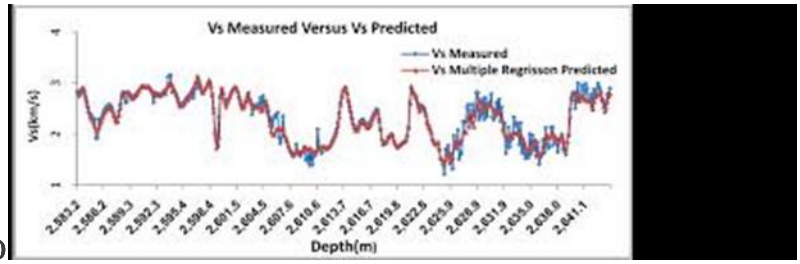
STATISTICS

CORRELATION AND REGRESSION

SESSION 10

STATISTICS
SESSION 10

SESSION 10



Correlation and Regression

SIMULTANEOUSLY EQUATION MODELS

Correlation and linear regression are the most commonly used techniques for investigating the relationship between two quantitative variables.

The goal of a correlation analysis is to see whether two measurement variables co vary, and to quantify the strength of the relationship between the variables, whereas regression expresses the relationship in the form of an equation.

For example, in students taking a Maths and English test, we could use correlation to determine whether students who are good at Maths tend to be good at English as well, and regression to determine whether the marks in English can be predicted for given marks in Maths.

What a Scatter Diagram Tells Us

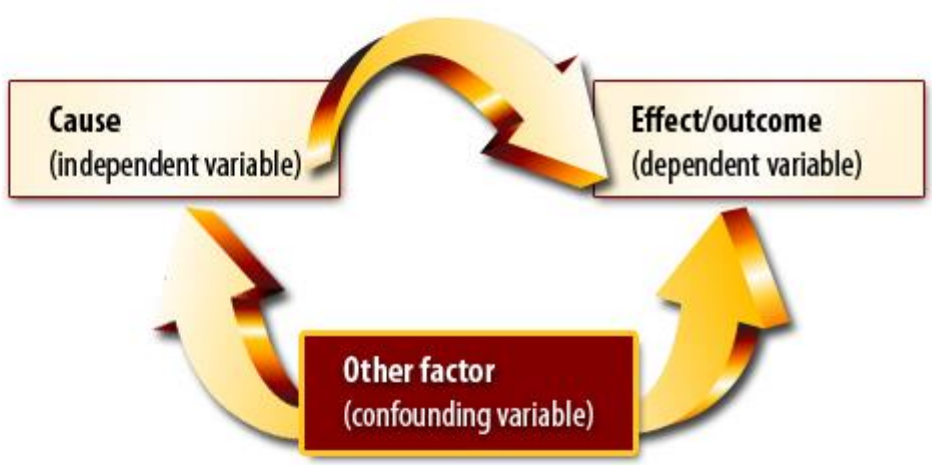
The starting point is to draw a scatter of points on a graph, with one variable on the X-axis and the other variable on the Y-axis, to get a feel of the relationship (if any) between the variables as suggested by the data. The closer the points are to a straight line, the stronger the linear relationship between two variables.

Why Use Correlation?

We can use the correlation coefficient, such as the Pearson Product Moment Correlation Coefficient, to test if there is a linear relationship between the variables. To quantify the strength of the relationship, we can calculate the correlation coefficient (r). Its numerical value ranges from +1.0 to -1.0. $r > 0$ indicates positive linear relationship, $r < 0$ indicates negative linear relationship while $r = 0$ indicates no linear relationship.

A Caveat

It must, however, be considered that there may be a third variable related to both of the variables being investigated, which is responsible for the apparent correlation. Correlation does not imply causation. Also, a nonlinear relationship may exist between two variables that would be inadequately described, or possibly even undetected, by the correlation coefficient.



Why Use Regression

In regression analysis, the problem of interest is the nature of the relationship itself between the dependent variable (response) and the (explanatory) independent variable.

The analysis consists of choosing and fitting an appropriate model, done by the method of least squares, with a view to exploiting the relationship between the variables to help estimate the expected response for a given value of the independent variable. For example, if we are interested in the effect of age on height, then by fitting a regression line, we can predict the height for a given age.

Assumptions

Some underlying assumptions governing the uses of correlation and regression are as follows.

