EC403, Part 1

Oliver Linton

October 14, 2002

Contents

Chapter 1

Linear Regression

1.1 The Model

- The first part of the course will be concerned with estimating and testing in the linear model. We will suggest procedures and derive their properties under certain assumptions. The linear model is the basis for most of econometrics and a firm grounding in this theory is essential for future work.

- We observe the following data

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad ;$$

$$X = \begin{pmatrix} x_{11} & \cdots & x_{K1} \\ \vdots & & \vdots \\ x_{1n} & & x_{Kn} \end{pmatrix} = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix},$$

  where $\text{rank}(X) = K$. Note that this is an assumption, but it is immediately verifiable from the data in contrast to some other assumptions we will make.

- It is desirable for statistical analysis to specify a model of how these data were generated. We suppose that there is a random mechanism which is behind everything - the data we have is one realisation of

an infinity of such potential outcomes. We shall make the following assumptions regarding the way y, X were generated:

- *Fixed Design Linear Model*
  - (A1) X is fixed in repeated samples
  - (A2) $\exists \beta = (\beta_1, \ldots, \beta_K)'$ such that $E(y) = X\beta$.
  - (A3) $\text{Var}(y) = \sigma^2 I_{n \times n}$.

- We stick with fixed design for most of the linear regression section. Fixed design is perhaps unconvincing for most economic data sets, because of the asymmetry between y and x. That is, in economic datasets we have no reason to think that some data were randomly generated while others were fixed. This is especially so in time series when one regressor might be a lagged value of the dependent variable.

- A slightly different specification is *Random Design Linear Model*
  - (A1r) X is random with respect to repeated samples
  - (A2r) $\exists \beta$ s.t. $E(y|X) = X\beta$
  - (A3r) $\text{Var}(y|X) = \sigma^2 I_{n \times n}$,

where formally A2r and A3r hold with probability one.

- However, one can believe in a random design model, but want to conduct inference in the conditional distribution [given X]. This is sensible at least in the cross-section case where there are no lagged dependent variables. In this case, we are effectively working in a fixed design model. So the real distinction in this case is whether one evaluates quantities in the conditional or unconditional distribution.

- Finally, we write the regression model in the more familiar form. Define $\varepsilon = y - X\beta = (\varepsilon_1, \ldots, \varepsilon_n)'$, then

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

where [in the fixed design]

$$\begin{aligned} E(\varepsilon) &= 0 \\ E(\varepsilon\varepsilon') &= \sigma^2 I_n. \end{aligned}$$

The linear regression model is more commonly stated like this with statistical assumptions made about the unobservable ε rather than directly on the observable y. The assumptions about the vector ε are quite weak in some respects - the observations need not be independent and identically distributed, since only the first two mometns of the vector are specified - but strong in regard to the second moments themselves.

- It is worth discussing here some alternative assumptions made about the error terms. For this purpose we shall assume a random design, and moreover suppose that (x_i, ε_i) are i.i.d. In this case, we can further assume that
  - $E(\varepsilon_i x_i) = 0$
  - $E(\varepsilon_i | x_i) = 0$, denoted $\varepsilon_i \perp x_i$
  - ε_i are i.i.d. and independent of x_i, denoted $\varepsilon_i \perp\!\!\!\perp x_i$.
  - $\varepsilon_i \sim N(0, \sigma^2)$.
- The first assumption, called an unconditional moment condition, is the weakest assumption needed to 'identify' the parameter β.
- The second assumption, called a conditional moment restriction, is a little bit stronger. It is really just a rewriting of the definition of conditional expectation.
- The third assumption is much stronger and is not strictly necessary for estimation purposes although it does have implications about efficiency and choice of estimator.
- The fourth assumption we will sometimes make in connection with hypothesis testing and for establishing optimality of least squares.

1.2 The OLS Procedure

- In practice we don't know the parameter β and seek to estimate it from the data.
- For any b, define Xb and

$$u(b) = y - Xb.$$

  Then $u(b)$ is the vector of discrepancies between the observed y from the predicted by Xb.
- The Ordinary Least Squares (OLS) procedure chooses $\widehat{\beta}$ to minimize the quadratic form

$$S(b) = u(b)'u(b) = \sum_{i=1}^{n} u_i^2(b) = (y - Xb)'(y - Xb)$$

  with respect to $b \in \mathbb{R}^k$. This is perhaps the main estimator of β, and we shall study its properties at length.
- The first question is whether a minimum exists. Since the criterion is a continuous function of b, a minimum over any compact subset always exists.
- A necessary condition for the uniqueness of a solution is that $n \geq K$. If $n = K$, the solution essentially involves interpolating the data, i.e., the fitted value of y will be equal to the actual value.
- When the assumption that $\text{rank}(X) = K$ is made, $\widehat{\beta}$ is uniquely defined for any y and X independently of model; so there is no need for assumptions A1-A3 when it comes to computing the estimator.
- We now give two derivations of the well-known result that

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

- We suppose the answer is given by this formula and demonstrate that $\widehat{\beta}$ minimizes $S(b)$ with respect to b. Write

$$u(b) = y - X\widehat{\beta} + X\widehat{\beta} - Xb,$$

so that

$$\begin{aligned} S(b) &= (y-X\widehat{\beta}+X\widehat{\beta}-Xb)'(y-X\widehat{\beta}+X\widehat{\beta}-Xb) \\ &= (y-X\widehat{\beta})'(y-X\widehat{\beta}) \\ &\quad +(\widehat{\beta}-b)'X'X(\widehat{\beta}-b) \\ &\quad +(y-X\widehat{\beta})'X(\widehat{\beta}-b)+(\widehat{\beta}-b)'X'(y-X\widehat{\beta}) \\ &= (y-X\widehat{\beta})'(y-X\widehat{\beta})+(\widehat{\beta}-b)'X'X(\widehat{\beta}-b), \end{aligned}$$

because

$$X'(y-X\widehat{\beta}) = X'y - X'y = 0.$$

But

$$(\widehat{\beta}-b)'X'X(\widehat{\beta}-b) \geq 0,$$

and equality holds only when $b = \widehat{\beta}$.

- A minimizer of $S(b)$ must satisfy the vector of first order conditions:

$$\frac{\partial S}{\partial b} = 2X'(y-X\widehat{\beta}) = 0.$$

Therefore,

$$X'y = X'X\widehat{\beta}.$$

Now we use the assumption that X is of full rank. This ensures that $X'X$ is invertible, and

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

as required. To verify that we have found a local minimum rather than maximum it is necessary to calculate the second derivatives

$$\frac{\partial^2 S}{\partial b \partial b'} = 2X'X > 0.$$

- The vector derivatives follow by straightforward calculus

$$\frac{\partial}{\partial b_j}\sum_{i=1}^{n} u_i(b)^2 = 2\sum_{i=1}^{n} u_i(b)\frac{\partial u_i}{\partial b_j} = -2\sum_{i=1}^{n} u_i(b)x_{ij} \ ,$$

since

$$\frac{\partial u_i}{\partial b_j} = -x_{ij}.$$

- CHARACTERIZATION OF THE SOLUTION. Define the fitted value $\widehat{y} = X\widehat{\beta}$ and the OLS residuals

$$\widehat{u} = y - \widehat{y} = y - X\widehat{\beta}.$$

- The OLSE $\widehat{\beta}$ solves the normal equations $X'\widehat{u} = 0$, i.e.,

$$\begin{aligned} \sum_{i=1}^{n} x_{1i}\widehat{u}_i &= 0 \\ \sum_{i=1}^{n} x_{2i}\widehat{u}_i &= 0 \\ &\vdots \\ \sum_{i=1}^{n} x_{Ki}\widehat{u}_i &= 0. \end{aligned}$$

- We say that X is orthogonal to $\widehat{u}$, denoted $X \perp \widehat{u}$. Note that if, as usual $X_{1i} = 1$, then, we have $\sum_{i=1}^{n} \widehat{u}_i = 0$.

1.2.1 Some Alternative Estimation Paradigms

- We briefly mention some alternative estimation methods which actually lead to the same estimator as the OLS estimator in some special cases, but which are more broadly applicable.

- MAXIMUM LIKELIHOOD. Suppose we also assume that $y \sim N(X\beta, \sigma^2 I)$. Then the density function of y [conditional on X] is

$$f_{y|X}(y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2}(y - X\beta)'(y - X\beta)\right).$$

- The density function depends on the unknown parameters β, σ^2, which we want to estimate. We therefore switch the emphasis an call the following quantity the log likelihood function for the observed data

$$\ell(b, \omega^2 | y, X) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \omega^2 - \frac{1}{2\omega^2}(y - Xb)'(y - Xb),$$

where b and ω are unknown parameters.

- The maximum likelihood estimator $\widehat{\beta}_{mle}, \widehat{\sigma}^2_{mle}$ maximizes $\ell(b, \omega^2)$ with respect to b and ω^2. It is easy to see that

$$\begin{aligned} \widehat{\beta}_{mle} &= \widehat{\beta} \\ \widehat{\sigma}^2_{mle} &= \frac{1}{n}(y - X\widehat{\beta}_{mle})'(y - X\widehat{\beta}_{mle}). \end{aligned}$$

  Basically, the criterion function is the least squares criterion apart from an affine transformation involving only ω.

- Note however, that if we had a different assumption about the errors than A4, e.g., they were from a t-distribution, then we would have a different likelihood and a different estimator than $\widehat{\beta}$. In particular, the estimator may not be explicitly defined and may be a nonlinear function of y.

- METHOD OF MOMENTS. Suppose that we define parameters through some population moment conditions; this can arise from an economic optimization problem, see below.

- For example, suppose that we say that β is defined as the unique parameter that satisfies the K moment conditions [we need as many moment conditions as parameters]

$$E[x_i(y_i - x_i'\beta)] = 0.$$

  Note that this is the natural consequence of our assumption that $E(\varepsilon_i x_i) = 0$.

- Replacing the population by the sample average we must find b such that

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i'b) = 0.$$

  The solution to this is of course

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

  i.e., the MOM estimator is equal to OLS in this case. Thus, for the moment conditions above we are lead to the least squares estimator.

- However, if we chose some other conditions, then a different estimator results. For example, suppose that we assume that

$$E[x_i(y_i - x_i'\beta)^3] = 0,$$

we would be lead to a different estimator - any solution of

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i'b)^3 = 0.$$

In general, this would be more complicated to analyze.

- We emphasize here that the above estimation methods are all suggested or motivated by our assumptions, but of course we can always carry out the procedure without regard to underlying model - that is, the procedures only require data, not assumptions.

1.3 Geometry of OLS and Partitioned Regression

- We want to give a geometric interpretation to the OLS procedure.
- The data: y, $x_1, \ldots, x_K$, can all be viewed as elements of the vector space $\mathbb{R}^n$. Define the set

$$\mathcal{C}(X) = \{\alpha_1 x_1 + \cdots + \alpha_K x_K\} = \{X\alpha : \alpha \in \mathbb{R}^K\} \subseteq \mathbb{R}^n,$$

otherwise known as the column span of X.

- Then, $\mathcal{C}(X)$ is a linear subspace of $\mathbb{R}^n$ of dimension K assuming that the matrix X is of full rank. If it is only of rank K^* with $K^* < K$ then $\mathcal{C}(X)$ is still a linear subspace of $\mathbb{R}^n$ but of dimension K.
- The OLS procedure can equivalently be defined as finding the point in $\mathcal{C}(X)$ closest to y, where closeness is measured in terms of Euclidean distance, i.e.,

$$d(y, Xb) = \|y - Xb\|^2 = (y - Xb)'(y - Xb)$$

is the Euclidean distance of y to the point $Xb \in \mathcal{C}(\mathrm{X})$.

- This is an old problem in geometry, which is now given a key role in abstract mathematics.

- The projection theorem [Hilbert] says that there is a unique solution to the minimization problem, call it $\widehat{y}$, which is characterized by the fact that

$$\widehat{u} = y - \widehat{y}$$

is orthogonal to $\mathcal{C}$(X).

- Equivalently we can write uniquely

$$y = \widehat{y} + \widehat{u},$$

where $\widehat{y} \in \mathcal{C}(X)$ and $\widehat{u} \in \mathcal{C}^{\perp}(X)$ [the space $\mathcal{C}^{\perp}(X)$ is called the orthocomplement of $\mathcal{C}(X)$, and consists of all vectors orthogonal to $\mathcal{C}(X)$]. Essentially, one is dropping a perpendicular, and the procedure should be familiar from high school geometry.

- For example, let $n = 3$ and

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then $\mathcal{C}(X)$ is the set of all vectors in $\mathbb{R}^3$ with third component zero. What is the closest point for the example above with $y = (1, 1, 1)'$? This is $(1, 1, 0)' = X\widehat{\beta}$, $\widehat{u} = (0, 0, 1)'$ In fact $\widehat{u}$ is orthogonal to $\mathcal{C}(X)$, i.e., $\widehat{u} \in \mathcal{C}^{\perp}(X) = (0, 0, \alpha)'$.

- In general how do we find $\widehat{y}$? When X is of full rank we can give a simple explicit solution

$$\widehat{y} = P_X y,$$

where the Projector matrix

$$P_X = X(X'X)^{-1}X'$$

projects onto $\mathcal{C}(X)$.

- Let $\widehat{u} = y - \widehat{y} = M_X y$, where the Projector matrix

$$M_X = I - X(X'X)^{-1}X'$$

projects onto $\mathcal{C}^{\perp}(X)$. Thus for any y, we can write

$$y = \widehat{y} + \widehat{u} = P_X y + M_X y.$$

The matrices P_X and M_X are symmetric and idempotent, i.e.,

$$P_X = P_X' \text{ and } P_X^2 = P_X.$$

After applying P_X once you are ready in $\mathcal{C}(X)$. This implies that

$$P_X X = X \text{ and } M_X X = 0,$$

so that

$$P_X M_X y = 0 \text{ for all } y.$$

- Since $\widehat{y} \in \mathcal{C}(X)$, we can rewrite it as $\widehat{y} = X\widehat{\beta}$, so that $\widehat{\beta} = (X'X)^{-1}X'y$.

- The space $\mathcal{C}(X)$ is invariant to nonsingular linear transforms

$$X \longmapsto XA_{K\times K}\ , \quad \text{where } \det A \neq 0.$$

Let $z \in \mathcal{C}(X)$. Then there exists an $\alpha \in \mathbb{R}^K$ such that $z = X\alpha$. Therefore,

$$z = XAA^{-1}\alpha = XA\gamma,$$

where $\gamma = A^{-1}\alpha \in \mathbb{R}^K$, and vice-versa.

- Since $\mathcal{C}(X)$ is invariant to linear transformations, so are $\widehat{y}$ and $\widehat{u}$ (but not $\widehat{\beta}$). For example, rescaling of the components of X does not affect the values of $\widehat{y}$ and $\widehat{u}$.

$$y \text{ on } (\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3)(1)$$

$$y \text{ on } (\mathrm{x}_1 + \mathrm{x}_2,\ 2\mathrm{x}_2 - \mathrm{x}_3,\ 3\mathrm{x}_1 - 2\mathrm{x}_2 + 5\mathrm{x}_3)(2)$$

in which case the transformation is

$$A = \begin{pmatrix} 1 & 0 & 3 \\ 1 & 2 & -2 \\ 0 & -1 & 5 \end{pmatrix},$$

which is of full rank. Therefore, (1) and (2) yield the same $\widehat{y}$, $\widehat{u}$.

- Emphasizing $\mathcal{C}(X)$ rather than X itself is called the *coordinate free* approach. Some aspects of model/estimate are properties of $\mathcal{C}(X)$ choice of coordinates is irrelevant.

- When X is not of full rank

  - the space $\mathcal{C}(X)$ is still well defined, as is the projection from y onto $\mathcal{C}(X)$.
  - The fitted value $\widehat{y}$ and residual $\widehat{u}$ are uniquely defined in this case,
  - but there is no unique coefficient vector $\widehat{\beta}$.
  - This is the case commonly called multicollinearity.

- We next consider an important application of the projection idea. Partition
$$X = (X_{1n\times K_1}, X_{2n\times k_2}) \ , \quad K_1 + K_2 = K,$$
and suppose we are interested in obtaining the coefficient $\widehat{\beta}_1$ in the projection of y onto $\mathcal{C}(X)$.

- A key property of projection is that if X_1 and X_2 are orthogonal, i.e., if $X_1'X_2 = 0$, then
$$P_X = P_{X_1} + P_{X_2}.$$
This can be verified algebraically, but also should be obvious geometrically. In this case, write
$$\widehat{y} = X\widehat{\beta} = P_X y = P_{X_1} y + P_{X_2} y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2.$$
This just says that if X_1 and X_2 were orthogonal, then we could get $\widehat{\beta}_1$ by regressing y on X_1 only, and $\widehat{\beta}_2$ by regressing y on X_2 only.

- Very rarely are X_1 and X_2 orthogonal, but we can construct equivalent regressors that are orthogonal. Suppose we have general X_1 and X_2, whose dimensions satisfy $K_1 + K_2 = K$. We make the following observations:

  - (X_1, X_2) and (M_2X_1, X_2) span the same space. This follows because
$$X_1 = M_2X_1 + P_2X_1,$$

where $\mathcal{C}(P_2X_1) \subset \mathcal{C}(X_2)$. Therefore,

$$\mathcal{C}(M_2X_1,\, X_2) = \mathcal{C}(X_1,\, X_2).$$

– M_2X_1 and X_2 are orthogonal.

- This says that if we regress y on $(X_1,\, X_2)$ or y on $(M_2X_1,\, X_2)$ we get the same $\widehat{y}$ and $\widehat{u}$, and that if we wanted the coefficients on M_2X_1 from the second regression we could in fact just regress y on M_2X_1 only.

- What are the coefficients on M_2X_1? Recall that

$$\begin{aligned}
\widehat{y} &= X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 \\
&= (M_2 + P_2)X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 \\
&= M_2X_1\widehat{\beta}_1 + X_2[\widehat{\beta}_2 + (X_2'X_2)^{-1}X_2'X_1\widehat{\beta}_1] \\
&= M_2X_1\widehat{\beta}_1 + X_2\widehat{C},
\end{aligned}$$

  where

$$\widehat{C} = \widehat{\beta}_2 + (X_2'X_2)^{-1}X_2'X_1\widehat{\beta}_1.$$

- So the coefficient on M_2X_1 is the original $\widehat{\beta}_1$, while that on X_2 is some combination of $\widehat{\beta}_1$ and $\widehat{\beta}_2$. Note that M_2X_1 are the residuals from a regression of X_1 on X_2.

- Practical Implication. If K is large and primarily interested in first K_1 variables, then we can get $\widehat{\beta}_1$ by regressing y [or M_2y equivalently] on M_2X_1 only, i.e.,

$$\widehat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2y = (X_1'M_2M_2X_1)^{-1}X_1'M_2M_2y.$$

  This involves inversion of only $K_1 \times K_1$ and $K_2 \times K_2$ matrices, which involves less computing time than inverting $K \times K$ matrices, especially when K is large [this computation can be as bad as $O(K^3)$].

- Suppose that $X_2 = (1, 1, \ldots, 1)' = i$, then

$$M_2 = I_n - i(i'i)^{-1}i' = I_n - \frac{ii'}{n}$$

and

$$M_2 x_{1_{n\times 1}} = x_1 - \frac{1}{n}\sum_{i=1}^{n} x_{1i}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = \begin{pmatrix} x_{1i} - \overline{x}_1 \\ \vdots \\ x_{1n} - \overline{x}_1 \end{pmatrix}.$$

When regression includes an intercept, can first demean the X variables (and the y's) then do regression on the demeaned variables.

1.4 Goodness of Fit

- How well does the model explain the data? One possibility is to measure the fit by the residual sum of squares

$$RSS = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

In general, the smaller the RSS the better. However, the numerical value of RSS depends on the units used to measure y in so that one cannot compare across models.

- Generally used measure of goodness of fit is the R^2. In actuality, there are three alternative definitions in general.

  - One minus the ratio of the residual sum of squares to total sum of squares,

$$R_1^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

  - The sample correlation squared between y and $\widehat{y}$,

$$R_2^2 = \frac{[\sum_{i=1}^{n}(y_i - \overline{y})(\widehat{y}_i - \overline{\widehat{y}})]^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(\widehat{y}_i - \overline{\widehat{y}})^2}.$$

  - The ratio of explained sum of squares to total sum of squares

$$R_3^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Here, $\overline{y} = \sum_{i=1}^{n} y_i/n$ and $\overline{\widehat{y}} = \sum_{i=1}^{n} \widehat{y}_i/n$.

- Theorem. When an intercept is included, all three measures are the same.
- Proof of $R_1^2 = R_2^2$. Since an intercept is included, we have

$$\sum_{i=1}^{n} \widehat{u}_i = 0,$$

which implies that $\overline{\widehat{y}} = \overline{y}$. Therefore,

$$\sum_{i=1}^{n}(y_i - \overline{y})(\widehat{y}_i - \overline{\widehat{y}}) = \sum_{i=1}^{n}(y_i - \overline{y})(\widehat{y}_i - \overline{y}) = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2,$$

because

$$\sum_{i=1}^{n} \widehat{u}_i(\widehat{y}_i - \overline{y}) = 0.$$

- Proof of $R_1^2 = R_3^2$. Similarly,

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2.$$

■

- If an intercept is included, then $0 \leq R^2 \leq 1$. If not, then $0 \leq R_2^2 \leq 1$, but R_3^2 could be greater than one, and R_1^2 could be less than zero.
- If $y = \alpha + \beta x + u$, then R^2 is the squared sample correlation between y and x.
- The R^2 is invariant to some changes of units.
- If $y \mapsto ay + b$ for any constants a, b, then
  - $\widehat{y}_i \mapsto a\widehat{y}_i + b$ and
  - $\overline{y} \mapsto a\overline{y} + b$,
  - so R^2 is the same in this case.
  - Clearly, if $X \mapsto XA$ for a nonsingular matrix A, then $\widehat{y}$ is unchanged, as is y and $\overline{y}$.

- R^2 always increases with addition of variables. With $K = n$ we can make $R^2 = 1$.

- Theil's adjusted R^2 is defined as follows

$$\overline{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2).$$

This amounts to dividing the sum of squares by the appropriate degrees of freedom, so that

$$1 - \overline{R}^2 = \frac{\frac{1}{n-K}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

It follows that

$$\frac{\Delta\overline{R}^2}{\Delta K} = \underbrace{\frac{n-1}{n-K}}_{+}\frac{\Delta R^2}{\Delta K} - \underbrace{\frac{n-1}{(n-K)^2}}_{-}(1 - R^2).$$

This measure allows some trade-off between fit and parsimony.

