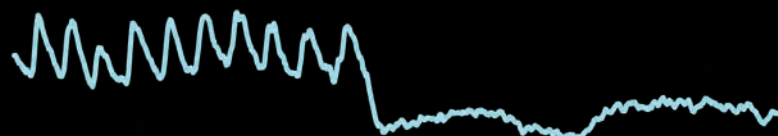
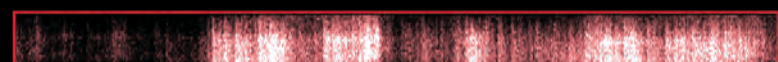
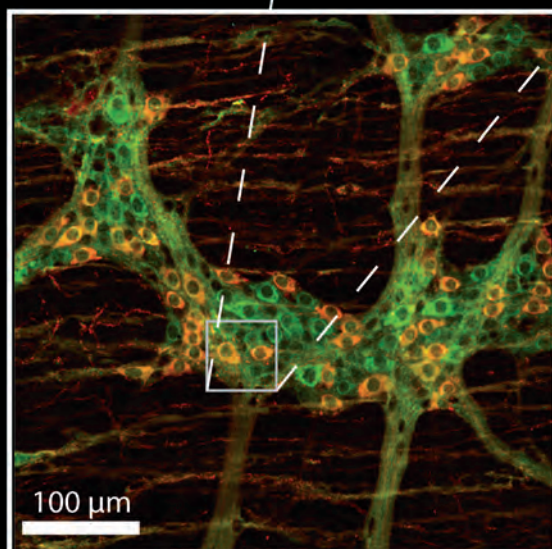
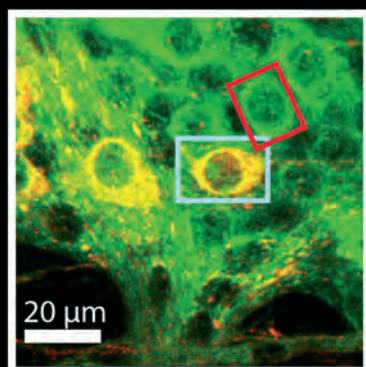


PHYSIOLOGY NEWS

autumn 2010 number 80



10s

Tim Biscoe presents a case for the importance of basic research
HIT to get fit: is it as good as high-volume continuous exercise?
Lymphatic vessels: absorptive sumps or leaky pumps?
The Society techniques workshops

500 µm



The Society's dog. 'Rudolf Magnus gave me to Charles Sherrington, who gave me to Henry Dale, who gave me to The Physiological Society in October 1942'

Published quarterly by The Physiological Society

Contributions and queries

Senior Production Editor
Jill Berriman

Editorial Administrator
Maev Fitzpatrick

The Physiological Society Publications Office
PO Box 502, Cambridge CB1 0AL, UK

Tel: +44 (0)1223 400180

Fax: +44 (0)1223 246858

Email: magazine@physoc.org

Website: www.physoc.org

Magazine Editorial Board

Editor

Austin Elliott

University of Manchester, Manchester, UK

Deputy Editor

Patricia de Winter

University College London, London, UK

Members

Angus Brown

University of Nottingham, Nottingham, UK

Sarah Hall

Cardiff University, Cardiff, UK

Munir Hussain

University of Bradford, Bradford, UK

John Lee

Rotherham General Hospital, Rotherham, UK

Thelma Lovick

University of Birmingham, Birmingham, UK

Samantha Passey

Bristol Heart Institute, Bristol, UK

Foreign Correspondents

John Hanrahan

McGill University, Montreal, Canada

John Morley

University of Western Sydney, NSW, Australia

Fiona Randall

Okinawa Institute of Science and Technology, Okinawa, Japan

© 2010 The Physiological Society

ISSN 1476-7996 (Print)

ISSN 2041-6512 (Online)

The Physiological Society is registered in England as a

company limited by guarantee: No 323575.

Registered office: Peer House, Verulam Street, London

WC1X 8LZ.

Registered Charity: No 211585.

