The use of statistical methods in management research: a critique and some suggestions based on a case study

30 March 2010

Michael Wood

University of Portsmouth Business School

SBS Department, Richmond Building

Portland Street, Portsmouth

PO1 3DE, UK

michael.wood@port.ac.uk .

http://userweb.port.ac.uk/~woodm/papers.htm

The use of statistical methods in management research: a critique and some suggestions based on a case study

Abstract

I discuss the statistical methods used in a paper in a respected management journal, in order to present a critique of how statistics is typically used in this type of research. Three themes emerge. The value of *any* statistical approach is limited by various factors, especially the restricted nature of the population sampled. The emphasis on null hypothesis testing may render conclusions almost meaningless: instead, I suggest deriving confidence intervals, or confidence levels for hypotheses – and suggest two approaches for doing this (one involving a bootstrap resampling method on a spreadsheet). Finally, the analysis should be made more user-friendly.

Keywords: Bootstrap resampling, Confidence, Management research, Null hypothesis significance test, Quantitative research, Statistics.

Introduction

The aim of this article is to consider the role which statistical methods can sensibly take in management research, and to look at some of the difficulties with typical uses of statistical methods and possible ways of reducing these difficulties. My approach is to focus on an article published in the *Academy of Management Journal* (Glebbeek and Bax, 2004), and to look at some of the problems with the analysis and at some alternative possibilities. My focus is management research, but many of the issues are likely to be relevant to other fields.

Glebbeek and Bax (2004) tested the hypothesis that there is an "inverted U-shape relationship" between two variables by deriving the linear and quadratic terms in a regression model, and their associated *p* values, and then checking whether these terms are positive or negative. This, however, ignores the fact that the pattern is a rather weak U-shape, and does not encourage scrutiny of the detailed relationship between the variables. My suggestion is to focus on this relationship by means of a graph (Figure 1 below) and parameters which, unlike the conventional standardized regression coefficients used by Glebbeek and Bax (2004), can be easily interpreted (Table 2 below). Furthermore, the evidence for the inverted U-shape hypothesis can be expressed as a confidence level (which comes to 65% as explained below) rather than in terms of the rather awkward, user-unfriendly, and inconclusive *p* values cited by Glebbeek and Bax. Finally, but perhaps most important of all, I discuss issues such as whether the target population is of sufficient intrinsic interest, and whether the variables analyzed explain enough, to make the research worthwhile.

The first two sections discuss the nature and value of statistical methods and some of their problems. Readers more interested in the analysis of the case study might prefer to go straight to the section on the case study.

The nature and value of statistical methods

According to the *New Fontana Dictionary of Modern Thought, s*tatistics, in the sense of statistical methods, is "the analysis of … data, usually with a probabilistic model as a background" (Sibson, 1999). This seems a good starting point, although the probabilistic model may be an implicit, possibly unrecognized, background. Statistical research methods typically work from a sample of data, and use this data to make inferences about whatever is of concern to the researchers. Other, non-statistical, approaches to research also make inferences from samples of data; the distinguishing feature of the statistical use of samples of data is that the results, the "statistics" derived (such as means, medians, proportions, *p* values, correlations or regression coefficients) depend on the *prevalence* of different types of individual in the sample – and these prevalences reflect probabilities.

To see what this might mean in a very simple situation, imagine that we have data on a sample of four individuals, and we then extend this sample by another two individuals from the same source. Suppose, further, that the two latest individuals are identical to two of the four in the original sample – in terms of the data we have, of course – let's call these four Type A. With the original sample we would estimate the probability of Type A as being 50% (two of the four), but with the extended sample the estimate of the probability would be 68% (four of the extended sample of six). From the statistical perspective the prevalence of Type A – measured by the proportion of the sample, which gives a natural estimate of the probability in the underlying population – is important. We might then compare this context with another context where Type A's are rarer – say 10% – and the comparison of the two contexts might give useful information about, for example, the causes of an individual being of Type A. This does not, of course enable us to predict with certainty about whether a particular individual will be of Type A: we can just talk about probabilities. (This obviously depends on suitable assumptions about the source of the sample and the context to which the probability applies.) The fact that the Type A individuals are identical from the point of view of our data does not mean they are identical from all points of view. All research, and statistical research in particular, has to take a simplified view of reality.

From a non-statistical point of view, finding the extra two examples of Type A would be of less interest because it would simply confirm what we already know. A second, and perhaps a third, identical case is helpful because it confirms that Type A is a possibility in several, doubtless slightly different, cases, but four might perhaps be considered a waste of time (although this would depend on the detailed context). This attitude to data has been dubbed "replication logic" (Yin, 2003): the point is not to count the cases, but just to confirm that similar things happen in different contexts, or to discover that they do not. "Qualitative" researchers who do take an interest in the proportion of Type A's are, in effect, using statistics.

Table 1 shows some data on postings on a web-based discussion board for six graduate-level university classes. The postings were categorised as "dominant" or "supportive" (agreeing with or encouraging another student).

Table 1: Frequencies of dominant and supportive postings on discussion boards

	Male	*Female*
Dominant	305	473
Supportive	50	308

(Data extracted from a conference presentation by Gregory, 2008.)

The obvious conclusion from this is that males are more likely to post dominant messages than females – 86% of postings from males are dominant compared with 61% of female postings. This suggests that there are things about males which makes them about 25% more likely than females to post dominant messages. However we do not know what these things are, and there are obviously many other variables in play here which we could almost certainly never fully understand. Despite this general tendency, there are some dominant postings from females, and some supportive ones from males. This table raises many further questions about the nature of the sample and the possibility of assessing the impact of further variables – age and culture are two explored by Gregory (2008).

Statistical conclusions like this involve concepts such as proportions of cases, or probabilities, or averages (if we had a numerical measure of dominance then an average would be a natural way of summarizing the overall difference between males and females), correlations, etc. The aim is to express a *tendency*, which may be partially obscured by noise, or nuisance, variables – the many other factors which may lead to differences between males and females.

Table 1 is a simplified view of the situation – all we are told is that there are 305 dominant male messages, 50 supportive male messages, 473 dominant female messages and 308 supportive female ones. The messages within each group are treated as identical or "exchangeable" (a term originally due to de Finetti – see Lindley and Novick, 1981);

any differences within the group are ignored with the focus of interest being on the frequency of occurrence of each group. Statistical methods analyze these frequencies to find out about the mysterious factors which lead to dominant male messages being more prevalent than dominant female ones. The metaphor of balls in an urn, often used in statistics texts, is a way of emphasizing the fact that cases in a statistical analysis are treated as being identical except for the measured variables (gender and dominant/supportive in this example).

This means, rather obviously, that statistical results are not deterministic; they are *partial* explanations because of the inevitable exclusion of noise factors. Some of these factors could be incorporated – e.g. Gregory (2008) produces breakdowns by age to incorporate this factor – but there are always likely to be noise factors not built into the model meaning that the estimated probabilities are rarely 0% or 100%. (The effect of these noise factors is often called error variation for obvious reasons.) This is a simple example of the value of incorporating context into research to enhance the precision of models, the value of which is widely acknowledged and supported by increasingly sophisticated statistical methods (e.g. Bamberger, 2008). However, even with extensive contextual factors included, it is rarely, if ever, possible to give deterministic predictions. This is probably inevitable given the difficulty of predicting the actions of individual human beings.