1.5 Functional Form

- Linearity can often be restrictive. We shall now consider how to generalize slightly the use of the linear model, so as to allow certain types of nonlinearity, but without fundamentally altering the applicability of the analytical results we have built up.

$$Wages = \alpha + \beta ed + \gamma UNION + u$$

$$Wages = \alpha + \beta ed + \gamma ab + \delta ed \cdot ab + u$$

$$Wages = \alpha + \beta ed + \gamma ex + \delta ex^2 + \rho ex^3 + u$$

$$\log Wages = \alpha + \beta ed + \gamma UNION + u$$

$$\log \frac{fs}{1 - fs} = \alpha + \beta inc + u$$

- These are all linear in the *parameters* model, i.e., can write

$$y = X\beta + u$$

for some X, some β, some y.

- Another interesting example is *Splines.* This is a Piecewise Linear function. For example, suppose we have a scalar regressor x, which is time, i.e., $x_t = t$, $t = 1, 2, \ldots, T$. Further suppose that

$$y = \begin{cases} \alpha_1 + \beta_1 x + u & \text{if } x \leq t_1^* \\ \alpha_2 + \beta_2 x + u & \text{if } t_1^* \leq x \leq t_2^* \\ \alpha_3 + \beta_3 x + u & \text{if } x \geq t_2^*. \end{cases}$$

- This can be expressed as follows:

$$y = \alpha_1 + \beta_1 x + \gamma_1 D_1 + \delta_1 D_1 \cdot x + \gamma_2 D_2 + \delta_2 D_2 \cdot x + u,$$

where

$$D_1 = \begin{cases} 1 & \text{if } x \geq t_1^*, \\ 0 & \text{else} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{if } x \geq t_2^*, \\ 0 & \text{else.} \end{cases}$$

- How do we impose that the function join up? We must have

$$\alpha_1 + \beta_1 t_1^* = \alpha_1 + \gamma_1 + (\beta_1 + \delta_1) t_1^*$$

$$\alpha_1 + \beta_1 t_2^* + \gamma_1 + \delta_1 t_2^* = \alpha_1 + \gamma_1 + (\beta_1 + \delta_1) t_1^* + \gamma_2 + \delta_2 t_2^*,$$

which implies that

$$\gamma_1 = -\delta_1 t_1^* \text{ and } \gamma_2 = -\delta_2 t_2^*,$$

which are two linear restrictions on the parameters, i.e.,

$$y = \alpha_1 + \beta_1 x + (D_1 x - D_1 t_1^*)\delta_1 + (D_2 x - D_2 t_2^*)\delta_2 + u.$$

- Some Nonlinear in Parameters Functions

- Box–Cox

$$y = \alpha + \beta \frac{x^\lambda - 1}{\lambda} + u,$$

where

$$\begin{aligned} \text{as } \lambda &\to 0, \ \frac{x^\lambda - 1}{\lambda} \to \ln(x); \\ \text{as } \lambda &\to 1, \ \frac{x^\lambda - 1}{\lambda} \to x - 1 \end{aligned}$$

- Real money demand

$$y = \beta_1 X_1 + \frac{\beta_2}{x_2 - \gamma} + u.$$

If there exists $\gamma > 0$, then we have a Liquidity trap.

- CES production function

$$Q = \beta_1 \left[\beta_2 K^{-\beta_3} + (1 - \beta_2) L^{-\beta_3}\right]^{\beta_4/\beta_3} + u.$$

Methods for treating these models will be considered below.

Chapter 2

Statistical Properties of the OLS Estimator

- We now investigate the statistical properties of the OLS estimator in both the fixed and random designs. Specifically, we calculate its exact mean and variance. We shall examine later what happens when the sample size increases.

- The first thing to note in connection with $\widehat{\beta}$ is that it is linear in y, i.e., there exists a matrix C not depending on y such that

$$\widehat{\beta} = (X'X)^{-1}X'y = Cy.$$

This property makes a lot of calculations simple.

- We want to evaluate how $\widehat{\beta}$ varies across hypothetical repeated samples. We shall examine both the fixed design and the random design case. The fixed design is the main setting we use in this course; it is simpler to work with and gives the main intuition. The random design approach is given here for completeness; it will be come more relevant later in the course.

- *Fixed Design.* First,

$$E(\widehat{\beta}) = (X'X)^{-1}X'E(y) = (X'X)^{-1}X'X\beta = \beta,$$

where this equality holds for all β. We say that $\widehat{\beta}$ is unbiased.

- Furthermore, we shall calculate the $K \times K$ covariance matrix of $\widehat{\beta}$,

$$\text{var}(\widehat{\beta}) = E\{(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)'\}.$$

This has diagonal elements $\text{var}(\widehat{\beta}_j)$ and off-diagonals $\text{cov}(\widehat{\beta}_j, \widehat{\beta}_k)$. We have

$$\begin{aligned} \text{var}((X'X)^{-1}X'y) &= (X'X)^{-1}X'\text{var}\, yX(X'X)^{-1} = E\{(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\} \\ &= (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} = \sigma^2(X'X)^{-1}. \end{aligned}$$

- *Random Design.* For this result we need $E[\varepsilon_i|x_i] = 0$. We first condition on the matrix X; this results in a fixed design and the above results hold. Thus, if we are after conditional results, we can stop here. If we want to calculate unconditional mean and variance we must now average over all possible X designs. Thus

$$E(\widehat{\beta}) = E\{E(\widehat{\beta}|X)\} = E(\beta) = \beta.$$

On average we get the true parameter β. Note that this calculation uses the important property called "The Law of Iterated Expectation". The most general version of this says that

$$E(Y|I_1) = E[E(Y|I_2)|I_1],$$

whenever $I_1 \subseteq I_2$ for two information sets I_1, I_2.

- Note that if only $E[x_i\varepsilon_i] = 0$, then the above calculation may not be valid. For example, suppose that $Y_i = X_i^3$, where X_i is i.i.d. standard normal. Then $\beta = 3$ minimizes $E[(Y_i - bX_i)^2]$. Now consider the least squares estimator

$$\widehat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n X_i^4}{\sum_{i=1}^n X_i^2}.$$

You can't show that this is unbiased, and indeed it isn't.

- As for the variance, we use another important property

$$\text{var}[y] = E[\text{var}(y|X)] + \text{var}[E(y|X)],$$

which is established by repeated application of the law of iterated expectation. We now obtain

$$\text{var}(\widehat{\beta}) = E\text{var}(\widehat{\beta}|X) = \sigma^2 E\{(X'X)^{-1}\}.$$

This is not quite the same answer as in the fixed design case, and the interpretation is of course different.

- The properties of an individual coefficient can be obtained from the partitioned regression formula

$$\widetilde{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2y.$$

- In the Fixed Design

$$\text{var}[\widehat{\beta}_1] = (X_1'M_2X_1)^{-1}X_1'M_2E\varepsilon\varepsilon'M_2X_1(X_1'M_2X_1)^{-1} = \sigma^2(X_1'M_2X_1)^{-1}.$$

- In the special case that $X_2 = (1, \ldots, 1)'$, we have

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

This is the well known variance of the least squares estimator in the single regressor plus intercept regression.

- We now turn to the distribution of $\widehat{\beta}$. This will be important when we want to conduct hypothesis tests and construct confidence intervals. In order to get the *exact* distribution we will need to make an additional assumption.
  - (A4) $y \sim N(X\beta, \sigma^2 I)$ or
  - (A4r) $y|X \sim N(X\beta, \sigma^2 I)$.

- Under A4,

$$\widehat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

in the fixed design case, because

$$\widehat{\beta} = (X'X)^{-1}X'y = \sum_{i=1}^n c_i y_i,$$

i.e., $\widehat{\beta}$ is a linear combination of independent normals.

- Under A4r, the conditional distribution of $\widehat{\beta}$ given X is normal with mean β and variance $\sigma^2(X'X)^{-1}$. However, the unconditional distribution will not be normal - in fact, it will be a scale mixture of normals meaning that, in the scalar case for simplicity, its density function is

$$f_{\widehat{\beta}}(z) = \int \frac{1}{\sigma \cdot v} \phi\left(\frac{z-\beta}{\sigma \cdot v}\right) g(v) dv,$$

where g is the density of $(\sum_{i=1}^n x_i^2)^{1/2}$ and ϕ is the standard normal density function.

2.1 Optimality

- There are many estimators of β. Consider the scalar regression $y_i = \beta x_i + \varepsilon_i$. The OLS estimator is $\widehat{\beta} = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2$. Also plausible are $\widetilde{\beta} = \overline{y}/\overline{x}$ and $\overline{\beta} = \sum_{i=1}^n y_i/x_i$, as well as nonlinear estimators such as the LAD procedure

$$\arg\min_{\beta} \sum_{i=1}^n |y_i - \beta x_i|.$$

- In fact, $\widehat{\beta}$, $\widetilde{\beta}$, and $\overline{\beta}$ are all linear unbiased. How do we choose between estimators? Computational convenience is an important issue, but the above estimators are all similar in their computational requirements. We now investigate statistical optimality.

- Definition: The mean squared error (hereafter MSE) matrix of a generic estimator $\widehat{\theta}$ of a parameter $\theta \in \mathbb{R}^p$ is

$$\begin{aligned} & E[(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)'] \\ = \; & E[(\widehat{\theta}-E(\widehat{\theta})+E(\widehat{\theta})-\theta)(\widehat{\theta}-E(\widehat{\theta})+E(\widehat{\theta})-\theta)'] \\ = \; & \underbrace{E[(\widehat{\theta}-E(\widehat{\theta}))(\widehat{\theta}-E(\widehat{\theta}))']}_{\text{variance}} \\ & + \underbrace{[E(\widehat{\theta})-\theta][E(\widehat{\theta})-\theta]'}_{\text{squared bias}}. \end{aligned}$$

- The MSE matrix is generally a function of the true parameter θ. We would like a method that does well for all θ, not just a subset of parameter values - the estimator $\widehat{\theta} = 0$ is an example of a procedure that will have MSE equal to zero at $\theta = 0$, and hence will do well at this point, but as θ moves away, the MSE increases quadratically without limit.

- MSE defines a complete ordering when $p = 1$, i.e., one can always rank any two estimators according to MSE. When $p > 1$, this is not so. In the general case we say that $\widehat{\theta}$ is better (according to MSE) than $\widetilde{\theta}$ if

$$B \geq A$$

(i.e., $B - A$ is a positive semidefinite matrix), where B is the MSE matrix of $\widetilde{\theta}$ and A is the MSE of $\widehat{\theta}$.

- For example, suppose that

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 0 \\ 0 & 1/4 \end{bmatrix}.$$

In this case, we can not rank the estimators. The problem is due to the multivariate nature of the optimality criterion.

- One solution is to take a scalar function of MSE such as the trace or determinant, which will result in a complete ordering. However, different functions will rank estimators differently [see the example above].

- Also note that no estimator can dominate uniformly across θ according to MSE because it would have to beat all constant estimatros which have zero MSE at a single point. This is impossible unless there is no randomness.

- One solution is to change the criterion function. For example, we might take $\max_\theta \operatorname{tr}(MSE)$, which takes the most pessimistic view. In this case, we might try and find the estimator that minimizes this criterion - this would be called a minimax estimator. The theory for this class of estimators is very complicated, and in any case it is not such a desirable criterion because it is so pessimistic about nature trying to do the worst to us.

- Instead, we reduce the class of allowable estimators. If we restrict attention to unbiased estimators then this rules out estimators like $\widehat{\theta} = 0$ because they will be biased. In this case there is some hope of an optimality theory for the class of unbiased estimators.

- We will now return to the linear regression model and make the further restriction that the estimators we consider are linear in y. That is, we suppose that we have the set of all estimators $\widetilde{\beta}$ that satisfy

$$\widetilde{\beta} = Ay$$

for some fixed matrix A such that

$$E(\widetilde{\beta}) = \beta, \quad \forall \beta.$$

This latter condition implies that $(AX - I)\beta = 0$ for all β, which is equivalent to $AX = I$.

- Gauss Markov Theorem. Assume that A1–A3 hold. The OLS estimator $\widehat{\beta}$ is Best Linear Unbiased (BLUE), i.e.,

$$\text{var}(\widehat{\beta}) \leq \text{var}(\widetilde{\beta})$$

for any other LUE.

- Proof. $\text{var}(\widehat{\beta}) = \sigma^2(X'X)^{-1}$; $\text{var}(\widetilde{\beta}) = \sigma^2 AA'$ and

$$\begin{aligned}
\text{var}(\widetilde{\beta}) - \text{var}(\widehat{\beta}) &= \sigma^2[AA' - (X'X)^{-1}] \\
&= \sigma^2[AA' - AX(X'X)^{-1}X'A'] \\
&= \sigma^2 A[I - X(X'X)^{-1}X']A' \\
&= \sigma^2 AMA' \\
&= \sigma^2(AM) \cdot (M'A') \\
&\geq 0.
\end{aligned}$$

- – Makes no assumption about the distribution of the errors; it only assumes 0 mean and $\sigma^2 I$ variance.
  – Result only compares *linear* estimators; it says nothing about for example $\sum_{i=1}^{n} |y_i - \beta x_i|$.

  - Result only compares *unbiased* estimators [biased estimators can have 0 variances]. In fact, although the OLS estimator is admissible with respect to MSE, it is inadmissible with respect to trace mean squared error when the number of regressors is at least three. The Stein estimator is better according to trace mean squared error. Of course in large samples this is all irrelevant.
  - There are extensions to consider affine estimators $\widetilde{\beta} = a + Ay$ for vectors a. There are also equivalent results for the invariant quantity $\widehat{y}$.

- If we dispense with the unbiasedness assumption and add the model assumption of error normality we get the well-known result.
- Cramèr–Rao Theorem. Under A1-A4, $\widehat{\beta}$ is Best Unbiased (statement is for MLE's).
- By making the stronger assumption A4, we get a much stronger conclusion. This allows us to compare say LAD estimation with OLS.
- Asymptotically, a very large class of estimators are both unbiased and indeed linear so that the Gauss-Markov and Cramèr–Rao Theorems apply to a very broad class of estimators when the words "for large n" are inserted.

Chapter 3

Hypothesis Testing

- In addition to point estimation we often want to know how good our estimator is and whether it is compatible with certain preconceived 'hypotheses' about the data.
- Suppose that we observe certain data (y, X), and there is a true data distribution denoted by f, which is known to lie in a family of models $\mathcal{F}$. We now suppose there is a further reduction called a Hypothesis $H_0 \subseteq \mathcal{F}$. For example, H_0 could be the:
  - Prediction of a scientific theory. For example, the interest elasticity of demand for money is zero; the gravitational constant is 9.
  - Absence of some structure, e.g., independence of error term over time, homoskedasticity etc.
  - Pretesting (used as part of model building process).
- We distinguish between a *Simple hypothesis* (under H_0, the data distribution is completely specified) and a *Composite hypothesis* (in which case, H_0 does not completely determine the distribution, i.e., there are 'nuisance' parameters not specified by H_0).
- We also distinguish between *Single* and *Multiple* hypotheses (one or more restriction on parameters of f).
- We shall also introduce the alternative hypothesis H_A, which will be the complement of H_0 in $\mathcal{F}$, i.e., $\mathcal{F} = H_0 \cup H_A$. That is, the choice of

$\mathcal{F}$ is itself of some significance since it can restrict the range of values taken by the data distribution. We shall also distinguish between *one-sided* and *two-sided* alternatives; when we have a single real-valued parameter this is an easy notion to comprehend.

- Examples

  - The theoretical model is the Cobb–Douglas production function

    $$Q = AK^{\alpha}L^{\beta}.$$

    Empirical version: take logs and add an error term to give a linear regression

    $$q = a + \alpha k + \beta \ell + \varepsilon.$$

    It is often of interest whether constant returns to scale operate, i.e., would like to test whether

    $$\alpha + \beta = 1$$

    is true. We may specify the alternative as $\alpha + \beta < 1$, because we can rule out increasing returns to scale.

  - Market efficiency

    $$r_t = \mu + \gamma' I_{t-1} + \varepsilon_t,$$

    where r_t are returns on some asset held between period $t-1$ and t, while I_t is public information at time t. Theory predicts that $\gamma = 0$; there is no particular reason to restrict the alternative here.

  - Structural change

    $$y = \alpha + \beta x_t + \gamma D_t + \varepsilon_t$$

    $$D_t = \begin{cases} 0\ , & t < 1974 \\ 1\ , & t \geq 1974. \end{cases}$$

    Would like to test $\gamma = 0$.

3.1 General Notations

- A hypothesis test is a rule [function of the data] which yields either reject or accept outcomes.
- There are two types of mistakes that any rule can make:
  - Type I error is to reject when the null hypothesis is true
  - Type II error is of accepting a false hypothesis.
- We would like to have as small a Type I and Type II error as possible. Unfortunatley, these are usually in conflict. The traditional approach is to fix the Type I error and then try to do the best in terms of the Type II error.
- We choose $\alpha \in [0,1]$ called the size of the test [magnitude of Type I error]. Let T(data) be a test statistic, typically scalar valued. Then, find acceptance region C_α of size α such that

$$\Pr[T \notin C_\alpha | H_0] = \alpha.$$

  The rule is to reject H_0 if $T \notin C_\alpha$ and to accept otherwise. The practical problem is how to choose T so that C_α [or equivalently the rejeciton region R_α] is easy to find.
- Define also Power of test:

$$\pi = \Pr[T \notin C_\alpha | H_A] = 1 - \text{TypeII}.$$

  It is desirable, ceteris paribus, to have a test that maximizes power for any given size.
- Optimal testing. Neyman-Pearson Lemma. Suppose you have a parametric model with parameter θ and consider the simple null hypothesis against a one-sided alternative:

$$H_0 : \theta = \theta_0, \;\; H_A : \theta > \theta_0 \text{ or } \theta < \theta_0.$$

  The likelihood ratio test is Uniformly Most Powerful UMP provided the parametric model has the monotone likelihood ratio (MLR) property. Examples: One parameter exponential families, e.g., Normal,Poisson, and Binomial.

- Testing against two-sided alternatives, UMP's do not exist.
- EXAMPLE. $X \sim N(\mu, 1)$; $H_0 : \mu = 0$ vs. $\mu > 0$. In this case, the best rejection region is $\{\mathcal{X}^n : \bar{X} > z_\alpha / n^{1/2}\}$. For any $\mu > 0$, this test is most powerful $\mu = 0$ vs. μ. Region and rule distribution is independent of μ. In the two-sided test

  $$\{\mathcal{X}^n : |\bar{X}| > \frac{z_{\alpha/2}}{n^{1/2}}\}$$

  is less powerful than

  $$\{\mathcal{X}^n : \bar{X} > \frac{z_\alpha}{n^{1/2}}\} \text{ when } \mu > 0,$$

  and less powerful than

  $$\{\mathcal{X}^n : \bar{X} < \frac{z_\alpha}{n^{1/2}}\} \text{ when } \mu < 0.$$

- Unbiased and Invariant Tests. Just like in estimation it can help to reduce the class of tests. An unbiased test satisfies

  $$\pi(\theta) \geq \alpha \text{ for all } \theta \in \Theta_1.$$

  Clearly the one-sided interval is biased because when $\mu < 0$ power is zero. The above two-sided normal test is UMP unbiased. Alternatively can eliminate some tests by requiring invariance under a group of transformations.

3.2 Examples

- Hypothesis Testing in Linear Regression: $y \sim N(X\beta, \sigma^2 I)$.
  - Single (Linear) Hypothesis:

    $$c'\beta = \gamma \in \mathbb{R},$$

    e.g., $\beta_2 = 0$ (t–test).
  - Multiple (Linear) Hypothesis:

    $$R_{q\times K}\beta_{K\times 1} = r_{q\times 1}, \quad q \leq K,$$

    e.g., $\beta_2 = \beta_3 = \cdots = \beta_K = 0$.

– Single Non-linear Hypothesis:

$$\beta_1^2 + \beta_2^2 + \cdots + \beta_K^2 = 1.$$

Note that these are all composite hypotheses, i.e., there are nuisance parameters like σ^2 that are not specified by the null hypothesis.

3.3 Test of a Single Linear Hypothesis

- We wish to test the hypothesis

$$c'\beta = \gamma,$$

e.g., $\beta_2 = 0$. Suppose that $y \sim N(X\beta, \sigma^2 I)$. Then,

$$\frac{c'\widehat{\beta} - \gamma}{\sigma(c'(X'X)^{-1}c)^{1/2}} \sim N(0,1).$$

We don't know σ and must replace it by an estimate. There are two widely used estimates:

$$\widehat{\sigma}^2_{mle} = \frac{\widehat{\varepsilon}'\widehat{\varepsilon}}{n}$$

$$s^2 = \frac{\widehat{\varepsilon}'\widehat{\varepsilon}}{n-K}$$

The first estimate is the maximum likelihood estimator of σ^2, which can be easily verified. The second estimate is a modification of the MLE, which happens to be unbiased. Now define the test statistic

$$T = \frac{c'\widehat{\beta} - \gamma}{s(c'(X'X)^{-1}c)^{1/2}}.$$

- Theorem Under H_0,

$$T \sim t(n-K).$$

- Proof. We show that:

– (1) $\frac{n-K}{\sigma^2}s^2 \sim \chi^2_{n-K}$

– (2) s^2 and $c'\widehat{\beta} - \gamma$ are independent.

This establishes the theorem by the defining property of a t-random variable.

- Recall that
$$\frac{\varepsilon'\varepsilon}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{\varepsilon_i}{\sigma}\right)^2 \sim \chi^2_n.$$
But $\widehat{\varepsilon}$ are residuals that use K parameter estimates. Furthermore,
$$\widehat{\varepsilon}'\widehat{\varepsilon} = \varepsilon' M_X \varepsilon$$
and
$$\begin{aligned} E[\varepsilon' M_X \varepsilon] &= E[\text{tr } M_X \varepsilon\varepsilon'] \\ &= \text{tr } M_X E(\varepsilon\varepsilon') \\ &= \sigma^2 \text{tr } M_X \\ &= \sigma^2(n - \text{tr } P_X) \\ &= \sigma^2(n - K) \end{aligned}$$
because
$$\begin{aligned} \text{tr}(X(X'X)^{-1}X') &= \text{tr } X'X(X'X)^{-1} \\ &= \text{tr } I_K = K. \end{aligned}$$
These calculations show that
$$E\widehat{\varepsilon}'\widehat{\varepsilon} = \sigma^2(n - K),$$
which suggests that $\widehat{\varepsilon}'\widehat{\varepsilon}/\sigma^2$ cannot be χ^2_n [and incidentally that $Es^2 = \sigma^2$].

- Note that M_X is a symmetric idempotent matrix, which means that it can be written
$$M_X = Q\Lambda Q',$$
where $QQ' = I$ and Λ is a diagonal matrix of eigenvalues, which in this case are either zero (K times) or one ($n - K$ times). Furthermore, by a property of the normal distribution,
$$Q\varepsilon = \varepsilon^*$$

has exactly the same normal distribution as ε [it has the same mean and variance, which is sufficient to determine the normal distribution]. Therefore,

$$\frac{\widehat{\varepsilon}'\widehat{\varepsilon}}{\sigma^2} = \sum_{i=1}^{n-K} z_i^2$$

for some i.i.d. standard normal random variables z_i. Therefore, $\widehat{\varepsilon}'\widehat{\varepsilon}/\sigma^2$ is χ^2_{n-K} by the definition of a chi-squared random variable.

