University College London

Department of Economics

M.Sc. in Economics

MC3: Econometric Theory and Methods

Course Notes 2

Notes on estimation and inference in regression models

Andrew Chesher

6/10/2005

1. Introduction

In the first set of notes we considered econometric models for investment in schooling (S) and a labour market outcome (e.g. the log wage), W, given an observable characteristic of an individual, X. The linear model we considered had the following form.

$$W = \alpha_0 + \alpha_1 S + \alpha_2 X + \varepsilon_1 + \lambda\varepsilon_2 \quad (1.1)$$

$$S = \beta_0 + \beta_1 X + \varepsilon_2 \quad (1.2)$$

We saw that without further restrictions, even if we knew the values of ε_1 and ε_2, knowledge of values of W, S, and X could not provide information about the values of the parameters α_0, α_1, α_2 and λ for the data generating process - structure - which produced the data[1]. However, if the values of X that we obtain have sufficient variation and we could observe the values of ε_2 then we could deduce the values of β_0 and β_1 in the data generating process which produced the data[2].

In practice we do *not* observe the values of ε_1 and ε_2 - values which we will think of as realisations of random variables with a probability distribution conditional upon X.

Considering the equation for S, data on S and X cannot then be informative about β_0 and β_1 without some restriction on the way in which this probability distribution depends upon X. To see this, suppose that ε_2 tends to take large values when X takes large values, and small values when X takes small values, that is ε_2 is positively correlated with X. In this situation, if, as is always the case, we cannot observe the values of ε_2 then we cannot distinguish the effect of ε_2 on S from the direct effect of X on S (which is what the parameter β_1 measures) and so the value of β_1 is not identifiable.

In constructing econometric models various sorts of restriction on the covariation of unobservable and observable variables are employed in order to achieve identification. A particularly severe restriction, but one we sometimes find employed, is that the unobservable random variables (ε_1 and ε_2 above) and X are *independently* distributed. Another restriction on covariation we find used is that *conditional medians* of unobservable random variables are independent of X.

By far the most common restriction employed in practice requires *conditional expected values* of unobservables given X to be invariant with respect to changes in the value of X. Consider the example of the structural equation (1.2). With the restriction $E[\varepsilon_2|X = x] = c_2$, a constant, there is

$$E[S|X = x] = \beta_0 + \beta_1 X + c_2$$

and β_1 is identified as the cofficient on X in the regression function of S given X. Typically we find that the conditional expectations of unobservables are normalised to be zero which serves, in the example above, to identify the value of β_0 in (1.2). These notes are concerned with the issues that arise when we work with econometric models in which there are such conditional

[1] We saw that with the additional restrictions $\alpha_2 = 0$, $\beta_1 \neq 0$, the parameters α_0, α_1 and λ could be identified.

[2] Just two observations $(S_1, X_1, \varepsilon_{21})$ and $(S_2, X_2, \varepsilon_{22})$ would suffice if $X_1 \neq X_2$ because then we have

$$S_1 - \varepsilon_{21} = \beta_0 + \beta_1 X_1$$
$$S_2 - \varepsilon_{22} = \beta_0 + \beta_1 X_2$$

which is a pair of simultaneous equations which can be solved for unique values of β_0 and β_1.

expected value restrictions on unobservable random variables. In such cases we are lead to consider estimation of (mean) regression functions.

1.1. Regression functions

Let us now use Y to denote an outcome and $x = \begin{bmatrix} x_1 & \dots & x_k \end{bmatrix}'$ to denote a $k \times 1$ vector of variables whose effect on Y is of interest, and consider an econometric model as follows

$$Y = x'\beta + \varepsilon$$

where "$'$" denotes matrix (vector) transposition, $\beta = \begin{bmatrix} \beta_1 & \dots & \beta_k \end{bmatrix}'$ is a vector of parameters, and ε is a scalar random variable with $E[\varepsilon|x] = 0$.[3] I will call the variables X *covariates*. You will also find the expressions *regressors* and *explanatory variables* used in the literature.

In this *linear regression model* for Y,

$$E[Y|x] = x'\beta \tag{1.3}$$

which, written without using matrix notation is:

$$E[Y|X = x] = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

We start with some preliminary marks about the interpretation and scope of this linear regression model. Then we consider ways of estimating β and how we can make probability statements about the unknown value of β. Finally we consider estimation of the parameters of non-linear regression models.

1.2. Intercepts and the location and scale of covariates

The model (1.3) does not include an intercept, or constant, term explicitly. If the expectation of Y given x is not zero when $x = 0$, then there must be a constant term (or non-zero intercept) in the regression function, that is:

$$E[Y|x] = \beta_0 + x'\beta.$$

This can be accommodated in the general formulation (1.3) by letting one element of x always equal 1, and unless noted we will assume that this has been done if a non-zero intercept is to be identified and estimated.

The value of the intercept is sensitive to the origin from which the elements of x are measured. If we measure x via $z = x - a$ then (1.3) becomes

$$E[Y|x] = x'\beta = (z + a)'\beta = \beta_0 + z'\beta$$

where $\beta_0 = a'\beta$, modifying the value of the intercept if one was originally in the model and introducing an intercept otherwise. It is quite common to find x measured as a deviation from the mean of its observed values.

[3] A more expansive notation would write $Y = X'\beta + \varepsilon$ and $E[\varepsilon|X = x] = 0$, leading to $E[Y|X = x] = x'\beta$. I abbreviate this as in the main text. The equation $Y = x'\beta + \varepsilon$ should be taken to mean that, when $X = x$, $Y = x'\beta + \varepsilon$.

There is always the freedom to choose alternative units of measurement for x. For example if the regression function for Y given a scalar X_1 is

$$E[Y|x_1] = \beta_0 + \beta_1 x_1 \tag{1.4}$$

when X_1 is measured in pounds, then, if Z_1 is the same magnitude measured in thousands of pounds, we have $X_1 = 1000 \times Z_1$, and so

$$E[Y|z_1] = \beta_0 + \beta_1 \times 1000 z_1$$

and the coefficient on z_1 in this latter regression equation is 1000 times the coefficient on x_1 in the regression function (1.4).

In a matrix formulation a change of scale is achieved by measuring vector x via $z = \Lambda x$ where Λ is a square diagonal matrix. This changes (1.3) into

$$E[Y|x] = x'\beta = \left(\Lambda^{-1}z\right)'\beta = z'\gamma$$

where $\gamma = \Lambda^{-1}\beta$..

Finally note that we can express the model in terms of full rank linear combinations of conditioning variables without changing the meaning of the model. Suppose that x is measured via $z = Ax$ where A has rank k.[4] Then (1.3) becomes

$$E[Y|x] = x'\beta = \left(A^{-1}z\right)'\beta = z'\gamma$$

where $\gamma = A^{-1\prime}\beta$. If we knew γ and A we could deduce β.

1.3. Linearity

Most of the discussion below is concerned with *linear models*. The use of the term "linear" here is a little deceptive because in what follows it is perfectly in order for some elements of x to be nonlinear functions of other elements or of some variable not appearing in x at all. In practice it is quite common to find x containing squares, and cross-products of conditioning variables when non-linearity is suspected in the dependence of Y on x. What is important, as will become clear, is that the regression function is *linear in parameters*, β, above. Later we will consider estimation when regression functions are nonlinear in parameters.

1.4. Regime and group specific regression functions

The linear regression model we will study is flexible in another way too. Suppose the model is constructed to capture the dependence of Y on x in a situation where the dependence may vary across groups of agents (e.g. urban and rural households), or, in a time series context, across periods of time (regimes - e.g. quarters of the year, phases of the business cycle).

Suppose there are G groups, or regimes. Define G binary random variables, $D_1, \ldots, D_G$, with $D_g = 1$ when group g is encountered, or in a time series context, when regime g prevails, and $D_g = 0$ otherwise. These binary or indicator variables are called "dummy variables" in

[4] Which means that A can be inverted.

some of the textbooks. Suppose there is a different regression function in each group or regime, thus:

$$E[Y|X = x, D_g = 1] = x'\beta_g$$

in which the regression coefficients are regime specific. This can be accommodated in the formulation (1.3) by writing

$$\begin{aligned} E[Y|X = x, D = d] &= d_1 x'\beta_1 + d_2 x'\beta_2 + \cdots + d_G x'\beta_G \\ &= \begin{bmatrix} d_1 x' & d_2 x' & \dots & d_G x' \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_G \end{bmatrix}. \end{aligned}$$

Note that this model is *linear in the parameters* $\beta_1, \dots \beta_G$ but *nonlinear in the conditioning variables*, which involve cross-products of the values, d_g, taken by the binary indicators and the elements of x. Since it is linear in parameters it can be handled in the linear regression model framework which we study below.