The statistical methods which are used in published management research are varied, but there are three widely used categories of methods which stand out from an informal survey (the reasons for not undertaking a more formal survey are discussed below):

- Descriptive statistics such as means, standard deviations, correlations and proportions are widely cited.
- Null hypothesis tests are very common, and are at the core of the logical structure of many research studies and the papers reporting them. Confidence intervals and Bayesian methods, which can be used as alternatives, are rare.
- Regression, in its many forms, is widely used.

Methods such as these are clearly of enormous potential value. Statistical analysis is of clear use for many tasks – e.g. predicting the quality of wine from the weather when the grapes are harvested, predicting house prices, and predicting which potential customers are most likely to buy something (Ayres, 2007). In these examples the influence of noise variables is substantial, but the statistics still enables us to peer through the fog and discern a pattern which is sufficiently reliable to be useful.

However, there are difficulties with the way statistics is used in some management research. Some common practices may have as little logic behind them as the "methodological myths" described by Vandenberg (2006).

Difficulties and critiques of statistical methods as used in management research

There is a large literature on the pros and cons of the different approaches to statistics (especially the Bayesian approach and how it compares with conventional alternatives), on the importance of particular methods and problems with their use (e.g. Becker, 2005; Cashen and Geiger, 2004; Vandenberg, 2002), on the importance and difficulties of educating users of statistics and readers of their conclusions, and, of course, on the derivation of new methods. However, there is surprisingly little by way of critique of statistical methods as a whole, or discussion of the characteristic features of statistical approaches to research.

One paper which purports to give such a critique of statistical modelling – in management science – is Mingers (2006). He claims that statistics, in practice, adopts "an impoverished and empiricist viewpoint", by which he means that it largely fails to go "beneath the surface to explain the mechanisms that give rise to empirically observable events". This is similar to the view of statistics as providing a partial explanation discussed in the previous section. If a satisfactory deterministic explanation is available, then a statistical model is not called for; in this sense, statistics is "the method of last resort" (Wood, 2006), but still a potentially useful approach for helping to make predictions when we do not fully understand what is happening. Mingers' critique also drew responses from Ormerod (2006) and Chiasson et al (2006) pointing out that Mingers' view of statistics is itself very restricted and impoverished.

A far more general attack on statistics aimed at the general public is contained in the recent book, *The black swan: the impact of the highly improbable* (Taleb, 2008). The statistical normal distribution is dismissed as “that great intellectual fraud”, and similar disparaging comments are made about many of the other standard ideas of statistics. Taleb’s point is that these ideas cannot cope with understanding the influence of occasional, extreme events, the “black swans”, which, Taleb claims with some credibility, may have a disproportionate influence on the course of history. However, this is hardly news to statisticians who have always taken a keen interest in outliers (Westfall and Hilbe, 2007). The message of the book is in effect that statistical methods cannot usefully model everything; they are useful in their place, but other approaches are necessary for trying to understand black swans. Few in management research would quarrel with this, although it is arguable that the useful scope of statistics is sometimes exaggerated – I will return to this below.

One commonly reported problem with statistics is that many people – including some researchers and some of their readers – find it difficult to understand. Even some articles published in respected journals use statistical methods in inappropriate ways (e.g. see the discussion below of Grinyer et al, 1994). This is particularly true of null hypothesis testing – which is a convoluted concept which involves trying to demonstrate “significance” by assuming the truth of a probably false null hypothesis (see Appendix 1). The obvious approach to dealing with problems of understanding is to call for more, and better, statistical education, and there is a very large literature, and several journals, on this topic.

An alternative approach to the education problem is to acknowledge that there are too many complicated techniques for researchers and readers to learn about (Simon, 1996, points out that people generally have about 10 years to become a specialist and this imposes a limit to the amount of expertise that can be mastered), so efforts should be made to present results that can be understood with the minimum level of technical understanding which is possible without sacrificing the rigour and usefulness of the analysis (Wood, 2002; Wood et al, 1998). This could be on the level of redefining output measures to make them more user-friendly, or using methods whose rationale is closer to common sense than conventional methods based on probability theory – this is one of the advantages of resampling methods such as bootstrapping (e.g. Diaconis and Efron, 1983;

Simon, 1992; Wood, Capon and Kaye, 1999; Wood, 2005a and 2005b). However, in practice, these opportunities are very rarely taken.

The user-friendliness issue is one, minor, strand in the debate about the pros and cons of different approaches to statistics. Another issue which deserves mention here is the debate about the role of null hypothesis significance tests because there are very strong arguments, put forward in numerous books and articles over the years, that these are often used in an inappropriate manner (e.g. Gardner and Altman, 1986; Kirk, 1996; Lindsay, 1995; Morrison and Henkel, 1970; Nickerson, 2000; Wood, 2003). Nickerson (2000) provides a thorough, recent review, but for our purposes here the three areas of possible difficulty highlighted by Wood (2003) are particularly relevant. (There is a more detailed discussion of significance tests and their alternatives in Appendix 1.)

First, the *p* values (significance levels) may not be correctly interpreted: the main problems being the incorrect assumption that a *p* value corresponds to the probability of the null hypothesis being true (which leads to the assumption that a high *p* value means that the null hypothesis is probably true), or the incorrect assumption that a *p* value is a measure of the strength of the effect, with a highly significant result being assumed to be a very important one. The second possible problem with null hypothesis tests arises if the null hypothesis is so obviously unlikely that evidence that it is false is of little value. For example, Grinyer et al (1994) tested the hypothesis that people are equally likely to express agreement, disagreement or neither with a given statement. This assumption is so arbitrary and unlikely as to render the test meaningless. And the third problem is that the focus on the hypothesis test may distract attention from the size of the effect.

One recommendation coming out of the debate is to formulate results as interval estimates instead of tests of hypotheses. Gardner and Altman (1986), writing in the *British Medical Journal,* recommended citing confidence intervals instead of *p* values, and a similar recommendation is among those advocated by Cashen and Geiger (2004) and Cortina and Folger (1998) for management research. Confidence intervals are less liable to serious misinterpretatation (although their correct interpretation is subtle as explained in Nickerson, 2000: 278-9), and avoid the focus on possibly irrelevant hypotheses in favour of the size of the effect. However, confidence intervals are very rare in management research (although they are used more widely in medical research): Nickerson (2000: 279-80) reviews suggestions which have been made about some of the possible reasons for this. I will look at the problems with *p* values in the case study below, and see how the results can be reformulated to show the findings in a more useful manner.

The potential misconception that a high, non-significant, *p* value confirms the null hypothesis can be clarified, in principle, by computing the statistical power. If the sample is small, the test has less *power* and the result may not be significant even though the null hypothesis is false. The difficulty is that doing this requires setting an essentially arbitrary minimum effect size for the null hypothesis to be false. Such power levels are very rarely computed for management research studies. Cashen and Geiger (2004) computed such power levels for a sample of studies and recommended researchers to provide this information as well as *p* values. However, they also recommended (following Cortina and Folger, 1998 and Nickerson, 2000) including confidence intervals, which "provide further detail that the hypothezised null is not trivial due to sampling error" (p. 162). In practice confidence intervals are likely to provide all the information needed, as I will demonstrate in the case study below, where I will also derive confidence levels for ranges of possibilities which are not intervals.

