Technical note

'Professor Regressor': A computer programme for rapid processing of large sets of data for pairwise regression analyses in palaeolontological contexts†

Susan J. Dykes[1]* & Richard D. Dykes[2]

[1]*Evolutionary Studies Institute, University of the Witwatersrand, Private Bag 3, WITS 2050, Gauteng, South Africa*

[2]*P.O. Box 81, Cramerview, 2060 South Africa*

Received 6 February 2015. Accepted 1 April 2015

INTRODUCTION

Pairwise regression analyses of cranio-dental and other skeletal measurements are useful for showing similarity/dissimilarity metrics between specimens within a sample. Applied versions of regression methodologies, such as 'Log se_m' (Log-Transformed Standard Error of the Slope) (Thackeray 1997; Thackeray *et al.* 1997; Braun *et al.* 2004; Thackeray & Odes 2013); ATD (Average Taxonomic Distance) (Aiello *et al.* 2000; Richmond & Jungers 1995); STET (Standard Error Test) (Wolpoff & Lee 2001); and S_{LR} (Standard Deviation of Logged Ratios) (Gordon & Wood 2013) are used in the palaeontological sciences to assess:

- the range of metric variability between specimens representing any one species. With fossil species, the range of metric variability is established *a priori* using extant species that are argued to be an analogue to the selected fossil species;
- the similarity that exists between any two specimens and whether such similarity falls within the established range of variability for a single species;
- central tendency of log se_m values for many vertebrate species in relation to mean, range and standard deviation of log se_m values, which could potentially define a typical or 'benchmark' species in terms of its expected metric variability;
- the statistical probability that one specimen belongs to the same species as any other specimen.

This technical note details a project concerning the use of Microsoft VBA® (Visual Basic for Applications) code for the rapid processing of large volumes of pairwise regression analyses (matrices of several hundred results to tens of thousands of results at a time) in Microsoft Excel®. The complete annotated script for the 'Professor Regressor' programme is provided in the Supporting Online Materials (SOM).

The objective is to create a series of matrices of variables that are useful not only for general regression and ANOVA analyses but more specifically for Log Se_m analyses, for which no existing coding has been written. A user interface built into the spreadsheet enables specific rows or vectors from any one of the matrices to be aligned with corresponding rows or vectors from any other of the matrices, to produce bivariate or multivariate plots.

THE CHOICE OF VBA® AS THE PROGRAMMING LANGUAGE USED

There are a number of statistical coding tools in the market. The choice of Microsoft VBA®, Visual Basic for Applications, largely came down to familiarity with the software (Microsoft Excel®) amongst the users and the authors, which lowered the barriers to adoption within the community. Table 1 sets out some of the strengths of each programme considered.

Microsoft's built-in graphics and analytics capability made Excel® and VBA® an excellent end-to-end product for the purposes of this iteration of the programme. The only limitation is speed, which was not an issue when considering the applied scope of this version.

REGRESSION ANALYSES AND LOG se_m ANALYSES: AN OVERVIEW

Previous analyses involving regression or pairwise comparisons in palaeontological contexts have been undertaken by Thackeray (1997), Thackeray *et al.* (1997), Thackeray (2007), Braun *et al.* (2004); Thackeray & Odes (2013) and Dykes (2014).

Log se_m analyses are applied regression analyses wherein the log-transformed standard error of the slope *m* of the regression line reflects the degree of scatter of data points around the line: near-identical specimens will be so highly correlated that all data points will be plotted on or near the regression line, with a low standard error value; dissimilar specimens will have data points plotted away from the regression line, with a high standard error value in at least one of the two possible regression analyses for each pairwise comparison. Figure 1 shows pairwise comparisons of cranial data for a conspecific pair of specimens (similar size and shape) and for an interspecific pair (different in both size and shape). In both instances, two regression lines are calculated, firstly with specimen 1 on the *x* axis and specimen 2 on the *y* axis, and secondly with specimen 2 on the *x* axis and specimen 1 on the *y* axis. If the

†Supporting online information for this article is permanently archived at: http://hdl.handle.net/10539/17371

*Author for correspondence. E-mail: susan.dykes@students.wits.ac.za

Table 1. Comparative table of various statistical coding tools and the strengths of each.

Name	Description	Strengths
Microsoft Excel® and VBA®	Microsoft's office platform is a well-socialized set of tools used on computers world-wide, encompassing names such as Microsoft Word®, Excel®, etc. VBA® is Microsoft's scripting language that was designed to be used within these programs.	• User(s)'s familiarity with Microsoft Office® (reducing barriers to use) • The compatibility of prior work and input data (already in Excel® format and ease of manipulation) • Excel's® established and built-in regression tool (increase in credibility) • Licences already acquired and compatible with stakeholders (no start-up cost) • Built-in graphical capabilities and further analysis (within Excel®)
R®	An updated version of the statistical software S®, R® gives developers a simple command line-like interface to manipulate raw data and perform statistical operations on it. It is a member of an open source community which continually improves the software base in the form of 'packages'	• Excellent statistical analytics suite • Multiplatform • A growing database of packages makes the uses more and more versatile
Matlab®	Mathematical language. Matlab® is a well-known programming language with a number of good freeware analogues. Matlab® also allows developers to develop and freely distribute modules of code to constantly improve the program.	• Excellent statistical analytics suite • Multiplatform • A growing database of modules makes the uses more and more versatile
Python®, C®, C#®, C++®	Compiler and scripting languages can be effectively used for machine learning techniques such as regression. They are well-accepted in the market and provide multiplatform support. They are, however, very specific once the program is compiled and users generally have no way of modifying the program once it has been released to them.	• Extremely versatile but require expert knowledge to develop in • Most languages include modules developed by other constituencies

specimens are of similar size and shape, the slopes of both lines will be close to a value of 1, the degree of scatter will be similar (because residuals measured along the *x* axis are almost identical to residuals measured along the *y* axis) and therefore both log se_m values will be similar. If the specimens are different in size and shape (as may be the case for interspecific pairs of specimens), the slopes will be very different from each other and the degree of scatter around the regression line will be different in each case. Thus, the two log se_m values for interspecific pairs are expected to be extremely dissimilar. The differential between the two log se_m values for each pairwise comparison is called the 'delta value' (Thackeray, 2014; Dykes, 2014). Low log se_m values generally indicate a high probability of conspecificity, on the condition that the delta value is also low (less than 0.3) (Thackeray, 2014).

