University of Hertfordshire

FITTING EQUATIONS TO DATA WITH THE PERFECT CORRELATION RELATIONSHIP

CHRIS TOFALLIS

Hertfordshire Business School Working Paper (2015)

This version: 23 December 2015

First version: September 2015

The Working Paper Series is intended for rapid dissemination of research results, work-in-progress, and innovative teaching methods, at the pre-publication stage. Comments are welcomed and should be addressed to the individual author(s). It should be noted that papers in this series are often provisional and comments and/or citations should take account of this.

Hertfordshire Business School Working Papers are freely downloadable from https://uhra.herts.ac.uk/dspace/handle/2299/5549 and also from the British Library: www.mbsportal.bl.uk

Hertfordshire Business School employs approximately 200 academic staff in a state-of-the-art environment located in Hatfield Business Park. It offers 17 undergraduate degree programmes and 21 postgraduate programmes; there are about 75 research students working at doctoral level. The University of Hertfordshire is the UK's leading business-facing university and an exemplar in the sector. It is one of the region's largest employers with over 2,600 staff and a turnover of almost £235 million. It ranks in the top 4% of all universities in the world according to the Times Higher Education World Rankings and is also one of the top 100 universities in the world under 50 years old.

FITTING EQUATIONS TO DATA WITH THE PERFECT CORRELATION RELATIONSHIP

This version: 23 December 2015

First version: September 2015

Chris Tofallis

Reader/Associate Professor of Decision Science

Hertfordshire Business School

University of Hertfordshire

College Lane

Hatfield

AL10 9AB

United Kingdom

c.tofallis@herts.ac.uk

ABSTRACT

We present a simple method for estimating a single relationship between multiple variables, which are all treated symmetrically i.e. there is no distinction between dependent and independent variables. This is of interest when estimating a law from observations in the natural sciences, although workers in the social sciences may also find this of interest when fitting relationships to data. All variables are assumed to have error but no information about the error is assumed. Unlike other symmetric methods, the weights or coefficients can be obtained easily – indeed, these can be expressed in terms of least squares coefficients. The approach has the important properties of providing a functional relationship which is scale invariant and unique.

Background

We are interested in summarising the relationship between a number of variables by means of a single equation. We wish to treat all variables on the same basis and so we make no distinction between dependent and independent variables. Furthermore, we assume no other knowledge about the data, such as error variances, which would enable more sophisticated methods to be applied, such as those described in Cheng and Van Ness, 1999. In addition, we prefer that the estimation process not be computationally intensive and that the weights or coefficients be directly expressible in simple terms (e.g. established statistics).

We begin with a consideration of ordinary least squares (OLS) regression for the case of two variables, x and y. The linear regression of y on x leads to an equation of the form:

$y = r(s_y / s_x)x + c$, where r is the correlation and s the standard deviation.

For convenience we now assume that the data have been standardized by subtracting the mean and dividing by the standard deviation. This equation then simplifies to $y = rx$. Now consider the regression of x on y; this leads to the relationship $x = ry$, or $y = x/r$.

These two lines do not agree with each other; the only point of agreement between them is where they intersect, which corresponds to the origin for standardized data, and which is the centroid, or point of means, for the original data. This centroid property is retained when fitting planes and hyperplanes in higher dimensions.

If we are trying to describe the relationship in the data, OLS unfortunately provides two inconsistent relations. Suppose we wanted to estimate the slope, here defined as the rate of change of y with x. One application would be to predict the extent of a change: in this case the change in y for a given change in x. The first regression estimates the slope as being equal to r, whereas the second regression estimates this as 1/r (using the same axes). Thus if the correlation were 0.71 one slope estimate would be twice as large as the other! Secondly, we would prefer that the level of noise in the data would not affect our estimation of the underlying relationship. Sadly this is not the case: any measurement error associated with the x-variable will cause the y on x regression slope to be biased downward in magnitude, i.e. closer to zero (Hausman, 2001). This effect is known as regression dilution, or regression attenuation.The slope estimate keeps falling as the noise increases. By contrast, we notice that the reverse OLS slope estimate will *increase* in magnitude as the noise in y rises.