Finally, although not strictly within the scope of this article, it is important to note the obvious fact that there are many alternatives to statistical methods. The simplest is to use case studies to illustrate and explore what is possible without any attempt to estimate population statistics (Yin, 2003, Wood and Christy, 1999). As will be explained below, this is essentially the method I am adopting in this paper. And there are several more elaborate possibilities. Fuzzy logic and its derivatives are intended to model uncertainties due to the vagueness in the meaning of words rather than those due to noise variables. The idea of chaos refers to the situation where a deterministic model exists but uncertainties in the intial conditions mean that the outcome is effectively random. Qualitative comparative analysis (e.g. Rihoux and Ragin, 2009) is intended for cross-case analysis with moderate sample sizes: the idea being to assess the combinations of values of variables which lead to various outcomes.

A case study of the use of statistical methods in a research project

In order to explore some of these issues I will look at one research project in some detail: Glebbeek and Bax (2004). This is chosen because it was published in a respected journal (the *Academy of Management Journal*), is clearly written and the statistical approach used is fairly typical involving regression and hypothesis testing. My aim is not to produce a critique of this paper, but to explore issues of wider concern for the use of statistics in management research. (There are several interesting issues concerning this paper that I will not explore because they are not relevant for this purpose.) I am very grateful to Dr. Arie Glebbeek for making the data available; this has enabled me to carry out some of the suggested methods of analysis discussed below.

Glebbeek and Bax (2004) wanted to test the hypothesis "that employee turnover and firm performance have an inverted U-shaped relationship: overly high or low turnover is harmful." To do this, they analysed the performance (profitability) of "110 offices of a temporary employment agency" by building regression models, using "net result per office" (p. 281) as the measure of performance, and staff turnover, and the square of turnover as independent variables, as well as three control variables. Elementary calculus shows that for an inverted U shape the regression coefficient for the squared turnover term should be negative, whereas for the linear (unsquared) turnover term the coefficient should be positive. They did four variations of this analysis – for example, they tried relating performance to current turnover and to turnover in the previous two years. The results are presented, in the conventional way, by means of tables of standardized regression coefficients for the various models, supplemented by symbols to denote different ranges of *p* values. In all cases the coefficients were as expected by the inverted U shape hypothesis. However, none of the coefficients for the linear terms were significantly greater than zero, although three of the four coefficients for the squared terms were significant ($p < 5\%$ in two cases and 10% in the third). The discussion in the article argues that this provides reasonable support, but with a few provisos, for the inverted U shape hypothesis in the context of the employment agency in question.

There are three broad issues which deserve discussion in relation to the statistical methods used this article (and other similar articles). It is convenient to discuss them in increasing order of importance, as this will enable us to discuss each of the issues on the assumptions that the preceding issue(s) have been dealt with. So I start with issues of user-friendliness, then progress to a discussion of the null hypothesis testing approach which has been adopted, and finally consider the most important issue – how useful any statistical approach is in this context.

User-friendliness of the statistics

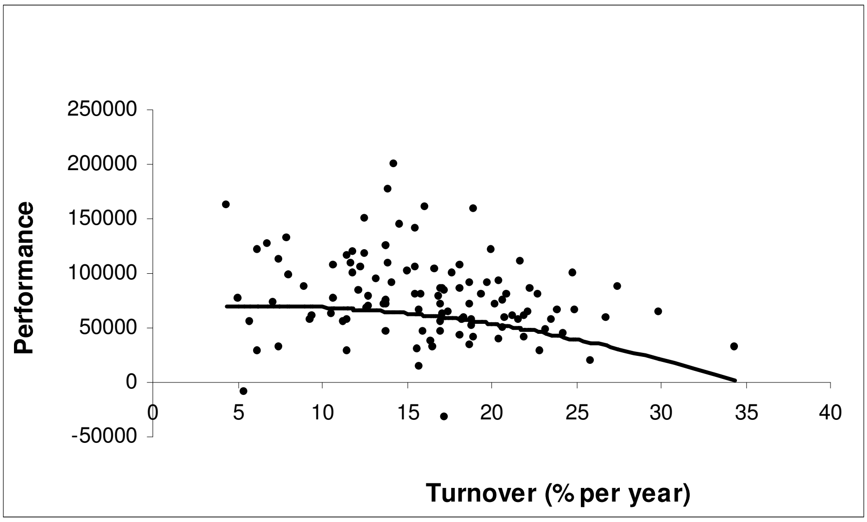
This covers both the user-friendliness of the way the statistical concepts are described, and the user friendliness of the concepts themselves. Readers may feel that readers of a technical research journal should be expected to understand technicalities without help. However, the extent of the expertise required, and its possible lack in many readers, means that it does seem reasonable to present results in as user-friendly manner as possible provided this does not lead to the article being substantially longer, or to a sacrifice in the rigour and value of the analysis.

Glebbeek and Bax give values for *Adjusted* R^2 and upper limits for p, and the regression coefficients are described in the header for the tables simply as "results of regression analysis" – all of which readers are expected to understand without further explanation. For example, the simplest linear model without control variables in Panel A of Table 2 in Glebbeek and Bax (2004) has the regression result as –0.23*, with the * indicating that $p < 0.05$, and at the bottom of the table we learn that *Adjusted* R^2 is 0.05.

This seems unnecessarily uninformative: the *Adjusted* R^2 could be described, perhaps in brackets, as "proportion of variation explained", and the p values as "probability of chance occurrence". Describing the regression coefficients themselves is more difficult because the tables in Glebbeek and Bax give *standardised* regression coefficients. If *unstandardised* coefficients were given (as they are in Shaw et al, 2005, a later article on the same theme in the same journal), these could be described as "predicted impact of an extra 1% staff turnover": in the example this is –1552. In other words, the regression predicts that each additional 1% per annum on staff turnover will lead to a reduction in performance of 1552 units. This interpretation would be even clearer if a graph were included.

The argument for a graph is even stronger in the case of the curvilinear model which is the focus of the article. Glebbeek and Bax (2004) analyse four such curvilinear models; the first (Model 4 of Panel A in Table 2) gives the linear turnover coefficient as +0.17 and the turnover squared coefficient as –0.45. This is in line with the hypothesis of an inverted U shape but tells the reader little about the details of the relationship between the two variables, both because the coefficients are standardised, and because the reader may not have the necessary background in quadratic functions to interpret even unstandardised parameters. This can be remedied by means of the graph in Figure 1 (which Glebbeek and Bax, 2004, mention but do not show). This illustrates the main conclusion from the article in a clear and user-friendly manner. (Graphs of curvilinear models are shown in two subsequent articles on the same theme – Shaw et al, 2005, and Siebert and Zubanov, 2009.)

Figure 1: Results and curvilinear predictions for Region 1 and mean absenteeism and age



As well as drawing a graph like Figure 1, it is also possible to make the parameters for the curvilinear model slightly more user-friendly. There is no obvious, natural measure of curvature, so the square coefficient can be taken as a measure of curvature and readers told that negative values correspond to an inverted U and positive values to a U the "right" way up. The other, linear, term gives information about the location of the turning point. If the linear term is positive and the square term negative, then the maximum of the graph will occur for a positive turnover figure. This is necessary for the inverted U hypothesis because if the maximum occurs for a negative turnover, a graph like Figure 1 will just show a declining performance with turnover. More precisely, elementary calculus shows that if the unstandardised regression coefficient for the square term is a and for the linear term is b, the turning point for performance will occur when the turnover is $-b/2a$. In the model in Figure 1, the two unstandardised regression coefficients are

a = -86.743836 and b = 1097.49998

and this formula predicts that the best level of performance will occur with a turnover of 6.3%. There seems little reason why this, more useful, parameter should not be given in the results table in preference to the linear regression term. These suggestions are incorporated in Table 2.