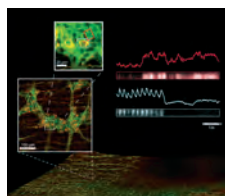
Printed by The Lavenham Press Ltd

PHYSIOLOGY NEWS

Editorial	3
Meetings	
Physiology 2010, University of Manchester	4
Mechanosensitivity: from transduction to sensation <i>Guy Bewick, Bob Banks</i>	6
Three months in the life of Professor Ole Petersen, Director of the Cardiff School of Biosciences	
<i>Ole Petersen, Sarah Hall</i>	7
Noticeboard	8
Soapbox	
<i>Tim Biscoe</i>	9
Techniques	
Multiple regression <i>Peter Cahusac</i>	12
Science News and Views	
Lymphatic vessels – absorptive sumps or leaky pumps? <i>Joshua Scallan, Virginia Huxley</i>	16
Electrical synapses synchronize motor output for tadpole swimming <i>Hong-Yan Zhang, Wen-Chang, Li, William Heitler, Keith Sillar</i>	19
HIT to get fit: metabolic adaptations to low-volume high- intensity interval training <i>Jonathan Little, Martin Gibala</i>	22
Can growth hormone strengthen the connective tissue of muscle and tendon? <i>Simon Doessing, Michael Kjaer</i>	25
Generation of complex neuronal behaviour in a mammalian nervous system <i>Peter Bayguinov, Grant Hennig, Terence Smith</i>	29
'Sensible to feeling as to sight' Some recent progress on mechanosensory transduction in mammals <i>Guy Bewick, Bob Banks</i>	33
Book Review	37
Reports	
Sense about Science Media Workshop <i>Laura McCallum</i>	38
Society for Biology <i>Mark Downs</i>	39
Unbelievable!	40
From the archives <i>Austin Elliott</i>	41
Society	
Parliamentary Links Day – science and the new Parliament <i>Liz Bell</i>	42
High jinks at Cheltenham <i>Liz Bell</i>	43
The Physiological Society Techniques Workshops <i>John Winpenny</i>	44
YPS at Physiology 2010 <i>Parini Mankad</i>	45
Mary Cotter awarded the first Otto Hutter Physiology Teaching Prize at Physiology 2010	47
Ask a physiologist!	48
The Society's journals	49
Obituary	
Richard Darwin Keynes <i>Christopher L-H Huang</i>	51



Advancing the science of life



Cover image: Structural and functional analysis of myenteric neurons in the myenteric plexus, from Bayguinov et al. p. 29.

Multiple regression

In this Techniques article Peter Cahusac explains multiple regression, a much used statistical procedure, but one that is frequently misunderstood or misused. Multiple regression allows the effects of many explanatory (independent) variables on the measured (dependent) variable to be analysed simultaneously for situations when a single explanatory variable fails to account for most of the variation in the dependent variable – a common occurrence. If you have ideas for future Techniques articles please email magazine@physoc.org.

We all know that correlation is a useful statistical technique widely used to assess the linear relationship between two variables. The correlation coefficient tells us the extent to which points in a scatter plot conform to a straight line, and by its polarity whether the relationship is positive or negative. The closely related technique of regression quantifies the relationship between the variables by providing an equation for the linear relationship in terms of slope and intercept, and therefore allows the prediction of values. Multiple regression takes the analysis one step further by allowing more than one independent variable (IV) to be used to predict (or explain) the dependent variable (DV). Using multiple predictors reflects more accurately the true relationship between variables – few phenomena

Table 1.

Heart rate	Intensity	Activity	Athlete
30	7	walk	Elite
35	8	walk	Elite
50	9	walk	Elite
32	8	walk	Elite
45	10	jog	Elite
50	11	jog	Elite
55	12	jog	Elite
52	11	jog	Elite
58	10	jog	Elite
65	11	sprint	Elite
70	12	sprint	Elite
75	14	sprint	Elite
70	4	walk	Couch potato
75	6	walk	Couch potato
85	7	jog	Couch potato
110	9	sprint	Couch potato
95	9	sprint	Couch potato
100	8	sprint	Couch potato

are dependent on a single IV. Multiple regression is a flexible statistical technique with wide general applicability. However, perhaps because of this flexibility, it is open to misuse and abuse. Tabachnick & Fidell (2007) is an excellent reference for multiple regression and other multivariate methods.

