NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

ISBN 0-315-53224-6

Canada

THE GEOMETRY OF REGRESSION ANALYSIS

by

Ian Campbell

(218804)

Thesis presented to the Department of Economics at the University of Ottawa for the obtainment of a Master's degree.

Thesis director: Professor Camilo Dagum

Ottawa, Ontario

October, 1989

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Table of Contents

SUMMARY

This paper analyzes geometric aspects of single-equation regression techniques as they occur in the space of n x k matrices.

It shows that all matrices fitting onto an error-free regression equation belong to a smooth, connected, non-linear sub-set of the space of n x k matrices. This sub-set is shown to be of dimension $(n+1)(k-1)$. The paper examines how observations, errors and fitted values act in and around this set under various assumptions, such as ordinary least squares, orthogonal regression, generalized least squares and absolute value regressions.

The shapes of error variables are analyzed through the level curves of their density functions under conditions of independent, identically distributed variables, heteroskedasticity, serial correlation, and errors in variables.

The error variables in any model lead to the definition of a distance function that is used to estimate the model. It is shown that the level curves of the distance function are the same as the level curves of the density function of the error. The geometric operation of estimation techniques is examined, including ordinary least squares and absolute value estimations. In particular, geometric structures are used to provide a new proof for the solution to orthogonal regression.

The paper provides a summary of the literature on geometric approaches to regression and suggestions for future research.

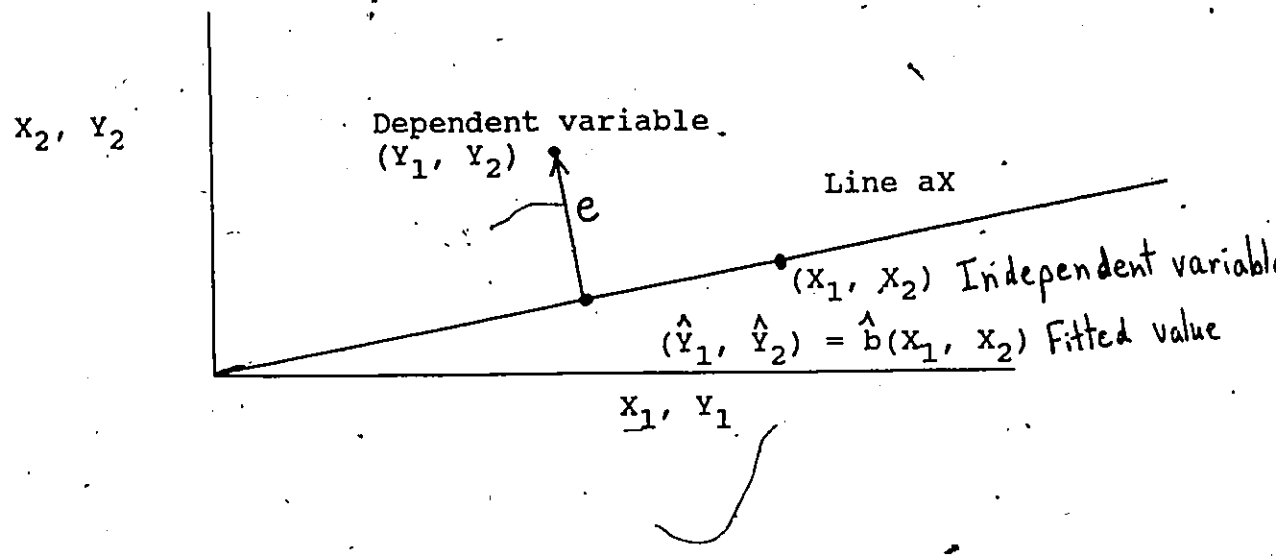
OVERVIEW

The purpose of this study is to analyze the geometric structure of regression analysis. Geometry provides an interesting and robust way of interpreting regression techniques, and leads to new insights. Geometric analysis has tremendous intuitive appeal. It can be easily visualized because of its parallels with the three-dimensional world of our daily experience. We will see that regression has a rich geometric structure that applies to all regression techniques.

The Oxford dictionary defines geometry as the "Science of properties and relations of magnitudes (as lines, surfaces, solids) in space." In this sense, regression has many geometric aspects. For example, a regression technique can be represented as occurring in the "space" of n x k matrices. We will see that every regression model implicitly defines a line or surface within this space. In fact, the whole purpose of regression is to find the most appropriate point on this surface given an observed point that is not on the surface.

Another important geometric aspect of regression is the use of distances: all regression techniques choose the optimal fitted values through the minimization of the distance between the observation and the surface defined by the model.

The simplest example of this geometric structure is found in ordinary least squares. Consider the simplest of all cases, where we estimate the equation $Y = bX + e$ with no constant term and two observations on both Y and X. In this case, the set of points that fits the model perfectly is the line spanned by the vector $X' = (X_1, X_2)$. The estimation procedure finds the point on this line of minimum distance to the observed point $Y' = (Y_1, Y_2)$. The estimated point is $\hat{Y}' = (\hat{Y}_1, \hat{Y}_2) = \hat{b}\ (X_1, X_2)$. The error is the vector $Y - \hat{Y}$, whose length is minimized among all possible points on the line aX.



This paper will show that all regression techniques have a geometric structure similar to the simple form outlined above. We will show that regression techniques are distinguished from each other by their distance function.

The first part of this paper will give an overview of geometric aspects of regression and a brief review of the literature on geometric approaches to regression. The rest of the paper will develop a general model for the geometry of all linear regression techniques, with examples examined throughout to illustrate particular cases.

1. Introduction

In this section, we will attempt to answer the question "What is geometric about regression?" by examining what regression methods do, and looking at some common examples. This introductory analysis will provide motivation for the more detailed analysis of the rest of the paper, and will hopefully spark some interest on the part of the reader. We will see that regression has a rich geometric structure, and that this geometric structure can provide interesting insights into estimation techniques.

1.1 What is Regression Analysis?

Regression analysis is a branch of statistics that estimates the parameters of a relationship between two or more variables. In order to perform any regression estimation, one needs three basic ingredients: a model, a set of data and an error structure. The Model consists of one or more equations with unknown parameters that describe the relationship between the variables involved.

An equation describes how the variables are linked to each other. For example, the equation $Y = aX + bZ + e$ says that the variable Y is roughly a linear combination of the variables X and Z. The parameters a and b are values that stay the same

for all the different observations in the data set. The equation $Y = aX + bZ + e$ says that the estimate of Y is always the same linear combination of X and Z, even though we do not know which linear combination.

The model also includes assumptions about error variables in the equation. For example, the model may assume that the errors in each observation are independent, identically distributed normal random variables with mean zero.

1.2 The Set of Data

The data are several series of real numbers where each series represents one of the variables in the model. For example, if the model involves variables X, Y, and Z, the data consist of three series $(X_1,\ldots, X_n)$, $(Y_1,\ldots, Y_n)$ and $(Z_1,\ldots, Z_n)$. The data can also be represented as a series of "observations" on the variables, where an observation is the triplet (X_i, Y_i, Z_i), for $i = 1,\ldots,n$. In general, the data can be represented in a matrix where each column represents the series of values observed on one of the variables. If there

are k variables and n observations, the matrix will have k columns and n rows:

$$\begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & & \vdots \\ X_{n1} & & X_{nk} \end{bmatrix}$$

1.3 Putting the Model and the Data Together

The equation or equations are assumed to represent the relationship between the variables for each and every observation. For example, if the equation of the model is $Y = aX + bZ$, then ideally every observation would fit this equation exactly. We would write:

$$\begin{aligned} Y_1 &= aX_1 + bZ_1 \\ Y_2 &= aX_1 + bZ_2 \\ &\vdots \\ Y_n &= aX_n + bZ_n \end{aligned}$$

or equivalently $Y = aX + bZ$ where $X' = (X_1, \ldots, X_n)$
$Y' = (Y_1, \ldots, Y_n)$
$Z' = (Z_1, \ldots, Z_n)$

or equivalently $Y = \left[\begin{array}{c|c} & \\ X & Z \\ & \end{array}\right] \begin{bmatrix} a \\ b \end{bmatrix}$

If this were the case, regression estimation would be relatively easy; we could simply plug a few of the observations into the equation and solve for the unknown parameters.

The problem in regression is that the data that we observe in reality never quite fit the equations exactly, whatever parameters we choose. Thus the problem in regression estimation is twofold: in addition to not knowing the values of the parameters of the equations, we do not even have data that fit the equations. In order to estimate the parameters of the equation(s), we have to alter the values of one or more of the series of variables to make them fit the equation(s). This is why we need the third element mentioned above.

1.4 The Error Structure

The error structure serves to explain why the data do not fit the equations exactly, and indicates how to adjust the variables so that they fit the equation. For example, suppose that the error structure assumes that the data do not fit the equations $Y_i = aX_i + bZ_i$ because there is one unobservable random error e_i in every equation. The equations should read $Y_i = aX_i + bZ_i + e_i$, or equivalently

$Y_i - e_i = aX_i + bZ_i$. In this case the variable Y_i has to be adjusted by the amount e_i in order to fit the model $aX_i + bZ_i$.

The above broad description of regression analysis fits virtually all regression estimation techniques. What distinguishes one technique from another are the ingredients that one begins with, that is the model, the data and the error structure. We will see that the most important distinguishing ingredient from a geometric point of view is the error structure.

1.5 What is Geometric About Regression?

We are now in a position to sketch the basic geometric structure that underlies all regression techniques. We have a set of data that can be represented in a matrix where each row is an observation of the variables in the model. Let the variables in the model be $X_1,\ldots,X_k$, and suppose we have n observations on each of these variables, say $(X_{i1},\ldots,X_{ik})$, $i = 1,\ldots,n$. In matrix form, this can be written as

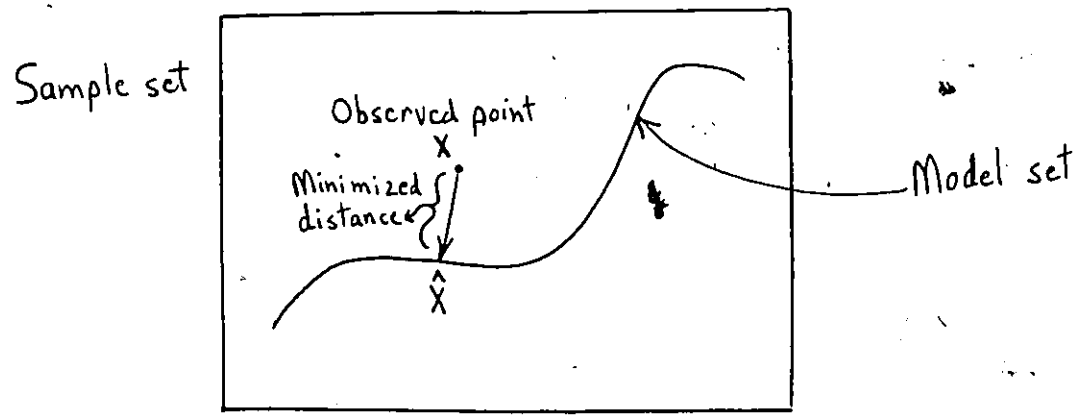
$$X = \begin{vmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & & \vdots \\ X_{n1} & & X_{nk} \end{vmatrix}$$

This matrix is simply one point in the n x k -dimensional set

of n x k matrices. Let us call this set the "sample set" S because it is the set of all possible n x k matrices of observations given the variables in the model. The problem with the point X is that it does not fit the model whose parameters we wish to estimate. A natural question to ask is: What points in the sample set actually do fit the model? We shall assume that such points exist. Let M be the set of all n x k matrices that do fit the model for all possible values of the parameters. We shall call this set the "model set M", because we wish to end up with a point on this set. We will see that for most points in the model set there is a unique set of parameters for the equations in the model. The set M is a subset of the sample set S. The problem of regression is to find which point in the model set M we should pick, given that we have observed the point X. We will see that the criteria for choosing the appropriate point in M is simply to choose the point in M <u>closest</u> to the observed point X, or equivalently the point of minimum distance from X. It is the error structure of the regression that determines how precisely we go about measuring distances between points in the sample set in order to pick the point on the model of minimum distance from the observed point. The vector matrix between the observed point X and its estimate $\hat{X}$ on the model set is the estimated error: $\hat{E} = X - \hat{X}$.

In a nutshell, this is the geometric structure underlying

all regression estimations: We begin with an observed point in a large sample set, and a model set contained in the sample set. The model set is the set of all "nice" points that fit exactly onto the model. The estimation procedure decomposes the observed point X into two points $\hat{X}$ and $\hat{E}$ such that $X = \hat{X} + \hat{E}$. The point $\hat{X}$ is on the model set and is the closest point to the observed point. This is a remarkably simple structure, considering that it encompasses all regression techniques, whatever their assumptions with respect to the distributions of the errors, number of variables, etc. It can be represented by a simple conceptual diagram as follows:



1.6 Geometric Interpretations of Regression Estimation

This section will briefly summarize the econometric literature on geometric approaches to regression estimation. We will begin with an extract from an article by David Herr, "On the History of the Use of Geometry in the General Linear

Model". (Herr, 1980, p. 43) The extract gives an excellent summary of the geometric approach to ordinary least squares, as compared to the standard "algebraic" approach. The standard approach could also be called "ordinary" ordinary least squares:

> Although there is a good deal more involved in the general linear model than least squares estimation, the fundamental ideas in least squares estimation are the fundamental ideas in the general linear model. In both cases, one considers a space (set) in which data must lie, a subspace (subset) of this space that corresponds to some assumptions on the data, and the relationship of the observed data to this subspace (subset).
>
> There are two general points of view taken with respect to linear models - an algebraic one and a geometric one. Consideration of least squares estimation from each point of view will illustrate these different perspectives.
>
> We suppose we have data $y = (y_1, \ldots, y_n)'$, which lie in Euclidian n-space, R^n. We further assume that there is a parameter vector $\beta = (\beta_1, \beta_2, \ldots, \beta_k)'$, which lies in R^k, $k \leq n$, and an $n \times k$ matrix X, so that $y = X\beta + \text{error}$. We wish to estimate β.

2. LEAST SQUARES ESTIMATION - ALGEBRAIC VIEWPOINT

Following (Gauss 1857) and (Legendre 1806), we choose as estimates of the β_i's those values β_i that minimize

$$Q = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} x_{ij}\beta_j)^2, \qquad (2.1)$$

where x_{ij} is the (i,j) element of X. To minimize Q, we consider the necessary conditions for a local minimum:

$$\partial Q(\beta_1,\ldots,\beta_k)/\partial\beta_p$$

$$= \sum_{i=1}^{n} 2(y_i - \sum_{j=1}^{k} x_{ij}\beta_j)(-x_{ip}) = 0 \quad (2.2)$$

for p = 1,...,k. We are thus led to the normal equations

$$\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ip}x_{ij}\beta_j = \sum_{i=1}^{n} y_i x_{ip},\ p = 1,\ldots,k, \quad (2.3)$$

or in matrix form

$$X'X\beta = X'y. \quad (2.4)$$

If $\hat{\beta}$ is the solution of these equations, $\hat{\beta}$ is called a least squares estimate of β.

3. LEAST SQUARES ESTIMATION - GEOMETRIC VIEWPOINT

To find values β, that minimize Q, we rewrite Q as

$$Q = \|y - X\beta\|^2 \quad (3.1)$$

and notice that Q is the squared distance of y from [X], the subspace of R^n spanned by the columns of X. Minimizing Q corresponds, then, to finding the point in [X] closest to y. The answer is readily visualized as the "point in [X] directly below y," that is, the perpendicular projection of y on [X]. If X denotes the perpendicular projection of y on [X], then $X\hat{\beta}$ is unique and satisfies

$$y = X\hat{\beta} + Z,\ Z \text{ perpendicular to } [X]. \quad (3.2)$$

Multiplying by X' we have

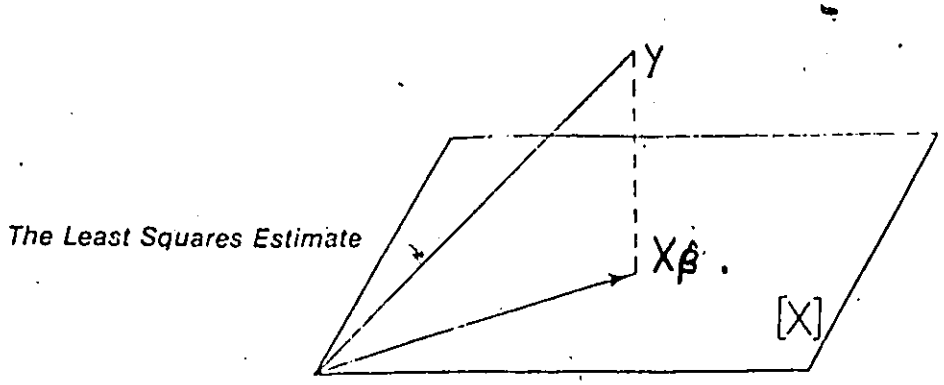
$$X'y = X'X\hat{\beta}. \quad (3.3)$$

since $X'Z = \emptyset$. Thus $\hat{\beta}$ must satisfy the normal equations. We appear to have arrived at the same place as in the

algebraic viewpoint [...] At this point we know there is a solution, $\hat{\beta}$, that $X\hat{\beta}$ is unique and that for any solution, $\hat{\beta}$,

$$\|y - X\hat{\beta}\|^2 \leq \|y - X\beta\|^2, \qquad (3.4)$$

for all β in R^k. Thus $\hat{\beta}$ yields the global minimum of Q.



The geometric approach to Ordinary Least Squares, as outlined above, is often mentioned in econometric texts as an interesting aside to the standard approach that uses algebra and calculus to derive estimators. Examples include Malinvaud (1970; pp. 19-23), Scheffé (1959; pp. 10-12) and Rao (1973; pp. 228-220).

Some authors have used a geometric approach to analyze particular estimation methods in order to enrich the understanding of the subject and to obtain further implications. Examples of these include:

1) Pollock, The Algebra of Econometrics, (1979) which emphasizes geometric notions in the exposition of common econometric methods in regression estimation;

2) Afriat, Regression and Projection: Studies in the Algebra and Geometry of Statistical Correlation, (1975) which gives a detailed analysis of ordinary least squares, correlation between variables and principal components using geometric notions such as orthogonality, angles and projections.

3) Gordon Fisher and Michael McAleer, The Geometry of Specification Error, (1984) which uses a geometric

approach to analyze the effect of the incorrect inclusion or exclusion of variables from the regression equation.

4) Fisher, A Co-ordination-free approach to linear estimation in Econometrics, (1973) which gives a geometric exposition of significance tests, restrictions on variables, standard assumptions of linear estimation and other topics.

Another interesting use of geometry in regression is given by Lester D. Taylor in Frontiers in Econometrics (1974, pp. 187-188). Taylor contrasts the geometry of Ordinary Least Squares (OLS) with the geometry of estimation by minimizing the sum of absolute errors (LAE):

> It is a well-known property of least squares estimation that its estimator can be interpreted as a projection - in the spherical disturbance case, as the orthogonal projection of the vector lying in T-dimensional space defined by the observations on the dependent variable into the n-dimensional subspace spanned by the vectors defined by the observations on the independent variables. In particular, the least squares estimator can be viewed as the one that is obtained by expanding a hypersphere with centre at y until it touches the hyperplane defined by the independent variables. The point of tangency yields the vector $\hat{Y} = X\hat{\beta}_{LS}$, where $\hat{\beta}_{LS} = (X'X)^{-1} X'y$.
>
> With LAE, the hypersurfaces of constant distance about y are not hyperspheres as with least squares, but rather regular polyhedrons whose corners lie in the

> rectangular coordinate axes emanating from y as the origin. The LAE estimator will then be defined as the $\hat{Y} = X\hat{\beta}$ obtained by expanding the polyhedron of equal distance until it touches the hyperplane defined by the independent variables.

The above-mentioned articles all use geometric ideas to analyze particular regression estimation techniques. The purpose of this paper is to construct a general geometric model that encompasses all regression techniques as special cases.

The geometric approach to ordinary least squares provides a simple, straightforward and intuitive way of interpreting OLS estimation. This paper is an attempt to generalize this approach to all single-equation regression techniques, including Generalized Least Squares, Absolute Value regressions and Orthogonal Regression. To the best knowledge of the author, this geometric overview of regression has not been attempted previously. The results presented in the remainder of this paper are original, unless specifically referenced in the text.

1.7 Scope of the Study

Having reviewed the literature on geometric approaches to ordinary least squares, we will proceed with an examination of the geometry of regression in general. We will see that a general geometric model of regression can be constructed, and

that special cases of the model will cover virtually all estimation techniques. Many of the special cases have interesting geometric properties which will be examined.

