Comparing Classical and Alternative Methods of Regression in the Context of Metabolic Scaling

Nathan Huey '13, Dr. Brad Hartlaub

Introduction:

Studies of metabolic scaling often ultimately seek to gain insight into how one characteristic of an organism (e.g. standard metabolic rate, nitrogen assimilation, etc.) changes as the size of the organism increases. Regression offers a formal way of modeling these relationships, and providing parameter estimates. These parameter estimates allow predictions of a "response" variable (Y) from one or more "explanatory" variables (X).

Although regression can take other forms, often simple linear regression is employed to examine a statistical relationship between two variables. "Simple" refers to the use of only one explanatory and one response, while "linear" means that some measure of the location of the distribution of response values at a given value of the explanatory varies in a linear way with the explanatory variable. In the context of allometric scaling projects, a linear relationship is often modeled between the logarithm of a response such as metabolic rate and the logarithm of body mass.

Ordinary least squares (OLS) regression is by far the most common method of parameter estimation for reasons of consistency and general effectiveness. However, like all forms of regression, OLS has a particular environment where it provides optimal estimates. In particular this is when the "errors" of the OLS model are like random draws from a single normal distribution. The method has been shown to break down outside this environment, especially in the presence of outlying observations. Interestingly, often scaling data sets from *Manduca sexta*, the model organism used for the scaling project at Kenyon, have moderate to high numbers of outliers. This, as well as the observation that alternative regression models can give models noticeably different than OLS, provides an impetus for a comparative study.

The goal of this study was to compare the performance of several alternative methods of regression thorough Monte Carlo simulation and the bootstrapping of parameter estimates generated by each method and by comparing the coverage probabilities of confidence intervals generated by several of these alternative methods.

Methods :

OLS

The parameter estimates derived from OLS are defined such that the sum of the squared errors (the vertical distances from the points to the line) is minimized. That is:

$$min \sum_{i=1}^{n} (yi - \hat{\alpha} - \hat{\beta} xi)^2$$

These estimates, derived using calculus, have been shown to be the best linear unbiased estimators (BLUE) when the error distribution takes any shape (Gauss-Markov Theorem) and to be the best unbiased estimators when the error distribution is normal (Birkes and Dodge, 1993). "Best" in this sense refers to the estimates being the most precise, that is, the range of parameter estimates will be smallest for OLS.

Least Absolute Deviations (LAD)

LAD regression also minimizes a so called "loss function", but instead of minimizing the sum of the squared error terms, the sum of the absolute value of the error terms is minimized:

$$min \sum_{i=1}^{n} |yi - \hat{\alpha} - \hat{\beta} xi|$$

Although this may seem a more intuitive approach, the general lack of closed form solutions for parameter estimates made this approach much less popular historically. To calculate the parameter estimates, an iterative algorithm must be used (Birkes and Dodge, 1993) and without the help of a computer, this can be very tedious. Unlike OLS, LAD can produce non-unique estimates, but the major benefit of LAD is that it is "robust" in the presence of the outliers that can skew OLS estimates.

M-Regression

These methods are any that include the minimization of a loss function. While technically this includes the previous two methods, here we refer to it as a sort of "compromise" method between OLS and LAD. Three methods of regression of this type were considered here:

Huber's M-estimation

$$\rho(e) = \begin{cases} .5e^2 & \text{if } |e| \le s \\ s|e| - .5s^2 & \text{if } |e| > s \end{cases}$$

Tukey's Bisquare estimation

$$\rho(e) = \begin{cases} \frac{s^2}{6}\left\{1 - \left[1 - \left(\frac{e}{s}\right)^2\right]^3\right\} & \text{if } |e| \le s \\ \frac{s^2}{6} & \text{if } |e| > s \end{cases}$$

The rho-functions are the respective loss functions that are minimized; *e* is the vertical error term associated with the models.

Yohai MM-estimation

The loss function is similar to bisquare, but initial estimate is a high breakdown regression model.

