CHAPTER 6: MULTIVARIATE DATA ANALYSIS.

6. 0. INTRODUCTION

6. 0. 1. Chapter 5 calculated bivariate correlates for variables believed to influence export performance, identified those for carrying forward, and modified the research hypotheses. The shortcomings of bivariate analysis and especially its likely accumulation of Type 1 error (erroneously rejecting the null hypothesis) were noted, as was the consequent need for multivariate analysis of the remaining data. (See Paras 5. 9.1.1 et seq). Accordingly, this chapter describes the use of these methods, and the outcomes. Main topics covered, after preliminary issues, are chief assumptions, selection of techniques, data reduction (IVs), analysis of data, and main results.

6. 1. PRELIMINARY ISSUES

6. 1. 1 There are three factors in particular which it is wise to keep in mind as we move towards multivariate analysis. These are: the generally low levels of correlation coefficients that characterise social science research (para 5. 0. 7); the inadvisability of imputing causality even to strong correlations (para 5. 0. 9); and, that while `export performance' is now widely recognised to be a multi-dimensional concept, there is not yet agreement on how it can best be measured (para 5. 0. 5).

6. 1. 2. As a prelude to the selection of specific multivariate techniques for data analysis, we need to consider the appropriate broad multivariate approaches to be adopted. Two basic criteria for choosing the most suitable of these are whether the research merits:

i. Dependence or inter-dependence analysis, and whether the data are suitable for:

ii. Parametric or Non-parametric techniques. *(Hair, J.F., et al, 1992, Figure 1.1, p12).* Where feasible, parametric techniques are generally to be preferred because of their greater statistical power *(Diamantopoulos, A., et al, 2000, p142)*

6. 1. 3. 1. Turning first to (i), an interdependence technique (as used e.g.in factor analysis), is one where the variables concerned are not categorised as dependent or independent. Variables appropriate for the use of dependence techniques, such as Multiple Regression, are those that can readily be classified as dependent or independent, in line with theory *(Hair, op. cit. p11, 13).*

6. 1. 3. Our research question and data together appear to facilitate a clear choice on both aspects. First, given the form of our export performance model (Chapter 1), our hypotheses-testing approach and the results of the foregoing bivariate analysis, it seems clear that a dependence technique is appropriate. Second, while much of our survey data are categorical in nature, it is of course dichotomously coded as dummy variables. The remaining variables are either numerical or interval in form. Overall, the data are compatible with parametric methods. Thus they qualify for analysis by dependence techniques that are parametric. But before proceeding to the selection of specific techniques, it is convenient at this point to consider the major assumptions underlying the main multivariate analytical techniques and the extent to which our data do now or can be made to satisfy them.

6. 2. MAIN ASSUMPTIONS OF MULTIVARIATE ANALYSIS

6. 2. 1. If the multivariate data analysis is to produce accurate and reliable output it must of course satisfy the related conditions. Common to the many of the main

multivariate techniques for consideration here are six such assumptions: independence of observations; linearity of relationship; limited intercorrelation; normality of variable distribution; homoscedasticity; and adequate sample size. The extent to which the variables for use in the analysis satisfy these requirements is discussed below. Where relevant and feasible, further adjustments are made to data to make them more consonant with the assumptions. Violations of the same can have serious consequences for the accuracy and validity of analyses. (See also para 6.2.10).

6. 2. 2. (a) Independence of observations**:** Observations are independent if the outcome of one observation is not dependent on another observation. Non-independent observations include those from before/after, panel studies and paired data comparisons (NCSU, PA 765, p19) and for Kenny they fall into one or more of three categories: groups, sequence (eg time series) and space (eg proximity of houses) *(See Kenny, D.A. et al, 1986, pp422-431*). (Breaching) the assumption has a substantial effect on significance level and the power of a test *(Sharma, S., 1996, p387).* Independence can best be assured by proper study design *(Osborne, J.W. et al, 2002, p1*) and, in particular, the selection of random samples of data for analysis. Perhaps the best known formal test for independence of observations is the Durbin –Watson where it is confirmed if the output statistic lies between 1.5 and 2.5 *(PA 765, p19).* Our survey sample was a random one, individually administered, and the independence of its observations also appeared to be confirmed by its Durbin-Watson test score *(1.61*). There are not, however, any sophisticated tests by means of which independence may be verified *(Sharma, S., op.cit, P388*)

6. 2. 3. (b) Linearity of relationship**:** While the relationships between variables may take a variety of forms, including the curvilinear, linearity is essential to the multivariate

analysis techniques for use here. Depending upon the technique used, lack of linearity may lead to under or over-estimation of relationships and the risk of Type 1 or Type 11 errors. Theory, the findings of previous research, and the examination of the standardised residual plots are main ways of detecting non-linearity *(Osborne, op.cit, p2*); as are scatterplots**.**

6. 2. 4. The outcome of our Chapter 5 bivariate correlations and the associated scatterplots gave no reason for doubting the linearity of correlations at that level. Additionally, most of our independent variables are dichotomous and, as we are reminded, this category can have only linear relationships with other variables *(Tabachnick, B. G. et al, 2001, p6)* ; the residuals plots are, however, examined below as a more searching test of the other metrics' linearity. It will be seen from Figure 6.1 (also relevant to some of the other assumptions) that linearity appears to be confirmed.