- Furthermore, under H_0,

$$c'\widehat{\beta} - \gamma = c'(X'X)^{-1}X'\varepsilon \text{ and } \widehat{\varepsilon} = M_X\varepsilon$$

are mutually uncorrelated since

$$E[M_X\varepsilon\varepsilon'X(X'X)^{-1}c] = \sigma^2 M_X X(X'X)^{-1}c = 0.$$

Under normality, uncorrelatedness is equivalent to independence.

- We can now base test of H_0 on

$$T = \frac{c'\widehat{\beta} - \gamma}{s(c'(X'X)^{-1}c)^{1/2}},$$

using the t_{n-k} distribution for an exact test under normality. Can test either one-sided and two-sided alternatives, i.e., reject if

$$|T| \geq t_{n-K}(\alpha/2)$$

[two-sided alternative] or if

$$T \geq t_{n-K}(\alpha)$$

[one-sided alternative].

- Above is a general rule, and would require some additional computations in addition to $\widehat{\beta}$. Sometimes one can avoid this: if computer automatically prints out results of hypothesis for $\beta_i = 0$, and one can redesign the null regression suitably. For example, suppose that

$$H_0 \,:\, \beta_2 + \beta_3 = 1.$$

Substitute the restriction in to the regression $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + u_i$, which gives the restricted regression

$$y_i - z_i = \beta_1 + \beta_2(x_i - z_i) + u_i.$$

Now test whether $\beta_3 = 0$ in the regression

$$y_i - z_i = \beta_1 + \beta_2(x_i - z_i) + \beta_3 z_i + u_i.$$

3.4 Test of a Multiple Linear Hypothesis

- We now consider a test of the multiple hypothesis $R\beta = r$. Define the quadratic form

$$\begin{aligned} F &= (R\widehat{\beta} - r)' \left[s^2 R(X'X)^{-1}R'\right]^{-1} (R\widehat{\beta} - r)/q \\ &= \frac{(R\widehat{\beta} - r)' \left[R(X'X)^{-1}R'\right]^{-1} (R\widehat{\beta} - r)/q}{(n-K)s^2/(n-K)}. \end{aligned}$$

- If $y \sim N(X\beta,\, \sigma^2 I)$, then

$$F = \frac{\chi^2_q/q}{\chi^2_{n-K}/(n-K)} \sim F(q,\, n-K)$$

under H_0. The rule is that if

$$F \geq F_\alpha(q,\, n-K),$$

then reject H_0 at level α. Note that we can only test against a two-sided alternative $R\beta \neq r$ because we have squared value above.

- Examples
  - Standard F–test, which is outputed from computer, is of the hypothesis

$$\beta_2 = 0, \ldots, \beta_K = 0,$$

where the intercept β_1 is included. In this case, $q = K - 1$, and

$$H_0 :\, R\beta = 0,$$

where

$$R = \begin{bmatrix} 0 & \\ \vdots & I_{K-1} \\ 0 & \end{bmatrix}.$$

The test statistic is compared with critical value from the $F(K-1,\, n-K)$ distribution.

– Structural Change. Null hypothesis is

$$y = X\beta + u.$$

Alternative is

$$\begin{aligned} y_1 &= X_1\beta_1 + u_1, \quad i \leq n_1, \\ y_2 &= X_2\beta_2 + u_2, \quad i \geq n_2, \end{aligned}$$

where $n = n_1 + n_2$. Let

$$\begin{aligned} y &= \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \; X^* = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}, \\ \beta^* &= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}_{2K\times 1}, \; u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_{n\times 1}. \end{aligned}$$

Then, we can write the alternative regression as

$$y = X^*\beta^* + u.$$

Consider the null hypothesis $H_0 : \beta_1 = \beta_2$. Let

$$R_{K\times 2K} = [I_K \vdots -I_K].$$

Compare with $F(K,\, n-2K)$.

- Confidence interval is just critical region centred not at H_0, but at a function of parameter estimates. For example,

$$c'\widehat{\beta} \pm t_{\alpha/2}(n-K)s\{c'(X'X)^{-1}c\}^{1/2}$$

is a two-sided confidence interval for the scalar quantity $c'\beta$. Can also construct one-sided confidence intervals and multivariate confidence intervals.

3.5 Test of Multiple Linear Hypothesis Based on fit

- The idea behind the F test is that under H_0,

$$R\widehat{\beta} - r$$

should be stochastically small, but under the alternative hypothesis it will not be so.

- An alternative approach is based on fit. Suppose we estimate β subject to the restriction $R\beta = r$, then the sum of squared residuals from that regression should be close to that from the unconstrained regression when the null hypothesis is true [but if it is false, the two regressions will have different fitting power].

- To understand this we must investigate the restricted least squares estimation procedure.

  - Unrestricted regression:

$$\min_b (y - Xb)'(y - Xb)$$

$\widehat{\beta}$, $\widehat{u} = y - X\widehat{\beta}$, $Q = \widehat{u}'\widehat{u}$.

  - Restricted regression:

$$\min_b (y - Xb)'(y - Xb) \text{ s.t. } Rb = r.$$

β^* , $u^* = y - X\beta^*$, $Q^* = u^{*'}u^*$

- The idea is that under H_0, $Q^* \sim Q$, but under the alternative the two quantities differ. The following theorem makes this more precise.

- Theorem. Under H_0,

$$\frac{Q^* - Q}{Q}\frac{n-K}{q} = F \sim F(q,\, n-K).$$

- **Proof**. We show that

$$Q^* - Q = (R\widehat{\beta} - r)' \left[R(X'X)^{-1}R'\right]^{-1} (R\widehat{\beta} - r)$$

Then, since

$$s^2 = Q/(n-K)$$

the result is established.

- To solve the restricted least squares problem we use the Lagrangean method. We know that β^* and λ^* solve the first order condition of the Lagrangean

$$\mathcal{L}(b, \lambda) = \frac{1}{2}(y - Xb)'(y - Xb) + \lambda'(Rb - r).$$

The first order conditions are

$$\begin{aligned} -X'y + X'X\beta^* + R'\lambda^* &= 0 (1) \\ R\beta^* &= r. (2) \end{aligned}$$

Now, from (1)

$$R'\lambda^* = X'y - X'X\beta^* = X'u^*,$$

which implies that

$$\begin{aligned} (X'X)^{-1}R'\lambda^* &= (X'X)^{-1}X'y - (X'X)^{-1}X'X\beta^* \\ &= \widehat{\beta} - \beta^* \end{aligned}$$

and

$$R(X'X)^{-1}R'\lambda^* = R\widehat{\beta} - R\beta^* = R\widehat{\beta} - r.$$

Therefore,

$$\lambda^* = \left[R(X'X)^{-1}R'\right]^{-1} (R\widehat{\beta} - r)$$

and

$$\beta^* = \widehat{\beta} - (X'X)^{-1}R' \left[R(X'X)^{-1}R'\right]^{-1} (R\widehat{\beta} - r).$$

This gives the restricted least squares estimator in terms of the restrictions and the unrestricted least squares estimator. From this relation we can derive the statistical properties of the estimator β^*.

- We now return to the testing question. First, write

$$\beta^* = \widehat{\beta} + \beta^* - \widehat{\beta}$$

and

$$\begin{aligned}
& (y - X\beta^*)'(y - X\beta^*) \\
= \; & [y - X\widehat{\beta} - X(\beta^* - \widehat{\beta})]'[y - X\widehat{\beta} - X(\beta^* - \widehat{\beta}] \\
= \; & (y - X\widehat{\beta})'(y - X\widehat{\beta}) + (\widehat{\beta} - \beta^*)'X'X(\widehat{\beta} - \beta^*) \\
& -(y - X\widehat{\beta})'X(\beta^* - \widehat{\beta}) \\
= \; & \widehat{u}'\widehat{u} + (\widehat{\beta} - \beta^*)'X'X(\widehat{\beta} - \beta^*)
\end{aligned}$$

using the orthogonality property of the unrestricted least squares estimator. Therefore,

$$Q^* - Q = (\widehat{\beta} - \beta^*)'X'X(\widehat{\beta} - \beta^*).$$

Substituting our formulae for $\widehat{\beta} - \beta^*$ and λ^* obtained above and cancelling out, we get

$$Q^* - Q = (R\widehat{\beta} - r)' \left[R(X'X)^{-1}R'\right]^{-1} (R\widehat{\beta} - r)$$

as required.

- An intermediate representation is

$$Q^* - Q = \lambda^{*'} R(X'X)^{-1}R'\lambda^*.$$

This brings out the use of the Lagrange Multipliers in defining the test statistic, and lead to the use of this name.

- Importance of the result: the fit version was easier to apply in the old days, before fast computers, because one can just do two separate regressions and use the sum of squared residuals. Special cases:

– Zero restrictions

$$\beta_2 = \cdots = \beta_K = 0$$

Then restricted regression is easy. In this case, $q = K - 1$. Note that the R^2 can be used to do an F–test of this hypothesis. We have

$$R^2 = 1 - \frac{Q}{Q^*} = \frac{Q^* - Q}{Q},$$

which implies that

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-k)}.$$

– Structural change. Allow coefficients to be different in two periods. Partition

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \begin{matrix} n_1 \\ n_2 \end{matrix}$$

$$\left.\begin{aligned} y_1 &= X_1\beta_1 + u_1 \\ y_2 &= X_2\beta_2 + u_2 \end{aligned}\right\} \text{ or } y = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + u.$$

Null is of no structural change, i.e.,

$$H_0 : \beta_1 = \beta_2,$$

with

$$R = (I \vdots -I).$$

- Consider the more general linear restriction

$$\begin{aligned} \beta_1 + \beta_2 - 3\beta_4 &= 1 \\ \beta_6 + \beta_1 &= 2. \end{aligned}$$

Harder to work with. Nevertheless, can always reparameterize to obtain restricted model as a simple regression. Partition X, β, and R

$$\begin{aligned} X &= \underset{n\times(k-q)}{(X_1} \; , \underset{n\times q}{X_2}) \quad ; \quad R = (\underset{q\times(k-q)}{R_1} \; , \underset{q\times q}{R_2}) \quad ; \\ \beta &= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \end{aligned}$$

where

$$X_1\beta_1 + X_2\beta_2 = X\beta \quad ; \quad R_1\beta_1 + R_2\beta_2 = r,$$

where R_2 is of full rank and invertible.

- Therefore,

$$\begin{aligned} \beta_2 &= R_2^{-1}(r - R_1\beta_1) \\ \\ X\beta &= X_1\beta_1 + X_2[R_2^{-1}(r - R_1\beta_1)] \\ &= (X_1 - X_2R_2^{-1}R_1)\beta_1 + X_2R_2^{-1}r, \end{aligned}$$

  so that

$$y - X_2R_2^{-1}r = (X_1 - X_2R_2^{-1}R_1)\beta_1 + u.$$

- In other words, we can regress

$$y^* = y - X_2R_2^{-1}r$$

  on

$$X_1^* = (X_1 - X_2R_2^{-1}R_1)$$

  to get β_1^*, and then define

$$\beta_2^* = R_2^{-1}(r - R_1\beta_1^*).$$

  We then define

$$u^* = y - X_1\beta_1^* - X_2\beta_2^* \text{ and } Q^*$$

  accordingly.

3.6 Examples of F–Tests, t vs. F

- Chow Tests: Structural change with intercepts. The unrestricted model is

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} i_1 & 0 & x_1 & 0 \\ 0 & i_2 & 0 & x_2 \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

  and let $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2)$. Different slopes and intercepts allowed.

- The first null hypothesis is that the slopes are the same, i.e.,

$$H_0 : \beta_1 = \beta_2 = \beta.$$

The restricted regression is

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} i_1 & 0 & x_1 \\ 0 & i_2 & x_2 \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

The test statistic is

$$F = \frac{(u^{*'}u^* - \widehat{u}'\widehat{u})/\dim(\beta_1)}{\widehat{u}'\widehat{u}/(n - \dim(\theta))},$$

which is compared with the quantiles from the

$$F(\dim(\beta_1), n - \dim(\theta))$$

distribution.

- The second null hypothesis is that the intercepts are the same, i.e.,

$$H_0 : \alpha_1 = \alpha_2 = \alpha.$$

Restricted regression (α, β_1, β_2)

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} i_1 & x_1 & 0 \\ i_2 & 0 & x_2 \end{bmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

Note that the unrestricted model can be rewritten using dummy variables:

$$y_i = \alpha + \beta x_i + \gamma D_i + \delta x_i D_i + u_i,$$

where

$$D_i = \begin{cases} 1 & \text{in period 2} \\ 0 & \text{else.} \end{cases}$$

Then, in period 1

$$y_i = \alpha + \beta x_i + u_i,$$

while in period 2

$$y_i = \alpha + \gamma + (\beta + \delta)x_i + u_i.$$

The null hypothesis is that $\gamma = 0$.

- But now suppose that $n_2 < K$. The restricted regression is ok, but the unrestricted regression runs into problems in the second period because n_2 is too small. In fact, $\widehat{u}_2 \equiv 0$. In this case we must simply acknowledge the fact that the degrees of freedom lost are n_2 not K. Thus
$$F = \frac{(Q^* - Q)/n_2}{Q/(n_1 - K)} \sim F(n_2,\, n_1 - K)$$
is a valid test in this case.

3.7 Likelihood Based Testing

- We have considered several different approaches which all led to the F test in linear regression. We now consider a general class of test statistics based on the Likelihood function. In principle these apply to any parametric model, but we shall at this stage just consider its application to linear regression.

- The Likelihood is denoted $L(y, X; \theta)$, where y, X are the observed data and θ is a vector of unknown parameter. The maximum likelihood estimator can be determined from $L(y, X; \theta)$, as we have already discussed. This quantity is also useful for testing.

- Consider again the linear restrictions
$$H_0 \,:\, R\theta = r.$$
  - The unrestricted maximum likelihood estimator of θ is denoted by $\widehat{\theta}$
  - the restricted MLE is denoted by θ^*, [this maximizes L subject to the restrictions $R\theta - r = 0$].

- Now define the following test statistics:
$$\begin{aligned}
\text{LR} &: 2\left[\log \frac{L(\widehat{\theta})}{L(\theta^*)}\right] = 2\{\log L(\widehat{\theta}) - \log L(\theta^*)\} \\
\text{Wald} &: (R\widehat{\theta} - r)' \left\{ RH(\widehat{\theta})^{-1} R' \right\}^{-1} (R\widehat{\theta} - r) \\
\text{LM} &: \frac{\partial \log L}{\partial \theta}\Big|'_{\theta^*} H(\theta^*)^{-1} \frac{\partial \log L}{\partial \theta}\Big|_{\theta^*},
\end{aligned}$$

where

$$H(\theta) = -\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\Big|_\theta$$

- The Wald test only requires computation of the unrestricted estimator
- the Lagrange Multiplier only requires computation of the restricted estimator.
- The Likelihood ratio requires computation of both.
- There are circumstances where the restricted estimator is easier to compute, and there are situations where the unrestricted estimator is easier to compute. These computational differences are what has motivated the use of either the Wald or the LM test.
- When it comes to nonlinear restrictions $g(\theta) = 0$, the LR test has the advantage that it is invariant to the parameterization, while the Wald test is affected by the way in which the restrictions are expressed.

- In the linear regression case, $\theta = (\beta, \sigma^2)$, and the restrictions only apply to β, so that $R\beta = r$. Therefore, we can replace the derivatives with respect to θ by derivatives with respect to β only.
- The log-likelihood is

$$\log L(\theta) = \frac{-n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}u(\beta)'u(\beta)$$

and its derivatives are

$$\begin{aligned}
\frac{\partial \log L}{\partial \beta} &= \frac{1}{\sigma^2}X'u(\beta) \\
\frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}u(\beta)'u(\beta) \\
\frac{\partial^2 \log L}{\partial\beta\partial\beta'} &= \frac{-1}{\sigma^2}X'X \\
\frac{\partial^2 \log L}{(\partial\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{2}{2\sigma^6}u(\beta)'u(\beta) \\
\frac{\partial^2 \log L}{\partial\beta\partial\sigma^2} &= -\frac{1}{\sigma^4}X'u(\beta).
\end{aligned}$$

- The Wald test is

$$\begin{aligned} W &= (R\widehat{\beta} - r)' \left[R(X'X)^{-1}R'\widehat{\sigma}^2\right]^{-1} (R\widehat{\beta} - r) \\ &= \frac{Q^* - Q}{(Q/n)}, \end{aligned}$$

  where

$$\widehat{\sigma}^2 = Q/n$$

  is the MLE of σ^2.

  - The Wald test statistic is the same as the F–test apart from the use of $\widehat{\sigma}^2$ instead of s^2 and a multiplicative factor q. In fact,

$$W = qF\frac{n}{n-k}.$$

    This is approximately equal to qF when the sample size is large.

- The Lagrange Multiplier or Score or Rao test statistic is

$$LM = \frac{u^{*'}X}{\sigma^{*2}} \left\{\frac{X'X}{\sigma^{*2}}\right\}^{-1} \frac{X'u^*}{\sigma^{*2}},$$

  where

$$\sigma^{*2} = Q^*/n.$$

  - Recall that

$$X'u^* = R'\lambda^*.$$

    Therefore,

$$LM = \frac{\lambda^{*'}R(X'X)^{-1}R'\lambda^*}{\sigma^{*2}},$$

    where λ^* is the vector of Lagrange Multipliers evaluated at the optimum.

  - Furthermore, we can write the score test as

$$LM = \frac{Q^* - Q}{(Q^*/n)} = n\left(1 - \frac{Q}{Q^*}\right).$$

    When the restrictions are the standard zero ones, the test statistic is n times the R^2 from the unrestricted regression.

- The Likelihood Ratio

$$\log L(\widehat{\beta},\widehat{\sigma}^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\widehat{\sigma}^2 - \frac{1}{2\widehat{\sigma}^2}\widehat{u}'\widehat{u} = -\frac{n}{2}\log 2\pi - \frac{1}{2\widehat{\sigma}^2}\log\frac{\widehat{u}'\widehat{u}}{n} - \frac{n}{2}$$

and

$$\log L(\beta^*,\sigma^{*2}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^{*2} - \frac{1}{2\sigma^{*2}}u^{*\prime}u^* = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\frac{u^{*'}u^*}{n} - \frac{n}{2}.$$

These two lines follow because

$$\widehat{\sigma}^2 = \widehat{u}'\widehat{u}/n \text{ and } \sigma^{*2} = u^{*\prime}u^*/n.$$

Therefore,

$$LR = 2\log\frac{L(\widehat{\beta},\,\widehat{\sigma}^2)}{L(\beta,\,\sigma^2)} = n\left[\log\frac{Q^*}{n} - \log\frac{Q}{n}\right] = n[\log Q^* - \log Q].$$

- Note that W, LM, and LR are all *monotonic* functions of F, in fact

$$\begin{aligned} W &= F\frac{qn}{n-k}, \\ LM &= \frac{W}{1+W/n}\,, \\ LR &= n\log\left(1+\frac{W}{n}\right). \end{aligned}$$

If we knew the exact distribution of any of them we can obtain the exact distributions of the others and the test result will be the same.

- However, in practice one uses asymptotic critical values, which lead to differences in outcomes.We have

$$LM \leq LR \leq W,$$

so that the Wald test will reject more frequently than the LR test and the LM tests, supposing that the same critical values are used.

- Also,

$$qF \leq W$$

Chapter 4

Further Topics in Estimation:

4.1 Omission of Relevant Variables

- Suppose that
$$y = X_1\beta_1 + X_2\beta_2 + u,$$
where the error term obeys the usual conditions.

- Suppose however that we regress y on X_1 only. Then,
$$\begin{aligned}\widehat{\beta}_1 &= (X_1'X_1)^{-1}X_1'y \\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'u,\end{aligned}$$
so that
$$\begin{aligned}E(\widehat{\beta}_1) &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \\ &= \beta_1 + \beta_{12},\end{aligned}$$
where
$$\beta_{12} = (X_1'X_1)^{-1}X_1'X_2\beta_2.$$
In general $\widehat{\beta}_1$ is biased and inconsistent; the direction and magnitude of the bias depends on β_2 and on $X_1'X_2$.

- Example. Wages on education get positive effect but are omitting ability. If ability has a positive effect on wages and is positively correlated with education this would explain some of the positive effect. Wages on race/gender (discrimination). Omit experience/education.

- What about variance? In fixed design, variance is $\sigma^2(X_1'X_1)^{-1}$, which is smaller than when X_2 is included. Therefore, if MSE is the criterion, one may actually prefer this procedure - at least in finite samples.

- Estimated variance is
$$s^2(X_1'X_1)^{-1},$$
where
$$\begin{aligned} s^2 &= \frac{y'M_1y}{n-K_1} \\ &= \frac{(X_2\beta_2+u)'M_1(X_2\beta_2+u)}{n-K_1}, \end{aligned}$$
which has expectation
$$\begin{aligned} E(s^2) &= \sigma^2 + \frac{\beta_2'X_2'M_1X_2\beta_2}{n-K_1} \\ &\geq \sigma^2, \end{aligned}$$
since M_1 is a positive semi-definite matrix.

  - Therefore, the estimated variance of $\widehat{\beta}_1$ is upwardly biased.

- If $X_1'X_2 = 0$, then $\widehat{\beta}$ is unbiased, but standard errors are still biased with expectation
$$\sigma^2 + \frac{\beta_2'X_2'X_2\beta_2}{n-K_1}.$$
In this special case, the t–ratio is downward biased.

- More generally, t–ratio could be upward or downward biased depending of course on the direction of the bias of $\widehat{\beta}_1$.

- Some common examples of omitted variables

  - Seasonality
  - Dynamics
  - Nonlinearity.

- In practice we might suspect that there are always going to be omitted variables. The questions is: is the magnitude large and the direction unambiguous? To address this question we first look at the consequences of including too many variables in the regression.

4.2 Inclusion of irrelevant variables

- Suppose now that
$$y = X_1\beta_1 + u,$$
where u obeys the usual conditions.

- However, we regress y on both X_1 and X_2. Then
$$\begin{aligned} \widehat{\beta}_1 &= (X_1'M_2X_1)^{-1}X_1'M_2y \\ &= \beta_1 + (X_1'M_2X_1)^{-1}X_1'M_2u \end{aligned}$$

- Therefore
$$\begin{aligned} E(\widehat{\beta}_1) &= \beta_1 \text{ all } \beta_1 \\ \text{var}(\widehat{\beta}_1) &= \sigma^2(X_1'M_2X_1)^{-1}. \end{aligned}$$

- Compare this with the variance of y on X_1, which is only $\sigma^2(X_1'X_1)^{-1}$. Now
$$X_1'X_1 - X_1'M_2X_1 = X_1'P_2X_1 \geq 0$$
which implies that
$$(X_1'X_1)^{-1} - (X_1'M_2X_1)^{-1} \leq 0.$$
Always better off, as far as variance is concerned, with the smaller model.

