Regressions and the SMATR Package in R

Loren Albert

5-2-2012

Roadmap…

- Types of ‘regression’
- When to use, and not use, the different techniques (Warton et al. 2006)
- Example (Reich et al. 2010)
- SMATR tutorial

Regression Types-Summary

“Linear regression, MA and SMA are all least squares methods—the line is estimated by minimising the sum of squares of residuals from the line….The differences in methods of estimation of the lines are due to differences in the direction in which errors from the line are measured”

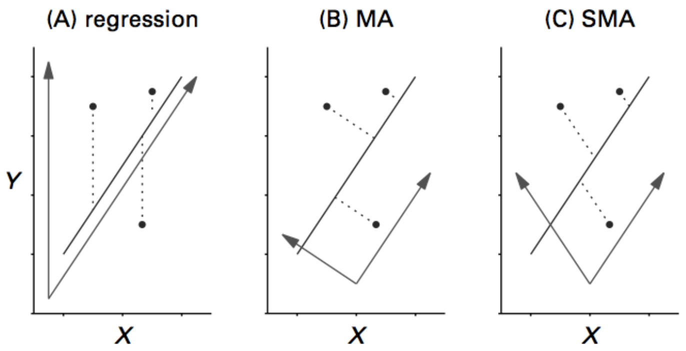


Fig. 4. The direction in which residuals are measured is (A) vertical for linear regression (B) perpendicular to the line for major axis estimation (C) the fitted line reflected about the Y axis for standardised major axis estimation. Axes are plotted on the same scale. The broken lines indicate residuals, and the arrows represent the fitted and residual axes, which are useful for understanding methods of estimation and inference about these lines.

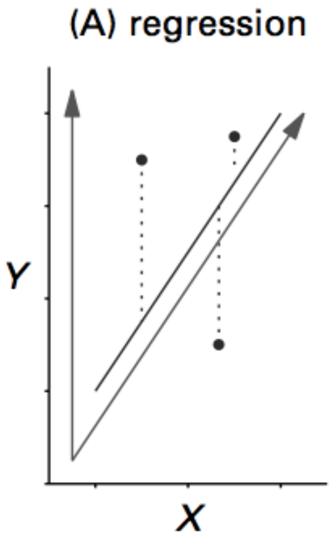
Warton et al. 2006

Regression Types

Linear regression (Model I regression)

$$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Minimize the residuals above, where y-hat is the fitted/predicted value of y_i.

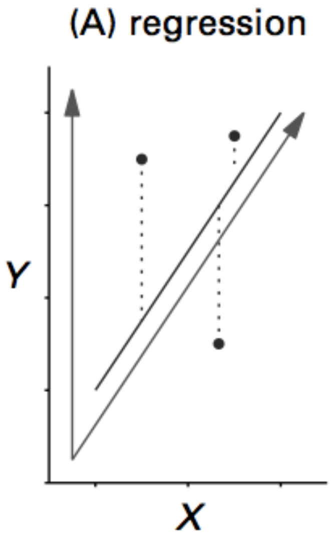


Warton et al. 2006

Regression Types

Linear regression (Model I regression)

Regression is useful to…
- Predict Y from X
- Determine if Y is associated with X (whether the slope is 0)
- Determine how strongly Y and X are associated (what is the r^2, the variation in Y that can be explained by linear regression on X)



Warton et al. 2006

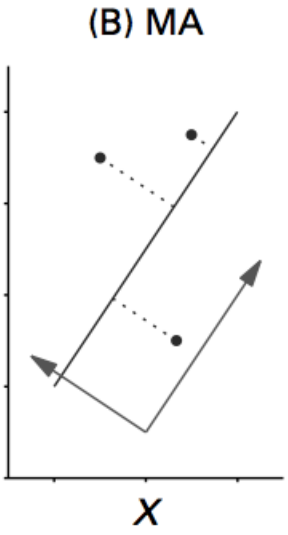
Regression Types

Major Axis

(a.k.a. First principal component axis of the covariance matrix)

(a.k.a "model II regression")

"The major axis is the line that minimises the sum of squares of the shortest distances from the data points to the line. The shortest distance from a point to a line is perpendicular to it, so in this method residuals are measured perpendicular to the line."



Warton et al. 2006

Regression Types

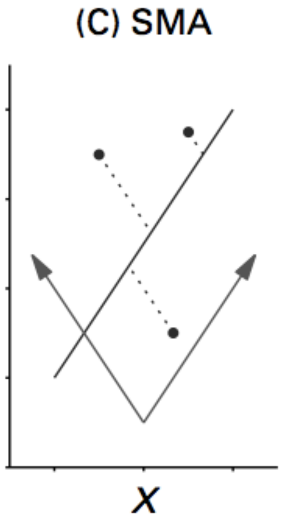
Standardized Major Axis (MA)

(a.k.a Geometric Mean Functional Relationship)
(a.k.a Reduced Major Axis)
(Along with MA, “model II regression”)

“The standardised major axis is the major axis calculated on standardised data, then rescaled to the original axes. This is typically done when two variables are not measured on comparable scales, in which case it might not seem reasonable to give the X and Y directions equal weight when measuring departures from the line.”