6. 2. 5. (c) Multicollinearity**:** While some correlation between independent variables is desirable and indeed necessary, the level of such intercorrelations can also be too high. Perfect multicollinearity (equal to 1), the limiting case, leads to infinite standard errors and indeterminant coefficients. But even high intercorrelations, the more common case, create problems in that the effects of the various IVs cannot be separated. In multiple regression a high level will lead to large variances and covariances, large confidence intervals, insignificant significance coefficients, low power and a high R-square *(North Carolina State University /NCSU, PA 765, p17).* It is therefore important to establish the acceptable upper limit. In this context, Pallant states that `multicollinearity exists when the independent variables are highly correlated, at *r =.9*

and above' *(Pallant, J , 2001, p137)*, but others consider that *r* =.8 or above signal a possible problem *(NCSU*, *PA 765, p17*). Even levels that fall short of these can be problematical. Thus Tabachnick et al advise against including in the same analysis variables with a correlation coefficient of 0.7 or higher. *(Tabachnick, B. G., et al, op cit. p84).* Among our selected independent variables the single highest correlation is 0.605. (See Table 6.4***)***. Multicollinearity does not therefore appear to be a problem for this analysis.

6. 2. 6 <u>(d) Normality of distribution</u>**:** In general the valid use of parametric statistical techniques is seen to require variables whose values are normally or near-normally distributed – that is those whose spread of values approximate to the normal or Gaussian curve. Dichotomous (0/1) variables, which by definition are non-normal (NCSU, .Edu, p5), are an exception to this requirement. As most of our variables are dichotomous the normality requirement is relevant only to the small group that are measured at the interval or ratio level, four of which are for use as dependent variables in the forthcoming multivariate analyses. Tabachnick et al point out that, other than their relationship with the DV, no distributional assumptions are made about IVs. But they also add that ` *…a prediction equation often is enhanced if the IVs are normally distributed…', (Tabachnick, B.G. et al, 2001, p119)*- implying benefits from transforming non-normal IVs.[1]

6. 2. 7 <u>Outliers</u>**:** measured by Mahalanobis[2] distances . With our (final 10) IVs taken as the degrees of freedom *(Pallant, J. op.cit, p144),* the critical chi square values are

[1] Normality of distribution is, however, an important assumption in Canonical Correlation Analysis.
[2] ` Mahalanobis Distance is defined as the statistical distance between two points that takes into account the covariance or the correlation among the variables' (*Sharma, S., 1996, P44).*Tabachnick et al define it as`*…the distance of a case from the centroid of the remaining cases where the centroid is the point created at the intersection of the means of all the variables' (op.cit., 2001,P68).*

8.31 (alpha =.05) and 23.21 (alpha =.01). (Distribution tables: statsoft; www.statsoft.com/textbook /sstable.html; and Richland *(Richland; Table; 2007*). In rank order , our 5 highest cases had distances of 22,10, 6, 3, and 3, and therefore do not exceed the critical value of 23.21. Thus we may conclude that outliers should not cause problems for our analysis.*(Pallant, J., op.cit., p145).* More generally, the quite close clustering of the regression standardised residuals to the diagonal straight line and in the scatterplot, the roughly rectangular distribution of standardised residuals suggest that there has been little if any violation of the major assumptions (See Figure 6.1)..

6. 2. 8. Experts' views differ on how robust the main parametric statistical techniques are to violations of the normality (or indeed of other key) assumptions and their opinions also vary on whether any departure from normality is permissible. Thus Tabachnick et al advise the use of data transformations to correct any abnormality found – unless there is some reason not to **(***op cit.,2001, p81* **).** Others take a more permissive view. Using the statistics of skewness and kurtosis as useful indicators of non-normality, they suggest that when we divide each of these by its standard error, the variable is acceptably normal if the result does not exceed +/- 2.- and comment that some researchers extend this range to +/-3 *(NCSU, PA 765, op cit, p4; SPSS v12 Help screen).*

Table 6. 1[3]	.Dependent Variable+	Skewness*	Kurtosis*
(a)	Trend in Turnover	0.35	-1.51
(b)	Share of Turnover Exported (%)~	1.49	-2.33
(c)	Trend in share exported	0.11	-1.98
(d)	Overall Profitability of Exports	-0.18	- 2.14

Notes: * These figures are SPSS skewness/ kurtosis statistics divided by their SEs

~ This variable had earlier had a number of outliers neutralised.

+ Figure 6.1. includes the histogram for the composite variable (a+b+c+d).

6. 2. 9. We would be reluctant to use transformations to normalise dependent variables because of the difficulties for interpretation which this process may create

[3] Re suitability for combining, *a* and *c* are measured on 11-point low/high scales, *d* on a 1-5 scale, with *b* of course a percentage. In all cases higher scores indicate higher performance and vice versa

(Tabachnick, B.G., et al, op cit. p80)- a form of interpretation on which there does not in any case appear to be much if any authoritative guidance available. Skewness is often regarded as a greater threat to validity than kurtosis. But, it will be seen that all of our values lie within the widest +/-3 limit, all but 2 within the narrower +/-2 limit and that skewness is generally low. Thus, while these 4 variables could be used without change, we have we believe increased their usefulness by combining them in a composite (standardised) variable, for use in multiple regression.

6. 2. 10. <u>(e) Homoscedasticity</u>**:** Refers to a state where the variance on a dependent variable is the same (homogeneous) across all levels of the independent variables (*SAS. 1998, P3.***).** Pallant says that `the variability in scores for variable X should be similar at all values of variable Y.- producing a fairly even cigar shape (Pallant, J, op.cit, P141). And Hair explains that it exists when the error terms (e) seem to be constant over a range of predictor variables (Hair, J.F.op.cit., P21). Homoscedasticity is a key assumption underlying the proper application of linear regression**.**

6. 2. 10.1. When the variance of errors differ at different values of the independent variables, there is hetereoscedasticity. Where it is only slight, it does not have much effect on significance tests, but where it is pronounced, it can seriously distort findings and increase the possibility of a Type 1 error. But the assessment of the residuals shows that hetereoscedasticity does not appear to be a problem here (Figure 6. 1).