Table 2. Parameters for the model in Figure 1

	Best estimate
Location of optimum (annual % staff turnover)*	6.3
Curvature**	-87
Predicted impact of 1% increase in absenteeism	-3,330
Predicted impact of 1 year increase in average age	-831
Predicted difference between neighbouring regions (with Region 1 having the lowest performance)	15,465
Proportion of variation explained (Adjusted R^2)	13%
Predicted optimum performance for Region 1, and mean absenteeism (3.8%) and mean age (28)	69,575

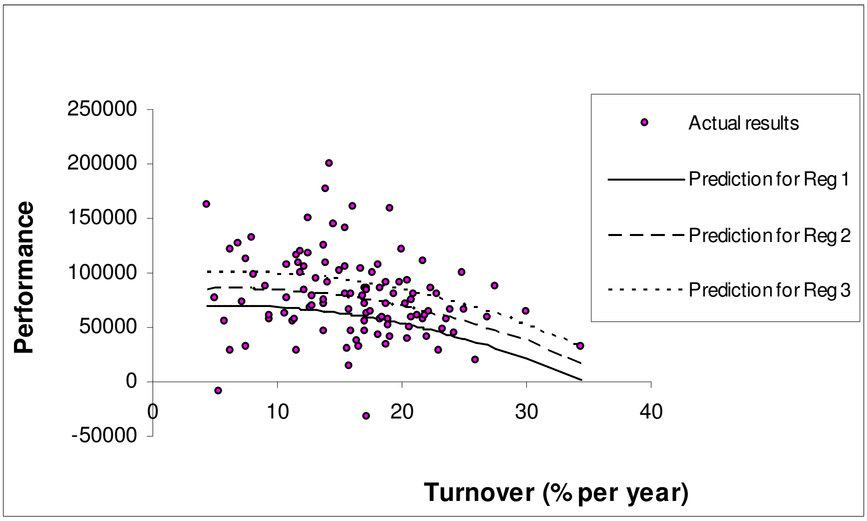
* Location is – b/2a, where a and b are the unstandardised regression coefficients for Turnover squared and Turnover respectively.

** Curvature is the unstandardised regression coefficients for Turnover squared. Negative values correspond to an inverted U shape. See text for more details of interpretation.

A little simple artithmetic will then allow the HR managers envisaged as users of the research by Glebbeek and Bax (2004: 285) to be more precise about the predicted effects of any particular level of turnover. The curvature, a, corresponds to the additional performance predicted from moving one unit away from the turning point. If the curvature is negative this additional performance is negative because any movement from the optimum brings a decrease in performance. Moving 2 units away from the optimum will change performance by 4*a*, and so on. If, for example, annual staff turnover were 15%, the departure from the optimum is 8.7% so this method predicts that performance would be 6,566 units below (= 8.7^2 x -86.743836) the optimum (69,575), which is 63,009. This should be clear by expanding the appropriate part of the Figure 1. However, it is important to remember that these predictions only incorporate an understanding of 13% of the variation in performance (which should be obvious from the scatter of the data in Figure 1).

Glebbeek and Bax use three control variables in the model in Figure 1. Figure 1 shows the predictions assuming Region 1 and mean levels of the other control variables. To show the effect of the control variables graphically we can draw separate lines for different levels. For example, Figure 2 shows predictions for the three different regions in the study. Again, this is information that can be deduced by the reader with appropriate expertise from Glebbeek and Bax (2004), but making it explicit in this way might be helpful.

Figure 2: Results and curvilinear predictions for three regions and mean absenteeism and age



Problems with null hypothesis testing, and suggested alternatives

The first suggestion here, stemming from the observation (above) that these tests are widely misinterpreted by readers, and even the occasional writer, of management research, is simply to describe *p* values in terms such as "the probability of chance occurrence", as suggested above. However, there are further difficulties (see above), which suggest a more radical solution.

The difficulties are that hypothesis tests often yield results which are trivial, and fail to provide the information that is really required. In the present example the aim is to test the hypothesis that "that employee turnover and firm performance have an inverted U-shaped relationship". If this means that the best performance is hypothesised to occur at some level of turnover greater than 0% and and then decline for greater levels of turnover, then this is *obviously* likely to be true. An organisation in which annual turnover is zero – nobody has ever left – is obviously likely to have problems, as is one where the turnover rate is 100% or more. The best level of turnover is likely to be somewhere between these two extremes. The hypothesis says nothing about the size of the differences – even the relatively slight inverted curvature in Figure 1 are sufficient to confirm the inverted U-shape hypothesis.

Strictly, the hypothesis that is tested is the *null hypothesis* that there is no relationship between turnover and performance exists, and the regression coefficients would all be zero. However, this is such an unlikely hypothesis that finding evidence against it adds little of value to our knowledge. It is arguable that the logic of the inference would be better if the null hypothesis were the currently accepted "best guess" – perhaps that performance is negatively related to turnover. However, because this is vague hypothesis – it does not specify how negative the relationship is – it cannot be used as a null hypothesis and as the basis for *p* values. Even if it could be it is doubtful that the hypothesis would be rejected because Figure 1 does show a very clear negative relationship, except for low values of turnover.

What is of clear interest are the details of the relationship – what level of turnover leads to the optimum performance, and how much difference do departures from this optimum make? The hypothesis testing format does not require us to work out such details; all that is required is a simple statement of whether or not the null hypothesis has been rejected. Glebbeek and Bax's article focuses on the hypothesis, and the fact that standardized regression coefficients are given in the tables without further explanation means that readers would have difficulty in answering questions about optimum turnover levels and the impact of departures from this, although these issues are discussed in the text.

This suggests that the helpful aim for a study such as this is not to test a hypothesis, but rather *to assess or measure a relationship between two, or more, variables*. The main answer is then given by graphs such as Figures 1 and 2 above, or by parameters such as the optimum turnover and curvature in Table 2.

Neither of these, however, addresses the sampling error issue: to what extent can we be sure that the results would be similar with another sample? As discussed above, the main alternative to null hypothesis tests for this purpose is to derive interval estimates for the important parameters. Deriving confidence intervals for the curvilinear model in Table 2 is a little difficult, so I will use a slightly different approach for this model. However, for a linear model, confidence intervals can easily be estimated by standard software. Table 3 incorporates such confidence intervals for Model 3 in Panel A of Table 2 of Glebbeek and Bax (2004). This model differs from the model in Figure 1 only by the omission of the square term.

Table 3. Linear Model (3 in Panel A of Table 2 in Glebbeek and Bax, 2004)

	Best estimate	*Lower limit of 95% CI*	*Upper limit of 95% CI*
Predicted impact of 1% increase in staff turnover	-1,778	-3,060	-495
Predicted impact of 1% increase in absenteeism	-3,389	-6,767	-10
Predicted impact of 1 year increase in average age	-731	-3,716	2,254
Predicted difference between neighbouring regions (with Region 1 having the lowest performance)	15,066	5,607	24,525
Proportion of variation explained (Adjusted R^2)	12%		

In Table 3, for example, the best estimate for the impact of staff turnover is that each extra 1% will reduce performance by 1,778. However, the exact amount is uncertain: the confidence intervals suggests that the true impact is somewhere between a reduction of 495 units and 3,060 units with 95% confidence. This interval *excludes* zero, which means that the significance level must be less than 5% (100% - 95%); in fact the significance level, *p,* is less than 1%, meaning that the 99% confidence interval would also include only negative values. On the other hand, the 95% confidence interval for the impact of age includes both positive and negative values, which means that it is not possible to reject the null hypothesis that age has no impact at the 5% significance level. In this way, all the information provided by the *p* values can be derived from the confidence intervals. The confidence intervals provide a direct and user-friendly assessment of the size of each impact, as well as the scale of the uncertainty due to sampling error.