This inquiry will limit itself to the discussion of single-equation linear regression models with a predetermined set of variables for which data are available on all observations. Excluded from the discussion are multiple equation models and estimations with missing observations. In other words, this inquiry is concerned with the classical problem of fitting n observations on each of K variables to a single equation with unknown parameters.

This paper will approach the construction of a general model for regression estimation by examining what all estimation techniques have in common. Once we have constructed a model that exhibits all the properties held in common by regression techniques, the geometric properties that distinguish one technique from another will be examined.

2. Observations, Parameters and Variables

All regression estimations begin with a matrix of data. Each row of the matrix is a vector containing one observed value on each variable in the model; each column of the matrix contains all of the observations on one of the variables. Let X be any n x k matrix of observations, which we will refer to as the "sample matrix" or the "sample point".

The column vectors of X will be referred to as X_i, with only one subscript, since most of this analysis concerns the columns of X.

2.1 The Sample Set

Let S, the "sample set", be the set of all possible sample matrices, that is the set of all n x k matrices. In most cases, the set of all possible sample matrices is actually much smaller than the set of all n x k matrices. For example, many economic variables can only take values greater than zero. However, this has no influence on the estimation method or on the geometric structure of the estimation.

The sample set is important in the geometric structure of regression because it is within this set that all of the "action" takes place; it is akin to the simple piece of paper on which all two-dimensional geometric analysis takes place.

Regression estimation isolates the systematic portion of the observation from the portion of the observation caused by the errors in the model. We will now define the elements of the process that takes us from a sample point to its corresponding error-free component.

2.2 The Model

The model consists of an equation and a set of assumptions about the variables in the equation.

In this paper, equations will usually be given in their homogeneous form:

$$\emptyset = a_1X_1 + a_2X_2 + \ldots + a_kX_k$$

where the X_i represent economic variables and the a_i represent fixed parameters. If a constant term is included in the equations, it can be represented by a variable with the value 1 for all observations.

2.3 Parameters

The parameters $a_1, \ldots, a_k$ in an equation form a k-vector $\mathbf{a} = (a_1, \ldots, a_k)$ in the k-dimensional parameter space R^k. A non-zero parameter vector that fits a homogeneous equation of the form $\emptyset = a_1 X_1 + a_2 X_2 + \ldots a_k X_k$ is not unique. It can be multiplied by any non-zero constant and still fit the same homogeneous equation. In order to avoid confusion about multiple values for parameters, we will sometimes adopt the conventional normalization $\mathbf{a}'\mathbf{a} = 1$. This restricts the parameter vector to the hypersphere of radius 1 within the k-dimensional parameter space R^k. This hypersphere is essentially a k-1-dimensional set, since knowing k-1 of the values of a vector in this hypersphere will tell us what the kth value is. We will refer to this k-1-dimensional hypersphere as S^{k-1}.

There is still a degree of redundancy in the parameter sphere S^{k-1}, since a parameter vector $\mathbf{a}$ is equivalent to the parameter vector $-\mathbf{a}$ because they both fit the same equations. Where necessary, we will refer to the half-sphere of S^{k-1}, which restricts one of the parameters to positive values, eliminating the redundancy of S^{k-1}.

2.4 Systematic and Random Variables

There are two types of variables in a regression equation: systematic variables and random variables.

Systematic variables represent economic factors that can only take on one fixed value at any given point in time.

Random variables, on the other hand, can take on any value at a point in time. All that we can assume about a random variable is the probability distribution of these values, as specified by its distribution function.

All regression equations contain at least two systematic variables because regression deals with the relationships between variables. Regression equations all contain at least one random variable because the data never exactly fit the systematic equation variables. The random variable explains the divergence between the data and the systematic variables in
the model.

We can write a generic equation whose special cases include all single equation regression models as follows:

(1) $\emptyset = a_1 (X_1-E_1) + \ldots + a_k(X_k-E_k)$.

where the X_i are systematic variables and the E_i are random variables. This equation appears to exclude the possibility of an error associated with the equation itself, since all of the error variables are associated with a particular variable. However, any one of these error variables can be split off and

treated as an error in the equation. If the model requires an error variable associated with every systematic variable as well as an error in the equation, one of the error variables can be treated as the sum of the two random variables, one associated with the systematic variable and one associated with the equation. The reason for incorporating any error variable in the equation to the errors associated with the systematic variables is that in practice, regression estimation techniques treat error variables in this way.

In matrix form, this equation is:

$\mathbf{O} = [X - E]\mathbf{a}$

where $X = [X_1 X_2 \ldots X_k]$, $E = [E_1 E_2 \ldots E_k]$ and $a' = (a_1, \ldots, a_k)$.

It will be useful to note the error-free version of the generic equation. If all of the random variables are zero, the equation of the model is:

$\emptyset = a_1 X_1 + \ldots + a_k X_k$

In matrix form, this gives:

$\emptyset = Xa$

We will refer to this as the error-free equation.

2.5 Observed and Unobservable Variables

All random variables are by definition unobservable. Systematic variables, on the other hand, may or may not be observable depending on the assumptions of the model. The assumption that the variable X_i is observable is equivalent to assuming that its associated error E_i is always zero. In this case, the values of X_i are identical to its values in the observed matrix X. The assumption that the error E_i associated with X_i is non-zero means that the variable X_i is not directly observable. This means that the values of the observed variable are actually observations of the sum of the systematic variable and its associated random variable. Letting X^o_i be the observed variable, we can write $X^o_i = X_i + E_i$, where X_i is the (unobservable) systematic component and E_i is the (unobservable) random component.

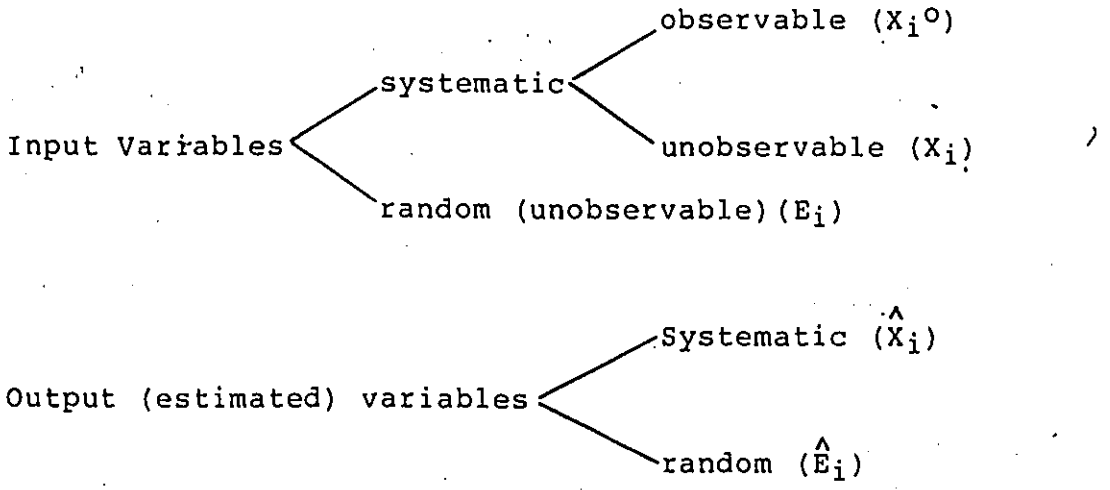
Since the purpose of regression is to estimate the systematic relationship between the variables, regression estimation must estimate the unobservable values both of the systematic component X_i and the error component E_i when the variable X_i is non-observable.

2.6 Estimated Variables

Estimated variables are observed variables whose associated errors have been estimated and removed. They are sometimes called fitted values. The model that results from a

regression estimation contains estimates of all variables that have an associated error. These estimates are obtained by subtracting the estimated error from the observed variable. In the above equation, if X^o_i is an observed variable containing the estimated $\hat{E}_i$, we obtain the estimated variable $\hat{X}_i = X^o_i - \hat{E}_i$. A variable observed with error, $X^o_i = X_i + E_i$, is decomposed into two estimated variables, one for the systematic component and one for the random component, such that $X^o_i = \hat{X}_i + \hat{E}_i$. When all the variables have been separated into estimates of their systematic and random components, the whole observed matrix can be written $X^o = \hat{X} + \hat{E}$ where $\hat{X}$ is a matrix of estimated systematic variables and $\hat{E}$ is a matrix of estimated random errors. We shall call the point $\hat{X}$ the estimate of X^o and $\hat{E}$ the estimated error of X^o.

The following table summarizes the types of variables found in regression models:



2.7 Examples

We will now show how the above definitions apply to the common parent models of econometrics.

2.7.1 Ordinary Least Squares

Let X_1 be the dependent variable and $X_2,\ldots,X_k$ be the independent variables of an O.L.S. estimation. Since the independent variables are assumed to be error-free, their associated errors $E_2,\ldots,E_k$ are zero. This leaves us with the model:

$$a_1(X_1-E_1) + a_2X_2 + \ldots + a_kX_k = \emptyset.$$

Dividing by a_1 and rearranging the variables gives the traditional form of the O.L.S. equation:

$$X_1 = b_2X_2 + .. + b_kX_k + 1/a_1\ E_1 = \emptyset$$

Once the equation is estimated, we end up with the fitted equation:

$$X_1 - \hat{E}_1 = \hat{X}_1 = \hat{b}_2X_2 + \ldots + \hat{b}_kX_k.$$

In matrix form, we have:

$$X = \hat{X} + \hat{E} \quad \text{where:} \quad \hat{X} = [\hat{X}_1, X_2, \ldots, X_k] \text{ and}$$

$$\hat{E} = [\ \hat{E}_1, \emptyset, \ldots, \emptyset]$$

and where $\hat{X}\hat{b} = \emptyset$, $\hat{b} = \begin{vmatrix} -1 \\ \hat{b}_2 \\ \cdot \\ \cdot \\ \hat{b}_k \end{vmatrix}$

Alternatively, we could write $\hat{X}_1 = [X_2,\ldots,X_k] \begin{bmatrix} \hat{b}_2 \\ \vdots \\ \hat{b}_k \end{bmatrix}$

which is the traditional form for O.L.S.

For the purposes of estimation, it does not matter whether the error variable E_1 is assumed to be an error in the equation, as in O.L.S., or an error in the variable X_1. Either way, all of the adjustment made to fit the sample point to the model equation is borne by the variable X_1. The estimation procedures under these two sets of assumptions are identical.

The equation as specified above also applies to Generalized Least Squares, whose distinguishing features emerge only in the distribution of the error variable.

2.7.2 Orthogonal Regression

Orthogonal Regression, sometimes called Pure Errors in Variables, assumes that the observed data do not fit the model equation because all of the observed variables contain random errors. There is no error associated with the equation itself. Thus the model for Orthogonal regression is

$$a_1(X_1-E_1) + \ldots + a_k(X_k-E_k) = \emptyset$$

where the E_i are all independent, identically distributed

random variables. Once the systematic and random components of each variable have been estimated, we obtain the fitted equation:

$$\hat{a}_1\hat{X}_1 + \ldots \hat{a}_k\hat{X}_k = \emptyset$$

In matrix form, this gives

$$X = \hat{X} + \hat{E} \text{ where } \hat{X} = [\hat{X}_1,\ldots,\hat{X}_k]$$

$$\hat{E} = [\hat{E}_1,\ldots,\hat{E}_k]$$

and where $\hat{X}\,\hat{a} = \emptyset$

2.7.3 Hybrid Model

Some situations call for a combination of the O.L.S. and Orthogonal Regression Models. Let us assume that some (more than one as is the case in O,L.S.) but not all (as is the case in Orthogonal Regression) of the variables have an associated error. In this case, the model becomes:

$$a_1(X_1-E_1) + \ldots + a_i\,(X_i-E_i) + a_jX_j + \ldots + a_kX_k = \emptyset.$$

Once estimated, the fitted equation is:

$$\hat{a}_1\hat{X}_1 + \ldots + \hat{a}_i\hat{X}_i + \hat{a}_jX_j + \ldots + \hat{a}_kX_k = \emptyset$$

In matrix form, this gives:

$$X = \hat{X} + \hat{E} \text{ where } \hat{X} = [\hat{X}_1,\ldots,\hat{X}_i,X_j,\ldots,X_k]$$

$$\hat{E} = [\hat{E}_1,\ldots,\hat{E}_i,\emptyset,\ldots,\emptyset]$$

and where $\hat{X}\,\hat{a} = \emptyset$.

Now that we have defined the variables and equations that make up a regression, we will examine how they form a geometric figure within the sample space.

3. The Model Set

The model set is the set of possible sample points that fit the model's error-free equation. They are the "nice" points we would always like to observe when estimating a regression, although this virtually never happens.

This set is essential to the geometric analysis of regression because the estimated matrix of any regression is always a point on the model set. All regression techniques estimate the error component and subtract it from the observation in order to obtain a point that fits the model's error-free equation. This point is needed in order to determine the parameters of the estimated equation. These parameters reveal the systematic structure underlying the observed data point.

Definition

The model set M is the subset of the sample set consisting of points X such that $Xa = \emptyset$ for some non-zero vector of parameters $a = (a_1, \ldots, a_k)$. In set terminology, $M = \{X \in S \mid Xa = \emptyset \text{ for some } a \in R^k, a \neq \emptyset\}$

The above definition can be given in several alternate forms, which we will use from time to time. They are:

(1) $M = \{X \in S | \hat{X} = X\}$, that is the set of all points that, when estimated and fitted to the model, are equal to themselves.

(2) $M = \{X \in S | \emptyset = \sum_{i=1}^{k} a_i X_i$ for some $a_i \neq \emptyset\}$, that is the set of nxK matrices where there is linear dependence among the columns.

(3) M is the set of matrices with rank less than or equal to k-1, since the rank of a matrix is the number of independent column vectors it contains.

(4) $M = \{X \in S | \det(X'X) = \emptyset\}$, that is the set of matrices where the determinant of X'X is equal to zero. This is true because the determinant of X'X is zero if and only if there is linear dependence among the columns of X (where $n \geq K$).

It is interesting to note that the observation of a point in the model set does not guarantee that the point is free of error. An observed point X^o is the sum of its systematic and error components, $X^o = X+E$. If the error E is zero, then $X^o = X$ and X^o must be in the model set. The converse is not necessarily true. If we observe X^o where $X^o a = \emptyset$ for some parameter vector a, X^o may still contain a component of error, that is $X^o = X + E$ where X is in M and

$E \neq \emptyset$. For this to occur, all that is required is for the vector E to take us from one point in the model set to another. In other words, it is possible, though unlikely, to observe economic data that fit perfectly onto a model equation yet contain a component of error. As an example, consider the observation:

$$X^o = \begin{bmatrix} 10 & 5 \\ 12 & 6 \\ 14 & 7 \\ 16 & 8 \end{bmatrix}$$

which is in the model set for n = 4, k = 2, because it fits the equation $X_1 - 2X_2 = \emptyset$ without error. It is possible for the observation X^o to have been generated by the components:

$$X = \begin{bmatrix} 9 & 3 \\ 12 & 4 \\ 15 & 5 \\ 18 & 6 \end{bmatrix} \qquad \text{and} \qquad E = \begin{bmatrix} 1 & 2 \\ \emptyset & 2 \\ -1 & 2 \\ -2 & 2 \end{bmatrix}$$

The actual underlying model in this case is $X_1 - 3X_2 = \emptyset$. The problem with errors of this sort that occur within the model set is that there is no way to practice statistical inference on them, so they will not be dealt with in detail.

We will now examine some of the geometric characteristics of the model set to get a feel for how it sits within the

sample space.

Lemma 1

The Model Set consists entirely of lines through the origin (i.e. the matrix of zeros).

Proof

Pick any point X in the model set besides the origin. If there is linear dependence among the columns of X, then $Xa = \emptyset$ for some $a \neq \emptyset$. But if $Xa = \emptyset$, then $(\alpha X)\, a = \emptyset$ for any $-\infty < \alpha < \infty$, so all points on the line from the origin through X are also in the model set. In particular, when $\alpha = \emptyset, \alpha X = \emptyset$, so the origin is in the model set.

It is worth noting some of particular lines that belong to the model set. All of the nxk coordinate axes of R^{nxk} belong to the model set. Let X be a matrix with the entry 1 in the ith row, jth column, and with zeros in all other positions. This point is obviously in the model set, as is any linear multiple of this point. This shows that lines in the model set point out in many different directions.

This also shows that the model set is not a linear subspace of the sample set. A linear subspace is a set that contains all linear combinations of points within it. The set

of all linear combinations of points on the nxk coordinate axes of $R^{n}xk$ is the whole sample space $R^{n}xk$. However, it is clear that the model set does not include all points in $R^{n}xk$, since many points in $R^{n}xk$ do not have linear dependence among their columns. Thus the model set is not a linear subspace.

This fact has important implications for the estimation of the model. Many regression estimation techniques use the properties of linear subspaces to estimate the parameters of the model because linear spaces are easier to deal with. We will see in later sections how these methods reduce the problem of estimation in the non-linear model set to the simpler problem of estimating in a linear environment.

Lemma 2

The model set consists of n(k-1)-dimensional hyperplanes. Each hyperplane is the null space of a vector of parameters.

Proof

Let X be any point in the model set. Then $Xa = \emptyset$ for some non-zero $a \in R^{k}$. (The vector $\emptyset$ is in R^{n}.) The vector a is a linear transformation from $R^{n}xk$ to R^{n}. The null set of this transformation is the set of Y in M that are mapped onto the zero vector in R^{n}, that is $N(a) = \{Y \in M \mid Ya = \emptyset\}$. The null set of any linear transformation is a linear subspace whose

dimension is equal to the dimension of the domain of the transformation minus the dimension of the range of the transformation. In this case, the domain is of dimension nxk. The range of a is R^n. Thus the null space of a is a hyperplane through the origin of dimension $(nxk) - n = n(k-1)$.

The hyperplanes just discussed are important because each hyperplane represents a distinct economic structure. Within each hyperplane, all the points fit the same equation with the same parameters. Points on a different hyperplane fit an equation with different parameters, and so must be generated by a different economic structure. For example, consider a simple two-variable model of consumption as a function of income. All economies, big or small, where consumption is 90% of income, would generate systematic components on the same hyperplane. Economies where consumption was only 70% of income would be on another hyperplane, and so on for other parameter values.

We can express the whole model set as the union of these hyperplanes, each representing a distinct possible relationship among the variables:

$$M = \bigcup_{a \in R^k} N(a) \quad , a \neq \emptyset$$

Another important thing about these hyperplanes is their relationship to multicollinearity. The intersections of the hyperplanes are precisely the points in the model set where the problem of multicollinearity occurs. This aspect will be examined after we have finished with the model set as a whole.

There are as many different hyperplanes N(a) as there are possible economic structures, that is one for every line through the hypersphere S^{k-1}.

Lemma 3

The Model Set is a connected subset of the sample space. Furthermore, any two points of the model set can be connected without passing through the origin.

Discussion

Prior to proving this result, the mathematical notion of connectedness will be discussed briefly.

As its name implies, a connected set is a set that consists only of one piece, rather than two or more pieces separated from each other by points not belonging to the set. Examples of connected and non-connected sets in R^2 are given below.

Connected

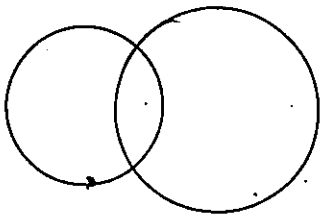


Figure 1

Not Connected

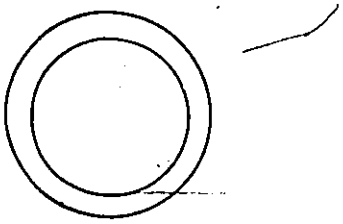


Figure 2

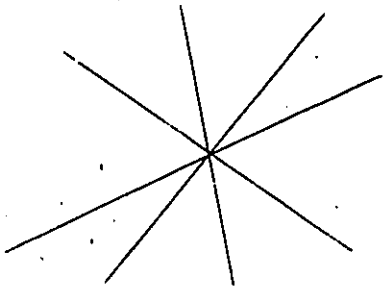


Figure 3

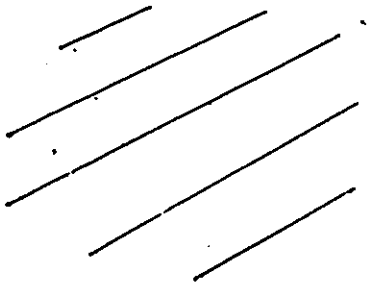


Figure 4

In order for a set to be connected, we must be able to find a continuous string of points, called a path, from any point in the set to any other point in the set without having to leave the set.

Proof

It is easy to show that the Model set is connected by using Lemma 1, since every point in the set is connected to the origin through its ray, and thus to every other point in the set. However, the model set is not simply a set of discrete rays emanating from the origin, as in figure 3 above. Even if the origin is omitted, the model set is still connected, as we will see.

