Comments on the Spirit Level Controversy

Hugh Noble

When Richard Wilkinson and Kate Pickett published "**The Spirit Level**", in 2009 [1], the book attracted a great deal of both praise and criticism. What the authors claimed in that book, is that across many populations, there is a significant statistical correlation between rising levels of income inequality and rising levels of social problems - such as unplanned teenage pregnancy, the levels of drug abuse and the proportion of a population in prison. To arrive at that conclusion, Wilkinson and Pickett (W&P) analysed the statistical data relating to over 20 different types of social ill and the result was established mainly by drawing straight-line regression curves through the relevant scatter diagrams.

Those of a left-wing persuasion saw in this thesis, and in these research results, a vindication of their own long-held views about the desirability of redistribution designed to achieve wealth equality. Those of the right saw in it a threat to what they regarded as the economic efficiency of merited inequality.

I declare partiality. I consider myself to be a supporter of the Wilkinson and Pickett thesis. But it is not the political merit or de-merit of W&P's thesis which I want to address in this article. I want, instead, to focus on the very poor quality of the discussion which followed publication of the book, and the disappointing level of understanding of basic statistics theory which characterised most of the critical arguments. It was also the case that those criticisms were delivered with an affected air of intellectual superiority, a characteristic which then carried over (without serious examination) into much of the press reporting which we can suppose would form the basis for an understanding of the issues by the lay public.

I am not a professional statistician. I am a retired scientist who has, over the course of a varied career, applied many different statistical techniques, in diverse research projects, within a range of scientific disciplines. Over the years I have learned (and been strongly reminded) that there are hidden assumptions built into most statistical tests and methodologies, and that adopting these assumptions without due consideration and in inappropriate circumstances is a mistake that should be avoided as a Mastodon should avoid a tar-pit.

W&P used two different datasets. The first was drawn from the published statistics relating to some 23 different countries all of which were relatively rich in terms of GDP per head of population. The second dataset related to all of the individual states of the USA. These two datasets yielded broadly similar results.

The datasets were restricted to relatively rich populations because it is well established that for countries below a certain level of GDP per head, there is a strong positive social benefit obtained by increasing GDP. The W&P thesis is that beyond a "knee-bend" point (of which Cuba, China, Uruguay and Romania are representative examples), a law of diminishing returns sets in. Beyond that point further increases in GDP may be desirable for many reasons but, according to Wilkinson and Pickett, increases in GDP have little beneficial effect on the incidence of certain social ills which the authors examined. For countries (and for the US states) which are all above that knee-bend point, the authors argue that those wishing to improve social conditions should instead direct their attention to measures which reduce income inequality. The surprising part of this result is that W&P claimed to have found that it was not only the poor who suffered in countries (or states) with the greatest income inequality but the whole population. It was that part of their conclusion which galvanised the subsequent debate.

The most conveniently accessible criticism of The Spirit Level comes from an organisation called "Policy Exchange". This is a right-wing "think-tank" which provides political advice to various other organisations including the current British Conservative party. The criticism they offer in this case, takes the form of an article entitled "**Beware of False Prophets**". The paper was written by Peter Saunders, who is an advisor to the organisation and an emeritus professor of Sociology. The document is available, free of charge, on the Internet [2].

The arguments which Wilkinson and Pickett have themselves offered to counter these criticisms, concentrate on the fact that the data on which their thesis is based, come from a variety of academically respectable studies and peer-reviewed publications. Those comments are also available, free of charge, on the Internet [3]. In a paper published in the peer-reviewed journal Social Science and Medicine [4] W&P listed some 168 analyses in 155 papers which address the issue of health and income inequality. Of those, the majority (70%) were supportive of the W&P thesis. The strength of that additional evidence from sources other than W&P, seems to have been generally ignored by critics and by subsequent press reports.