This approach is not so easy for assessing the confidence in the conclusion that the curve is an inverted U shape because this is measured by two parameters – location and curvature in Table 2. Some software packages will plot confidence bands (e.g. Minitab, but not SPSS) but only for regression with a single independent variable (i.e. without taking account of the control variables).

Bootstrapping provides an easy approach to this problem. The procedure is to take successive resamples, with replacement, from the original sample (see, for example, Diaconis and Efron, 1983; Simon, 1992; Wood, 2005a). Each resample can then be taken as representing a possible population from which the original sample was drawn, and the variation between the resamples can be used as the basis for confidence statements. The bootstrap approach normally gives fairly accurate results; the argument is particularly convincing if the resample distribution is reasonably symmetrical about the actual sample results (Wood, 2003: 111; Wood, 2005a). The bootstrap procedure for the present data is implemented on the Excel spreadsheet at http://userweb.port.ac.uk/~woodm/BRQ.xls. (There is a slightly more detailed discussion of bootstrapping in Appendix 2.)

Pressing the F9 key on the Resample sheet of this spreadsheet produces another resample – another "guessed" sample from the same source, representing another possible population – and another graph representing the relationship between turnover and performance like Figure 1. Some of these are obviously inverted U shapes, whereas others are not. Figure 3 shows predictions derived from 5 of these resamples. The line in Figure 1 based on the actual data is clearly somewhere in the "middle" of these lines; the resample lines are distributed reasonably symmetrically about this line giving an indication of the results that might have been obtained from different (random) samples.

Figure 3: Predictions from data (bold) and 5 resamples for the model in Figure 1

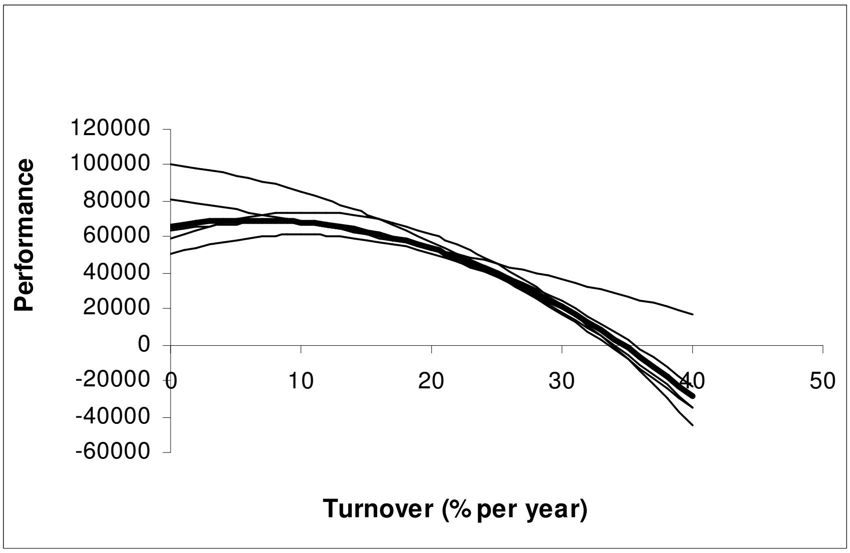
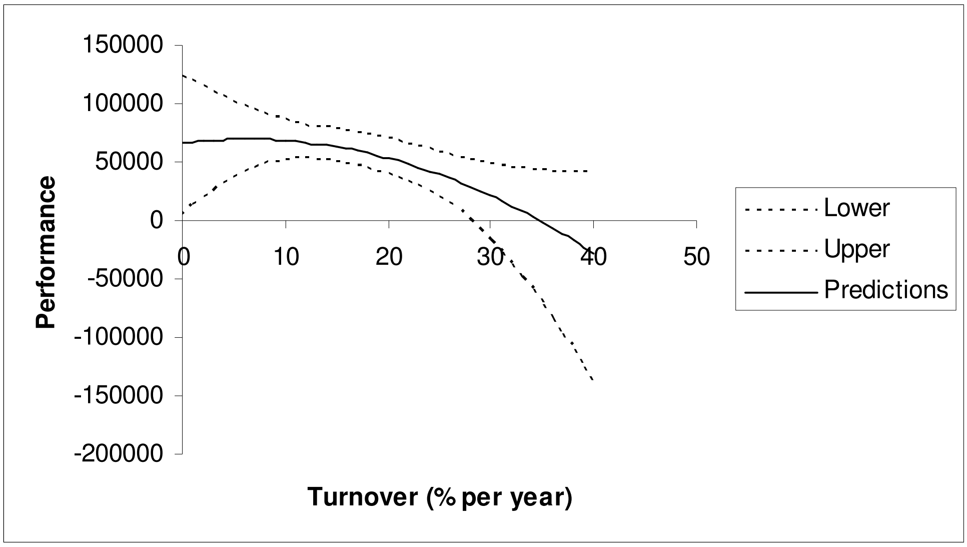


Figure 3 gives a clear demonstration of the fact that Figure 1 may be misleading, simple because two of the five lines are not inverted U shapes. With 10,000 resamples, 65% produced an inverted U shape (with a negative curvature and positive value for the location of the optimum). *This suggests a confidence level for the inverted U shape hypothesis of 65%.*

Figure 3 can also be presented in confidence band format: Figure 4 shows a 95% confidence band corresponding to the 2.5 and 97.5 percentiles of the predictions from 10,000 resamples at each level of Turnover. (This confidence band should be clearly distinguished from *prediction intervals,* which also take account of the additional uncertainties involved in predicting individual observations.)

Figure 4: Confidence bands for the model in Figure 1 based on 10,000 resamples



These bootstrap results are roughly consistent with the formulae from conventional software. For example, if the bootstrap spreadsheet is adjusted so that it follows the linear model in Table 3 (http://userweb.port.ac.uk/~woodm/BRL.xls), the bootstrap confidence interval (using 10,000 resamples) for the predicted impact of 1% increase in staff turnover extends from

– 3,119 to – 452

which is similar to the results in Table 3.

It is also possible to use p values to derive confidence statements. Because of the method by which conventional confidence intervals are constructed, a (two tail) significance level of p corresponds to a 1-p confidence interval extending from zero (or whatever the null hypothesis level of the parameter is) either upwards or downwards. This means that our confidence level for the parameter being either above or below zero is $1\text{-}p+ p/2$ (the extra $p/2$ being the other end of the confidence interval), or

$1\text{-}\ p/2$

For example, the p value for the predicted impact of turnover for the model in Table 3 is 0.7% (from the Excel Regression Tool or SPSS). *This corresponds to a confidence level that this impact is negative of 100% – 0.7%/2 or 99.65%.* (The corresponding result from the bootstrap method is 99.54%.) This statement avoids the counter-intuitive approach of the *p* value. (There is a slightly more detailed discussion of this in Appendix 1, and a similar idea in Hopkins, 2002.)

However, unlike the confidence interval in Table 3, this gives little indication of the size of the effect. This is also true of the confidence level of 65% (see above) for the curve being an inverted U shape. In both cases the result does not distinguish between small and possibly trivial effects, and large and potentially interesting ones.