I have made up a small data set to illustrate how the technique can be useful (see Table 1). Let us say I recruited some athletes, mainly elite sporty types (Elite), but to make up numbers, a few others whose only interest in sport is watching it on TV (couch potatoes). I am interested in heart rate changes (DV) in response to the type of activity and intensity of exercise (IVs). Each participant was randomly selected to carry out a specified type of activity (walk, jog, sprint), while intensity was measured by the exercise machine on a scale from 0 to 20. Heart rate was



Peter Cahusac

measured at the end of the specified activity.

A scatter plot of the data, heart rate against exercise intensity, is shown in Fig. 1. The types of activity are indicated by different symbols in the plot. The line is the regression line (line of 'best fit') for heart rate on exercise intensity. Simple correlations between all four variables are shown in Table 2.

The simple correlation between heart rate and exercise intensity

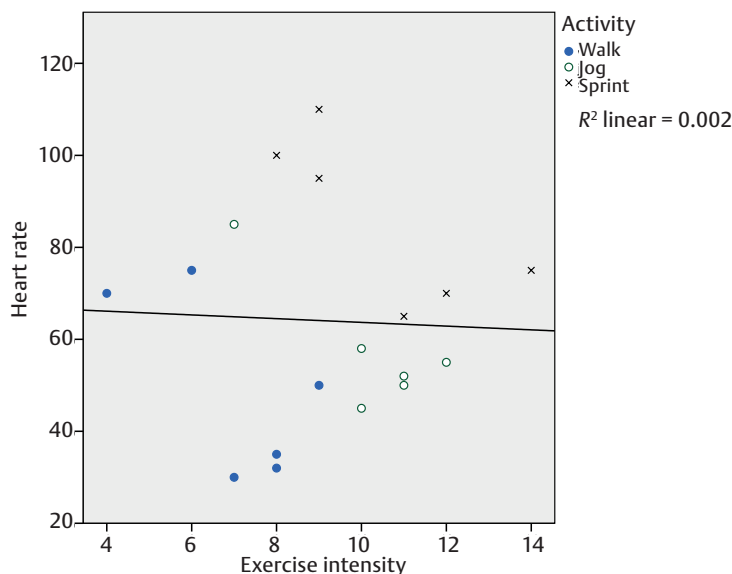


Figure 1. Scatter plot.

Table 2. Simple correlation coefficients for the data in Table 1. For all analyses, type of activity is coded as: 1, walk; 2, jog; 3, sprint; and type of athlete is coded as: 0, elite; 1, couch potato

Correlations				
		Exercise intensity	Activity	Type of athlete
Heart rate	Pearson Correlation	-.043	.672	.789
	p value (2-tailed)	.867	.002	.000
Exercise intensity	Pearson Correlation		.603	-.613
	p value (2-tailed)		.008	.007
Activity	Pearson Correlation			.144
	p value (2-tailed)			.568

Table 3. SPSS output for all the data given in Table 1.

Model Summary ^b					ANOVA ^b						
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Model	Sum of Squares	df	Mean Square	F	Sig.	
1	.978 ^a	.957	.947	5.320	1	Regression	8767.734	3	2922.578	103.254	.000 ^a
						Residual	396.266	14	28.305		
						Total	9164.000	17			

a. Predictors: (Constant), Type of athlete, Activity, Exercise intensity
 b. Dependent Variable: Heart rate