6. 2. 11. <u>(f) Sample size</u>**:** Turning now to the specific sample size requirements of multiple regression, we have a total of 69 cases and we wish to regress 19 independent variables on the dependent variable. A general danger with any relatively large number

of independent variables, numerous parameter calculations and a small sample is that the degrees of freedom may be exhausted – producing larger standard errors and wider confidence intervals. But the main risk in the present case may well be that of `overfitting' the data, which occurs when the ratio of cases (observations) to predictors is too low. As one expert explains, the outcome can be serious. `The issue at stake here is generalisability.' *(Pallant, J.,2001, p136).*

6. 2. 12. So, our next requirement is to ascertain what this ratio needs to be and whether our sample can meet it. There is, however, no consensus on this issue. For example, Stevens holds that about 15 cases are required for each predictor if the regression equation is to be reliable (*quoted in Pallant,J. 2001, p136).* Tabachnick's suggested formula is >50 cases + 8m where `m' is the number of predictor variables *(Tabachnick B.G., et al, 1996, p132).* To satisfy these recommendations would require sample sizes of respectively 270 and 194. But Hair points out that the several rules of thumb which have been proposed range from 10 to 15 observations to an absolute minimum of 4 observations per predictor- and he also stresses the risk of` overfitting' the data *(Hair, J. F., et al, 1992, p46).* With our 19 predictor variables, we fall short of this absolute minimum (76 cases required) for e. g. multiple regression. To ensure the validity of our analyses, we shall have to improve the predictor /cases ratio and we return below to the selection of suitable data reduction methods.

6. 3. SELECTION OF MULTIVARIATE TECHNIQUES FOR DATA ANALYSIS

6. 3. 1**.** On the broad basis of compatibility together with the advantages of the increasing sophistication in multivariate analytic techniques, it was decided to analyse the data successively using the Multiple Regression, Canonical Correlation and Partial

Least Squares techniques. Fuller details of each of these methods and the main reasons underlying their selection follow.

6. 3. 2. <u>(i) Multiple Regression</u>**:** In contrast with Pearsonian bivariate, one-on-one correlation, Multiple Regression is `many- to-one., having a single dependent variable (which must be continuous) and two or more independent variables ( which can be continuous or dichotomous) that are regressed upon it. It produces an equation of the form: $Y' = A + B_1 X_1 + B_2 X_2 + \ldots + B_k X_k$; and fits a `line of best fit'[4] to the data.

[<u>Note</u>: Where Y'= the predicted value on the DV; A= the Y intercept (value of Y when all Xs are zero); the Xs represent the various IVs (of which there are k); and the Bs are the Coefficients assigned to each of the IVs during Regression (*Tabachnick, B.G. et al, 2001, p111).].*

Pallant summarises the *modus operandi* of MR as having 3 elements: it indicates how much of the variance in the DV is explained by the IVs; it gives an indication of the relative contribution made by each IV; and it indicates the statistical significance of the results for the model overall, as well as for the individual IVs. *(Pallant, J., 2001, p139)***.** Noting that the technique can be used both to predict and to explain, Osborne observes that in the second of these roles it explores the relationship between a number of variables to illuminate some phenomenon with the aim of generalising the results to a population *(Osborne, J. W., p1: D/L June 2007).* Columbia university sees its general purpose as being that of ascertaining more about the relationships between a dependent variable and several predictor variables. *(QMSS e-Lessons, Columbia .edu: undated* **).**

[4] Using bivariate simple linear regression as an example, Tabachnich, et al explain that the best fitting straight line goes through the means of X and Y , minimising the sum of the squared deviations *(Tabachnick, B G et al, 1992***,** *p54.)*

6. 3. 3. Its main merit in the context of this research is probably that it should produce a more accurate picture of the relationship by offsetting, in part at least, the effects of inter-correlation between the independent variables and , by its simultaneity, the build up of Type 1 error that may have resulted from the many separate correlations in Chapter 5.

---

6. 3. 4.. (ii) Canonical Correlation Analysis (CCA): Whereas Multiple Regression investigates the relationship between two or more IVs and a single DV, CCA analyses that between (usually) two sets of variables – one of which is commonly designated the independent and the other the dependent set. As Hair et al explain, it can be expressed in the general form as *:* $Y_1 + Y_2 + Y_3 + \ldots + Y_n = X_1 + X_2 + X_3 + \ldots X_n$ *(Hair, J. et al, 1992, p193).* Given an appropriate theory, the X variables on the right of the equation might be cast in the role of predictors of those on the left, but such designations are not necessary *(Sharma, S, op. cit, p391).* CCA's main advantage is that it can take into account the complexity of real world data including the possibility that the sets of variables are significantly related along more than one dimension.

6. 3. 5. CCA does this by first creating linear combinations of the variables in each set to produce what are called variates or roots, so as to maximise the correlation between the two variates. This is done by computing a weighted average for each set that will maximise its correlation with its fellow, producing the first canonical correlation coefficient. With this technique we can also create another set of weightings that are unrelated to the first and maximise the intervariate correlation to produce the second canonical correlation coefficient. And the canonical analysis continues till the number

of Canonical Correlation coefficients equals the number of variables in the smaller set. (*Clark, M.,Unt.Edu. p2: D/L 21 June `07)*

6. 3. 6. Negatively, it might be argued that the technique is redundant here because the composite standardised DV we are using in Multiple Regression includes all four of the DV variables. This is not, however, the case because all four are locked together; in CCA they are free to vary. This freedom to vary is the positive aspect of the use of CCA for the survey sample data analysis. It will be recalled that export performance was operationalised by using these 4 variables with the aim of overcoming the shortcomings associated by scholars with a single item measurement (Chapter 1), One often mentioned disadvantage of CCA is that its output can be difficult to interpret.