We should also note that, strictly, a confidence level for an interval or a hypothesis is *not* the same as a probability of the truth of the hypothesis or of the true parameter value being in an interval (Nickerson, 2000: 278-280). Like null hypothesis tests, confidence intervals are based on probabilities of sample data given the truth about a parameter. To reverse these probabilities and find the probability of a hypothesis given sample data, we need to use Bayes theorem and take account of prior probabilities. However, for many parameters, including the slope of a regression line and the difference of two means, the Bayesian equivalent of a confidence interval, the credible interval, is identical to the conventional confidence interval (Bolstad, 2004: 214-5, 247) provided we use "flat" priors (i.e. we assume a uniform prior probability distibution) for the Bayesian analysis. *This means that it is often reasonable to interpret confidence intervals and levels in terms of probabilities*: the only loss (from the Bayesian perspective) is that any prior information is not incorporated. (There is more detail of this in Appendix 1.)

The usefulness of statistical results in this context

Having looked at how the statistical analysis is done, I now turn to the question of whether it is worthwhile even if it is done in the best possible manner, with sensibly chosen samples and so on. The first thing to note here is the distinctive style of statistical reasoning – discussed above. The conclusions are not deterministic, but are qualified by probabilities or equivalent concepts (and there is an assumption that there will be no "black swan" events). This can either be viewed as an inevitable limitation, or as the key strength of the approach. In the case of Glebbeek and Bax (2004), the situation is too complex for a complete, deterministic explanation of all the ways in which staff turnover has an impact on performance to be possible: the statistical approach is therefore worth considering. Figure 1 illustrate both the complexity of the situation – by the scatter in the diagram – and the way statistics can be used to give us the prediction line and assess how much of the variability is explained by the line (13%). Given this, it is helpful to distinguish four issues with a bearing on the usefulness of statistical methods in a given context.

The first issue is that statistical research has to focus on *easily measurable variables*. Otherwise it is not practical to obtain useful sample sizes. It is in a sense the opposite of "qualitative" research which "can provide thick, detailed descriptions in real-life contexts" (Gephart, 2004: 455). This means that the information analysed by statistical methods may be superficial in the sense that it is close to surface, althought his does not, of course, mean that it is necessarily uninteresting. Glebbeek and Bax's data says nothing about *how* high turnover influences performance. This is not an argument against using a statistical analysis, but it may be an argument for supplementing a statistical analysis with a more detailed qualitative study of a smaller sample. This principle seems to be widely accepted in theory, although possibly not always in practice.

The second issue concerns the *generality of the target context*. Glebbeek and Bax's results are based on data from a single organisation in a single country (the Netherlands) in one time period (1995-1998 during an economic boom). If we define the *target context* as the context the research is targetting and to which the results can reasonably be generalised, this is perhaps similar organisations in the same country at a similar time in history. Obviously, a different context may lead to a different pattern of relationship between staff turnover and performance, so their conclusions are qualified by words such as "can". The inverted U shape hypothesis is in no sense demonstrated in general terms, but they have shown that it is a possibility because it applies to this one target context.

The scope of the target context is a key issue in this, and most other empirical management research. The difficulty with making the target context too broad is that it becomes difficult to obtain reasonable samples, and the specific contextual factors are likely to add to the noise factors with the result that the proportion of variation explained is likely to be lower, although new methods of analysis may help (Bamberger, 2008). On the other side, if the context is too narrow this may lead many to conclude that the research is of little relevance.

The notion of a target context becomes more subtle when we consider the time dimension, or when we extend the idea to include what is possible. Most management research, and Glebbeek and Bax's is no exception, has the ultimate purpose of improving some aspect of management in the future. The aim of empirical research is to try and ascertain what works and what does not work. Let's imagine, for the sake of argument, that we had a similar data from a representative sample of a broader target context: all large organizations in Europe over the last ten years. This would certainly be useful, but the context might change over the next few years so we would have to be cautious in generalising to the future. The difficulty with almost any target context for statistical management research is that it depends crucially on contextual factors which may change in the future. Although it is perhaps a worthy aim to extend our theories to incorporate these contextual factors, this may make the resulting theories messy and unmanageable. Perhaps we should try to focus on the core, immutable, truths instead? This difficulty, of course, is that there may may be no such immutable truth other than the fact that it varies from situation to situation – in which cases statistical analysis would be of limited interest.

A comparison with medical research may be instructive. The target context here might be people, perhaps of a particular age or gender. With a management study the target context would typically be organisations or people in a particular business context. The problem with the business context, but not the medical one, is that it is a man-made context which may differ radically between different places or different time periods, making extrapolation from one context to another very difficult. Can conclusions about how staff turnover affects one employment agency's performance in an economic boom in the Netherlands be assumed to apply to universities England in 2009, or to social networking websites in California in 2010? Almost certainly not. In medicine, research with a restricted local sample may be of wider value simply because people are less variable from a medical point of view than business environments are from a management point of view.

The third issue concerns the *proportion of variation accounted for* by the model, or the strength of the effect relative to the effect of noise variables. The model in Figure 1 and Table 2 only explains 13% of the variation (variance) in performance between the branches. The question then arises of whether a model which explains so little of the variation is worth the effort of deriving. The answer obviously depends on the theoretical and practical context, but the question is certainly worth posing.

The final issue concerns the relationship between the results of a statistical analysis and other hypotheses and knowledge about the topic. How much *added value* does the research provide? In the present case, the prevailing wisdom is probably that performance has a tendency to fall as staff turnover rises. Figure 1 is largely consistent with this: it is only an inverted U shape in a marginal sense because there is insufficient data with a low level of turnover to demonstrate more than a small part of the rising (left hand side) part of the curve (only 8 of the data values are less than the optimum turnover level of 6.3%). In other words, the analysis has modified the prevailing wisdom only to a very small extent. Furthermore, as pointed out above, there is a sense in which the inverted U shape would be expected on the basis of "common sense" – both zero percent staff turnover, and 100% turnover, are likely to lead to relatively poor performance, so the optimum value must be somewhere between these extremes.

Reasons for the lack of statistical survey data in this article about statistical methods

The argument in the present article does not draw on statistical surveys of methods used in published research, or on statistically analysed experiments, or on empirical data of readers' understanding of statistical research. Obviously this means that no conclusions can be drawn about how typical some of the flaws described are; the argument is simply one of exploring various possible problems and how they might be resolved. There would certainly be a case for including some statistical research of this kind, but its value would be limited for three of the four reasons outlined in the previous section.

First the argument needs a "thick, detailed" analysis of the methods used in a particular project so that this can be used as demonstration of what *might* be possible in other contexts as well as the one studied. For example, it is necessary to demonstrate the alternative to null hypothesis testing to make the case that it is viable, for example. It would be difficult to incorporate this sort of detail into a statistical survey. This type of "illustrative" inference (Wood and Christy, 1999; Christy and Wood, 1999) is very different to statistical inference, but it is still an approach to generalising results from the sample studied.

Second, there is the problem of the target context: any such research would be based on a restricted sample of people and publications, and although it might be useful in context, how generalizable the results would be to very different contexts which may arise in the future would be inevitably debatable.

And third, statistical analysis of such things as methods used in publications (similar to the survey in Becker, 2005), user comprehension, and so on, would add a little to the value of this article, but probably not that much. The argument depends on readers' accepting that, for example, misuse of null hypothesis tests is sufficiently widespread to be a problem, but it does not hinge on any detailed data on how prevalent this problem is. This is dismissed as "exampling" by Gephart (2004), but examples which are reasonably obviously typical of at least part of the population can surely have real value.