Let X, Y be any two non-zero points in the model set.

In order to show that a path between X and Y exists, we proceed as follows. The columns of X, say X_1, ..., X_k, span a hyperplane of R^n of dimension d_y. Take the lesser of d_x and d_y, say d_x. From the properties of matrices, we can choose d_x independent vectors from the columns of both X and Y, say X^1, ..., X^j and Y^1, ..., Y^j where $d_x = j$. Now let us define a family of j-dimensional hyperplanes in R^n as follows:

$$H(\alpha) = \text{Span } ([X^1 + \alpha(Y^1 - X^1)], \ldots, [X^j + \alpha(Y^j - X^j)])$$

When $\alpha = 0$, $H(\alpha)$ is the span of X^1, ..., X^j, which is the hyperplane spanned by the columns of X. When $\alpha = 1$, $H(\alpha)$ is the span of Y^1, ..., Y^j, which belongs to the hyperplane spanned by the columns of Y. As α goes from zero to one, the hyperplane $H(\alpha)$ moves from the span of the columns of X to the span of the columns of Y.

Now let $E_\alpha(Xi)$ be the orthogonal projection of the vector X_i onto the hyperplane $H(\alpha)$. When $\alpha = 0$, $E_\alpha(X_i) = X_i$. When $\alpha = 1$, $E_\alpha(X_i)$ is a vector in the column span of Y. Now define the path:

$$P_\alpha[X] = [E_\alpha(X_1), \ldots, E_\alpha(X_k)]$$

where $o < \alpha < 1$. When $\alpha = \emptyset$, $P_\alpha[X]$ simply projects each column of X onto itself, so $P_o(X) = X$. As α moves towards 1, $P_\alpha(X)$ gives a matrix whose columns all belong to a d_X-dimensional hyperplane. Since $d_X < k$, there must be linear dependence among the columns, so the matrix $P_\alpha(X)$ clearly is in the model set. When $\alpha = 1$, $P_\alpha(X)$ consists of K columns that all belong to the column span of Y. By moving along the line segment $\overline{P_\alpha(X), Y}$, whose points all have columns contained in the column span of Y, we complete the path from X to Y. Of course, all points in the column span of Y are in the model set since the columns of Y are linearly dependent.

So far, we have examined some of the components of the model set and how they fit together. We have found that the model set is composed of linear hyperplanes through the origin, but that as a whole it is not a linear set. We have found that the set is connected, and that it does not consist of lines or hyperplanes that are connected only at the origin. We now turn our attention to the question of how big the whole model set is within the larger set of nxk matrices.

The size of the model set can be measured by its dimension. A subset of another set is said to be of dimension m if there is a continuous, one-to-one correspondence between the set and the set R^m. In order to give some intuitive appeal

to the following discussion, let us give a few examples of one-dimensional subsets contained in the two-dimensional plane:

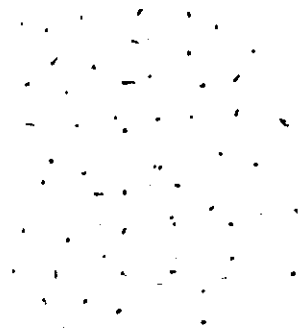


Figure 1

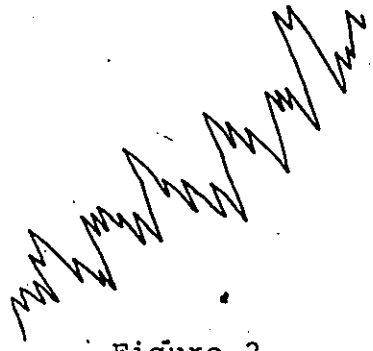


Figure 2

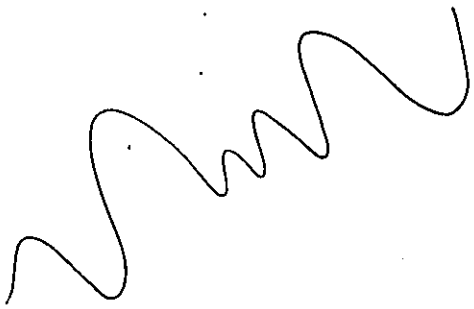


Figure 3

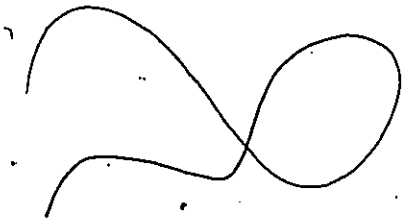


Figure 4

All of these sets are subsets of the plane. The scatter of points in Figure 1 is not a one-dimensional set because it is not a continuous image of a straight line. The other figures are one-dimensional because there is a continuous, one-to-one correspondence between them and the one-dimensional line. We can easily imagine taking the line like a piece of string and shaping it as in the above figures.

The property of dimension is a local property. It applies to pieces of the curve, but not necessarily to the curve as a whole. This is illustrated in Figure 4. At the points where the curve crosses itself, we cannot find an exact one-to-one correspondence between the set and the one-dimensional line. If we were to straighten out the line, there would be two points on the straight line corresponding to only one point in the set in the drawing. These points require special attention.

Lemma 2 gives us an indication of the dimension of the model set. This Lemma showed that the model set is composed of $n(k-1)$-dimensional hyperplanes. There are as many hyperplanes as there are lines through the half-hypersphere S^{k-1}. The hyperspere S^{k-1} is a set of dimension $k-1$. Thus it seems likely that the whole model set is of dimension $n(k-1) + (k-1) = (n+1)(k-1)$. In the following lemma, we show that this is indeed the case by constructing a direct, one-to-one correspondence between neighbourhoods of the model set and $R^{(n+1)(k-1)}$.

Lemma 4

The model set has dimension $(n+1)(k-1)$.

Proof

Let X be a point in the model set. Then $Xa = \emptyset$ for some $a \in S^k$. Let us further assume that $a_1 \neq \emptyset$. Then we can write X_1 as a linear combination of the column vectors $X_2, \ldots, X_k$:

$$X_1 = \sum_{i=2}^{k} -\frac{a_i}{a_1} X_i$$

Letting $c_i = \frac{a_i}{a_1}$ for $i = 2, \ldots, k$ gives us a $(k-1)$-dimensional vector of parameters $c = (c_2, \ldots, c_k)$. The vector c is a solution to the equation:

$$[X_2 \ldots X_k]\, c = X_1$$

Let us assume for a moment that the column vectors $X_2, \ldots, X_k$ are independent. The matrix $[X_2 \ldots X_k]$ can be reduced to its row-echelon form through elementary operations on its rows. If the columns of the matrix are independent, the row-echelon form is:

$$\begin{bmatrix} I_{k-1} \\ \cdots \\ 0 \end{bmatrix}$$

Elementary row operations can be executed by pre-multiplying a matrix by an elementary matrix. The operations that reduce $[X_2 \ldots X_k]$ to its row-echelon form are the product of a series of elementary matrices, say

$P = P_1P_2 \ldots P_p$. Applying these operations to the equation above gives:

$$P[X_2 \ldots X_k]\, c = PX_1$$

$$\begin{bmatrix} & \\ I_{k-1} & \\ \cdots & \\ 0 & \end{bmatrix} c = PX_1$$

$$\begin{bmatrix} & \\ c & \\ \emptyset & \\ \cdot & \\ \cdot & \\ \emptyset & \end{bmatrix} = PX_1$$

We will now show that the vector c is a continuous function of X.

Each matrix P_i represents one elementary row operation, either multiplying a row by a scalar, adding a scalar multiple of one row to another row or interchanging two rows of the matrix. For example, in order to row-reduce the first column of the matrix, we multiply the first row by $1/x_{12}$ so that the first entry equals one, then we subtract x_{i2} times the first row from the following rows, for $i = 2, 3, \ldots, n$. This gives a first column with a 1 in the first position and zeros elsewhere. The elementary matrices for these operations are as follows:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \cdot \\ \cdot & & & \cdot \\ -x_{2n} & \cdots & & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -x_{21} & 1 & \ddots & \cdot \\ \cdot & & & \cdot \\ 0 & \cdots & & 1 \end{bmatrix} \quad \begin{bmatrix} 1/x_{12} & 0 & \cdots & 0 \\ 0 & 1 & & \\ & \cdot & \cdot & \cdot \\ & \cdot & \cdot & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1/x_{12} & 0 & \cdots & 0 \\ -x_{22}/x_{12} & 1 & \cdots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ -x_{n2}/x_{12} & 0 & \cdots & 1 \end{bmatrix}$$

If the entry x_{12} happens to be zero, we must change the order of the rows before reducing the first column to row-echelon form. This is done by multiplying by the matrix:

$$\begin{array}{c} \\ i \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \overset{\quad i \quad}{} \\ \begin{bmatrix} 0 \cdots 1 \cdots 0 \\ .1. \quad\quad . \\ \cdot \quad \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \quad \cdot \\ 1 \quad\quad\quad \cdot \quad \cdot \\ \cdot \quad\quad\quad\quad \cdot \quad \cdot \\ \cdots\cdots\cdots 1 \end{bmatrix} \end{array}$$

in order to switch the ith row with the first row. We can be sure that at least one entry in the first column is non-zero since the column vectors are independent, and the zero vector is not an independent vector. Matrices similar to the above can be used to reduce the other columns to their row-reduced form. The matrix P is the product of all of these matrices. Every entry in the matrix P is the quotient of polynomials involving entries in the matrix X. Similarly, the product of

P with X_1 gives an n x 1 column vector, where each entry is the quotient of polynomials involving entries in the matrix X. We could write $P = P(x_ij)$. A quotient of polynomials is a continuous function as long as the denominator is not zero. We now have all the elements we need to construct a continuous function from the model set to $\mathbb{R}^{(n+1)(k-1)}$.

Now consider the function defined as follows:

$f(X) = (c, X_2, \ldots, X_k)$

This function simply takes the first column out of the matrix X and replaces it with the parameters of the first column written in terms of the other k-1 columns of the matrix. The range of the function has points consisting of a matrix with n rows and k-1 columns, plus a (k-1) - dimensional vector of parameters. The total dimension of the range-space is $(nx(k-1)) + (k-1) = (n+1)(k-1)$, as specified in the lemma.

The function f(X) is surjective, since every set of columns $(X_2, \ldots, X_k)$ and every linear combination c of these columns corresponds to some point in the model set.

The function f is continuous, because its component parts are continuous. The vector c is composed of continuous

functions, as we have just seen, and the rest of the function is simply the identity function, which is continuous.

To show that the function is one-to-one, we must show that $f^{-1}(f(x)) = X$. The inverse of f is:

$$f^{-1}(b, y_2, \ldots, y_k) = [\sum_{i=2}^{k} b_i y_i, y_2, \ldots, y_k]$$

Where b is in R^{k-1} and $y_2, \ldots, y_k$ are in R^n.

Plugging f(X) into this gives

$$\begin{aligned} f^{-}(f(X)) &= [\sum_{i=2}^{k} c_i x_i, x_2, \ldots, x_k] \\ &= [x_1, x_2, \ldots, x_k] \\ &= X \end{aligned}$$

Now that we have constructed a continuous, one-to-one correspondence between the model set and $R^{(n+1)(k-1)}$, we will return to take care of the loose ends that have been left hanging during this proof.

First, we began by assuming that $a_1 \neq \emptyset$. If $a_1 = \emptyset$, we could construct a similar function for some other non-zero parameter. We are assured that not all of the a_i are zero since $a \neq \emptyset$. The rest of the proof would follow as above. Since the property of dimension is a local property, it is legitimate to use different correspondences between M and $R^{(n+1)(k-1)}$ for different neighbourhoods of M.

The second assumption we made was that the vectors

$X_2, \ldots, X_k$ are independent. We assumed that the first column X_1 could be written as a linear combination of the k-1 independent vectors $X_2, \ldots, X_k$. We could have chosen any column of the matrix and constructed the same function as long as the column could be written in terms of the other k-1 independent column vectors of the matrix X. This is always possible when the rank of the matrix is k-1.

If the rank of X is less than k-1, then the vectors $X_2, \ldots, X_k$ cannot be independent. As we will see later, matrices with rank less that k-1 fit more than one parameter equation $Xa = \emptyset$. This causes a problem similar to that illustrated in Figure 4, page 33, because these matrices cannot be matched one-to-one to a point in $R^{(n+1)(K-1)}$ by the function f(X). We will see that such points are points of multicollinearity. The geometry of multicollinearity will be examined in detail in the next section.

We have shown that the model set in regression estimation has dimension $(n+1)(k-1)$, and that the only points that cannot be matched one-to-one to points in $R^{(n+1)(k-1)}$ are points of multicollinearity. If we exclude the multicollinear points, we can obtain a stronger characterization of the model set. We can show that it is an $(n+1)(k-1)$-dimensional manifold.

A manifold is a smooth image of a Euclidean space. The

mathematical property of smoothness is similar to its everyday meaning. Intuitively, a smooth set is one that is free of sudden bumps, corners, holes and peaks. These properties are expressed mathematically through the partial derivatives of the function f that matches neighbourhoods of the model set to $R^{(n+1)(k-1)}$.

Figures 3 and 4 on page 33 are examples of smooth sets, while Figure 2 shows a set that is not smooth.

Lemma 5

The non-multi-collinear points of the model set form an $(n+1)(k-1)$ - dimensional manifold. (The set of non-multicollinear points of the model set is the same as the set of nxk matrices with rank equal to k-1, since multicollinear matrices have rank less than k-1.)

Proof

To prove the lemma, we must show that the function $f:M \rightarrow R^{(n+1)(k-1)}$ is a diffeomorphism. A diffeomorphism is a function that is one-to-one and onto, which has already been shown for f(X). In addition, we must show that f(X) can be extended to a map from R^{nxk} to $R^{(n+1)(k-1)}$ which has continuous partial derivatives of all orders, and that the inverse of f(X) has continuous partial derivatives of all orders.

In order to extend f(X) to a map of all nxk matrices, let $F(X) = (c, X_2, \ldots, X_k)$ for all nxk matrices X,

where $c = \begin{vmatrix} - & - \\ c_2 & \\ \cdot & \\ \cdot & \\ \cdot & \\ c_k & \\ - & - \end{vmatrix}$ and $\begin{vmatrix} - & - \\ c & \\ 0 & \\ \cdot & \\ \cdot & \\ 0 & \\ - & - \end{vmatrix} = PX_1$

and P is the product of the elementary matrices that reduce $[X_2, \ldots, X_k]$ to row-echelon form. This is the same definition we used in defining f(X) on the model set. The only difference here is that, for points outside the model set, the first column is not a linear combination of the other columns, as weighted by the parameters in the vector c. Outside the model set, the vector c has no importance other than helping to define an extension of f(X) to all nxk matrices.

We have already seen that the function f(X) is continuous because it is composed of continuous components, namely the identity of $[X_2, \ldots, X_k]$ and a quotient of polynomials in the X_ij where the denominator is not zero for the elements of the vector c. The same is true of F(X). Similarly, F(X) has continuous partial derivatives of all orders. The partial derivatives of the identity function are either zero or one. The partial derivatives of the elements of c are all continuous quotients of polynomials in the variables X_ij.

Now let us consider the partial derivative of the inverse

function. Recalling that $f^{-1}(b, Y_2, \ldots, Y_k) = (\sum_{i=2}^{k} b_i Y_i, Y_2, \ldots, Y_k) = (Y_1, Y_2, \ldots, Y_k)$, the matrix of first order partial derivatives is:

	Range: $Y_1 = \sum_{i=2}^{k} b_i Y_i$, $Y_1' = (y_{11}, \ldots, y_{n1})$	$Y_2' \ \ldots \ Y_k'$
Domain: b_2	$\frac{\partial y_{i1}}{\partial b_j} = y_{ij}$	$0 \ \ldots \ 0$
b_k		
Y_2	$\frac{\partial y_{pl}}{\partial y_{mi}} = \begin{cases} b_i & \text{for } m=p \\ 0 & \text{for } m \neq p \end{cases}$ for $m,p=1,\ldots,n$; $i=2,\ldots,k$	$I_n \ 0 \ \ldots \ 0$
Y_k		$0 \ \ldots \ 0 \ I_n$

Clearly these partial derivatives are all continuous, as are higher order partial derivates.

This completes the necessary conditions for the non-multicollinear points of the model set to form an $(n+1)(k-1)$ - dimensional manifold.

Examples

In order to give the preceding results a more concrete interpretation, we will examine how big the model set is,

given specific sample sizes and numbers of variables.

Since regression is concerned with relationships between variables, the smallest number of variables in a regression is two. If we have only one observation on each of two variables, we can always find a linear relationship between the two because they are simply scalar multiples of each other. This is a trivial case, as there is no room for an error term to explain the relationship. As we would expect in such a case, the model set fills the whole sample space. The model set has dimension $(n+1)(k-1) = (1+1)(2-1) = 2$, as does the sample set of 2 x 1 matrices.

Non-trivial cases of regressions in two variables begin when $n = 2$. Here the sample set is the set of 2 x 2 matrices which has dimension 4. The model set is the subset of 2 x 2 matrices where the two columns are multiples of each other. It has dimension $(2+1)(2-1) = 3$, so there is one dimension left in which errors may range.

Consider a two-variable regression where $n = 20$. In this case, the model set is of dimension $(20+1)(2-1) = 21$. This is where all systematic variables would occur. The observed variables, containing components of error, can range over the full sample set whose dimension is 40. We see here how small

the model set is within the total sample space. Every point in the model set is associated with a 19-dimensional set of sample points not in the model set.

Now consider a multiple regression model with five variables. The minimum sample for a non-trivial regression in five variables is n = 5. In this case, the model set, containing all 5 x 5 matrices with linear dependence among the columns, would have dimension (5+1)(5-1) = 24. It is contained in the total sample set of dimension 25.

A regression in five variables with n = 40 would have a model set of dimension (40+1)(5-1) = 164, contained within a total sample space of dimension 200.

In all the above cases, the dimensions of the sample set and model set hold whatever the specific regression estimation technique used. We can see that as the sample size grows, the model set gets smaller and smaller in comparison to the sample set.

Non-Convexity

We have seen that the model set is not a linear subspace, and hence that methods of linear estimation cannot be applied directly to the set. Often in economics, we use convex

programming on non-linear sets where linear methods do not apply. For example, convex programming is used to find the maximum value of a utility function under budget constraints, or maximum profits under the limits of a production function. We will now show that convex programming cannot apply directly to regression because the model set is not convex.

In order to show that the model set is not convex, we need the following more general result.

Lemma 6

A non-linear set of measure zero is non-convex.

Proof

We will prove this result by showing that a convex set of measure zero must be a linear set, which is equivalent to proving the lemma.

By the term linear set, we mean a set that can be contained in a linear subspace that is strictly smaller than the ambient vector space.

Let W be any convex set of measure zero in any Euclidean space R^n. Pick any x, y in W. By convexity, the line segment $\overline{xy}$ is in W. The segment $\overline{xy}$ is contained in the line spanned by $\overline{xy}$, say $H_1 = \{x+a(x-y), -\infty < a < \infty\}$

If this line contains all points of W, we are done. If not, let z be in W, and not in H_1. By convexity, the segments $\overline{xz}$ and $\overline{yz}$ are in W, as is the triangle $\overline{xyz}$. The triangle $\overline{xyz}$ is contained in the plane that is spanned by the sides of the triangle $\overline{xyz}$, $H_2 = \{x+a(x-y)+b(x-z), -\infty < a,b < \infty\}$. If this plane contains all of W, we are done. If not, we pick a point w in W but not in H_2, and use it to construct a hyperplane $H_3 = \{x+a(x-y) + b(x-z) + c(x-w), -\infty < a,b,c < \infty\}$, and so on until we have a hyperplane that contains all of W.

If the hyperplane H_{n-1} does not contain all of W, the set W cannot have measure zero, so the assumption of the lemma has been violated. To see this, pick a point v in W that is not in H_{n-1}. By convexity, the n-pointed polyhedron $\overline{xyzw...v}$ must be contained in W, as are all points within it. But this polyhedron obviously has non-zero volume, so the set W has measure more than zero.

To apply this lemma to the model set, we note that the model set is a set of measure zero. This is true because the dimension of the model set is smaller than the dimension of the sample space, except for the trivial case when n<k. We also note that the model set is non-linear. This is true because it contains matrix vectors that generate the entire

sample space, as we noted earlier. Thus, the lemma confirms that the model set is not a convex set.

The fact that the model set in regression estimation is not convex is important because it sets apart regression estimation from most other types of optimization problems encountered in economics. Regression estimation is a problem of optimization under a constraint condition, similar to the problems of optimization under resource constraints encountered throughout economics. In regression, the objective function to be optimized is the distance function, and the solution is constrained to the model set. Regression is peculiar in that the constraint condition cannot be expressed as a linear or convex set, so that regression is one of very few problems in the class of non-convex programming.