One way to think about it (from Smith 2009):
Minimize $\Sigma(X-\overline{X})(Y-\overline{Y})$

See Warton et al. 2006 Appendix for more derivations



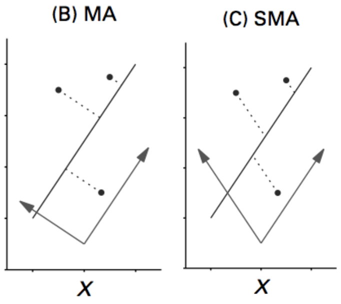
Warton et al. 2006

Regression Types

MA and SMA

MA and SMA are useful to…

- Summarize the relationship between two variables (a single dimension describes the two dimensions of the data)
- Contexts:
  - Allometry
  - ‘law-like’ relationships
  - Testing agreement of two methods of measurement



Warton et al. 2006

Regression Types

Summary of when to use what

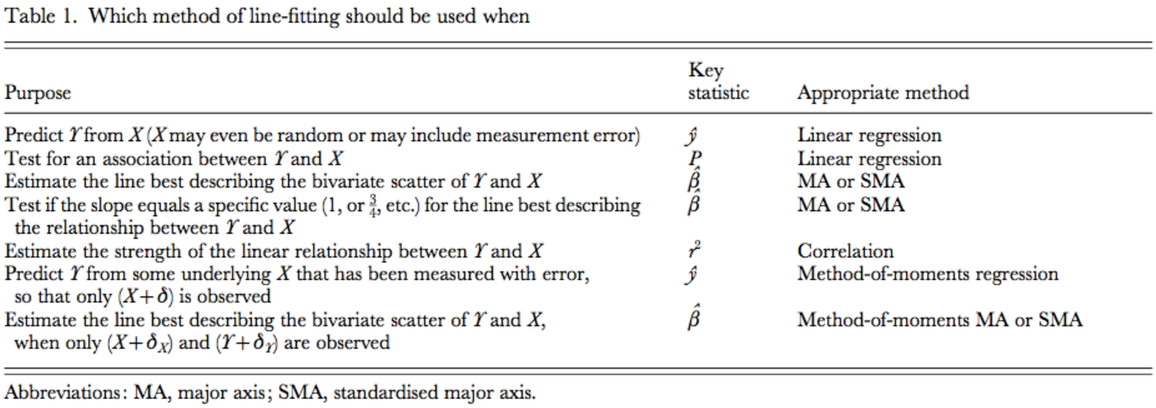


Table 1. Which method of line-fitting should be used when

Purpose	Key statistic	Appropriate method
Predict Y from X (X may even be random or may include measurement error)	$\hat{y}$	Linear regression
Test for an association between Y and X	P	Linear regression
Estimate the line best describing the bivariate scatter of Y and X	$\hat{\beta}$	MA or SMA
Test if the slope equals a specific value (1, or $\frac{3}{4}$, etc.) for the line best describing the relationship between Y and X	$\hat{\beta}$	MA or SMA
Estimate the strength of the linear relationship between Y and X	r^2	Correlation
Predict Y from some underlying X that has been measured with error, so that only $(X+\delta)$ is observed	$\hat{y}$	Method-of-moments regression
Estimate the line best describing the bivariate scatter of Y and X, when only $(X+\delta_X)$ and $(Y+\delta_Y)$ are observed	$\hat{\beta}$	Method-of-moments MA or SMA

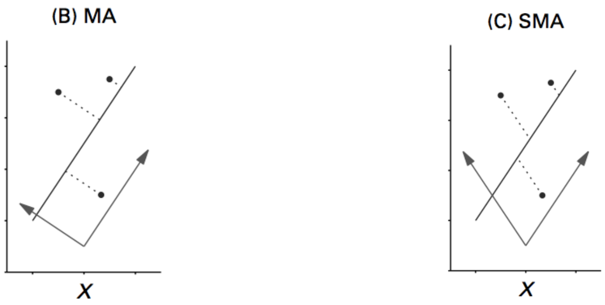
Abbreviations: MA, major axis; SMA, standardised major axis.

Warton et al. 2006

Regression Types

MA and SMA

But when do you use MA versus SMA?



Warton et al. 2006

But when do you use MA versus SMA?