Coefficients ^a									
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics
		B	Std. Error	Beta			Lower Bound	Upper Bound	VIF
1	(Constant)	.519	9.677		.054	.958	-20.236	21.274	
	Exercise intensity	3.271	1.436	.344	2.278	.039	.191	6.352	7.369
	Activity	9.061	3.328	.328	2.723	.017	1.923	16.199	4.695
	Type of athlete	45.572	5.822	.952	7.827	.000	33.084	58.059	4.790

a. Dependent Variable: Heart rate

gives $r = -0.043$ ($P = 0.867$). That's right, like the line, the relationship is slightly negative. How can this be? Well, if we examine the scatter plot you will notice that there are two clusters of points – and these clusters correspond to the two types of athlete (couch potatoes, upper left; elite, lower right). For bivariate correlation and regression we have here a problem known as *heterogeneity of subsamples*, which would invalidate the analysis. However, if we include the type of athlete, and for good measure the specified activity, into a multiple regression analysis, then we can examine the effects of more than one IV on the DV heart rate. What we see in the plot is that within each cluster there is a clear positive relationship between heart rate and exercise intensity. However, elite athletes have lower heart rates and can work the exercise machine harder. So the apparent negative relationship between heart rate and exercise intensity is due to including two different (heterogeneous) types of athlete in our sample. The R^2 statistic measures how much variability is explained by the relationship between the variables, and here it is negligible at 0.2% (Fig. 1). We need to improve our analysis...

There are three general ways of performing a multiple regression:

standard, sequential (hierarchical) and statistical (stepwise). Here, for simplicity, we will use standard multiple regression, entering all IVs into the analysis simultaneously. All statistical packages will do the analysis, and I will use SPSS (also known as PASW) to illustrate. The output, with various options selected, looks something like Table 3.

SPSS provides a lot of information, but I am going to concentrate only on the essentials. The overall relationship is statistically significant now (see Table 3, ANOVA box) with $P < 0.001$, and the R^2 has dramatically improved to 95.7% (Table 3, Model summary). In the Coefficients box you can see that each of the IVs is now statistically significant ($P < 0.05$), and reassuringly, there is now a positive relationship between heart rate and exercise intensity (unstandardized coefficient, B). The coefficient for type of athlete is large and positive, and represents the large 'step' difference between elite athletes and couch potatoes (coded 0 and 1, respectively). Just like with simple regression we write out the equation, here it would be:

$$\text{Heart rate} = 0.519 + 3.271(\text{Exercise intensity}) + 9.061(\text{Activity}) + 45.572(\text{Type of athlete})$$

What multiple regression does is to calculate the optimal relationship between a combination of predictor IVs and the DV (by minimizing the squared residuals). Each variable's coefficient in the equation tells us how much the DV changes for each unit increase of that IV, while keeping all other IVs constant. Saying an IV is a 'predictor' does not mean that it has a causal relationship with the DV.

Regression analyses are characterised by numerous diagnostic tests. Initially these appear to be tedious formalities; however, with time and experience their usefulness and importance is increasingly valued. One such diagnostic is the *collinearity* (aka *multicollinearity*) statistics given in the last column of the Coefficients box. VIF stands for variance inflation factor, and informs us that the standard errors for the variable coefficients are inflated by between 4.695 and 7.369 times. An alternative statistic usually given is the *tolerance* which is merely the reciprocal of the VIF. Inflation of standard errors indicates instability of the regression equation, and arises because two or more IVs are strongly correlated (+ve or -ve) with each other. A rule of thumb is that you should be very concerned if there is a $VIF > 10$ (or $\text{tolerance} < 0.1$), and act to remove

Table 4. SPSS output for the data in Table 1, with Activity omitted.

Model Summary ^b					ANOVA ^b						
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Model	Sum of Squares	df	Mean Square	F	Sig.	
1	.966 ^a	.934	.925	6.357	1	Regression	8557.915	2	4278.957	105.900	.000 ^a
						Residual	606.085	15	40.406		
						Total	9164.000	17			

a. Predictors: (Constant), Type of athlete, Exercise intensity
 b. Dependent Variable: Heart rate