Often some of the parameters of the regression function are restricted to be common across groups or regimes. On many occasions, all parameters are so restricted except the intercepts

2. Estimation in linear models

2.1. Introduction

Now we consider how data can be used to estimate regression functions. Suppose we have n records of values of Y and x. We model these as realisations of n random variables, $Y_1, \dots, Y_n$, which we arrange in a $n \times 1$ vector, y. We arrange the associated values of x in a $n \times k$ matrix X. Unless noted otherwise we will assume that the matrix X has rank k. We return to this point in Section 2.2. Here x_{ij} is value i of the variable x_j. We will call the x variables "covariates".

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

Here x_i' is the ith observation on all the covariates arranged as a row vector - that is the ith row of the matrix X.

In some cases it makes sense to regard the Y_i's as mutually independent given the x_i's. This is a quite common position to take in microeconometric work where data may arise by random sampling of households, people or firms from some population. Sometimes it is clear *a priori* that there is dependence among the Y_i's, for example in time series analysis, where often it is precisely the dependence between successive Y_i's that is of primary interest, perhaps because it opens the possibility of developing informative forecasts. As another example, in microeconometric work we might have data recording responses of many members of a number of households. Then we might expect there to be dependence amongst responses from members

of the same household, but perhaps independence among responses from members of different households. We will consider both cases, y's elements independent and dependent.

In some cases the values of x can reasonably be thought of as realisations of a vector random variable[5]. We can think in these terms, for example, when our data are a sample of households, randomly chosen from a population, and x records a household's characteristics. Then the x data carry meaningful information about the population distribution of X.

Sometimes we meet data for which the values of x cannot be thought of in this way. For example, sometimes, in order to obtain a good range of values of x, we purposively seek groups of agents with interesting values of x and observe the realisations of Y that they generate. And sometimes we even assign a value of x, for example when attempting to measure the impact of an intervention where we might give some households a treatment, for example, an income subsidy, leaving others untreated, with unchanged income.

As long as we are only interested in regression functions conditional on x these alternative mechanisms for generating data do not affect our choice of estimation method. The *principle of conditionality*, commonly regarded as a useful principle[6], tells us that, in the interests of accuracy, inferences about parameters should be made conditional on all values of random variables whose marginal distribution is invariant with respect to the parameters of interest.

In the case being considered here, even when x values are informative about the marginal distribution of X, they are rarely informative about the conditional distribution of Y given X and in particular about the regression function, except to the extent that different values of x give a wider view of the sensitivity of the regression function of Y on X to values realised for X. So even when x values are obtained by some form of random sampling, we will conduct inference conditional on the realised x values - an approach which is the only one available when x values are purposively chosen.

We write the linear regression model as

$$y = X\beta + \varepsilon$$

where ε is a $n \times 1$ vector of values of the unobservable, that is:

$$\varepsilon = \left[\begin{array}{cccc} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{array} \right]'.$$

We now seek statistics (that is, functions of y and X) which serve as useful estimators of β.

2.2. Analogue estimation and the ordinary least squares estimator

One way to proceed is by what we will loosely call the *analogue principle*[7]. In the simple form considered here this involves expressing the parameter to be estimated (here β) as a function of expected values of random variables and replacing expected values by sample data based analogues of them.

[5] By this I mean that it makes sense to think in terms of the probability of observing particular values or ranges of values of x.

[6] There is a good discussion in "Theoretical Statistics", D.R. Cox and D.V. Hinkley, Chapman and Hall, London: 1974.

[7] An excellent book devoted to this topic and its wide-ranging applications is: *Analog Estimation Methods in Econometrics*, C.F. Manski, New York, Chapman and Hall, 1988. The book is freely available on the internet, along with a few others, at *http://emlab.berkeley.edu/books.html*.

To see how this can lead to an estimator of β, note that[8] $E[\varepsilon|X] = 0$ implies that

$$E[y - X\beta|X] = 0$$

and therefore

$$\begin{aligned} X'E[(y - X\beta)|X] &= E[X'(y - X\beta)|X] \\ &= E[X'y] - X'X\beta \\ &= 0 \end{aligned} \tag{2.1}$$

and so,

$$\beta = (X'X)^{-1} E[X'y|X]$$

as long as $X'X$ has full rank (k) so that its inverse exists. Replacing $E[X'y|X]$ by $X'y$ leads to an estimator

$$\hat{\beta} = (X'X)^{-1} X'y$$

which is clearly unbiased, that is, $E[\hat{\beta}|X] = E[\hat{\beta}] = \beta$. This estimator is known as the *ordinary least squares* (OLS) estimator, for reasons that will become clear.

Note that to calculate this estimator it must be the case that the rank of X is equal to k. In this case, for all non-zero $k \times 1$ vectors, c, $Xc \neq 0$. When the rank of X is less than k, there exists a non-zero vector c such that $Xc = 0$. In words, there is a linear combination of the columns of X which is a vector of zeros. In this situation the OLS estimator cannot be calculated. Looking back to (2.1) we see that really what is going on here is that β cannot defined by using the information contained in X. Perhaps one could obtain other values of x and then be in a position to define β. But sometimes this is not possible, and then β is not *identifiable* given the information in X. Perhaps we could estimate functions (e.g. linear functions) of β that would be identifiable even without more x values..

The variance of the OLS estimator (conditional on X) is

$$\begin{aligned} Var(\hat{\beta}|X) &= E[\left(\hat{\beta} - E[\hat{\beta}|X]\right)\left(\hat{\beta} - E[\hat{\beta}|X]\right)'|X] \\ &= E[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'|X] \\ &= E[(X'X)^{-1} X'\varepsilon\varepsilon' X (X'X)^{-1}|X] \\ &= (X'X)^{-1} X'E[\varepsilon\varepsilon'|X]X (X'X)^{-1} \\ &= (X'X)^{-1} X'\Sigma X (X'X)^{-1} \end{aligned}$$

where $\Sigma = Var[\varepsilon|X]$. Here, at line 2 we have used the result on the unbiasedness of $\hat{\beta}$, at line 3

[8] Expectations here and later in this section are taken conditional on X. Note that this would not make much sense in a time series context in which X contained lagged values of elements of y, e.g. if we had, for $t = 1, \ldots, n$,

$$y_t = \beta y_{t-1} + \varepsilon_t$$

and $E[\varepsilon_t|y_{t-1}] = 0$. So the arguments here will not apply in a time series context - a setting considered later in the course.

we have used

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta+\varepsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon\end{aligned}$$

and throughout the fact that conditional on X, the matrix X and functions of it alone are constant when taking expected values.

If $\Sigma = \sigma^2 I_n$, which means that the unobservables, ε_i, (and therefore the Y_i's) have, conditional on X, common variance, independent of X, and are pairwise uncorrelated, then

$$Var(\hat{\beta}|X) = \sigma^2 \left(X'X\right)^{-1}. \tag{2.2}$$

Pairwise uncorrelatedness would arise if the Y_i's were mutually independently distributed given X.

The diagonal elements of $Var(\hat{\beta}|X)$ are the variances of the OLS estimators of the elements of β. Their square roots (standard deviations) are commonly referred to as *standard errors*. Note that these involve a parameter, σ^2, which will usually be unknown. We will shortly consider ways of estimating σ^2. Using an estimator of σ^2 in (2.2) produces an estimated variance of the OLS estimator,

$$\widehat{Var}(\hat{\beta}|X) = \hat{\sigma}^2 \left(X'X\right)^{-1}.$$

The square roots of the diagonal elements of this matrix are often referred to as *estimated standard errors*.

Inspecting (2.2) we see that the OLS estimator is less variable the smaller is σ^2 (i.e. the less variation there is in Y around its regression function), and the smaller are the elements of $(X'X)^{-1}$. In situations where we can choose the values of X then it is well to choose them to make the elements of $\left(X'X\right)^{-1}$ small. There is a whole statistical literature on this topic under the heading "design of experiments". In many cases it is best to have a wide range of variation for each element[9].

We can write (2.2) as

$$Var(\hat{\beta}|X) = \frac{\sigma^2}{n}\left(n^{-1}X'X\right)^{-1}$$

and if we think of the rows in X as being obtained by sampling from some probability distribution then we might expect $n^{-1}X'X$ (which contains the average squares and cross-products of the values of the covariates) to remain fairly constant as the sample size (n) increases. It is clear then that larger samples (larger n) tend to lead to more accurate (lower variance) OLS estimators.

2.3. (*)[10] Alternative analogue estimators

The operation done in the previous Section, that is obtaining an expression for β in terms of moments and then replacing moments by their sample analogues to produce an estimator, can

[9] When would that be a bad thing to have?