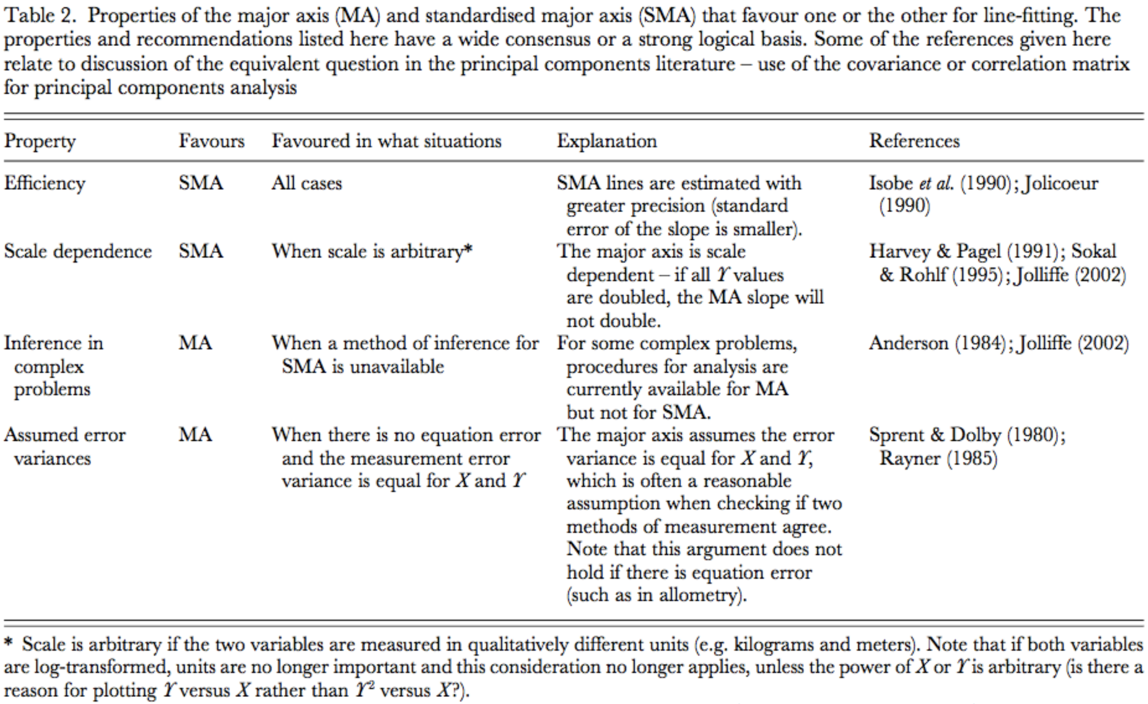


Table 2. Properties of the major axis (MA) and standardised major axis (SMA) that favour one or the other for line-fitting. The properties and recommendations listed here have a wide consensus or a strong logical basis. Some of the references given here relate to discussion of the equivalent question in the principal components literature – use of the covariance or correlation matrix for principal components analysis

Property	Favours	Favoured in what situations	Explanation	References
Efficiency	SMA	All cases	SMA lines are estimated with greater precision (standard error of the slope is smaller).	Isobe *et al.* (1990); Jolicoeur (1990)
Scale dependence	SMA	When scale is arbitrary*	The major axis is scale dependent – if all Y values are doubled, the MA slope will not double.	Harvey & Pagel (1991); Sokal & Rohlf (1995); Jolliffe (2002)
Inference in complex problems	MA	When a method of inference for SMA is unavailable	For some complex problems, procedures for analysis are currently available for MA but not for SMA.	Anderson (1984); Jolliffe (2002)
Assumed error variances	MA	When there is no equation error and the measurement error variance is equal for X and Y	The major axis assumes the error variance is equal for X and Y, which is often a reasonable assumption when checking if two methods of measurement agree. Note that this argument does not hold if there is equation error (such as in allometry).	Sprent & Dolby (1980); Rayner (1985)

* Scale is arbitrary if the two variables are measured in qualitatively different units (e.g. kilograms and meters). Note that if both variables are log-transformed, units are no longer important and this consideration no longer applies, unless the power of X or Y is arbitrary (is there a reason for plotting Y versus X rather than Y^2 versus X?).

Warton et al. 2006

Note: Also see table 3 of Warton et al. for more controversial relative merits of MA and SMA

Regression Types-Formulas

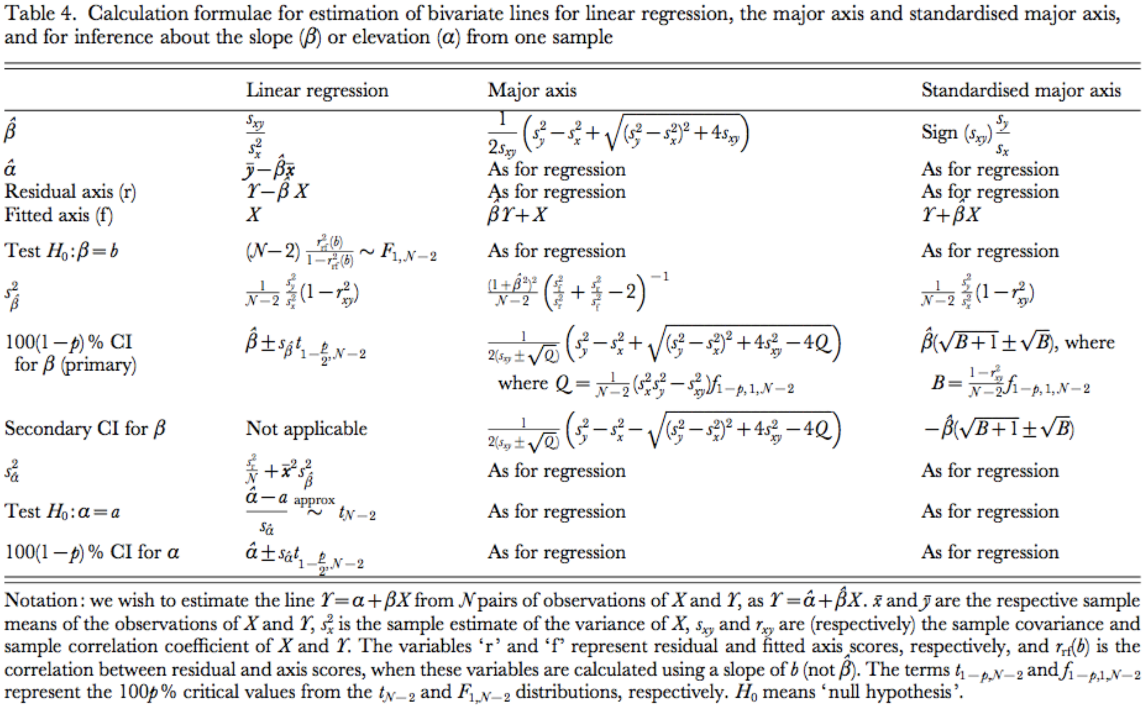


Table 4. Calculation formulae for estimation of bivariate lines for linear regression, the major axis and standardised major axis, and for inference about the slope (β) or elevation (α) from one sample