Log se_m values can be compared and mean, range and confidence limits calculated for both within-species (conspecific) and between-species (interspecific) pairwise comparisons. To avoid Type I and Type II errors occurring (the erroneous assumption of conspecificity where there is none, or of non-specificity where there is such), the delta value (Thackeray 2014; Dykes 2014) should always be considered in conjunction with the two log se_m values for each pairwise comparison. Figure 2 presents a bivariate plot showing log se_m values along the *x* axis, plotted against delta values along the *y* axis for conspecific and interspecific pairs of specimens. For conspecific pairs (*Colobus* vs *Colobus*), both the log se_m values and the delta values are low. For interspecific pairs (*Colobus* vs *Gorilla*), at least one of the log se_m values for each pairwise comparison is extremely high and does not overlap with the log se_m values for conspecific pairs along the *x* axis; additionally the delta values for all comparisons between *Colobus* and *Gorilla* are high, with no overlap area along the *y* axis between conspecific comparisons and interspecific comparisons.

PROGRAMME SYNOPSIS

This programme is an automated approach to pairwise linear regression analyses to be applied to large data sets. Manual calculation of these values is onerous and subject to input error once the number of specimens in the analysis exceeds 20. For example, the 'Professor Regressor' programme reduces the time required for an analysis of 20 specimens (requiring 380 pairwise regressions to be carried out sequentially and the output of data to individual 20 × 20 matrices for each variable in the analysis) from approximately six hours to two minutes, and for an analysis of 200 specimens (39 800 pairwise regressions and 200× 200 matrices for each variable in the analysis) from 444 hours to just over three hours.

An objective of the VBA® Coding Project was to construct an inbuilt Microsoft Excel® macro to manage these large-scale pairwise regression analyses and the output of large data matrices calculating for variable such as: regression statistics; ANOVA calculations; intercept and slope values; error values; *t*-stats; *P*-values and confidence intervals. (This output capacity is standard with Data Analysis Add-In to Microsoft Excel®.) A further objective