For these reasons no statistical study of research methods employed by actual researchers was undertaken for this article. Hopefully, this article about statistical methods will itself serve as an illustration of the value of research which does not employ statistical methods.

Conclusions and recommendations

I have looked in detail at just one research article. The conclusions which might be derived from another article would doubtless be slightly different. However, it is possible to formulate the following general suggestions, which are illustrated by Glebbeek and Bax (2004), but which are likely to be of wider applicability. These are described as suggestions, and prefaced by the word "consider", to emphasize their tentative status.

1 Consider whether any statistical approach is likely to be useful. The advantage of statistical methods is that they enable us to peer through the fog of noise variables to see patterns such as the curve in Figure 1. But for a statistical analysis to be worthwhile it is necessary to check four issues: whether the necessity to focus on easily measurable variables damages the credibility of the results, whether the target population is likely to be of lasting interest, whether the amount of variation explained is likely to be sufficient to justify

the effort, and, taking these factors into consideration, whether the research makes a useful addition to existing knowledge (including "common sense"). In the case of Glebbeek and Bax (2004), the variables – staff turnover and performance – are certainly of interest. However, it is arguable that the target context is a little restricted, that the proportion of variation explained is on the low side, and that the extent of the addition to current knowledge is debatable.

2 Consider avoiding research aims which simply comprise a series of hypotheses to test. Instead, the aim should be to assess the relationship between variables – as numerical statistics and/or in the form of graphs. In the case of Glebeek and Bax (2004) this would lead to Figures 1 and 2, and Table 2 above. These have the advantage of provided a far more direct and user-friendly answer to questions about the effects of turnover, and the control variables, on performance. Conventional quantitative research based on null hypothesis tests is often strangely non-quantitative because readers are told very little about the *size* of impacts, differences or effects. Figure 1 is an inverted U shape (which supports Glebbeek and Bax's hypothesis), but because the left hand side is largely missing it is, in practical terms, very similar to a negative (downhill) relationship.

3 Consider using confidence intervals to assess the uncertainties due to sampling error – e.g. Table 3 and Figure 4 above. These tell readers about the size of the effects (impacts on performance), *and* the likely uncertainties due to sampling error, in a far more useful way than null hypotheses and *p* values. (And, for the purists, problems over the interpretation of confidence intervals can be resolved by viewing them in terms of Bayesian credible intervals with flat priors, so quibbles about the distinction between confidence and probability can be ignored.)

4 Alternatively, if formal hypotheses are to be evaluated, consider using confidence levels for this purpose. For example, in the model in Table 3, the confidence level for the hypothesis that the impact of Turnover is negative is 99.65%. This can easily be worked out from the corresponding *p v*alue. Such

confidence levels avoid a hypothetical and potentially confusing null hypothesis.

5 Bootstrap methods may be helpful for estimating confidence levels in a transparent and flexible manner. The confidence intervals in Table 3 can be produced by bootstrapping, as can the graphs in Figures 3 and 4 (which were produced by spreadsheets linked to http://userweb.port.ac.uk/~woodm/BRLS.xls – these spreadsheets are designed to be flexible so that they can adapted to other models and data). Figure 3, based on five bootstrap resamples, shows in a very direct manner how much the results from several samples from the same source might have varied. Bootstrap methods may be more powerful in the sense that they are sufficiently flexible to be adapted to the question at hand: it is difficult to see, for example, how the 65% confidence level for the inverted U shape hypothesis for the model in Figure 1 could have been obtained from conventional software and methods. The same method could obviously be extended to other hypotheses about the shape of functional relationships (e.g. those in Shaw et al, 2005).

6 How easy to understand are readers likely to find the results? Consider reformulating results to make them more user-friendly – perhaps by giving brief descriptions of the interpretation of parameters, or by using parameters which have an interpretation which is as simple and directly useful as possible (e.g. unstandardized instead of standardized regression coefficients), or perhaps by using graphs more than is customary. I have tried to do this in relation to Glebbeek and Bax (2004) in Tables 2-4 and Figures 1-4.

Acknowledgment

I am very grateful to Dr. Arie Glebbeek for making a version of his data available to me.

References

Ayres, I. (2007). *Super crunchers: how anything can be predicted*. London: John Murray.

Bamberger, P. (2008). From the editors. Beyond contextualization: using context theories to narrow the micro-macro gap in management research. *Academy of Management Journal, 51*(5), 839-846.

Becker, T. E. (2005). Potential problems in the statistical control of variables in organizational research: a qualitative analysis with recommendations. *Organizational Research Methods, 8*(3), 274-289.

Bolstad, William M. (2004). *Introduction to Bayesian statistics (2nd edition).* Hoboken, New Jersey: Wiley.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypotheses: a review of contemporary management research and recommendations for future studies. *Organizational Research Methods, 7*(2), 151-167.

Chiasson, M., Fildes, R., & Pidd, M. (2006). Intelligent thinking instead of critical realism. *Journal of the Operational Research Society, 57*, 1373-1375.

Christy, R., & Wood, M. (1999). Researching possibilities in marketing. *Qualitative Market Research, 2*(3), 189-196.

Cortina, J. M., & Folger, R. G. (1998). When is it acceptable to accept a null hypothesis: no way Jose? *Organizational Research Methods, 1*(3), 334-350.

Diaconis, P., & Efron, B. (1983, May). Computer intensive methods in statistics. *Scientific American, 248*, 96-108.

Gardner, M., & Altman, D. G. (1986, March 15). Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal, 292*, 746-750.

Gephart, R. P. J. (2004). Editor's note: Qualitative research and the Academy of Management Journal. *Academy of Management Journal, 47*(4), 454-462.

Glebbeek, A. C., & Bax, E. H. (2004). Is high employee turnover really harmful? An empirical test using company records. *Academy of Management Journal, 47*(2), 277-286.

Gregory, V. L. (2008). Effectiveness of varying communication media in enhancing learning. *Paper presented at the Third Interdisciplinary Social Sciences Conference*, Prato, Italy., Grinyer, J. R., Collison, D. J., & Russell, A. (1994). The impact of financial reporting on revenue investment: theory and evidence. *British Accounting Review, 26*, 123-136.

Hopkins, W. G. (2002). *A New View of Statistics.* http://www.sportsci.org/resource/stats/pvalues.html, accessed 19 March 2010.

Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement*, 56(5), 746- 759.

Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability in inference. *The Annals of Statistics, 9*(1), 45-58.

Lindsay, R. M. (1995). Reconsidering the status of tests of significance: an alternative criterion of adequacy. *Accounting, Organizations and Society, 20*(1), 35-53.

Mingers, J. (2006). A critique of statistical modelling in management science from a critical realist perspective: its role within multimethodology. *Journal of the Operational Research Society, 57*(2), 202-219.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy*. London: Butterworths.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Ormerod, R. (2006). The OR approach to forecasting: comments on Mingers' paper. *Journal of the Operational Research Society, 57*, 1371-1373.

Rihoux, B., & Ragin, C. C. (2009). *Configurational comparative methods: qualitative comparative analysis (QCA) and related techniques*. Los Angeles: Sage.

Shaw, J. D., Gupta, N., & Delery, J. E. (2005). Alternative conceptualizations of the relationship between voluntary turnover and organizational performance. *Academy of Management Journal, 48*(1), 50-68.