In the section on distance functions in regression, we will examine how various regression techniques handle the problem of non-convex programming. We will see that in some cases, the problem is reduced to one of linear estimation, while other methods tackle the non-convex problem directly.

4. Multicollinearity

Multicollinearity is an important problem in regression estimation. It prevents the modeller from choosing a model for the observed data because it prevents the choice of parameters for the equation. In many cases, estimation techniques do not work when multicollinearity is present.

The model set consists of matrices whose columns fit homogeneous equations. Multicollinearity occurs when a matrix of data fits two or more different equations. This occurs when there is more than one relationship tieing together the systematic variables in the matrix. We will see the multicollinear matrices have a particular geometric configuration that can be used to determine the likelihood of multicollinearity occurring and how it can be eliminated.

Definition

A sample point is multicollinear if it fits two or more different error-free equations. The set of all multicollinear points, which we will call the collinear set, is:

$C = \{X \mid Xa = \emptyset \text{ and } Xb = \emptyset \text{ for } a, b \neq \emptyset, a \neq rb, r \in R\}$

A multicollinear matrix has at most k-2 independent column vectors. To see this, let $Xa = \emptyset$ and $Xb = \emptyset$. Then $X_i a_i = \emptyset$ and $X_i b_i = \emptyset$. Since $a \neq \emptyset$, there must be some

$1 \leq j \leq k$ such that $a_j \neq \emptyset$. Then we can divide the equation $\Sigma X_i a_i = \emptyset$ by a_j, and re-arrange the terms such that

$$X_j = \sum_{i \neq j} \frac{a_i}{a_j} X_i .$$

Substituting this into the equation $\Sigma b_i X_i = \emptyset$

gives $$\emptyset = \sum_{i \neq j} b_i X_i + b_j \left(\sum_{i \neq j} \frac{a_i}{a_j} X_i \right)$$

$$= \sum_{i \neq j} \left(b_i + b_j \frac{a_i}{a_j} \right) X_i$$

Thus there is linear dependence among the k-1 column vectors $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots X_k$. In other words, a multicollinear matrix has rank less than or equal to k-2.

It is interesting to relate the occurrence of multicollinearity to the hyperplanes of the model set identified in Lemma 2. In that lemma, we noted that the model set could be divided into n(k-1)-dimensional hyperplanes N(a), each representing a particular economic structure as indicated by the parameter vector a. What are the multicollinear points within one such hyperplane? They are the points where the hyperplane N(a) intersects any another hyperplane N(b) where $N(a) \neq N(b)$.

Furthermore, the multicollinear points in N(a) can be

characterized using what we know about the model set within the sample space. The set N(a) can be viewed as a sample space of dimension n(k-1). The multicollinear points within N(a) are matrices with rank less than or equal to k-2. Applying lemma 4 on the dimension of the model set in a sample space, we find that the dimension of the set of multicollinear points in N(a) is equal to (n+1)(k-2).

The following lemma highlights the main problem posed by multicollinearity, namely that it is impossible to pick one parameter vector for a matrix whose columns fit more than one equation.

Lemma 7

Each multicollinear point fits an infinite number of different equations. The parameter vectors of these equations form a circle contained in the (k-1)-dimensional parameter sphere.

Proof

Let X be multicollinear with $Xa = \emptyset$ and $Xb = \emptyset$, and with the vectors **a** and **b** normalized so the $a'a = 1$ and $b'b = 1$.

Take any vector in the parameter set of the form $ra + sb$.

$$\begin{aligned} \text{Then } X(ra + sb) &= X(ra) + X(sb) \\ &= rX(a) + sX(b) \\ &= r\emptyset + s\emptyset \\ &= \emptyset \end{aligned}$$

So $\tilde{X}$ fits any model with a vector of parameters $ra + sb$. The set of all vectors $ra + sb$ is the plane spanned by the vectors **a** and **b**. Recalling that the parameter space is a k-1 dimensional sphere in R^k, we see that this sphere intersects the plane spanned by **a** and **b** in a circle centred at the origin and including the points **a** and **b**. The circle is a one-dimensional set contained in the k-1 dimensional parameter sphere.

The following lemmas describe other geometric properties of the collinear set.

Lemma 8

The collinear set consists entirely of lines through the origin.

Proof

If X is in the collinear set, then $Xa = \hat{\emptyset}$ and

$Xb = \emptyset$ for $a \neq rb$. The line through the origin and through X consists of points αX, $-\infty < \alpha < \infty$. Since $Xa = \emptyset$ and $Xb = \emptyset$, we have:

$$(\alpha X)a = \alpha(Xa) = \alpha(\emptyset) = \emptyset \quad \text{and} \quad (\alpha X)b = \alpha(Xb) = \alpha(\emptyset) = \emptyset$$

So every point on the line through X and the origin is in the collinear set.

Lemma 9

The collinear set is composed of $nx(k-2)$ hyperplanes. Each hyperplane is the null space of a kx2 parameter matrix.

Proof

Let X be any point in the collinear set. Then $Xa = \emptyset$ and $Xb = \emptyset$ where a and b are non-zero, independent vectors of parameters. The kx2 matrix [ab] is a linear transformation from $R^n xk$ to $R^n x2$. Since $Xa = \emptyset$ and $Xb = \emptyset$, $X[ab] = [\emptyset,\emptyset]$ when $[\emptyset,\emptyset]$ is an nx2 matrix of zeros. The null set of this transformation is the set of points in the collinear set that are mapped onto the zero matrix, that is $N(a,b) = \{Y \in C \mid Y[ab] = \emptyset\}$

Recalling that the dimension of the null set is the dimension of the domain minus the dimension of the range, we see that:

$$\dim N(a,b) = (nxk) - (nx2) = nx(k-2).$$

Many other characteristics of the collinear set can be proved using approaches similar to those used in the previous section to describe the model set. For example, we could show that the collinear set is connected and non-linear. These results are not very useful for this research and will not be pursued. However, it is important to know how big the collinear set is within the model set. This result is dealt with in the following lemma.

Lemma 10

The collinear set has dimension $(n+2)(k-2)$.

Proof

In order to prove this result, we must show that there is a continuous, one-to-one function from the neighbourhoods of the collinear set to $R^{(n+2)(k-2)}$. The construction of this function is similar to the function used to determine the dimension of the model set.

Let $\underline{X}$ be a point in the collinear set. Assume for the moment that the vectors $X_3, \ldots, X_k$ are independent. Then, since the column rank of X is k-2, the columns X_1 and X_2 can be written as linear combinations of the columns $X_3, \ldots, X_k$: $X_1 = \sum_{i=3}^{k} a_i X_i$ and $X_2 = \sum_{i=3}^{k} b_i X_i$.

The (k-2)-dimensional parameter vectors $\mathbf{a} = (a_3, \ldots, a_k)$ and

$\mathbf{b} = (b_3, \ldots, b_k)$ are solutions to the matrix equation:

$[X_3 \ldots X_k][\mathbf{ab}] = [X_1\ X_2]$

Premultiplying this equation by elementary matrices that reduce $(X_3 \ldots X_k)$ to row-echelon form gives:

$$\begin{bmatrix} & \\ & I_{k-2} \\ & \\ & 0 \end{bmatrix} (\mathbf{ab}) = P[X_1\ X_2]$$

$$\begin{bmatrix} \mathbf{a} & \mathbf{b} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} = P(X_1\ X_2)$$

where P is the product of elementary matrices determined by the entries in $[X_3 \ldots X_k]$.

We can now form the function:

$g(X) = (\mathbf{a}, \mathbf{b}, X_3, \ldots, X_k)$

This function takes the first two columns out of the matrix X and replaces them with the two (k-2)-parameter vectors for these columns written in terms of the other k-2 columns of X. The range of the function has points consisting of an nx(k-2) matrix and two k-2 vectors. The total dimension of the range-space is

$nx(k-2) + 2(k-2) = (n+2)(k-2)$, as specified in the lemma.

The function g(X) is surjective, since every set of

columns (X_3, ..., X_k) and every linear combination of these columns corresponds to some point in the collinear set.

The function g(X) is continuous, since its components are continuous. The parameter vectors **a** and **b** are continuous functions of the entries in X, as we saw in the analysis of the products of elementary matrices (Lemma 4). The rest of the function is the (continuous) identity function.

The function is also one-to-one. Consider the inverse of g(X):

$$g^{-1}(\mathbf{c}, \mathbf{d}, Y_3, \ldots, Y_k) = \left(\sum_{i=3}^{k} c_i Y_i, \sum_{i=3}^{k} d_i Y_i, Y_3, \ldots, Y_k\right)$$

where **c** and **d** are in R^{k-2} and Y_3, ..., Y_k are in R^n.

Applying this to g(X) gives:

$$\begin{aligned} g^{-1}(g(X)) &= \left(\sum_{i=3}^{k} a_i X_i, \sum_{i=3}^{k} b_i X_i, X_3, \ldots, X_k\right) \\ &= (X_1, X_2, X_3, \ldots, X_k) \\ &= X \end{aligned}$$

The function g(X) specified above does not apply to all collinear points. If the first two columns of X are not linear combinations of the other columns, or if the matrix P contains discontinuities, we can adjust the function g(X) as we did for the function f(X) in Lemma 4. The only cases where g(X) cannot be adjusted to give a continuous, one-to-one correspondence with $R^{(n+2)(k-2)}$ are points where the column span of the matrix X is less than k-2. These are points of

extreme multicollinearity, which will be examined next.

Before passing on, we will note the extension of the preceeding lemma to manifolds. The proof of this extension is not included, since it is very similar to the case of the model set, as shown in Lemma 5.

Lemma 11

The set of nxk matrices with column rank equal to k-2 is a manifold of dimension (n+2)(k-2).

Degrees of Multicollinearity

All nxk matrices where $n \geq k$ can be classified according to how many equations $Xa = \emptyset$ they fit:

If the matrix has rank equal to k, it fits only the trivial equation $X\emptyset = \emptyset$.

If the matrix has rank k-1, it fits a one-dimensional set of equations corresponding to a line in R^k spanned by one parameter vector **a**. This is the case of non-multicollinear points in the model set, which has been examined in detail.

If the matrix has rank k-2, it fits a two-dimensional set of equations corresponding to a plane in R^k spanned by two independent parameter vectors **a** and **b**. This is the case of

most multicollinear points, as has just been discussed. This condition could be called first-order multicollinearity. It can be corrected by removing one of the multicollinear columns from the matrix.

If the matrix has rank k-3, it fits a three-dimensional set of equations corresponding to a hyperplane in R^k spanned by three independent parameter vectors. This condition could be called second-order multicollinearity. It can be corrected by removing two of the multicollinear columns from the matrix. We can continue in this manner to classify increasing degrees of multicollinearity, up to the point where the matrix has rank one. If the matrix has rank one, it fits a k-1-dimensional set of equations corresponding to a k-1 dimensional hyperplane in R^k spanned by k-1 independent parameter vectors. The only nxk matrix that fits k independent equations is the zero matrix, which has rank zero.

The geometry of a multicollinear point tells us something about which variable or variables need to be removed to correct for multicollinearity. In order to eliminate multicollinearity, a column vector that is linearly dependent on the other columns must be removed. This will reduce the number of variables in the matrix without reducing the rank of the matrix.

It does not matter which dependent variable is removed since they all span the same hyperplane. One must not remove a column vector that is independent of the other vectors, because this reduces the rank of the matrix. Independent column vectors can be recognized because their corresponding parameter entry is zero. We can be sure that a multicollinear matrix has no more than k-3 such columns since the rank of a multicollinear matrix is at most k-2.

Examples

A few examples of collinear sets with specific sample sizes and numbers of variables will now be examined.

The smallest of all possible regressions, where n=2 and k=2, can have no multicollinear points because it has only two variables. As specified in the theorem, the dimension of the collinear set is $(n+2)(k-2) = (2+2)(2-2) = \emptyset$.

The smallest regression with a collinear set is where n=3 and k = 3. The collinear set has dimension $(3+2)(3-2) = 5$. It is contained in a model set of dimension $(n+1)(k-1) = (3+1)(3-1) = 8$, which is in turn contained in a sample set of dimension $nxk = 9$.

Consider a regression in three variables with twenty observations. Here, the collinear set has dimension $(20+2)(3-2) = 22$, the model set has dimension $(20+1)(3-1) = 42$, and the sample set has dimension 60. We can see from this that the collinear set, although of considerable size in itself, is very small within the model set, which is itself very small in the sample set.

Now consider a regression with $n = 40$ and $k=5$, as examined in the examples of model sets. Here the sample space has dimension 200. It contains a model set of dimension $(40+1)(5-1) = 164$, which in turn contains a collinear set of dimension $(40+2)(5-2) = 126$.

Now let us examine how the sizes of the model sets and collinear sets respond to changes in sample sizes and variables.

As we would expect, increasing the number of variables in a model increases the size of the model set and the collinear set faster than the size of the whole sample set. If we add one variable to an nxk regression matrix, the sample space increases in dimension by n, but the model set increases in dimension by n+1 and the collinear set increases in dimension by n+2. In other words, the likelihood of collinearity

increases with the number of variables, other things being equal. If we add enough variables, the model set will end up filling the whole sample set.

Conversely, increasing the sample size of a regression increases the dimension of the sample set faster than the dimensions of the model set and collinear set. By increasing the sample size by one, the sample set increases in dimension by k, while the model set increases in dimension by k-1 and the collinear set by k-2. In other words, the likelihood of multicollinearity decreases with the sample size.

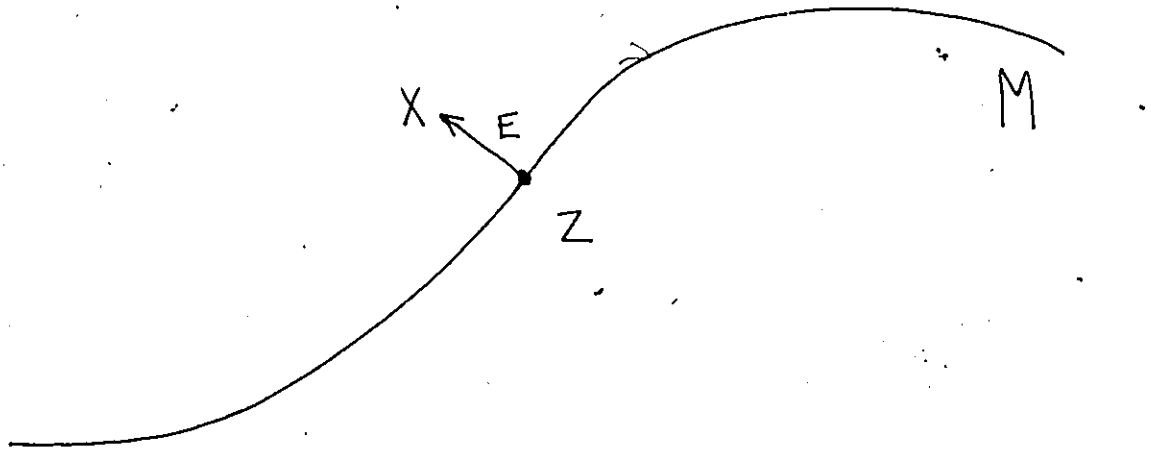
This concludes the examination of the model set and the collinear set, which are the points generated by the systematic variables in a regression model. The next step is to examine the geometry of the errors that are observed in addition to the systematic components.

5. The Error Space

In econometrics, we virtually never encounter an observed matrix in the model set because the observed matrix is almost always beset with errors. Geometrically, this means that rather than observing a point on the model set, we observe a point in the sample space that is not in the model set. We shall refer to the set of points not belonging to the model set as the error space R. In set terminology, we could write:

$$R = M^C$$

Any observed point X in the error space is generated by a point on the model set Z, and an error E : X = Z + E.



The problem of regression is to get back onto the model set by estimating and removing the error component. In order to do this, it is necessary to know what direction the error is likely to go. By knowing how the error is distributed, we can trace back to the point on the model set that is most likely to

have generated the observation.

The distribution of the error can be largely specified by its variance-covariance matrix. This matrix tells us what dimensions of the sample set the error occurs in, and what direction it is likely to point.

Let us write the nxk error matrix E as one long column vector by stacking the k column vectors of E one on top of the other:

$$E^C = \begin{vmatrix} E_1 \\ E_2 \\ \vdots \\ E_k \end{vmatrix}$$

The n x 1 column vector E_1 is the vector of errors associated with the variable in column one of the sample matrix; E_2 is the vector of errors associated with the variable in column two, and so on for E_3, E_4, ..., E_k. The variance-covariance matrix of the column vector E^C is the (n x k) x (n x k) matrix:

$$Q = \begin{bmatrix} \mathrm{Var}(E_1) & \mathrm{Cov}(E_1,E_2) & \cdots & \mathrm{Cov}(E_1,E_k) \\ \mathrm{Cov}(E_1,E_2)^T & \mathrm{Var}(E_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(E_1,E_k)^T & & & (\mathrm{Var}\ E_k) \end{bmatrix}$$

where $\mathrm{Var}(E_i)$, $i = 1,2,\ldots,k$ is the symmetric, n x n variance-covariance matrix of E_i with itself, and $\mathrm{Cov}(E_i,E_j)$, $i = 1,2,\ldots,k$, $j = 1,2,\ldots,k$ is the symmetric, n x n covariance matrix of E_i with E_j. In fully explicit form,

$$\mathrm{Var}(E_i) = \begin{bmatrix} \mathrm{Var}(e_{1i}) & \mathrm{Cov}(e_{1i},e_{2i}) & \cdots & \mathrm{Cov}(e_{1i},e_{ni}) \\ \mathrm{Cov}(e_{1i},e_{2i}) & \mathrm{Var}(e_{2i}) & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(e_{1i},e_{ni}) & \cdots & & \mathrm{Var}(e_{ni}) \end{bmatrix}$$

for $i = 1,2,\ldots k$, $j = 1,2,\ldots,k$.

Putting these sub-matrices all together gives the following matrix. It allows for any possible configuration of variances and covariances of an n x k matrix of error variables:

$$Q=\begin{bmatrix} \mathrm{Var}(e_{11}) & \cdots & \mathrm{Cov}(e_{11},e_{n1}) & \cdots & \mathrm{Cov}(e_{11},e_{1k}) & \cdots & \mathrm{Cov}(e_{11},e_{nk}) \\ \vdots & & \vdots & & & & \vdots \\ \mathrm{Cov}(e_{11},e_{n1}) & \cdots & \mathrm{Var}(e_{n1}) & \cdots & \mathrm{Cov}(e_{n1},e_{1k}) & \cdots & \mathrm{Cov}(e_{n1},e_{nk}) \\ \vdots & & \vdots & & \vdots & & \vdots \\ \mathrm{Cov}(e_{11},e_{1k}) & \cdots & \mathrm{Cov}(e_{n1},e_{1k}) & \cdots & \mathrm{Var}(e_{1k}) & \cdots & \mathrm{Cov}(e_{1k},e_{nk}) \\ \vdots & & \vdots & & & & \vdots \\ \mathrm{Cov}(e_{11},e_{nk}) & \cdots & \mathrm{Cov}(e_{n1},e_{nk}) & \cdots & \mathrm{Cov}(e_{1k},e_{nk}) & \cdots & \mathrm{Var}(e_{nk}) \end{bmatrix}$$

Before doing an analysis of the general case, it is useful to examine of few special cases in two-dimensional spaces to illustrate the geometric properties of the error and how they are indicated by the variance-covariance matrix.

The first characteristic of the error to consider is the dimension within which it may occur. In some econometric models, such as orthogonal least squares, there are errors associated with every variable in the observed matrix. In these cases, the error may occur in any dimension of the

sample space. In other cases, such as ordinary least squares, the error is only associated with a subset of the variables in the observed matrix. In these cases, the error may only occur in those dimensions of the sample space associated with those variables that have an associated error.

We can illustrate these two types of models in the two-dimensional case as follows. Although the two-dimensional case has no application in practice, the underlying structure applies to higher-dimensional cases.