But, as remarked above, my own concern is with the validity of the statistical analysis which was offered by those critics and by Saunders

in particular. In an article, which has been published on the Equality Trust website [5] (and of which this article is a shortened version), I analysed several technical mistakes in Saunders' analysis. My article was aimed at those who have little expertise in statistics and for that reason I eschewed mathematical analysis and concentrated instead on easily understood examples which illustrate with clarity the problems I identified.

In this article I will do the same, but because of editorial space restrictions, I want to concentrate on just three issues – (1) Linear regression, (2) correlation (and what it can tell us about causality), and (3) the problem of identifying so-called "outliers".

Consider first the issue of linear regression. It is always possible to draw a best-straight-line through any scatter diagram. The example I used in that earlier article to illustrate that point, concerned some invented (but plausible) datum-points giving the death rate in a population during a succession of heat waves. The graph, as we might expect, indicated that the number of deaths rises in an upward curve as ambient temperature rises beyond blood heat. But the curve is more or less flat in the region below that critical temperature. If we project the graph to very high temperatures our expectation is that the graph will become vertical indicating that at some improbably high temperature, everybody would die. But even with such an obviously non-linear graph, it is still possible to draw a linear regression line through the scatter diagram (for the section of data which we have) and we are still justified in drawing, from the calculated significance of that regression line, the conclusion that there is a significant relationship between temperature and death rate.

Peter Saunders, however, in his critique, wants to deny the significance of the correlation found by W&P. He is adamant that the drawing of a linear regression line is valid only if the underlying relationship is in fact genuinely and strictly linear.

"... regression techniques are quite demanding. They not only require that the slope of the trend line should not be distorted by a few extreme cases, but also that the association between variables be linear. (i.e. as the value of X increases, so the value of Y should increase or decrease at a fairly steady rate across the whole distribution) ..." [PS: 55]

Pointing out that within some of the graphs used by W&P to illustrate their thesis, a uniform slope of line is not obvious, he declared that …

"… *a key requirement of regression analysis has been violated*" [PS 57]

I dispute that claim. In doing so, I rely not on my own version of statistical theory, but on the words of one of the grandmasters of statistical theory – M.G. Kendal, using quotations from the "The

Dictionary of Statistical Terms" which he wrote with W.R. Buckland [6].

"Regression Curve: A diagrammatic exposition of a regression equation. .... The term is sometimes interpreted to mean a regression equation of a higher degree than first, [i.e. not a straight line] *the emphasis then lying on the word "curve" as opposed to a straight line."*

[Kendal and Buckland 1957]

"Regression Line: In general this is synonymous with regression curve, but is sometimes (and rather ambiguously) used to denote a linear regression."

[ibid]

Saunders' insistence that strict linearity is a key requirement of linear regression analysis is clearly not shared by Kendal. Regression lines can be curved. The drawing of a best-straight-line is merely a first approximation which may or may not indicate the existence of a significant relationship. If we have some idea of the nature of that relationship we may explore the idea by plotting best-fit-lines using higher degree equations.

However, most curve fitting procedures are based on the assumption that residuals (the differences between the actual positions of plotted points and the corresponding positions on the regression curve) are due to random errors of measurement, and that the magnitudes of these errors have a normal distribution. That is the rationale behind the method of least squares approach which is fundamental to most curve-fitting procedures. The justification of that assumption is based on the Theorem of Central Limits. That theorem and its conclusions assume that the total error of any measurement is the sum total of a very large number of very small random errors. In these circumstances a binomial distribution can be assumed. In practical circumstances and for large numbers of trials, the binomial distribution is so close to a normal distribution that we can ignore the discrepancy.

These assumptions are all valid when we are dealing with measurements made using scientific instruments or similar. But they are not obviously valid when we are dealing with measurements made using questionnaires or any of the other ways commonly used to gather data in the sociological field. So are we justified in accepting the data analysed by W&P, or the objections raised by Saunders using boxplots, which (as we shall see later) are based on similar assumptions concerning the normal distribution of errors?