6. 3. 7. (iii) Partial Least Squares Regression**.** ` Partial Least Squares is a linear regression method that forms components (factors, or latent variables) as new independent variables (explanatory variables, or predictors) in a regression model. The components in partial least squares are determined by both the response variable(s) and the predictor variables. A regression model from partial least squares can be expected to have a smaller number of components without an appreciably smaller R-square value'. (http://www.statsoft.com/textbook/glosp.html : d/loaded 21 Nov07 *).* It is probably the least restrictive, in terms of the assumptions it makes, of all the main multivariate techniques available. Among its several major advantages are two that are especially relevant here: its components are orthogonal and it functions happily with small sample sizes. (On the negative side, it is reported to be better at prediction (not our main interest) than explanation and its output can be difficult to interpret). It is also

employed in our analysis to buttress CCA in the context of the `normality' assumption problem.

6. 3. 8. The three multivariate techniques for use in our survey data analysis are fully described and assessed in a wide range of publications including: Multiple Regression and Canonical Correlation Analysis); Tabachnick, B. G., et al, Chapters 5 and 6): Hair, J. et al, Chapters 2, 5; and Pallant, J, 2001, Ch 13 (Regression only). PLS is described in detail in the Umetrics (UK) Ltd textbook *(Eriksson, L. et al, 2006) a*nd doubtless in many other publications also. (The PLS data analysis, not catered for in SPSS, was conducted for the writer by *Umetrics UK Ltd*).

6. 4. DATA REDUCTION: INDEPENDENT VARIABLES.

6. 4. 1. As our sample size cannot be increased, the only solution available appears to be to reduce the number of predictor variables. Given that dropping any of the 19 could entail the loss of key data, this is not seen as an option. We must seek to reduce the number of variables by techniques entailing minimal loss of data. Two methods potentially well-suited to this task are Principal Component Analysis, and the formation of composite variables- the latter one of Tabachnick's suggestions *(op cit. p117).*

(a) Principal Component Analysis

6. 4. 2. Principal Component Analysis, a subdivision of Factor Analysis, is a useful means of transforming a large number of related variables into a smaller set of linear combinations – a more manageable number of dimensions or components, with all of the variance in the variables being used *(Pallant, J. 2001, p110, 151,152).*

6.. 4. 3 Data suitability for factor analysis turns largely on sample size and the strength of the inter-variable relationships. While there is no agreement on the minimum sample size, some leading authorities claim it is not the overall size of the sample that is important but the ratio of cases to variables, suggesting that 5 cases per item is adequate in most circumstances. On the second, Tabachnick et al advise that if few correlations of .3 or above are found in the matrix, then Factor Analysis may not be appropriate *(quoted in Pallant, op.cit., p153).*

TABLE 6. 2 INDEPENDENT VARIABLES: PCA CORRELATION MATRIX

	Component					
Variable	1	2	3	4	5	6
Man-Days Abroad pa	.709					
Forecasts Export Sales	.701					-.360
Formal Export Market Plans	.652					
Number of Export Markets	.622	-.515				
Number of Geo Areas Served	.617	-.576				
Number of Sales Channels Used	.580			-.363		.383
Select Markets Both Methods	.570		.302			
Main Info Sources: TFs & Exhibs	.549					
- In/Out Trade Missions	.546		-.453			
- Visit Foreign Retailers	.514	.312	-.324		-.322	
Committed/ Passive Exporter	.526	.375				
Commission Agent	.525		-.430		.380	.307
Special Staff Skills	.516		.393	-.422		
Number Quotation Terms Used	.492			.363		
Analyse Profitability	.513	.540				
Cost Plus & Market Based Pricing	.374			.610		
Total FTF Meetings Annually	.455			-.522	.372	
Should UK Join the Euro		.422	.461		.560	
Continuity Of Exports	.425	.420				.470

[Extraction Method: Principal Component Analysis: 6 components extracted]

6. 4. 4. The Kaiser-Meyer-Olkin *(Sharma, S., 1996, p116)* and Bartlett's Test of Sphericity *(Sharma, S., op.cit., p76, 123)* (SPSS), employed to test the factorability of our 19 variables, produced a score of .748 (against 0.6 or above required) for the former and a significance level of .000 (0.05 or less) for the latter, comfortably confirming suitability – and many of the correlations had scores above the recommended 0.3 level.

6. 4. 5. With the Eigenvalue set at a minimum of 1, PCA produced 6 components which together accounted for 66.39 per cent of the total variance. However, this initial solution did not reveal any simple structure among the components. While almost all factors loaded heavily on Component 1, loadings as a whole were, as can be seen, quite significantly split and there were some quite high values loaded on the later components.

6. 4. 6. The scree test gave a cut-off at 2 components and Varimax rotation was used, followed by orthogonal (Direct Oblimin *(See Pallant, J., 2001, P155 et seq*). The final 2-component output showed an almost simple structure. But there are several significant problems with this solution: first, the groups of variables loading on each of the two components do not seem clearly differentiated from one another by the types of variables they contain- and mainly for this reason, it has not been possible to find general descriptive titles for either of them; second, although together they account for 38.66 per cent of the total variance., even if they had generated a simple structure, relating our 8 hypotheses to them might not have been readily achievable**.** Finally, there are some doubts about the appropriateness of applying PCA to largely categorical predictor variables. Overall, it does not appear that PCA has succeeded in bringing about the desired level of data reduction with minimum information loss. Therefore we turn now to our second option: combining our predictor variables.