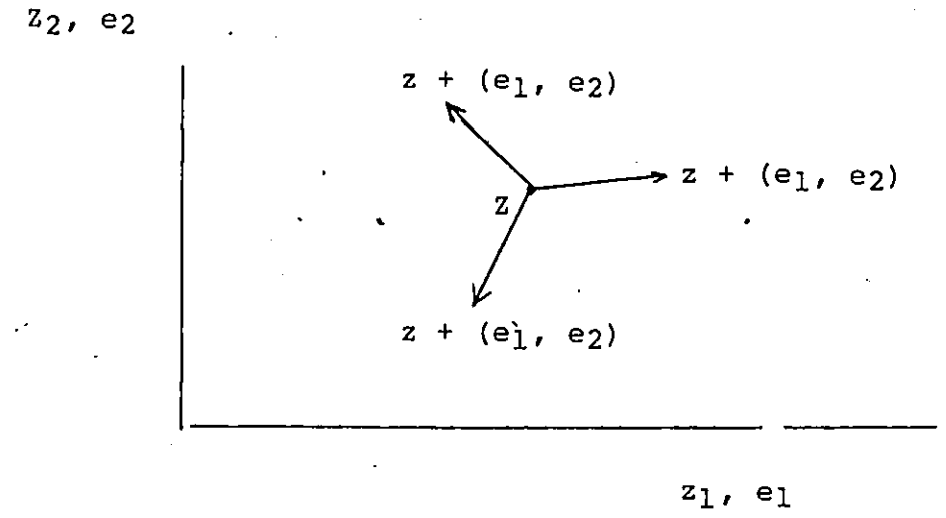
Suppose we have a two-dimensional sample space. Any observed point $X = (X_1, X_2)$ in the sample space is the sum of its systematic component Z and its error component E. The variance-covariance matrix of the error is the 2 x 2 matrix:

$$Q = \begin{bmatrix} \text{var } e_1 & \text{cov}(e_1, e_2) \\ \text{cov}(e_1, e_2) & \text{var } e_2 \end{bmatrix}$$

Let us assume that the errors are uncorrelated, so $\text{cov}(e_1, e_2) = \emptyset$. First consider the case where there is an error associated with both variables in the observed matrix, that is that var $e_1 \neq \emptyset$ and var $e_2 \neq \emptyset$. For the sake of simplicity, let var e_1 = var e_2 = 1. The variance-covariance matrix is:

$$Q = \begin{vmatrix} 1 & \emptyset \\ \emptyset & 1 \end{vmatrix}$$

Since the variances of both of the errors are more than zero, the error can project out from the systematic component Z in any direction within the sample space. Three possible errors are shown below:

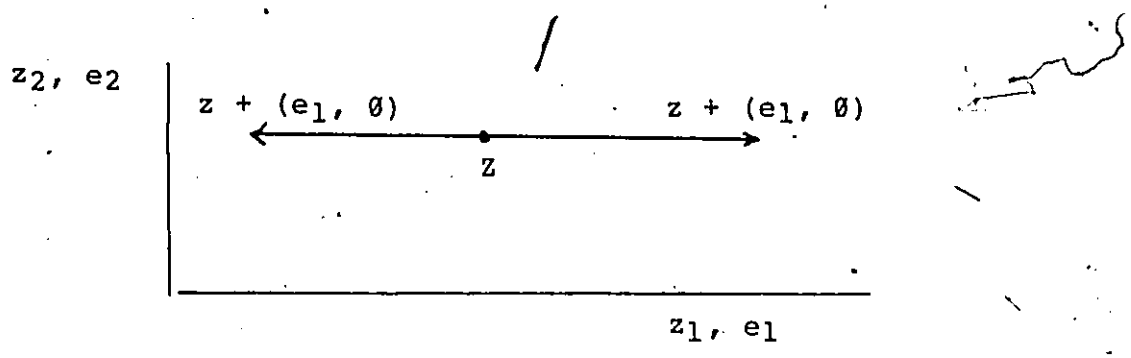


Now consider the case where the error is only associated with the first entry, but not the second, so var $e_1 = 1$ and var $e_2 = \emptyset$. The variance-covariance matrix is this case is:

$$Q = \begin{vmatrix} 1 & \emptyset \\ \emptyset & \emptyset \end{vmatrix}$$

Since the variance of the second error is zero, the error can only project out from the systematic component in the

dimension of the first error. Two possible errors are shown below:

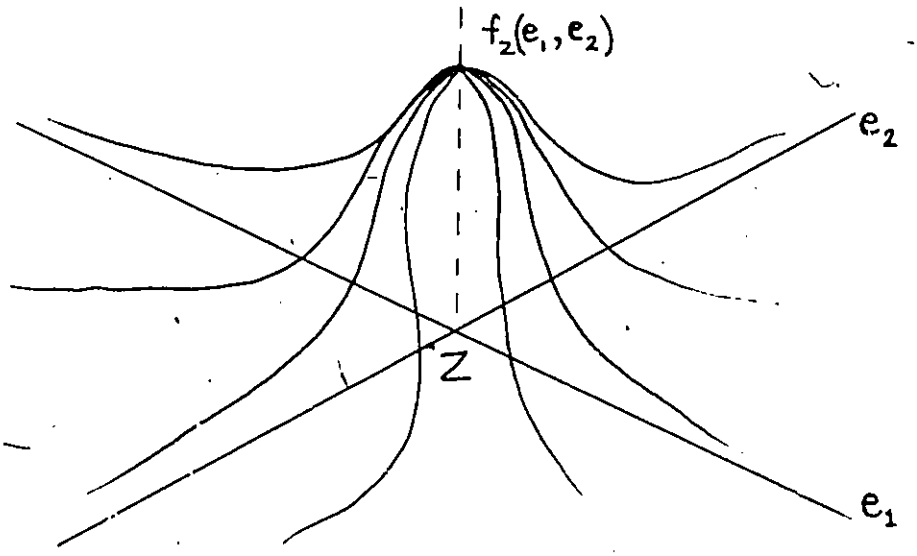


The distinction between the two cases illustrated above applies to regression models in the general nxk sample space. Some models, such as orthogonal least squares, assume that the error rooted at a systematic point can point in any direction within the sample space. Other models, such as ordinary least squares, assume that the error can only occur within particular subspaces of the sample space.

The second important characteristic of the error is the direction and length it is likely to take within whatever dimensions of the sample space it occurs. The two-dimensional case can be used to illustrate the main properties of the error. First consider the case of a two-dimensional error where the two error variables are identically distributed and uncorrelated. The variance-convariance matrix of the error is:

$$Q = \begin{vmatrix} \overline{1} & \overline{\emptyset} \\ \emptyset & 1 \end{vmatrix}$$

In this case the error is evenly distributed in all directions around the systematic point. Assuming that the errors are normally distributed, the density function would look like this:

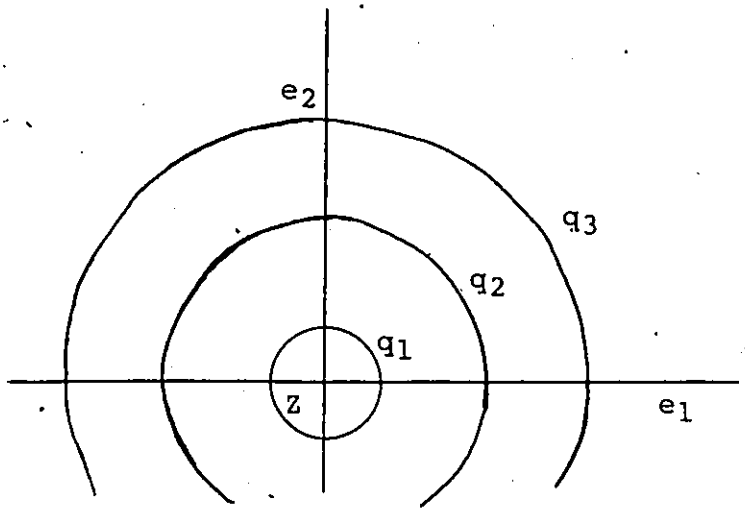


To simplify the exposition, we will present the density function through its level curves in the sample space. Let:

$$L_Z(q) = \{X = Z + E \mid f_Z(E) = q\}$$

be the level-q set of points around the point Z, where $f_Z(E)$ is the density function of the error E given the systematic point Z. This is the set of errors around Z having an equivalent chance of occurring. In the two-dimensional case, the level sets resemble the level curves of a topographical

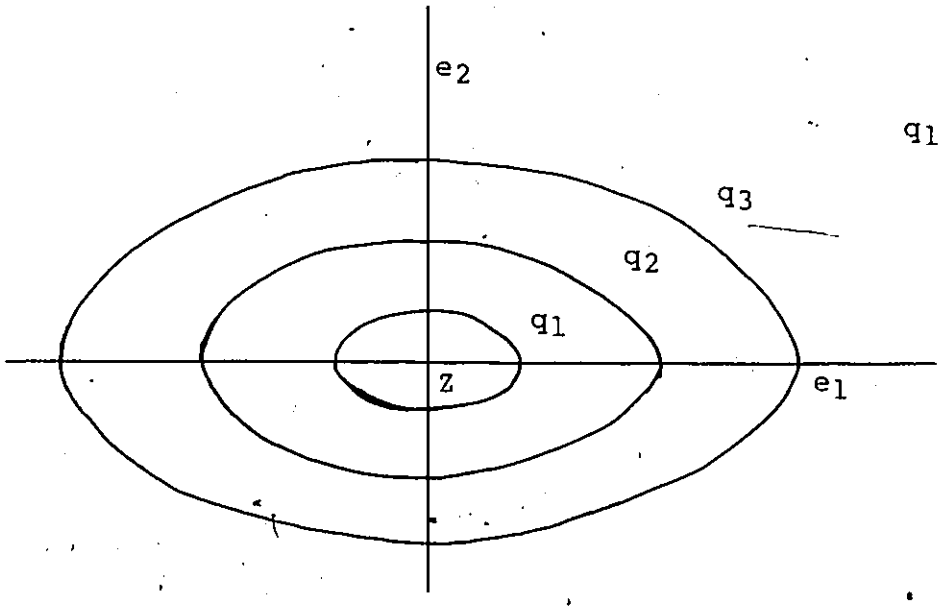
map of a mountain whose summit is at the point Z.



The lower the level curve, the lower the probability that the error will occur since larger errors have a smaller chance of occurring than smaller errors. Now consider the case where the variance of e_1 is larger than the variance of e_2 i.e. heteroskedasticity. Let the variance-convariance matrix be:

$$Q = \begin{vmatrix} 2 & \emptyset \\ \emptyset & 1 \end{vmatrix}$$

The level curves of this error would be an follows:

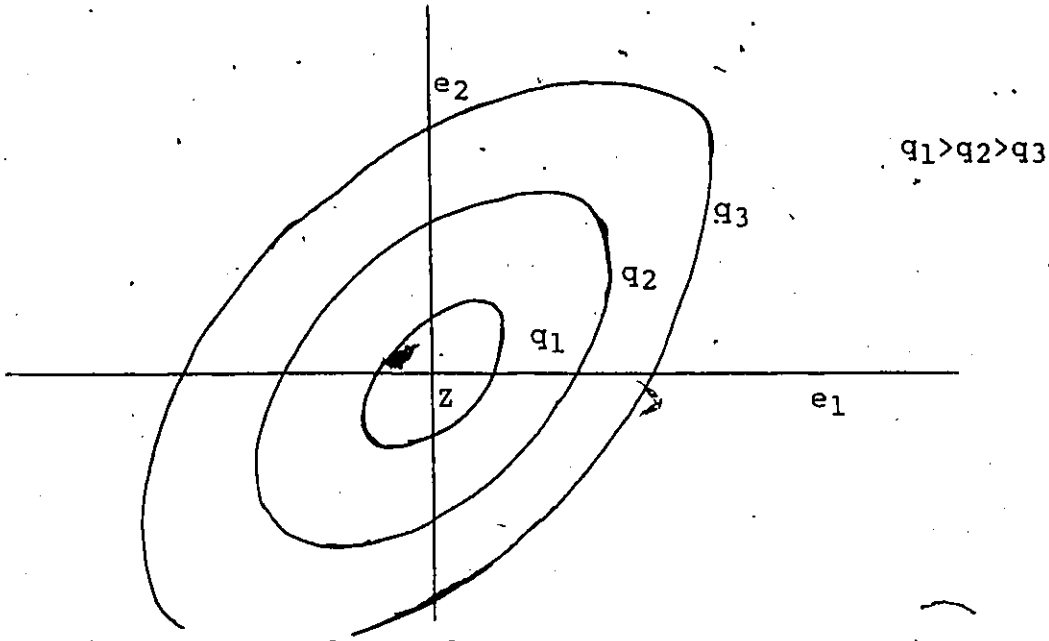


Since the first error is usually larger than the second with this form of distribution, the level curves will be ellipsoids stretched in the direction of the error component with the highest variance.

The case of models where the components of the error are correlated with each other gives another type of level curve (eg. serial correlation). Suppose the errors are positively correlated with each other. Let the variance-covariance matrix be:

$$Q = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

The level curves would be as follows:



Since large values of e_1 would tend to occur in conjunction with large values of e_2, the level curves would be ellipsoids stretched in the direction of the line $e_1 = e_2$. Similarly, if the errors are negatively correlated, the level curves would be ellipsoids stretched in the direction of the line $e_1 = -e_2$.

These simple two-dimensional examples illustrate the important characteristics of the error distribution in the general nxk-dimensional space. We will now examine the distribution of errors in the general case.

5.1 Ordinary Least Squares

Ordinary least squares assumes that the errors of the

independent variables have zero variance and that the errors associated with the dependent variable are independent and identically distributed. In terms of the matrix Q, assuming that X_1 is the dependent variable, we have:

$Var(E_1) = I_n$, $Var(E_2) = Var(E_3) = \ldots = Var(E_k) = \emptyset$ and $Cov(E_i, E_j) = \emptyset$ for all i, j. Putting this all together gives:

$$Q = \begin{bmatrix} 1 & \emptyset & \cdots & \emptyset & \cdots & \emptyset \\ \emptyset & 1 & & & & \\ & & & & & \\ & & & 1 & & \\ \emptyset & \cdots & \emptyset & 1 & & \\ & & & \emptyset & & \\ & & & & \emptyset & \\ & & & & \emptyset & \\ \emptyset & \cdots & & & & \emptyset \end{bmatrix}$$

In ordinary least squares the error occurs only in the n-dimensional subspace corresponding to the n entries in the first column of the sample matrix. Its density function is evenly distributed in all n directions within this subspace, so its level curves are n-dimensional spheres centred at the systematic component and imbedded in the nxk-dimensional

sample space.

5.2 Hybrid Model

The hybrid model, where the first j variables ($1 < j < k$) are assumed to have associated errors, all of which are independent and identically distributed, gives the matrix Q as follows:

$$Q = \begin{vmatrix} i & \cdots & j & \cdots & k \\ I_n & \emptyset & & & \emptyset \\ \emptyset & & & & \\ & & & & \\ & & I_n & & \\ & & & \emptyset & \\ & & & & \\ \emptyset & & & & \emptyset \end{vmatrix}$$

In the hybrid model, the error can occur in the nxj ($j < k$) dimensions of the sample space associated with the first j columns of the sample matrix. Its density function is distributed evenly in all nxj directions within this subspace, so its level curves are nxj-dimensional spheres centred at the systematic component and imbedded in the nxk dimensional sample space.

5.3 Orthogonal Regression

Orthogonal Regression assumes that all of the entries in

the sample matrix have an associated error variable, all of which are independent and identically distributed. In terms of the matrix Q, this gives Var (E_1) = Var(E_2) = ... = Var(E_k) = I_n and Cov(E_i, E_j) = Ø for all i, j. Putting these sub-matrices together gives:

$$Q = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{bmatrix}$$

which is the identity matrix $I_n$xk. In orthogonal least squares, the error can occur in all nxk dimensions of the sample space. Its density function is distributed evenly in all nxk directions within the sample space, so its level curves are nxk-dimensional spheres centred at the systematic component.

5.4 Generalized Orthogonal Regression

The standard model for orthogonal regression assumes the error variables are all independent and identically distributed. This model can be generalized to accommodate cases where the variances differ between variables or where the errors are correlated to each other. The variance-covariance matrix Q is adjusted accordingly with the appropriate variances and covariances.

Let us examine the case where each economic variable in the model is assumed to be associated with an error variable having a distinct variance. (The error variables are assumed to have zero covariance.) For example, consider an orthogonal regression model of the relationship between consumption, permanent income and the price index. We could reasonably assume that the variance of the error variable associated with the price index is smaller than the variance associated with consumption, since absolute fluctuations in consumption levels are far greater than absolute fluctuations in price levels. We could also reasonably assume that the variance associated with permanent income is higher than the variance associated with consumption because measurements of a hypothetical variable like permanent income are bound to be less accurate than measurements of something more concrete like consumption.

Given a model of this type, let σ_1^2, σ^2_2, ..., σ_k^2 be the variances of the errors associated with the K variables X_1, X_2, ..., X_k. Then the variance-covariance matrix of the model is

$$Q = \begin{bmatrix} \sigma_1^2 & & & & \\ & \sigma_1^2 & & & \\ & & \sigma_2^2 & & \\ & & & \sigma_2^2 & \\ & & & & \sigma_k^2 \\ & & & & & \sigma_k^2 \end{bmatrix}$$

where all entries off the main diagonal are zero.

In this case, the level curves are nxk-dimensional ellipsoids whose longest axes are in the dimensions of the highest variance, and whose shortest axes are in the dimensions of the lowest variance.

If the variance of one variable, say X_1, is assumed to be relatively greater than the variances of the other variables, then the value σ_1^2, will be relatively greater than the other

variances σ_2^2, σ_3^2, ...,σ_k^2. If we push this case to the limit, where the variances σ_2^2, σ_3^2, ...,σ_k^2 tend to zero, we arrive at the variance-covariance matrix of ordinary least squares, where the errors of the model are all absorbed by one variable. In other words, Ordinary Least Squares is the limiting case of a particular type of generalized orthogonal regression. (Malinvaud (1970) p. 42; Pollock (1979), p. 76) We can imagine the nxk-dimensional ellipsoids of generalized orthogonal regression being squeezed flatter and flatter until they are the simple spheres of ordinary least squares, like pumpkins that are squeezed until they are as flat as a pancake.

There are many other ways to generalize the orthogonal regression squares model. For example, one can assume that the errors associated with a particular variable are not all identically distributed, and that some have larger variances than others. One can also assume that the errors are correlated in some way with each other. The assumptions of these models are accommodated by inserting the appropriate parameters in the variance-covariance matrix Q.

5.5 Generalized Least Squares

Generalized least squares can be represented by letting all of the variances and covariances in the matrix Q be zero

except for those in the sub-matrix Var (E_1). We can accommodate any set of assumptions about the errors associated with the dependent variable by inserting the variances and covariances in their appropriate places in the matrix Q.

For example, consider the generalized least squares model with heteroskedastic error variables, where the variances of the n error variables are $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. The variance-convariance matrix of the whole sample matrix would be:

$$Q = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{bmatrix}$$

where the blank entries are zeros. The level curves in this case are n-dimensional ellipsoids imbedded in the nxk-dimensional sample space and stretched to reflect the n different variances.

The generalized least squares model for first-order serial correlation where the coefficient of correlation is p gives the following variance-covariance matrix:

$$Q = \begin{bmatrix} 1 & p & p^2 & \cdots & p^{n-1} \\ p & 1 & p & & p^{n-2} \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ p^{n-1} & & & & 1 \end{bmatrix}$$

where the blank entries indicate zeros. The level curves in this case are n-dimensional ellipsoids which are stretched mainly in the direction of sequential values, i.e. along the axes $x_{11} = x_{12}$, $x_{12} = x_{13}$, $x_{13} = x_{14}$, ..., $x_{1,n-1} = x_{1,n}$, because the highest degree of correlation is between these values as indicated by the parameter p in the matrix Q. The ellipsoids are also stretched secondarily in the directions $x_{11} = x_{13}$, $x_{12} = x_{14}$, ..., $x_{1,n-2} = x_{1,n}$ reflecting the effect of the parameter p^2, and so on for the lower correlations of further-removed values.

The main point of discussing these examples is to show that all possible assumptions about a regression model, from the choice of ordinary least squares or orthogonal regression, to the choice of the dependent variable, to the variance and covariances of the errors, can be reduced to the specification of one matrix. This matrix indicates the dimensions within

the sample space where the error occurs and the direction it is likely to go within these dimensions.

One can often work backwards from a matrix Q to the assumptions that it represents: If the matrix Q has only zeros off the main diagonal, then the model assumes all the error variables to be independent. If the matrix Q has zeros in the block corresponding to the ith variable, then the model assumes that the ith variable is measured without error.

However, the matrix Q only tells us about the dispersion of the errors in the model; it does not tell us why the errors are where they are. For example, if the matrix Q has zeros in all blocks except the one corresponding to the first variable, then the model assumes that the entire weight of adjustment for the errors will fall on the first variable. It does not tell us why the errors are only associated with this variable. It could be because this variable is the dependent variable in an ordinary least squares regression, or because this variable is the only one assumed to be measured with error in an errors-in-variables regression, or both.

In spite of this, the matrix Q gives enough information to estimate the model, whatever the underlying assumptions.

6. The Estimation Function

The preceding sections have examined, from a geometric point of view, how a sample point is generated. It is the sum of a systematic point in the model set and a random component whose direction and length depends on the variance-covariance matrix of its distribution. We will now deal with estimation, which tries to untangle a sample point into its systematic and random parts.