	Linear regression	Major axis	Standardised major axis
$\hat{\beta}$	$\frac{s_{xy}}{s_x^2}$	$\frac{1}{2s_{xy}}\left(s_y^2-s_x^2+\sqrt{(s_y^2-s_x^2)^2+4s_{xy}}\right)$	Sign $(s_{xy})\frac{s_y}{s_x}$
$\hat{\alpha}$	$\bar{y}-\hat{\beta}\bar{x}$	As for regression	As for regression
Residual axis (r)	$Y-\hat{\beta}X$	As for regression	As for regression
Fitted axis (f)	X	$\hat{\beta}Y+X$	$Y+\hat{\beta}X$
Test $H_0:\beta=b$	$(N-2)\frac{r_{rf}^2(b)}{1-r_{rf}^2(b)}\sim F_{1,N-2}$	As for regression	As for regression
$s_{\hat{\beta}}^2$	$\frac{1}{N-2}\frac{s_y^2}{s_x^2}(1-r_{xy}^2)$	$\frac{(1+\hat{\beta}^2)^2}{N-2}\left(\frac{s_f^2}{s_r^2}+\frac{s_r^2}{s_f^2}-2\right)^{-1}$	$\frac{1}{N-2}\frac{s_y^2}{s_x^2}(1-r_{xy}^2)$
$100(1-p)\%$ CI for β (primary)	$\hat{\beta}\pm s_{\hat{\beta}}t_{1-\frac{p}{2},N-2}$	$\frac{1}{2(s_{xy}\pm\sqrt{Q})}\left(s_y^2-s_x^2+\sqrt{(s_y^2-s_x^2)^2+4s_{xy}^2-4Q}\right)$ where $Q=\frac{1}{N-2}(s_x^2s_y^2-s_{xy}^2)f_{1-p,1,N-2}$	$\hat{\beta}(\sqrt{B+1}\pm\sqrt{B})$, where $B=\frac{1-r_{xy}^2}{N-2}f_{1-p,1,N-2}$
Secondary CI for β	Not applicable	$\frac{1}{2(s_{xy}\pm\sqrt{Q})}\left(s_y^2-s_x^2-\sqrt{(s_y^2-s_x^2)^2+4s_{xy}^2-4Q}\right)$	$-\hat{\beta}(\sqrt{B+1}\pm\sqrt{B})$
$s_{\hat{\alpha}}^2$	$\frac{s_r^2}{N}+\bar{x}^2s_{\hat{\beta}}^2$	As for regression	As for regression
Test $H_0:\alpha=a$	$\frac{\hat{\alpha}-a}{s_{\hat{\alpha}}}\overset{\text{approx}}{\sim}t_{N-2}$	As for regression	As for regression
$100(1-p)\%$ CI for α	$\hat{\alpha}\pm s_{\hat{\alpha}}t_{1-\frac{p}{2},N-2}$	As for regression	As for regression

Notation: we wish to estimate the line $Y=\alpha+\beta X$ from N pairs of observations of X and Y, as $Y=\hat{\alpha}+\hat{\beta}X$. $\bar{x}$ and $\bar{y}$ are the respective sample means of the observations of X and Y, s_x^2 is the sample estimate of the variance of X, s_{xy} and r_{xy} are (respectively) the sample covariance and sample correlation coefficient of X and Y. The variables 'r' and 'f' represent residual and fitted axis scores, respectively, and $r_{rf}(b)$ is the correlation between residual and axis scores, when these variables are calculated using a slope of b (not $\hat{\beta}$). The terms $t_{1-p,N-2}$ and $f_{1-p,1,N-2}$ represent the $100p\%$ critical values from the t_{N-2} and $F_{1,N-2}$ distributions, respectively. H_0 means 'null hypothesis'.