[10] Starred sections can be omitted at a first reading.

be done in many ways. Let H be a $n \times k$ matrix containing elements which are functions of the elements of X then, arguing as above[11]:

$$\begin{aligned} E[H'\varepsilon|X] &= E[H'(y - X\beta)|X] \\ &= E[H'y|X] - E[H'X|X]\beta \\ &= E[H'y|X] - (H'X)\,\beta \\ &= 0. \end{aligned}$$

If the matrix $H'X$ has rank k then

$$\beta = (H'X)^{-1} E[H'y|X].$$

So, β can be written as a simple function of conditional moments involving H, X, and y.

We can get from here to a family of estimators, $\hat{\beta}_H$ by replacing $E[H'y|X]$ by $H'y$ itself leading to

$$\hat{\beta}_H = (H'X)^{-1} H'y$$

which has the unbiasedness property

$$E[\hat{\beta}_H|X] = \beta = E[\hat{\beta}_H].$$

The variance (matrix) of $\hat{\beta}_H$.is, arguing as in the last Section

$$V[\hat{\beta}_H|X] = (H'X)^{-1} H'\Sigma H\,(X'H)^{-1}$$

where $\Sigma = V[\varepsilon|X]$ which might depend upon X. In the special case in which Σ does not depend on X and elements of y are uncorrelated with one another and have common variance (all conditional on X) we can write $\Sigma = \sigma^2 I$ where I is a $n \times n$ identity matrix and $\sigma^2 > 0$ is a constant. Then

$$V[\hat{\beta}_H|X] = \sigma^2\,(H'X)^{-1} H'H\,(X'H)^{-1}.$$

Different choices of H generally lead to different estimators. One has to choose between these, for which one needs a criterion that ranks estimators. One criterion commonly used considers the accuracy of the estimator as measured by its variance. This will be taken up further in Section 2.7 where we show that, when $V[\varepsilon|X] = \sigma^2 I_n$, the estimator with smallest variance (in a sense to be defined) is got by choosing $H = X$ which is what we did in the previous Section, a choice that produced the OLS estimator.

2.4. Misspecification

Suppose that the regression model (1.3) is not correct. How does the OLS estimator perform? One way it could be incorrect is if the true regression function is non-linear. Suppose

$$E[Y|X = x] = g(x, \theta),$$

[11] In the previsou Section we used $H = X$.

equivalently

$$\begin{aligned} Y &= g(x,\theta) + \varepsilon \\ E[\varepsilon|X = x] &= 0, \end{aligned}$$

and define the n element vector

$$G(X,\theta) = \begin{bmatrix} g(x_1,\theta) \\ \vdots \\ g(x_n,\theta) \end{bmatrix}.$$

Then

$$\begin{aligned} E[\hat{\beta}|X] &= E[(X'X)^{-1}X'y|X] \\ &= (X'X)^{-1}X'G(X,\theta) \\ &\neq \beta. \end{aligned}$$

The OLS estimator is biased and its bias depends upon the values of x and the value of the parameter θ. Different researchers faced with different values of x will come to different conclusions about the value of β using the OLS estimator if they use a linear regression model and a non-linear model is appropriate.

The variance of the OLS estimator is

$$\begin{aligned} Var(\hat{\beta}|X) &= E[\left(\hat{\beta} - E[\hat{\beta}|X]\right)\left(\hat{\beta} - E[\hat{\beta}|X]\right)'|X] \\ &= E[(X'X)^{-1}X'\left(y - G(X,\theta)\right)\left(y - G(X,\theta)\right)'X\left(X'X\right)^{-1}|X] \\ &= E[(X'X)^{-1}X'\varepsilon\varepsilon'X\left(X'X\right)^{-1}|X] \\ &= (X'X)^{-1}X'\Sigma X\left(X'X\right)^{-1} \end{aligned}$$

exactly as it is when the regression function is correctly specified. Here $Var[\varepsilon|X] = \Sigma$. When $\Sigma = \sigma^2 I_n$, $Var(\hat{\beta}|X) = \sigma^2\left(X'X\right)^{-1}$.

2.5. Omitted regressors

In most of the standard textbooks there is an analysis of the bias of the OLS estimator when there are "omitted regressors". Suppose that conditional on X and on a matrix Z the expected value of y is $Z\gamma$, equivalently

$$\begin{aligned} y &= Z\gamma + \varepsilon \\ E[\varepsilon|X,Z] &= 0. \end{aligned}$$

Consider the OLS estimator, $\hat{\beta} = (X'X)^{-1}X'y$, calculated using data X. It is possible that X and Z have common columns. The expected value of the OLS estimator conditional on X and

Z is:

$$\begin{aligned} E[\hat{\beta}|X,Z] &= E[(X'X)^{-1}X'y|X,Z] \\ &= E[(X'X)^{-1}X'(Z\gamma+\varepsilon)|X,Z] \\ &= (X'X)^{-1}X'Z\gamma + (X'X)^{-1}X'E[\varepsilon|X,Z] \\ &= (X'X)^{-1}X'Z\gamma \end{aligned}$$

It is worth drawing out one implication of this. Let

$$Z = \left[\begin{array}{c|c} X & Q \end{array}\right], \qquad \gamma' = \left[\begin{array}{c|c} \gamma_X' & \gamma_Q' \end{array}\right]$$

so that the matrix X containing the values of the covariates used in calculating $\hat{\beta}$ is a part of the matrix Z. In the fitted model the variables Q have been "omitted". Then

$$\begin{aligned} E[\hat{\beta}|X,Z] &= E[\hat{\beta}|Z] \\ &= (X'X)^{-1}X'Z\gamma \\ &= (X'X)^{-1}\left[\begin{array}{c|c} X'X & X'Q \end{array}\right]\gamma \\ &= \left[\begin{array}{c|c} I & (X'X)^{-1}X'Q \end{array}\right]\gamma \\ &= \gamma_X + (X'X)^{-1}X'Q\gamma_Q. \end{aligned}$$

If $X'Q = 0$ or $\gamma_Q = 0$ then $E[\hat{\beta}|X,Z] = \gamma_X$, that is, when one or both of these conditions hold, the OLS estimator in the fitted model is an unbiased estimator of the coefficient on X in the extended model that includes the additional variables, Q. If values in the columns of X and/or Q are measured as deviations about column means then, if $X'Q = 0$ the values in X are uncorrelated[12] with the values in Q. So, omitting Q can lead to no bias in estimating the coefficient on X in the regression of y on X and Q if X and Q are uncorrelated.

2.6. Estimation of linear functions of β

Suppose we are interested in a particular linear combination of the elements of β, say $c'\beta$. For example the first element of β is obtained by setting

$$c' = [\ 1 \quad 0 \quad 0 \quad \ldots \quad 0\ ],$$

the sum of the first two elements of β is got by setting

$$c' = [\ 1 \quad 1 \quad 0 \quad \ldots \quad 0\ ].$$

[12] Let $\bar{X}$ be a matrix with the same number of rows and columns as X containing the column means of X repeated in each row and let

$$Z = \left[\begin{array}{c|c} X - \bar{X} & Q \end{array}\right]$$

which changes the value of any intercept in the equation for y. Since $(X-\bar{X})'Q = (X-\bar{X})'(Q-\bar{Q})$ where $\bar{Q}$ contains the column means of Q (check that $(X-\bar{X})'\bar{Q} = 0$), the condition $(X-\bar{X})'Q = 0$ implies that $Cov(X,Q) = 0$, since $Cov(X,Q) \equiv n^{-1}(X-\bar{X})'(Q-\bar{Q})$

The expected value of Y when $x = [\ x_1^* \quad x_2^* \quad x_3^* \quad \ldots \quad x_k^*\ ]$, which we might use in predicting the value of Y at $X = x^*$, is got by setting

$$c' = [\ x_1^* \quad x_2^* \quad x_3^* \quad \ldots \quad x_k^*\ ].$$

An obvious estimator of $c'\beta$ is $c'\hat{\beta}$ whose variance is

$$Var(c'\hat{\beta}|X) = \sigma^2 c' \left(X'X\right)^{-1} c.$$

2.7. The minimum variance property of OLS

The OLS estimator possesses an optimality property when $Var[\varepsilon|X] = \sigma^2 I_n$, namely that among the class of *linear* functions of y that are *unbiased* estimators of β the OLS estimator has the smallest variance, in the sense that, considering any other estimator, $\tilde{\beta} = Q(X)y$ (note, a linear function of y, with $Q(X)$ chosen so that $\tilde{\beta}$ is unbiased), $Var(c'\tilde{\beta}) \geq Var(c'\hat{\beta})$ for all c.

