allometric estimation of metabolic rates in animals. *Comparative Biochemistry and Physiology A* 158, 346-357.

White, C. R. and Kearney, M. (2014). Metabolic Scaling in Animals: Methods, Empirical Results, and Theoretical Explanations. *Comprehensive Physiology* 4, 231-256.