The Assumptions of the Linear Regression Model

Author(s): Michael A. Poole and Patrick N. O'Farrell

Source: *Transactions of the Institute of British Geographers*, Mar., 1971, No. 52 (Mar., 1971), pp. 145-158

Published by: The Royal Geographical Society (with the Institute of British Geographers)

Stable URL: https://www.jstor.org/stable/621706

The assumptions of the linear regression model

MICHAEL A. POOLE
(*Lecturer in Geography, The Queen's University of Belfast*)

AND PATRICK N. O'FARRELL
(*Research Geographer, Research and Development, Coras Iompair Eireann, Dublin*)

Revised MS received 10 July 1970

ABSTRACT. The paper is prompted by certain apparent deficiences both in the discussion of the regression model in instructional sources for geographers and in the actual empirical application of the model by geographical writers. In the first part of the paper the assumptions of the two regression models, the 'fixed X' and the 'random X', are outlined in detail, and the relative importance of each of the assumptions for the variety of purposes for which regression analysis may be employed is indicated. Where any of the critical assumptions of the model are seriously violated, variations on the basic model must be used and these are reviewed in the second half of the paper.

THE rapid increase in the employment of mathematical models in planning has led R. J. Colenutt to discuss 'some of the problems and errors encountered in building linear models for prediction'.[1] Colenutt rightly points out that the mathematical framework selected for such models 'places severe demands on the model builder because it is associated with a highly restrictive set of assumptions . . . and it is therefore imperative that, if simple linear models are to be used in planning, their limitations should be clearly understood'.[2]

These models have also been widely used in geography, for descriptive and inferential purposes as well as for prediction, and there is abundant evidence that, like their colleagues in planning, many geographers, when employing these models, have not ensured that their data satisfied the appropriate assumptions. Thus many researchers appear to have employed linear models either without verifying a sufficient number of assumptions or else after performing tests which are irrelevant because they relate to one or more assumptions not required by the model. Furthermore, many writers, reporting geographical research, have completely omitted to indicate whether any of the assumptions have been satisfied. This last group is ambiguous, and it is clearly not possible, unless the values of the variables are published, to judge whether the correct set of assumptions has been tested or, indeed, to ascertain whether any such testing has been performed at all.

This problem partially arises from certain shortcomings in material which has been published with the specific objective, at least *inter alia*, of instructing geographers on the use of quantitative techniques. All of these sources make either incomplete or inaccurate specifications of the assumptions underlying the application of linear models, although it is encouraging to note that there has been a considerable improvement in the quality of this literature in recent years. Thus, there were four books and two articles published in the early and mid-1960s which may be classified as belonging to this body of literature,[3] yet, in five of these six sources, only one of the assumptions of the model is mentioned and, even

in the other, only two are referred to. Many of these writers, it is true, discuss regression analysis too briefly to allow space for a comprehensive treatment of the assumptions of the model, but it is unfortunate that none of them did the same as J. B. Cole and C. A. M. King in 1968 who at least warned of the existence of unspecified dangers in the use of regression analysis.[4] Except for this qualification, the work of Cole and King is similar to the earlier volumes, for only one of the model's assumptions is mentioned.[5] However, M. H. Yeates's volume, published in 1968, represents a significant improvement, for three of the assumptions are referred to.[6] This improvement has since continued, for much the most comprehensive coverage of these assumptions presented so far by a geographer has been that of L. J. King, published in 1969: he, in fact, alludes to each one of the model's seven assumptions.[7] Nevertheless, L. J. King's account must be criticized for its unsystematic exposition of the assumptions, for its inaccurate or ambiguous treatment of three of them and for its failure to distinguish basic assumptions from rather less critical ones. Further, he fails to discuss either the task of testing to discover whether the assumptions are satisfied or the problem of finding an alternative method to overcome the difficulty encountered when testing reveals that an assumption is not satisfied.

There is a close parallel between this work of L. J. King, directed towards geographers, and that of Colenutt in the field of planning. Both writers have felt it necessary to warn their colleagues in their respective professions that the correct use of the regression model requires that several critical assumptions be satified. This at least implies that the model has been used carelessly in the past. In fact, King has explicitly pointed out that geographers have tended to employ correlation and regression analysis without showing sufficient awareness of the technical problems involved, prominent among which are the assumptions on which such analysis is based. The similarity between Colenutt's paper and the account of King even extends to the fact that neither of them presents a fully adequate or accurate account of the assumptions of the model, though it should be pointed out that the work of King is, in this respect, superior to that of Colenutt: thus the latter totally ignores two of the most critical assumptions, and his treatment of a third is in error.

Such inadequate treatment of the topic by planners and geographers suggests the need for a concise review of the assumptions of linear models, especially as the elementary statistical texts, such as those of M. J. Moroney and M. R. Spiegel,[8] generally concentrate on outlining the computational procedures and ignore the underlying assumptions. Moreover, even the more advanced and specialized sources are rarely comprehensive in their treatment of these assumptions and their implications; and they tend necessarily, too, to submerge the assumptions in the details of the theory of mathematical statistics.

Therefore, the objective of this paper is to bring together, from many of the less elementary sources, material on two major topics relating to what is probably the most frequently applied of the linear models, the regression equation. The two topics on which attention is focused are:

(1) The fundamental assumptions which must be satisfied if the application of the classical linear regression model is to be totally valid.

(2) The alternative techniques which may be employed when these assumptions are not satisfied in any specific empirical problem; treatment of this second topic necessarily involves the discussion of tests designed to discover the extent to which the assumptions are satisfied, and an indication of how severe a deviation from each assumption may be tolerated without having to resort to the alternative techniques.