I was worried about the validity of the "normal distribution" assumption. So I took a set of datum-points straight from a couple of graphs I found in The Spirit Level and applied to them the non-parametric test called The Kendal Rank Correlation test. This does not

use numerical parametric values but instead compares only the rank ordering of data. We take the numeric values and write them down in a sequence, largest value at the top and lowest value at the bottom. We then test the null hypothesis – is it plausible that we could have got that degree of similarity in two orderings, by a random roll of the dice? By doing that, of course, we throw away information about the degree of numeric difference, but we also throw away any dependence on assumptions about error distribution. We are dealing then only with the ordering of data pairs – this one is larger than that one with respect to this factor. The analysis is based (again) on the binomial theorem in the same way as we would analyse the likelihood of getting two sequences of coin-toss results which match to some extent.

When I did this test I got a result which confirmed the conclusion reached by Wilkinson and Pickett. The null hypothesis was improbable. The two orderings were significantly similar. Income inequality does indeed correlate in a significant way with the social ills I tested.

Before we leave the issue of linear regression however, it is worth considering this point – in all complex systems, biological, electronic, economic or social, non-linearity is the norm. Where a linear relationship can be found between any two variables (say X and Y with a possible Z as well), the linearity of that relationship will be only approximate and will hold only over a limited range of circumstances. Beyond those limits, feedback loops, time-delays and saturation effects will set in and destroy the apparent linearity. So a trend line can *look* linear, and may actually *be* linear over the range being analysed, while still being non-linear when we try to stretch the limits beyond that normal range.

When a system contains a great many inter-related variables, it is almost impossible to find two variables which are not causally linked in some way. Usually those connections are in the form of a network. In those circumstances, by tracing connections through the network along different pathways, a case can be made for saying that X causes Y, Y causes X, and Z causes both - all at the same time, (and all mediated by other unconsidered variables).

Consider this also – any system which contains within its own mechanisms a memory of its own past experience, cannot ever repeat any previous circumstance exactly, because on each subsequent repetition there will be present a memory of a previous experience which was not present during that past experience. In view of that we must abandon any simplified notions of “clockwork” determinism in a system which contains a memory of itself and which is influenced by unpredictable external events. The notion that any economic system

we can devise could ever be optimum and remain so without need of constant revision and re-adjustment, is naïve utopianism.

The second point I wish to address is this - when we do identify a correlation between two quantities, say Y on X, what valid conclusions may we draw from that? Various commentators have been quick, and apparently pleased to notice (and delighted to point out), that "*a statistical correlation does not imply a causal relationship"*. That seems to be the popular way of expressing this particular caveat.

But is that strictly correct? By using the phraseology "*causal relationship*" they seem to mean a relationship such that *X causes Y directly (without intermediate variables)*. If that is what is meant then they are of course correct. The literature is full of counter examples of apparently absurd causal connections. My own favourite is the correlation between the size of a child's big toe and the quality of that child's handwriting. The hidden variable which produces this relationship is, of course, increasing age.

But when we see a relationship of that kind which probably involves some third hidden variable, I would argue that there is, still present, *a kind* of causal relationship. It is perhaps arguable that we cannot say that increasing age actually "*causes*" an increase in toe size (that is a philosophical question which I hesitate to raise in this short article) but we cannot argue legitimately that causal connectivity is not involved at all. Body growth is deeply and intimately involved with the passage of time. So we are entitled to say that causality is involved in some way. I submit that the only valid restriction is that it is not necessarily a *direct* causal relationship or one in which we must necessarily have a particular interest.