References:

- Hussain, S.S and Sprent, P., "Nonparametric Regression."1983. Journal of the Royal Statistical Society. 146(2):182-191
- Birkes, D. and Dodge, Y., "Alternative Methods of Regression." 1993. New York, NY. Wiley
- Fox J., "Bootstrapping Regression Models." 2002. An R and S-Plus Companion to Applied Regression: Appendix.
- R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Nonparametric Regression

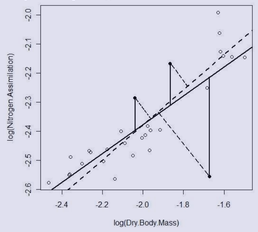
In general, nonparametric methods require fewer assumptions than parametric methods. Indeed, while all of the previous methods work well when OLS assumptions are approximately met, nonparametric regression only assumes that the error terms are symmetrically distributed with a median of zero (Hussain and Sprent, 1983). In this study, we considered Theil's nonparametric estimators. The slope estimate is simply the median of all the pairwise slope estimates. The intercept is the median of the set of intercepts generated by fitting a line with the calculated slope straight through each point.

Least-Trimmed Squares and Least-Median of Squares

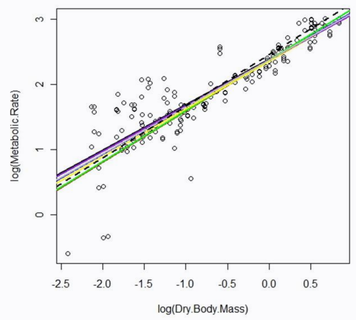
Least-Trimmed Squares (LTS) takes a specified percentage of the data points (>50%) and fits the best possible least-squares line to a subset of this size. Least-Median of Squares estimates are such that the median of the squared error terms is minimized. While both methods can handle especially large numbers of outliers, they have very poor precision under the optimal conditions of OLS.

Standardized Major Axis (SMA)

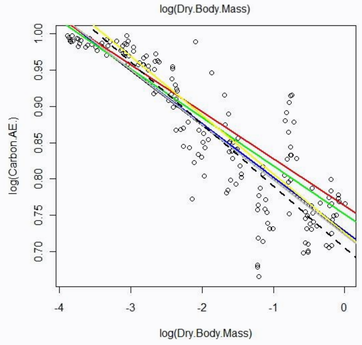
This method of regression is used primarily when there is not only uncertainty in the response variable, but also uncertainty in the explanatory variable. Unlike OLS regression where the "errors" are derived from the vertical displacements of the points from the line, SMA minimizes the straight line distances to the line.

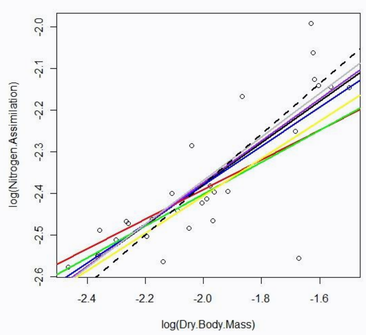


Graphical Comparisons of Alternative Regression Models:

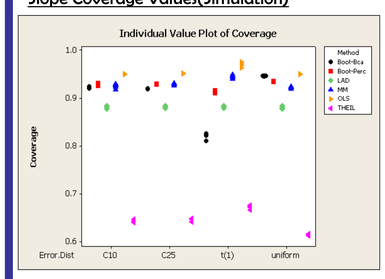


OLS	Black
LAD	Blue
LMS	Red
LTS	Green
SMA	Dotted
THEIL	Yellow
HUBER	Purple
BISQ	Gray
MM	Orange

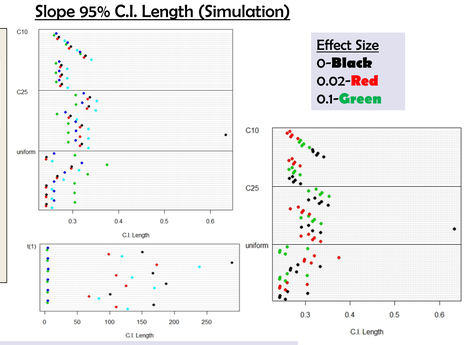




Slope Coverage Values(Simulation)



Slope 95% C.I. Length (Simulation)