Every regression estimation technique is essentially a procedure that takes a sample matrix, performs a series of calculations on it, and provides an estimate of the error-free component. This procedure can be viewed as a function whose domain is the sample space and whose range is the model set. We will refer to this procedure as the estimation function. The estimate of the systematic component of a sample point X will be denoted in the usual way as $\hat{X}$.

It should be kept in mind that $\hat{X}$ is not the only estimate that is a function of X. The estimate of the parameter vector and the estimate of the error are also functions of the sample point.

The estimate of X is the solution to a problem in optimal programming under a constraint condition, as is common in many

branches of economics. In regression, the objective function is the distance between the sample point and the estimate. This distance is to be minimized. The estimate is constrained to belonging to the model set. In general terms, the estimation function can be expressed as follows:

$$\hat{X} = Z \in M \text{ where } D(X,Z) < D(X,Y) \text{ for any } Y \in M,\ Y \neq Z.$$

This states that the estimate of X is the point Z in M whose distance from X, D(X,Z), is smaller than the distance from X to any other point in M.

6.1 The Distance Function

The distance between the sample point and the estimate is the key element of the estimation function. It is the way these distances are measured that varies from one regression technique to another. This is because different assumptions about the error contained in the sample point have different implications about what distance should be minimized.

In this section, we will examine why the minimum distance criteria is used and what particular distance function should be matched to a particular error distribution. First, we will examine what a distance function is, and look at a few common examples.

A distance is a function whose domain is the Cartesian product of a set with itself and whose range is the set of non-negative real numbers. It is usually assumed to have the following properties for all points a and b in the set on which it is defined (Fréchet, 1953, pp. 13-14):

1) $D(a,b) = D(b,a)$

2) $D(a,b) = 0$ if and only if $a = b$

3) $D(a,b) \leq D(a,c) + D(c,b)$ ("triangle inequality")

4) $D(a,b) \geq 0$

A function with these properties is sometimes called a metric.

In vector spaces, the most common distance is the Euclidean metric, defined as $D(a,b) = \sqrt{(a-b)'(a-b)} = \sqrt{\sum(a_i-b_i)^2}$ where a and b are column vectors in the same vector space. In regular three-dimensional space, this distance is the same as the distance between a and b as measured by a ruler.

Another common distance is the generalized Euclidean metric defined as $D(a,b) = \sqrt{(a-b)'Q(a-b)}$, where Q is a fixed full-rank positive definite matrix. If Q is the inverse of the variance-covariance matrix of the random variable (a-b),

this distance is sometimes called the Mahalanobis distance (Sastry and Tintner, 1977).

In order to ensure the distances are positive, the inputs are usually squared, as in the Euclidean metric and the generalized Euclidean metric. An alternate approach is to take the absolute value of the inputs. This is sometimes used in economic applications where the error variables are assumed to have non-finite variances, since squaring tends to put exaggerated importance on larger values. The absolute value distance between two points in n-space is defined as:

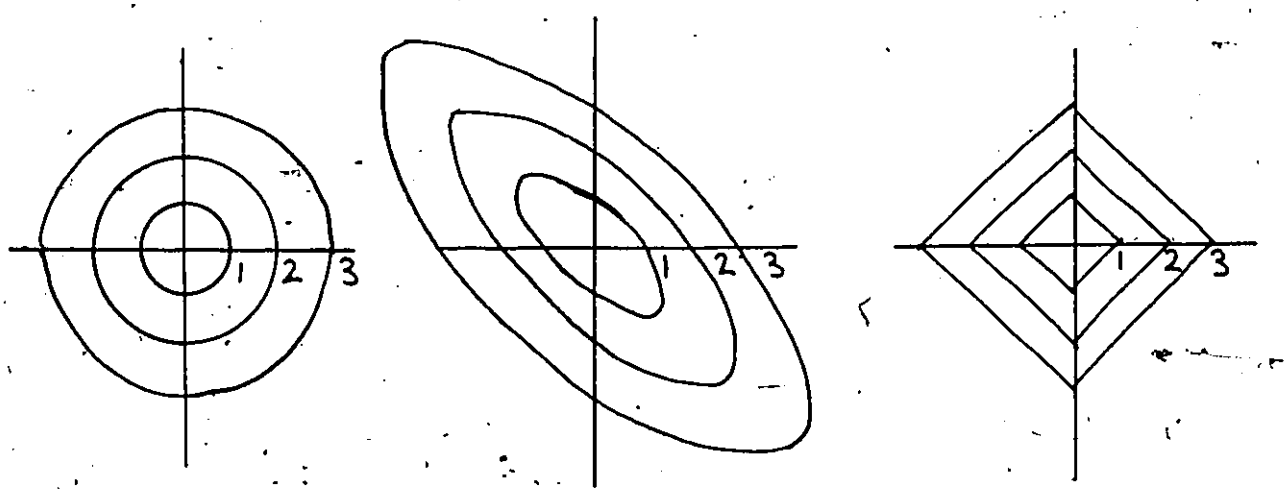
$$D(a,b) = \sum_{i=1}^{n} |a_i - b_i|$$

A distance can be used to define level curves around a point a by taking the set of points in the set equidistant to a:

$$L_a(d) = \{b | D(a,b) = d\}$$

Geometrically, the level curves of the Euclidean metric are concentric hyperspheres around the given point. The level curves of the generalized Euclidean metric are concentric ellipsoids around the given point. The axes of the ellipsoids are determined by the matrix Q. The level curves of the absolute value metric are regular polyhedrons. The following illustrates these three common cases for three level curves

$L_a(1)$, $L_a(2)$, and $L_a(3)$.



Euclidean Metric

Generalized Euclidean Metric with:

Absolute Value Metric

$$Q = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

Regression estimation minimizes the distance between the observed point and the model set. In this problem, the actual value of the distance does not matter in finding the minimum; only the order of the distances matters. In regression, the

most common distance used is the square of the Euclidean metric, from which we get the expression least squares. Squaring the Euclidean metric leaves the order and shape of the level curves of the distance function unchanged, so it does not change the minimum value.

Using the square of the Euclidean distance has the odd property of violating the "triangle inequality" that is commonly associated with distance. However, this has no consequence on the estimation function of a regression, since regression is only concerned with the distance between a particular sample point and a particular point in the model set.

Distance in some metrics is closely related to the notion of length. As we mentioned earlier, the Euclidean metric from point a to point b is identical to the length of the line segment $\overline{ab}$. In regression, if a standard least squares metric is used, minimizing the sum of the squares is equivalent to minimizing the length of the line segment from the observed point to the estimated point on the model set.

6.2 Why Minimize the Distance?

We will now address the question of why an estimation function should minimize the distance between the observed

matrix and the model set. This section uses the geometric shapes of the model set, the level curves of the distance function and the level curves of the density function to show why a minimum distance rule provides a solution to the problem of regression estimation.

The purpose of regression estimation is to find the systematic matrix most likely to have given rise to the observed matrix. The use of a minimum distance criterion for this purpose may be justified with or without reference to a probability space of random variables.

The minimum distance criterion can be justified by simply appealing to intuition and common sense. If we want a matrix X to fit onto an equation of the form $Za = \emptyset$, while maintaining whatever structure underlies X as intact as possible, then X should be changed as little as possible. Thus we should simply find the point $\hat{X}$ that is closest to X yet that also fits some linear equation $\hat{X}a = \emptyset$. As stated by Malinvaud (1970, p. 7) in a chapter on "Econometrics Without Stochastic Models":

> It is obviously desirable that the average distance between the points [i.e. observations] and the line should be minimized... [This] common sense criterion is completely defined only when we have shown how to measure a distance between a point and the line, and how to calculate the average of the distances from the different

points.

> A quadratic mean is generally chosen, since such a formula involves the simplest calculations. So we look for a line such that the sum of the squares from the points to the line should be as small as possible... [Several] different measures of distance can be, and in fact are, used...

This justification has the advantage of being simple and avoiding the need for knowledge of the distribution function of the error. It has the disadvantage of offering no help in choosing the most appropriate distance criteria, as well as offering no insight into why the data do not fit the equation or how accurate the estimate may be. For these reasons, regression estimations are usually based on assumptions about the probability space underlying the observed matrix.

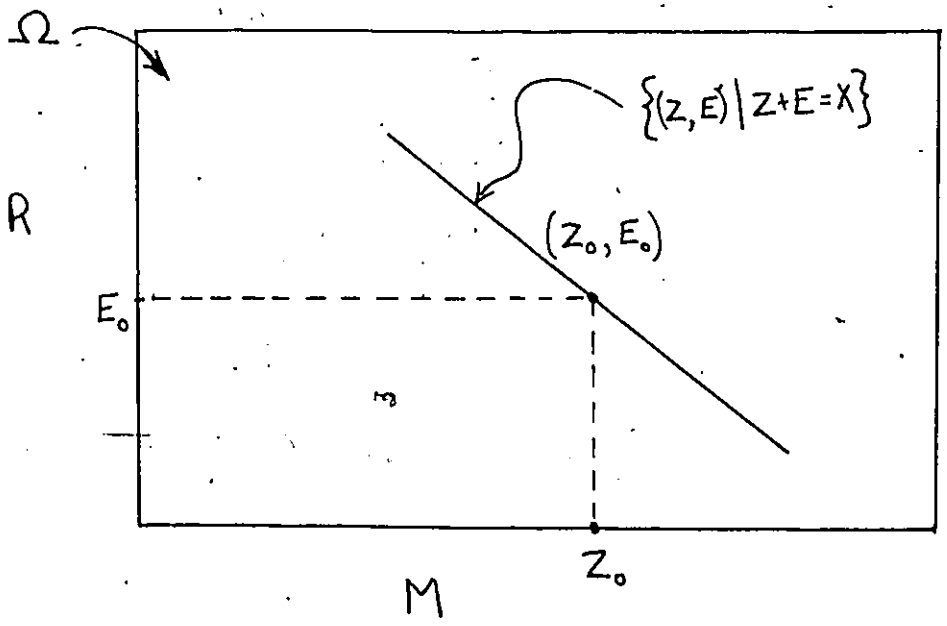
These assumptions generally include knowledge of the distribution function and density function of the error variables, for example that they are normally distributed with variances and covariances that are known or can be estimated.

Usually regression estimation techniques are justified by the properties of their parameter estimates. For example, likelihood functions maximize the probability of a parameter estimate being the true parameter value given a particular sample point.

The following discussion will focus on the estimate of

the variable matrix as opposed to the estimated parameters of the equation, since this paper is mainly concerned with the geometry of the sample space.

When we observe a matrix of data X, we are implicitly observing the sum of a systematic component Z in the model set, and an error E emanating from the point Z. Let us represent this underlying event as the couplet (Z,E). The universe of all possible events (Z,E) is the set $\Omega = M \times R$, that is the cross product of the model set with the set of errors that could be rooted at any systematic point. The observed matrix X could have been produced by any point (Z,E) in Ω such that Z+E=X. The set of all such points forms a subset of the universe of events Ω. We can represent Ω by the following diagram.



In order to find the Z most likely to have caused the observation X, we consider the distribution of events in the probability space Ω.

The assumptions of the model tell us the distributions of the error variables based at each systematic point Z in the model set and their density functions, as outlined in Section 5. In the context of the space Ω, these are marginal density functions of E given a particular Z, which we denote $f_Z(E)$.

The assumptions of the model implicitly assume that all of the systematic points in the model set have equal chances of being the true systematic component, at least within certain areas of the model set. Putting this together with what is assumed about the marginal density of E given a particular value of Z, the probability density function over the space Ω is of the form:

$$f(Z,E) = \frac{1}{k} f_Z(E), \text{ k constant.}$$

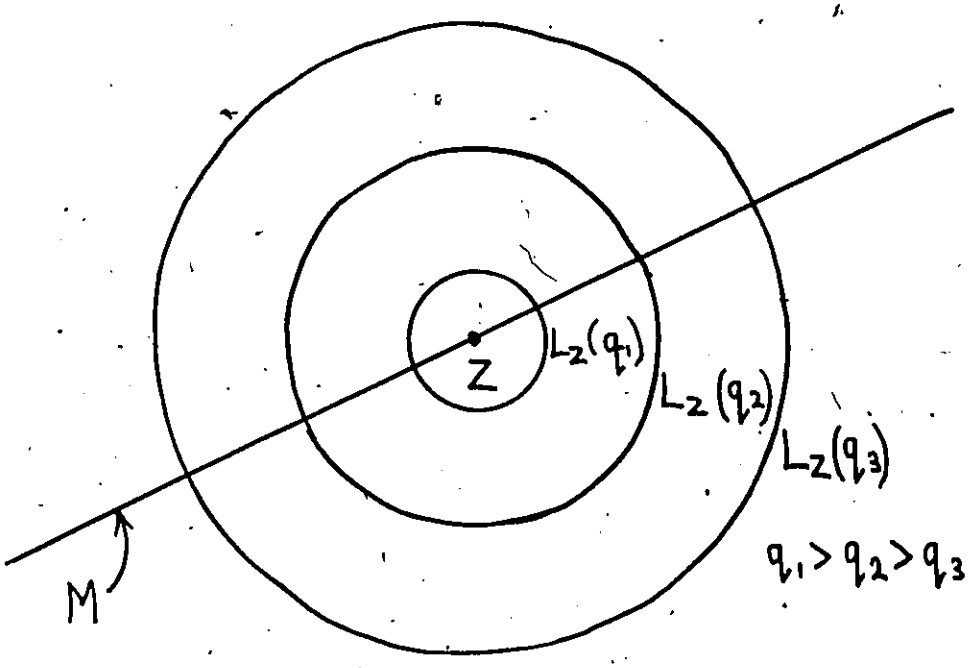
We want to find the point Z that maximizes the probability density function given that a particular point X has been observed. This is the conditional probability function of E=X-Z given X. The density function of this conditional probability is:

$$\frac{\frac{1}{k} f_Z(X-Z)}{\int_{((Y,E)|Y+E=X)} \frac{1}{k} f_Z(X-Y)\,dw}$$

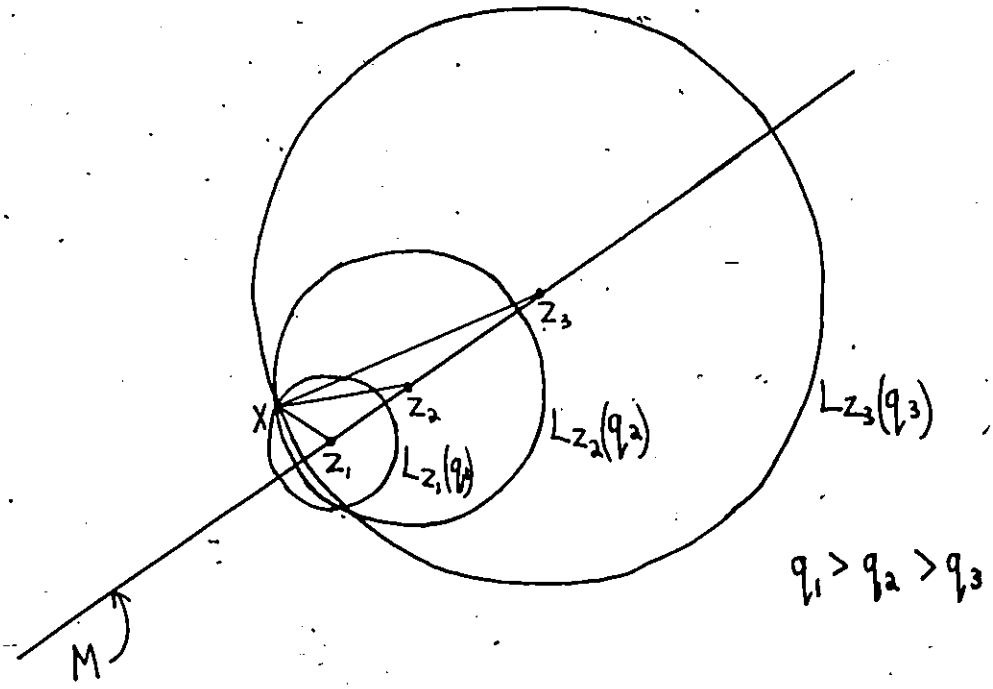
where the integral is understood to be multi-dimensional, and to integrate over all points (Y,E) in Ω such that Y + E = X. This expression can be simplified by dropping the constant term 1/k from both numerator and denominator. Next we note that the denominator is a constant, given any particular X. Given this, the maximum of this expression occurs where $f_Z(X-Z)$ is a maximum. Thus we have reduced the estimation problem to one of finding:

$$\underset{Z \in M}{\text{Max}}\ f_Z(X-Z)$$

Let us recall that the density function of the error based at a point Z in the model set can be represented in the sample space by a series of level curves $L_Z(q) = \{X = Z + E | f_Z(E) = q\}$. The value $f_Z(X-Z)$ is simply the index of the level curve based at Z that passes through the point X:



Maximizing the value $f_Z(X-Z)$ is the same as maximizing the index of the level curve passing through X over all base points in the model set. The level curves around a point Z are nested such that higher levels are closer to the point Z. This situation can be represented in the following diagram:



Now suppose a distance function is defined such that its level curves around Z coincide with the level curves of the density function of errors around Z. Then maximizing the level of the density function's curve is the same as minimizing the level of the distance function's curve. This is because the levels of the density function are monotone increasing as they approach Z, while the levels of the distance function are monotone decreasing as they approach Z. Maximizing the value $f_Z(X-Z)$ is the same as minimizing the distance D(X,Z) if the distance function and the density function have the same level curves.

This exposition gives a very visual justification of how minimizing the distance to the model set leads from the error-ridden sample point back to the systematic point that is most likely to have generated it.

It is worth noting that the level curves of the distance function can be viewed as surrounding the sample point instead of surrounding the estimated point on the model set. The result of choosing the minimum distance is the same, because distances are defined such that $D(X,Y) = D(Y,X)$.

6.3 Which Distance Suits Which Model?

We will now examine how different distance functions reflect the assumptions of the regression model that they are used to estimate. The preceding discussion showed why the distance function should reflect the distribution of the errors emanating from a systematic point. As we saw in the discussion of error variables, the distribution of the errors can be largely represented by the matrix Q of variances and covariances between all elements in the error matrix. This section will look at how these characteristics are taken into account in the distance function.

6.3.1 Orthogonal Regression

First let us consider the case where the errors are assumed to occur in all dimensions of the sample space, as in orthogonal regression. If the error variables are all independent and identically distributed, the variance-covariance matrix Q is the identity matrix. In this case, the squared Euclidean metric is used because its level curves are hyperspheres, as are the level curves of the density function.

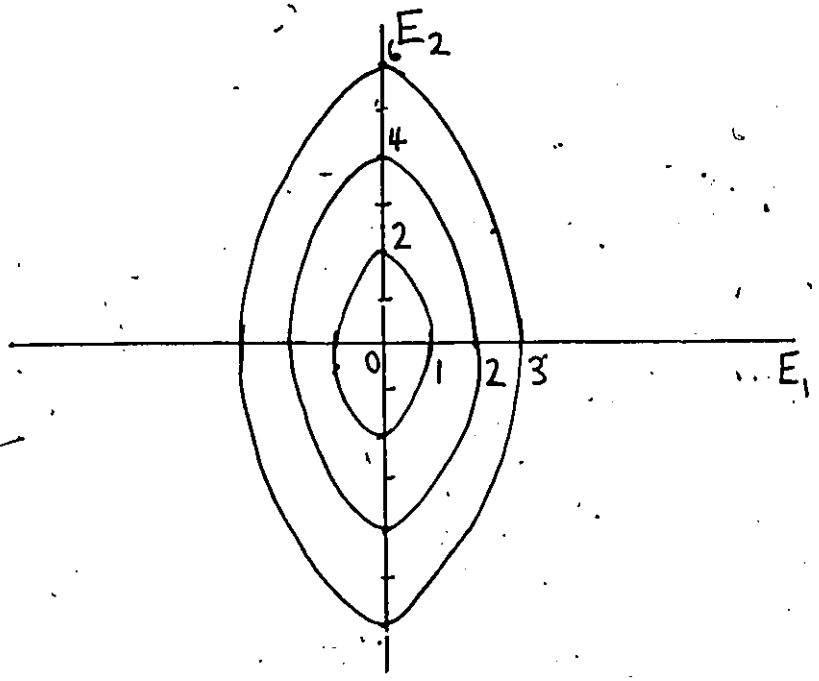
6.3.2 Generalized Orthogonal Regression

If the errors are not all independent and identically distributed, the level curves of the distance function can be made to correspond to the level curves of the density function by using the distance function $D(X,Y) = (X-Y)'Q^{-1}(X-Y)$. For

example, consider a two-dimensional normal error $E = (E_1, E_2)$ distributed with variance-covariance matrix

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

that is where the variance of the second variable is twice the variance of the first. Its level curves look like:



The distance $D(E,O) = (E-0)'Q^{-1}(E-0) = E'\begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} E = E_1^2 + 1/2 E_2^2$ has the same level curves. The effect of the matrix Q^{-1} is to reduce the weight given by the distance function to the error variable with the higher variance. Similarly, correlations between variables can be accounted for by weighting the distance function with Q^{-1}. In general, the result of using the distance $(X-Y)'Q^{-1}(X-Y)$ is to weight down errors

that tend to be larger so that errors with an equivalent chance of occurring are put on the same level curve of the distance function. This ensures them all of equal treatment during the estimation procedure.