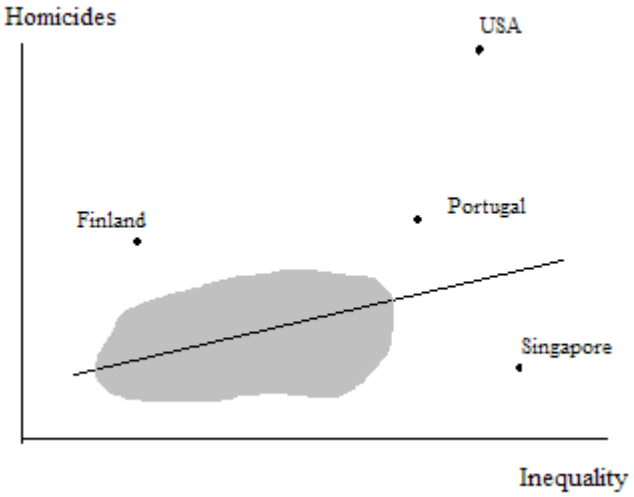
In the case of the data used by W&P, a significant correlation between, for example, income inequality and teenage pregnancy, may not be indicative of a direct causal connection, but it does indicate that something is going on which we should try to understand. The two factors are playing roles in that complex network of cause and effect that we call society. Each may well have an effect on the other in some way which, though it may not be direct, is still interesting and worthy of exploration. I suggest that to dismiss the correlation as uninteresting simply because it need not be a directly causal one, is just as much a mistake as assuming that the relationship must be a direct causal one.

However, and this is one reservation which I have about the Spirit Level analysis, the relationships between those various social ills and income inequality, even if they are each significant, may differ quite markedly from one another. For example the time-course of the inter-relationship may be quite dissimilar.

It is not immediately obvious why income inequality rather than simple poverty may be a cause of poor educational attainment, until one realises that income inequality will give some parents the financial clout to buy better educational facilities and that, in a competitive market environment, can deny resources to their poorer compatriots. Why would a well qualified teacher choose to work in a poorly-resourced school if a much more congenial environment was readily available? No doubt some altruistic teachers may do so, but there will also be others who would not.

These interrelationships and their outcomes are complicated, difficult to analyse and, over a period of extended time, almost impossible to predict. Non-linear relationships, as was mentioned earlier, are typically difficult to predict, but they may exhibit "attractor" points – circumstances towards which the system appears to exhibit a kind of gravitational attraction. While several of these attractor points may be easily identified, it is usually not possible to predict which of them will be the end-result of an on-going dynamic system. The conclusion we may draw from that is that we need to monitor the situation constantly and make small adjustments in an effort to arrive at desirable conditions.

The third point I want to explore here is the issue of "outliers". This



was the issue which was taken up and emphasised by most of the press reports on The Spirit Level I have read. We can see why outliers are considered important by looking at the graph shown here. That diagram is a simplified reconstruction of a typical example used in The Spirit Level.

The diagram shows the relationship between income inequality and homicides (per unit of population). The grey blob represents a cluster of datum points which correspond to Sweden, France, Canada, Australia, UK, Norway, Germany, Greece, New Zealand, Italy,

Switzerland, Belgium, Japan, Austria, Denmark, Israel, Spain and Ireland. I have not shown these individually because the exact position they occupy is not important to the point I am making here.

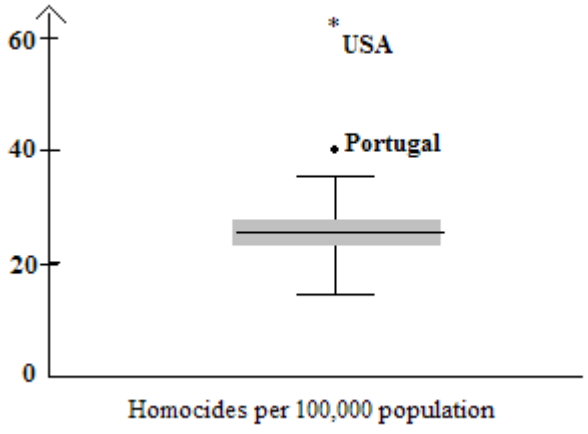
In his critique Saunders draws our attention to the position of the USA and Portugal. It is his contention that the USA is an anomalous "outlier" which should not have been included in the dataset.