In seeking a *single* equation to summarise the data we are faced with a situation where we have two lines, neither of which is acceptable. Clearly the 'true line', however defined, lies somewhere in between and we need to develop an argument for estimating it. Suppose that the data are affected by noise (measurement error) in the data, and that there is a true underlying linear relationship, possibly due to a physical law or other law of nature. Now imagine turning down the noise (more accurate instruments) so that the scatter is reduced and the correlation rises. The remarkable thing is that the above two OLS lines will gradually converge, and eventually coincide when the correlation is perfect (zero noise). The equation of the resulting unique line will take the form:

$y = \pm x$ (the sign is given by the correlation).

In terms of the original unstandardized data the slope is the ratio of standard deviations:

$\pm(s_y / s_x)$.

Because of the way we have arrived at this line, we shall refer to it as the perfect correlation line. We choose this name because it provides the key to unlocking the fitting process in higher dimensions.

This line is also known as the geometric mean functional relationship, because the slope is the geometric mean of the two OLS slopes. Notice that by taking the geometric mean of the slopes the presence of the correlation in the slope formulae is removed. This would seem to be appropriate for a situation where it is believed that a genuine functional relationship exists. In such cases the correlation would be less than perfect as a consequence of noise or measurement error in the data. But in general we would not expect this to affect the slope or underlying relationship in a systematic way.

Looked at in another way, any researcher who regresses y on x and is later able to obtain more accurate (less noisy) data will always conclude that their previous slope estimate was too low. Whereas any researcher who regresses x on y will later find that their original slope estimate was too high, and that more accurate data provided a lower estimate. It is therefore natural to follow the path of taking some form of mean of these two estimators, which is what the perfect correlation line provides.

This line has a number of attractive properties. Greenall (1949) proved that for a large class of distributions, including the bivariate normal, this line pairs of X and Y values such that the proportion of x-values below X is the same as the proportion of y-values below Y. In other words, it pairs of values with equal percentiles. One application for this is when matching scores on two tests which are designed to measure the same aspect of performance. Another application is to relate measurements from two different instruments which measure the same quantity.

By definition, the OLS line has the least sum of squared residuals in the y-direction. Similarly, no other line has a lower sum of squares in the x-direction than the reverse OLS line. In general, no line can simultaneously achieve both of these optimal properties. Greenall (1949) proved that what we are calling the perfect correlation line increases both sums of squares by the same factor or percentage. This implies symmetry in the errors of estimation.

Treating variables in the same way

When measurement error affects both variables it is possible to derive a maximum likelihood estimate of the slope under the assumption of a bivariate normal distribution. Unfortunately this requires knowledge regarding the variance of the errors. Such information can be obtained if one has the luxury of replicated observations, which is not usually the case. Nevertheless, we can now pose the question: under what circumstance does the above perfect correlation line arise?

Assuming the measurement errors (e_x) are uncorrelated with the observed values (x), then

$x = x_{true} + e_x$

leads to $var(x) = var(x_{true}) + var(e_x)$

Note that since the mean error is zero, $var(e_x)$ is simply the mean square error, and the square root of this is the standard error.

The reliability (coefficient) for x is defined as $var(x_{true})/(var(x)$, i.e. $[var(x) – var(e_x)]/var(x)$

or $1– [var(e_x)/var(x)]$

So that as the error variance tends to zero, the reliability tends to unity.

Now the maximum likelihood estimator for the slope corresponds to the perfect correlation (geometric mean) slope when the ratio of the error variances is equal to the ratio of the total variances (see for example McArdle, 2003):

$var(e_y)/var(e_x) = var(y)/var(x)$

i.e. $var(e_x)/var(x) = var(e_y)/var(y)$

It then becomes clear that the perfect correlation slope is the maximum likelihood slope when both variables are measured with equal reliability.

McArdle (2003) gives practical advice for dealing with the situation when we have no information about error variances. The two OLS lines provide the extreme limits and arise from attaching all the uncertainty to just one of the variables. If the correlation in the data is high then these limits will be close to each other and the uncertainty in the line position will be low. Conversely, if there is much noise in the data then, as expected the range of uncertainty in the slope will be high. How the uncertainty associated with the location of each data point is decomposed between x and y determines how close to these limits the true line is located. If there is no reason to suppose that one variable is measured more reliably than the other, then it would seem reasonable to use the approach we have been discussing.

Aims

The main contribution of this paper will be to extend the perfect correlation or geometric mean approach to multiple variables. Previous attempts have focused on a geometric property of the geometric mean functional relationship in two dimensions. Namely, the fact that the resulting line minimizes the sum of triangular areas formed by the data points and the line (when each point is connected to the line by vertical and horizontal segments). Draper and Yang (1997) generalized this aspect in a particular way. They considered the geometric mean of the distances to the (hyper)plane in each dimension and applied least squares to that quantity. The resulting optimization problem is nonlinear and no expression is available for the coefficients. They proved that in the space of coefficients the result lies in the simplex defined by all the least squares solutions (taking each variable in turn as the dependent variable.