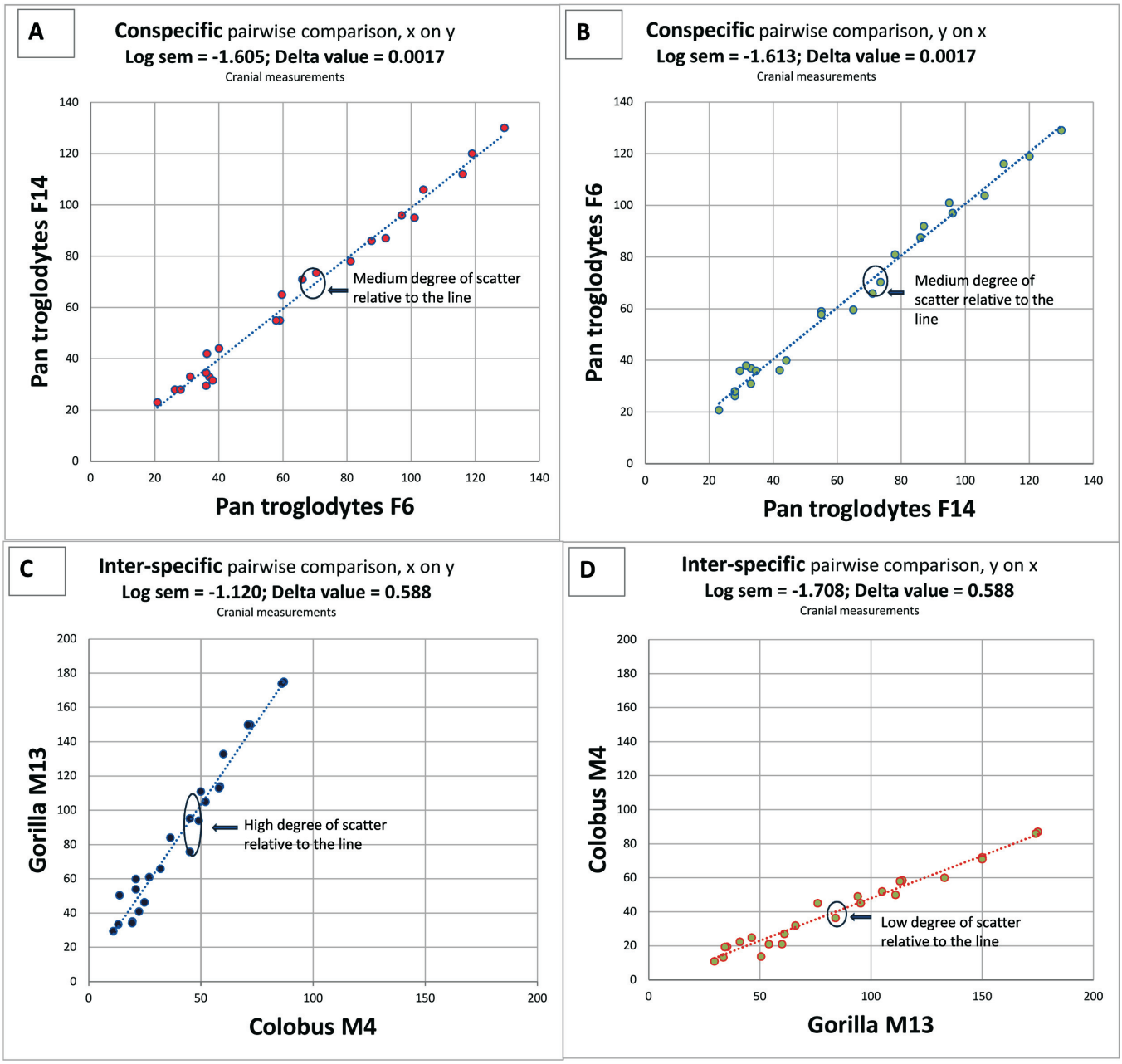


Figure 1. Examples of regression analyses in the context of log se_m analyses. **A**, Pairwise regression of conspecific specimens with specimen '*Pan troglodytes* F6' on the *x* axis and '*Pan troglodytes* F14' on the *y* axis; **B**, pairwise regression of conspecific specimens, using the same two specimens with the axes reversed. (These conspecific specimens are similar in size and shape, with the result that the two slopes are similar and the log se_m values are not only low but there is little differential (delta value) between them); **C**, pairwise regression of inter-specific specimens with specimen '*Colobus* M4' on the *x* axis and '*Gorilla* M13' on the *y* axis; **D**, pairwise regression of inter-specific specimens using the same two specimens with the axes reversed. (These specimens from different genera are significantly different in size and shape, with the result that the two slopes are different and the log se_m values are high in image C but low in image D. The differential (delta value) between the two log se_m values is high). Source of data: Gordon & Wood Supporting Online Material 1 (2013).

involved the delivery of additional matrices specific to log se_m analyses that had the capacity to calculate:

- The $\log_{10}$ value of the output value for the standard error of the slope (two log se_m values per pairwise comparison)
- The minimum, maximum and average values of the two log se_m values calculated for each pairwise comparison
- The differential or range ('delta value') between the minimum and maximum value of the two log se_m values produced for each pairwise comparison.

An interactive screen was added to select individual rows or columns/vectors from selected matrices and output these in alignment with corresponding rows or vectors from different matrices. For example, the average log se_m value for a series of pairwise comparisons could be aligned with the corresponding delta values enabling a bivariate plot to be constructed with ease.

Lastly, the facility to modify the input data to a 'standardized' format was included, wherein the original measurements for each specimen are scaled so that each data set begins at 0 and ends at 100, with the intervening measurements falling between these extremes on a pro-rata basis. This methodology is advised for additional confirmation of Type I and Type II error avoidance (as

Figure 2. Bivariate plot of conspecific and interspecific pairwise comparisons (log se_m values on the *x* axis and delta values on the *y* axis). By combining both values on the same chart, all possible overlaps between conspecific and interspecific log se_m values are removed and the two comparisons (conspecific *vs* interspecific) appear in separate clouds of data points, removing the likelihood of Type I and Type II errors. Source of data: Gordon & Wood Supporting Online Material 1 (2013)

highlighted by Gordon & Wood (2013)), due to mathematical artefacts caused by the differences in *y*-intercept values and gradient values for the two lines produced for each pairwise comparison (in such instances it is possible that a spuriously low log se_m value will be produced for either one of the paired (*x*-on-*y* or *y*-on-*x*) values produced for two specimens attributed to two different species).

STEP-THROUGH OF PROGRAMME

1. Follow the instructions on the 'Home' page of the 'Professor Regressor' spreadsheet. This involves pasting the data to be analysed/regressed in column format into the 'Input' worksheet of the spreadsheet, selecting the Type of regression output required (standardized, normal, or both) from a drop-down menu found in cell B3 of the 'Home' page and then clicking on the macro-linked button 'I Regress' (Fig. 3).
2. The initial 'sub' (sub-routine) that is executed is a decision shell. This determines whether the data are to be standardized or unaltered or whether both options are required, also whether the user requests the updating of an existing workbook. This makes use of the input specified in the workbook (SOM line 16).
3. A new workbook is created (preloaded excess MS Excel® sheets are removed) and sheets are added, renamed and given headers as appropriate (SOM line 60).
4. The programme copies the raw/standardized data into a clean sheet (SOM line 86)
5. The copied data are used to perform various functions (SOM line 147):
   a. The programme executes the MS Excel® Regression Analysis function for each pairwise input. It places regression output blocks (20 by 20 cells of statistical data output for each pairwise comparison) in block-matrix format in a new sheet named 'Regression Matrix' (Fig. 4; SOM line 265).
   b. Outputs are copied or calculated from the regression output block matrix and placed into individual matrices for each parameter (SOM line 273)
   c. Slope 'm' of the regression line (copied from the Regression Matrix to a new matrix on a sheet called 'Slope M')
   d. *R*-squared value (copied from the Regression Matrix to a new matrix on a sheet called 'R squared')
   e. *P*-value from the regression analysis (copied from the Regression Matrix to a new matrix on a sheet called 'P value')
   f. Intercept 'c' of the regression line (copied from the Regression Matrix to a new matrix on a sheet called 'Intercept')