The observations are assumed to be independent. For correlation, both variables should be random variables, but for regression only the dependent variable Y must be random. In carrying out hypothesis tests, the response variable should follow Normal distribution and the variability of Y should be the same for each value of the predictor variable. A scatter diagram of the data provides an initial check of the assumptions for regression.

Uses of Correlation and Regression

There are three main uses for correlation and regression.

- One is to test hypotheses about cause-and-effect relationships. In this case, the experimenter determines the values of the X-variable and sees whether variation in X causes variation in Y. For example, giving people different amounts of a drug and measuring their blood pressure.

- The second main use for correlation and regression is to see whether two variables are associated, without necessarily inferring a cause-and-effect relationship. In this case, neither variable is determined by the experimenter; both are naturally variable. If an association is found, the inference is that variation in X may cause variation in Y, or variation in Y may cause variation in X, or variation in some other factor may affect both X and Y.
- The third common use of linear regression is estimating the value of one variable corresponding to a particular value of the other variable.

Regression and correlation analysis:

Regression analysis involves identifying the relationship between a dependent variable and one or more independent variables. A model of the relationship is hypothesized, and estimates of the parameter values are used to develop an estimated regression equation. Various tests are then employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given values for the independent variables.

Regression model.

In simple linear regression, the model used to describe the relationship between a single dependent variable y and a single independent variable x is y = a0 + a1x + k. a0and a1 are referred to as the model parameters, and is a probabilistic error term that accounts for the variability in y that cannot be explained by the linear relationship with x. If the error term were not present, the model would be deterministic; in that case, knowledge of the value of x would be sufficient to determine the value of y.

Least squares method.

Either a simple or multiple regression model is initially posed as a hypothesis concerning the relationship among the dependent and independent variables. The least squares method is the most widely used procedure for developing estimates of the model parameters.

As an illustration of regression analysis and the least squares method, suppose a university medical centre is investigating the relationship between stress and blood pressure. Assume that both a stress test score and a blood pressure reading have been recorded for a sample of 20 patients. The data are shown graphically in the figure below, called a scatter diagram. Values of the independent variable, stress test score, are given on the horizontal axis, and values of the dependent variable, blood pressure, are shown on the vertical axis. The line passing through the data points is the graph of the estimated regression equation: y = 42.3 + 0.49x. The parameter estimates, b0 = 42.3 and b1 = 0.49, were obtained using the least squares method

Correlation.

Correlation and regression analysis are related in the sense that both deal with relationships among variables. The correlation coefficient is a measure of linear association between two variables. Values of the correlation coefficient are always between -1 and +1. A correlation coefficient of +1 indicates that two variables are perfectly related in a positive linear sense, a correlation coefficient of -1 indicates that two variables are perfectly related in a negative linear sense, and a correlation coefficient of 0 indicates that there is no linear relationship between the two variables. For simple linear regression, the sample correlation coefficient is the square root of the coefficient of determination, with the sign of the correlation coefficient being the same as the sign of b1, the coefficient of x1 in the estimated regression equation.

Neither regression nor correlation analyses can be interpreted as establishing cause-and-effect relationships.

They can indicate only how or to what extent variables are associated with each other. The correlation coefficient measures only the degree of linear association between two variables. Any conclusions about a cause-and-effect relationship must be based on the judgment of the analyst.

Regression and correlation analysis:

Regression analysis involves identifying the relationship between a dependent variable and one or more independent variables. A model of the relationship is hypothesized, and estimates of the parameter values are used to develop an estimated regression equation. Various tests are then employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given values for the independent variables.

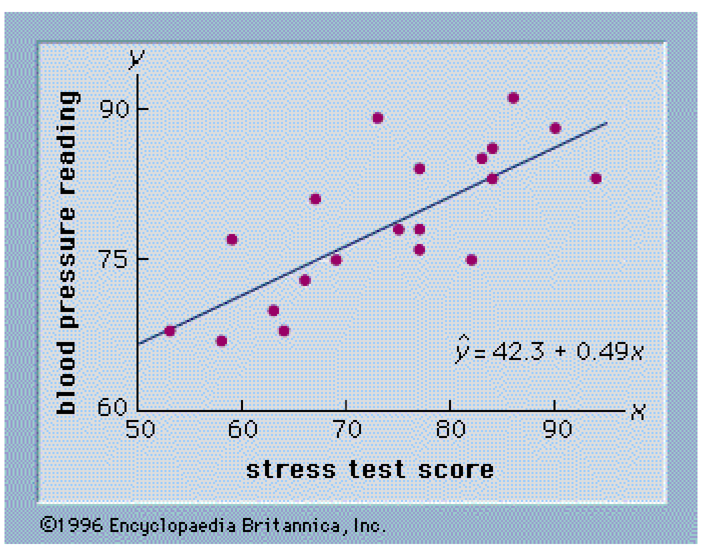
Regression model.