To show this result (embodied in what is known as the *Gauss-Markov theorem*), let $Q(X) = \left(X'X\right)^{-1} X' + R'$, where R may be a function of X, and note that

$$E[\tilde{\beta}|X] = \beta + R'X\beta.$$

This is equal to β for all β only when $R'X = 0$. This condition is required if $\tilde{\beta}$ is to be a linear unbiased estimator. Imposing that condition,

$$Var[\tilde{\beta}|X] - Var[\hat{\beta}|X] = \sigma^2 R'R,$$

and

$$Var(c'\tilde{\beta}) - Var(c'\hat{\beta}) = \sigma^2 d'd = \sigma^2 \sum_{i=1}^{k} d_i^2 \geq 0$$

where $d = Rc$.

2.8. (*) More data good, less data bad

The matrix $\left(X'X\right)^{-1}$, which along with σ^2 determines the variance of the OLS estimator, can be made smaller in a well defined sense by increasing the sample size. To see this, we use the following useful result[13] for nonsingular A and column vectors U and V.

$$\left(A + UV'\right)^{-1} = A^{-1} - \frac{A^{-1}UV'A^{-1}}{1 + U'A^{-1}V}$$

Let $A = X'X$ and let $U = V = X_a$ be a value of x at which an additional realisation of Y is obtained. Then the variance of $\hat{\beta}$ with the additional data is

$$\begin{aligned} Var(\hat{\beta}|X, X_a) &= \sigma^2 \left(X'X + X_a X_a'\right)^{-1} \\ &= \sigma^2 \left(X'X\right)^{-1} - \frac{\left(X'X\right)^{-1} X_a X_a' \left(X'X\right)^{-1}}{1 + X_a' \left(X'X\right)^{-1} X_a}, \end{aligned} \tag{2.3}$$

[13] Many of the econometric text books have a collections of matrix algebra results useful in developing estimatoirs and their properties. The following books contain many results not found in the standard texts. *Linear Statistical Inference and its Applications*, C. Radhakrshna Rao, John Wiley, New York: 1973, *The Algebra of Econometrics*, D.S.G. Pollock, John Wiley, New York: 1979, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, J.R. Magnus and H. Neudecker, John Wiley, New York: 1988.

and the variance of a linear combination of $\hat{\beta}$, $c'\hat{\beta}$ is

$$\begin{aligned} Var(c'\hat{\beta}|X, X_a) &= \sigma^2 c' \left(X'X + X_a X_a'\right)^{-1} c \\ &= \sigma^2 c' \left(X'X\right)^{-1} c - \frac{c' \left(X'X\right)^{-1} X_a X_a' \left(X'X\right)^{-1} c}{1 + X_a' \left(X'X\right)^{-1} X_a} \\ &= \sigma^2 c' \left(X'X\right)^{-1} c - \frac{\left(c' \left(X'X\right)^{-1} X_a\right)^2}{1 + X_a' \left(X'X\right)^{-1} X_a} \\ &\leq \sigma^2 c' \left(X'X\right)^{-1} c = Var(c'\hat{\beta}|X). \end{aligned}$$

So, more data never hurts and can reduce the variance of the estimator. In situations where data collection is costly (drug trials, oil exploration, industrial testing) one may want to know at what value of x to collect an additional realisation of Y in order to give a maximal reduction in the variance of an estimator of $c'\beta$. The result (2.3) is useful in this context. In survey design one might use the result to determine what sorts of e.g., households to oversample to get a better estimate of some important policy relevant magnitude.

2.9. M estimation

A completely different strategy for developing an estimator of β is to seek a statistic which results in an estimated regression function which "fits the data" as well as possible in some sense. At first sight this might seem an obvious strategy to adopt, but deeper thought raises some doubts.

There are some situations in which variation around the regression function is an *essential* part of the problem we are modelling, perhaps because luck, chance, or measurement error, are essential features of the process whereby data are generated.

In these situations, seeking the "best fitting" model can result in "overfitting" and misleading estimates of the true regression function. This is a particular problem in time series data analysis where searching for a good fitting model for the data available at one point in time can lead to disappointing predictions of future values of the response, Y. The so-called (and perhaps misnamed) "technical" financial trading literature abounds with examples of trading rules (sell after a "head and shoulders" profile in a share price) which seem to work when applied to historical data but fail to make money when applied in practice. Put simply you can always find a pattern in enough data - whether it will be repeated is another matter. In the old macroeconometric literature which was notable for the continual reanalysis of the same data set by different researchers, we saw searches for the best fitting model leading to some results in which quite strange patterns of dependence of a response on its historical values appear in fitted models.

Despite these caveats, let us consider what a "search for the best fit" strategy produces. The first problem to deal with is what is meant by the "best fit". Actually the earliest approach to this problem sought a statistic, say $\bar{\beta}$,

$$\bar{\beta} = \arg\min_b \sum_{i=1}^{n} \left|Y_i - b'x_i\right|.$$

Estimators obtained as the solution to an optimisation problem are given the generic name M-estimators.

This particular M-estimator, known as the *least absolute deviation* (LAD) estimator, was first used by the French mathematician, Laplace. He was attempting to measure the curvature of the Earth's surface, in particular the departure from sphericity of the Earth, using data in which there were many apparently aberrant observations, caused by the poor quality of early astronomical measuring devices[14]. The estimator has some remarkable robustness properties. It turns out that it can be completely insensitive to quite substantial changes in values of particular realisations of Y. As a result it is sometimes used when there is the possibility that response data are seriously contaminated by wild measurement errors. The estimator is a little awkward to compute (one method involves solution of a large linear programming problem) and it fell into disuse until recent advances in computing technology, since when it has made something of a comeback.

Another, and mathematically simpler, approach to the "best fit" problem involves using an estimator which minimises a smooth quadratic criterion, namely the estimator

$$\hat{\beta} = \arg\min_{b} \sum_{i=1}^{n} (Y_i - b'x_i)^2.$$

This "method of least squares" was discovered by the French mathematician, Legendre, who was using error contaminated data to calculate a measure of the length of a meridian quadrant, the distance from the equator to the north pole[15].

Differentiating with respect to b gives

$$-2\sum_{i=1}^{n} (Y_i - b'x_i)x_i' = -2\left(y'X - b'X'X\right)$$

The unique solution obtained for b when this is set equal to zero is (recall $rank(X) = k$)

$$\hat{\beta} = \left(X'X\right)^{-1} X'y,$$

the OLS estimator, just obtained by the analogue principle. The second derivative of the objective function is

$$\frac{\partial^2}{\partial b \partial b'} = 2\sum_{i=1}^{n} x_i x_i' = 2X'X.$$

This is positive definite (recall X has full rank) so the solution, $\hat{\beta}$, does locate the minimum of the objective function.

[14] *Mécanique Céleste Volumes 1 - 4*, Pierre Simon Laplace, trans N. Bowditch, 1829-1839, Hilliard, Gray, Little and Wilkins: Boston. See *The History of Statistics*, Stephen M Stigler, Harvard University Press: Harvard, 1986, for an excellent account of the development of the method of least squares.

[15] *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*, Adrien Marie Legendre, Coursier: Paris, 1805.

2.10. The Frisch-Waugh-Lovell Theorem

Suppose X is partitioned into two blocks, thus:

$$X = \left[\begin{array}{cc} X_1 & X_2 \end{array} \right]$$

so that

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

where β_1 and β_2 are elements of the conformable partition of β. Let

$$M_1 = I - X_1(X_1'X_1)^{-1}X_1'.$$

Then the Frisch-Waugh-Lovell Theorem tells us that, $\hat{\beta}_2$, the OLS estimator of β_2, that is the relevant part of

$$\hat{\beta} = (X'X)^{-1}X'y$$

can be written as

$$\begin{array}{rcl} \hat{\beta}_2 & = & \left((M_1X_2)'(M_1X_2)\right)^{-1}(M_1X_2)'M_1y \\ & = & \left(X_2'M_1X_2\right)^{-1}X_2'M_1y \end{array}$$

the second line following because M_1 is idempotent[16]. The following argument shows that the result is true.

Writing $X'y = (X'X)\hat{\beta}$ in partitioned form:

$$\left[\begin{array}{c} X_1'y \\ X_2'y \end{array} \right] = \left[\begin{array}{cc} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{array} \right] \left[\begin{array}{c} \hat{\beta}_1 \\ \hat{\beta}_2 \end{array} \right]$$

gives the following two matrix equations.

$$X_1'y = X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 \quad (2.4)$$

$$X_2'y = X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 \quad (2.5)$$

From (2.4)

$$\hat{\beta}_1 = \left(X_1'X_1\right)^{-1}X_1'y - \left(X_1'X_1\right)^{-1}X_1'X_2\hat{\beta}_2$$

substituting in (2.5)

$$X_2'y - X_2'X_1\left(X_1'X_1\right)^{-1}X_1'y = X_2'X_2\hat{\beta}_2 - X_2'X_1\left(X_1'X_1\right)^{-1}X_1'X_2\hat{\beta}_2$$

which after some rearrangement is

$$X_2'M_1y = \left(X_2'M_1X_2\right)\hat{\beta}_2$$

from which the result follows directly.