Warton et al. 2006

Example: Reich et al. 2010 paper

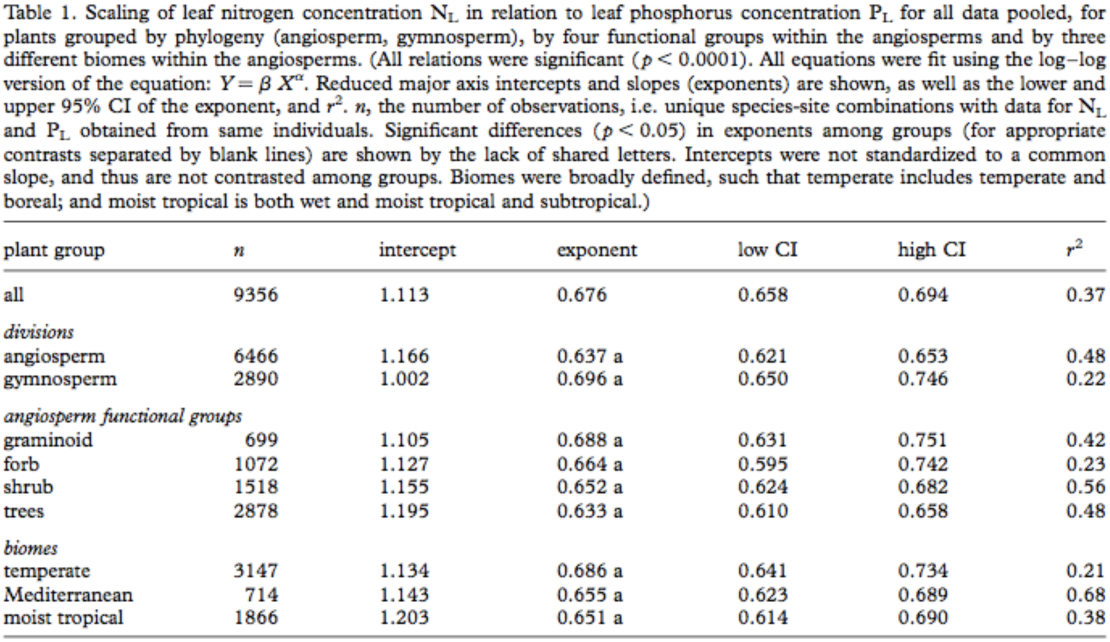
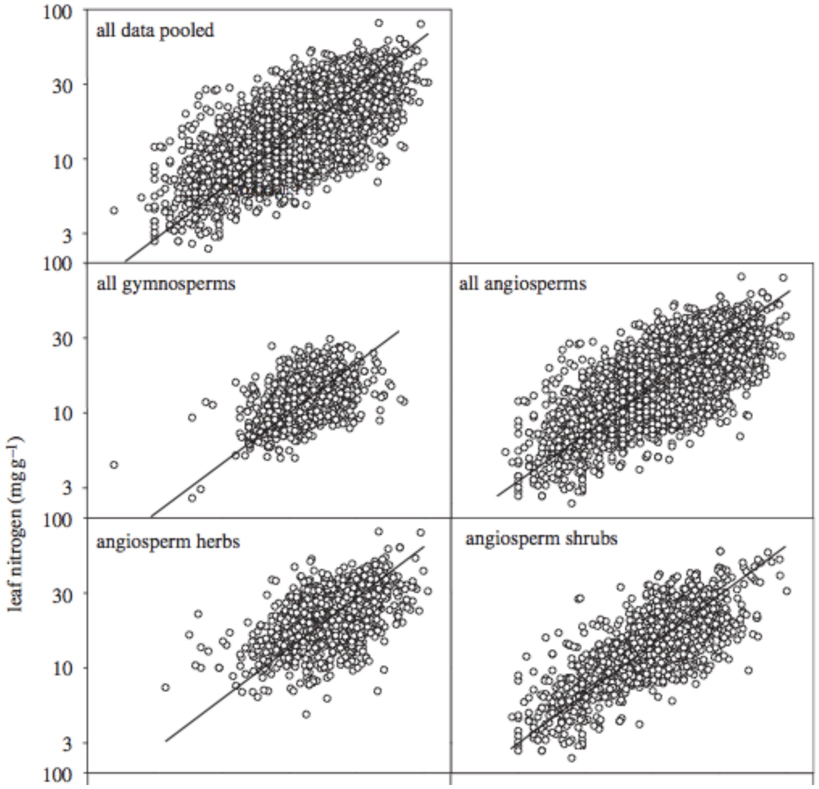


Table 1. Scaling of leaf nitrogen concentration N_L in relation to leaf phosphorus concentration P_L for all data pooled, for plants grouped by phylogeny (angiosperm, gymnosperm), by four functional groups within the angiosperms and by three different biomes within the angiosperms. (All relations were significant ($p < 0.0001$). All equations were fit using the log–log version of the equation: $Y = \beta X^{\alpha}$. Reduced major axis intercepts and slopes (exponents) are shown, as well as the lower and upper 95% CI of the exponent, and r^2. n, the number of observations, i.e. unique species-site combinations with data for N_L and P_L obtained from same individuals. Significant differences ($p < 0.05$) in exponents among groups (for appropriate contrasts separated by blank lines) are shown by the lack of shared letters. Intercepts were not standardized to a common slope, and thus are not contrasted among groups. Biomes were broadly defined, such that temperate includes temperate and boreal; and moist tropical is both wet and moist tropical and subtropical.)

plant group	n	intercept	exponent	low CI	high CI	r^2
all	9356	1.113	0.676	0.658	0.694	0.37
divisions						
angiosperm	6466	1.166	0.637 a	0.621	0.653	0.48
gymnosperm	2890	1.002	0.696 a	0.650	0.746	0.22
angiosperm functional groups						
graminoid	699	1.105	0.688 a	0.631	0.751	0.42
forb	1072	1.127	0.664 a	0.595	0.742	0.23
shrub	1518	1.155	0.652 a	0.624	0.682	0.56
trees	2878	1.195	0.633 a	0.610	0.658	0.48
biomes						
temperate	3147	1.134	0.686 a	0.641	0.734	0.21
Mediterranean	714	1.143	0.655 a	0.623	0.689	0.68
moist tropical	1866	1.203	0.651 a	0.614	0.690	0.38