- We can generalize the above discussion to the case where we have some linear restrictions $R\beta = r$. In which case, the restricted estimator is
$$\beta^* = \widehat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta} - r)$$
If we estimate by restricted least squares we get smaller variance but if the restriction is not true, then there is a bias.

- There is clearly a trade-off between bias and variance.
- The above arguments suggest that including irrelevant variables never leads to bias, but this is not correct. We relied above on the assumption that the included regressors are all fixed and therefore the error term is uncorrelated with them. Clearly, if one of the included right hand side variables was say y, then you would definitely get a biased estimate of the coefficient on the remaining variables.

4.3 Model Selection

- Let $\mathcal{M}$ be a collection of *linear* regression models obtained from a given set of K regressors $X = (X_1, \ldots, X_K)$, e.g., $X, X_1, (X_2, X_{27}), etc.$ Suppose that the true model lies in $\mathcal{M}$. There are a total of $(2^K - 1)$ different subsets of X, i.e., models.
- Let K_j be the number of explanatory variables in a given regression. The following criteria can be used for selecting the 'best' regression:

$$\overline{R}_j^2 = 1 - \frac{n-1}{n-K_j}(1-R_j^2) = 1 - \frac{n-1}{n-K_j}\frac{\widehat{u}_j\widehat{u}_j}{\overline{u}'\overline{u}},$$

$$PC_j = \frac{\widehat{u}_j'\widehat{u}_j}{n-K_j}\left(1+\frac{K_j}{n}\right)$$

$$AIC_j = \ln\frac{\widehat{u}_j'\widehat{u}_j}{n} + \frac{2K_j}{n}$$

$$BIC_j = \ln\frac{\widehat{u}_j'\widehat{u}_j}{n} + \frac{K_j\log n}{n}.$$

  The first criterion should be maximized, while the others should be minimized. Note that maximizing $\overline{R}_j^2$ is equivalent to minimizing the unbiased variance estimate $\widehat{u}_j\widehat{u}_j/(n-K_j)$.
- It has been shown that all these methods have the property that the selected model is larger than or equal to the true model with probability tending to one; only BIC_j correctly selects the true model with probability tending to one.
- $\mathcal{M}$ may be large and computing $2^K - 1$ regressions infeasible.

- True model may not be in $\mathcal{M}$, but procedure is guaranteed to find a best model (data mining).
- Other criteria are important, especially for nonexperimental data.
  - Consistency with economic theory elasticities the right sign? Demand slopes down?
  - Consistency with data, e.g., suppose dependent variable is food share $\notin [0,1]$, then ideally don't want a model that predicts outside this range.
  - Residuals should be approximately random, i.e., pass diagnostic checks for serial correlation, heteroskedasticity, nonlinearity, etc.
  - How well model performs out-of-sample. (Often used in time series analysis.)
  - Correlation is not causation.
- An alternative strategy is to choose a large initial model and perform a sequence of t–tests to eliminate redundant variables. Finally, we give a well known result that links the properties of the regression t test and the R squared.
- $\overline{R}^2$ falls (rises) when the deleted variable has $t > (<)1$

4.4 Multicollinearity

- Exact multicollinearity: $X'X$ is singular, i.e., there is an *exact, linear*, relationship between variables in X. In this case, cannot define least squares estimates
$$\widehat{\beta} = (X'X)^{-1}X'y.$$
Solution: Find a minimal (not unique) basis X^* for $\mathcal{C}(X)$ and do least squares.
- *Example*: Seasonal dummies
$$\begin{aligned} D1 &= 1\text{if Quarter } 1,\ 0 \text{ else} \\ D2 &= 1\text{if Quarter } 2,\ 0 \text{ else} \\ D3 &= 1\text{if Quarter } 3,\ 0 \text{ else} \\ D4 &= 1\text{if Quarter } 4,\ 0 \text{ else.} \end{aligned}$$

Define the regressor matrix

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ \vdots & 0 & 1 & 0 & 0 \\ \vdots & 0 & 0 & 1 & 0 \\ \vdots & 0 & 0 & 0 & 1 \\ 1 & \vdots & \vdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

In this case, for all observations:

$$Col2 + Col3 + Col4 + Col5 = Col1$$

- Solution
  - Drop $D4$, and run
    $$y = \alpha + \beta_1 D1 + \beta_2 D2 + \beta_3 D3 + u$$
  - Drop intercept, and run
    $$y = \gamma_1 D2 + \gamma_2 D2 + \gamma_3 D3 + \gamma_4 D4 + u.$$
    Gives same $\widehat{y}$ and $\widehat{u}$, but different parameters. Intuitively, the same vector space is generated by both sets of regressors.

- 'Approximate Multicollinearity', i.e., $\det(X'X) \approx 0$. Informally, if the columns of X are highly mutually correlated then it is hard to get their separate effects. This is really a misnomer and shan't really be treated as a separate subject. Arthur Goldberger in his text on econometrics illustrated this point by having a section on 'micronumerosity', a supposed problem where one has too few observations. The consequence of this is that the variance of the parameter estimates is large - precisely the symptom of 'Approximate Multicollinearity'.

4.5 Influential Observations

- At times one can suspect that some observations are having a large impact on the regression results. This could be a real influence, i.e., just part of the way the data were generated, or it could be because some observations have been misrecorded, say with an extra zero added on by a careless clerk.

- How do we detect influential observations? Delete one observation at a time and see what changes. Define the leave-one-out estimator and residual

$$\begin{aligned} \widehat{\beta}(i) &= [X(i)'X(i)]^{-1}X(i)'y(i) \\ \widehat{u}_j(i) &= y_j - X_j'\widehat{\beta}(i), \end{aligned}$$

  where

$$y(i) = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)'$$

  and similarly for $X(i)$. We shall say that observation $(X_i,\, y_i)$ is influential if $\widehat{u}_i(i)$ is large.

- Note that

$$\widehat{u}_i(i) = u_i - X_i'(\widehat{\beta}(i) - \beta),$$

  so that

$$\begin{aligned} E[\widehat{u}_i(i)] &= 0 \\ \text{var}[\widehat{u}_i(i)] &= \sigma^2[1 + x_i'(X'(i)X(i))^{-1}x_i]. \end{aligned}$$

  Then examine standardized residuals

$$T_i = \frac{\widehat{u}_i(i)}{\dot{s}(1 + x_i'(X'(i)X(i))^{-1}x_i)^{1/2}}.$$

- Large values of T_i, in comparison with standard normal, are evidence of extreme observations or outliers. Unfortunately, we do not learn whether this is because the error distribution has a different shape from the normal, e.g., t-distribution, or whether the observation has been misrecorded by some blundering clerk.

4.6 Missing Observations

- In surveys, responders are not representative sample of full population. For example, we don't have information on people with $y > \$250,000$, $y \leq \$5,000$. In this case, $\frac{1}{n}\sum_{i=1}^{n} y_i$ is biased and inconsistent as an estimate of the population mean.

- In regression, parameter estimates are biased if selection is:

  - On dependent variable (or on error term);
  - Non-random. For example, there is no bias [although precision is affected] if in a regression of inc on education, we have missing data when edu ≥ 5 years.

- We look at 'ignorable' case where the process of missingness is unrelated to the effect of interest.

- Missing y

$$\begin{array}{ccc} y_A & X_A & n_A \\ ? & X_B & n_B. \end{array}$$

- What do we do? One solution is to impute values of the missing variable. In this case, we might let

$$\widehat{y}_B = X_B\widehat{\beta}_A, \text{ where } \widehat{\beta}_A = (X_A'X_A)^{-1}X_A'y_A.$$

  We can then recompute the least squares estimate of β using 'all the data'

$$\begin{aligned} \widehat{\beta} &= (X'X)^{-1}X' \begin{bmatrix} y_A \\ \widehat{y}_B \end{bmatrix}, \\ X &= \begin{pmatrix} X_A \\ X_B \end{pmatrix}. \end{aligned}$$

  However, some simple algebra reveals that there is no new information in $\widehat{y}_B$, and in fact

$$\widehat{\beta} = \widehat{\beta}_A.$$

- Start from

$$(X'X)^{-1}X'\begin{bmatrix} y_A \\ X_B\widehat{\beta}_A \end{bmatrix} = (X_A'X_A)^{-1}X_A'y_A,$$

and pre-multiply both sides by

$$X'X = (X_A'X_A + X_B'X_B).$$

- Ee have

$$\begin{aligned} X'\begin{bmatrix} y_A \\ X_B\widehat{\beta}_A \end{bmatrix} &= X_A'y_A + X_B'X_B\widehat{\beta}_A \\ &= X_A'y_A + X_B'X_B(X_A'X_A)^{-1}X_A'y_A \end{aligned}$$

and

$$X'X(X_A'X_A)^{-1}X_A'y_A = X_A'y_A + (X_B'X_B)(X_A'X_A)^{-1}X_A'y_A.$$

Therefore, this imputation method has not really added anything. It is not possible to improve estimation in this case.

- Now suppose that we have some missing X. For example, $X = (x, z)$, and x_B is missing, i.e., we observe (x_A, z_A, y_A) and (z_B, y_B). The model for the complete data set is

$$y = \beta x + \gamma z + u$$

with $\text{var}(u) = \sigma_u^2$, and suppose also that

$$x = \delta z + \epsilon$$

with ϵ being an iid mean zero error term with $\text{var}(\epsilon) = \sigma_u^2$.

- There are a number of ways of trying to use the information in period B. First, predict x_B by regressing x_A on z_A

$$\widehat{x}_B = z_B(z_A'z_A)^{-1}z_A'x_A.$$

Then regress y on

$$\widehat{X} = \begin{pmatrix} x_A & z_A \\ \widehat{x}_B & z_B \end{pmatrix}.$$

- The second approach is to write now

$$\begin{aligned} y_A &= \beta x_A + \gamma z_A + u_A \\ x_A &= \delta z_A + \epsilon_A \\ y_B &= (\gamma + \beta\delta) z_B + u_B + \beta\epsilon_B, \end{aligned}$$

  where we have substituted out the x_B, which we don't observe. Now we can estimate β, γ, and δ from the A observations, denoting these estimates by $\widehat{\beta}_A, \widehat{\gamma}_A$, and $\widehat{\delta}_A$. Then we have a new regression with

$$y_B - \widehat{\beta}_A \widehat{\delta}_A z_B = \gamma z_B + e_B$$

  for some error term e that includes $u_B + \beta\epsilon_B$ plus the estimation error in $\widehat{\beta}_A \widehat{\delta}_A$. This regression can be jointly estimated with the

$$y_A - \widehat{\beta}_A x_A = \gamma z_A + e_A.$$

- This sometimes improves matters, but sometimes does not! The answer depends on relationship between x and z. In any case, the effect β of x is not better estimated; the effect of z maybe improved. Griliches (1986) shows that the (asymptotic) relative efficiency of this approach to the just use A least squares estimator is

$$(1 - \lambda)\left(1 + \lambda\beta^2 \frac{\sigma_\epsilon^2}{\sigma_u^2}\right),$$

  where λ is the fraction of the sample that is missing. Efficiency will be improved by this method when

$$\beta^2 \frac{\sigma_\epsilon^2}{\sigma_u^2} < \frac{1}{1 - \lambda},$$

  i.e., the unpredictable part of x from z is not too large relative to the overall noise in the y equation.

- Another approach. Let $\theta = \gamma + \beta\delta$. Then clearly, we can estimate θ from the B data by OLS say, call this $\widehat{\theta}_B$. Then let $\widehat{\gamma}_B = \widehat{\theta}_B - \widehat{\beta}_A \widehat{\delta}_A$. Now consider the class of estimators

$$\widehat{\gamma}(\omega) = \omega\widehat{\gamma}_A + (1 - \omega)\widehat{\gamma}_B,$$

as ω varies. In the homework 2 we showed that the best choice of ω is

$$\omega_{opt} = \frac{\sigma_A^2 - \sigma_{AB}}{\sigma_A^2 + \sigma_B^2 - 2\sigma_{AB}},$$

where in our case σ_A^2, σ_B^2 are the asymptotic variances of the two esitmators and σ_{AB} is their asymptotic vcovariance. Intuitively, unless either $\sigma_A^2 = \sigma_{AB}$ or $\sigma_B^2 = \sigma_{AB}$, we should be able to improve matters.

- What about the likelihood approach? Suppose for convenience that z is a fixed variable, then the log likelihood function of the observed data is

$$\sum_A \log f(y_A, x_A|z_A) + \sum_B \log f(y_B|z_B).$$

Suppose that u, ϵ are normally distributed and mutually independent, then

$$\begin{pmatrix} y_A \\ x_A \end{pmatrix} \sim N\left[\begin{pmatrix} (\gamma + \beta\delta)z_A \\ \delta z_A \end{pmatrix}, \begin{pmatrix} \sigma_u^2 + \beta^2\sigma_\epsilon^2 & \beta\sigma_\epsilon^2 \\ & \sigma_\epsilon^2 \end{pmatrix}\right]$$

$$y_B \sim N\left[(\gamma + \beta\delta)z_B, \sigma_u^2 + \beta^2\sigma_\epsilon^2\right],$$

which follows from the relations $x = \delta z + \epsilon$ and $y = (\gamma + \beta\delta)z + u + \beta\epsilon$. There are five unknown parameters $\theta = (\gamma, \beta, \delta, \sigma_u^2, \sigma_\epsilon^2)$. The likelihood follow from this.

- The MLE is going to be quite complicated here because the error variances depend on the mean parameter β, but it going to be more efficient than the simple least squares that only uses the A data.

Chapter 5

Asymptotics

5.1 Types of Asymptotic Convergence

- Exact distribution theory is limited to very special cases [normal i.i.d. errors linear estimators], or involves very difficult calculations. This is too restrictive for applications. By making approximations based on large sample sizes, we can obtain distribution theory that is applicable in a much wider range of circumstances.

- Asymptotic theory involves generalizing the usual notions of convergence for real sequences to allow for random variables. We say that a real sequence x_n converges to a limit x_∞, denoted $\lim_{n\to\infty} x_n = x_\infty$, if for all $\epsilon > 0$ there exists an n_0 such that

$$|x_n - x_\infty| < \epsilon$$

  for all $n \geq n_0$.

- DEFINITION: We say that a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges in probability to a random variable X, denoted,

$$X_n \xrightarrow{P} X \text{ or } \operatorname*{p\,lim}_{n\to\infty} X_n = X.$$

  if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr[|X_n - X| > \varepsilon] = 0.$$

  X could be a constant or a random variable.

- We say that a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges almost surely or with probability one to a random variable X, denoted

$$X_n \xrightarrow{a.s.} X,$$

if

$$\Pr[\lim_{n\to\infty} X_n = X] = 1.$$

- DEFINITION: We say that a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges in distribution to a random variable X, denoted,

$$X_n \xrightarrow{D} X,$$

if for all x,

$$\lim_{n\to\infty} \Pr[X_n \leq x] = \Pr[X \leq x].$$

Specifically, we often have

$$n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} N(0, \sigma^2) \ .$$

- DEFINITION: We say that a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges in mean square to a random variable X, denoted

$$X_n \xrightarrow{m.s.} X,$$

if

$$\lim_{n\to\infty} E[|X_n - X|^2] = 0.$$

  - This presumes of course that $EX_n^2 < \infty$ and $EX^2 < \infty$.
  - When X is a constant,

$$\begin{aligned} E[|X_n - X|^2] &= E[|X_n - EX_n|^2] + |EX_n - X|^2 \\ &= \text{var}(X_n) + |EX_n - X|^2, \end{aligned}$$

and it is necessary and sufficient that

$$EX_n \to X \text{ and } \text{var}(X_n) \to 0.$$

- Mean square convergence implies convergence in probability. This follows from the Chebychev inequality

$$\Pr[|X_n - X| > \varepsilon] \leq \frac{E[|X_n - X|^2]}{\varepsilon^2}.$$

- Note that convergence in probability is stronger than convergence in distribution, but they are equivalent when X is a constant (i.e., not random). Almost sure convergence implies convergence in probability, but there is no necessary relationship between almost sure convergence and convergence in mean square. Examples where convergence in distribution does not imply convergence in probability.

5.2 Laws of Large Numbers and Central Limit Theorems

- (Kolmogorov Law of Large Numbers) Suppose that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.). Then a necessary and sufficient condition for

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu \equiv E(X_1),$$

is that

$$E(|X_i|) < \infty.$$

- (Lindeberg-Levy Central Limit Theorem) Suppose that $X_1, \ldots, X_n$ are i.i.d. with $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$. Then

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \xrightarrow{D} N(0,\, 1).$$

- These results are important because many estimators and test statistics can be reduced to sample averages or functions thereof. There are now many generalizations of these results for data that are not i.i.d., e.g., heterogeneous, dependent weighted sums. We give one example

- (Lindeberg-Feller) Let $X_1, \ldots, X_n$ be independent random variables with $E(X_i) = 0$ and $\text{var}(X_i) = \sigma_i^2$. Suppose also that Lindeberg's condition holds: for all $\epsilon > 0$,

$$\frac{1}{\sum_{i=1}^n \sigma_i^2} \sum_{i=1}^n E\left[X_i^2 1\left(X_i^2 > \epsilon \sum_{j=1}^i \sigma_j^2\right)\right] \to 0.$$

  Then

$$\frac{1}{(\sum_{i=1}^n \sigma_i^2)^{1/2}} \sum_{i=1}^n X_i \xrightarrow{D} N(0,\, 1).$$

- A sufficient condition for the Lindeberg condition is that

$$E[|X_i|^3] < \infty.$$

5.3 Additional Results

- Mann–Wald Theorem.
  - If $X_n \xrightarrow{D} X$ and if g is continuous, then
    $$g(X_n) \xrightarrow{D} g(X).$$
  - If $X_n \xrightarrow{P} \alpha$, then
    $$g(X_n) \xrightarrow{P} g(\alpha).$$
- Slutsky Theorem. If $X_n \xrightarrow{D} X$, $y_n \xrightarrow{P} \alpha$, then:
  - $X_n + y_n \xrightarrow{D} X + \alpha$;
  - $X_n y_n \xrightarrow{D} \alpha X$; and
  - $X_n / y_n \xrightarrow{D} X/\alpha$, provided $\alpha \neq 0$.
- Vector random variables. Consider the vector sequence
  $$X_n = (X_{n1}, \ldots, X_{nk})'.$$
  We have the result that
  $$\|X_n - X\| \xrightarrow{P} 0,$$
  where $\|x\| = (x'x)^{1/2}$ is Euclidean norm, if and only if
  $$|X_{nj} - X_j| \xrightarrow{P} 0 \quad \text{for all } j = 1, \ldots, k.$$
  - The if part is no surprise and follows from the continuous mapping theorem. The only if part follows because if
    $$\|X_n - X\| < \varepsilon$$
    then there exists a constant c such that
    $$|X_{nj} - X_j| < \varepsilon/c$$
    for each j.
- Cramers Theorem. A vector X_n converges in distribution to a normal vector X if and only if $a'X_n$ converges in distribution to $a'X$ for every vector a.

5.4 Applications to OLS

- We are now able to establish some results about the large sample properties of the least squares estimator. We start with the i.i.d. random design case because the result is very simple.

- If we assume that:

  - x_i, ε_i are i.i.d. with $E(x_i\varepsilon_i) = 0$
  - $0 < E\left[x_i x_i'\right] < \infty$ and $E\left[\|x_i\varepsilon_i\|\right] < \infty$.
  - Then,
$$\widehat{\beta} \xrightarrow{P} \beta.$$

- The proof comes from applying laws of large numbers to the numerator and denominator of
$$\widehat{\beta} - \beta = \left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i.$$
These regularity conditions are often regarded as unnecessary and perhaps strong and unsuited to the fixed design.

- We next consider the 'bare minimum' condition that works in the fixed design case and is perhaps more general since it allows for example trending variables.

- Theorem. Suppose that A0-A2 hold and that
$$\lambda_{\min}(X'X) \to \infty \quad \text{as} \quad n \to \infty. \qquad (\star)$$
Then,
$$\widehat{\beta} \xrightarrow{P} \beta.$$

- Proof. First,
$$E(\widehat{\beta}) = \beta$$
for all β. Then,
$$\text{var}(\widehat{\beta}) = \sigma^2(X'X)^{-1},$$

where

$$\left\|(X'X)^{-1}\right\| = \lambda_{\max}((X'X)^{-1}) = \frac{1}{\lambda_{\min}(X'X)},$$

and provided $(\star)$ is true,

$$\mathrm{var}(\widehat{\beta}) \to 0.$$

- Suppose that $x_i = i^\alpha$ for some α, then

$$\mathrm{var}(\widehat{\beta}) = \frac{\sigma^2}{\sum_{j=1}^n j^{2\alpha}} = \begin{cases} O(n^{-(2\alpha+1)}) & \text{if } \alpha \neq -1/2 \\ O(1/\log n) & \text{if } \alpha = -1/2 \end{cases}.$$

  Therefore, consistency holds if and only if $\alpha \geq -1/2$.

- If we have a random design then the conditions and conclusion should be interpreted as holding with probability one in the conditional distribution given X. Under the above random design assumptions, $(\star)$ holds with probability one.

5.5 Asymptotic Distribution of OLS

- We first state the result for the simplest random design case.
- Suppose that
  - x_i, ε_i are i.i.d. with ε_i independent of x_i,
  - $E(\varepsilon_i^2) = \sigma^2$
  - $0 < E\left[x_i x_i'\right] < \infty$.
  - Then,

$$n^{1/2}(\widehat{\beta} - \beta) \xrightarrow{D} N(0,\, \sigma^2 \left\{E\left[x_i x_i'\right]\right\}^{-1}).$$

- Proof uses Mann–Wald Theorem and Slutsky Theorems.
- We next consider the fixed design case [where the errors are i.i.d. still]. In this case, it suffices to have a vector central limit theorem for the weighted i.i.d. sequence

$$\widehat{\beta} - \beta = (\sum_{i=1}^n x_i x_i')^{-1} \sum_{i=1}^n x_i \varepsilon_i = \sum_{i=1}^n w_i \varepsilon_i,$$

for some weights w_i depending only on the X data. That is, the source of the heterogeneity is the fixed regressors.

- A sufficient condition for the scalar standardized random variable

$$T_n = \frac{\sum_{i=1}^n w_i \varepsilon_i}{(\sum_{i=1}^n w_i^2 \sigma^2)^{1/2}}$$

to converge to a standard normal random variable is the following condition

$$\frac{\max_{1 \le i \le n} w_i^2}{\sum_{i=1}^n w_i^2} \to 0.$$

This is a so-called negligibility requirement, which means that no one of the weights

dominates every other term.

- Therefore,

$$\left(\sum_{i=1}^n x_i x_i' / \sigma^2 \right)^{1/2} (\widehat{\beta} - \beta) \xrightarrow{D} N(0, 1),$$

provided the following negligibility condition holds:

$$\max_{1 \le i \le n} x_i (X'X)^{-1} x_i' \to 0 \text{ as } n \to \infty.$$

Actually it suffices for the diagonal elements of this matrix to converge to zero. This condition is usually satisfied.