(b) Composite variables

6. 4. 7. It will be recalled that four main general areas together embraced all of the independent variables whose correlations rose above the r = .25 cut-off level, viz:

- Managements’ Export Commitment, Positive Attitudes And Perceptions;

- Export Planning And Research (to include special staff skills);
- Export Markets: Selection Methods And Market Numbers And Regions Served;
- Marketing Mix (Commission Agents and Number Of Sales Terms). [Table 5.12]

Our aim is therefore primarily to reduce the numbers of predictor variables- variously dichotomous, interval and ratio- that together currently represent each of these broader topic groups- by forming composite variables. This cannot, however, be achieved entirely uniformly. Quantitative and qualitative variables cannot readily be combined. Therefore we are restricted to combining, *ceteris paribus,* numerical with numerical and dichotomous with dichotomous- and these last of course predominate.

6. 4. 8. Independent variables to be combined should ideally have roughly equal effects on the dependent variable(s). They ought also to be correlated but not excessively so. On the first of these, the effects are approximately similar in our case but their combined impact tends to be somewhat lower than in the individual cases. This may perhaps indicate that because they are intercorrelated, their bivariate correlations overstated somewhat their real impact on the dependent variable.

6. 4. 9. The approach adopted here to data reduction by combining variables is three-sided. Where existing variables do not have the necessary affinity with any of the others, they remain unchanged. Numerical variables are combined by addition when they are compatible. And this compatibility and affinity is clearest where the variables concerned are similarly scaled and measure different aspects of an underlying construct (eg point of sale advertising and meetings with retailers -both being aspects of promotion). Finally, in the case of dichotomous predictor variables, they are combined, in line with Amine's methodology *(Amine, L.S. 1976, pp301-303- PhD thesis),* by

scoring a new variable 0/1; the score is 1 when and only when all of the `combinees' possess the positive quality in question and 0 if any of them do not. (As this procedure entails dropping some *1s* in the original variables, it would not be surprising to find a rather lower value for the composite variable's correlation with the dependent variable. This coding process has reduced our group of 19 independent variables to a new total of 10, giving a safer case-to -independent ratio of 6.9 which is comfortably above Hair's absolute minimum of 4 cases per independent variable. The original and combined variables are set out in Table 6.3. Correlation coefficients with the DV (Sharto- share of turnover exported) are shown on the right (original IVs) and on the left (combined & single IVs).

6. 5. ANALYSIS OF DATA AND MAIN RESULTS

6. 5. 1. Company size**:** It will be recalled that the bivariate correlations in Chapter 5 did not find any statistically significant relationships between the DV (export ratio) and either turnover or number of employees (r = -.106 and -.201 respectively: Table 5.2*);* and the subsequent partial correlation showed that these weak results owed nothing to the high level of correlation between the two IVs . Prior to undertaking the main analysis it seemed worthwhile to test again, this time seeking a multivariate relationship. The resulting regression output was not statistically significant (*p =.176)* and the model accounted for only 1.6% of the variance in the DV.

Table 6.3.[5] Reducing The Number Of Independent Regressor Variables

New Variable Title	(*r* =)	Variables Original Subtitles	(*r* =)
Managements' Export Commitment,			
Positive Attitudes & Perceptions			
		Continuity of exporting	(0.382**)
X1 MGTCOM	(0.369**)	Committed or passive	(0.288*)
		Should the UK join the Euro	(0.320**)

[5] Where composite IVs shown in this table are later prefixed by LOG or LUG this indicates that they have been log-transformed. LOG and LUG thus mean the same thing.

Export Planning, Research and	
Special Staff Skills	
	Formal Export Market Plans (0.278*)
X2 COMPLAN (0.305*)	Forecasts Export Sales (0.278*)
	Analyses Export Profitability (0.453**)
X3 Special Staff Skills (0.274*)	Special Staff Skills (0.274*)
Export Markets: Selection, Numbers,	
Dispersion.	
X4 Formal & Inf'l Sel. (0. 372**)	Formal & Informal Selection (0.372**)
X5+ EXSPRED (0. 347**)	Number of Markets (0.344**)
	Number of Regions (0.320**)
The Marketing Mix	
X6 Cost+ & M-Based Pricing (0.275*)	Cost + & M-Based Pricing (0.275*)
X7 NOQTSC (0.401**)	No of Sales Terms (0.261*)
	No of Sales Channels (0.367**)
X8 Commission Agent (0.313**)	Commission Agent (0.313**)
X9 VISITS (0. 302*)	Trade Missions (0. 292*)
	Trade Fairs & Exhibitions (0.397**)
	Foreign Retailers (0.281*)
X10+ MDAANDTFT (0.480**)	Annual Man-days Abroad (0.473**)
	No of FTF Meetings p.a. (0.304*)

Note: + Denotes metric/ numerical variable which sums the original variables. All of the other new variables are dummies. They have been scored 1 only where **all** of their constituent original variables were also scored 1 (eg each 1 score in X1 represents 1 scores in each of the 3 originals); otherwise 0.
* Significant at the 0.05 level.; ** at the 0.01 level. All correlations are with the export ratio (DV).

It thus seems clear that we are justified in dropping company size as a possible predictor of knitted apparel export performance in the rest of this analysis.

6. 5. 2. The three selected multivariate data analysis techniques are deployed sequentially below. Multiple Regression analysis is applied to the data at two levels. The first two analyses deal with main effects only.

(i) Multiple Regression**:** We begin by using multiple regression to test individual hypotheses. In line with Green, S.B. et al (*2000, p264*), Multiple Regression may be used, inter alia, with unordered sets of IVs such as our 8 Hypotheses and the IVs, above the cut-off level, associated with them (Table 5. 12). As will be seen from Table 6.4, all hypotheses submodels apart from H6 (Pricing methods) are significant below our alpha = 0.05. Not having specified it beforehand, we are not of course entitled to invoke the 0.10 level of significance, but we may still take account of significance results that fall at or below that level here –because of probable limitations associated with these unordered sets, that are subsets of the full model.