Since the aim is to present a concise review of these topics, theoretical proofs are not presented, nor are the computational procedures outlined; however, references to more detailed sources are provided.

THE CLASSICAL LINEAR REGRESSION MODEL

The assumptions of the model

The general single-equation linear regression model, which is the universal set containing simple (two-variable) regression and multiple regression as complementary subsets, may be represented as

$$Y = a + \sum_{i=1}^{k} b_i X_i + u$$

where Y is the dependent variable; $X_1, X_2 \ldots X_i \ldots X_k$ are k independent variables; a and b_i are the regression coefficients, representing the parameters of the model for a specific population; and u is a stochastic disturbance-term which may be interpreted as resulting from the effect of unspecified independent variables and/or a totally random element in the relationship specified.[9]

So far, the discussion has proceeded as if there was only one general regression model. It is important to distinguish two distinct models, each of which is expressed in the form of the equation above. The critical difference between these two models concerns the nature of the independent variables, X_i; in one, the X_i are held constant experimentally at certain fixed values while, in the other, the X_i values are selected at random.[10] Therefore, since Y is random in both models, the 'fixed X' model is characterized by a contrast between X_i and Y, for the former are fixed, while the latter is random; on the other hand, in the 'random X' model, either sets of X_i and Y values are selected at random from a multivariate population, or else pairs of X and Y values are drawn at random from a bivariate population. The implication of this last difference is that, since the concept of correlation is appropriate only for bivariate or multivariate populations, it follows that correlation analysis is valid only in the case of the 'random X' model.[11]

There are several research objectives for which the regression model may be used, but they may be classified into three groups: (i) the computation of point estimates, (ii) the derivation of interval estimates, and (iii) the testing of hypotheses. The assumptions to be satisfied for the proper application of the model vary with the research objective: in particular, the computation of point estimates requires a less restrictive set of assumptions than do the others, and it is therefore proposed to commence by considering such estimates.[12]

Four principal point estimates may be required. First, estimates are generally wanted for a and for the b_i in order to allow the derivation of a regression equation containing specific numerical constants. Secondly, it may be required to predict the expected value of Y corresponding to specific values of X_i; thirdly, a point estimate of the variance of Y must be computed as an intermediate step in the deriving of interval estimates and in the testing of hypotheses; and fourthly, a point estimate of the correlation coefficient, r, may be obtained. It may be shown that the best (minimum variance) linear unbiased estimates of the regression parameters are derived by applying the least-squares method of computation to the sample data to be analysed. Moreover, the least-squares principle allows the derivation both of the best linear unbiased predictor of the expected value of Y for specific values of X_i and also of unbiased estimates of the variance of Y and of r.[13]

However, these results are conditional upon six critical assumptions being satisfied:

(1) Each value of X_i and of Y is observed without measurement error.[14]

Alternatively, this first assumption may be partially relaxed to state that only X_i must be observed without measurement error but, in this case, the interpretation of u must be expanded to include not only the effect of unspecified independent variables and of an essentially random element in the relationship but also the error incurred in measuring Y.[15]

The second assumption is that of linearity.

(2) The relationships between Y and each of the independent variables X_i are linear in the parameters of the specific functional form chosen. Three of the four remaining assumptions relate to the attributes of the disturbance-term, u: the first two of them relate specifically to the nature of the conditional distribution of u (i.e., the set of frequency distributions of u, each corresponding to specific values of X_i) and, therefore, by implication, are both concerned with the conditional distribution of Y.[16]

(3) Each conditional distribution of u has a mean of zero.

(4) The variance of the conditional distribution of u is constant for all such distributions; this is the homoscedasticity assumption.

(5) The values of u are serially independent; thus the values of u are independent of each other and their covariance is accordingly zero. If the fourth assumption is not satisfied in a specific empirical situation, heteroscedasticity is said to be present in the data, while, if the fifth assumption is not satisfied, autocorrelation is said to be present.

It should be noted that the requisite properties of the conditional distribution of the disturbances need hold only for certain specific values in the 'fixed X' model, whereas, in the 'random X' model, these properties must be satisfied for every possible value of X_i.[17]

All of these five assumptions are critical for both simple and multiple regression, but the sixth of the fundamental assumptions is relevant only to the multiple regression model since it is concerned with the relationships between the independent variables.

(6) The independent variables, X_i, are linearly independent of each other. If this assumption is not satisfied in a specific case, multicollinearity is said to be present.[18]

These six assumptions, which are all critical for point estimation in regression analysis, must also all be satisfied if the model is to be used for the purpose either of interval estimation or of hypothesis testing. However, if the regression model is to be employed for such inferential purposes, then these six assumptions are not sufficient for the valid application of the model, for one further assumption is needed. The precise form of this further assumption differs according to whether the model being employed is of the 'fixed X' or 'random X' type:

(7) The fixed X model requires that the conditional distribution of the disturbance-term must be normal in form, which clearly implies that the dependent variable, Y, has a normal conditional distribution. **The random X model requires that both the conditional and marginal distributions of each variable are normal:** this model thus requires not only conditional normality for Y, but also for X_i, and, in addition, the overall frequency distribution of each variable must be normal.