- If also $X'X/n \to M > 0$, then

$$n^{1/2}(\widehat{\beta} - \beta) \xrightarrow{D} N(0, \sigma^2 M^{-1}),$$

- Suppose $k = 1$, then the negligibility condition is

$$\max_{1 \le i \le n} \frac{x_i^2}{\sum_{j=1}^n x_j^2} \to 0.$$

For example, if $x_i = i$,

$$\frac{\max_{1 \le i \le n} i^2}{\sum_{j=1}^n j^2} = \frac{n^2}{O(n^3)} \to 0.$$

In this case, even though the largest element is increasing with sample size many other elements are increasing just as fast.

- An example, where the CLT would fail is

$$x_i = \begin{cases} 1 & \text{if } i < n \\ n & \text{if } i = n. \end{cases}$$

In this case, the negligibility condition fails and the distribution of the least squares estimator would be largely determined by the last observation.

5.6 Order Notation

- In the sequel we shall use the Order notation:

$$X_n = o_p(\delta_n) \text{ if } \frac{X_n}{\delta_n} \xrightarrow{P} 0$$

and

$$X_n = O_p(\delta_n) \text{ if } X_n/\delta_n \text{ is stochastically bounded,}$$

i.e., if for all K,

$$\lim_{n\to\infty} \Pr\left[\left|\frac{X_n}{\delta_n}\right| > K\right] < 1.$$

The latter means that X_n is of no larger order than δ_n, while the first one is stronger and says that X_n is of smaller order than δ_n. These concepts correspond to the $o(\cdot)$ and $O(\cdot)$ used in standard real analysis.

- The order symbols obey the following algebra, which is really just the Slutsky theorem:

$$\begin{aligned} O_p(1)o_p(1) &= o_p(1) \\ O_p(a_n)O_p(b_n) &= O_p(a_n b_n) \\ O_p(a_n) + O_p(b_n) &= O_p(\max\{a_n,\, b_n\}). \end{aligned}$$

5.7 Standard Errors and Test Statistics in Linear Regression

- We first consider the standard error. We have

$$s^2 = \frac{\widehat{u}'\widehat{u}}{n-k}$$

$$\begin{aligned} &= \frac{1}{n-k}\{u'u - u'X(X'X)^{-1}X'u\} \\ &= \left(\frac{n}{n-k}\right)\frac{u'u}{n} - \frac{1}{n-k}\frac{u'X}{n^{1/2}}(X'X/n)^{-1}X'u/n^{1/2}. \end{aligned}$$

- Theorem. Suppose that u_i are i.i.d. with finite fourth moment, and that the regressors are from a fixed design and satisfy $(X'X/n) \to M$,where M is a positive definite matrix. Then

$$n^{1/2}(s^2 - \sigma^2) \xrightarrow{D} N(0,\ \mathrm{var}[u^2 - \sigma^2]).$$

- PROOF. Note that

$$\mathrm{var}(\frac{u'X}{n^{1/2}}) = \sigma^2\frac{X'X}{n},$$

which stays bounded by assumption, so that $(u'X/n^{1/2}) = O_p(1)$. Therefore the second term in s^2 is $O_p(n^{-1})$. Furthermore, $u'u/n$ converges in probability to σ^2 by the Law of Large Numbers. Therefore,

$$\begin{aligned} s^2 &= [1 + o_p(1)]\sigma^2 - \frac{1}{n-k}O_p(1) \\ &= \sigma^2 + o_p(1). \end{aligned}$$

- What about the asymptotic distribution of s^2?

$$\begin{aligned} & n^{1/2}(s^2 - \sigma^2) \\ = \ & [1 + o_p(1)]\frac{1}{n^{1/2}}\sum_{i=1}^{n}(u_i^2 - \sigma^2) - \frac{n^{1/2}}{n-k}O_p(1) \\ = \ & \frac{1}{n^{1/2}}\sum_{i=1}^{n}(u_i^2 - \sigma^2) + o_p(1) \\ & \xrightarrow{D} N(0,\ \mathrm{var}[u^2 - \sigma^2]), \end{aligned}$$

provided the second moment of $(u_i^2 - \sigma^2)$ exists, which it does under our assumption. When the errors are normally distributed,

$$\mathrm{var}[u^2 - \sigma^2] = 2\sigma^4.$$

- Now what about the t statistic:

$$\begin{aligned} t &= \frac{n^{1/2}c'\widehat{\beta}}{s(c'\frac{(X'X)^{-1}}{n}c)^{1/2}} \\ &= \frac{n^{1/2}c'\widehat{\beta}}{\sigma(c'M^{-1}c)^{1/2}} + o_p(1) \\ \xrightarrow{D} \frac{N(0,\sigma^2 c'M^{-1}c)}{\sigma(c'M^{-1}c)^{1/2}} &\equiv N(0,1) \text{ under } H_0. \end{aligned}$$

- As for the Wald statistic

$$W = n(R\widehat{\beta} - r)' \left[s^2 R \left(\frac{X'X}{n} \right)^{-1} R' \right]^{-1} (R\widehat{\beta} - r).$$

Theorem. Suppose that R is of full rank, that u_i are i.i.d. with finite fourth moment, and that the regressors are from a fixed design and satisfy

$$(X'X/n) \to M,$$

where M is a positive definite matrix. Then,

$$W \xrightarrow{D} N(0, \sigma^2 RM^{-1}R') \times [\sigma^2 RM^{-1}R']^{-1} \times N(0, \sigma^2 RM^{-1}R') = \chi^2_q.$$

5.8 The delta method

- Theorem. Suppose that

$$n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} N(0,\ \Sigma)$$

and that f is a continuously differentiable function. Then

$$n^{1/2}(f(\widehat{\theta}) - f(\theta)) \xrightarrow{D} N\left(0,\ \frac{\partial f}{\partial \theta}\Sigma\frac{\partial f}{\partial \theta'}\right).$$

- Proof (Scalar case). By the mean value theorem

$$f(\widehat{\theta}) = f(\theta) + (\widehat{\theta} - \theta)f'(\theta^*),$$

i.e.,

$$n^{1/2}(f(\widehat{\theta}) - f(\theta)) = f'(\theta^*) \cdot n^{1/2}(\widehat{\theta} - \theta).$$

- Furthermore,
$$\widehat{\theta} \xrightarrow{P} \theta \Rightarrow \theta^* \xrightarrow{P} \theta,$$
which implies that
$$f'(\theta^*) \xrightarrow{P} f'(\theta),$$
where
$$f'(\theta) \neq 0 < \infty.$$

- Therefore,
$$n^{1/2}(f(\widehat{\theta}) - f(\theta)) = [f'(\theta) + o_p(1)]n^{1/2}(\widehat{\theta} - \theta),$$
and the result now follows.

- EXAMPLE 1. $f(\beta) = e^{\beta}$, what is the distribution of $e^{\widehat{\beta}}$ (scalar)
$$\begin{aligned} n^{1/2}(e^{\widehat{\beta}} - e^{\beta}) &= e^{\beta} n^{1/2}(\widehat{\beta} - \beta) \\ &\xrightarrow{D} N(0,\ e^{2\beta}\sigma^2 M^{-1}). \end{aligned}$$

- EXAMPLE 2. Suppose that
$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$
What about $\widehat{\beta}_2/\widehat{\beta}_3$? We have
$$n^{1/2}\left(\frac{\widehat{\beta}_2}{\widehat{\beta}_3} - \frac{\beta_2}{\beta_3}\right) \xrightarrow{D} N\left(0,\ \sigma^2 \frac{\partial f}{\partial \beta} M^{-1} \frac{\partial f}{\partial \beta'}\right),$$
where
$$\frac{\partial f}{\partial \beta} = \begin{pmatrix} 0 \\ 1/\beta_3 \\ -\beta_2/\beta_3^2 \end{pmatrix},$$
so that the limiting variance is
$$\sigma^2 \left\{ \begin{pmatrix} \frac{1}{\beta_3} \\ \frac{-\beta_2}{\beta_3^2} \end{pmatrix} \begin{pmatrix} M^{22} & M^{23} \\ M^{32} & M^{33} \end{pmatrix} \begin{pmatrix} \frac{1}{\beta_3} \\ \frac{-\beta_2}{\beta_3^2} \end{pmatrix} \right\}.$$

Chapter 6

Errors in Variables

- Measurement error is a widespread problem in practice, since much economics data is poorly measured. This is an important problem that has been investigated a lot over the years.
- One interpretation of the linear model is that
  - there is some unobservable y^* satisfying
  $$y^* = X\beta$$
  - we observe y^* subject to error
  $$y = y^* + \varepsilon,$$
  where ε is a mean zero stochastic error term satisfying
  $$\varepsilon \perp y^*$$
  [or more fundamentally, $\varepsilon \perp X$].
- Combining these two equations
  $$y = X\beta + \varepsilon,$$
  where ε has the properties of the usual linear regression error term. It is clear that we treat X, y asymmetrically; X is assumed to have been measured perfectly.

- What about assuming instead that

$$y = X^*\beta + \varepsilon,$$

where

$$X = X^* + U.$$

We might assume that X is stochastic but X^* is fixed or that both are random. The usual strong assumption is that $U, \varepsilon \perp X^*$ in any case, and that U, ε are mutually independent. Clearly a variety of assumptions can be made here, and the results depend critically on what is assumed.

- Together these equations imply that

$$y = X\beta + \varepsilon - U\beta = X\beta + \nu,$$

where

$$\nu = \varepsilon - U\beta$$

is correlated with X because $X(U)$ and $\nu(U)$.

- In this case, the least squares estimator has an obvious bias. We have

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

Take expectations [note that X is now a random variable, although X^* may not be]

$$\begin{aligned} E(\widehat{\beta}) &= \beta - E\left\{(X'X)^{-1}X'E[U|X]\right\}\beta \\ &= \beta - E\left\{(X'X)^{-1}X'[X - X^*]\right\}\beta \end{aligned}$$

In general this is not equal to β, but it is difficult to calculate the bias exactly. Instead it is better to work with asymptotic approximation and to obtain an asymptotic bias.

- The denominator of $\widehat{\beta}$ satisfies

$$\frac{X'X}{n} = \frac{X^{*\prime}X^*}{n} + 2\frac{X^{*\prime}U}{n} + \frac{U'U}{n}.$$

- We shall suppose that for

$$\begin{aligned}
\frac{X^{*\prime}X^*}{n} &\xrightarrow{P} Q^* \\
\frac{X^{*\prime}U}{n} &\xrightarrow{P} 0 \\
\frac{U'\varepsilon}{n} &\xrightarrow{P} 0 \\
\frac{X^{*\prime}\varepsilon}{n} &\xrightarrow{P} 0 \\
\frac{U'U}{n} &\xrightarrow{P} \Sigma_{UU}
\end{aligned}$$

  which would be justified by the Law of Large Numbers under some assumptions on U, ε, X^*. Therefore,

$$\frac{X'X}{n} \xrightarrow{P} Q^* + \Sigma_{UU}.$$

- The numerator of $\widehat{\beta}$ satisfies

$$\frac{X'\varepsilon}{n} \xrightarrow{P} 0$$

$$\frac{X'U}{n} = \frac{X^{*\prime}U}{n} \xrightarrow{P} 0 + \frac{U'U}{n} \xrightarrow{P} \Sigma_{UU}$$

  by similar reasoning.

- Therefore,

$$\widehat{\beta} \xrightarrow{P} \beta - \left[Q^* + \Sigma_{UU}\right]^{-1} \Sigma_{UU}\beta = \left\{\left[Q^* + \Sigma_{UU}\right]^{-1} Q^*\right\} \cdot \beta \equiv C\beta.$$

- In the scalar case,

$$C = \frac{q}{q + \sigma_u^2} = \frac{1}{1 + \sigma_u^2/q},$$

  where σ_u^2/q is the noise to signal ratio;

  - when

$$\frac{\text{noise}}{\text{signal}} = 1,$$

    $\widehat{\beta}$ is unbiased.

- When
$$\frac{\text{noise}}{\text{signal}} \uparrow,$$
|bias| increases and β shrinks towards zero.

- In the vector case
$$|| \operatorname*{p\,lim}_{n\to\infty} \widehat{\beta}|| \leq \|\beta\|,$$
but it is not necessarily the case that each element is shrunk towards zero.

- Suppose that $K > 1$, but only one regressor is measured with error, i.e.,
$$\Sigma_{UU} = \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & 0 \end{bmatrix}.$$
In this case, all $\widehat{\beta}$ are biased; that particular coefficient estimate is shrunk towards zero.

- The downward bias result is specific to the strong assumptions case. For example, suppose that $(X_i^*, U_i, \varepsilon_i)$ are normally distributed but that U_i, ε_i are mutually correlated with covariance $\sigma_{u\varepsilon}$. Then
$$\widehat{\beta} \xrightarrow{P} \frac{\beta q}{q+\sigma_u^2} + \frac{\sigma_{u\varepsilon}}{q+\sigma_u^2},$$
and if $\sigma_{u\varepsilon}$ is large enough the bias can even be upward.

- If X^* is trending, then measurement error may produce no bias. For example, suppose that
$$x_t^* = t \text{ and } x_t = x_t^* + U_t.$$
Now
$$\begin{aligned} X^{*\prime}X^* &= \sum_{t=1}^{T} t^2 = O(T^3), \\ U'U &= \sum_{t=1}^{T} U_t^2 = O_p(T). \end{aligned}$$

Therefore,

$$\begin{aligned} & \frac{X'X}{T^3} \xrightarrow{P} \frac{X^{*\prime}X^*}{T^3} \\ \frac{X'U}{T^{3/2}} &= \frac{X^{*\prime}U}{T^{3/2}} + \frac{U'U}{T^{3/2}} \xrightarrow{P} 0. \end{aligned}$$

Therefore,

$$\widehat{\beta} \xrightarrow{P} \beta.$$

This is because the signal here is very strong and swamps the noise.

6.1 Solutions to EIV

- Assume knowledge of signal to noise ratio q/σ_u^2 and adjust $\widehat{\beta}$ appropriately. This is hard to justify nowadays because we rarely are willing to specify this information.

- Orthogonal regression.

- Instrumental variables. Let $Z_{n\times k}$ be instruments; that is, we have

$$\begin{aligned} \frac{Z'X}{n} &\xrightarrow{P} Q_{ZX} \\ \frac{Z'\nu}{n} &\xrightarrow{P} 0, \end{aligned}$$

  or equivalently $Z'\varepsilon/n \xrightarrow{P} 0$, $Z'U/n \xrightarrow{P} 0$.

- Then define the instrumental variables estimator (IVE)

$$\widehat{\beta}_{IV} = (Z'X)^{-1}Z'y.$$

- We have

$$\widehat{\beta}_{IV} = \beta + \left(\frac{Z'X}{n}\right)^{-1} \frac{Z'\nu}{n} \xrightarrow{P} \beta,$$

  using the above assumptions.

- Suppose that ν_i are i.i.d. with mean zero and variance σ_ν^2 and that in fact
$$\frac{Z'\nu}{n^{\frac{1}{2}}} \xrightarrow{D} N(0, \sigma_\nu^2 Q_{ZZ}).$$
Then, we can conclude that
$$n^{\frac{1}{2}}\left(\widehat{\beta}_{IV} - \beta\right) = \left(\frac{Z'X}{n}\right)^{-1} \frac{Z'\nu}{n^{\frac{1}{2}}}$$
$$\xrightarrow{D} N\left(0, \sigma_\nu^2 Q_{ZX}^{-1} Q_{ZZ} Q_{ZX}^{-1}\right).$$

- Where do the instruments come from?
  - Suppose that measurement errors affects cardinal outcome but not ordinality, i.e.,
  $$x_i < x_j \Leftrightarrow x_i^* < x_j^*.$$
  Then take as z_i the rank of x_i.
  - A slightly weaker restriction is to suppose that measurement error does not affect whether a variable is below or above the median, although it could affect other ranks. In this case,
  $$z_i = \begin{cases} 1 \text{ if } x_i > \text{ median } x_i \\ 0 \text{ if } x_i < \text{ median } x_i \end{cases}$$
  would be the natural instrument.
  - Method of grouping Wald (1940). The estimator is $\overline{y}_1/\overline{x}_1$.
  - Time series examples, z are lagged variables.
  - Specific examples. Month of birth.

6.2 Other Types of Measurement Error

- Discrete Covariates. Suppose that the covariate is discrete, then the above model of measurement is logically impossible. Suppose instead that
$$X_i = \begin{cases} X_i^* & \text{with prob } \pi \\ 1 - X_i^* & \text{with prob } 1 - \pi. \end{cases}$$
We can write this as $X_i = X_i^* + U_i$, but U_i is not independent of X_i^*.

- Magic Numbers. Suppose that there is rounding of numbers so that X_i^* is continuous, while X_i is the closest integer to X_i^*.

6.3 Durbin-Wu-Hausman Test

- We now consider a well known test for the presence of measurement error, called the Durbin-Wu-Hausman test. Actually, the test is applicable more generally.

- Suppose that our null hypothesis is

$$H_0: \text{ no measurement error.}$$

  This is equivalent to

$$\sigma_U^2 = 0,$$

  which may be a difficult test to contemplate by our existing methods.

- Instead consider the test statistic

$$H = \left(\widehat{\beta}_{OLS} - \widehat{\beta}_{IV}\right)' \widehat{V}^{-1} \left(\widehat{\beta}_{OLS} - \widehat{\beta}_{IV}\right),$$

  and reject the null hypothesis for large values of this statistic.

  - The idea is that $\widehat{\beta}_{OLS}$ and $\widehat{\beta}_{IV}$ are both consistent under H_0, but under H_A, $\widehat{\beta}_{OLS}$ is inconsistent. Therefore, there should be a discrepancy that can be picked up under the alternative.

- What is the null asymptotic variance? We have

$$\widehat{\beta}_{OLS} - \widehat{\beta}_{IV} = \left\{(Z'X)^{-1}Z' - (X'X)^{-1}X'\right\}\nu = A\nu$$

  with variance $V = \sigma_\nu^2 AA'$.

- In fact, AA' simplifies

$$\begin{aligned} AA' &= (Z'X)^{-1}Z'Z(Z'X)^{-1} \\ &\quad -(Z'X)^{-1}Z'X(X'X)^{-1} \\ &\quad -(X'X)^{-1}X'Z(Z'X)^{-1} \\ &\quad +(X'X)^{-1} \\ &= (Z'X)^{-1}Z'Z(Z'X)^{-1} - (X'X)^{-1} \\ &\geq 0, \end{aligned}$$

  where the inequality follows by the Gauss Markov Theorem.

- So we use

$$\widehat{V} = s_\nu^2 \left\{ (Z'X)^{-1} Z'Z (Z'X)^{-1} - (X'X)^{-1} \right\} = s_\nu^2 (Z'X)^{-1} Z' M_X Z (X'Z)^{-1},$$

where

$$s_\nu^2 = \frac{1}{n-k} \sum_{i=1}^{n} \widehat{\nu}_i^2,$$

where

$$\widehat{\nu}_i = y_i - X\widehat{\beta}_{IV}.$$

- Thus

$$\widehat{V}^{-1} = s_\nu^{-2} X' Z \left(Z' M_X Z \right)^{-1} Z' X.$$

- Under H_0,

$$H \xrightarrow{D} \chi_K^2,$$

and the rule is to reject for large values of H.

Chapter 7

Heteroskedasticity

- We made the assumption that

$$Var(y) = \sigma^2 I$$

in the context of the linear regression model. This contains two material parts:

  - off diagonals are zero (independence), and
  - diagonals are the same.

- Here we extend to the case where

$$Var(y) = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix} = \Sigma,$$

i.e., the data are heterogeneous.

- We look at the effects of this on estimation and testing inside linear (nonlinear) regression model, where $E(y) = X\beta$. In practice, many data are heterogeneous.

7.1 Effects of Heteroskedasticity

- Consider the OLS estimator

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

- In the new circumstances, this is unbiased, because

$$E(\widehat{\beta}) = \beta, \quad \forall \beta.$$

- However,

$$\begin{aligned} Var(\widehat{\beta}) &= (X'X)^{-1}X'\Sigma X(X'X)^{-1} \\ &= \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i x_i' \sigma_i^2 \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \\ &\neq \sigma^2 (X'X)^{-1}. \end{aligned}$$

- As sample size increases,

$$Var(\widehat{\beta}) \to 0,$$

  so that the OLSE is still consistent.

- The main problem then is with the variance.

  - Least squares standard errors are estimating the wrong quantity. We have

$$\begin{aligned} s^2 &= \frac{1}{n-k}\widehat{u}'\widehat{u} = \frac{1}{n}\sum_{i=1}^{n} u_i^2 + o_p(1) \\ &\xrightarrow{P} \overline{\sigma}^2 \equiv \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2, \end{aligned}$$

    but

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \sigma_i^2 - \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \nrightarrow 0$$

    in general.

- OLS is inefficient. Why?

$$y^* = \Sigma^{-1/2} y = \Sigma^{-1/2} X\beta + \Sigma^{-1/2} u = X^*\beta + u^*,$$

  where u^* are homogeneous. Therefore,

$$\widehat{\beta}^* = \left(X^{*\prime}X^*\right)^{-1} X^{*\prime} y^* = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1} y$$

is efficient by Gauss-Markov. So

$$\widehat{\beta}_{GLS} = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}y$$

is the efficient estimator here; this is not equal to

$$\widehat{\beta}_{OLS} = (X'X)^{-1}X'y,$$

unless

$$\Sigma = \sigma^2 I$$

(or some more complicated conditions are satisfied).

- Can show directly that

$$\left(X'\Sigma^{-1}X\right)^{-1} \leq (X'X)^{-1}X'\Sigma X(X'X)^{-1}.$$

In some special cases $OLS = GLS$, but in general they are different. What to do?