In simple linear regression, the model used to describe the relationship between a single dependent variable y and a single independent variable x is $y = a_0 + a_1x + k$. $a_0$and a_1 are referred to as the model parameters, and is a probabilistic error term that accounts for the variability in y that cannot be explained by the linear relationship with x. If the error term were not present, the model would be deterministic; in that case, knowledge of the value of x would be sufficient to determine the value of y.

Least squares method.

Either a simple or multiple regression model is initially posed as a hypothesis concerning the relationship among the dependent and independent variables. The least squares method is the most widely used procedure for developing estimates of the model parameters.

As an illustration of regression analysis and the least squares method, suppose a university medical centre is investigating the relationship between stress and blood pressure. Assume that both a stress test score and a blood pressure reading have been recorded for a sample of 20 patients. The data are shown graphically in the figure below, called a scatter diagram. Values of the independent variable, stress test score, are given on the horizontal axis, and values of the dependent variable, blood pressure, are shown on the vertical axis. The line passing through the data points is the graph of the estimated regression equation: y = 42.3 + 0.49x. The parameter estimates, b0 = 42.3 and b1 = 0.49, were obtained using the least squares method.



Correlation.

Correlation and regression analysis are related in the sense that both deal with relationships among variables. The correlation coefficient is a measure of linear association between two variables. Values of the correlation coefficient are always between -1 and +1. A correlation coefficient of +1 indicates that two variables are perfectly related in a positive linear sense, a correlation coefficient of -1 indicates that two variables are perfectly related in a negative linear sense, and a correlation coefficient of 0 indicates that there is no linear relationship between the two variables. For simple linear regression, the sample correlation coefficient is the square root of the coefficient of determination, with the sign of the correlation coefficient being the same as the sign of b1, the coefficient of x1 in the estimated regression equation.

Neither regression nor correlation analyses can be interpreted as establishing cause-and-effect relationships. They can indicate only how or to what extent variables are associated with each other. The correlation coefficient measures only the degree of linear association between two variables. Any conclusions about a cause-and-effect relationship must be based on the judgment of the analyst.

Multiple Regression Analysis

Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables- also called the predictors.

For example the yield of rice per acre depends upon quality of seed, fertility of soil, fertilizer used, temperature, rainfall. If

one is interested to study the joint affect of all these variables on rice yield, one can use this technique.

An additional advantage of this technique is it also enables us to study the individual influence of these variables on yield.

Dependent and Independent Variables

By multiple regression, we mean models with just one dependent and two or more independent (exploratory) variables. The variable whose value is to be predicted is known as the dependent variable and the ones whose known values are used for prediction are known independent (exploratory) variables.

The Multiple Regression Model

In general, the multiple regression equation of Y on $X_1, X_2, \ldots, X_k$ is given by:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + b_k X_k$$

Interpreting Regression Coefficients

Here b_0 is the intercept and $b_1, b_2, b_3, \ldots, b_k$ are analogous to the slope in linear regression equation and are also called regression coefficients. They can be interpreted the same way as slope. Thus if $b_i = 2.5$, it would indicates that Y will increase by 2.5 units if X_i increased by 1 unit.

The appropriateness of the multiple regression model as a whole can be tested by the F-test in the ANOVA table. A significant F indicates a linear relationship between Y and at least one of the X's.

How Good Is the Regression?

Once a multiple regression equation has been constructed, one can check how good it is (in terms of predictive ability) by examining the coefficient of determination (R2). R2 always lies between 0 and 1.

R_2 - coefficient of determination

All software provides it whenever regression procedure is run. The closer R_2 is to 1, the better is the model and its prediction.

A related question is whether the independent variables individually influence the dependent variable significantly. Statistically, it is equivalent to testing the null hypothesis that the relevant regression coefficient is zero.