A different generalization was developed by Tofallis (2002a, 2003) which was to view the least areas in the two dimensional case as a minimization of the sum of products of distances to the line in each dimension. Thus in three dimensions the problem becomes one of minimizing the sum of volumes of the tetrahedra created by each data point and the plane. Once again this involved nonlinear optimization. Another variation was the least sum of geometric mean deviations (Tofallis, 2002b).

This leads to a linear objective function with all constraints linear apart from one which sets the product of coefficients to unity.

In the following sections we shall show how to fit a functional relationship without having to solve any nonlinear optimization problems. We assume that no knowledge is available regarding the errors in each of the variables.

Extension to three variables: Fitting a plane

We now describe how to fit a plane to data on three variables, x, y, z in a symmetric fashion. To begin with, imagine the data points are scattered about but still lying perfectly on a plane. If we view a scatterplot of all data points showing just two variables (x,y) we will not see a straight line because the points have different values of z. Hence we cannot impose conditions saying that the bivariate correlations should be perfect, as we did before.

However, if we only plot points with the same z-value we will see evidence of a straight line. This is because we have taken a cross-section through the plane. The way forward is thus to impose the condition that the correlation between x and y for fixed z, be perfect. We denote this partial correlation by $r_{xy.z}$. Naturally, the perfect correlation condition also applies to the other two partial correlations, and this will permit us to deduce the equation of the plane being fitted to the data.

The theory of partial correlation was initially developed by Yule (1907) where he defines this quantity as $r_{xy.z} = (b_{xy}\, b_{yx})^{½}$ (1)

where b_{yx} is the OLS coefficient of x when y is regressed on x, keeping z constant, and b_{xy} is the coefficient of y when x is regressed on y, keeping z constant. Yule used the notation $b_{xy.z}$ to emphasize that it was a partial coefficient i.e. keeping z constant, but we shall follow modern notation where OLS coefficients are always taken to be partials.

In one regression the rate of change of y with x for fixed z is estimated by b_{yx}, whereas in the other it is estimated by $1/b_{xy}$. Since we seek a single estimate we require these two estimates to agree with each other. So we set

$b_{yx} = 1/b_{xy}$ (2)

If we substitute this into (1) it leads to $r_{xy.z} = 1^{½} = \pm 1$. Thus equation (1) demonstrates that the OLS slopes will only agree when there is perfect partial correlation. This provides support for our perfect correlation approach.

If we multiply through (2) by b_{yx} we get $b^2_{yx} = b_{yx}/b_{xy}$ (3)

Denoting the resulting estimate of the 'perfect slope' in the x-y plane by b^*_{yx} gives us

$b^*_{yx} = (b_{yx}/b_{xy})^{½}$ (4)

The sign of this square root should be taken as the sign of either b_{yx} or b_{xy}, which will always be the same sign because the left side of (3) is positive.

Note that the reciprocal of b_{xy} refers to the rate of change of y with x. Thus (4) shows that our perfect correlation 'slope' is the geometric mean of the two OLS slopes. Fortunately, and very conveniently, this is the same connection as in the two dimensional case.

We next show that these 'perfect' regression coefficients can be expressed in terms of simple correlations. Spiegel (1972, p.270) shows that OLS regression coefficients can be written as follows:

$$b_{yx} = (r_{yx} - r_{yz}\, r_{xz})/(1 - r^2_{xz})$$

$$b_{xy} = (r_{yx} - r_{xz}\, r_{yz})/(1 - r^2_{yz})$$

Substituting these into (4) gives:

$$b^*_{yx} = [(1 - r^2_{yz})/(1 - r^2_{xz})]^{½}$$

When taking the square root, the sign of this partial coefficient should match the sign of the associated partial correlation, $r_{yx.z}$

Similarly, the rate of change of y with z is estimated by

$$b^*_{yz} = [(1 - r^2_{yx})/(1 - r^2_{xz})]^{½}$$

Thus, in terms of standardized variables, the equation of the fitted plane can be written very simply as

$$x(1 - r^2_{yz})^{½} \pm y(1 - r^2_{xz})^{½} \pm z(1 - r^2_{yx})^{½} = 0$$

<u>The general case with multiple variables</u>

When there are multiple (*p*) variables Yule (1907) gives the general relationship between partial correlations and OLS coefficients (b):

$$r_{12.3\ldots p} = (b_{12}\, b_{21})^{½}$$

where the left hand side is the correlation between variables x_1 and x_2 when the remaining variables are held constant .