7.2 Plan A: Eicker-White

- Use OLS but correct standard errors. Accept inefficiency but have correct tests, etc.
- How do we do this? Can't estimate σ_i^2, $i = 1, ..., n$ because there are n of them. However, this is not necessary - instead we must estimate the sample average $\frac{1}{n}\sum_{i=1}^n x_i x_i' \sigma_i^2$. We estimate $Var(\widehat{\beta}) =$

$$V = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\sigma_i^2\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}$$

by

$$\widehat{V} = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\widehat{u}_i^2\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}.$$

Then under regularity conditions

$$\frac{1}{n}\sum_{i=1}^n x_i x_i'(\widehat{u}_i^2 - \sigma_i^2) \xrightarrow{P} 0,$$

which shows that

$$\widehat{V} - V \xrightarrow{P} 0.$$

- Typically find that White's standard errors [obtained from the diagonal elements of $\widehat{V}$] are larger than OLS standard errors.
- Finally, one can construct test statistics which are robust to heteroskedasticity, thus
$$n(R\widehat{\beta} - r)'[R\widehat{V}R']^{-1}(R\widehat{\beta} - r) \xrightarrow{D} \chi^2_J.$$

7.3 Plan B: Model Heteroskedasticity

- Sometimes models are suggested by data. Suppose original observations are by individual, but then aggregate up to a household level. Homogeneous at the individual level implies heterogeneous at the household level, i.e.,
$$u_i = \frac{1}{n_i}\sum_{j=1}^{n_i} u_{ij}.$$
Then,
$$Var(u_i) = \frac{1}{n_i}Var(u_{ij}) = \frac{\sigma^2}{n_i}.$$
Here, the variance is inversely proportional to household size. This is easy case since apart from single constant, σ^2, σ_i^2 is known.
- General strategy. Suppose that
$$\Sigma = \Sigma(\theta).$$
Further example $\sigma_i^2 = e^{\gamma}x_i$ or $\sigma_i^2 = \gamma x_i^2$ for some parameters.
- Suppose we have a normal error and that $\theta \cap \beta = \phi$. Then,
$$\begin{aligned} L(\beta,\theta) &= -\frac{1}{2}\ln|\Sigma(\theta)| - \frac{1}{2}(y - X\beta)'\Sigma(\theta)(y - X\beta) \\ &= -\frac{1}{2}\sum_{i=1}^{n}\ln\sigma_i^2(\theta) - \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\sigma_i^{-2}(\theta). \end{aligned}$$
In this case,
$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial\beta} &= \sum_{i=1}^{n} x_i\frac{(y_i - x_i'\beta)}{\sigma_i^2(\theta)} \\ \frac{\partial\mathcal{L}}{\partial\theta} &= -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial\ln\sigma_i^2}{\partial\theta}(\theta)\left[\frac{(y_i - x_i'\beta)^2}{\sigma_i^2(\theta)} - 1\right]. \end{aligned}$$

- The estimators $(\widehat{\beta}_{MLE}, \widehat{\theta}_{MLE})$ solve this pair of equations, which are nonlinear in general.

- Note that the equation for β is conditionally linear, that is suppose that we have a solution $\widehat{\theta}_{MLE}$, then

$$\widehat{\beta}_{MLE} = \left[\sum_{i=1}^{n} x_i x_i' \sigma_i^{-2}(\widehat{\theta}_{MLE}) \right]^{-1} \sum_{i=1}^{n} x_i y_i \sigma_i^{-2}(\widehat{\theta}_{MLE}).$$

  Iterate. Start with $\widehat{\beta}_{OLS}$, which is consistent, this gives us $\widehat{\theta}$, which we then use in the GLS definition. See below for a proper treatment of nonlinear estimators.

- Example. Suppose that

$$\sigma_i^2 = \frac{1}{\theta\,(x_i' x_i)}$$

  for some positive constant θ. In this case

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\theta} + \frac{1}{2} \sum_{i=1}^{n} u_i^2(\beta) x_i' x_i.$$

  Therefore, we have a closed form solution

$$\widehat{\theta} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i^2(\widehat{\beta}) x_i' x_i},$$

  where

$$\widehat{u}_i = y_i - x_i' \widehat{\beta}_{MLE}.$$

7.4 Properties of the Procedure

- Firstly, under general conditions not requiring y to be normally distributed,

$$\begin{bmatrix} n^{\frac{1}{2}} \left(\widehat{\beta}_{MLE} - \beta \right) \\ n^{\frac{1}{2}} \left(\widehat{\theta}_{MLE} - \theta \right) \end{bmatrix} \xrightarrow{D} N(0, \Omega)$$

  for some Ω.

- If y is normal, then $\Omega = I^{-1}$, the information matrix,

$$I = \begin{bmatrix} \lim_{n\to\infty} n^{-1} X'\Sigma^{-1}X & o \\ 0 & ? \end{bmatrix}.$$

  In this case, $\widehat{\beta}$ is asymptotically equivalent to

$$\widehat{\beta}_{GLS} = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}y.$$

  We say that $\widehat{\beta}_{ML}$ is asymptotically Gauss-Markov efficient, BLAUE.

- Often people use ad hoc estimates of θ and construct

$$\widehat{\beta}_{FGLS} = \left(X'\Sigma^{-1}(\widehat{\theta}_{AH})X\right)^{-1} X'\Sigma^{-1}(\widehat{\theta}_{AH})y.$$

  Provided $\widehat{\theta} \xrightarrow{P} \theta$ and some additional conditions, this procedure is also asymptotically equivalent to $\widehat{\beta}_{GLS}$.

7.5 Testing for Heteroskedasticity

- The likelihood framework has been widely employed to suggest tests of heteroskedasticity. Suppose that

$$\sigma_i^2(\theta) = \alpha e^{\gamma x_i}$$

$$H_0 : \gamma = 0 \quad \text{vs.} \quad \gamma \neq 0.$$

- The LM tests are simplest to implement here because we only have to estimate under homogeneous null. We have

$$\frac{\partial \mathcal{L}}{\partial \gamma} = -\frac{1}{2}\sum_{i=1}^{n} x_i \left(\frac{u_i^2}{\alpha} - 1\right).$$

  Under normality,

$$Var\left(\frac{u_i^2}{\alpha}\right) = 2.$$

  Therefore,

$$LM = \sum_{i=1}^{n}\left(\frac{\widehat{u}_i^2}{\widehat{\alpha}} - 1\right) x_i \left[2\sum_{i=1}^{n} x_i x_i'\right]^{-1} \sum_{i=1}^{n}\left(\frac{\widehat{u}_i^2}{\widehat{\alpha}} - 1\right) x_i,$$

where

$$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i^2,$$

where $\widehat{u}_i^2$ are the OLS residuals from the restricted regression.

- Under H_0

$$LM \xrightarrow{D} \chi_1^2.$$

Reject for large LM.

Chapter 8

Nonlinear Regression Models

- Suppose that
$$y_i = g(x_i, \beta) + \varepsilon_i, \quad i = 1, 2, ..., n,$$
where ε_i are i.i.d. mean zero with variance σ^2.

- In this case, how do we estimate β? The main criterion we shall consider is the Nonlinear least squares, which is of course the MLE when $y \sim N(g, \sigma^2 I)$. In this case one chooses β to minimize
$$S_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} [y_i - g(x_i, \beta)]^2$$
over some parameter set B. Let
$$\widehat{\beta}_{NLLS} = \arg\min_{\beta \in B} S_n(\beta).$$

- If B is compact and g is continuous, then the minimizer exists but is not necessarily unique. More generally, one cannot even guarantee existence of a solution.

- We usually try to solve a first order condition, which would be appropriate for finding interior minima in differentiable cases. In general, the first order conditions do not have a closed form solution. If there are multiple solutions to the first order condition, then one can end up with different answers depending on the way the algorithm is implemented. Statistical properties also an issue, $\widehat{\beta}_{NLLS}$ is a nonlinear function of y, so we cannot easily calculate mean and variance.

- If S_n is globally convex, then there exists a unique minimum for all n regardless of the parameter space. Linear regression has a globally convex criterion - it is a quadratic function. Some nonlinear models are also known to have this property.

8.1 Computation

- In one-dimension with a bounded parameter space B, the method of line search is effective. This involves dividing B into a grid of, perhaps equally spaced, points, computing the criterion at each point and then settling on the minimum. There can be further refinements - you further subdivide the grid around the minimum etc. Unfortunately, this method is not so useful in higher dimensions d because of the 'curse of dimensionality'. That is, the number of grid points required to achieve a given accuracy increases exponentially in d.

- 'Concentration' or 'Profiling' can sometimes help: some aspects of the problem may be linear, e.g.,

$$g(x, \theta) = \beta \frac{x^\lambda - 1}{\lambda}.$$

If λ were known, would estimate β by

$$\widehat{\beta} = [X(\lambda)'X(\lambda)]^{-1} X(\lambda)'y,$$

where

$$X(\lambda) = \begin{bmatrix} \frac{x_1^\lambda - 1}{\lambda} \\ \vdots \\ \frac{x_n^\lambda - 1}{\lambda} \end{bmatrix}.$$

Then write

$$S_n\left(\widehat{\beta}(\lambda), \lambda\right) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \widehat{\beta}(\lambda) \frac{x_i^\lambda - 1}{\lambda} \right]^2,$$

which is the concentrated criterion function. Now find $\widehat{\lambda}$ to min this, e.g., by line search on $[0, 1]$.

- Derivative based methods. We are trying to find a root of

$$\frac{\partial S_n}{\partial \beta}\left(\widehat{\beta}_{NNLS}\right) = 0.$$

- Can evaluate S_n, $\partial S_n/\partial\beta$, $\partial^2 S_n/\partial\beta\partial\beta'$, for any β. Suppose we take an initial guess β_1 and then modify it - which direction and how far?

  - If $\partial S_n(\beta_1)/\partial\beta > 0$, then we are to the right of the minimum, should move left.
  - We fit a line tangent to the curve $\partial S_n/\partial\beta$ at the point β_1 and find where that line intersects the zero.

- The tangent at β_1 is $\partial^2 S_n(\beta_1)/\partial\beta^2$ and the constant term is $\partial S_n(\beta_1)/\partial\beta - \partial^2 S_n(\beta_1)/\partial\beta^2\beta_1$.

- Therefore,

$$0 = \frac{\partial^2 S_n}{\partial\beta^2}(\beta_1)\beta_2 + \frac{\partial S_n}{\partial\beta}(\beta_1) - \frac{\partial^2 S_n}{\partial\beta}(\beta_1)\beta_1,$$

  which implies that

$$\beta_2 = \beta_1 - \left[\frac{\partial^2 S_n}{\partial\beta^2}(\beta_1)\right]^{-1}\frac{\partial S_n}{\partial\beta}(\beta_1).$$

  Repeat until convergence. This is Newton's method.

- In practice the following criteria are used

$$\left|\beta_{r+1} - \beta_r\right| < \tau \text{ or } \left|S_n(\beta_{r+1}) - S_n(\beta_r)\right| < \tau$$

  to stop the algorithm.

- In k-dimensions

$$\beta_2 = \beta_1 - \left[\frac{\partial^2 S_n}{\partial\beta\partial\beta'}(\beta_1)\right]^{-1}\frac{\partial S_n}{\partial\beta}(\beta_1).$$

- There are some modifications to this that sometimes work better. Outer produce (OPE) of the scores

$$\beta_2 = \beta_1 - \left[\sum_{i=1}^{n} \frac{\partial S_i}{\partial \beta}(\beta_1)\frac{\partial S_i}{\partial \beta'}(\beta_1)\right]^{-1} \sum_{i=1}^{n} \frac{\partial S_i}{\partial \beta}(\beta_1).$$

- Variable step length λ

$$\beta_2(\lambda) = \beta_1 - \lambda \left[\frac{\partial^2 S_n}{\partial \beta \partial \beta'}(\beta_1)\right]^{-1} \frac{\partial S_n}{\partial \beta}(\beta_1),$$

  and choose λ to max $S_n(\beta_2(\lambda))$.

- There are some issues with all the derivative-based methods:

  - If there are multiple local minima, need to try different starting values and check that always converge to same value.
  - When the criterion function is not globally convex one can have overshooting, and the process may not converge. The variable step length method can improve this.
  - If their criterion is flat near the minimum, then the algorithm may take a very long time to converge. The precise outcome depends on which convergence criterion is used. If you use the change in the criterion function then the chosen parameter value may actually be far from the true minimum.
  - If the minimum is at a boundary point, then the derivative-based methods will not converge.
  - In some problems, the analytic derivatives are difficult or time consuming to compute, and people substitute them by numerical derivatives, computed by an approximation. These can raise further problems of stability and accuracy.

8.2 Consistency of NLLS

- Theorem. Suppose that

  (1) The parameter space B is a compact subset of $\mathbb{R}^K$;

(2) $S_n(\beta)$ is continuous in β for all possible data;

(3) $S_n(\beta)$ converges in probability to a non-random function $S(\beta)$ uniformly in $\beta \in B$, i.e.,

$$\sup_{\beta \in B} |S_n(\beta) - S(\beta)| \xrightarrow{P} 0.$$

(4) The function $S(\beta)$ is uniquely minimized at $\beta = \beta_0$.

- Then

$$\widehat{\beta} \xrightarrow{P} \beta_0.$$

- Proof is in Amemiya (1986, Theorem 4.1.1). We just show why (3) and (4) are plausible. Substituting for y_i, we have

$$\begin{aligned} S_n(\beta) &= \frac{1}{n}\sum_{i=1}^{n} [\varepsilon_i + g(x_i, \beta_0) - g(x_i, \beta)]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^{n} [g(x_i, \beta) - g(x_i, \beta_0)]^2 + 2\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i [g(x_i, \beta) - g(x_i, \beta_0)]. \end{aligned}$$

- With i.i.d. data, by the Law of Large numbers

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \xrightarrow{P} \sigma^2$$

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i [g(x_i, \beta) - g(x_i, \beta_0)] \xrightarrow{P} 0 \text{ for all } \beta$$

and for all β

$$\frac{1}{n}\sum_{i=1}^{n} [g(x_i, \beta) - g(x_i, \beta_0)]^2 \xrightarrow{P} E\left([g(x_i, \beta) - g(x_i, \beta_0)]^2\right).$$

Therefore,

$$S_n(\beta) \xrightarrow{P} \sigma^2 + E\left([g(x_i, \beta) - g(x_i, \beta_0)]^2\right) \equiv S(\beta).$$

- Need convergence in probability to hold uniformly in a compact set containing β_0 (or over B), which requires a domination condition like

$$\sup_{\beta \in B} |S_n(\beta)| \leq Y \text{ with } E(Y) < \infty.$$

- Now

$$S(\beta_0) = \sigma^2 \text{ and } S(\beta) \geq \sigma^2 \text{ for all } \beta.$$

So, in the limit, β_0 minimizes S. Need $S(\beta)$ to be uniquely minimized at β_0 (identification condition).

- Example where (4) is satisfied is where g is linear, i.e., $g(x_i, \beta) = \beta' x_i$. Then

$$S(\beta) = \sigma^2 + (\beta - \beta_0)' E[x_i x_i'](\beta - \beta_0),$$

which is a quadratic function of β. (3) also holds in this case under mild conditions on x_i.

8.3 Asymptotic Distribution of NLLS

- Theorem. Suppose that:

1. $\widehat{\beta}$ is such that

$$\frac{\partial S_n(\widehat{\beta})}{\partial \beta} = 0$$

and satisfies

$$\widehat{\beta} \xrightarrow{P} \beta_0,$$

where β_0 is an interior point of B;

2. $\partial^2 S_n(\beta)/\partial\beta\partial\beta'$ exists and is continuous in an open convex neighbourhood of β_0;

3. $\partial^2 S_n(\beta)/\partial\beta\partial\beta'$ converges in probability to a finite nonsingular matrix $A(\beta)$ uniformly in β over any shrinking neighbourhood of β_0;

4. For some finite matrix B,

$$n^{\frac{1}{2}} \frac{\partial S_n(\beta_0)}{\partial \beta} \xrightarrow{D} N(0, B).$$

- Then,

$$n^{\frac{1}{2}}(\widehat{\beta} - \beta_0) \xrightarrow{D} N0, V),$$

where $V = A^{-1}BA^{-1}$ and $A = A(\beta_0)$.

- Proof. We have

$$0 = n^{\frac{1}{2}}\frac{\partial S_n}{\partial \beta}(\widehat{\beta}) = n^{\frac{1}{2}}\frac{\partial S_n}{\partial \beta}(\beta_0) + \frac{\partial^2 S_n}{\partial \beta \partial \beta'}(\beta^*)n^{\frac{1}{2}}(\widehat{\beta} - \beta_0),$$

where β^* lies between β_0 and $\widehat{\beta}$ by the multivariate mean value theorem. Applying assumptions (1)-(3) we get

$$n^{\frac{1}{2}}(\widehat{\beta} - \beta) = -A^{-1}n^{\frac{1}{2}}\frac{\partial S_n}{\partial \beta}(\beta_0) + o_p(1).$$

Finally, apply assumption (4) we get the desired result.

- We now investigate the sort of conditions needed to satisfy the assumptions of the theorem. In our case

$$\begin{aligned}\frac{\partial S_n}{\partial \beta}(\beta_0) &= -2\frac{1}{n}\sum_{i=1}^{n}\left[y_i - g(x_i, \beta_0)\right]\frac{\partial g}{\partial \beta}(x_i, \beta_0) \\ &= \frac{-2}{n}\sum_{i=1}^{n}\varepsilon_i \cdot \frac{\partial g}{\partial \beta}(x_i, \beta_0).\end{aligned}$$

- Suppose that (x_i, ε_i) are i.i.d. with

$$E(\varepsilon_i | \, x_i) = 0$$

with probability one. In this case, provided

$$E\left[\left\|\varepsilon_i^2 \frac{\partial g}{\partial \beta}(x_i, \beta_0)\frac{\partial g}{\partial \beta'}(x_i, \beta_0)\right\|\right] < \infty,$$

we can apply the standard central limit theorem to obtain

$$n^{\frac{1}{2}}\frac{\partial S_n}{\partial \beta}(\beta_0) \overset{D}{\to} N\left(0, 4E\left[\varepsilon_i^2 \frac{\partial g}{\partial \beta}(x_i, \beta_0)\frac{\partial g}{\partial \beta'}(x_i, \beta_0)\right]\right).$$

- What about (3)?

$$\frac{\partial^2 S_n}{\partial\beta\partial\beta'}(\beta) = -2\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 g}{\partial\beta\partial\beta'}(x_i,\beta)\left[y_i - g(x_i,\beta)\right] + 2\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial\beta}\frac{\partial g}{\partial\beta'}(x_i,\beta),$$

which in the special case $\beta = \beta_0$ is

$$\frac{\partial^2 S_n}{\partial\beta\partial\beta'}(\beta) = \frac{-2}{n}\sum_{i=1}^{n}\varepsilon_i\frac{\partial^2 g}{\partial\beta\partial\beta'}(x_i,\beta_0) + \frac{2}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial\beta}\frac{\partial g}{\partial\beta'}(x_i,\beta_0).$$

- Provided

$$E\left[\left\|\varepsilon_i\frac{\partial^2 g}{\partial\beta\partial\beta'}(x_i,\beta_0)\right\|\right] < \infty$$

and

$$E\left[\left\|\frac{\partial g}{\partial\beta}\frac{\partial g}{\partial\beta'}(x_i,\beta_0)\right\|\right] < \infty,$$

we can apply the law of large numbers to obtain

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{\partial^2 g}{\partial\beta\partial\beta'}(x_i,\beta_0) \xrightarrow{P} 0$$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial\beta}\frac{\partial g}{\partial\beta'}(x_i,\beta_0) \xrightarrow{P} E\left[\frac{\partial g}{\partial\beta}\frac{\partial g}{\partial\beta'}(x_i,\beta_0)\right].$$

- These conditions need to be strengthened a little to obtain uniformity over the neighbourhood of β_0. For example, suppose that we have additional smoothness and

$$\frac{\partial^2 g}{\partial\beta^2}(x_i,\beta^*) = \frac{\partial^2 g}{\partial\beta^2}(x_i,\beta_0) + (\beta^* - \beta_0)\frac{\partial^3 g}{\partial\beta^3}(x_i,\beta^{**})$$

for some intermediate point β^{**}. Then, provided

$$\sup_{\beta\in B}\left\|\frac{\partial^3 g}{\partial\beta^3}(x,\beta^{**})\right\| \le D(x)$$

for some function D for which

$$ED(X) < \infty,$$

condition (2) will be satisfied.

- Similar results can be shown in the fixed design case, but we need to use the CLT and LLN for weighted sums of i.i.d. random variables.
- Note that when ε_i are i.i.d. and independent of x_i, we have

$$A = 4E\left[\frac{\partial g}{\partial \beta}\frac{\partial g}{\partial \beta'}(x_i, \beta_0)\right] = \frac{B}{\sigma^2}$$

and the asymptotic distribution is

$$n^{\frac{1}{2}}(\widehat{\beta} - \beta_0) \xrightarrow{D} N\left(0, \sigma^2 A^{-1}\right).$$

- Standard errors. Let

$$\widehat{A} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial \beta}\frac{\partial g}{\partial \beta'}(x_i, \widehat{\beta})$$

$$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial \beta}\frac{\partial g}{\partial \beta'}(x_i, \widehat{\beta})\widehat{\varepsilon}_i^2,$$

where $\widehat{\varepsilon}_i = y_i - g(x_i, \widehat{\beta})$.Then

$$\widehat{V} \xrightarrow{P} V.$$

8.4 Likelihood and Efficiency

- These results generalize to the likelihood framework for i.i.d. data

$$\ell(\text{data}, \theta) = \sum_{i=1}^{n}\ell_i(\theta).$$

Let $\widehat{\theta}$ maximize $\ell(\text{data}, \theta)$.

- Then under regularity conditions

$$\widehat{\theta} \xrightarrow{P} \theta_0$$

and

$$n^{\frac{1}{2}}(\widehat{\theta} - \theta_0) \xrightarrow{D} N\left(0, \mathcal{I}^{-1}(\theta_0)\right),$$

where the information matrix

$$\mathcal{I}(\theta_0) = E\left[\frac{\partial \ell_i}{\partial \theta}\frac{\partial \ell_i}{\partial \theta'}(\theta_0)\right] = -E\left[\frac{\partial^2 \ell_i}{\partial \theta \partial \theta'}\partial \ell_i(\theta_0)\right].$$

This last equation is called the information matrix equality.

- Asymptotic Cramér-Rao Theorem. The MLE is asymptotically "efficient" amongst the class of all asymptotically normal estimates (stronger than Gauss-Markov).

Chapter 9

Generalized Method of Moments

- We suppose that there is i.i.d. data $\{Z_i\}_{i=1}^n$ from some population.
- It is known that there exists a unique θ_0 such that

$$E\left[g(\theta_0, Z_i)\right] = 0$$

for some $q \times 1$ vector of known functions $g(\theta_0, \cdot)$.

  - For example, g could be the first order condition from OLS or more generally maximum likelihood, e.g., $g(\beta, Z_i) = x_i\left(y_i - x_i'\beta\right)$.
  - Conditional moment specification. Suppose in fact we know for some given function ρ that

$$E[\rho(\theta_0, Z_i)|X_i] = 0,$$

    where X_i can be a subset of Z_i.Then this implies the unconditional moment given above when you take

$$g(\theta_0, Z_i) = \rho(\theta_0, Z_i) \otimes h(X_i)$$

    for any function h of the 'instruments' X_i. This sort of specification arises a lot in economic models, which is what really motivates this approach.
  - The functions g can be nonlinear in θ and Z.

– The distribution of Z_i is unspecified apart from the q moments.

- For any θ, let
$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(\theta, Z_i).$$

- There are several cases:

  - $p > q$ unidentified case
  - $p = q$ exactly identified case
  - $p < q$ overidentified case.

- In the exactly identified case, we define our estimator as any solution to the equations
$$G_n(\widehat{\theta}) = 0.$$
Since we have p equations in p-unknowns, we can expect a solution to exist under some regularity conditions. However, the equations are nonlinear and have to be solved by numerical methods.