6. 5. 3. The more detailed results may be summarised as follows; A disposition explicitly stated by many of the managing directors, Commitment to exporting (0.1 alpha) is the only significant IV in support of H1 (Management commitment). H2 (Export planning) is supported by two IVs: Analyses profitability (0.05) and formal export market planning (0.1). For H3 (Use of market research information) only visits to trade fairs & exhibitions (0.05) is significant. Formal & informal market selection methods (0.05) is the only IV associated with H4 (Market selection methods); and H5 (Markets & regions), represented solely by composite IV LOGEXSPRED, is significant at alpha = 0.05. As noted above, neither of the IVs representing H6 (Pricing methods) is significant at either alpha level. Number of sales channels is the only one of the IVs associated with H7 (Sales channels used) that is significant and then only at the 0.1 alpha level. But composite IV (man-days plus annual face-to-face meetings) unequivocally supports H8 (promotional intensity) at the alpha 0.05 level. Overall, if each of these unordered subsets can properly be regarded as a complete model in itself, then, based on the *Beta* coefficients and including significant and non-significant IVs, it will be seen that the amount of variance in the DV accounted for by the IVs linked to each hypothesis ranges from 26 per cent (H6) to 74 per cent (H2).

6. 5. 4. These initial multiple regression results appear to support usefully all of our hypotheses except that relating to pricing methods (H6). However, we note that the unordered subsets used to test the hypotheses did not- and could not- simultaneously

take account of the various inter-correlations between all of the IVs (See Table 6.5) and may therefore be less accurate on that account. It is thus appropriate now to proceed to regress all 10 IVs of the full model on the composite dependent variable.

(ii) Multiple Regression: full model

6. 5. 5. The results of the run in standard regression, whose aim is to ascertain the impact of the 10 IVs on the DV, are evaluated sequentially below. The model as a whole is statistically significant (F =4.482: Sign= .000). And R^2 in the full model summary is seen to explain 44 per cent of the variance in our composite DV. But in small samples such as this, R^2 tends to over–estimate the value for the parent population and Adjusted R^2 provides a truer estimate of the population value. Hence we opt for the Adjusted R^2 value, as Pallant advises (*(Pallant, J. op. cit., pp 144/ 145,* which is here a quite reasonable 34 per cent.

6. 5. 6. Two other aspects are of particular interest. First, turning to the significance of these results, the ANOVA test of the null hypothesis (.0005) indicates that they are very unlikely to have occurred by chance. Second, it is interesting and relevant to ask which of the IVs contributed and by how much to the amount of variance in the DV that is explained by the model.

6. 5. 7. The contributions made by the various IVs can be established by examining the table of coefficients (Table 6. 5). Although views differ among scholars on the relative merits of using the B and/or *B* values for this purpose (See for instance Professor Swank, who is critical of *Beta [Swank, P.R., Allstat Jiscmail archive, August 2007]* ***),*** we show both here but opt for the *B* on the grounds that its standardised data

enable comparisons to be readily made between the different values; the largest being the most important irrespective of sign.

6. 5. 8. It will be seen that, in rank order, the five main contributors to variance explanation are: Man-days abroad and FTF meetings; Number and spread of export markets; Management Commitment ; use of a Commission Agent; and Market Visits. Only the first of these made a statistically unique contribution at the chosen alpha level of 0.05, but it is interesting that the second and third were statistically significant at the 0.1 level. All of the related *sr squared* values are shown in the table.

6. 5. 9. Now that the Multiple Regression process has been completed, we may note two things: first, the overall result should be better than those reported above because our sample was, at above 40 per cent, a large share of the total population of knitted apparel exporters; the Finite Population Correction factor is in fact 0.7482: and, second, the analysis of residuals, shown in the histogram, scatterplot and Normal P-P plots together confirm that the assumptions underlying multivariate analysis have been reasonably satisfied. (See Figure 6.1). Reverting to the first of these, there does not seem to be any reliable and authoritative scholarly literature on how to apply the FPC in this analysis. That was the advice received direct from Professor S. Simon[6] in 2006 nor has the writer been able to locate any. It appears then that we are confined to noting that the benefit should be appreciable and leaving it at that.

---

[6] Professor D..Simon, Childrens' Mercy Hospital. ssimon@cmh.edu. Personal communication in 2006

(iii) Canonical Correlation Analysis (CCA)

6. 5. 10. In the canonical correlation analysis (CCA) our aim was to establish more precisely the nature of the relationship between our two sets of variables (including finding out how many dimensions are required to do so). Because theory and our export model (Chapter 1: and paras 6.3, 6.4 above) specify a dependence relationship, the two sets comprise respectively 10 predictor and 4 criterion/ dependent variables and the specific objective was to ascertain how much of the variance in the dependent set is explained by the predictor set.

6. 5. 11. The CCA was run (as with MR) in SPSS v14, this time using the MANOVA macro[7]. From the earlier analyses we know that there is a relationship (which is linear) between the two sets, so we do not need to check specifically for this. Table 6.6. shows the results of the multivariate tests of significance, using the conventional four measures and although Hotellings gives the strongest significance value (.008), we shall mostly follow below the normal practice of centring on Wilks (.033).

Table 6.6: Results of Multivariate Tests of Significance: CCA Analysis

Test Name	Value	Approx. F	Hypoth. DF	Error DF	Sign of F
Pillais	.68626	1.35763	36.00	236.00	.095
Hotellings	1.15978	1.75578	36.00	218.00	.008
Wilks	.41737	1.54341	36.00	211.60	.033
Roys	.48267				

6. 5. 12. From the output it is useful next to focus on the number of dimensions required to account for the association between the two sets of variables and this is done below.

[7] Which does not, however, output either the raw canonical correlation scores or cross-loadings (i.e individual variable with opposite canonical variate) that are so useful for the interpretation of CCA results.