Sibson, R. (1999). Statistics. in A. Bullock, & S. Trombley (eds), *The new Fontana dictionary of modern thought*. London: HarperCollins.

Siebert, W. S., & Zubanov, N. (2009). Searching for the optimal level of employee turnover: a study of a large UK retail organization. *Academy of Management Journal, 52*(2), 294-313.

Simon, H. A. (1996). *The sciences of the artificial*. Cambridge, Massachusetts: MIT Press.

Simon, J. L. (1992). *Resampling: the new statistics*. Arlington, VA: Resampling Stats, Inc.

Taleb, N. N. (2008). *The black swan: the impact of the highly improbable*. London: Penguin

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.

Vandenberg, R. J. (2006). Statistical and methodological myths and urban legends: where, pray tell, did they get this idea? *Organizational Research Methods, 9*(2), 194-201.

Westfall, P. H., & Hilbe, J. M. (2007, August). The black swan: praise and criticism. *The American Statistician, 61*(3), 193-194.

Wood, M. (2002). Maths should not be hard: the case for making academic knowledge more palatable. *Higher Education Review, 34*(3), 3-19.

Wood, M. (2003). *Making sense of statistics: a non-mathematical approach*. Basingstoke: Palgrave.

Wood, M. (2005a). Bootstrapped confidence intervals as an approach to statistical inference. *Organizational Research Methods, 8*(4), 454-470.

Wood, M. (2005b). The role of simulation approaches in statistics. *Journal of Statistics Education, 13*(3), http://www.amstat.org/publications/jse/v13n3/wood.html.

Wood, M. (2006). Viewpoint: Statistics and management science. *Journal of the Operational Research Society, 57*(11), 1369-1370.

Wood, M., Capon, N., & Kaye, M. (1998). User-friendly statistical concepts for process monitoring. *Journal of the Operational Research Society, 49*(9), 976-985.

Wood, M., & Christy, R. (1999). Sampling for possibilities. *Quality & Quantity, 33*, 185-202.

Wood, M., Kaye, M., & Capon, N. (1999). The use of resampling for estimating control chart limits. *Journal of the Operational Research Society, 50*, 651-659.

Yin, R. K. (2003). *Case study research: design and methods (3rd edition).* Thousand Oaks, CA: Sage.

Appendix 1: The construction and interpretation of p values (significance levels), confidence intervals and Bayesian credible intervals

The *p* value for the predicted impact of turnover (i.e. the regression coefficient) in the model in Table 3 is 0.7% (from the Excel Regression Tool, or SPSS). This means that, if we assume the truth of the null hypothesis that the population value of this parameter is 0 corresponding to the assumption that turnover has no impact, positive or negative, on turnover, then chance fluctuations would lead to a value of -1,778 or less or +1,778 or more, with a probability of 0.7%. This low probability means that the null hypothesis is most unlikely to be true, so we can assert that the evidence shows there is a real association which would be likely to occur again if we took further samples.

It is important to be aware that this *p* is not in any sense the probability of the null hypothesis being true. The null hypothesis is that the true impact (regression coefficient) is *exactly* 0: i.e. not 1 or -1, or even 0.1, but exactly zero. Given how may possible numerical values there are, the probability of the value being exactly zero must be very low, for all practical purposes, zero. This would apply even if the *p* value was actually fairly high – e.g. the *p* value for the impact of age in Table 3 is 63%, but this is certainly not the probability of age having *exactly zero* impact.

Despite this, something like a probability of a hypothesis being true would be useful. The usual way of achieving is by using *confidence* intervals. Here, we take not a single value of our parameter, but a range of values, and work out a confidence level of the truth being in this range. Normally we start with a conventional confidence level, and then work out the size of the corresponding interval – these are the 95% confidence intervals in Table 3. The probability calculations involved here are of the same type as those used to work out *p* values – we start by imagining that particular value for a parameter is the true one, and then work out the probabilities of various possible sample results from this. The similarity of the two methods should be obvious if we work out the 100 – *p* or 99.3% confidence interval for the impact of turnover (which is easy with the

Regression Tool in Excel). This interval extends from zero down to -3,555: zero, of course, being the null hypothesis value of the parameter.

The idea of confidence can be extended beyond conventional confidence intervals. For example, we have seen that we have 99.3% confidence that the true impact of turnover lies between -3,555 and zero. Because of the way confidence intervals are constructed, the 0.7% confidence that the true value is not in this interval is equally split between the two ends: 0.35% confidence that the true value will be greater than 0, and 0.35% confidence that it will less than -3,555. This gives a total confidence of the impact being negative of 99.65, and 0.35% confidence that the true impact is positive. This seems a clear, and potentially useful, statement. However, this step seems to be rarely, if ever, taken.

The use of the word "confidence" stems from the fact that "95% confident" is *not* normally seen as meaning "with a 95% probability". However, the difference is subtle (Nickerson, 2000), and arguably of little real importance. A *p* value is a probability of particular data values *given* the truth of a hypothesis (a null one). The problem is to reverse this and derive a probability for the hypothesis *given* particular data values. Mathematically, this can be achieved by using a result from probability theory known as Bayes theorem: this is the basis of the Bayesian approach to statistics and can be used to derive a Bayesian equivalent of confidence intervals called credible intervals (e.g. see Bolstad, 2004).

In practice there is a problem. Bayes theorem requires knowledge of *prior probabilities*. These reflect beliefs prior to obtaining the data, and are essential for the estimation of credible intervals. They may be difficult to ascertain in a meaningful sense, and tend to be viewed by conventional statisticians as bringing unwelcome subjectivity into the process. However, in practice, it is possible to use "flat priors", corresponding to an assumption that all values are equally likely, in Bayes theorem. If we do this, in many cases, including our regression example, the Bayesian credible intervals are identical to conventional confidence intervals (Bolstad, 2004). This means that confidence intervals can be interpreted in terms of probabilities if we make "flat" assumptions about prior knowledge. *This assumption is an innocuous one, so it is reasonable to ignore the distinction between confidence and probability, and think of a 95% confidence interval as* *being an interval which has a 95% probability of including the true value of the parameter.*

Appendix 2: the bootstrap method for deriving confidence intervals

The bootstrap approach used here is the simplest bootstrapping methods for estimating confidence intervals – the bootstrap percentile interval. There are 110 records in the data on staff turnover. Resampling with replacement means that we choose one of these at random, then replace it so that we are starting again with the original sample, and then choose another at random, and so on until we have a "resample" of 110. All records in the resample come from the sample, but some records in the sample may appear several times in the resample, and others not all. We then follow this procedure repeatedly to generate multiple resamples.

Now, imagine that the population from which the real sample is drawn follows the same pattern as the sample. This means that 0.91% (=1/110) of the population will be like the first record in the sample, 0.91% like the second, and so on. This, in turn, means that to take a random sample from this population we want to choose a record like the first member of the sample with a probability of 0.91%, and similarly for the second, third and so on. But this is exactly what resampling with replacement achieves, so these resamples can be regarded as random samples from this "guessed" population. This guessed population is not the real population, but it is reasonable to assume it is similar, so that the variation between the resamples gives a good idea of sampling error when sampling a real population, and the distribution of statistics worked out from the resamples should be a reasonable surrogate for a confidence distribution. In practice, experience (including the comparisons made here) shows that bootstrapping generally gives very similar results to conventional methods where these are possible. For more details, readers are advised to refer to the citations above, and experiment with the spreadsheets at http://userweb.port.ac.uk/~woodm/BRLS.xls.