It may be added that, in relation to the calculation of point estimates, the assumption of the normal conditional distribution of the disturbance-term would allow the derivation of maximum likelihood estimates of the regression coefficients. However, these, in fact, turn out to be identical to the least-squares estimates.[19] Therefore, neither form of the normality assumption is necessary for point estimation.[20] The reason that it is necessary, on the other

hand, for inferential problems is that the two statistics commonly used, Student's t and Snedecor's F, both require that the data be normally distributed. Even in relation to inferential problems, however, this assumption is not binding, provided that the sample is very large. Interval estimates are most often made when computing confidence intervals for the individual regression coefficients, a and b_i, and when calculating interval estimates for predicted values of Y corresponding to specific values of X_i: in both these cases Student's t is used. Sometimes, however, the aim is to establish confidence regions for both or all regression coefficients simultaneously and, in this case, Snedecor's F is used. The most frequently performed tests relate to hypotheses on the value of individual regression coefficients and on the values of the entire set of such coefficients: when the hypothesized values are zero, this second test is equivalent to testing the significance of the regression model as a whole. For the first of these tests, either the t or F statistics may be used, but the second test requires Snedecor's F.[21]

In the case of the random X model, inferential analysis may also be carried out upon the correlation coefficient. Such inference includes the establishment of confidence intervals about r by means either of David's tables, based on the density function of r, or of Fisher's z transformation and the distribution of Student's t. It also includes both the test of the hypothesis that r equals zero, using Student's t, and also the test of the hypothesis that r equals some specific value other than zero, for which either David's tables or else Student's t with Fisher's z transformation may be employed. Such tests and estimation procedures are applicable not only to simple and multiple correlation coefficients but also to partial correlation coefficients.[22]

The geographical literature on the assumptions of the regression model

It is claimed that the justification for this paper is the inadequate attention given to the assumptions of the regression model in the geographical literature. Except for a passing mention given to those books written by geographers which purport to provide instruction on the use of regression analysis, no evidence has yet been presented to demonstrate the inadequacy of the attention given to the assumptions of the model. Therefore, now that each assumption has been described, it is proposed to examine briefly the extent to which each one has been alluded to by geographers when reporting specific applications of regression analysis. In the course of this examination, the comments made in the introduction about the treatment of the assumptions in the literature purporting to give instruction to geographers and planners on the use of the regression model will also be elaborated.

It is rare to find explicit reference to the assumption that there are no measurement errors in the data, though it might be argued that the need for accurate data is so obvious that it is taken for granted: A. H. Robinson and R. A. Bryson are among the few geographers who have referred explicitly to the problem.[23] Clearly, it is difficult to test for measurement error, even though A. R. Hill has conducted an experiment to isolate operator variability in pebble-orientation analysis.[24] However, there appears to be little awareness in the geographical literature of the fact that the presence of measurement error in the independent variables is a much more serious problem than its presence in the dependent variable. The exception to this statement is found in the book by L. J. King, for he does refer quite correctly to this measurement error assumption.[25]

The linearity assumption is the only one which is mentioned by every single geographer giving instruction on the use of the regression model.[26] It is therefore no surprise that many

geographers, when using regression analysis, have been conscious of the need for the relationships investigated to be linear if the linear model is to be fitted: H. G. Kariel, and also J. F. Hart and N. E. Salisbury, are examples of geographers who have actually tested for the presence of linearity.[27]

The assumption relating to the nature of the disturbance-term has been given much less attention in the geographical literature than has the linearity assumption. The assumption that each conditional distribution has a mean of zero appears to have been almost totally ignored, and the homoscedasticity assumption has not been given much more attention. B. J. L. Berry and H. G. Barnum have performed a logarithmic transformation in order to ensure greater homoscedasticity,[28] but there are few other explicit references in the geographical literature to testing for homoscedasticity, and it seems reasonable to conclude that most geographers have not verified this assumption of the model.

The third of the assumptions relating to the disturbance-term is that there should be no autocorrelation. Almost all the discussion of this problem in the literature of econometrics and statistics has dealt with the presence of autocorrelation in time-series data. Clearly, however, since geographical analysis is more concerned with spatial variation than with temporal variation, much of the data subjected to regression analysis by geographers refer to cross-sections through time, so the problem of time-series autocorrelation does not arise. But the correlation between values corresponding to successive time-periods, which is such a common feature of time-series data, has its parallel in the analysis of spatial variation at a cross-section through time, for there is frequently correlation between the values of the disturbance-term corresponding to contiguous spatial units. This problem of spatial autocorrelation is even more complex than the temporal autocorrelation problem, because there is more than one dimension involved in the spatial case. For a long time the only contribution on the spatial autocorrelation problem was that of the statistician R. C. Geary,[29] but some geographers have recently become aware of the problem: L. Curry has alluded to it, and M. F. Dacey has derived methods to test for the presence of spatial autocorrelation in data measured on a nominal scale.[30] However, it still appears to be rare for geographers using regression analysis to test for spatial autocorrelation.

Of all the geographers who have provided instruction on the use of regression analysis in their books, L. J. King is the only one who has mentioned any of these three assumptions relating to the characteristics of the disturbance-term.[31] He refers to each of the three although, in stating the assumptions that the disturbances have a mean of zero, he fails to state that this relates only to the conditional distribution. This is in contrast to his treatment of the homoscedasticity assumption, which follows immediately after his reference to the zero mean assumption, for he does make it clear that homoscedasticity implies equal variance for all conditional distributions.

The last of the six basic assumptions of the regression model, the absence of multicollinearity, has been recognized by many geographers; and E. N. Thomas, R. A. Mitchell and D. A. Blome and Shue Tuck Wong are examples of writers who have examined their correlation matrix for the presence of multicollinearity.[32] Moreover, of the instructional sources, two of them, the books by Yeates and L. J. King, mention this assumption.[33] Multicollinearity, in fact, is probably second only to linearity among the six basic assumptions in the frequency with which it is alluded to by geographers using regression analysis. The remaining four of these six assumptions have been referred to much less frequently.