Once again, since we are seeking a single relationship, we require the partial slopes to agree when we regress x_1 on x_2 , and x_2 on x_1 . Thus the rate of change of x_1 with x_2, (b_{12}), will be equal to $1/b_{21}$. Notice that this makes the above partial correlation perfect, just as with the two and three dimensional cases.

Using the same argument which led to (4), we can deduce an expression for this rate of change in terms of OLS coefficients:

$$b^*_{12} = (b_{12}/b_{21})^{½}$$

Once again we obtain the geometric mean of the two associated OLS rates of change. The other b^* coefficients can be found similarly:

$$b^*_{ij} = (b_{ij}/b_{ji})^{½} \qquad (5)$$

This shows that the (partial) rate of change of x_i with x_j is the geometric mean of the estimates of this quantity from the OLS regressions using x_i and x_j as dependent variables. In this sense the property which gives the name to the geometric functional relationship in two dimensions is now seen to persist in higher dimensions.

It is now clear that we can easily fit a single symmetric equation to the data if we have the separate OLS equations. This will often be the most practical route. This development will hopefully allow more widespread application of symmetric fitting to estimate functional relationships.

The fact that the required coefficients are directly expressible in terms of OLS coefficients brings with it the valuable property of uniqueness: provided we have more observations than parameters and assuming there is no degeneracy in the data, then the fitted relationship will be unique.

Coefficients expressed in terms of conditional standard deviations

Kendall and Stuart (1973, p.338) show that an OLS regression coefficient can be expressed as the ratio of a conditional covariance to a conditional variance:

$$b_{ij} = s_{ij.k} / s^2_{j.k}$$

where the suffix notation indicates that all variables apart from those on the left of the point are held fixed.

Applying this to (5) we obtain:

$$b^*_{ij} = \pm (s_{i.k} / s_{j.k})$$

Note that this is a generalization from the bivariate case where we the slope was $\pm(s_y / s_x)$. The only difference is that in three or more dimensions we have to use conditional standard deviations.

Thus in the fitted equation $\Sigma b^*_j x_j = 0$ the coefficients are reciprocals of conditional standard deviations.

Scale invariance

If we re-scale by changing the measurement units of a variable we would expect the associated coefficient to adjust accordingly. For example, if we switch a variable from metres to kilometres then the coefficient of that variable should become a thousand times larger, so that their product stays the same. This valuable feature is present in OLS regression. Since the coefficients in the proposed method can be directly obtained from those in OLS by (5), it follows that the equation relating all the variables will be scale invariant, as required.

Mention should perhaps also be made of orthogonal regression. This is the approach of minimizing the sum of squares of perpendicular distances to the fitted plane. While this can produce a single equation relating all variables, and appears not to treat any variable preferentially, it is well-known that it is not scale invariant (Kermack and Haldane, 1950), and so does not satisfy the above property. It is therefore not an attractive approach unless all variables happen to be measured in the same units. It is also more difficult computationally compared to the method presented here.

Discussion

When fitting equations by treating all variables on the same basis, the name 'symmetric regression' is sometimes used, thus OLS is considered asymmetric. Let us review some of the arguments for using symmetric models. In their chapter on this topic, Von Eye and Schuster (1998) state that "One of the most striking and counterintuitive results from employing two asymmetric regression lines for prediction and inverse prediction is that back-prediction does not carry one back to the point where the predictions originated." (p.214). They provide a numerical example showing predicted values of Performance with IQ as the predictor. These Performance values are then used as the predictor of IQ. The resulting IQ values differ from the starting values. Moreover, they differ systematically: they increase with distance from the mean. The differences also rise with falling correlation.

In his book 'Making social sciences more scientific' (2008), Taagepera includes a chapter entitled 'Why we should shift to symmetric regression'. He argues that OLS models "cannot form a system of interlocking models, because they are not unique, cannot be reversed, and lack transitivity. Scale-independent symmetric regression avoids these problems by offering a single, reversible, and transitive equation." Reversibility requires that if $y = f(x)$ then $x = f^{-1}(y)$, that is, we should be able to rearrange the resulting relationship just as we do with any algebraic equation or scientific law. By contrast OLS models are unidirectional.