Reich et al. 2010



Reich et al. 2010

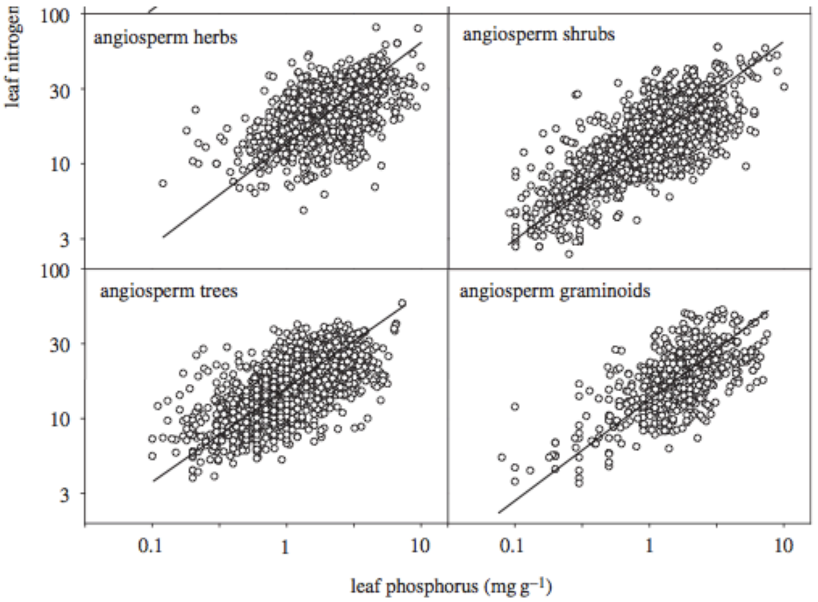


Figure 1. Relationships of leaf N (N_L) to leaf P (P_L) for all data pooled, and for plants grouped by both phylogeny (angiosperm, gymnosperm) and life form within the angiosperm group. The details of these relations using reduced major axis (RMA) regressions are presented in table 1.

Reich et al. 2010

Example: Reich et al. 2010 paper

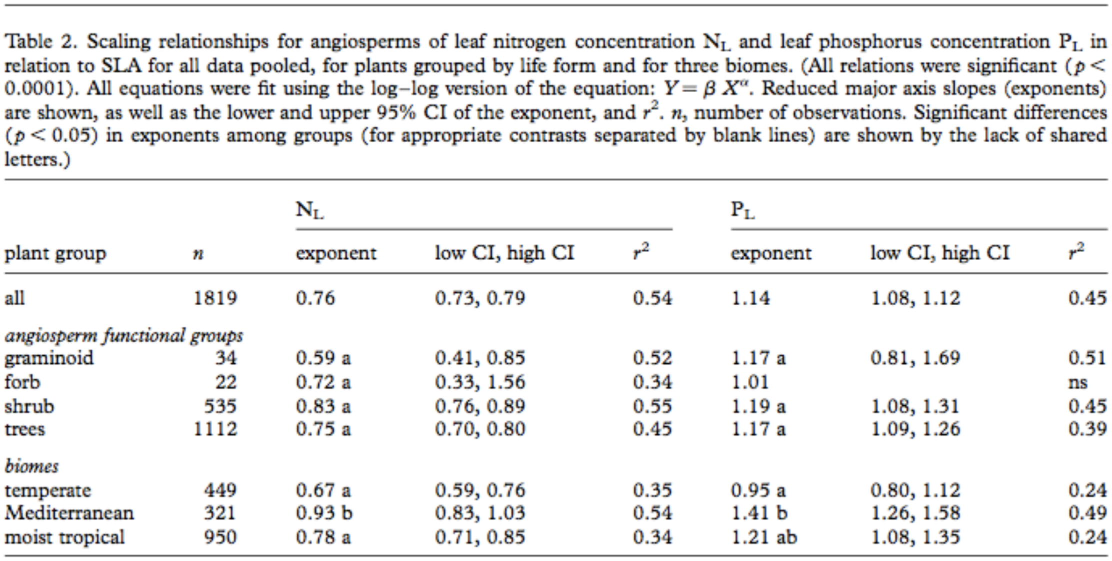


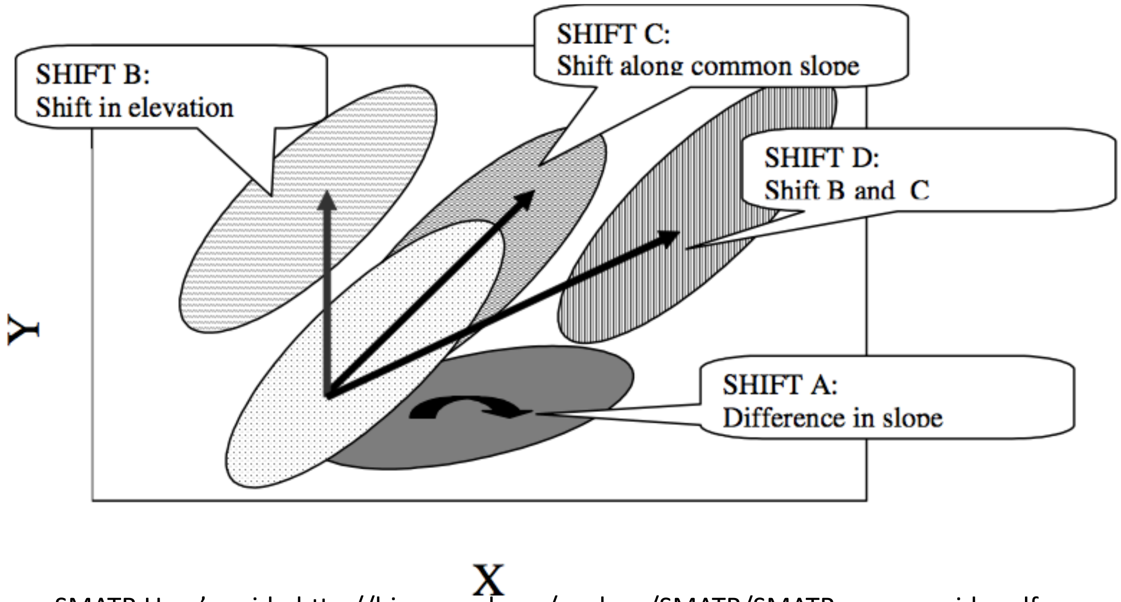
Table 2. Scaling relationships for angiosperms of leaf nitrogen concentration N_L and leaf phosphorus concentration P_L in relation to SLA for all data pooled, for plants grouped by life form and for three biomes. (All relations were significant ($p <$ 0.0001). All equations were fit using the log–log version of the equation: $Y = \beta X^{\alpha}$. Reduced major axis slopes (exponents) are shown, as well as the lower and upper 95% CI of the exponent, and r^2. *n*, number of observations. Significant differences ($p < 0.05$) in exponents among groups (for appropriate contrasts separated by blank lines) are shown by the lack of shared letters.)