The most curious feature, however, of the geographical literature on the assumptions

of the regression model is that neither the presence of linearity, nor the absence of multicollinearity, nor any of the other basic assumptions of the regression model has been alluded to as frequently by geographers as has the property of normality. It is the presence of normality which has been most often stated by geographers to be a critical assumption of the model and it is normality whose presence has been most often tested for. Yet, as we have seen, this property is relevant only for interval estimation and hypothesis testing: it is not one of the six basic assumptions necessary for the initial point estimation.[34]

Even more important, geographers, when testing distributions prior to using regression analysis, appear almost invariably to have examined the marginal distribution and to have ignored the conditional distribution. True, this has seldom been made absolutely explicit, for the term 'marginal distribution' never seems to have been employed. However, when geographers, reporting the use of regression analysis, write of testing 'the distribution of the individual variables'[35] or of 'normalizing the data by means of log transformation',[36] it seems likely that they are referring implicitly to the marginal distribution of the variables concerned. Yet, at least in the case of the fixed X model, the form of the marginal distribution is totally irrelevant. Clearly, in most instances in geographical research, it is, in fact, the random X model which is the appropriate one to employ, and this, of course, does require that the marginal distribution be normal. However, a normal marginal distribution is not alone sufficient, for it is essential that the conditional distribution also be normal.

Those sources purporting to provide instruction on the use of regression analysis are no better, for McCarty and Lindberg and also Yeates fail to state, or even imply, that the conditional distribution should be normal, and L. J. King, in referring to the normality assumption, does not make it clear whether he is referring to the marginal or the conditional distribution of the disturbance-term.[37] Moreover, none of these sources points out that the normality assumption is relevant only for interval estimation and hypothesis testing: true, King does state that the assumptions of the regression model are of varying importance, depending on whether or not the work has an inferential purpose, but he fails to state which specific assumptions this statement refers to.[38]

Before concluding this discussion of the extent to which geographers have shown an awareness of the assumptions of the regression model, detailed mention should be made of the paper written on this model by Colenutt for, although directed primarily towards planners, it is published in a source familiar to many geographers. Of the six basic assumptions of the model, four are alluded to by Colenutt, but he omits to state that the conditional distributions of the disturbance-term should each have a mean of zero and a constant variance. Moreover, in relation to the normality assumption, he makes the same error that so many geographers commit by omitting to say that the conditional distribution of the variables should be normal.[39]

Thus, although L. J. King and Colenutt have provided much better instruction on the assumptions necessary for the valid use of regression analysis than have previous writers in the disciplines of geography and planning, there are deficiencies even in their accounts. Since the geographical and planning sources on the use of the model thus exhibit inadequacies and since the record of application of the model in these disciplines has been rather unsatisfactory, there appears to be some need for a paper whose objectives are to state clearly to geographers the assumptions of the regression model, to indicate how these assumptions may be tested for and to describe alternative models which may be used when certain assumptions are not satisfied.

ALTERNATIVES TO THE CLASSICAL LINEAR REGRESSION MODEL

Since the application of classical linear regression analysis, to be totally valid, requires that so many assumptions are satisfied, it follows that the testing of these assumptions is a critical part of any such analysis. Yet the geographical literature on regression analysis contains few detailed references to such testing: thus the only instructional sources including such references are the volumes by L. J. King and by Cole and C. A. M. King. Both of them contain descriptions of testing for normality, though not in relation to regression analysis specifically,[40] and L. J. King also describes the tests for autocorrelation devised by Geary and Dacey.[41] Reports on specific empirical applications of regression analysis by geographers often simply state that some assumption has been tested for, without indicating the method used. It is clear, however, from those reports which do specify the actual test used, that geographers have made little use of statistical inference procedures when performing such tests: in fact, most of the assumption-testing which has been done has consisted of the visual inspection of graphs, such as the fractile diagrams used by P. D. La Valle to test for normality and the scattergrams employed by Kariel to test for linearity.[42]

If testing reveals that a particular assumption of the classical regression model is not satisfied, then some alternative to the straightforward application of this classical model must be resorted to. Those alternative methods, which still basically involve regression analysis, are of two main types: either the input data may be tranformed, most frequently by applying a logarithmic, reciprocal, power or arcsin transformation,[43] or else a variation on the classical regression model may be applied. Geographers have frequently used the first of these methods, transformation, though almost entirely in order to satisfy the normality and linearity assumptions. Thus almost all the instructional sources refer to the possibility of transforming the data to achieve linearity and D. S. Knos is an example of a geographer who has done this when working on a specific empirical problem.[44] Far fewer of the instructional sources allude to transformation as a way to achieve normality, but many geographers, such as G. Olsson and A. Persson and La Valle, have, in fact, done this in empirical work.[45] Berry and Barnum, in contrast, are two of the few geographers who have transformed data in order to ensure greater homoscedasticity.[46] The second of the alternative ways of satisfying the assumptions of the model, by using a variant on classical regression analysis, appears to have been almost totally ignored by geographers; among the few references of this type are the allusions by P. Haggett and R. J. Chorley to the use of polynomials when the linearity assumption cannot be satisfied, even by transformation.[47] In this paper, on the other hand, a brief outline will be given of both types of method and of the circumstances in which they may be used. The topic will be approached by discussing each of the seven assumptions in turn.