- When $p < q$, we cannot simultaneously solve all equations, and the most we can hope to do is to make them close to zero.

- Let
$$Q_n(\theta) = G_n(\theta)' W_n G_n(\theta),$$
where W_n is a $q \times q$ positive definite weighting matrix. For example, $W_n = Q_{q\times q}$. Then let
$$\widehat{\theta}_{GMM} \text{ minimize } Q_n(\theta)$$
$$\text{over } \theta \in \Theta \subseteq \mathbb{R}^p.$$

- This defines a large class of estimators, one for each weighting matrix W_n.

- It is generally a nonlinear optimization problem like nonlinear least squares; various techniques are available for finding the minimizer.

- GMM is a general estimation method that includes both OLS and more general MLE as special cases!!

- Thus consider the sample log likelihood

$$\sum_{i=1}^{n} \ell(Z_i, \theta),$$

where $\exp(\ell)$ is the density function of Z_i. The MLE maximises the log likelihood function or equivalently finds the parameter value that solves the score equations:

$$\sum_{i=1}^{n} \frac{\partial \ell}{\partial \theta}(Z_i, \theta) = 0.$$

- This is exactly identified GMM with

$$g(\theta, Z_i) = \frac{\partial \ell}{\partial \theta}(Z_i, \theta).$$

- What is different is really the model specification part, that is the specification of models through conditional moment restrictions.

9.1 Asymptotic Properties in the iid case

- We now turn to the asymptotic properties. Under some regularity conditions we have

$$\widehat{\theta}_{GMM} \xrightarrow{P} \theta_0.$$

Namely, we need that the criterion function converges uniformly to a function that is uniquely minimized by θ_0.

- Under further regularity conditions, we can establish

$$n^{\frac{1}{2}} \left(\widehat{\theta}_{GMM} - \theta \right) \xrightarrow{D} N\left(0, (\Gamma' W \Gamma)^{-1} \Gamma' W \Omega W \Gamma (\Gamma' W \Gamma)^{-1}\right),$$

where:

$$\begin{aligned} \Omega(\theta_0) &= \text{Var } n^{\frac{1}{2}} G_n(\theta_0), \\ \Gamma &= p\lim_{n\to\infty} \frac{\partial G_n(\theta_0)}{\partial \theta}. \\ 0 &< W = p\lim_{n\to\infty} W_n. \end{aligned}$$

- Special case of exactly identified case: weights are irrelevant and

$$n^{\frac{1}{2}}(\widehat{\theta} - \theta_0) \xrightarrow{D} N\left(0, \Gamma^{-1}\Omega\Gamma^{-1\prime}\right).$$

- What is the optimal choice of W in the overidentified case?
  - In fact W_n should be an estimate of Ω^{-1}.
  - In the iid case we take

$$\widetilde{\Omega} = \Omega(\widetilde{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(\widetilde{\theta}, Z_i)g(\widetilde{\theta}, Z_i)',$$

    where $\widetilde{\theta}$ is a preliminary estimate of θ_0 obtained using some arbitrary weighting matrix, e.g., I_q.

- In sum, then, the full procedure is
  - $\widetilde{\theta} = \arg\min G_n(\theta)' G_n(\theta)$
  - $\widehat{\theta} = \widehat{\theta}^{opt}_{GMM} = \arg\min G_n(\theta)' \widetilde{\Omega}^{-1} G_n(\theta)$.

- The asymptotic distribution is now normal with mean zero and variance

$$n^{\frac{1}{2}}(\widehat{\theta} - \theta_0) \xrightarrow{D} N\left(0, (\Gamma'\Omega^{-1}\Gamma)^{-1}\right).$$

- This estimator is efficient in the sense that it has minimum asymptotic variance among all GMM estimators.

- Can estimate the asymptotic variance of $\widehat{\theta}$ by

$$\widehat{V} = \left[\widehat{\Gamma}'\widehat{\Omega}^{-1}\widehat{\Gamma}\right]^{-1},$$

  where

$$\widehat{\Omega} = \Omega(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(\widehat{\theta}, Z_i)g(\widehat{\theta}, Z_i)',$$

  and

$$\widehat{\Gamma} = \frac{\partial G_n(\widehat{\theta})}{\partial \theta},$$

  are consistent estimates of Γ and Ω.

9.2 Test Statistics

- t-test. Consider the null hypothesis

$$c'\theta = \gamma$$

for some vector c and scalar γ. Then

$$\frac{n^{\frac{1}{2}}[c'\widehat{\theta} - \gamma]}{(c'\widehat{V}c)^{\frac{1}{2}}} \xrightarrow{D} N(0,1)$$

under the null hypothesis. Can do one-sided and two-sided tests.

- Consider the null hypothesis

$$R\theta = r,$$

where r is of dimensions m. Then

$$n(R\widehat{\theta} - r)'[R\widehat{V}R']^{-1}(R\widehat{\theta} - r) \xrightarrow{D} \chi^2_m.$$

- Reject for large values.
- Nonlinear restrictions. Suppose that

$$n^{\frac{1}{2}}(\widehat{\theta} - \theta) \xrightarrow{D} N(0, V(\theta))$$

for some variance V.

- By a Taylor series expansion

$$f(\widehat{\theta}) \simeq f(\theta) + F(\theta)(\widehat{\theta} - \theta),$$

where

$$F(\theta) = \frac{\partial f(\theta)}{\partial \theta'}.$$

- Therefore,

$$n^{\frac{1}{2}}\left(f(\widehat{\theta}) - f(\theta)\right) \xrightarrow{D} N\left(0, F(\theta)V(\theta)F(\theta)'\right).$$

- This is called the delta method. If f is linear, then this is obvious.

- Application to hypothesis testing. Consider the null hypothesis

$$f(\theta) = 0$$

for some m vector nonlinear function f.

- Let

$$\widehat{f} = f(\widehat{\theta}) \text{ and } \widehat{F} = F(\widehat{\theta}).$$

Then

$$n\widehat{f}' \left[\widehat{F}\widehat{V}\widehat{F}'\right]^{-1} \widehat{f} \xrightarrow{D} \chi^2_m$$

under H_0.

9.3 Examples

- Linear regression

$$y = X\beta + u,$$

with some error vector u.

- Suppose also that it is known that for some unique β_0 we have

$$E\left[x_i u_i(\beta_0)\right] = 0.$$

There are K conditions and K parameters and this is an exactly identified case.

  - In this case, there exists a unique $\widehat{\beta}$, the OLS estimator in fact, that satisfies the empirical conditions

$$\frac{1}{n}X'(y - X\widehat{\beta}) = 0.$$

- Suppose now

$$E\left[x_i u_i(\beta_0)\right] \neq 0,$$

i.e., the errors are correlated with the regressors. This could be because

  - omitted variables. There are variables in u that should be in X.
  - The included X variables have been measured with error

(a) Simultaneous Equations. Demand and supply

$$Q^S = S(P;\ w,\ r,\ t)$$

$$Q^D = D(P;\ P^*,\ y)$$

In equilibrium $Q^S = Q^D$ determines Q, P given w, r, t, P^*, and y. The econometric model

$$\ln Q = \alpha + \beta \ln P + \delta w + \rho r + \tau t + e\ ,\quad \text{supply}$$

$$\ln Q = \alpha' + \beta' \ln P + \xi P^* + \eta y + u\ ,\quad \text{demand}$$

Parameters of interest β, β' price elasticities, ξ cross-price, η income. This is a simultaneous system. P, Q endogenous variables. w, r, t, P^* and y exogenous variables. Because P and Q simultaneously determined, expect

$$\mathrm{Cov}(P, e) \neq 0 \neq \mathrm{Cov}(Q, u)$$

$$Q(P),\ P(u) \Rightarrow Q(u)$$
$$P(Q),\ Q(e) \Rightarrow P(e)$$

Simultaneity means we can't usually use OLS to estimate parameters.

- Suppose however that there exists some instruments z_i such that

$$E\left[z_i u_i(\beta_0)\right] = 0 \tag{9.1}$$

for some instruments $z_i \in \mathbb{R}^J$.

- Suppose that there are many instruments, i.e., $J > K$. In this case, we can't solve uniquely for $\widehat{\beta}_{IV}$ because there are too many equations which can't all be satisfied simultaneously.

- Now take

$$\begin{aligned} G_n(\beta) &= \frac{1}{n} Z'(y - X\beta) \\ &= \frac{1}{n} \sum_{i=1}^{n} z_i (y_i - x_i'\beta). \end{aligned}$$

A GMM estimator can be defined as any minimizer of

$$Q_n(\beta) = (y - X\beta)' \, ZW_nZ'(y - X\beta)$$

for some $J \times J$ weighting matrix W_n. What is the estimator?

- We shall suppose that W_n is a symmetric matrix and define the real symmetric matrix

$$A = ZW_nZ'$$

and its square root $A^{\frac{1}{2}}$. Letting

$$y^* = A^{\frac{1}{2}}y \text{ and } X^* = A^{\frac{1}{2}}X$$

we see that

$$Q_n(\beta) = (y^* - X^*\beta)'(y^* - X^*\beta)$$

with solution

$$\begin{aligned} \widehat{\beta}_{GMM} &= (X^{*\prime}X^*)^{-1} X^{*\prime}y^* \\ &= (X'AX)^{-1} X'Ay \\ &= (X'ZW_nZ'X)^{-1} X'ZW_nZ'y. \end{aligned}$$

- The question is, what is the best choice of W_n? Suppose also that u has variance matrix $\sigma^2 I$ independent of Z, and that Z is a fixed variable. Then

$$var\left[n^{\frac{1}{2}}G_n(\beta_0)\right] = var\frac{1}{n^{\frac{1}{2}}}Z'u = \sigma^2\frac{Z'Z}{n}.$$

Therefore, the optimal weighting is to take

$$W_n \propto (Z'Z)^{-1}$$

in which case

$$\widehat{\beta}_{GMM} = (X'P_ZX)^{-1} X'P_Zy,$$

where

$$P_Z = Z(Z'Z)^{-1}Z'$$

i.e., it is the two-stage least squares estimator.

- Suppose instead that u_i is heteroskedastic, then the optimal weighting is by

$$W_n = \frac{1}{n}\sum_{i=1}^{n} z_i z_i' \widehat{u}_i^2.$$

9.4 Time Series Case

- We next suppose that the data is stationary and mixing.
- CONDITIONAL MOMENT RESTRICTIONS. We suppose that for some $m \times 1$ vector of known functions ρ, with probability one

$$E\left[\rho(\theta_0, Y_t)\,|\mathcal{F}_t\right] = 0$$

where $\theta_0 \in \mathbb{R}^p$ is the true parameter value and $\mathcal{F}_t$ is some information set containing perhaps contemporaneous regressors and lagged variables. Many economic models fall into this framework for example Euler equations. In finance applications ρ could be some excess return, and the efficient markets hypothesis guarantees that this is unforecastable given certain sorts of information.

- Examples.
  - Static time series regression

$$y_t = \beta' x_t + \varepsilon_t, \text{ where } E(\varepsilon_t|x_t) = 0.$$

    In this case, the error term ε_t can be serially correlated.
  - Time series regression

$$y_t = \gamma y_{t-1} + \varepsilon_t, \text{ where } E(\varepsilon_t|y_{t-1}, y_{t-2}\ldots) = 0.$$

    In this case, the error term is serially uncorrelated.
  - Same model but instead suppose only that

$$E(\varepsilon_t|y_{t-1}) = 0.$$

    This is strictly weaker than the earlier assumption.
  - Same model but instead suppose that

$$E(\varepsilon_t|x_t, y_{t-1}) = 0.$$

    This is strictly weaker than the earlier assumption.

- Estimation now proceeds by forming some UNCONDITIONAL MOMENT RESTRICTIONS using valid instruments, i.e., variables from $\mathcal{F}_t^* \subset \mathcal{F}_t$. Thus, let

$$g(\theta, Z_t) = \rho(\theta_0, Y_t) \otimes X_t,$$

where $X_t \in \mathcal{F}_t$ and $Z_t = (Y_t, X_t)$. We suppose that g is of dimensions q with $q \geq p$. Then

$$E[g(\theta, Z_t)] = 0 \Longleftrightarrow \theta = \theta_0.$$

We then form the sample moment condition

$$G_T(\theta) = \frac{1}{T} \sum_{i=1}^{T} g(\theta, Z_t).$$

- If $q = p$, the estimator solves $G_T(\theta) = 0$. If $q > p$, let

$$Q_T(\theta) = G_T(\theta)' W_T G_T(\theta),$$

where W_T is a $q \times q$ positive definite weighting matrix. For example, $W_T = I_{q \times q}$. Then let

$$\begin{array}{ccc} \widehat{\theta}_{GMM} & \text{minimize} & Q_T(\theta) \\ & \text{over} & \theta \in \Theta \subseteq \mathbb{R}^p. \end{array}$$

- In the regression case $E(\varepsilon_t | x_t) = 0$ means that

$$E(\varepsilon_t \cdot h(x_t)) = 0$$

for any measurable function h. Therefore, take

$$g(\theta, Z_t) = h(x_t) \cdot (y_t - \beta' x_t)$$

In the autoregression case $E(\varepsilon_t | y_{t-1}, \ldots) = 0$ means that

$$E(\varepsilon_t \cdot h(y_{t-1}, \ldots)) = 0$$

for any measurable function h. Therefore, take

$$g(\theta, Z_t) = h(y_{t-1}, \ldots) \cdot (y_t - \gamma y_{t-1}).$$

In this case there are many functions that work.

9.5 Asymptotics

- As before we have

$$T^{\frac{1}{2}}\left(\widehat{\theta}_{GMM}-\theta\right)\xrightarrow{D}N\left(0,(\Gamma'W\Gamma)^{-1}\Gamma'W\Omega W\Gamma(\Gamma'W\Gamma)^{-1}\right),$$

  where:

$$\begin{aligned}\Omega(\theta_0) &= \text{Var } n^{\frac{1}{2}}G_n(\theta_0),\\ \Gamma &= p\lim_{n\to\infty}\frac{\partial G_n(\theta_0)}{\partial\theta}.\\ 0 &< W = p\lim_{n\to\infty}W_n.\end{aligned}$$

  Now, however

$$\begin{aligned}\Omega(\theta_0) &= \lim_{T\to\infty} var T^{\frac{1}{2}}G_T(\theta_0)\\ &= \lim_{T\to\infty} var\left[\frac{1}{T^{\frac{1}{2}}}\sum_{t=1}^{T}g(\theta_0,Z_t)\right]\\ &= \lim_{T\to\infty} E\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}g(\theta_0,Z_t)g(\theta_0,Z_s)'\right].\end{aligned}$$

- In the special case where $g(\theta,Z_t)$ is a martingale with respect to past information, i.e., $E[g(\theta,Z_t)|\mathcal{F}_{t-1}\}=0$, where $Z_t\in\mathcal{F}_t$, then

$$\Omega(\theta_0)=\lim_{T\to\infty} E\left[\frac{1}{T}\sum_{t=1}^{T}g(\theta_0,Z_t)g(\theta_0,Z_t)'\right].$$

- In general though, you have to take account of the covariance terms. If the vector time series $U_t=g(\theta_0,Z_t)$ is stationary, then

$$\Omega(\theta_0)=\gamma_0+\sum_{k=1}^{\infty}\left(\gamma_k+\gamma_k'\right),$$

where

$$\gamma_k = E\left[g(\theta_0, Z_t)g(\theta_0, Z_{t-k})'\right]$$

$$\gamma_k' = E\left[g(\theta_0, Z_{t-k})g(\theta_0, Z_t)'\right]$$

is the covariance function of U_t.

- For standard errors and optimal estimation we need an estimator of Ω. The Newey-West estimator

$$\widehat{\Omega}_T = \sum_{t,s:|t-s|\leq n(T)} \sum w\left(|t-s|\right) g\left(\widetilde{\theta}, Z_t\right) g\left(\widetilde{\theta}, Z_s\right)',$$

  where

$$w(j) = 1 - \frac{j}{n+1},$$

  and where $\widetilde{\theta}$ is a preliminary estimate of θ_0 obtained using some arbitrary weighting matrix, e.g., I_q. This ensures a positive definite covariance matrix estimate. Provided $n = n(T) \to \infty$ but $n(T)/T \to 0$ at some rate

$$\widehat{\Omega}_T \xrightarrow{P} \Omega.$$

- This is used to construct standard errors.

- The optimal choice of W should be an estimate of Ω^{-1}. We take $W_T = \widehat{\Omega}_T^{-1}$.

9.6 Example

- Hansen and Singleton, *Econometrica* (1982). One of the most influential econometric papers of the 1980s. Intertemporal consumption/Investment decision:

  - c_t consumption
  - $u(\cdot)$ utility $u_c > 0$, $u_{cc} < 0$.
  - $1 + r_{i,t+1}$, $i = 1, \ldots, m$ is gross return on asset i at time $t+1$.

- The representative agent solves the following optimization problem

$$\max_{\{c_t, w_t\}_{t=0}^{\infty}} \sum_{\tau=0}^{\infty} \beta^{\tau} E\left[u(c_{t+\tau}) \,|I_t\right],$$

where

- w_t is a vector of portfolio weights.
- β is the discount rate with $0 < \beta < 1$.
- I_t is the information available to the agent at time t.

- We assume that there is a unique interior solution; this is characterized by the following condition

$$u'(c_t) = \beta E\left[(1 + r_{i,t+1}) u'(c_{t+1}) \,|I_t\right],$$

for $i = 1, \ldots, m$.

- Now suppose that

$$u(c_t) = \begin{cases} \frac{c_t^{1-\gamma}}{1-\gamma} & \text{if } \gamma > 0, \gamma \neq 1, \\ \log c_t & \gamma = 1. \end{cases}$$

Here, γ is the coefficient of relative risk aversion.

- In this case, the first order condition is

$$c_t^{-\gamma} = \beta E\left[(1 + r_{i,t+1}) c_{t+1}^{-\gamma} \,|I_t\right]$$

for $i = 1, \ldots, m$.

- This implies that

$$E\left[1 - \beta\left\{(1 + r_{i,t+1})\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right\} |I_t^*\right] = 0$$

for $i = 1, \ldots, m$, where

$$I_t^* \subset I_t$$

and I_t^* is the econometrician's information set.

- We want to estimate the parameters $\theta_{p\times 1} = (\beta, \gamma)$ and test whether the theory is valid given a dataset consisting of

$$\{c_t, r_{i,t+1}, I_t^*\}_{t=1}^T .$$

- Define the $q \times 1$ vector

$$g(\theta, x_t) = \begin{bmatrix} \vdots \\ \left[1 - \beta\left\{(1 + r_{i,t+1})\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right\}\right] z_{jt} \\ \vdots \end{bmatrix},$$

where

$$z_t \in I_t^* \subset \mathbb{R}^J$$

are 'instruments', $q = mJ$, and

$$x_t = (z_t, c_t, c_{t+1}, r_{1,t+1}, \ldots, r_{m,t+1})' .$$

- Typically, z_t is chosen to be lagged variables and are numerous, so that $q \geq p$.
- The model assumption is that

$$E\left[g(\theta_0, x_t)\right] = 0$$

for some unique θ_0.

- This is a nonlinear function of γ.
- Exercise. Show how to consistently estimate Γ and Ω in this case.

Chapter 10

Time Series

10.1 Some Fundamental Properties

- We start with univariate time series $\{y_t\}_{t=1}^{T}$. There are two main features:

  - stationarity/nonstationarity
  - dependence

- We first define stationarity.

- Strong Stationarity. The stochastic process y is said to be strongly stationary if the vectors

$$(y_t, \ldots, y_{t+r})$$

  and

$$(y_{t+s}, \ldots, y_{t+s+r})$$

  have the same distribution for all t, s, r.

- Weak Stationarity. The stochastic process y is said to be weakly stationary if the vectors

$$(y_t, \ldots, y_{t+r})$$

  and

$$(y_{t+s}, \ldots, y_{t+s+r})$$

  have the same mean and variance for all t, s, r.

- Most of what we know is restricted to stationary series, but in the last 20 years there have been major advances in the theory of nonstationary time series, see below. In Gaussian [i.e., linear] time series processes, strong and weak stationarity coincide.

- Dependence. One measure of dependence is given by the covariogram [or correlogram]
$$cov(y_t, y_{t-s}) = \gamma_s; \quad \rho_s = \frac{\gamma_s}{\gamma_0}.$$

- Note that stationarity was used here in order to assert that these moments only depend on the gap s and not on calendar time t as well.

- For i.i.d. series,
$$\gamma_s = 0 \text{ for all } s \neq 0,$$
while for positively (negative) dependent series $\gamma_s > (<)0$. Economics series data often appear to come from positively dependent series.

- Mixing. (Covariance) If $\gamma_s \to 0$ as $s \to \infty$.

- This just says that the dependence [as measured by the covariance] on the past shrinks with horizon. This is an important property that is possessed by many models.

- ARMA Models. The following is a very general class of models called $ARMA(p, q)$:
$$\begin{aligned} y_t &= \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} \\ &\quad + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}, \end{aligned}$$
where ε_t is i.i.d., mean zero and variance σ^2.

  - We shall for convenience usually assume that $\mu = 0$.
  - We also assume for convenience that this model holds for $t = 0, \pm 1, \ldots$.

- It is convenient to write this model using lag polynomial notation. We define the lag operator
$$Ly_t = y_{t-1}$$

so that we can now define

$$A(L)y_t = B(L)\varepsilon_t,$$

where the lag polynomials

$$\begin{aligned} A(L) &= 1 - \phi_1 L - \ldots - \phi_p L^p \\ B(L) &= 1 - \theta_1 L - \ldots - \theta_q L^q. \end{aligned}$$

The reason for this is to save space and to emphasize the mathematical connection with the theory of polynomials.

- Special case $AR(1)$. Suppose that

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

Here,

$$A(L) = 1 - \phi L.$$

  - We assume $|\phi| < 1$, which is necessary and sufficient for y_t to be a stationary process.
  - Now write

$$y_{t-1} = \phi y_{t-2} + \varepsilon_{t-1}.$$

    Continuing we obtain

$$\begin{aligned} y_t &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2 y_{t-2} \\ &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \ldots \\ &= \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}, \end{aligned}$$

    which is called the $MA(\infty)$ representation of the time series;
  - this shows that y_t depends on all the past shocks.

- Now we calculate the moments of y_t using the stationarity property. We have

$$E(y_t) = \phi E(y_{t-1}),$$

which can be phrased as

$$\mu = \phi\mu \Leftrightarrow \mu = 0,$$

where

$$\mu = E(y_t) = E(y_{t-1}).$$

- Furthermore,
$$var(y_t) = \phi^2 var(y_{t-1}) + \sigma^2,$$
which implies that
$$\gamma_0 = \frac{\sigma^2}{1-\phi^2},$$
where
$$\gamma_0 = var(y_t) = var(y_{t-1}).$$
This last calculation of course requires that $|\phi| < 1$, which we are assuming for stationarity.