[16] An idempotent matrix A has the property $AA = A$. Check this is true for M_1.

The result has a nice interpretation and is useful in practice. The term M_1X_2 is just the matrix of residuals from OLS estimation of

$$X_2 = X_1\Psi + \Gamma$$

where Ψ is a matrix of coefficients and Γ is a matrix of "errors". The term M_1y is just the residuals from OLS estimation of

$$y = X_1\gamma + \nu.$$

So to get the OLS estimate of β_2 in the full model we can perform OLS estimation using residuals as left and right hand side variables. This also motivates the study of residual-residual plots when studying model specification.

3. Generalised least squares estimation

The simple result $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$ is true when $Var(\varepsilon|X) = \sigma^2 I_n$ which is independent of X. There are many situations in which we would expect to find some dependence on X so that $Var[\varepsilon|X] \neq \sigma^2 I_n$.

For example in a household expenditure survey we might expect to find people with high values of time purchasing large amounts infrequently (e.g. of food, storing purchases in a freezer) and poor people purchasing small amounts frequently. If we just observed households' expenditures for a week (as in the British National Food Survey[17] conducted by the Department for Environment, Food and Rural Affairs) then we would expect to see that, conditional on variables X that are correlated with the value of time, the variance of expenditure depends on X. When this happens we talk of the disturbances, ε, as being *heteroskedastic*.

In other contexts we might expect to find correlation among the disturbances, in which case we talk of the disturbances as being *serially correlated*. An example, again in the context of a household survey, was given at the start of Section 2. Serial correlation of a variety of forms frequently arises in time series analysis, to be considered later in the course.

The BLU property of the OLS estimator does not usually apply when $Var[\varepsilon|X] \neq \sigma^2 I_n$. To get some insight into why this is the case, suppose that Y has a much larger conditional variance at one value of x, x^*, than at other values. Realisations produced at x^* will less informative about the location of the regression function than realisations obtained at other values of x. It seems natural to give realisations obtained at x^* less weight when estimating the regression function.

We know how to produce a BLU estimator when $Var[\varepsilon|X] = \sigma^2 I_n$. Our strategy for producing a BLU estimator when this condition does not hold is to transform the original regression model so that the conditional variance of the transformed Y is proportional to an identity matrix and apply the OLS estimator in the context of that transformed model.

Suppose $Var[\varepsilon|X] = \Sigma$ is positive definite. Of course it is always symmetric. Then we can find[18] a matrix P such that $P\Sigma P' = I$. Consider the random vector $z = Py$ which

[17] Since April 2001 merged with the Family Expenditure Survey to produce the Expenditure and Food Survey. Annual reports are available at *http://www.defra.gov.uk/esg/m_publications.htm.*

[18] Let Λ be a diagonal matrix with the (positive valued) eigenvalues of Σ on its diagonal, and let C be the matrix of associated orthonormal eigenvectors. Then $C\Sigma C' = \Lambda$ and so $\Lambda^{-1/2}C\Sigma C'\Lambda^{-1/2} = I$. The required matrix P is $\Lambda^{-1/2}C$.

has conditional expected value given X equal to $PX\beta$ and conditional variance, $P\Sigma P' = I$. Therefore we can write

$$z = Py = PX\beta + u$$

where $u = P\varepsilon$ and $Var[u|X] = I$, and in the context of this model the OLS estimator,

$$\breve{\beta} = (X'P'PX)^{-1}X'P'Py,$$

does possess the BLU property. Further, its conditional variance given X is $(X'P'PX)^{-1}$. Since $P\Sigma P' = I$, it follows that $\Sigma = P^{-1}P'^{-1} = (P'P)^{-1}$, so that $P'P = \Sigma^{-1}$. The estimator $\breve{\beta}$, and its conditional mean and variance can therefore be written as[19]

$$\begin{aligned} \breve{\beta} &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \\ E[\breve{\beta}|X] &= \beta \\ Var[\breve{\beta}|X] &= (X'\Sigma^{-1}X)^{-1} \end{aligned}$$

The estimator is known as the *generalised least squares* (GLS) estimator. Obviously the estimator cannot be calculated unless Σ is known which is rarely the case. However sometimes it is possible to produce a well behaved estimator $\hat{\Sigma}$ in which case the *feasible GLS estimator*

$$\breve{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}y$$

could be used. To study the properties of this estimator requires the use of asymptotic approximations and we return to this later.

3.1. Feasible GLS estimation and robust standard errors

To produce the feasible GLS estimator we must impose some structure on the variance matrix of the unobservables, Σ. If we did not then we would have to estimate $n(n+1)/2$ parameters (the number of distinct elements of Σ) using data containing just n observations which is infeasible.

One way to proceed is to impose the restriction that the diagonal elements of Σ are constant and allow nonzero off diagonal elements but only close to the main diagonal of Σ. This requires ε to have homoskedastic variation with X but allows a degree of correlation between values of ε for observations that are close together (e.g. in time if the data are in time order in the vector y). One could impose a parametric model on the variation of elements of Σ. You will learn more about this in the part of the course dealing with time series.

Where heteroskedasticity is likely to arise, a parametric approach is occasionally employed, using a model that requires σ_{ii} (the ith main diagonal element of Σ) to be a parametric function of x_i (the ith row of the X matrix). For example if the model $\sigma_{ii} = \gamma' x_i$ is thought to be appropriate then one could estimate γ, for example by calculating an OLS estimator of γ in the model with equation

$$\hat{\varepsilon}_i^2 = \gamma' x_i + u_i$$

where $\hat{\varepsilon}_i^2$ is the squared ith residual from an OLS estimation[20]. Then an estimate of Σ could be produced using $\hat{\gamma}' x_i$ as the ith main diagonal element.

[19] Note that this is an "analogue estimator" as in Section 2.3 with $H = \Sigma^{-1}X$.

[20] Of course σ_{ii} must be non-negative, so an alternative functional form might be appropriate. Also, if we use OLS residuals $E[\hat{\varepsilon}_i^2|X] = \sigma^2 + r(n)$ where $r(n)$ is not in general zero but tends to zero as n increases. Feasible GLS using a parametric model of this sort for the elements of Σ may not work well when samples are small.

Economics rarely suggests suitable parametric models for variances of unobservables, though graphical analysis of residuals may provide some guidance. One may therefore not which to pursue the gains in efficiency that GLS in principle offers. But, if the OLS estimator is used and $\Sigma \neq \sigma^2 I_n$ one must still be aware that the formula yielding standard errors, $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$, correct if $\Sigma = \sigma^2 I_n$, is generally *incorrect* if $\Sigma \neq \sigma^2 I_n$, in which case

$$Var(\hat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}.$$

One popular strategy is to proceed with the OLS estimator but to use an estimate of the matrix $(X'X)^{-1}X'\Sigma X(X'X)^{-1}$ to construct standard errors. In models in which the off-diagonal elements of Σ are zero but heteroskedasticity is potentially present this can be done by using

$$Var(\hat{\beta}) = (X'X)^{-1}X'\hat{\Sigma}X(X'X)^{-1}.$$

where $\hat{\Sigma}$ is a diagonal matrix with squared OLS residuals, $\hat{\varepsilon}_i^2$, on its main diagonal. The (heteroskedasticity) *robust standard errors* that result, in large samples in which there are is no extreme variation in the elements of X, generally give a good indication of the accuracy of OLS estimation.[21]

4. Inference

4.1. Sampling distributions

For now suppose that y given X (equivalently ε given X) is *normally distributed*.

The OLS estimator is a *linear* function of y and is therefore, conditional on X, normally distributed. The same argument applies to the GLS estimator (employing Σ, rather than $\hat{\Sigma}$), so we have, when $Var[\varepsilon|X] = \sigma^2 I$, for the OLS estimator

$$\hat{\beta}|X \sim N_k[\beta, \sigma^2(X'X)^{-1}]$$

and when $Var[\varepsilon|X] = \Sigma$, for the GLS estimator,

$$\breve{\beta}|X \sim N_k[\beta, (X'\Sigma^{-1}X)^{-1}].$$

Consider a linear combination of β, $c'\beta$. From now on unless noted we suppose that $Var[\varepsilon|X] = \sigma^2 I$.