Table 6.7+ : Canonical Dimensions Test

Dimension	Canonical Corr.	Multi. F.	df1	df2	p-value
1	0.695	1.358	40.00	210.41	0.033
2	0.360	0.659	27.00	164.19	0.899
3.	0.242	0.544	16.00	114.00	0.918
4.	0.124	0.488	7.00	58.00	0.839

\+ This table has a very similar format and content to the sample write-up of analysis used by UCLA ATS in its SPSS Data analysis examples: Canonical correlation analysis

6. 5. 13. As Table 6.7. indicates, the canonical correlation between the first pair of canonical variates (Dimension 1) is, at 0.70, quite high. It is twice or more that of the second and later dimensions and this pair is also the only one of the four pairs of variates that is statistically significant (alpha =0.05). Therefore this first pair of canonical variates accounts for all of the correlation between the sets of predictor and criterion variables *(Sharma, S., 1996, p406)* and by convention only this pair needs to be interpreted

6. 5. 14. But, overall, the first run of the CCA produced unexpected and unclear results. It was found that, while the predictor and dependent variable correlations were of the broad magnitudes and rankings expected, most of them were seen to have negative signs when positive ones were looked for. Through a process of re-running the CCA, beginning with only the 3 continuous IVs in play, then gradually adding, one at a time, the others, in most plausible order, a form of suppressor variable (or similar), PrcCPMB, was identified with the final entry. It will be recalled that this variable made a very small contribution in Multiple Regression (.009), in which it did not achieve statistical significance. (Also, it seems to be the weakest result in PLS. See Figure 6. 2).

After its deletion from the CCA data, leaving 9 predictors, all variables in the analysis assumed their expected (predominantly positive) signs.

6. 5. 15. In the broad format recommended by Sherry, A. et al **(***2005, p48*), the outcome may be reported as follows. A CCA embracing 9 independent variables as predictors of the 4 turnover-related dependent variables was conducted to evaluate the multivariate shared relationship between the two variable sets (variates).

6. 5. 16. As expected, CCA identified 4 functions (roots or dimensions), whose squared canonical correlations (R^2) were, in sequential order, respectively: .483, .130, .059 and .015. Overall, the full model, across the four functions and using Wilks lamda (/\), was statistically significant: .417 criterion F(36, 211.60)= 1.5434, p= .033. Because lamda represents the variance **not** explained by the model, 1- /\ gives the full model effect size, in r^2 format. Therefore, taking our set of 4 canonical functions, the size of the r^2 effect was .517, showing that the full model explained a substantial proportion (52%) of the total variance shared between the variable sets.

6. 5. 17. While this full model was statistically significant, such was not the case with 3 of its constituent functions. The dimension reduction analysis showed that functions 2 to 4, 3 to 4 and 4 alone [F(24, 165.92)= .531, p=.965: F (14, 116.00)= .320, p=.990: F(6, 59) =.153, p.988] were all statistically non-significant. Thus, in line with normal practice, only the first function (52% of shared variance, p=.033) will be interpreted *(Sharma, S., op cit., p403)*

6. 5. 18. The standardised canonical function coefficients, structure coefficients and squared structure coefficients for Function 1 are presented in Table 6. 8. As there is only one function for interpretation, the communalities coefficient (h^2) (sum of r^2 across all functions) would appear to be either redundant or identical with r^2 .

Table 6. 8

Variable	Coeff.	Function 1. *r*	r^2(%)	h^2 (%)
TnvrStrSnd	.290	.595	.354	.354
Sharto	.705	. 905	.819	.819
StrSharto	.211	.732	.536	.536
ExPfyOall	.060	.575	.331	(as r^2)
LUGMDAFTF	.542	.868	.753	(as r^2)
LUGEXSPRED	.223	.650	. 423	..
VISITS	.173	.500	. 250	..
MGTCOM	. 309	. 497	.247	
Both14J05	. 189	. 607	.368	
OrgExp4	. 088	. 490	.240	
COMPLAN	-.176	. 471	. 222	
CommAgt	. 205	..461	.213	
LUGQTSC	-.042	.602	.362	

Note: Based on Sherry, A.et al, 2005, p44

6. 5. 19. From the Function 1 Coefficients, it will be seen that all four of the dependent (criterion) variables, all with positive signs and scores ranging downwards from Sharto to ExPfyOall, made important primary contributions to the dependent (criterion) variate**,** and that Sharto (export ratio) had a value more than double that of the largest of the others. Thus this variable made the most important contribution. The squared structure coefficients also fortify this conclusion- though the differentials are not as striking.

6. 5. 20. Function 1 predictors were predominantly positive-signed, but with two negatives. In rank order of the importance of their contributions to the predictor variate

were LUGMDAFTF, LUGEXSPRED, Both14J05, VISITS, MGTCOM, OrgExp4, and CommAgt. With LUGMDAFTF (75%) the top rating, and the others ranked downwards to 21%, the corresponding seven squared structural correlations also match the order of the foregoing. LUGQTSC and COMPLAN have negative signed correlation coefficients- and it is not entirely clear why this is so.

6. 5. 21. We do need to keep in mind the relative smallness of our sample, CCA's general requirement for normally distributed data whereas most of our predictor variables are dummies, and the possible need for bootstrapping (i.e repeated resampling from own data to establish confidence intervals for any test statistic) of this CCA output[8]. We have also the disadvantage that SPSS, in MANOVA macro mode, does not provide either raw canonical correlations, or the cross-loadings which could cast further light on the weight of the relationship between the predictor and dependent variables used in this CCA model (See above).