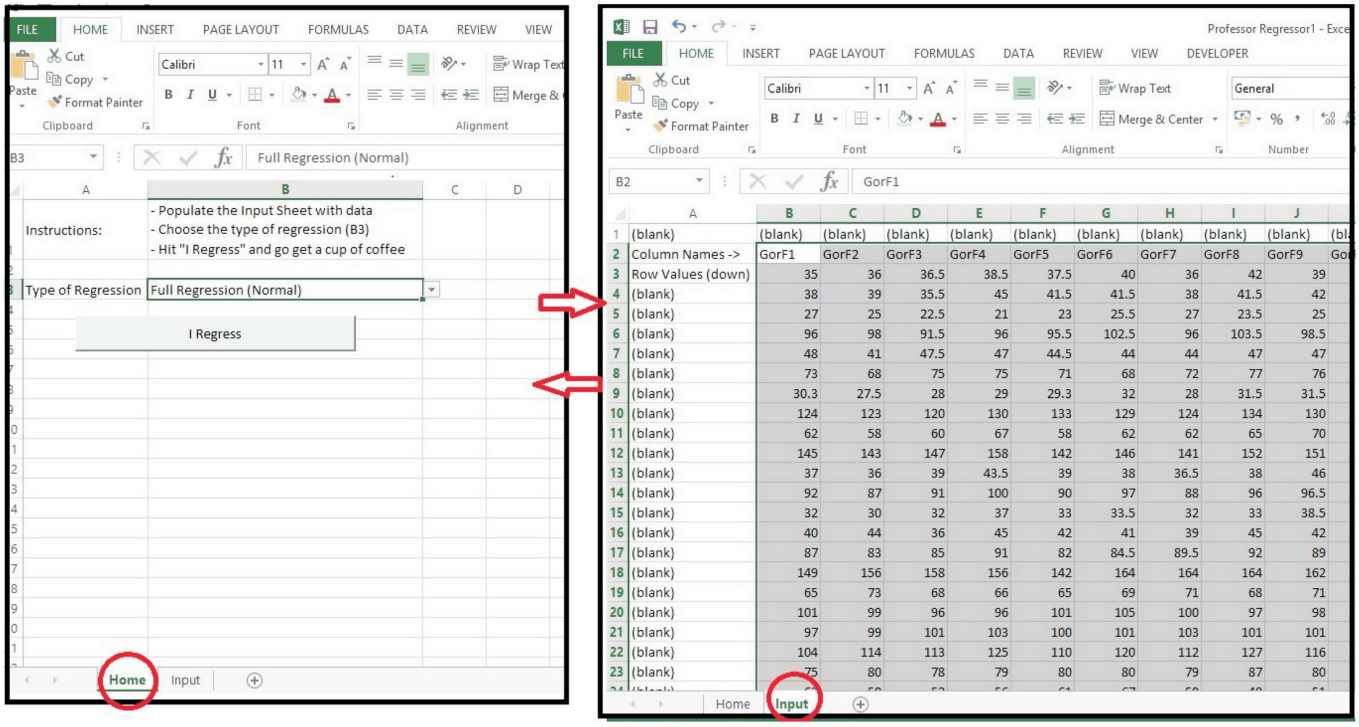


Figure 3. 'Home' page and 'Input' page of the Professor Regressor spreadsheet. Instructions for input of data are given and a choice is made by the user as to which kind of regression output is required (normal, standardized or both). The macro-enabled 'I Regress' button is found on the Home page.

g. Log se_m values (both x-on-y and y-on-x for each pairwise comparison) (calculated by taking the log transformed value of the standard error output on the Regression Matrix and placing these into a new matrix on a sheet called 'Log sem')

h. Minimum log se_m value (calculated by taking the two paired log se_m values (e.g. cells C2 and B3) from the Log se_m matrix and placing the lower of the two values into a new matrix on a sheet called 'Minimum')