We have, for the OLS estimator

$$c'\hat{\beta}|X \sim N[c'\beta, \sigma^2 c'(X'X)^{-1}c].$$

Let $Z \sim N[0,1]$ and let $z_L(\alpha)$ and $z_U(\alpha)$ be the closest pair of values such that $P[z_L(\alpha) \leq Z \leq z_U(\alpha)] = \alpha$. Since the standard normal density function is symmetric around zero, $z_L(\alpha) = -z_U(\alpha)$ and $z_L(\alpha)$ is the $(1-\alpha)/2$ quantile of the standard normal distribution. Choosing $\alpha = 0.95$ gives $z_U(\alpha) = 1.96$, $z_L(\alpha) = -1.96$.

[21] References on this topic can be found in Davidson and McKinnon's text book. There are similar procedures available when off diagonal elements of Σ are potentially non-zero.

The result above concerning the distribution of $c'\hat{\beta}$ implies that

$$P[z_L(\alpha) \leq \frac{c'\hat{\beta} - c'\beta}{\sigma\left(c'(X'X)^{-1}c\right)^{1/2}} \leq z_U(\alpha)] = \alpha \tag{4.1}$$

which in turn implies that

$$P[c'\hat{\beta} - z_U(\alpha)\sigma\left(c'(X'X)^{-1}c\right)^{1/2} \leq c'\beta \leq c'\hat{\beta} - z_L(\alpha)\sigma\left(c'(X'X)^{-1}c\right)^{1/2}] = \alpha.$$

Consider the interval

$$[c'\hat{\beta} - z_U(\alpha)\sigma\left(c'(X'X)^{-1}c\right)^{1/2}, c'\hat{\beta} - z_L(\alpha)\sigma\left(c'(X'X)^{-1}c\right)^{1/2}].$$

This random interval covers the value $c'\beta$ with probability α. This is known as a $100\alpha\%$ confidence interval for $c'\beta$. Note that this interval cannot be calculated without knowledge of σ. In practice here and in the tests and interval estimators that follow one will use an estimator of σ^2. In many of problems considered in econometrics the effect of using an estimator is negligible.

4.2. Estimation of σ

Note that

$$\sigma^2 = n^{-1}E[(y - X\beta)'(y - X\beta)\,|X]$$

which suggests the analogue estimator

$$\begin{aligned}\hat{\sigma}^2 &= n^{-1}\left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right)\\ &= n^{-1}\hat{\varepsilon}'\hat{\varepsilon}\\ &= n^{-1}y'My\end{aligned}$$

where $\hat{\varepsilon} = y - X\hat{\beta} = My$ and $M = I - X(X'X)^{-1}X'$ where note $MX = 0$.

The elements of $\hat{\varepsilon}$ are called residuals. They measure how far each element of y is from the associated value on the estimated regression function, $X\hat{\beta}$. The OLS estimator minimises the sum of squared residuals.

It is in some sense because the OLS has this *minimising* property that $\hat{\sigma}^2$ is a biased estimator and that the bias is in the *downward* direction. In fact

$$E[\hat{\sigma}^2] = \frac{n-k}{n}\sigma^2 < \sigma^2$$

but note that the bias is negligible unless k the number of covariates is large relative to n the sample size. It is obviously possible to correct the bias using the estimator $(n-k)^{-1}\hat{\varepsilon}'\hat{\varepsilon}$ but the effect is small in most economic data sets.

The expected value of $\hat{\sigma}^2$ is obtained as follows. First note that $My = M\varepsilon$ because $MX = 0$. So $\hat{\sigma}^2 = n^{-1}y'My = n^{-1}\varepsilon'M\varepsilon$. Then there is the following

$$\begin{aligned}E[\hat{\sigma}^2|X] &= n^{-1}E[\varepsilon'M\varepsilon|X]\\ &= n^{-1}E[\text{trace}(\varepsilon'M\varepsilon)|X]\\ &= n^{-1}E[\text{trace}(M\varepsilon\varepsilon')|X]\\ &= n^{-1}\,\text{trace}(ME[\varepsilon\varepsilon'|X])\\ &= n^{-1}\,\text{trace}(M\Sigma)\end{aligned}$$

and when $\Sigma = \sigma^2 I_n$,

$$\begin{aligned} n^{-1}\,\text{trace}(M\Sigma) &= n^{-1}\sigma^2\,\text{trace}(M) \\ &= n^{-1}\sigma^2\,\text{trace}(I_n - X(X'X)^{-1}X') \\ &= \sigma^2\frac{n-k}{n}. \end{aligned}$$

Under certain conditions to be discussed shortly the estimator $\hat{\sigma}^2$ is *consistent*. This means that in large samples the inaccuracy of the estimator is small and that if in the tests described below the unknown σ^2 is replaced by $\hat{\sigma}^2$ the tests are still approximately correct. This will be elucidated in the notes on asymptotic approximations.

4.3. Confidence regions

Sometimes we need to make probability statements about the values of more than one linear combination of β. We can do this by developing *confidence regions*. For j linear combinations a $100\alpha\%$ confidence region is a subset of $\Re^j$ which covers the unknown (vector) value of the j linear combinations with probability α.

Continue to work under the assumption that y given X (equivalently ε given X) is *normally distributed*.

Let the j linear combinations of interest be $R\beta = r$, say, where R is $j \times k$ with rank j. The OLS estimator of r is $R\hat{\beta}$ and

$$R\hat{\beta} \sim N[r, \sigma^2 R(X'X)^{-1}R']$$

which implies that

$$\left(R\hat{\beta} - r\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\beta} - r\right)/\sigma^2 \sim \chi^2_{(j)} \tag{4.2}$$

where $\chi^2_{(j)}$ denotes a *Chi-square* random variable with parameter (degrees of freedom) j.

4.4. The Chi-square distribution

Let the ν element vector $Z \sim N(0, I_\nu)$. Then $\xi = Z'Z = \sum_{i=1}^{\nu} Z_i^2$ (positive valued) has a distribution known as a Chi-square distribution, written $Z \sim \chi^2_{(\nu)}$. The probability density function associated with the $\chi^2_{(\nu)}$ distribution is positively skewed. For small ν its mode is at zero. The expected value[22] and variance of a Chi-square random variable are

$$\begin{aligned} E[\chi^2_{(\nu)}] &= \nu \\ Var[\chi^2_{(\nu)}] &= 2\nu. \end{aligned}$$

For large ν, the distribution is approximately normal.

[22] If $Z_i \sim N(0,1)$ then $V[Z_i] = E[Z_i^2] = 1$. Therefore $E[\Sigma_{i=1}^{v} Z_i^2] = v$. The variance result is more difficult to get. But you could try.

The following result will be used below. Let $A \sim N_\nu[\mu, \Sigma]$ and let P be such that $P\Sigma P' = I$, which implies that $P'P = \Sigma^{-1}$. Then $Z = P(A - \mu) \sim N_\nu[0, I]$ so that

$$\xi = Z'Z = (A - \mu)'\Sigma^{-1}(A - \mu) \sim \chi^2_{(\nu)}.$$

To obtain the result (4.2) this is applied with $\mu = r$ and $\Sigma = \sigma^2 R(X'X)^{-1}R'$.

4.5. Confidence regions continued

Let $q_{\chi^2(j)}(\alpha)$ denote the $\alpha-$quantile of the $\chi^2_{(j)}$ distribution. Then

$$P[\chi^2_{(j)} \leq q_{\chi^2(j)}(\alpha)] = \alpha$$

implies that

$$P[\left(R\hat{\beta} - r\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\beta} - r\right) / \sigma^2 \leq q_{\chi^2(j)}(\alpha)] = \alpha.$$

The region in $\Re^j$ defined by

$$\{r : \left(R\hat{\beta} - r\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\beta} - r\right) / \sigma^2 \leq q_{\chi^2(j)}(\alpha)\}$$

is a $100\alpha\%$ confidence region for r, covering r with probability α. The boundary of the region is an ellipsoid centred on the point $R\hat{\beta}$.

Setting R equal to a vector c' (note then $j = 1$) and letting $c^* = c'\beta$, produces

$$\begin{aligned}
\alpha &= P[\left(c'\hat{\beta} - c^*\right)' \left(c'(X'X)^{-1}c\right)^{-1} \left(c'\hat{\beta} - c^*\right) / \sigma^2 \leq q_{\chi^2(1)}(\alpha)] \\
&= P[\frac{\left(c'\hat{\beta} - c^*\right)^2}{\sigma^2 c'(X'X)^{-1}c} \leq q_{\chi^2(1)}(\alpha)] \\
&= P[-\left(q_{\chi^2(1)}(\alpha)\right)^{1/2} \leq \frac{\left(c'\hat{\beta} - c^*\right)}{\sigma \left(c'(X'X)^{-1}c\right)^{1/2}} \leq \left(q_{\chi^2(1)}(\alpha)\right)^{1/2}] \\
&= P[z_L(\alpha) \leq \frac{\left(c'\hat{\beta} - c^*\right)}{\sigma \left(c'(X'X)^{-1}c\right)^{1/2}} \leq z_U(\alpha)]
\end{aligned}$$

where we have used the relationship $\chi^2(1) = N(0,1)^2$. The last line here agrees with (4.1), so when $j = 1$, the confidence region just derived agrees with the confidence interval derived earlier.