		N_L			P_L		
plant group	*n*	exponent	low CI, high CI	r^2	exponent	low CI, high CI	r^2
all	1819	0.76	0.73, 0.79	0.54	1.14	1.08, 1.12	0.45
angiosperm functional groups							
graminoid	34	0.59 a	0.41, 0.85	0.52	1.17 a	0.81, 1.69	0.51
forb	22	0.72 a	0.33, 1.56	0.34	1.01		ns
shrub	535	0.83 a	0.76, 0.89	0.55	1.19 a	1.08, 1.31	0.45
trees	1112	0.75 a	0.70, 0.80	0.45	1.17 a	1.09, 1.26	0.39
biomes							
temperate	449	0.67 a	0.59, 0.76	0.35	0.95 a	0.80, 1.12	0.24
Mediterranean	321	0.93 b	0.83, 1.03	0.54	1.41 b	1.26, 1.58	0.49
moist tropical	950	0.78 a	0.71, 0.85	0.34	1.21 ab	1.08, 1.35	0.24

Reich et al. 2010

Possible tests with SMATR

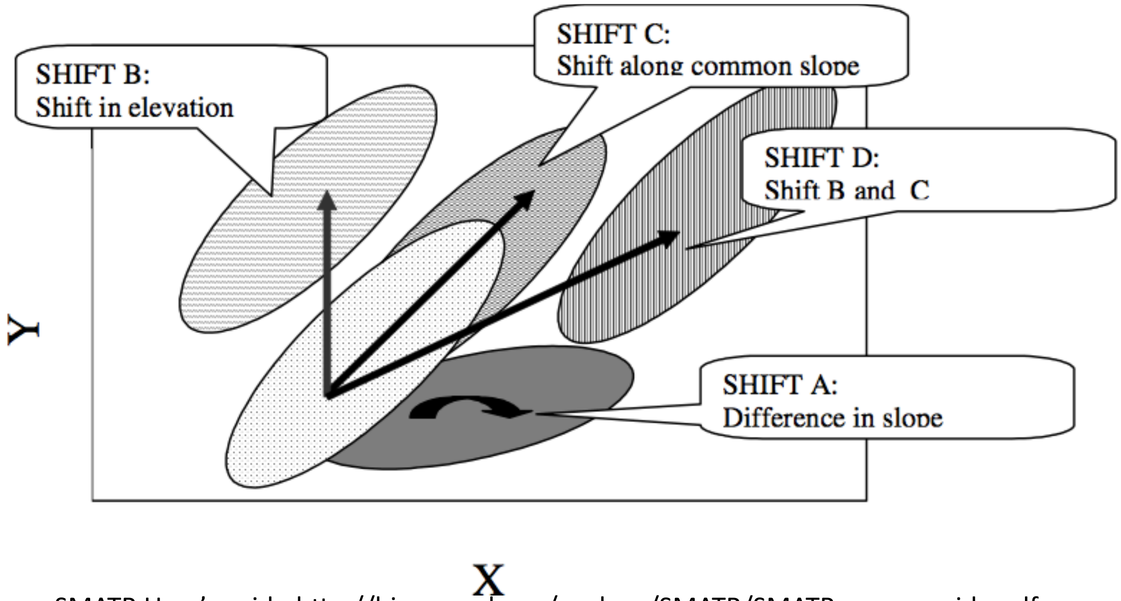
1. Test for differences among two or more fitted slopes (shift A)



From SMATR User’s guide:http://bio.mq.edu.au/ecology/SMATR/SMATR_users_guide.pdf

Possible tests with SMATR

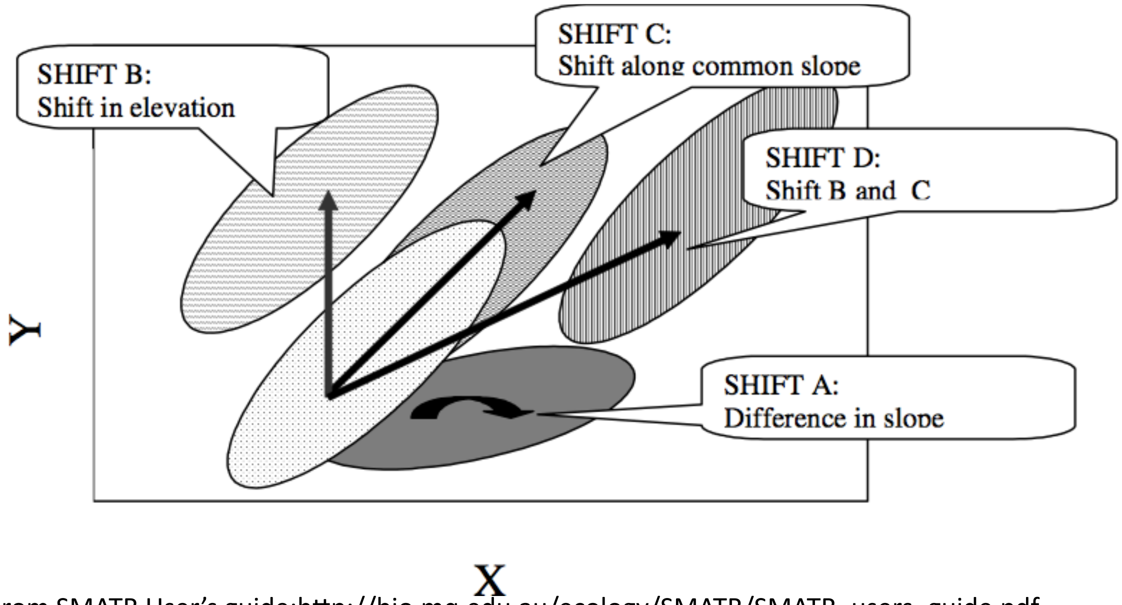
1. Test for differences among two or more fitted slopes (shift A)
2. If fitting a common slope can be justified (i.e., Test 1 is n.s.), test whether the fitted slopes share a common elevation (shift B)