This can be done using t-test. If the t-test of a regression coefficient is significant, it indicates that the variable is in question influences Y significantly while controlling for other independent explanatory variables.

Assumptions

Multiple regression technique does not test whether data are linear. On the contrary, it proceeds by assuming that the relationship between the Y and each of X_i's is linear. Hence as a rule, it is prudent to always look at the scatter plots of

(Y, X_i), i= 1, 2,…,k. If any plot suggests non linearity, one may use a suitable transformation to attain linearity.

Another important assumption is nonexistence of multicollinearity- the independent variables are not related among themselves. At a very basic level, this can be tested by computing the correlation coefficient between each pair of independent variables.

Other assumptions include those of homoscedasticity and normality.

Multiple regression analysis is used when one is interested in predicting a continuous dependent variable from a number of independent variables. If dependent variable is dichotomous, then logistic regression should be used.

Suppose there are m regression equations of the form

y_{it} = y_{-i,t}'\gamma_i + x_{it}'\;\!\beta_i + u_{it}, \quad i=1,\ldots,m,

where i is the equation number, and t = 1, ..., T is the observation index. In these equations xit is the ki×1 vector of exogenous variables, yit is the dependent variable, y−i,t is the ni×1 vector of all other endogenous variables which enter the ith equation on the right-hand side, and uit are the error terms. The "−i" notation indicates that the vector y−i,t may contain any of the y's except for yit (since it is already present on the left-hand side). The regression coefficients βi and γi are of dimensions ki×1 and ni×1 correspondingly. Vertically stacking the T observations corresponding to the ith equation, we can write each equation in vector form as

$$y_i = Y_{-i}\gamma_i + X_i\beta_i + u_i, \quad i=1,\ldots,m,$$

where y_i and u_i are $T\times 1$ vectors, X_i is a $T\times k_i$ matrix of exogenous regressors, and Y_{-i} is a $T\times n_i$ matrix of endogenous regressors on the right-hand side of the ith equation. Finally, we can move all endogenous variables to the left-hand side and write the m equations jointly in vector form as

$$Y\Gamma = X\Beta + U.\,$$

This representation is known as the structural form. In this equation $Y = [y_1\ y_2\ \ldots\ y_m]$ is the $T\times m$ matrix of dependent variables. Each of the matrices Y_{-i} is in fact an n_i-columned submatrix of this Y. The $m\times m$ matrix Γ, which describes the relation between the dependent variables, has a complicated structure. It has ones on the diagonal, and all other elements of each column i are either the components of the vector $-\gamma_i$ or zeros, depending on which columns of Y were included in the matrix Y_{-i}. The $T\times k$ matrix X contains all exogenous regressors from all equations, but without repetitions (that is, matrix X should be of full rank). Thus, each X_i is a k_i-columned submatrix of X. Matrix B has size $k\times m$, and each of its columns consists of the components of vectors β_i and zeros, depending on which of the regressors from X were included or excluded from X_i. Finally, $U = [u_1\ u_2\ \ldots\ u_m]$ is a $T\times m$ matrix of the error terms.

Postmultiplying the structural equation by Γ^{-1}, the system can be written in the reduced form as

$$Y = X\Beta\Gamma^{-1} + U\Gamma^{-1} = X\Pi + V.\,$$

This is already a simple general linear model, and it can be estimated for example by ordinary least squares. Unfortunately, the task of decomposing the estimated matrix $\scriptstyle\hat\Pi$ into the individual factors Β and Γ^{-1} is quite complicated, and therefore the reduced form is more suitable for prediction but not inference.

Assumptions[edit]

Firstly, the rank of the matrix X of exogenous regressors must be equal to k, both in finite samples and in the limit as $T \to \infty$ (this later requirement means that in the limit the expression $\scriptstyle \frac1TX'\!X$ should converge to a nondegenerate $k \times k$ matrix). Matrix Γ is also assumed to be non-degenerate.

Secondly, error terms are assumed to be serially independent and identically distributed. That is, if the tth row of matrix U is denoted by $u_{(t)}$, then the sequence of vectors $\{u_{(t)}\}$ should be iid, with zero mean and some covariance matrix Σ (which is unknown). In particular, this implies that $E[U] = 0$, and $E[U'U] = T\,\Sigma$.