6. 5. 22. Given these actual and possible limitations, it does, however, seem clear that the results do confirm, in a reasonably strong form, our Chapter 5 amended hypotheses relating to the IVs that are clearly associated with effective export performance. It is considered that the use of CCA has fortified the more limited positive results from Multiple Regression- and that this is owing to its fully multivariate nature.

(iv) <u>Partial Least Squares:</u>

---

[8] Dr Bernard North, a statistician based at Imperial College, London, did not consider bootstrapping was necessary in this case. Personal communication (January 2008)

6. 5. 23. As noted above, Partial Least Squares is better at prediction than explanation (our main interest) and its output can be difficult to interpret. Less often employed in the social sciences than elsewhere, PLS was fielded here with the main aim of fortifying both the MRA and CCA analyses (respectively in respect of sample size and normally distributed data requirements).

6. 5. 23. 1. The PLS output has been subjected to jack-knife testing and all but one of the IVs are significant. It will be seen from Figure 6.2 (R2) that the model accounts for a useful 35 per cent or so of the variance in the DV, comparing favourably with the others. Moreover, the pattern of the Coefficients also resonates. The PLS analysis seem clearly to buttress and support all of the MRA and CCA findings.

6. 5. 24. Indeed there is a strong similarity between the ranking of our independent variables given by all three techniques. This can be seen in Table 6.9 where the main IVs are ranked by the three methods as follows.

Table 6. 9. Rankings of the main IVs in MRA, CCA and PLS

Variable	Rank No:	MRA	CCA	PLS
LOGMDAFTF		1	1	1
LOGEXSPRED		2	2	2
SELECTBOTH		3	3	3
VISITS		4	4	5
MGTCOM		5	5	6
S.STAFF SKILLS		6	6	6
C.AGENT		7	7	6

Note: PLS ranked LOGQTSC 4th

6. 6. MULTIVARIATE DATA ANALYSIS RESULTS:

Multiple Regression, Canonical Correlation & Partial Least Squares

6. 6. 1. Introduction. A two-part multiple regression analysis (main effects only) was performed then Canonical Correlation, using all remaining variables, and finally Partial Least Squares..

6. 6. 2. Multiple Regression: The standard multiple regression was performed between a composite dependent variable DV (trend in turnover, share of turnover exported; trend in share of turnover exported; overall profitability of exporting) and the 10 Independent variables: Management commitment; Company planning; VISITS: Formal plus Informal Market Selection methods; cost plus and market based pricing; numbers of quotation terms and sales channels; number of markets and regions; special staff skills; Commission Agents; Visits; number of man-days abroad and annual face-to-face meetings. The analysis was carried out with the aid of SPSS (v 14) Regression and SPSS Frequencies for the evaluation of the assumptions.

6. 6. 3. (3.) Transformations**:** Arising from this process, several variables were transformed to reduce skewness and number of outliers**;** and generally to enhance normality, linearity and homoscedasticity. A square-root transformation was used on special staff skills. Logarithimic transformations were employed on: number of markets and regions; numbers of quotation terms and sales channels; and numbers of man-days abroad and of annual face-to-face meetings. Four DVs: (trend in turnover, share of turnover exported; trend in share of turnover exported; and profitability of exports), derived from Likert scales running from low to high, were standardised and added to produce the composite DV. All of these transformations effected desired improvements in the variables concerned.

6. 6. 4. (4). Outliers: Taking the 10 IVs as the degrees of freedom, the critical chi square values were 8.31 (alpha=.05) and 23.21 (alpha=.01) and Mahalanobis distance was found not to exceed 16, thus no further outliers among the cases were identified.

There was no missing data, nor were any obvious suppressor variables identified. N (number of cases) totalled 69.

6. 6. 5. (5.) The inter-variable correlations, the intercept, the unstandardised regression coefficients (B), the standardised regression coefficients *(B)*, the semi-partial correlations (*sr2)*, and adjusted R-squared are shown in Table 6.5.

6. 6. 6. (6). R for regression was significantly greater than zero: F (10, .361.1= 4.214, p <.001. For the regression coefficient that differed significantly from zero the 95 per cent confidence limits calculated for the (square root of) man-days abroad and face-to-face meetings ranged from .366 to 3.584 – the sole range that did not pass through zero.

6. 6. 7. (7). Only one of the IVs (LUGMDAFTF) contributed significantly to the prediction/ explanation of export performance, in its standardised measure (at the chosen alpha level of 0.05 (-though 2 others: MGTCOM and CommAgt were <alpha =0.1). In total 44 % (Adjusted R^2 =34%) of the variability in export performance was predicted/ explained by knowing the scores on the one significant and nine insignificant IVs. The tentative results from the PLS analysis is also to be taken into account here.

6. 7. BROAD CONCLUSION FROM PEARSONIAN, MULTIPLE REGRESSION, CANONICAL CORRELATION, AND PARTIAL LEAST SQUARES ANALYSES

6. 7. 1. Given these actual and possible limitations, it does, however, seem clear that the CCA results do confirm, in a reasonably strong form, our Chapter 5 amended hypotheses relating to the IVs that are clearly associated with effective export

performance. It is considered that the use of CCA has fortified the more limited positive results from Multiple Regression- and that this is owing to its fully multivariate nature; and that PLS has also made a positive contribution to the findings.

Overall, the results from Pearsonian, Multiple Regression and Canonical Correlation are broadly mutually consistent and reinforcing. Bearing in mind the generally low level of correlations in the social sciences they may perhaps be regarded as positive and encouraging. Table 6.9 suggests the worth of the contribution made by Partial Least Squares[9]

6. 7. 2. This completes the analysis of the survey data . The next and final chapter conducts some final tests of the hypotheses and presents findings and conclusions from this research project as a whole.

---

[9] Following further gratis work by Umetrics UK Ltd the contribution made to the research findings by the PLS data analysis has increased significantly.