4.6. Tests of hypotheses

The statistics developed to construct confidence intervals can also be used to conduct tests of hypotheses. For example, suppose we wish to conduct a test of the null hypothesis $H_0 : R\beta - r = 0$ against the alternative $H_1 : R\beta - r \neq 0$. The statistic

$$S = \left(R\hat{\beta} - r\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\beta} - r\right) / \sigma^2. \tag{4.3}$$

has a $\chi^2(j)$ distribution under the null hypothesis. Under the alternative, let

$$R\beta - r = \delta \neq 0.$$

Then

$$R\hat{\beta} - r \sim N[\delta, \sigma^2 R(X'X)^{-1}R']$$

and the statistic S will tend to be larger than we would expect to obtain from a $\chi^2(j)$ distribution. So we reject the null hypothesis for large values of S. The following test procedure has size[23] λ.

Decision rule: Reject H_0 if $S > q_{\chi^2(j)}(1-\lambda)$, otherwise do not reject H_0.

Here $q_{\chi^2(j)}(1-\lambda)$ is the $(1-\lambda)$ quantile of the $\chi^2(j)$ distribution. Note that we do *not* talk in terms of accepting H_0 as an alternative to rejection. The reason is that a value of S that does not fall in the rejection region of the test is consonant with many values of $R\beta - r$ that are close to but not equal to 0.

To obtain a test concerning a *single* linear combination of β, $H_0 : c'\beta = c^*$, we can use the procedure above with $j = 1$, giving

$$S = \frac{\left(c'\hat{\beta} - c^*\right)^2}{\sigma^2 c'(X'X)^{-1}c}$$

and the following size λ test procedure.

Decision rule: Reject H_0 if $S > q_{\chi^2(1)}(1-\lambda)$, otherwise do not reject H_0.

Alternatively we can proceed directly from the sampling distribution of $c'\hat{\beta}$. Since, when H_0 is true,

$$\frac{\left(c'\hat{\beta} - c^*\right)}{\sigma\left(c'(X'X)^{-1}c\right)^{1/2}} \sim N(0,1),$$

we can obtain $z_L(\alpha)$, $z_U(\alpha)$, such that

$$P[z_L(\alpha) < N(0,1) < z_U(\alpha)] = \alpha = 1 - \lambda.$$

The following test procedure has size (probability of rejecting a true null hypothesis) equal to λ.

Decision rule: Reject H_0 if $S > z_U(\alpha)$ or $S < z_L(\alpha)$, otherwise do not reject H_0.

Because of the relationship between the standard normal $N(0,1)$ distribution and the $\chi^2_{(1)}$ distribution the tests are identical.

[23] The *size* of a test of H_0 is the probability of rejecting H_0 when H_0 is true. The *power* of a test against a specific alternative H_1 is the probability of rejecting H_0 when H_1 is true.

4.7. Restricted least squares

Another way of testing the hypothesis $H_0 : R\beta - r = 0$ is to force the estimate $\hat{\beta}_R$ to obey the restriction $R\hat{\beta}_R - r$ and then to estimate β unrestrictedly and to compare the quality of the fit of the two estimated models, as captured by the mean squared residuals. Let the residuals from the restricted and unrestricted estimation be respectively $\hat{\varepsilon}_R$ and $\hat{\varepsilon}_U$. Then, under the null hypothesis, the difference in the sum of squared residuals divided by σ^2 has a $\chi^2_{(j)}$ distribution.[24]

$$\frac{\left(\hat{\varepsilon}'_R\hat{\varepsilon}_R - \hat{\varepsilon}'_U\hat{\varepsilon}_U\right)}{\sigma^2} \sim \chi^2_{(j)}.$$

We now show that this statistic is identical to (4.3).

Restricted estimation can be done by defining $\hat{\beta}_R$ as follows.

$$\hat{\beta}_R = \arg\min_b (y - Xb)'(y - Xb), \qquad \text{subject to: } Rb = r$$

Define the Lagrangian

$$L = (y - Xb)'(y - Xb) - 2\lambda'(Rb - r)$$

leading to the following first order conditions.

$$\begin{aligned} -2X'y + 2X'X\hat{\beta}_R - 2R'\hat{\lambda} &= 0 \qquad (4.4)\\ R\hat{\beta}_R - r &= 0 \qquad (4.5)\end{aligned}$$

The first equation gives[25]

$$\hat{\beta}_R = \hat{\beta}_U + (X'X)^{-1}R'\hat{\lambda}$$

where $\hat{\beta}_U = (X'X)^{-1}X'y$ is the unrestricted OLS estimator, and multiplying by R and using the first order condition (4.5) gives

$$R\hat{\beta}_R = r = R\hat{\beta}_U + R(X'X)^{-1}R'\hat{\lambda} \qquad (4.6)$$

which can be solved for $\hat{\lambda}$, giving

$$\hat{\lambda} = -\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta}_U - r\right).$$

Substituting for $\hat{\lambda}$ in (4.6) gives the restricted least squares (RLS) estimator.

$$\hat{\beta}_R = \hat{\beta}_U - (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta}_U - r\right).$$

The RLS residual vector is

$$\begin{aligned}\hat{\varepsilon}_R &= y - X\hat{\beta}_R\\ &= y - X\hat{\beta}_U + X(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta}_U - r\right)\\ &= \hat{\varepsilon}_U + X(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta}_U - r\right).\end{aligned}$$

[24] In practice we replace σ^2 by the consistent estimator $\hat{\sigma}^2$.

[25] Multiply both sides by $(X'X)^{-1}$.

Therefore, after some simplification

$$\hat{\varepsilon}_R'\hat{\varepsilon}_R = \hat{\varepsilon}_U'\hat{\varepsilon}_U + \left(R\hat{\beta}_U - r\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\beta}_U - r\right)$$

- compare with (4.3).

4.8. Detecting structural change

A common application of this testing procedure in econometrics arises when attempting to detect "structural change". In a time series application one might imagine that up to some time T_s the vector $\beta = \beta_b$ and after T_s, $\beta = \beta_a$, that is that there are two regimes with switching occurring at time T_s. This situation can be captured by specifying the model

$$y = \begin{bmatrix} y_b \\ y_a \end{bmatrix} = \begin{bmatrix} X_b & 0 \\ 0 & X_a \end{bmatrix} \begin{bmatrix} \beta_b \\ \beta_a \end{bmatrix} + \begin{bmatrix} \varepsilon_b \\ \varepsilon_a \end{bmatrix} = X\beta + \varepsilon$$

where X_b contains data for the period before T_s and X_a contains data for the period after T_s. The null hypothesis of no structural change is expressed by $H_0 : \beta_b = \beta_a$. If all the coefficients are allowed to alter across the structural break then

$$\hat{\varepsilon}_U'\hat{\varepsilon}_U = \hat{\varepsilon}_b'\hat{\varepsilon}_b + \hat{\varepsilon}_a'\hat{\varepsilon}_a$$

where, e.g., $\hat{\varepsilon}_b'\hat{\varepsilon}_b$ is the sum of squared residuals from estimating

$$y_b = X_b\beta_b + \varepsilon_b.$$

The test statistic developed above, specialised to this problem can then be written

$$S = \frac{\left(\hat{\varepsilon}'\hat{\varepsilon} - \left(\hat{\varepsilon}_b'\hat{\varepsilon}_b + \hat{\varepsilon}_a'\hat{\varepsilon}_a\right)\right)}{\sigma^2}$$

where $\hat{\varepsilon}'\hat{\varepsilon}$ is the sum of squared residuals from estimating with the constraint $\hat{\beta}_a = \hat{\beta}_b$ imposed and σ^2 is the common variance of the errors. When the errors are identically and independently normally distributed S has a $\chi^2_{(k)}$ distribution under H_0. In practice an estimate of σ^2 is used - for example there is the statistic

$$S^* = \frac{\left(\hat{\varepsilon}'\hat{\varepsilon} - \left(\hat{\varepsilon}_b'\hat{\varepsilon}_b + \hat{\varepsilon}_a'\hat{\varepsilon}_a\right)\right)}{\left(\hat{\varepsilon}_b'\hat{\varepsilon}_b + \hat{\varepsilon}_a'\hat{\varepsilon}_a\right)/n}$$

where n is the total number of observations in the two periods combined. S has approximately a $\chi^2_{(k)}$ distribution under H_0. This application of the theory of tests of linear hypotheses is given the name, "Chow test", after Gregory Chow who popularised the procedure some 30 years ago. The version of the test popularised by Chow employs degree of freedom corrections which lead to a test statistic with exactly an F distribution under H_0 when the errors are normally distributed. In practice the normal assumption is unlikely to hold so we do not give details here.