Assumptions on measurement error and linearity

The incurring of measurement error in the observation of the independent variables, X_i, would lead to biased estimates of the regression coefficients if the classical regression model were used.[48] However, any test of the degree of measurement error is clearly made difficult by the fact that the amount of error is unknown. In fact, for most purposes, it is frequently assumed, at least in econometrics, that measurement error is much less significant than errors resulting from incorrect equation specification, so the former is generally ignored.[49]

The problem of measurement error can also be ignored if the sole objective of the regression analysis is to predict the value of Y corresponding to a given set of X_i values.[50]

Prediction, however, is rarely the sole reason for performing regression analysis, and it sometimes happens that serious measurement error is suspected in the data, so that some other method than classical least-squares must be adopted. Since the mathematics of some of these methods are complex, most writers discuss them in relation only to simple regression[51] and this convention will be followed here. There is quite an easy method which is applicable when measurement error occurs in the independent variable, X, in simple regression, but Y is observed without error. The solution in this case is to reverse the roles of X and Y by using the former as the dependent variable and the latter as the independent variable: having made this adjustment, ordinary least-squares estimation is then valid, provided, of course, that the other assumptions of the model are satisfied.[52]

It is when serious measurement error occurs in the variables on both sides of the equation that more elaborate methods are required. The first such method is to assume that the measurement errors are serially independent and normally distributed and to use estimates of the error variances as weights in a modified application of the least-squares model.[53] The second method is to rank and group the data and, assuming that the errors are serially independent but not necessarily normally distributed, to derive the regression coefficients by manipulating the subgroup means for the respective variables.[54] The third and final method involves the use of instrumental variables, which are independent of the errors and highly correlated with the true values of the variables, and the manipulation of the deviations of individual values from mean values for both the original variables and the instrumental variables.[55]

Turning from measurement error to equation specification error, it is now necessary to consider tests relating to the linearity assumption of the regression model and to indicate the procedures available when this assumption is significantly violated. Testing for the linearity of a relationship may take one of three forms. Either, having fitted a high-order polynomial function, the regression coefficients for the terms of second or higher order may be tested for significant departures from zero; or, after stratifying the data on the basis of the X values, a regression equation may be calculated for each stratum and the significance of the differences between each of the slope coefficients may be tested; or else the sequence of residuals, arranged in order of increasing X, may be tested for randomness.[56]

If the application of any of these tests suggests that the linearity assumption is not satisfied in a specific instance, the input data are generally transformed to yield new data which satisfy this assumption more closely: ordinary linear regression can then be applied to these transformed data.[57] Alternatively, either the attempt may be made to fit a higher-order polynomial function to the original data which appear to be linked in a curvilinear relationship,[58] or else one of several iterative methods of estimating the parameters of other non-linear functions may be used.[59]

Assumptions on the pattern of disturbances

It is impossible to test directly, in any specific empirical example, the validity of the four assumptions relating to characteristics of the disturbances, for these characteristics are unknown because the disturbances are unobservable. However, tests may be carried out on the pattern of the residuals, using this as an estimate of the pattern of disturbances.[60]

The first of these assumptions relating to the nature of the disturbances is that the mean disturbance is zero for each value of X_i. In practice, the principal point is that the residual mean, $\bar{e}$, should be independent of X_i. However, the bias introduced into the model when this assumption is not satisfied is small, provided that the residual variance is small.[61] Testing thus involves measuring both the correlation between $\bar{e}$ and X_i and also the variance of $\bar{e}$. If such testing reveals, however, that the assumption is so poorly satisfied that a considerable bias is introduced into the estimation of the regression coefficients, then the form of the specification of the relationship between X_i and Y should be changed and an alternative equation used.

The homoscedasticity assumption states that the conditional disturbance distribution should have a variance which is constant for all X_i values but, again, the major requirement in practice is that the residual variance should be independent of X_i. In fact, if the variance of $\bar{e}$ is not constant, but is independent of X_i, the estimates of the regression coefficients are still unbiased, though the usual methods of statistical inference are invalid. However, if the variance of $\bar{e}$ is not only subject to variation, but is correlated with X_i, then the estimates of the regression coefficients are seriously biased, and valid inference is also impossible.[62]

No precise test of homoscedasticity is possible, because the tests available, such as those of Hartley or Bartlett, are highly sensitive to non-normality in the data.[63] However, if there does appear to be a correlation between X_i and the variance of $\bar{e}$, then either the input data may be transformed in order to try to reduce or eliminate the heteroscedasticity,[64] or else a modified form of the regression model, weighted regression, may be used. In this modified regression model, weights, which are proportional to the variance of $\bar{e}$, are applied to the variables; a frequent special case, used when the variance of $\bar{e}$ is proportional to X_i, arises when the ratio Y/X_i is used as the dependent variable instead of Y itself and the reciprocal of X_i is the independent variable.[65]

The third assumption states that the errors are serially independent. It may be shown that, although the presence of autocorrelated disturbances does not prevent the derivation of an unbiased estimate of the regression coefficients, it does lead to two serious consequences, especially if the autocorrelation coefficient is high: first, the estimates of the regression coefficients have an unduly large and inaccurately estimated variance, and, secondly, the procedures for statistical inference are inapplicable.[66]

The presence of autocorrelation in one-dimensional data, such as the values corresponding to a time series or to a cross-section through space, may be tested for by means of the Durbin-Watson d statistic: specifically, this tests for the existence of dependence between successive residuals, arrayed in order of temporal or spatial sequence and derived by the application of ordinary least-squares methods.[67] Testing for the presence of autocorrelation in the case of two-dimensional spatial data is more difficult, but a variant on the Durbin-Watson d statistic, called the contiguity ratio, was developed by Geary: essentially the use of this ratio tests for the similarity of the residuals corresponding to contiguous spatial units.[68] Dacey, on the other hand, in a suggested alternative way of testing for autocorrelation in two-dimensional data, proposes an extension of the conventional one-dimensional runs test: his method is to reduce the residual-term to a binary variable by distinguishing only between positive and negative residuals and then to test whether there is a significant tendency for contiguous areas to have residuals of the same sign.[69]