i. Maximum log se_m value (calculated by taking the

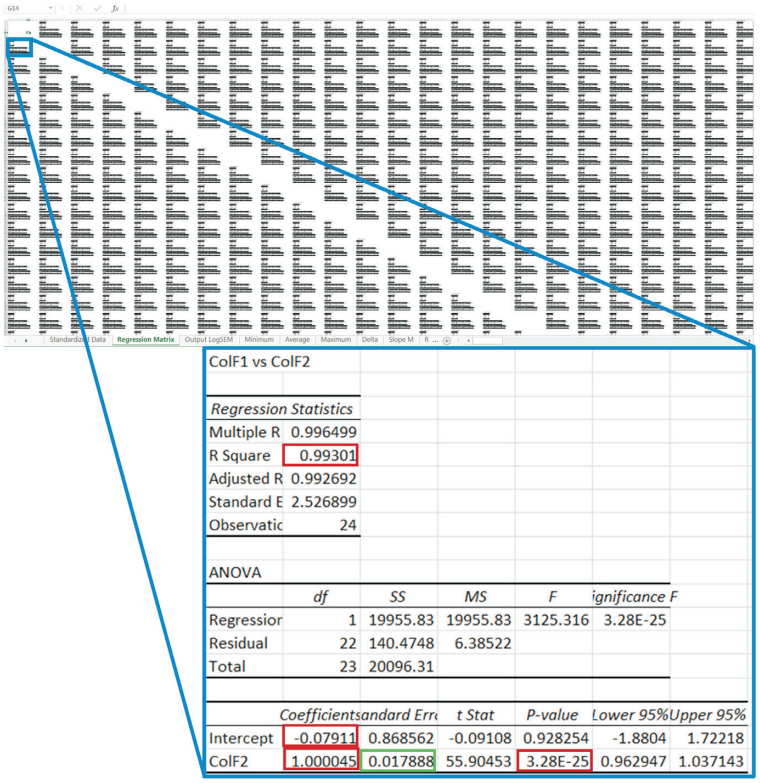
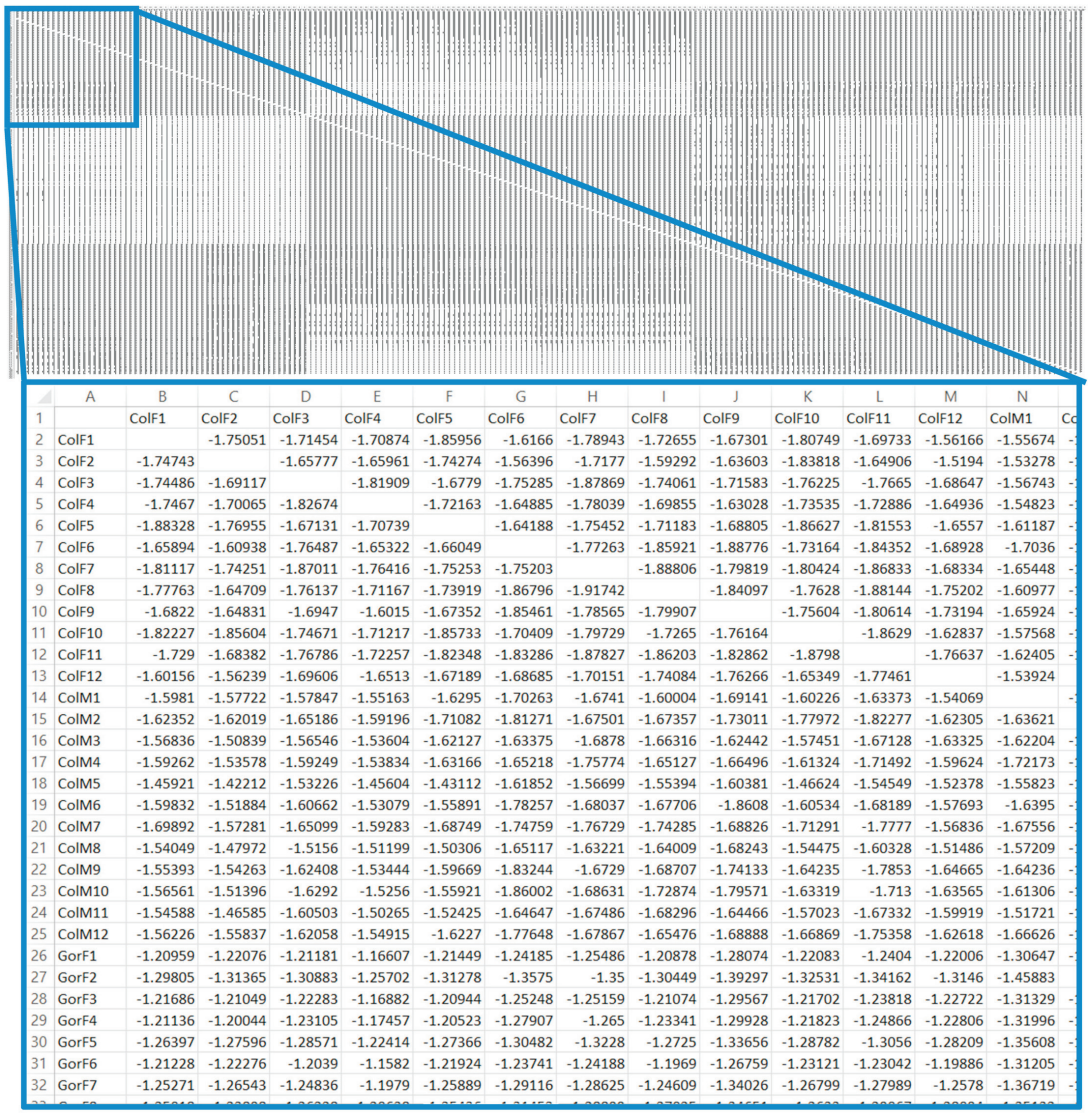