Of course the test can be modified in various ways. For example we might wish to keep some of the elements of β constant across regimes. The procedure can be modified to produce

a statistic with which to measure the quality of forecasts from a model. In macroeconometrics there is considerable interest in detecting the number and location of regime shifts as well as the changes in coefficients across regimes. This is an active area of research which we will not have time to cover. In microeconometrics the same procedure can be employed to test for differences across groups of households, firms etc.

We saw how pivotal statistics played a central role in developing confidence intervals. They play a similarly central role in the development of tests of hypotheses. In the classical, Fisherian, inference that we are studying, to conduct a hypothesis test we develop a statistic that is pivotal when the null hypothesis is true but which has a distribution dependent on the extent of departure from the null hypothesis when that hypothesis is false. Probability statements about the pivotal statistic are transformed into a "Reject - Do not reject" decision rule. A powerful test is one whose distribution under the alternative hypothesis shows large departures from the distribution of the pivotal statistic that obtains under the null hypothesis.

5. Estimation in non-linear regression models

An obvious extension to the linear regression model studied so far is the non-linear regression model[26]:

$$E[Y|X = x] = g(x, \theta)$$

equivalently, in regression function plus error form:

$$\begin{aligned} Y &= g(x, \theta) + \varepsilon \\ E[\varepsilon|X = x] &= 0. \end{aligned}$$

Consider M-estimation and in particular the non-linear least squares estimator obtained as follows.

$$\hat{\theta} = \arg\min_{\theta^*} n^{-1} \sum_{i=1}^{n} \left(Y_i - g(x_i; \theta^*)\right)^2$$

For now we just consider how a minimising value $\hat{\theta}$ can be found. Many of the statistical software packages have a routine to conduct non-linear optimisation and some have a non-linear least squares routine. Many of these routines employ a variant of Newton's method, which proceeds as follows.

5.1. (*) Numerical optimisation: Newton's method and variants

Write the minimisation problem as:

$$\hat{\theta} = \arg\min_{\theta^*} Q(\theta^*).$$

[26] This is not the most general form that can arise. For example models of the form

$$E[g(Y, x; \theta)|X = x] = 0$$

equivalently

$$\begin{aligned} g(Y, x; \theta) &= \varepsilon \\ E[\varepsilon|X = x] &= 0 \end{aligned}$$

do occur, but we will not study these sorts of models in his course.

Newton's method involves taking a sequence of steps, $\theta_0, \theta_1, \ldots, \theta_m, \ldots \theta_M$ from a starting value, θ_0 to an approximate minimising value θ_M which we will use as our estimator $\hat{\theta}$. The starting value is provided by the user. One of the tricks is to use a good starting value near to the final solution. This sometimes requires some thought.

Suppose we are at θ_m. Newton's method considers a quadratic approximation to $Q(\theta)$ which is constructed to be an accurate approximation in a neighbourhood of θ_m, and moves to the value θ_{m+1} which minimises this quadratic approximation. At θ_{m+1} a new quadratic approximation, accurate in a neighbourhood of θ_{m+1} is constructed and the next value in the sequence, θ_{m+2}, is chosen as the value of θ minimising this new approximation. Steps are taken until a convergence criterion is satisfied. Usually this involves a number of elements. For example one might continue until the following conditions is satisfied:

$$Q_\theta(\theta_m)' Q_\theta(\theta_m) \leq \delta_1, \qquad |Q(\theta_m) - Q(\theta_{m-1})| < \delta_2.$$

Convergence criteria vary form package to package. Some care is required in choosing these criteria. Clearly δ_1 and δ_2 above should be chosen bearing in mind the orders of magnitude of the objective function and its derivative.

The quadratic approximation used at each stage is a quadratic Taylor series approximation. At $\theta = \theta_m$,

$$Q(\theta) \simeq Q(\theta_m) + (\theta - \theta_m)' Q_\theta(\theta_m) + \frac{1}{2} (\theta - \theta_m)' Q_{\theta\theta'}(\theta_m) (\theta - \theta_m) = Q^a(\theta, \theta_m).$$

The derivative of $Q^a(\theta, \theta_m)$ with respect to θ is

$$Q^a_\theta(\theta, \theta_m) = Q_\theta(\theta_m) + Q_{\theta\theta'}(\theta_m) (\theta - \theta_m)$$

and θ_{m+1} is chosen as the value of θ that solves $Q^a_\theta(\theta, \theta_m) = 0$, namely

$$\theta_{m+1} = \theta_m - Q_{\theta\theta'}(\theta_m)^{-1} Q_\theta(\theta_m).$$

There are a number of points to consider here.

1. Obviously the procedure can only work when the objective function is twice differentiable with respect to θ. For example it doesn't work for the non-linear least absolute deviation estimator where
$$Q(\theta) = n^{-1} \sum_{i=1}^{n} |Y_i - g(x_i; \theta)|^2 .$$

2. The procedure as described above, will stop whenever $Q_\theta(\theta_m) = 0$, which can occur at a maximum and saddlepoint as well as at a minimum. The Hessian, $Q_{\theta\theta'}(\theta_m)$, should be positive definite at a minimum of the function, and this should be checked at any claimed minimising value.

3. When a minimum is found there is no guarantee that it is a global minimum. In problems where this possibility arises it is normal to run the optimisation from a variety of start points to guard against using an estimator that corresponds to a local minimum.

4. If, at a point in the sequence, $Q_{\theta\theta'}(\theta_m)$ is not positive definite then the algorithm may move away from the minimum and there may be no convergence. Many minimisation (maximisation) problems we deal with involve globally convex (concave) objective functions and for these there is no problem. For other cases, Newton's method is usually modified, e.g. by taking steps

$$\theta_{m+1} = \theta_m - A(\theta_m)^{-1} Q_\theta(\theta_m)$$

where $A(\theta_m)$ is constructed to be positive definite and in cases in which $Q_{\theta\theta'}(\theta_m)$ is in fact positive definite, to be a good approximation to $Q_{\theta\theta'}(\theta_m)$.

5. The algorithm may "overstep" the minimum to the extent that it takes an "uphill" step, i.e. so that $Q(\theta_{m+1}) > Q(\theta_m)$. This is guarded against in many implementations of Newton's method by taking steps

$$\theta_{m+1} = \theta_m - \alpha(\theta_m) A(\theta_m)^{-1} Q_\theta(\theta_m)$$

where $\alpha(\theta_m)$ is a scalar step scaling factor, chosen to ensure that $Q(\theta_{m+1}) < Q(\theta_m)$. Some implementations attempt to make an optimal choice of $\alpha(\theta_m)$ at each step. For example candidate values of $\alpha(\theta_m)$ can be employed, $\alpha_i(\theta_m)$, $i = 1, 2, 3$, and a quadratic approximation fitted to the resulting values of the objective function. Then $\alpha(\theta_m)$ is chosen as the value of the step scale factor which minimises this quadratic approximation.

6. In practice it may be difficult to calculate exact expressions for the derivatives that appear in Newton's method. In some cases symbolic computational methods can help. In others we can use a numerical approximation, e.g.

$$Q_{\theta_i}(\theta_m) \simeq \frac{Q_\theta(\theta_m + \delta_i e_i) - Q_\theta(\theta_m)}{\delta_i}$$

where e_i is a vector with a one in position i and zeros elsewhere, and δ_i is a small perturbing value, possibly varying across the elements of θ.

5.2. Inference using the NLS estimator and in non-normal models

The NLS estimator is a nonlinear function of y and this turns out to present very difficult problems when we come to develop methods for making exact inferences about the unknown parameter vector θ.

In our work on inference so far we have only considered cases in which y given X is normally distributed and only estimators that are simple functions of y - the OLS and GLS estimators, which are linear functions of y, and s^2 which is a quadratic function of y. In these cases it is fairly easy to develop the distributions of the estimators. Once we consider more complicated functions of y the development of an exact distribution theory is usually technically very demanding and often computationally infeasible. Further it is often the case that the resulting exact distribution depends upon the unknown parameters in a complicated way so that it is not possible to develop pivotal statistics to use in the production of exact confidence intervals and tests of hypotheses.

The same problem arises when we consider the OLS and GLS estimator, and s^2, and y given X is *non-normally* distributed. A further difficulty is that we usually do not know what the

distribution of y given X actually is, so even if we were able to develop exact inferential methods we would not know which distribution for y to base them on.

There are a number of ways of at least partially getting round this problem. We will look at one of these next, namely the use of approximation methods, specifically what are known as *large sample approximations*. They are universally applied in econometrics.