He points out that if one were testing a well-established linear law, the OLS slope would always be shallower than the true slope. Moreover, this unwanted artefact would arise whichever variable one plotted on the horizontal axis!

Transitivity is the property where calculating z from x directly gives the same result as the indirect calculation of z from y, and y from x. "This requirement is indispensable if one wants to construct a knowledge system consisting of equations that interlock". "As long as social sciences depend heavily on standard OLS and related unidirectional regression methods, they are bound to face disconnected bits and pieces of relationships, because with OLS, $x \rightarrow y \rightarrow z$ is NOT the same as $x \rightarrow z$" (p.167).

He too proposes the geometric mean functional relationship as the 'only scale-independent symmetric regression' that avoids all of the above difficulties of OLS. As we have seen, this line has slope s_y/s_x, whilst for OLS it is rs_y/s_x which is unfortunately a mixed measure of (lack of) noise and slope. Taagepera illustrates the difficulty by considering the question: How much does weight increase with height? With OLS the answer depends on how accurate the measuring instruments are; not in a random way, but in a systematic way: as the accuracy rises so does the rate of increase! This is obviously not a real effect, but an artefact of conventional least squares models.

He attempts to extend the geometric mean functional relationship beyond two dimensions (Taagepera, 2008, p.165 and 174) proposing that the fitted expression will still only require simple standard deviations as in the two dimensional case:

$\Sigma (\pm x_j / s_i)$ = constant, but this has turned out not be correct (personal communication from Taagepera).

We do not claim that the method discussed in this paper should replace OLS in all situations. However, it is hoped that we have provided a way forward in overcoming the issues raised above by

showing how to fit a compact formula to data on multiple variables, when one wishes to treat each variable in the same way. Moreover, this symmetric functional relationship is both unique and scale-invariant. The parameters in the formula can be estimated directly, notably by making use of OLS formulae. The key result is that the (partial) rates of change ('slopes' in each x_i-x_j plane) are the geometric means of this quantity as estimated by the two OLS regressions using x_i and x_j as dependent variables.

There are many aspects of the method that need exploring. For example, the associated inference theory is yet to be developed, but the bootstrap method can certainly be used as a means of obtaining confidence intervals for the coefficients.

In concluding their paper Draper and Yang (1997) stated that their work "provides a practical solution to a difficult problem". We hope that we have progressed sufficiently to claim to offer both a simple and practical solution to a difficult problem.

REFERENCES

Cheng, C-L, and. Van Ness, JW (1999). *Statistical regression with measurement error*. Arnold, London.

Draper, NR and Yang, YF (1997). Generalization of the geometric mean functional relationship. *Computational Statistics and Data Analysis*, 23, 355-372.

Greenall, PD (1949). The concept of equivalent scores in similar tests. *British Journal of Psychology, Statistical Section*, *2*(1), pp.30-40.

Hausman, J (2001). Mismeasured variables in econometric analysis: problems from the right and problems from the left. *Journal of Economic perspectives*, 57-67.

Kendall, MG and Stuart, A (1973). *The Advanced Theory of Statistics*, volume 2. Charles Griffin and Co, London.

Kermack, KA and Haldane, JBS (1950). Organic correlation and allometry. *Biometrika*, 37, 30-41.

McArdle, BH (2003). Lines, models, and errors: regression in the field. *Limnology and Oceanography*, *48*(3), pp.1363-1366.

Spiegel, MR (1972*). Theory and problems of statistics* (Schaum's Outline Series). McGraw Hill, NY.

Taagepera, R (2008). *Making social sciences more scientific*. Oxford University Press.

Tofallis, C (2002a). Model fitting using the least volume criterion. In *Algorithms for Approximation IV*. Eds. JC Mason and J Levesley. University of Huddersfield Press.

Tofallis, C (2002b). Model fitting for multiple variables by minimising the geometric mean deviation. In *Total least squares and errors-in-variables modeling: algorithms, analysis and applications*. Eds. S.Van Huffel and P.Lemmerling. Kluwer Academic, Dordrecht.

Tofallis, C (2003). Multiple Neutral Data Fitting. *Annals of Operations Research*, 124, 69-79.

Von Eye, A and Schuster, C (1998). *Regression analysis for social sciences*. Academic Press.

Yule, GU (1907). On the theory of correlation for any number of variables, treated by a new system of notation. *Proc. Royal Soc. A*, 79, 182-193.