Figure 4. 'Regression Matrix' sheet. A 20 × 20 block of cells for each pairwise comparison is output from the regression analysis function of Microsoft Excel's® Data Analysis Add-In. Each output block is placed in a matrix, and from here some variables are located and copied to individual matrices for each variable (variables outlined in red/black), while the standard error value (outlined in green/grey) is recalculated to form the basis of log se_m-specific matrices.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1		ColF1	ColF2	ColF3	ColF4	ColF5	ColF6	ColF7	ColF8	ColF9	ColF10	ColF11	ColF12	ColM1	Co
2	ColF1		-1.75051	-1.71454	-1.70874	-1.85956	-1.6166	-1.78943	-1.72655	-1.67301	-1.80749	-1.69733	-1.56166	-1.55674	
3	ColF2	-1.74743		-1.65777	-1.65961	-1.74274	-1.56396	-1.7177	-1.59292	-1.63603	-1.83818	-1.64906	-1.5194	-1.53278	
4	ColF3	-1.74486	-1.69117		-1.81909	-1.6779	-1.75285	-1.87869	-1.74061	-1.71583	-1.76225	-1.7665	-1.68647	-1.56743	
5	ColF4	-1.7467	-1.70065	-1.82674		-1.72163	-1.64885	-1.78039	-1.69855	-1.63028	-1.73535	-1.72886	-1.64936	-1.54823	
6	ColF5	-1.88328	-1.76955	-1.67131	-1.70739		-1.64188	-1.75452	-1.71183	-1.68805	-1.86627	-1.81553	-1.6557	-1.61187	
7	ColF6	-1.65894	-1.60938	-1.76487	-1.65322	-1.66049		-1.77263	-1.85921	-1.88776	-1.73164	-1.84352	-1.68928	-1.7036	
8	ColF7	-1.81117	-1.74251	-1.87011	-1.76416	-1.75253	-1.75203		-1.88806	-1.79819	-1.80424	-1.86833	-1.68334	-1.65448	
9	ColF8	-1.77763	-1.64709	-1.76137	-1.71167	-1.73919	-1.86796	-1.91742		-1.84097	-1.7628	-1.88144	-1.75202	-1.60977	
10	ColF9	-1.6822	-1.64831	-1.6947	-1.6015	-1.67352	-1.85461	-1.78565	-1.79907		-1.75604	-1.80614	-1.73194	-1.65924	
11	ColF10	-1.82227	-1.85604	-1.74671	-1.71217	-1.85733	-1.70409	-1.79729	-1.7265	-1.76164		-1.8629	-1.62837	-1.57568	
12	ColF11	-1.729	-1.68382	-1.76786	-1.72257	-1.82348	-1.83286	-1.87827	-1.86203	-1.82862	-1.8798		-1.76637	-1.62405	
13	ColF12	-1.60156	-1.56239	-1.69606	-1.6513	-1.67189	-1.68685	-1.70151	-1.74084	-1.76266	-1.65349	-1.77461		-1.53924	
14	ColM1	-1.5981	-1.57722	-1.57847	-1.55163	-1.6295	-1.70263	-1.6741	-1.60004	-1.69141	-1.60226	-1.63373	-1.54069		
15	ColM2	-1.62352	-1.62019	-1.65186	-1.59196	-1.71082	-1.81271	-1.67501	-1.67357	-1.73011	-1.77972	-1.82277	-1.62305	-1.63621	
16	ColM3	-1.56836	-1.50839	-1.56546	-1.53604	-1.62127	-1.63375	-1.6878	-1.66316	-1.62442	-1.57451	-1.67128	-1.63325	-1.62204	
17	ColM4	-1.59262	-1.53578	-1.59249	-1.53834	-1.63166	-1.65218	-1.75774	-1.65127	-1.66496	-1.61324	-1.71492	-1.59624	-1.72173	
18	ColM5	-1.45921	-1.42212	-1.53226	-1.45604	-1.43112	-1.61852	-1.56699	-1.55394	-1.60381	-1.46624	-1.54549	-1.52378	-1.55823	
19	ColM6	-1.59832	-1.51884	-1.60662	-1.53079	-1.55891	-1.78257	-1.68037	-1.67706	-1.8608	-1.60534	-1.68189	-1.57693	-1.6395	
20	ColM7	-1.69892	-1.57281	-1.65099	-1.59283	-1.68749	-1.74759	-1.76729	-1.74285	-1.68826	-1.71291	-1.7777	-1.56836	-1.67556	
21	ColM8	-1.54049	-1.47972	-1.5156	-1.51199	-1.50306	-1.65117	-1.63221	-1.64009	-1.68243	-1.54475	-1.60328	-1.51486	-1.57209	
22	ColM9	-1.55393	-1.54263	-1.62408	-1.53444	-1.59669	-1.83244	-1.6729	-1.68707	-1.74133	-1.64235	-1.7853	-1.64665	-1.64236	
23	ColM10	-1.56561	-1.51396	-1.6292	-1.5256	-1.55921	-1.86002	-1.68631	-1.72874	-1.79571	-1.63319	-1.713	-1.63565	-1.61306	
24	ColM11	-1.54588	-1.46585	-1.60503	-1.50265	-1.52425	-1.64647	-1.67486	-1.68296	-1.64466	-1.57023	-1.67332	-1.59919	-1.51721	
25	ColM12	-1.56226	-1.55837	-1.62058	-1.54915	-1.6227	-1.77648	-1.67867	-1.65476	-1.68888	-1.66869	-1.75358	-1.62618	-1.66626	
26	GorF1	-1.20959	-1.22076	-1.21181	-1.16607	-1.21449	-1.24185	-1.25486	-1.20878	-1.28074	-1.22083	-1.2404	-1.22006	-1.30647	
27	GorF2	-1.29805	-1.31365	-1.30883	-1.25702	-1.31278	-1.3575	-1.35	-1.30449	-1.39297	-1.32531	-1.34162	-1.3146	-1.45883	
28	GorF3	-1.21686	-1.21049	-1.22283	-1.16882	-1.20944	-1.25248	-1.25159	-1.21074	-1.29567	-1.21702	-1.23818	-1.22722	-1.31329	
29	GorF4	-1.21136	-1.20044	-1.23105	-1.17457	-1.20523	-1.27907	-1.265	-1.23341	-1.29928	-1.21823	-1.24866	-1.22806	-1.31996	
30	GorF5	-1.26397	-1.27596	-1.28571	-1.22414	-1.27366	-1.30482	-1.3228	-1.2725	-1.33656	-1.28782	-1.3056	-1.28209	-1.35608	
31	GorF6	-1.21228	-1.22276	-1.2039	-1.1582	-1.21924	-1.23741	-1.24188	-1.1969	-1.26759	-1.23121	-1.23042	-1.19886	-1.31205	
32	GorF7	-1.25271	-1.26543	-1.24836	-1.1979	-1.25889	-1.29116	-1.28625	-1.24609	-1.34026	-1.26799	-1.27989	-1.2578	-1.36719	

Figure 5. Example of the 'Log se_m' output matrix sheet produced by the programme. At top, the full matrix is shown (this particular set of data had 212 specimens, thus the matrix contained 44732 log se_m values calculated by the programme). At bottom is an enlarged portion of the first few cells of the full matrix. (Source of data: Gordon & Wood Supporting Online Material 1 (2013)).

two paired values from the Log se_m matrix and placing the higher of the two values into a new matrix on a sheet called 'Maximum')