If testing reveals that a set of data is autocorrelated, then two types of solution are available. First, since the autocorrelation probably results either from an error in the linearity

specification or from measurement error or from the effect of a variable excluded from the model, the attempt may be made to eliminate it by transforming the data or by introducing further independent variables into the model and then using ordinary least-squares methods.[70] Secondly, one of the several more complex methods available for the computation of the regression coefficients may be used, though their application is generally made difficult by the fact that an estimation of, or assumptions about, the form of the autocorrelation function must be made.[71]

The last of the assumptions relating to the conditional disturbance distributions is that these distributions should be normal, but, even when the intention is to perform significance tests and establish confidence intervals, this assumption may frequently be relaxed. This is because such statistical inference procedures are not particularly sensitive to departures from normality: if the disturbances are non-normally distributed, the tests and intervals are still approximately correct and, indeed, if the sample is large, the approximations are extremely good.[72] On the other hand, if the sample is small, it is very difficult to test for normality.

If the assumption of normality does appear to be seriously violated, the data may be transformed to derive more normal conditional disturbance distributions. However, the robustness of regression analysis with respect to the assumption of normality and the fact that there is a greater need to satisfy such other assumptions as homoscedasticity and linearity, together have the result that transformations specifically for the purpose of imposing normality are infrequent.[73]

The assumption of the absence of multicollinearity

The last of the assumptions of the classical linear regression model is that the independent variables, X_i, are linearly independent of each other. If this assumption is not satisfied and the independent variables are thus multicollinear, the result is that the individual regression coefficients for each variable are not identifiable: in fact, the closer the linear correlation between the independent variables, the less the certainty with which these coefficients may be identified. This imprecision in the estimate of the regression coefficients is generally revealed by the occurrence of high standard errors. However, if the data contain measurement error, it can happen that standard errors are low despite the presence of multicollinearity, and, in this case, confluence analysis (bunch-map analysis) may be necessary to reveal the existence of the multicollinearity.[74]

Because multicollinearity makes the regression coefficients quite unidentifiable, it is important, if the aim is to estimate the regression equation, to reduce it as much as possible. Either further data may be sought,[75] or certain variables may be omitted from the model. If the latter solution is adopted, however, care must be taken in interpreting the resulting equation, for it cannot be assumed that an omitted variable has no effect: it is simply that its separate effect could not be isolated.[76] It may be added, however, that, if the purpose of the regression analysis is only to predict the value of Y corresponding to a set of X_i values, then multicollinearity is not a serious problem, provided that the intercorrelations continue unchanged into the future.[77]

CONCLUSION

This paper has attempted to summarize the major properties and assumptions of the linear regression model, and has reviewed and commented upon the shortcomings revealed by

some geographers in employing this model. In the case of each of the seven assumptions of the least-squares regression model, methods have been developed to overcome the problems presented when these assumptions are not satisfied in specific empirical situations. However, when alternative models are proposed as the solution, the model developed to overcome any one problem often cannot simultaneously handle other problems too, because it is highly dependent upon the other assumptions being satisfied:[78] thus one of the methods for overcoming the problem of measurement error depends upon the assumption that these errors are not autocorrelated and have a normal conditional distribution with zero mean and constant variance.[79] On the other hand, in the case of data transformations, it frequently happens in practice that a transformation which is designed to overcome the problems arising when one of the assumptions is not satisfied, simultaneously solves problems relating to other assumptions.[80]

In addition to indicating methods of overcoming these problems when the assumptions of the simple model are not satisfied, it has also been shown that the assumptions vary considerably in their significance. They vary both according to the purpose for which the model is to be used, especially in relation to whether or not any significance testing or derivation of confidence limits is to be done, and according to whether the purpose of the analysis is explanation or prediction: in the case of the latter, it is not essential to satisfy the assumptions of measurement error or multicollinearity. The assumptions also vary in the degree to which they are robust for any particular purpose. In general, however, it may be concluded that the normality, measurement error and zero disturbance-mean assumptions may be given less attention than is necessary in the case of the other four assumptions: it is of paramount importance that the relationships between variables be linear, that the disturbances be homoscedastic and serially independent and, if multiple regression is being performed, that the independent variables are not linearly correlated.

ACKNOWLEDGEMENTS

The authors are indebted to Mr S. B. Essig, Lecturer in Economics, The Queen's University of Belfast, for commenting upon an earlier draft of this paper.

NOTES

1. R. J. COLENUTT, 'Building linear predictive models for urban planning', *Reg. Stud.* 2 (1968), 140
2. Ibid., 140
3. S. GREGORY, *Statistical methods and the geographer* (1963), 167–208; R. G. BARRY, 'An introduction to numerical and mechanical techniques' in F. J. MONKHOUSE and H. R. WILKINSON, *Maps and diagrams: their compilation and construction* (1963), 415–21; P. HAGGETT, *Locational analysis in human geography* (1965), 293–9; R. J. CHORLEY, 'The application of statistical methods to geomorphology' in G. H. DURY (ed.), *Essays in geomorphology* (1966), 340–56, 370–77; C. A. M. KING, *Techniques in geomorphology* (1966), 312–23; H. H. MCCARTY and J. B. LINDBERG, *A preface to economic geography* (New Jersey, 1966), 71–81
4. J. B. COLE and C. A. M. KING, *Quantitative geography: techniques and theories in geography* (1968), 263
5. Ibid., 138–46, 150–3, 287–94
6. M. H. YEATES, *An introduction to quantitative analysis in economic geography* (1968), 15–21, 81–2, 50–3, 100–6
7. L. J. KING, *Statistical analysis in geography* (New Jersey, 1969), 117–64
8. M. J. MORONEY, *Facts from figures* (1956), 276–320; M. R. SPIEGEL, *Theory and problems of statistics* (1961), 217–82
9. J. JOHNSTON, *Econometric methods* (1963), 5–6; F. A. GRAYBILL, *An introduction to linear statistical models*, vol. 1 (1961), 99–104