From SMATR User’s guide:http://bio.mq.edu.au/ecology/SMATR/SMATR_users_guide.pdf

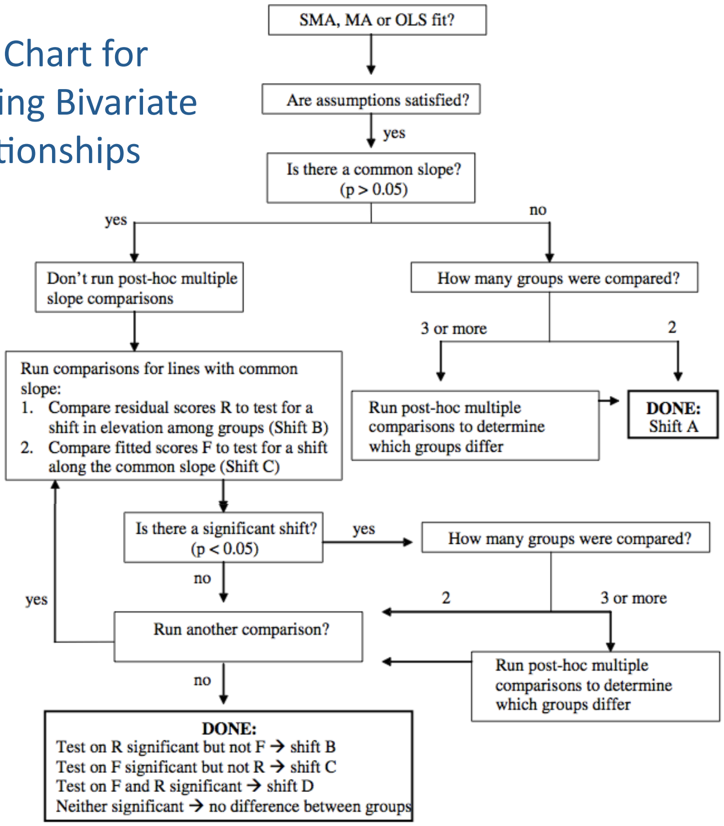
Possible tests with SMATR

1. Test for differences among two or more fitted slopes (shift A)
2. If fitting a common slope can be justified (i.e., Test 1 is n.s.), test whether the fitted slopes share a common elevation (shift B)
3. If fitting a common slope and elevation is justified, test whether the slopes fitted to each group are significantly separated along the common slope (shift C).



From SMATR User’s guide:http://bio.mq.edu.au/ecology/SMATR/SMATR_users_guide.pdf

Flow Chart for Comparing Bivariate Relationships



SMA, MA or OLS fit?

Are assumptions satisfied?

yes

Is there a common slope? (p > 0.05)

yes

Don’t run post-hoc multiple slope comparisons

Run comparisons for lines with common slope:

1. Compare residual scores R to test for a shift in elevation among groups (Shift B)
2. Compare fitted scores F to test for a shift along the common slope (Shift C)

no

How many groups were compared?

3 or more

Run post-hoc multiple comparisons to determine which groups differ

2

DONE: Shift A

Is there a significant shift? (p < 0.05)

yes

How many groups were compared?

2

3 or more

Run post-hoc multiple comparisons to determine which groups differ

no

Run another comparison?

yes

no

DONE:
Test on R significant but not F → shift B
Test on F significant but not R → shift C
Test on F and R significant → shift D
Neither significant → no difference between groups

Figure from SMATR User’s guide:http://bio.mq.edu.au/ecology/SMATR/SMATR_users_guide.pdf

Relevant Resources

Papers that use the R SMATR package:

http://bio.mq.edu.au/ecology/SMATR/pubs.html

Another R package for model II regression:

http://cran.r-project.org/web/packages/lmodel2/vignettes/mod2user.pdf

SMATR documentation:

http://cran.r-project.org/web/packages/smatr/smatr.pdf

Plasticity and Allometry protocol:

http://prometheuswiki.publish.csiro.au/tiki-index.php?page=Plasticity+and+Allometry+protocol

Works Cited

Reich et al. 2010. Evidence of a general 2/3-power law of scaliing leaf nitrogen to phosphorus among major plant groups and biomes. *Proc. R. Soc. B.* 277: 877-883

Smith 2009. Use and misuse of the Reduced Major Axis for line-fitting. *Am. J. Phys. Anth.* 140: 476-486.

Warton et al. 2006. Bivariate line-fitting methods for allometry. *Biol. Rev.* 81: 259-291.

Warton website documents online: http://bio.mq.edu.au/ecology/SMATR/