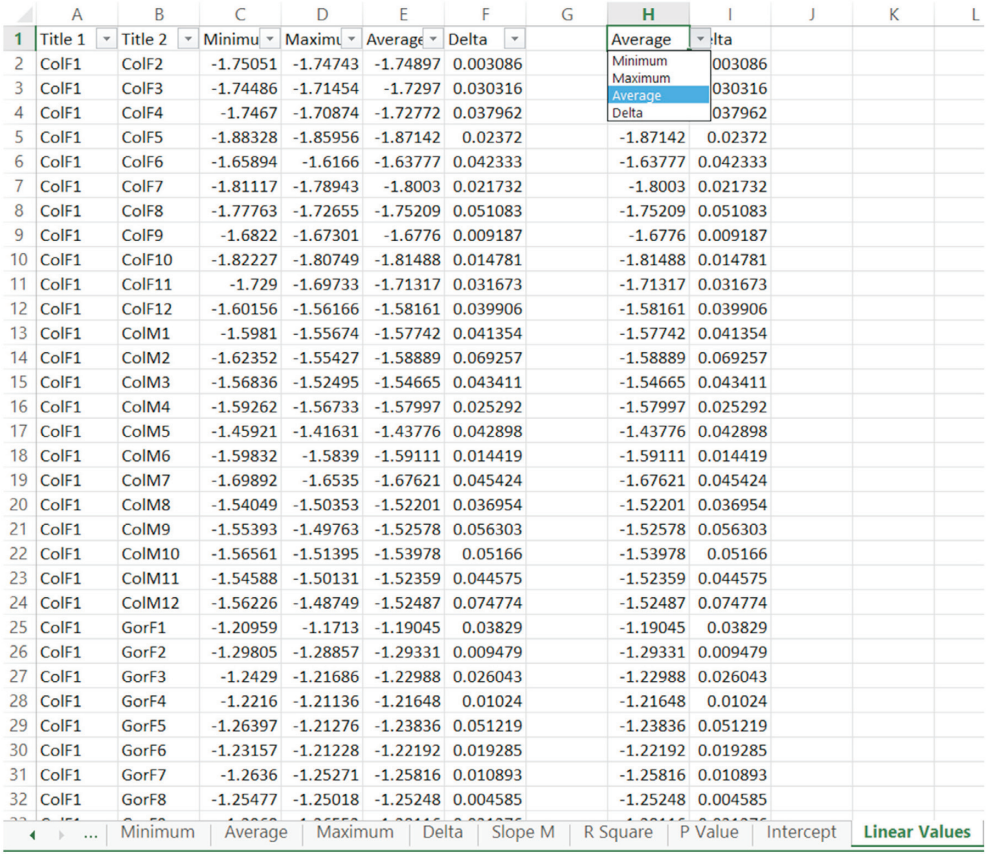
j. Average log se_m value (calculated by taking the two paired values from the Log se_m sheet and placing the average value into a new matrix on a sheet called 'Average')

k. Delta value (the differential (expressed as an absolute value) between the two log se_m values for each pairwise comparison) (calculated from the Log se_m sheet and placing the value into a new matrix on a sheet called 'Delta')

6. Figure 5 details the output matrix of calculations of log sem values for each pairwise comparisons showing the *y*-on-*x* and the *x*-on-*y* values (e.g. cells C2 and B3).
7. These results are loaded interactively into a linear version of the matrices. The purpose of this step is to transform the data from matrix format into column format for ease of use of Microsoft Excel's® graphing functions (see Fig. 6). The 'Linear Values' sheet of the spreadsheet allows any two variables (e.g. average log se_m value; delta value) to be aligned in column format for all specimens in the dataset, to produce scatter diagrams, frequency charts (histograms) and bivariate plots as desired (SOM line 303).
8. A blank sheet, the 'Sandbox' is loaded into the model for the user to carry out any related work without changing the original data in the model (SOM line 333).

LIMITATIONS AND SCOPE FOR FUTURE ADAPTATIONS

The programme has been applied to datasets of up to 200+ specimens, requiring consecutive processing in a pairwise manner against each other. For very large datasets, the macro will take a long time to run (several hours for outputs of 40 000+ pairwise comparisons), during which computer processing in general is slowed down and access to Microsoft Excel® is limited. Parts of the macro can, if necessary, be integrated with faster programming options (for instance, some sub-routines can be coded in R®, which is quicker than running a Microsoft Excel® macro), with user-friendly instructions as to how to operate the modified programme. However, for most