Coefficients ^a									
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics
		B	Std. Error	Beta			Lower Bound	Upper Bound	VIF
1	(Constant)	-17.571	8.406		-2.090	.054	-35.488	.346	
	Exercise intensity	6.731	.800	.707	8.410	.000	5.025	8.436	1.603
	Type of athlete	58.502	4.024	1.222	14.539	.000	49.926	67.079	1.603

a. Dependent Variable: Heart rate

one or more IVs. We should be a bit concerned about the 7.369 associated with exercise intensity. Since we are particularly interested in this variable as a predictor of heart rate we should look at other IVs for removal. Clearly, type of athlete is also crucial (that’s how we got into doing the multiple regression to avoid heterogeneous subsamples), so we could consider removing the type of activity. It is quite strongly correlated with exercise intensity ($r = 0.603$, Table 2); moreover, its standardized coefficient at 0.328 is the smallest – which means it is the weakest among the three predictors. Although it is often easy to find statistical reasons to include or remove IVs, the best reasons come from your understanding of the importance of particular variables for

the purpose of the analysis. Variables that are of theoretical importance, even if not statistically significant, should still be included in a multiple regression. Let us say here, for illustrative purposes, that we are primarily interested in exercise intensity and so remove the type of activity variable. The SPSS output is given in Table 4.

You can see that R^2 is still very large at 93.4% (Table 4, Model Summary box). The VIFs are now much lower, giving narrower 95% confidence intervals for the variable Bs (Coefficients box). In addition, the standardized coefficients (Betas) are higher and more statistically significant (both $P < 0.001$). Our equation is now:

$$\text{Heart rate} = -17.571 + 6.731(\text{Exercise intensity}) + 58.502(\text{Type of athlete})$$

We could predict the heart rate of a couch potato doing moderate exercise (say intensity of 9) as:

$$\text{Heart rate} = -17.571 + 6.731(9) + 58.502(1) = 101.510$$

With one IV, the relationship with the DV is two dimensional about a fitted line; with two IVs we can visualise the fit in three dimensions with data points scattered about a plane. With three or more IVs the points are scattered in hyperspace. In our example, two clusters of points occur at different places about a plane, each cluster determined by the type of athlete (see Fig. 2).

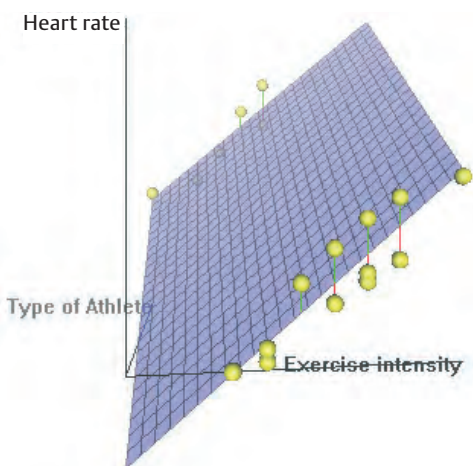


Figure 2. 3-D plot of the variables.

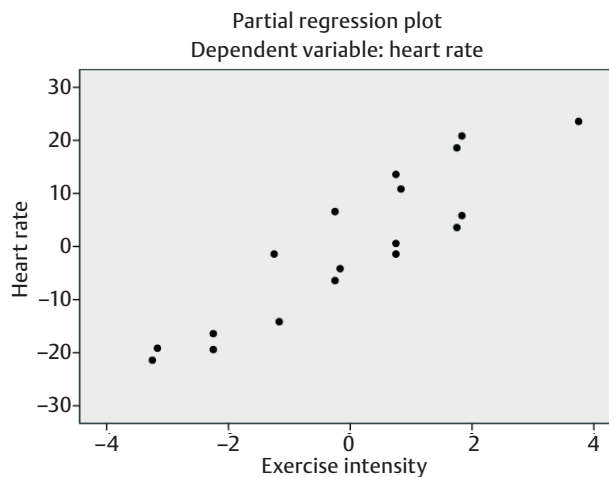


Figure 3. Partial regression plot. If the effect of different types of athlete is kept constant, then we can see that there is a clear positive linear relationship between heart rate and exercise intensity here. Scales are centred on the means.

In order to see the relationship of an individual IV and DV it is possible to do a *partial regression plot*. Of particular interest to us is intensity of exercise as a predictor of heart rate (see Fig. 3). Here we see a cigar-shaped and clear positive relationship between these variables (as we had expected).