10. It was with reference to the 'random X' model that the term 'regression' was originally used (G. SNEDECOR, *Statistical methods* (Ames, Iowa, 1956), 152–3), and some writers still restrict the use of the term in this way (F. A. GRAYBILL, op. cit., 101). However, it has become common to refer to the 'fixed X' model, too, as a regression model (N. R. DRAPER and H. SMITH, *Applied regression analysis* (1966), 6)

11. F. S. ACTON, *Analysis of straight-line data* (1959), 7; F. A. GRAYBILL, op. cit., 206–7

12. F. A. GRAYBILL, op. cit., 109–10

13. J. JOHNSTON, op. cit., 9–20, 34–6, 108–13; F. A. GRAYBILL, op. cit., 114–17; E. MALINVAUD, *Statistical methods of econometrics* (Amsterdam, 1966), 78–81, 84–6, 97–9

14. E. MALINVAUD, op. cit., 75

15. F. S. ACTON, op. cit. 8; F. A. GRAYBILL, op. cit., 103–4

16. J. JOHNSTON, op. cit., 7–9, 106–8; E. MALINVAUD, op. cit., 73–86, 98–9, 173–4; N. R. DRAPER and H. SMITH, op. cit., 17; F. A. GRAYBILL, op. cit., 108–9, 114–17

17. J. JOHNSTON, op. cit., 25–9, 133; F. A. GRAYBILL, op. cit., 204–6

18. J. JOHNSTON, op. cit., 107–8; E. MALINVAUD, op. cit., 174–6

19. J. JOHNSTON, op. cit., 20–21, 115–6; E. MALINVAUD, op. cit., 86–9; F. A. GRAYBILL, op. cit., 110–14

20. J. JOHNSTON, op. cit., 21–5, 116–27; E. MALINVAUD, op. cit., 77, 89–92; N. R. DRAPER and H. SMITH, op. cit., 59; G. E. V. LESER, *Econometric techniques and problems* (1966), 9

21. J. JOHNSTON, op. cit. 23–5, 36–7, 118–33; E. MALINVAUD, op. cit., 91–2, 99–100, 199–205; F. A. GRAYBILL, op. cit., 120–45; N. R. DRAPER and H. SMITH, op. cit., 18–26, 63–7; F. S. ACTON, op. cit., 23–53

22. F. A. GRAYBILL, op. cit., 208–16

23. A. H. ROBINSON and R. A. BRYSON, 'A method for describing quantitatively the correspondence of geographical distributions', *Ann. Ass. Am. Geogr.* 47 (1957), 388

24. A. R. HILL, 'An experimental test for the field technique of till macrofabric analysis', *Trans. Inst. Br. Geogr.* 45 (1968), 93–105

25. L. J. KING, op. cit., 122–3

26. S. GREGORY, op. cit., 203; R. G. BARRY, op. cit., 417–18; P. HAGGETT, op. cit., 294–6; R. J. CHORLEY, op. cit., 341, 371, 374; C. A. M. KING, op. cit., 312; H. H. MCCARTY and J. B. LINDBERG, op. cit., 71–2; J. B. COLE and C. A. M. KING, op. cit., 138; M. H. YEATES, op. cit., 15; L. J. KING, op. cit., 120

27. H. G. KARIEL, 'Selected factors areally associated with population growth due to net migration', *Ann. Ass. Am. Geogr.* 53 (1963), 215; J. F. HART and N. E. SALISBURY, 'Population changes in Middle Western villages: a statistical approach', *Ann. Ass. Am. Geogr.* 55 (1965), 151–2

28. B. J. L. BERRY and H. G. BARNUM, 'Aggregate relations and elemental components of central place systems', *J. reg. Sci.* 4 (1962), 36

29. R. C. GEARY, 'The contiguity ratio and statistical mapping', *Inc. Statist.* 5 (1954), 115–41

30. L. CURRY, 'Quantitative geography, 1967', *Canad. Geogr.* 11 (1967), 268–73; M. F. DACEY, 'A review on measures of contiguity for two and k-color maps', *Tech. Rep. 2* (Spatial Diffusion Study, Dept. of Geography, Northwestern University, Evanston, Illinois, 1965)

31. L. J. KING, op. cit., 121–3, 157–162

32. E. N. THOMAS, R. A. MITCHELL and D. A. BLOME, 'The spatial behavior of a dispersed non-farm population', *Pap. reg. Sci. Ass.* 9 (1962), 125–6; SHUE TUCK WONG, 'A multivariate statistical model for predicting mean annual flood in New England', *Ann. Ass. Am. Geogr.* 53 (1963), 299

33. YEATES, op. cit., 81; L. J. KING, op. cit., 162–3

34. Two examples of papers, in which variables have been transformed in order to achieve normality, despite the use of the entire population of data, are: H. ALDSKOGIUS, 'Vacation house settlement in the Siljan region', *Geogr. Annlr* 49 B (1967), 78; and G. OLSSON and A. PERSSON, 'The spacing of central places in Sweden', *Pap. reg. Sci. Ass.* 12 (1964) 90

35. M. H. YEATES, 'Some factors affecting the spatial distribution of Chicago land values, 1910–1960', *Econ. Geogr.* 41 (1965), 63

36. E. J. TAFFE, R. L. MORRILL and P. R. GOULD, 'Transport expansion in underdeveloped countries: a comparative analysis', *Geogrl Rev.* 53 (1963), 516