"… look at the scatter of countries on the vertical (y) axis in figure 5a. Most of them seem to have homicide rates which are compressed in a range between about 10 to 20 murders per 100,000. The glaring exception is the USA … with its homicide rate of over 60 per 100,000. Judging by this graph we might expect that the USA is a unique case, and that its exceptionally high homicide rate is being caused by factors which are specific to that one country alone (the laxity of gun control laws is an obvious explanation)." [PS p29]

It is also clear from the quotation above that Saunders bases his identification of the USA as an outlier, on the fact that the number of homicides is far in excess of the number for other countries. To emphasise that point he shows us a "boxplot" of the data.

This, he claims "identifies" Portugal as an "outlier" and the USA as an "extreme outlier". When these points are removed, he claims, the regression line becomes flat and all significance is lost.

There are several things wrong with this claim. The first is that a boxplot is merely a method of graphical presentation, like a pie-chart or a histogram. It cannot take away the responsibility for *our* decisions about what is an outlier and what is not. It merely draws our attention



to possible candidates.

The second error is that in his consideration of the datum points for the USA and Portugal, Saunders has compared the raw data (i.e. positions on the Y-axis) with those of other datum points. He should have been comparing the residuals – that is, the vertical distance

between the points and the corresponding point on the regression line. When that is done we find that a boxplot also identifies Finland and Singapore as possible outliers. If these points are also removed from the dataset, the slope of the regression line is returned to its original value (or close to it). That is obviously not what Saunders wants however.

But there is an even more serious error associated with Saunders' analysis and identification of "outliers". I can illustrate this point with a simple example. Let us say that a surveyor is trying to establish the true height of Ben Nevis (the highest mountain in Britain). He makes a total of six measurements. Five of these are clustered close to 4406 feet, while the sixth gives a value of 6044 feet. The difference is startling and we might suspect that that sixth measurement was caused by some simple error such as transcribing the measurement figures in the wrong sequence. We might reasonably decide that that particular measurement is an "outlier" and should therefore be removed from the dataset rather than allow it to distort the mean value of the other readings which are in close agreement. Note that by taking the average value of the five good readings we are once again making that assumption about the normal distribution of errors. This is justified in the case of measurements taken using standard survey instrumentation.

Now consider another scenario. Our surveyor is now trying to measure the heights of several different mountains in the foothills of the Himalayas. He makes only one measurement of each mountain. On returning to his office and computing the results, he finds that one mountain within the dataset appears to be at least four times the height of any of the others. Instead of being between 6000 and 7000 feet, this one is over 29,000 feet. Could that be an error like our rogue measurement of Ben Nevis? Or could it be that through a gap in the surrounding hills and a break in the clouds, our surveyor has glimpsed the summit of Mount Everest a long way further to the North?

To answer that question, we need to think about the reason why we were justified in removing the rogue measurement of Ben Nevis. In that example all the measurements were of a single mountain and so we were justified in expecting them to be very similar. In calculating the arithmetic mean value of the 5 closely grouped measurements, we are justified (by the Theorem of Central Limits) in regarding that average value as being more likely to be a more accurate measure of the height of Ben Nevis than any of the individual measurements in the group of 5.

In the second example all the measurements were of different mountains. We therefore have no prior reason to expect them all to be

the same and we have no justification for thinking that that very different measurement (of Mount Everest) is an "outlier" - hence no justification for removing that datum point from the dataset. If we calculate the arithmetic mean of all those heights of all those different mountains, it is not at all clear what that average value represents. We have no reason to suppose that it is a more accurate measure of anything and no reason to be suspicious of any value which does not have a value close to that average value. Why would the heights of mountains have a normal distribution?

Note that a boxplot, in "identifying" outliers, makes exactly that same assumption about normal distributions. As we have seen earlier, that is not necessarily a valid assumption in the case of sociological data.