	A	B	C	D	E	F	G	H	I	J	K	L
1	Title 1	Title 2	Minimu	Maximu	Average	Delta		Average	lta			
2	ColF1	ColF2	-1.75051	-1.74743	-1.74897	0.003086		Minimum	003086			
3	ColF1	ColF3	-1.74486	-1.71454	-1.7297	0.030316		Maximum	030316			
4	ColF1	ColF4	-1.7467	-1.70874	-1.72772	0.037962		Average	037962			
5	ColF1	ColF5	-1.88328	-1.85956	-1.87142	0.02372		Delta	-1.87142	0.02372		
6	ColF1	ColF6	-1.65894	-1.6166	-1.63777	0.042333			-1.63777	0.042333		
7	ColF1	ColF7	-1.81117	-1.78943	-1.8003	0.021732			-1.8003	0.021732		
8	ColF1	ColF8	-1.77763	-1.72655	-1.75209	0.051083			-1.75209	0.051083		
9	ColF1	ColF9	-1.6822	-1.67301	-1.6776	0.009187			-1.6776	0.009187		
10	ColF1	ColF10	-1.82227	-1.80749	-1.81488	0.014781			-1.81488	0.014781		
11	ColF1	ColF11	-1.729	-1.69733	-1.71317	0.031673			-1.71317	0.031673		
12	ColF1	ColF12	-1.60156	-1.56166	-1.58161	0.039906			-1.58161	0.039906		
13	ColF1	ColM1	-1.5981	-1.55674	-1.57742	0.041354			-1.57742	0.041354		
14	ColF1	ColM2	-1.62352	-1.55427	-1.58889	0.069257			-1.58889	0.069257		
15	ColF1	ColM3	-1.56836	-1.52495	-1.54665	0.043411			-1.54665	0.043411		
16	ColF1	ColM4	-1.59262	-1.56733	-1.57997	0.025292			-1.57997	0.025292		
17	ColF1	ColM5	-1.45921	-1.41631	-1.43776	0.042898			-1.43776	0.042898		
18	ColF1	ColM6	-1.59832	-1.5839	-1.59111	0.014419			-1.59111	0.014419		
19	ColF1	ColM7	-1.69892	-1.6535	-1.67621	0.045424			-1.67621	0.045424		
20	ColF1	ColM8	-1.54049	-1.50353	-1.52201	0.036954			-1.52201	0.036954		
21	ColF1	ColM9	-1.55393	-1.49763	-1.52578	0.056303			-1.52578	0.056303		
22	ColF1	ColM10	-1.56561	-1.51395	-1.53978	0.05166			-1.53978	0.05166		
23	ColF1	ColM11	-1.54588	-1.50131	-1.52359	0.044575			-1.52359	0.044575		
24	ColF1	ColM12	-1.56226	-1.48749	-1.52487	0.074774			-1.52487	0.074774		
25	ColF1	GorF1	-1.20959	-1.1713	-1.19045	0.03829			-1.19045	0.03829		
26	ColF1	GorF2	-1.29805	-1.28857	-1.29331	0.009479			-1.29331	0.009479		
27	ColF1	GorF3	-1.2429	-1.21686	-1.22988	0.026043			-1.22988	0.026043		
28	ColF1	GorF4	-1.2216	-1.21136	-1.21648	0.01024			-1.21648	0.01024		
29	ColF1	GorF5	-1.26397	-1.21276	-1.23836	0.051219			-1.23836	0.051219		
30	ColF1	GorF6	-1.23157	-1.21228	-1.22192	0.019285			-1.22192	0.019285		
31	ColF1	GorF7	-1.2636	-1.25271	-1.25816	0.010893			-1.25816	0.010893		
32	ColF1	GorF8	-1.25477	-1.25018	-1.25248	0.004585			-1.25248	0.004585		

... Minimum Average Maximum Delta Slope M R Square P Value Intercept **Linear Values**

Figure 6. The 'Linear Values' sheet of the Professor Regressor spreadsheet. This sheet has a drop-down menu that enables the user to select columns from an output matrix of choice (e.g. Average log se_m value) and to align corresponding values from the outputs from another matrix (e.g. Delta value) in paired column format rather than in matrix format, to enable bivariate plots and other graphs to be produced.

requirements, the simplicity of the 'I Regress!' button compensates for the length of 'down time' involved. For very large datasets (200+ specimens), the macro can be set to run overnight.

In future iterations of the programme, other variables will be included (e.g. ANOVA results) and additional options for interactive data selections could be added to the 'Linear Values' page.

REFERENCES

AIELLO, L.C., COLLARD, M., THACKERAY, J.F. & WOOD, B.A. 2000. Assessing exact randomization-based methods for determining the taxonomic significance of variability in the human fossil record. *South African Journal of Science* **96,** 179–183.

BRAUN, S., THACKERAY, J.F. & LOOTS, M. 2004. Scientific notes: a morphometric technique to assess probabilities of conspecificity in extant primates and Plio-Pleistocene hominids. *Annals of the Transvaal Museum* **41,** 93–95.

DYKES, S.J. 2014. A morphometric analysis of hominin teeth attributed to different species of *Australopithecus, Paranthropus* and *Homo*. M.Sc. dissertation, University of the Witwatersrand, Johannesburg.

GORDON, A.D. & WOOD, B.A. 2013. Evaluating the use of pairwise dissimilarity metrics in paleoanthropology. *Journal of Human Evolution* **65**, 465–477.

RICHMOND, B.G. & JUNGERS, W.L. 1995. Size variation and sexual dimorphism in *Australopithecus afarensis* and living hominoids. *Journal of Human Evolution* **29**(3), 229–245.

THACKERAY, J.F. 1997. Probabilities of conspecificity. *Nature* **390**, 30–31,

THACKERAY, J.F. 2007. Approximation of a biological species constant? *South African Journal of Science* **103,** 489.

THACKERAY, J.F. 2014. Palaeoanthropology: probabilities of conspecificity. *PalNews, Biannual Newsletter of the Palaeontological Society of Southern Africa* **19**(4), 35–37.

THACKERAY, J.F., BELLAMY, C.L., BELLARS, D., BRONNER, G., BRONNER, L., CHIMIMBA, C., FOURIE, H., KEMP, A., KRÜGER, M., PLUG, I., PRINSLOO, S., TOMS, R., VAN ZYL, A.J. & WHITING, M.J. 1997. Probabilities of conspecificity: application of a morphometric technique to modern taxa and fossil specimens attributed to *Australopithecus* and *Homo*. *South African Journal of Science* **93,** 195–196.

THACKERAY, J.F. & ODES, E. 2013. Morphometric analysis of early Pleistocene African hominin crania in the context of a statistical (probabilistic) definition of a species. *Antiquity* **87**. http://antiquity.ac.uk/projgall/thackeray335/

WOLPOFF, M.H. & LEE, S-H. 2001. The late Pleistocene human species of Israel. *Bulletins et Mémoirs de la Société d'Anthropologie de Paris* **13,** 291–310.