37. H. H. MCCARTY and J. B. LINDBERG, op. cit., 72; M. H. YEATES, *An introduction to quantitative analysis*, 20; L. J. KING, op. cit., 121–3

38. L. J. KING, op. cit., 123

39. R. J. COLENUTT, op. cit., 140–1

40. L. J. KING, op. cit., 82; J. B. COLE and C. A. M. KING, op. cit., 130–1

41. L. J. KING, op. cit., 158–60

42. P. D. LA VALLE, 'Some aspects of linear karst depression development in southcentral Kentucky', *Ann. Ass. Am. Geogr.* 57 (1967), 61; H. G. KARIEL, op. cit., 215

43. N. R. DRAPER and H. SMITH, op. cit., 131–4; F. S. ACTON, op. cit., 221–3; J. JOHNSTON, op. cit., 47–50

44. D. S. KNOS, 'The distribution of land values in Topeka, Kansas' in B. J. L. BARRY and D. F. MARBLE (eds.) *Spatial analysis: a reader in statistical geography* (New Jersey, 1968), 271–5

45. G. Olsson and A. Persson, op. cit., 96; P. D. La Valle, op. cit., 61
46. B. J. L. Berry and H. G. Barnum, op. cit., 36
47. P. Haggett, op. cit., 296; R. J. Chorley, op. cit., 347–8
48. J. Johnston, op. cit., 148–50; E. Malinvaud, op. cit., 331–3
49. E. Malinvaud, op. cit., 362
50. J. Johnston, op. cit., 162–4
51. E. Malinvaud, op. cit., 362
52. C. E. V. Leser, op. cit., 18
53. J. Johnston, op. cit., 150–62; E. Malinvaud, op. cit., 335–47
54. J. Johnston, op. cit., 164–5; E. Malinvaud, op. cit., 359–62
55. J. Johnston, op. cit., 165–6; E. Malinvaud, op. cit., 347–52
56. E. Malinvaud, op. cit., 268–71
57. J. Johnston, op. cit., 44–50; N. R. Draper and H. Smith, op. cit., 131–4
58. N. R. Draper and H. Smith, op. cit., 129–30, 150–55; G. Snedecor, op. cit., 452–4, 461–71; F. A. Graybill, op. cit., 165–84; F. S. Acton, op. cit., 193–218
59. N. R. Draper and H. Smith, op. cit., 267–85; E. Malinvaud, op. cit., 290–314
60. N. R. Draper and H. Smith, op. cit., 86–97; C. E. V. Leser, op. cit., 14–15
61. E. Malinvaud, op. cit., 258–60
62. Ibid., 254–7
63. C. E. V. Leser, op. cit., 13; F. S. Acton, op. cit., 89–90
64. F. S. Acton, op. cit., 90, 219
65. J. Johnston, op. cit., 208–11; E. Malinvaud, op. cit., 257–58; C. E. V. Leser, op. cit., 13–14
66. J. Johnston, op. cit., 179, 187–92; E. Malinvaud, op. cit., 433–9
67. J. Johnston, op. cit., 192; E. Malinvaud, op. cit., 421–5
68. R. C. Geary, op. cit., 115–41
69. M. F. Dacey, op. cit.
70. C. E. V. Leser, op. cit., 17
71. J. Johnston, op. cit., 179–87, 192–5; E. Malinvaud, op. cit., 439–45
72. E. Malinvaud, op. cit., 93
73. F. S. Acton, op. cit., 220
74. J. Johnston, op. cit., 201–7; C. E. V. Leser, op. cit., 27
75. J. Johnston, op. cit., 207
76. C. E. V. Leser, op. cit., 28
77. J. Johnston, op. cit., 207; C. E. V. Leser, op. cit., 28
78. J. Johnston, op. cit., 147
79. E. Malinvaud, op. cit., 329
80. F. S. Acton, op. cit., 221

Résumé. *Les hypothèses du modèle de régression linéaire.* Ce qui a inspiré cet exposé, c'est qu'il semble y avoir quelques insuffisances tant dans la discussion du modèle de régression dans les instructions fournies pour les géographes que dans l'application empirique elle-même du modèle par les écrivains de la géographie. Dans la première partie de l'exposé, les hypothèses des deux modèles de régression, le « X fixe » et le « X pris au hasard », sont indiquées en détail, et l'on a aussi indiqué l'importance relative de chacune des hypothèses dans tous les usages où l'on pourrait employer l'analyse de régression. Dans le cas où quelques-unes des hypothèses critiques du modèle portent gravement à faux, il faut employer des variations du modèle fondamental, et ces variations sont examinées dans la deuxième partie de l'exposé.

Zusammenfassung. *Die Annahmen des lineares Regressionsmodells.* Die Abhandlung beschäftigt sich mit scheinbaren Unzulänglichkeiten sowohl in der Besprechung des Regressionsmodells als auch in Lehrmaterial für Geographen und in der tatsächlichen erfahrungsmässigen Anwendung des Modells durch geographische Schriftsteller. Im ersten Teil der Abhandlung sind die Annahmen der zwei Regressionsmodelle, das ‚festgelegte X' und das ‚wahllose X' in Einzelheiten umrissen und die relative Wichtigkeit von jeder dieser beiden Annahmen, für die verschiedenen Möglichkeiten auf welche die Regressionsanalyse angewandt werden kann, ist angedeutet. Wo eine der kritischen Annahmen des Modells ernstlich übertreten wird, müssen Variationen des Ausgangsmodells benutzt werden und diese werden in der zweiten Hälfte der Abhandlung besprochen.