By using a boxplot to "identify" the USA datum-point as an "extreme outlier", Saunders has made that same hidden assumption - that the difference between the USA data and that of other countries, is due to some improbably large error of measurement. We can ask this question - Why would the measurements of social problems in different countries have a normal distribution? The idea, and the assumption being made by Saunders, in this argument about outliers, is simply absurd. The datum point for the USA is certainly anomalous, and it requires explanation, but it is not, in the sense the term is used in other more appropriate circumstances, an outlier.

There is no suggestion that the data for the USA are actually wrong. The USA is just very different from other countries with respect to social ills and income inequality. Saunders suggested that the reason why the USA has such a high rate of gun-related homicides is due to its lax form of gun control, and therefore, by implication, not due to its extreme income inequality as suggested by The Spirit Level authors. However, even if gun controls (or lack of them) is a contributing factor in this case, in view of the observations I made earlier about the network of causal connections implicit in any social order, we should ask this question:- In a country with a shockingly high rate of gun-related homicides – why should that society tolerate a lack of gun control laws? What causes that? Does its extreme form of income inequality contribute to that circumstance or are both caused by some other unidentified variable? One simply cannot discount a possible causal connectivity of one part of that network by pointing to another possible causal connection in another part of the network.

In the graph the datum point for Singapore is also anomalous. It too has a high level of income inequality but a much lower incidence of recorded social ills. That is a counter-example which seems to spoil the W&P thesis. If the Singapore datum point was removed the regression line would be much steeper and more significant. It would also be tempting to draw a regression line which curved upwards to

the right. So that datum point is certainly worthy of further examination (but not automatic elimination from the dataset).

One factor which has been drawn to my attention since I wrote on this topic previously, is that the work force in Singapore (which consists virtually of a single city) contains a high proportion of workers from other countries (mainly Malaya). These workers also occupy the lowest rungs on the income scale and are therefore included in (and influence) the statistics concerning inequality. However, the law in Singapore denies these workers many of the normal rights of Singapore citizenship - such as the right to get married or give birth to children. Any foreign worker who transgresses these rules is immediately deported and therefore does not appear in some of the recorded statistics concerning social problems - such as the statistics concerning teenage pregnancies.

If that information is correct it may go some way to explaining why Singapore seems able to buck the trend of the association between income inequality and increasing social ills. It does have these problems but it simply exports them elsewhere. I hasten to add however, that I have no special knowledge of Singapore and have no easy means of checking the validity of what I have been told. Clearly judgement should be reserved until more investigation has been carried out.

I offer that thought to justify my claim that rather than discard data, as Saunders suggests, because it does not fit comfortably with preconceived ideas, the correlation between social ills and income inequality, which Wilkinson and Pickett have identified, should prompt us to investigate more fully and without delay or prevarication. If the correlation which W&P claim to have identified does exist, it is a finding of central importance to political science and politicians in every country should pay attention, particularly those in the wealthier countries. It is much too important an issue for it to be discussed in a manner characterised by faulty statistical arguments.

References

[1] The Spirit Level – Wilkinson and Pickett, Penguin Books (2009/2010)

[2] Beware of False Prophets

http://www.policyexchange.org.uk/assets/Beware_False_Prophets_Jul_10.pdf

[3] W&P response

http://www.equalitytrust.org.uk/resources/other/response-to-questions

[4] Income inequality and population health: A review and explanation of the

evidence. Richard Wilkinson and Kate Pickett, Social Science and Medicine
62 (2006) 1768-1784
[5] The Spirit Level Revisited
http://www.equalitytrust.org.uk/docs/hughnobletslrevisited.pdf
[6] The Dictionary of Statistical Terms. Kendal and Buckland.
Oliver and Boyd 1957.

Hugh Noble
www.tartanhen.co.uk
hugh@tartanhen.co.uk