6.3.3 Ordinary Least Squares

Now let us consider the case where the variances of all but the first column tend to zero. In the limit, this is the case of ordinary least squares estimation where the first column contains the dependent variable. The inverse of the matrix:

$$Q = \begin{bmatrix} I_n & & & 0 \\ & aI_n & & \\ & & \ddots & \\ 0 & & & aI_n \end{bmatrix}$$

is:

$$Q^{-1} = \begin{bmatrix} I_n & & & 0 \\ & \frac{1}{a}I_n & & \\ & & \ddots & \\ 0 & & & \frac{1}{a}I_n \end{bmatrix}$$

As a tends to zero, the inverse tends to:

$$Q^{-1} = \begin{bmatrix} I_n & & & 0 \\ & \infty I_n & & \\ & & \ddots & \\ 0 & & & \infty I_n \end{bmatrix}$$

The distance function $(X-Y)'Q^{-1}(X-Y)$ in this case is $(X_1-Y_1)'(X_1-Y_1) + \infty (X_2-Y_2)'(X_2 - Y_2) + \ldots + \infty(X_k-Y_k)(X_k-Y_k)$. In other words, this distance gives an infinite weight to the distance between any two matrices where the columns 2, 3, ..., k are not identical. If X is the observed matrix and Y is a point in the model set, this distance assigns an infinite distance to any matrix Y whose columns 2, 3, ..., k differ from the corresponding columns of X. In the context of the estimation procedures, it effectively excludes any such matrices from the set of possible estimates. This is exactly what we would expect from ordinary least squares estimation, since the independent variables in ordinary least squares are not adjusted for error, and the entire weight of adjustment is borne by the dependent variable.

The level curves of this distance function based at a point Y on the model set are n-dimensional spheres imbedded in the nxk-dimensional sample space, as are the level curves of the density function at Y.

We will now relate this geometric view of ordinary least squares in the nxk sample space to its traditional view in an n-dimensional vector space. Given an observed sample matrix, the only level curves of the density function through it are based at points on the model set that have the same values as X for all of the independent variables. The set of all such points is an n-dimensional affine linear subspace of the sample set:

$A(X) = \{W \in M \mid W=[W_1, 0, \ldots, 0] + [0, W_2, \ldots, W_k]$ where $W_2 = X_2, \ldots, W_k = X_k\}$

(An affine linear subspace is a linear subspace that does not pass through the origin.)

The first column of these matrices can be any vector in R^n.

Within the affine set A(X), there are many points that are also in the model set. Since the model set consists of all points that have linear dependence among the columns, and the column $X_2, \ldots, X_k$ are arbitrary vectors in R^n, the first

columns of W must be linearly dependent on the other columns of W in order for W to be in the model set. The set of all model points in A(X) is the set of points where the first column belongs to the span of the vectors X_2, ..., X_k transposed to the space R^n.

Once the observed matrix X has been chosen, the rest of the estimation process of ordinary least squares occurs within the affine space A(X). The estimate $\hat{X}$ is the matrix whose first column belongs to the subspace spanned by the other columns of X and which is of minimum distance to the first column of X. This is why ordinary least squares can be represented as occurring in R^n, as was discussed in Section 5. The minimum distance estimate of X_1 is the orthogonal projection of X_1 onto the span of the other columns of X in R^n.

The advantage of viewing ordinary least squares estimation in the larger nxk-dimensional sample space is that it allows us to visualize all possible ordinary least squares regressions on an nxk matrix simultaneously, and to compare this estimation technique with other estimation techniques. Once the sample point has been chosen, we can narrow the focus to the smaller subset of the sample space where the estimation occurs, which we have called the space A(X).

6.3.4 Generalized Least Squares

The geometric structure of generalized least squares is identical to that of ordinary least squares, except that the distance function applied within the sub-space A(X) is the generalized squared Euclidean metric rather than the standard squared Euclidean metric. As with the generalized case of orthogonal least squares, the distance function is weighted with the inverse of variance-covariance matrix, Q^{-1}, so that the level curves of the distance function will correspond to the level curves of the density function of the errors. Of course, the variance-covariance matrix applies only to the errors in R^n.

6.3.5 Absolute Value Metric

The absolute value metric is used in econometric estimation where the distribution of the error is assumed to have a large, even infinite, variance. The distribution function of the error in these cases is not generally specified, beyond the assumption that the variance is large. (Taylor, 1974). This assumption distinguishes methods using the absolute value metric from other econometric methods which are usually based on a normal distribution with a finite variance and a thin tail. The reason for favouring the absolute value metric over the least squares metric is that the least squares metric is very sensitive to large deviations because they are squared. Even a single large deviation in an ordinary least squares estimation

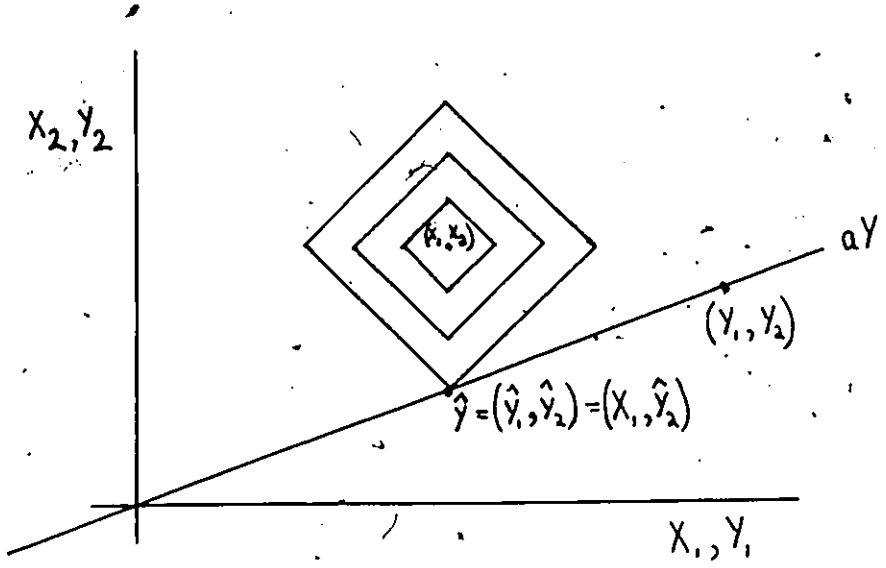
can perceptibly alter the estimated dependent variable and estimated parameter vector. The absolute value metric has the effect of being much less influenced by large deviations.

The absolute value metric has interesting geometric properties when applied to regression estimation. As we saw earlier, the level curves of the absolute value distance function are regular polyhedrons. Solutions to the estimation problem in this case are found by methods of linear programming, since the boundaries of the polyhedrons are portions of linear subspaces. The geometry of "least absolute value estimation"(LAE) is well described in the following passage by Taylor (1974, p. 188). It applies the absolute value metric to an ordinary least squares type of model with one dependent variable. In this passage, the polyhedrons of the distance function are described as surrounding the sample points in contrast to the approach used in most of this paper in which the level curves of the distance function surround the systematic point. However, as we saw in the discussion of distance functions the two approaches are equivalent since the distance from point a to point b is equal to the distance from point b to point a.

> ... If, for example, there were three observations, the polyhedron in question would have the shape of two identical regular four-sided pyramids butted base-to-base centred at the point $y' = (y_1, y_2,$

> y_3) and whose eight faces are isosceles triangles of equal area. If there is only a single independent variable, the LAE estimate will be given by the point on the line defined by the origin and the point $x' = (x_1, x_2, x_3)$ which is cut by the expanding "double pyramid." A moments (sic) thought will convince the reader that this will occur, with probability unity, on one of the edges of the expanding double pyramid. This being the case, one of the residuals will be zero. On the other hand, if there are two independent variables, the LAE estimate will be given by the point where the expanding polyhedron hits the plane defined by the origin and the two points $x_1' = (x_{11}, x_{21}, x_{31})$ and $x_2' = (x_{12}, x_{22}, x_{32})$. Again, reflection will show that this will occur, with probability unity, at one of the corners of the double pyramid. A corner corresponds to two of the three residuals being zero.
>
> Returning to the general case of T observations and n independent variables, it should, in view of the foregoing, be clear that the expanding polyhedron of equal distance will touch, with probability unity, the hyperplane defined by the independent variables at a point on the polyhedron that lies in as many edges of the polyhedron as there are dimensions to the hyperplane. Since each edge of the polyhedron corresponds to some residuals being zero, there will therefore be n zero residuals in all.

We can illustrate the functioning of the absolute value metric in regression estimation in the case of two observations on one independent variable Y and one dependent variable X, as follows:



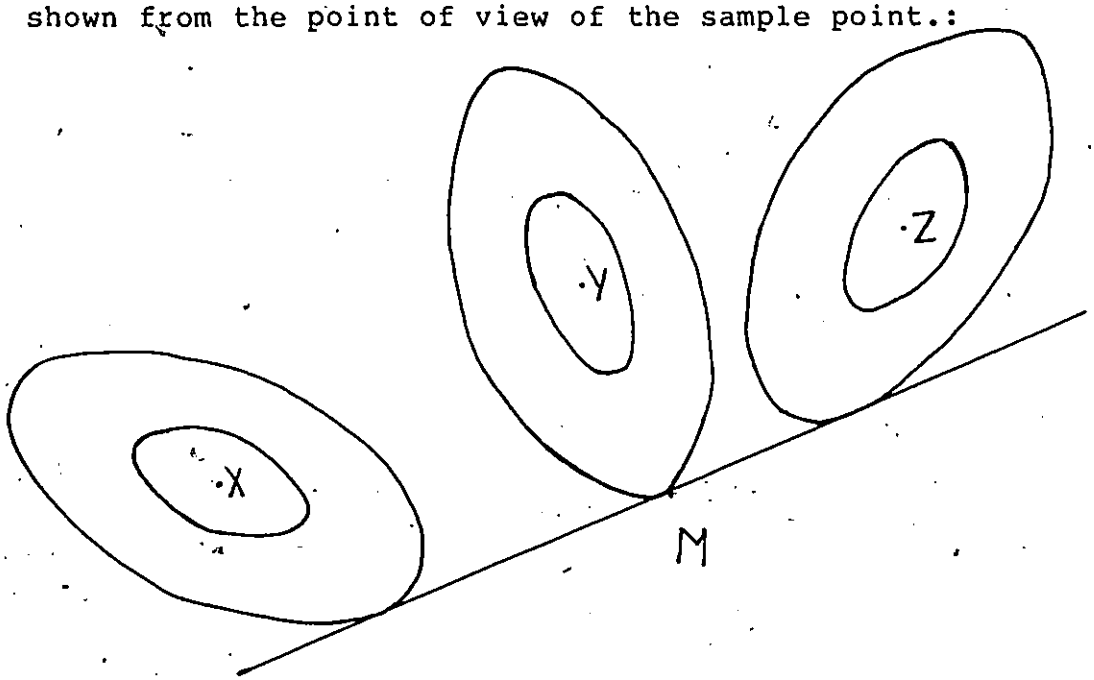
As noted above, one of the two residuals of the error vector is zero, in this case E_1.

6.3.6 Unknown Variance-Covariance Matrices

So far in the discussion of estimation techniques, we have assumed that the variance-covariance matrix of the error, Q, is completely known. Many estimation techniques do not assume complete knowledge of the variance-covariance matrix. Rather, they assume that the matrix has a certain parametric form, where the parameters are estimated using the data in the sample matrix. In these methods, the matrix Q is essentially a function of the sample matrix X. We could write Q = Q(X).

The use of a variance-covariance matrix Q(X) that is a function of the sample point has an interesting effect on the

geometric structure of the regression estimation. It means that the distance function varies from point to point in the sample space, since Q(X) is not generally equal to Q(Y) when X is not equal to Y. This situation can be illustrated in the following way for three distinct sample points X, Y and Z. For purposes of exposition, the level curves of the distance function are shown from the point of view of the sample point.:



This situation has the odd effect that the distance between two points is not identified with the length of the vector between the two points. Suppose we are measuring the distance between a sample point X and a point Y on the model set. The generalized squared Euclidean metric with a fixed matrix Q defines this as $D(X,Y) = (X-Y)'Q(X-Y)$. If we let $E = X-Y$ be the matrix vector between X and Y, the distance between any two points Z and W will be the same as D(X,Y) as long as Z-

W=E. However, if the matrix Q is a function of X, then the distance is defined as D(X,Y) = (X-Y)'Q(X)(X-Y). This is not necessarily equal to D(Z,W) = (Z-W)'Q(Z)(Z-W) even if Z-W = X-Y, since Q(X) may not be the same as Q(Z).

6.4 Distance Functions in Regression Analysis

This section examines the internal mechanics of distance functions as they are used in regression, and proposes a general form for all regression distance functions.

The distance function in regression has as domain the cross-product of the sample space with itself. Given a pair of points (Y,Z) in S x S, where both Y and Z are nxk matrices in S, the distance function gives a non-negative real number d:

$$D(Y,Z) = d$$

The distance D(Y,Z) is not always synonymous with the length of the line segment from Y to Z. This segment can be represented as the error matrix E = Y - Z. It is important to distinguish between distance and length in regression geometry. As we have seen, some regression distance functions assign different lengths to the same error matrix, depending on the end points that generated it.

Three steps are needed to transform the entries in the

matrix E into the distance d: the entries must be made non-negative, they must be weighted to reflect the assumptions of the model's error structure, and they must be combined into a single real number.

Let POS (E) be the operation that converts the entries into non-negative numbers. In theory, we could use any number of operations, such as squaring, quadrupling, taking absolute values or taking positive square roots. In nearly all of econometrics, the entries are squared, that is POS (E) = (e_{ij}^2). The only other operator encountered during this research on econometrics is the absolute value in which case POS (E) = $(|e_{ij}|)$. This is sometimes used in models where the error variables are assumed to have high variances, or "thick tails". (Taylor, 1974)

Let WEIGHT (E) be the operator that weights each entry in E to reflect the assumptions of the model's error structure. These weights reflect the variance associated with each error variable and the covariances of the error variables with each other and with other variables in the model. In particular, these weights reflect which error variables have zero variance, that is which variables are systematic and have no associated error component.

The third step of the distance function is to combine all of the non-negative, weighted entries of the error matrix E into a single number. In theory, we could use a variety of operations, such as multiplying them together, or taking their geometric mean but in practice we always simply add them up.

By combining these three operations into one, we arrive at a general form for all distances in regression:

$$D(X,Y) = \sum_{i,j} POS \cdot WEIGHT\ [X_{ij} - Y_{ij}]$$

The operations POS and WEIGHT are independent of each other and it does not matter which is executed first. These operations are important only because of their role in translating the model's assumptions into a measure of distance applicable to the sample space.

6.5 Solving the Estimation Problem

The geometric structure of the sample space created by the distance function leads directly to the computational method used to solve the estimation problem. As we have seen, the estimation function in regression is the solution to the optimization problem of minimizing the distance from the observed sample point to the model set. The factor that distinguishes one regression estimation technique from another is the distance function, which is determined by the assumptions of the model. We will now examine how the geometry of the

distance function leads to the method used to compute the estimated systematic point.

6.5.1 Ordinary Least Squares

In ordinary least squares, the estimation function minimizes the distance from the sample point to the affine linear subspace spanned by the k-1 independent variables within the n-dimensional space of the dependent variable. This is equivalent to minimizing the distance from the vector of the dependent variable to the linear subspace spanned by the independent variables within an n-dimensional space. The distances are defined by the square of the standard Euclidean metric. The closest point on a linear subspace to a given point not on the subspace is the orthogonal projection of the point onto the subspace. Geometrically, this solution is found by simply dropping a perpendicular line from the point to the subspace.

Assuming that the dependent variable is in the first column of the matrix X, the subspace to be projected onto is spanned by the vectors $X_2, \ldots, X_k$. Let $Y = [X_2, \ldots, X_k]$ be the $n \times (k-1)$ matrix of independent variables. Then the orthogonal projector onto the span of the columns of Y is $Y(Y'Y)^{-1}Y'$. The estimated systematic component of the dependent variable, then, is:

$$\hat{X}_1 = Y(Y'Y)^{-1}Y' X_1$$

The estimated matrix of the systematic component is $\hat{X} = (\hat{X}_1, X_2, \ldots, X_k)$ since the independent variables are assumed to be error-free.

This estimation can be easily illustrated in the space of the dependent variable, where the sample matrix is

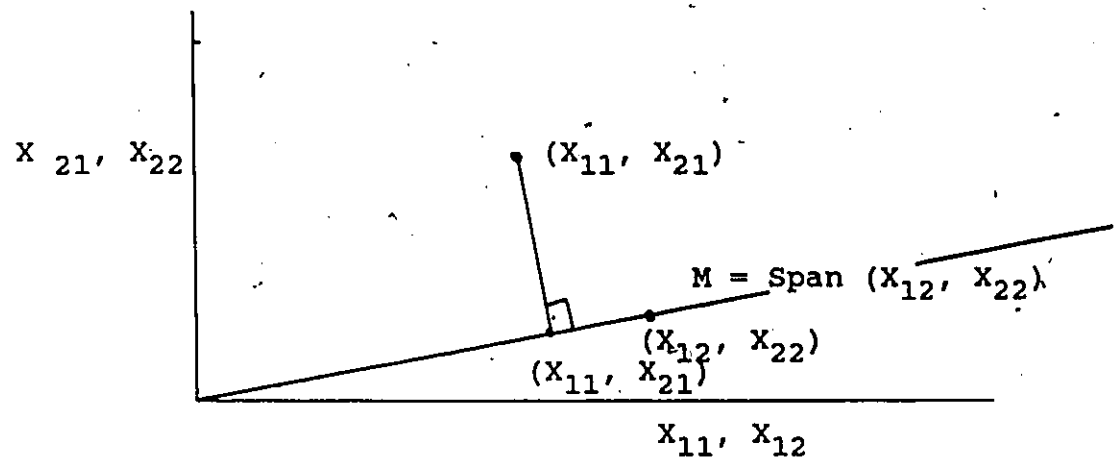
$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$$

In this graph, the origin can be viewed as the point

$$\begin{bmatrix} 0 & X_{12} \\ 0 & X_{22} \end{bmatrix}$$

which separates the affine space A(X) from the origin of the sample space:

$$\begin{bmatrix} \emptyset & \emptyset \\ \emptyset & \emptyset \end{bmatrix}.$$



The parameter vector for ordinary least squares is obtained from the orthogonal projection by solving the equation $Y\hat{b} = \hat{X}_1$. This gives $Y\hat{b} = [Y(Y'Y)^{-1}Y']\ X_1$ so $\hat{b} = (Y'Y)^{-1}Y'X_1$. The parameter vector $\hat{b}$ has k-1 entries, one for each independent variable, since the parameter for the dependent variable is conventionally normalized to 1.

The above estimation procedure only works if the column vectors $X_2, \ldots X_k$ are independent. Otherwise, the matrix Y'Y is singular and its inverse cannot be computed. This does not mean that the optimization problem as discussed above cannot be solved if Y'Y is singular. We can still find the point on the column span of Y closest to the point X_1. This can be done by reducing the set of vectors $X_2, \ldots, X_k$ to contain only independent vectors that span the column space, say $Y^o = [X_2^o, \ldots, X_j^o]$ where $j < k$. The orthogonal projection of X_1 on the span of Y^o in this case is:

$\hat{X}_1 = [Y^o(Y^{o\prime}Y^o)^{-1} Y^{o\prime}]X_1$

The problem with this estimate is that the parameter vector associated with X_1 is not unique. This is because the estimated matrix $X = [X_1, X_2, \ldots X_k]$ is in the collinear set, as discussed in Section 4 on multi-collinearity.

6.5.2 Orthogonal Regression

The estimation procedure for orthogonal regression is very different from that of ordinary least squares because the systematic component is not restricted to a linear subspace of the sample set. The distance function for orthogonal regression applies the standard squared Euclidean metric to all points in the model set. As we saw in the discussion of the model set, the model set is not a linear subspace, although it is composed of various linear subspaces all attached together. Consequently, we cannot directly apply the formula for the orthogonal projection onto a linear subspace.

Orthogonal regression can be solved using some of the geometric notions developed in this paper. The following proof follows a proof given in Pollock (1979,pp. 89-90), but is made shorter and more intuitive with geometry.