Lastly, the identification conditions require that the number of unknowns in this system of equations should not exceed the number of equations. More specifically, the order condition requires that for each equation $k_i + n_i \leq k$, which can be phrased as "the number of excluded exogenous variables is greater or equal to the number of included endogenous variables". The rank condition of identifiability is that $\operatorname{rank}(\Pi_{i0}) = n_i$, where Π_{i0} is a $(k - k_i) \times n_i$ matrix which is obtained from Π by crossing out those columns which correspond to the excluded endogenous variables, and those rows which correspond to the included exogenous variables.

Simultaneous equation models are a form of statistical model in the form of a set of linear simultaneous equations. They are often used in econometrics.

INTRODUCTION TO THE SIMULTANEOUS EQUATONS REGRESSION MODEL

When a single equation is embedded in a system of simultaneous equations, at least one of the right-hand side variables will be endogenous, and therefore the error term will be correlated with at least one of the right-hand side variables. In this case, the true data generation process is not described by the classical linear regression model, general linear regression model, or seemingly unrelated regression model; rather, it is described by a simultaneous equations regression model. If you use the OLS estimator, FGLS estimator, SUR estimator, or ISUR estimator, you will get biased and inconsistent estimates of the population parameters.

Specifying a Simultaneous Equation System

A simultaneous equation system is one of 4 important types of equation systems that are used to specify statistical models in economics. The others are the seemingly unrelated equations system, recursive equations system, and block recursive equation system. It is important to know the difference between these 4 types of equation systems when specifying statistical models of data generation processes.

The Identification Problem

Before you estimate a structural equation that is part of a simultaneous equation system, you must first determine whether the equation is identified. If the equation is not identified, then estimating its parameters is meaningless. This is because the estimates you obtain will have no interpretation, and therefore will not provide any useful information.

Classifying Structural Equations

Every structural equation can be placed in one of the following three categories.

1. Unidentified equation – The parameters of an unidentified equation have no interpretation, because you do not have enough information to obtain meaningful estimates.
2. Exactly identified equation – The parameters of an exactly identified equation have an interpretation, because you have just enough information to obtain meaningful estimates.
3. Overidentified equation – The parameters of an overidentified equation have an interpretation, because you have more than enough information to obtain meaningful estimates.

Exclusion Restrictions

The most often used way to identify a structural equation is to use prior information provided by economic theory to exclude certain variables from an equation that appear in a model. This is called obtaining identification through exclusion restrictions. To exclude a variable from a structural equation, you restrict the value of its coefficient to zero. This type of zero fixed value restriction is called an exclusion restriction because it has the effect of omitting a variable from the equation to obtain identification.

Rank and Order Condition for Identification

Exclusion restrictions are most often used to identify a structural equation in a simultaneous equations model. When using exclusion restrictions, you can use two general rules to check if identification is achieved. These are the rank condition and the order condition. The order condition is a necessary but not sufficient condition for identification. The rank condition is both a necessary and sufficient condition for identification. Because the rank condition is more difficult to apply, many economists only check the order condition and gamble that the rank condition is satisfied. This is usually, but not always the case.

Order Condition

The order condition is a simple counting rule that you can use to determine if one structural equation in a system of linear simultaneous equations is identified. Define the following:

G = total number of endogenous variables in the model (i.e., in all equations that comprise the model).

K = total number of variables (endogenous and exogenous) excluded in the equation being checked for identification.

The order condition is as follows.

If	$K = G - 1$	the equation is exactly identified
If	$K > G - 1$	the equation is overidentified
If	$K < G - 1$	the equation is unidentified

Rank Condition

The rank condition tells you whether the structural equation you are checking for identification can be distinguished from a linear combination of all structural equations in the simultaneous equation system. The procedure is as follows.

1. Construct a matrix for which each row represents one equation and each column represents one variable in the simultaneous equations model.
2. If a variable occurs in an equation, mark it with an X. If a variable does not occur in an equation, market it with a 0.
3. Delete the row for the equation you are checking for identification.
4. Form a new matrix from the columns that correspond to the elements that have zeros in the row that you deleted.
5. For this new matrix, if you can find at least (G – 1) rows and columns that are not all zeros, then the equation is identified. If you cannot, the equation is unidentified.

SPECIFICATION

A simultaneous equation regression model has two alternative specifications:

1. Reduced form
2. Structural form