Multiple regression is a useful general technique for analysing data where the DV (aka *outcome* or *criterion* variable) is associated with more than one IV (aka *explanatory* or *predictor* variable). It can handle continuous and dichotomous IVs (and indeed the DV can also be dichotomous, as in *logistic regression*). Regression can be done instead of ANOVA (by dummy variable coding the different levels of a factor), but ANOVA cannot necessarily be done using regression data if one or more variables are continuous (though can be done by converting to e.g. low/med/high – but with loss of information). ANOVA is a restricted form of regression. Actually our particular example could have been analysed in a between-participants ANOVA, entering type of athlete as a fixed factor and exercise intensity as a covariate, but it would not normally have produced the coefficients used to construct the equation.

Regression can include the outputs from other analyses. A good example is the use of output scores from a principal components analysis (PCA, see Patricia de Winter's article in the previous issue, PN79) to reduce the dimensionality of the variables used. In the above hypothetical study it may have been possible to derive a variable 'fitness' (known as a *latent* variable) if we had administered a questionnaire with numerous questions about participants' sporting activities (how often they train, how long, what type of activity, etc. etc.). The factor scores for the relevant component would then be entered as a variable into the regression analysis. More generally, if we had problems of collinearity among a set of variables

in our regression analysis, we could carry out a PCA on those variables, which would reduce them to a subset of uncorrelated factors. The factor scores would then be used in a regression analysis (however, we would need to be sure what each factor represents).

In practice, you should carefully choose which variables to enter, rather than just enter them all. Sometimes attempts are made to find the best equation by entering as many IVs as possible, regardless of their meaning. Unfortunately this is encouraged by stepwise procedures, and should be avoided, except perhaps for exploratory analyses. It is important to stress that for an accurate regression equation, all *relevant* IVs with respect to DV changes must be included in an analysis. Imagine if we had not included the type of athlete variable in our example (and had not recorded it), we would have been unable to interpret our data. If there are more IVs than cases then the regression equation predicts precisely the DV values. We could have included sex, type of sport, age, height, weight, health status, etc. etc. as predictors. This would lead to a better fit to the particular sample data, but paradoxically leads to a less useful result because of *overfitting* to the idiosyncrasies of our particular sample. This means that the results of our analysis might not be readily applicable to other data selected from the same population, i.e. there is poor generalization. One way to check how 'good' the regression equation is, is to apply cross-validation to the data. Here, a regression equation developed from a randomly selected large subset of the data is then used to predict scores from the remaining data. The predicted and actual scores are correlated, and the R^2 compared with the initial R^2 from the larger subset. We would expect the former value to be similar though slightly lower than the latter.

As mentioned above, regression analyses come with numerous

diagnostics. In order to check assumptions for the analysis, it is useful to look at residuals for unexplained variability, outliers or non-linearity. A plot of standardized residuals against standardized predicted values should show a random cloud of points. A departure from that pattern suggests a bad fit. It may indicate that an important IV is missing. Alternatively, a clear pattern among the points, such as a curved wave or increasing spread (funnel shaped) with increasing predicted values, could indicate that an IV or DV should be transformed (e.g. square and log transform, respectively). In some situations it may be appropriate to look at an interaction between IVs, by multiplying two or more of them together, and entering the resultant product, along with the individual IVs, into the analysis. Non-independence of residuals indicates another variable is in play (e.g. the order in which the data were collected), and needs to be taken into account. The residuals should also be normally distributed. An important assumption is linearity. If a factor with m levels is non-linear with respect to the DV, it can be converted into $m - 1$ *dummy* variables (coding each with 0 and 1), ensuring linearity since a straight line always connects between the two points $0, y_1$ and $1, y_2$. A number of diagnostics help detect outliers which might exert excessive leverage within the equation.

Finally, it is necessary to say something about sample size. Our fabricated example was clearly deficient. Generally, with medium-sized effects, you will need at least 50 participants + (8 x no. of IVs). So in our example we would need 50 + (3 x 8) = 74. Otherwise, the more the better.

Peter Cahusac
Stirling University

References

- de Winter P (2010). Principal components analysis, *Physiology News* 79, 17–19.
- Tabachnick BG & Fidell LS (2007). *Using Multivariate Statistics*. Pearson Education, Boston.