The distance function in orthogonal regression is the sum

of the squares of all of the entries in the matrix vector $X - \hat{X}$. This can be written as the sum of the inner products of each row vector with itself:

$$D(X,X) = \sum_{i=1}^{n} (X_i. - \hat{X}_i.)(X_i. - \hat{X}_i.)'$$

We wish to minimize this value for a given X over all possible points $\hat{X}$ in the model set.

Let us recall that the model set is composed completely of n(k-1)-dimensional hyperplanes. Each hyperplane is the null space N(a) of a parameter vector where a'a = 1. So the estimate $\hat{X}$ must belong to one of these null spaces. Let us assume for a moment that $\hat{X}$ belongs to N(a) for some parameter vector a. Since N(a) is a linear subspace, we can easily compute the orthogonal projection of any point in the sample set onto N(a). The orthogonal projection of X into N(a) is the closest point to X belonging to N(a), so it must be $\hat{X}$.

The orthogonal projection onto N(a) is I-aa'. Thus we have:

$$\hat{X} = X(I-aa')$$
$$= X-Xaa'$$

so: $X-\hat{X} = Xaa'$

The matrix $X-\hat{X}$ on the left is the error whose length we wish to minimize. Applying the distance function to the new expression for $X-\hat{X}$ gives:

$$D(X,\hat{X}) = \sum_{i=1}^{n}(X_i.aa')\ (X_i.aa')'$$

$$= \sum_{i=1}^{n}(X_i.aa')\ (aa'X_i.)$$

$$= \sum_{i=1}^{n}X_i.aa'aa'X'_i$$

$$= \sum_{i=1}^{n}X_i.aa'X'_i$$

$$= \sum_{i=1}^{n}(X_i.a)^2$$

$$= a'X'Xa$$

Thus the estimation problem can be reduced to the problem of minimizing the value a'X'Xa subject to a'a = 1. Essentially, the problem has been transformed by picking the closest point to X out of each of the hyperplanes N(a), so the remaining task is only to find which of these points is closest to X. This can be done through solving a simple Lagrangean procedure, minimizing a'X'Xa subject to the constraint a'a = 1. The Lagrangean is:

$$L = a'X'Xa - \lambda(a'a - 1)$$

Differentiating with respect to **a** and setting the result to zero gives the condition

$$(X'X - \lambda I)\ \mathbf{a} = \emptyset$$

A vector **a** satisfying this equation is a characteristic vector of X'X. The value λ corresponding to a characteristic vector is called a characteristic root. Premultiplying by **a'** gives $\mathbf{a}'X'X\mathbf{a} = \lambda \mathbf{a}'\mathbf{a} = \lambda$, which is the distance that is to be minimized. Thus the parameter vector **a** of the estimated solution is the characteristic vector corresponding to the smallest characteristic value of X'X. Once we know the estimated vector $\hat{\mathbf{a}}$, we can find the estimated matrix $\hat{X}$ with the orthogonal projection $I-\hat{\mathbf{a}}\hat{\mathbf{a}}'$ discussed above.

6.5.3 Absolute Value Regressions

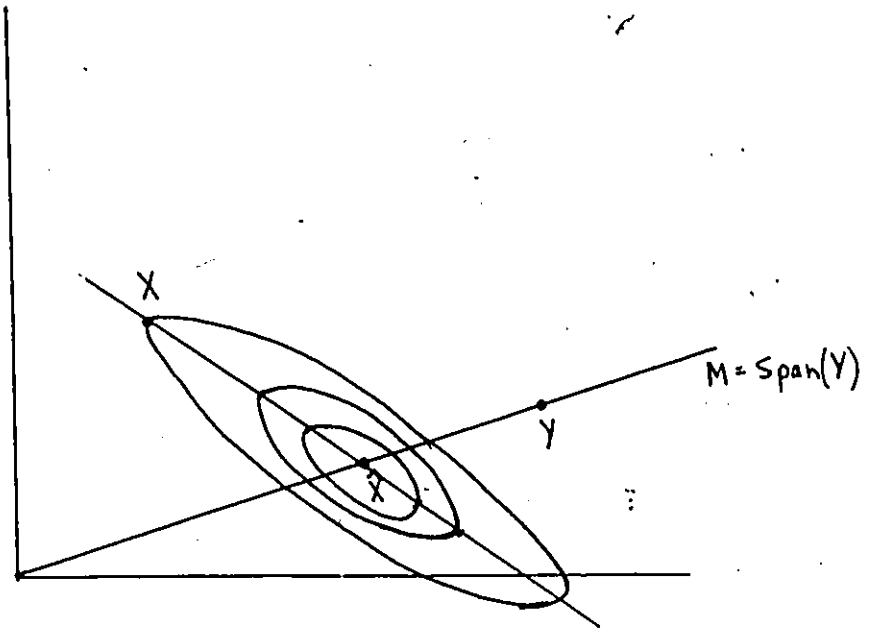
The estimation procedure for absolute value regressions is different from either ordinary least squares or orthogonal regression. Absolute value regressions are usually based on an ordinary least squares model equation, with one dependent variable and several independent variables. Because of this, they can be interpreted as a projection of the vector of the dependent variable onto the hyperplane spanned by the independent variables, as in ordinary least squares. As we have seen, the level curves of the absolute value distance are

very different from those of the least squares metric, so the computation of the estimate is also different. The solution to an absolute value regression is found using methods of linear programming, since both the distance function and the model set in this case are defined by linear functions. (Taylor, 1974, p. 171)

6.5.4 Generalized Regression Methods

The estimations discussed so far, i.e. Ordinary least squares, orthogonal regression and absolute value regression, have assumed that the variance-covariance matrix of the distance function is the identity matrix. The solution to the estimation problem in all three cases is found in the same way if the variance-covariance matrix is generalized to a non-identity matrix Q.

In generalized ordinary least squares, the generalized orthogonal projection $Y(Y'Q^{-1}Y)^{-1}Y'Q^{-1}$ is used to project the dependent variable onto the hyperplane spanned by the independent variables. The resulting estimate is the closest point to the dependent variable under the generalized least squares metric.



Generalized orthogonal regressions can be estimated using Lagrangean methods similar to the solution described earlier to the standard case of orthogonal regression. These solutions will not be discussed here since they have no particular geometric interest. The interested reader may consult Pollock (1979, pp. 94-98) for a discussion of several examples.

Absolute value regressions could theoretically be generalized to incorporate a non-identity variance-covariance matrix. Since no example of a generalized absolute value regression was encountered during this research, it will not be discussed any further.

7. Decomposition of the Sample Space

The estimation function of a regression technique can be used to specify a total decomposition of the sample space into two distinct spaces, the model set and the inverse image of the estimation function. Every point X in the sample space is the unique sum of one point, $\hat{X}$, in the model set and one point in the inverse image of $\hat{X}$ under the estimation function.

Consider the inverse image of the estimation function in the sample set. Let $h(X) = \hat{X}$ be the estimation function for any regression method. It maps all points in the sample space onto their estimate in the model set. Then:

$$h^{-1}(Z) = \{X \epsilon R^{nxk} | h(X) = Z\}$$

for any point $Z \epsilon M$. Taken together, the sets $h^{-1}(Z)$, $Z \epsilon M$, form a family of sets in the sample space, one for each point in the model set.

Several properties of $h^{-1}(Z)$ can be easily discerned. The point Z is in $h^{-1}(Z)$, since $h(Z) = \hat{Z} = Z$ for all points Z in the model set. Furthermore, Z is the only point in the model set belonging to $h^{-1}(Z)$. All points in the sample space must belong to some set $h^{-1}(Z)$ for some Z in M. The sets $h^{-1}(Z)$ are disjoint, that is $h^{-1}(Z) \cap h^{-1}(Y) = \emptyset$ if $Z \neq Y$.

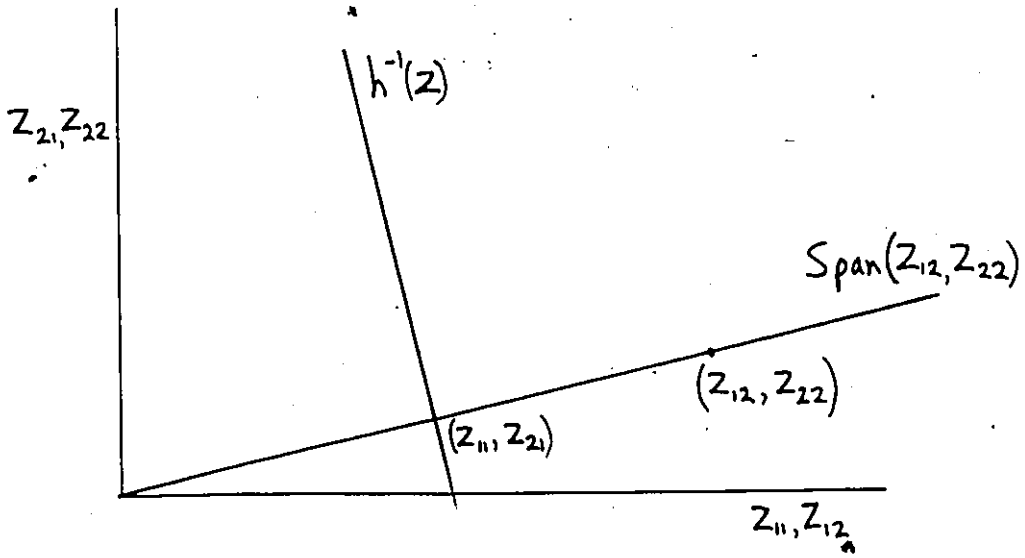
If they were not disjoint, then the estimation function would not be a function.

If the estimation function is based on the squared Euclidean metric, then the sets $h^{-1}(Z)$ are composed of lines emanating from the point Z. Consider a point $X \in h^{-1}(Z)$, and any point Y on the line segment $\overline{XZ}$. We want to show that Y is necessarily in $h^{-1}(Z)$. If the distance is based on the squared Euclidean metric, the distance between two points can be made equivalent to the length of the vector between the two points by taking the square root. Now if $X \in h^{-1}(Z)$, then $D(X,Z) < D(X,W)$ for all $W \neq Z$ in the model set. But $\sqrt{D(X,Z)} = \sqrt{D(X,Y)} + \sqrt{D(Y,Z)}$ since the matrix vector $X-Z = (X-Y) + (Y-Z)$. Suppose Y is not in $h^{-1}(Z)$. Then there is a point W in the Model Set such that $D(Y,W) < D(Y,Z)$. But this implies that $\sqrt{D(X,Z)} = \sqrt{D(X,Y)} + \sqrt{D(Y,Z)} > \sqrt{D(X,Y)} + \sqrt{D(Y,W)} = \sqrt{D(X,W)}$, which contradicts the assumption that X is in $h^{-1}(Z)$. Thus all points on the line from X to h(X) belong to $h^{-1}(X)$.

In some cases, we can get a complete characterization of the set $h^{-1}(Z)$. Consider an ordinary least squares regression model, and let Z be some point on the model set. The set of sample points whose estimate is Z is contained in the n-dimensional affine subspace, A(Z) corresponding to the dependent variable, as we saw earlier. The set $h^{-1}(Z)$ is the

orthogonal complement of the hyperplane spanned by the independent variables of Z, and passing through the dependent variable of Z. This is because the estimator is the orthogonal projection of the dependent variable onto the hyperplane spanned by the independent variables.

If Z $\begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix}$ and the dependent variables is in the first column, we can illustrate as follows:



The set h^{-1} (Z) in ordinary least squares is an affine subspace of the affine subspace A(Z). Its dimension is the dimension of the orthogonal complement of the span of the independent variables within a n-dimensional space. The dimension of an orthogonal complement of a subspace is the

dimension of the total space minus the dimension of the subspace. In this case, dim $(h^{-1}(Z)) = n-(k-1) = n-k+1$. This is an encouraging result, since it is also the dimension of the complement of the model set within the total sample space. Recalling that the sample set has dimension nxk, and that the model set has dimension $(n + 1)(k - 1)$, the complement of the sample set is of dimension $(n \times k) - (n + 1)(k - 1) = nk - k + n + 1 - nk = n - k + 1$.

In the case illustrated above, where $n = k = 2$, the model set is of dimension three and the sets $h^{-1}(Z)$ are of dimension one. For every point Z in the three-dimensional model set, there is a disjoint line $h^{-1}(Z)$ running through it. Every point in the sample set has such a line running through it, linking it to its estimate on the model set.

In the general nxk case of ordinary least squares, every point in the $(n+1)(k-1)$ -dimensional model set has a disjoint n-k+1-dimensional hyperplane of sample points running through it. Every point in the nxk-dimensional sample set belongs to one of these hyperplanes, which links it to its estimate on the model set.

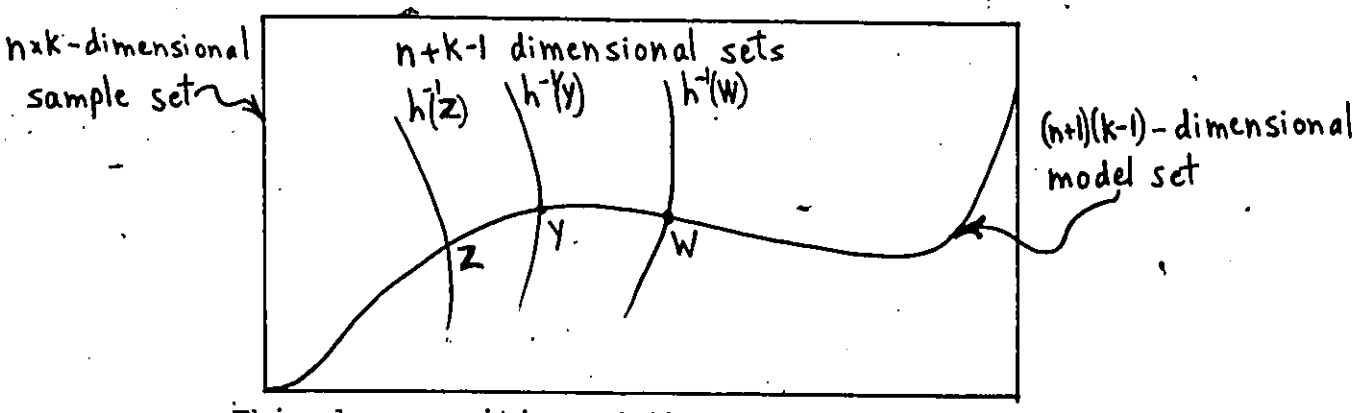
It seems likely that a similar result holds true for all regression estimation techniques. In all cases, there is a

disjoint set $h^{-1}(Z)$ containing the point Z in the sample set. All points in the sample set belong to one of these sets, which link the point to its estimate. Assuming that the error structure, and the resulting estimation function, are similar for all points in the sample space, it appears that the sets $h^{-1}(Z)$ should have the same dimension. If they have the same dimension, and given that the model set has dimension (n+1)(k-1), the sets $h^{-1}(Z)$ should all have the complementary dimension n-k+1. I have not been able to prove this result with the structure developed in this paper. Intuitively, however, it seems that for all regression techniques, the estimation function h(X) defines a disjoint, n-k+1-dimensional set $h^{-1}(Z)$ running through every point Z in the model set. Every point in the sample set belongs to one and only one of these sets.

The model set and the family of inverse images of the estimation function can be combined to provide a unique decomposition of every point in the sample set into two components. The first component is the estimate of the systematic portion of the point. It belongs to the (n+1)(k-1)-dimensional model set. The second component is the estimate of the error portion of the sample point. It belongs to the n-k+1-dimensional inverse image of the systematic component under the estimation function. Together, these two

components uniquely determine every point in the nxk sample space for a given regression technique.

This decomposition of the sample space can be illustrated as follows:



This decomposition of the sample space takes a similar form for all regression techniques. The model set is the same for all regressions. What differs from one technique to another is the family of inverse images $h^{-1}(Z)$ of the estimation function. The estimation function is determined by the distance function, which in turn reflects the assumptions of the model.

By having one global geometric characterization that encompasses all regression techniques, it is easy to compare the different techniques with each other. We have seen how different assumptions lead the sample space to be decomposed in different ways.

8. Conclusions

This paper has used geometric notions to analyze the general class of single-equation regression estimations. It has shown that regression estimation has a common geometric structure within the sample space of nxk matrices. It has examined how regression estimation techniques lead to variations in the general geometric structure of the sample space.

In sum, we have seen the following:

1) All regressions implicity define an estimation function that maps the observed data matrix onto a point in a subset of the sample space consisting of matrices with linearly dependent columns. This subset is connected, non-linear and has dimension (n+1) (k-1). The non-multicollinear points in this set form an (n+1) (k-1)-dimensional manifold in R^{nxk}. All points in the sample space are uniquely decomposed by a regression technique into two components, one on this subset, and one in the inverse image of the estimation function.

2) Multicollinear matrices form an (n+2) (k-2)-dimensional subset of the set of matrices with dependent columns. The geometric properties of the multicollinear set are examined with a view to correcting this problem if it occurs.

3) Regression estimations are based on minimizing a distance criterion. The minimum distance criterion can be justified using the geometries of the distance function and the distribution of the error variables through their respective level curves.

4) The geometric structure of the sample set is used to provide a proof for the solution of orthogonal regression.

All of these seem to be new results in the interpretation of regression analysis.

The paper shows the wide variety of geometric shapes and surfaces that occur in regression spaces, including lines, hyperplanes, spheres, ellipsoids, and polyhedrons. The use of these notions serves to bring the process of regression to life in a visual sense, and allows the reader to gain an intuitive feel for how a regression technique works and how it differs from other regression techniques.

Further Research

There are a great number of avenues for research that could be explored by applying a geometric approach to regression analysis. A few will now be suggested.

1) Geometric analysis could be applied to multi-equation models, since the basic structure is much the same. In particular, the identifiability or non-identifiability of a particular model could probably be given a geometric interpretation in terms of the model set outlined in this paper.

2) The characterization of multicollinearity as a concrete subset of the sample space could perhaps be developed into a test for multicollinearity. In practical applications, we never observe a matrix that is perfectly multicollinear. Knowledge of the multicollinear set could be used to define the set of "almost" multicollinear matrices as those that are within some fixed distance from the multicollinear set.

3) New estimation procedures could be inspired by a geometric approach. For example, the procedure for estimating orthogonal regressions is very cumbersome and unintuitive since it involves the iterative approximation of characteristic vectors and characteristic values of a matrix. It may be possible to find an easier solution to orthogonal regression by geometric manipulation. If it were possible to transform the sample space such that

the model set were made into a linear hyperplane, while the distance structure remained the same, it would be possible to solve orthogonal regression with a simple orthogonal projection.

4) The geometric structure of non-linear modelling could be examined, with a view to providing a global structure within which both linear and non-linear models could be examined.

BIBLIOGRAPHY

Afriat, S. (1975). "Regression and Projection", in Studies in Correlation, G Tintner, P.D. Thionet and H. Strecker (editors) Vanderhoech & Ruprecht, Goettingen.

Fisher, G.R. (1973). "A Coordinate Free Approach to Linear Estimation in Econometrics", Chapter 15 in Special University Lecture in Economics, University of London, October 1973.

Fisher, G. and McAleer, M. (1984). "The Geometry of Specification Error", Australian Journal of Statistics, Vol. 26, No. 3.

Fréchet, Maurice (1953). Pages choisies d'analyse générale, Gauthier-Villars, Paris.

Legendre, A.M. (1806) Nouvelles Méthodes pour la détermination des orbites des comètes, Paris, France: Courcier.

Guillemin, V., and Pollack, A. (1974). Differential Topology, Prentice-Hall, New Jersey.

Herr, David (1980) "On the History of the Use of Geometry in the General Linear Model", The American Statistician, Vol. 34, No. 1.

Gauss, Carl F. (1857), Theory of the Motion of the Heavenly Bodies..., trans. C.H. Davis, Boston: Little, Brown.

Malinvaud, Edmond (1970) Statistical Methods of Econometrics, North-Holland, Amsterdam.

Pindyck, R.S. and Rubinfeld, D.L. (1981). Econometric Models and Economic Forecasts, Second Edition, McGraw-Hill, New York.

Pollock, D.S.G. (1979). The Algebra of Econometrics, Wiley, New York.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications, Second Edition, Wiley, New York.

Sastry, M.V.R. and Tintner, G. (1975). Multivariate Analysis, Econometrics and Information Theory, in Studies in Correlation, G. Tintner, P.D. Thionet and H. Strecker (editors), Vandenhoeck & Ruprecht, Goettingen.

Scheffé, Henry (1959). The Analysis of Variance, Wiley, New York.

Taylor, Lester D. (1974). "Estimation by minimizing the sum of absolute errors", in Frontiers in Econometrics, Academic Press, New York.