Effect Size
0-Black
0.02-Red
0.1-Green

OLS	Light Blue	Percentile	Red
LAD	Green	MM	Dark Blue
BCa	Black		

	Carbon.AE	Nitro.Assim	Metabolic.Rate
OLS	0.01003	0.13301	0.13687
LAD	0.01547	0.10576	0.10771
LMS	0.04237	0.20559	0.11292
HUBER	0.01005	0.10289	0.11430
BISQ	0.01024	0.11524	0.13301
LTS	0.05903	0.23008	0.12614
THEIL	0.01454	0.13576	0.08595
MM	0.01018	0.11767	0.13553
SMA	0.01090	0.10219	0.15851

Table of Bootstrapped BCa 95% Confidence Interval Lengths (5,000 replications)

Comparisons:

To begin to compare the different methods in the context of metabolic scaling, each method (where possible) was fit to three data sets used in this study: a metabolic rate set, a nitrogen assimilation set, and a carbon assimilation efficiency set.

Bootstrap confidence intervals were then constructed for each type of regression. The bootstrap is a way to simulate the distribution of a statistic when only a limited set of data is available. To create the bootstrapped data sets a number of observations equal to the number of observations in the original data set were drawn one at a time, with replacement, from the original data set. Then the parameter estimates are generated for this new data set. This process was repeated 10,000 times.

One method of obtaining confidence intervals is the percentile method. This simply takes the empirical statistics at the 5th and 95th percentile of the ordered statistics as the endpoints of the confidence interval. Bias-corrected, accelerated (BCa) intervals also use empirical statistics as endpoints, but make a correction based on the number of estimates below that of the original data set (Fox, 2002). The lengths of the BCa 95% confidence intervals were then compared for the three data sets.

The final set of comparisons we did was through a Monte Carlo simulation study. We used OLS, LAD, MM, Theil, and a bootstrap for OLS in this study. We used the number of observations, the slope, intercept, and estimate of the standard deviation of the error distribution from the nitrogen assimilation data set, varying the type of error distribution and effect size.

Error distributions: 10% and 25% contaminated normal distributions, t-distribution with 1 degree of freedom, and a uniform distribution.

Effect Sizes: 0, 0.02, and 0.1-Before data were generated, an effect was added to the slope parameter. If coverage changed between these simulations, it would be due to the value of the slope parameter.

For each simulation we generated 100,000 new data sets and calculated a 95% confidence interval associated with each method (using 5,000 bootstrapped data sets per simulated data set). We then calculated the coverage associated with each method. This is the number of times the actual slope parameter fell within the interval generated divided by the number of replications (100,000). The expected value is 95%, but when conditions are not optimal, these values can change. Average confidence interval lengths were also computed for each method and compared.

Results/Conclusions:

- The exploratory graphs give little impression that the parameter estimates would differ significantly between the methods for these sets of scaling data.
- The bootstrapped C.I. lengths generated from the actual sets of data show that the method yielding the smallest interval varies between data sets, perhaps reflecting the different percentages of outliers present in the data. Although using the actual data is appealing, we have no idea how accurate the estimates are.
- Although the coverage associated with OLS remains consistently high, the t distribution with 1 degree of freedom yielded large enough errors to cause the associated confidence intervals to become so large as to be practically useless. Both robust methods (data for Theil C.I. length not available), LAD and MM, had much narrower intervals while retaining relatively coverage values. Although we expect the effect size to make no difference on the length of the C.I., effects showed patterns within both contaminated error distributions.
- Theil coverage values were consistently much lower (<70%) than expected, perhaps indicating that the assumptions of symmetric errors with median of zero is consistently violated.
- Overall, it would seem from this very limited view of the nature of scaling data that OLS can provide reasonably accurate and precise parameter estimates in the context of metabolic scaling, although it is essential to check the nature of the errors associated with each individual data set.
- It may be beneficial to look at how changing the shape of the error distribution and the parameters associated with them effects the coverage and confidence interval length of each method in more detail. Future studies could also look into methods of multiple regression and regression with repeated measurements.
- Graphs and simulations were created using the open-source statistical software package R and Minitab. This study was supported by the Kenyon College Summer Science Program and the InStaRs program funded by the NSF (Grant Number 0827208).