- Finally,
$$cov(y_t, y_{t-1}) = E(y_t y_{t-1}) = \phi E(y_{t-1}^2) + 0,$$
which implies that
$$\gamma_1 = \phi \frac{\sigma^2}{1-\phi^2},$$
while
$$cov(y_t, y_{t-2}) = E(y_t y_{t-2}) = \phi E(y_{t-1} y_{t-2}) = \phi^2 \frac{\sigma^2}{1-\phi^2}.$$

- In general
$$\gamma_s = \sigma^2 \frac{\phi^s}{1-\phi^2}; \quad \rho_s = \phi^s.$$
The correlation function decays geometrically towards zero.

- Exercise calculate correlogram for $AR(2)$.

- Moving Average $MA(1)$. Suppose that
$$y_t = \varepsilon_t - \theta \varepsilon_{t-1},$$
where as before ε_t are i.i.d. mean zero with variance σ^2.

  - In this case,
$$E(y_t) = 0,$$
and
$$var(y_t) = \sigma^2(1+\theta^2).$$

- Furthermore,

$$\begin{aligned} cov(y_t, y_{t-1}) &= E\left\{(\varepsilon_t - \theta\varepsilon_{t-1})\ (\varepsilon_{t-1} - \theta\varepsilon_{t-2})\right\} \\ &= -\theta E(\varepsilon_{t-1}^2) \\ &= -\theta\sigma^2. \end{aligned}$$

  Therefore,

$$\begin{aligned} \rho_1 &= \frac{-\theta}{1+\theta^2}, \\ \rho_j &= 0, \quad j = 2, \ldots \end{aligned}$$

- This is a 1-dependent series. $MA(q)$ is a q-dependent series.
- Note that the process is automatically stationary for any value of θ.
- If $|\theta| < 1$, we say that the process is invertible and we can write

$$\sum_{j=0}^{\infty} \theta^j y_{t-j} = \varepsilon_t.$$

- In general $ARMA(p, q)$, we can write

$$A(L)y_t = B(L)\varepsilon_t.$$

  - The stationarity condition for an $ARMA(p, q)$ process is just that the roots of the autoregressive polynomial

$$1 - \phi_1 z - \ldots - \phi_p z^p$$

    to be outside unit circle.
  - Likewise the condition for invertibility is that the roots of the moving average polynomial

$$1 - \theta_1 z - \ldots - \theta_q z^q$$

    lie outside the unit circle.

- Assuming these conditions are satisfied we can write this process in two different ways:

$$\frac{A(L)}{B(L)}y_t = \sum_{j=0}^{\infty} \gamma_j y_{t-j} = \varepsilon_t.$$

This is called the $AR(\infty)$ representation, and expresses y in terms of its own past. Or

$$y_t = \frac{B(L)}{A(L)}\varepsilon_t = \sum_{j=0}^{\infty} \delta_j \varepsilon_{t-j}.$$

This is called the $MA(\infty)$ representation, and expresses y in terms of the past history of the random shocks.

10.2 Estimation

In this section we discuss estimation of the autocovariance function of a stationary time series as well as the parameters of an ARMA model.

- Autocovariance. Replace population quantities by sample

$$\begin{aligned} \widehat{\gamma}_s &= \frac{1}{T-s}\sum_{t=s+1}^{T}(y_t - \overline{y})(y_{t-s} - \overline{y}) \\ \widehat{\rho}_s &= \frac{\widehat{\gamma}_s}{\widehat{\gamma}_0}. \end{aligned}$$

These sample quantities are often used to describe the actual series properties. Consistent and asymptotically normal.

- Box-Jenkins analysis: 'identification' of the process by looking at the correlogram. In practice, it is hard to identify any but the simplest processes, but the covariance function still has many uses.

- Estimation of ARMA parameters ϕ. Can 'invert' the autocovariance/autocorrelation function to compute an estimate of ϕ. For example in the AR(1) case, the parameter ρ is precisely the first order autocorrelation. In the

MA(1) case, can show that the parameter θ satisfies a quadratic equation in which the coefficients are the autocorrelation function at the first two lags. A popular estimation method is the Likelihood under normality. Suppose that

$$\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{pmatrix} \sim N(0, \sigma^2 I),$$

then

$$\begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix} \sim N(0, \Sigma)$$

for some matrix Σ.

- for an $AR(1)$ process

$$\Sigma = \frac{\sigma^2}{1-\gamma^2} \begin{bmatrix} 1 & \gamma & \gamma^2 & \cdots & \gamma^{T-1} \\ & & & \ddots & \vdots \\ & & & \ddots & 1 \end{bmatrix},$$

- for an $MA(1)$ process

$$\Sigma = \sigma^2(1+\theta^2) \begin{bmatrix} 1 & \frac{-\theta}{1+\theta^2} & & 0 \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}.$$

- For general ARMA then, the log likelihood function is

$$\ell = \frac{-T}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} y' \Sigma^{-1} y.$$

Maximize with respect to all the parameters ϕ.

- Distribution theory. The MLE is consistent and asymptotically normal provided the process is stationary and invertible.

$$T^{\frac{1}{2}}(\widehat{\phi} - \phi) \xrightarrow{D} N\left(0, I_{\phi\phi}^{-1}\right),$$

where $I_{\phi\phi}$ is the information matrix.

- In practice, $|\Sigma|$ and Σ^{-1} can be tough to find. We seek a helpful approach to computing the likelihood and an approximation to it, which is even easier to work with.

- The Prediction error decomposition is just a factorization of the joint density into the product of a conditional density and a marginal density,

$$f(x, z) = f(x \,|z\,) f(z).$$

We use this repeatedly and take logs to give

$$\ell\left(y_1, \ldots, y_T; \theta\right) = \sum_{t=p+1}^{T} \ell\left(y_t \,|y_{t-1}\,, \ldots, y_1\right) + \ell\left(y_1, \ldots, y_p\right).$$

- This writes the log likelihood in terms of conditional distributions and a single marginal distribution. In AR cases the distribution of $y_t \,|y_{t-1}\,, \ldots, y_1$ is easy to find:

$$y_t \,|y_{t-1}\,, \ldots, y_1 \sim N\left(\phi_1 y_{t-1} + \ldots + \phi_p y_{t-p}, \sigma^2\right).$$

- In the $AR(1)$ case

$$\ell_{t|t-1} \sim -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y_t - \phi_1 y_{t-1}\right)^2.$$

Also, $y_1 \sim N\left(0, \sigma^2/(1-\phi^2)\right)$, i.e.,

$$\ell(y_1) = -\frac{1}{2}\log\frac{\sigma^2}{1-\phi^2} - \frac{(1-\phi^2)}{2\sigma^2} y_1^2.$$

Therefore, the full likelihood in the $AR(1)$ case is

$$\ell = -\frac{T-1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - \phi y_{t-1})^2 - \frac{1}{2}\log\frac{\sigma^2}{1-\phi^2} - \frac{1-\phi^2}{2\sigma^2} y_1^2.$$

- Often it is argued that $\ell(y_1)$ is small relative to $\sum_{t=2}^{T} \ell\left(y_t \,|y_{t-1}\,, \ldots, y_1\right)$, in which case we use

$$-\frac{T-1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(y_t - \phi y_{t-1}\right)^2.$$

- This criterion is equivalent to the least squares criterion, and has unique maximum
$$\widehat{\phi} = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}.$$
This estimator is just the OLS on y_t on y_{t-1} [but using the reduced sample]. Can also interpret this as a GMM estimator with moment condition
$$E\left[y_{t-1}\left(y_t - \phi y_{t-1}\right)\right] = 0.$$
- The full MLE will be slightly different from the approximate MLE. In terms of asymptotic properties, the difference is negligible.
  - However, in finite sample there can be significant differences.
  - Also, the MLE imposes that $\widehat{\phi}$ be less than one - as $\phi \to \pm 1$, $\ell \to -\infty$. The OLS estimate however can be either side of the unit circle.

10.3 Forecasting

- Let the sample be $\{y_1, \ldots, y_T\}$. Suppose that
$$y_t = \gamma y_{t-1} + \varepsilon_t, \quad |\gamma| < 1,$$
where we first assume that γ is known.
- Want to forecast $y_{T+1}, y_{T+2}, \ldots, y_{T+r}$ given the sample information. We have
$$y_{T+1} = \gamma y_T + \varepsilon_{T+1}.$$
Therefore, forecast y_{T+1} by
$$\widehat{y}_{T+1|T} = E\left[y_{T+1}\,|\text{sample}\right] = \gamma y_T.$$
- The forecast error is ε_{T+1}, which is mean zero and has variance σ^2.
- What about forecasting r periods ahead?
$$y_{T+r} = \gamma^r y_T + \gamma^{r-1}\varepsilon_{T+1} + \ldots + \varepsilon_{T+r}.$$
Therefore, let
$$\widehat{y}_{T+r|T} = \gamma^r y_T$$
be our forecast.

- The forecast error $\widehat{y}_{T+r|T} - y_{T+r}$ has mean zero and variance

$$\sigma^2 \left(1 + \gamma^2 + \ldots + \gamma^{2r-2}\right).$$

- Asymptotically the forecast reverts to the unconditional mean and the forecast variance reverts to the unconditional variance.

- In practice, we must use an estimate of γ, so that

$$\widehat{y}_{T+r|T} = \widehat{\gamma}^r y_T,$$

  where $\widehat{\gamma}$ is estimated from sample data. If γ is estimated well, then this will not make much difference.

- Forecast interval

$$\widehat{y}_{T+r|t} \pm 1.96 \cdot DS,$$
$$SD = \sigma^2 \left(1 + \gamma^2 + \ldots + \gamma^{2r-2}\right).$$

  This is to be interpreted like a confidence interval. Again we must replace the unknown parameters by consistent estimates.

- This theory generalizes naturally to AR(2) and higher order AR processes in which case the forecast is a linear combination of the most recent observations. The question is, how to forecast for an MA(1) process?

$$y_t = \varepsilon_t - \theta\varepsilon_{t-1} = (1 - \theta L)\varepsilon_t.$$

  We must use the AR(∞) representation

$$\frac{y_t}{1 - \theta L} = y_t + \theta y_{t-1} + \ldots = \varepsilon_t.$$

  This means that the forecast for MA processes is very complicated and depends on all the sample $y_1, \ldots, y_T$.

10.4 Autocorrelation and Regression

- Regression models with correlated disturbances

$$y_t = \beta' x_t + u_t,$$

where x_t is *exogenous*, i.e., is determined outside the system; fixed regressors are an example. There are a number of different variations on this theme - strongly exogenous and weakly exogenous. A weakly exogenous process could include lagged dependent variables. We will for now assume strong exogeneity.

- We also suppose that

$$E(u_t u_s) \neq 0 \text{ for some } s \neq t.$$

- As an example, suppose that

$$\ln GNP = \beta_1 + \beta_2 \text{time } + u_t.$$

We expect the deviation from trend, u_t, to be positively autocorrelated reflecting the business cycle, i.e., not i.i.d. Recession quarter tends to be followed by recession quarter.

- We can write the model in matrix form

$$y = X\beta + u,$$

$$E(uu') = \Sigma = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{T-1} \\ & & & \ddots & \gamma_2 \\ & & & \ddots & \gamma_0 \end{bmatrix}.$$

- The consequences for estimation and testing of β are the same as with heteroskedasticity: OLS is consistent and unbiased, but inefficient, while the SE's are wrong.

- Specifically,

$$var(\widehat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1},$$

where

$$\psi_T = X'\Sigma X = \sum_{t=1}^{T}\sum_{s=1}^{T} x_t x_t' \gamma_{|t-s|}.$$

- A naive implementation of the White strategy is going to fail here, i.e.,

$$\widehat{\psi}_T = X' \begin{pmatrix} \widehat{u}_1^2 & \widehat{u}_1\widehat{u}_2 & \cdots & \widehat{u}_1\widehat{u}_T \\ & \widehat{u}_1^2 & & \\ & & \ddots & \\ & & & \widehat{u}_T^2 \end{pmatrix} X = \sum_{t=1}^{T}\sum_{s=1}^{T} x_t x_t' \widehat{u}_t \widehat{u}_s$$

  is inconsistent. This is basically because there are too many random variables in the sample matrix, in fact order T^2, whereas in the independent but heterogeneous case there were only order T terms.

- The correct approach is to use some downweighting that concentrates weight on a smaller fraction of their terms. Bartlett/White/Newey/West SE's: Replace by sample equivalents and use weights

$$w(j) = 1 - \frac{j}{n+1},$$

  so that

$$\widehat{\psi}_T = \sum_{t,s:|t-s|\leq n(T)} \sum X_t X_s' w\left(|t-s|\right) \widehat{u}_t \widehat{u}_s.$$

  This also ensures a positive definite covariance matrix estimate. Provides consistent standard errors.

- An alternative strategy is to parameterize u_t by, say, an ARMA process and do maximum likelihood

$$\ell = -\frac{1}{2}\ln\left|\Sigma(\theta)\right| - \frac{1}{2}\left(y - X\beta\right)'\Sigma(\theta)^{-1}\left(y - X\beta\right).$$

- Efficient estimate of β (under Gaussianity) is a sort of GLS

$$\widehat{\beta}_{ML} = \left(X'\Sigma(\widehat{\theta})^{-1}X\right)^{-1} X'\Sigma(\widehat{\theta})^{-1}y,$$

  where $\widehat{\theta}$ is the MLE of θ. This will be asymptotically efficient when the chosen parametric model is correct.

10.5 Testing for Autocorrelation

- Suppose that we observe u_t, which is generated from an $AR(1)$ process

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

where ε_t are i.i.d.

- The null hypothesis is that u_t is i.i.d., i.e.,

$$H_0 : \rho = 0 \quad \text{vs.} \quad H_A : \rho \neq 0.$$

This is used as (a) general diagnostic, and (b) efficient markets.

- General strategy: use LR, Wald or LM tests to detect departures.

- The LM test is easiest, this is based on

$$LM = T \left(\frac{\sum_t \widehat{u}_t \widehat{u}_{t-1}}{\sum_t \widehat{u}_{t-1}^2} \right)^2 = T r_1^2 \xrightarrow{D} \chi_1^2,$$

where $\widehat{u}_t$ are the OLS residuals. Therefore, we reject the null hypothesis when LM is large relative to the critical value from χ_1^2.

- This approach is limited to two-sided alternatives. We can however also use the signed version, $T^{\frac{1}{2}} r_1$, which satisfies

$$T^{\frac{1}{2}} r_1 \xrightarrow{D} N(0,1)$$

under the null hypothesis.

- The Durbin-Watson d is

$$d = \frac{\sum_{t=2}^{T} (\widehat{u}_t - \widehat{u}_{t-1})^2}{\sum_{t=1}^{T} \widehat{u}_t^2}.$$

This is always printed out by many regression packages.

- Using the approximation

$$d \approx 2(1 - r_1),$$

we have [under the null hypothesis]

$$T^{\frac{1}{2}} \left(1 - \frac{d}{2} \right) \to N(0,1).$$

- Generalization (test against $AR(p)$). Suppose that

$$u_t = \rho_1 u_{t-1} + \ldots + \rho_p u_{t-p} + \varepsilon_t,$$

where ε_t are i.i.d. The null hypothesis is that u_t is i.i.d., i.e.,

$$H_0 : \rho_1 = \ldots = \rho_p = 0$$

$$\text{vs.} \quad H_A \text{some } \rho_j \neq 0.$$

- Box-Pierce Q

$$Q = T \sum_{j=1}^{P} r_j^2 \xrightarrow{D} \chi_P^2.$$

10.6 Dynamic Regression Models

- We have looked at pure time series models with dynamic response and at static regression models. In practice, we may want to consider models that have both features.

- Distributed lag

$$y_t = \alpha + \sum_{j=0}^{q} \beta_j X_{t-j} + u_t,$$

[could have $q = \infty$], where for now

$$u_t \overset{iid}{\sim} 0, \quad \sigma^2.$$

Captures the idea of dynamic response: affect on y of change in x may take several periods to work through.

- Temporary change. Suppose that

$$x_t \to x_t + \Delta$$

but that future x_s are unaffected, then

$$\begin{aligned} y_t &\to y_t + \beta_0 \Delta \\ y_{t+1} &\to y_t + \beta_1 \Delta \text{ etc.} \end{aligned}$$

- Permanent change. Suppose that

$$x_s \to x_s + \Delta, \quad \forall s \geq t.$$

Then

$$\begin{aligned} y_t &\to y_t + \beta_0 \Delta \\ y_{t+1} &\to y_t + (\beta_0 + \beta_1)\Delta \text{ etc.} \end{aligned}$$

  - The impact effect is $\beta_0 \Delta$.
  - Long run effect is $\Delta \sum_{s=0}^{\infty} \beta_s$.

- When q is large (infinite) there are too many free parameters β_j, which makes estimation difficult and imprecise. To reduce the dimensionality it is appropriate to make restrictions on β_j.

  - The polynomial lag

$$\beta_j = \begin{cases} a_0 + a_1 j + \ldots + a_p j^p & \text{if } j \leq p \\ 0 & \text{else.} \end{cases}$$

  - The Geometric lag

$$\beta_j = \beta \lambda^j, \quad j = 0, 1, \ldots$$

for some $0 < \lambda < 1$. This implies that

$$\begin{aligned} y_t &= \alpha + \beta \sum_{j=0}^{\infty} \lambda^j x_{t-j} + u_t \\ &= \alpha + \beta \left[\sum_{j=0}^{\infty} (\lambda^j L^j) \right] x_t + u_t \\ &= \alpha + \beta \frac{1}{1 - \lambda L} x_t + u_t. \end{aligned}$$

Therefore,

$$(1 - \lambda L) y_t = \alpha(1 - \lambda L) + \beta x_t + (1 - \lambda L) u_t,$$

which is the same as

$$y_t = \alpha(1 - \lambda) + \lambda y_{t-1} + \beta x_t + u_t - \lambda u_{t-1}.$$

The last equation is called the lagged dependent variable representation.

- More generally [ADL model]

$$A(L)y_t = B(L)x_t + u_t,$$

where A, B are polynomials of order p, q, while

$$C(L)u_t = D(L)\varepsilon_t, \quad \varepsilon_t \text{ i.i.d. } 0, \sigma^2.$$

This is a very general class of models; estimation, forecasting, and testing have all been worked out at this generality, and one can find accounts of this in advanced time series texts.

10.7 Adaptive expectations

- Suppose that

$$\underbrace{y_t}_{\text{demand}} = \alpha + \beta \underbrace{x^*_{t+1}}_{\text{expected price}} + \varepsilon_t,$$

but that the expected price is made at time t and is unobserved by the econometrician. Let

- We observe x_t, where

$$\underbrace{x^*_{t+1} - x^*_t}_{\text{revised expectations}} = (1-\lambda) \underbrace{(x_t - x^*_t)}_{\text{forecast error}},$$

i.e.,

$$x^*_{t+1} = \underbrace{\lambda x^*_t}_{\text{old forecast}} + \underbrace{(1-\lambda)x_t}_{\text{news}}.$$

- Write

$$(1 - \lambda L)x^*_t = (1-\lambda)x_t,$$

which implies that

$$x^*_t = \frac{(1-\lambda)}{1-\lambda L}x_t = (1-\lambda)\left[x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \ldots\right].$$

- Therefore,

$$y_t = \alpha + \frac{\beta(1-\lambda)}{1-\lambda L}x_t + \varepsilon_t,$$

which implies that

$$y_t = \lambda y_{t-1} + \alpha(1-\lambda) + \beta(1-\lambda)x_t + \varepsilon_t - \lambda\varepsilon_{t-1}.$$

This is an ADL with an $MA(1)$ error term.

10.8 Partial adjustment

- Suppose that
$$y_t^* = \alpha + \beta x_t,$$
where y_t^* is the desired level.

- However, because of costs of adjustment
$$\underbrace{y_t - y_{t-1}}_{\text{actual change}} = (1-\lambda)(y_t^* - y_{t-1}) + \varepsilon_t.$$

- Substituting we get
$$\begin{aligned} y_t &= (1-\lambda)y_t^* + \lambda y_{t-1} + \varepsilon_t \\ &= \alpha(1-\lambda) + \lambda y_{t-1} + \beta(1-\lambda)x_t + \varepsilon_t. \end{aligned}$$
This is an ADL with an i.i.d. error term - assuming that the original error term was i.i.d.

10.9 Error Correction

- Suppose long run equilibrium is
$$y = \lambda x.$$

- Disequilibria are corrected according to
$$\Delta y_t = \beta\left(y_{t-1} - \lambda x_{t-1}\right) + \lambda \Delta x_{t-1} + \varepsilon_t,$$
where $\beta < 0$.

- This implies that
$$y_t = y_{t-1}(1+\beta) + \lambda(1-\beta)x_{t-1} - \lambda x_{t-2} + \varepsilon_t.$$

10.10 Estimation of ADL Models

- Suppose that
$$y_t = \theta_1 + \theta_2 y_{t-1} + \theta_3 x_t + \varepsilon_t,$$
where we have two general cases regarding the error term:

  (1) ε_t is i.i.d. 0, σ^2

  (2) ε_t is autocorrelated.

- In case (1), we can use OLS regression to get consistent estimates of θ_1, θ_2 and θ_3. The original parameters are related to the θ_j in some way, for example
$$\left.\begin{array}{l} \theta_1 = \alpha(1-\lambda) \\ \theta_2 = \lambda \\ \theta_3 = \beta(1-\lambda) \end{array}\right\}.$$
In this case, we would estimate the original parameters by indirect least squares
$$\begin{aligned} \widehat{\lambda} &= \widehat{\theta}_2 \\ \widehat{\alpha} &= \frac{\widehat{\theta}_1}{1-\widehat{\theta}_2} \\ \widehat{\beta} &= \frac{\widehat{\theta}_3}{1-\widehat{\theta}_2}. \end{aligned}$$

- In case (2), we must use instrumental variables or some other procedure because OLS will be inconsistent.

  - For example, if
  $$\varepsilon_t = \eta_t - \theta\eta_{t-1},$$
  then y_{t-1} is correlated with ε_t through η_{t-1}. In this case there are many instruments: (1) All lagged x_t, (2) $y_{t-2}, \ldots$.

  - However, when
  $$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t,$$
  η_t i.i.d. lagged y are no longer valid instruments and we must rely on lagged x.

- There are many instruments; efficiency considerations require that one has a good way of combining them such as in our GMM discussion.
- IV are not generally as efficient as ML when the error terms are normally distributed.

10.11 Nonstationary Time Series Models

- There are many different ways in which a time series y_t can be nonstationary. For example, there may be fixed seasonal effects such that

$$y_t = \sum_{j=1}^{m} D_{jt}\gamma_j + u_t,$$

where D_{jt} are seasonal dummy variables, i.e., one if we are in season j and zero otherwise. If u_t is an iid mean zero error term,

$$Ey_t = \sum_{j=1}^{m} D_{jt}\gamma_j$$

and so varies with time. In this case there is a sort of periodic movement in the time series but no 'trend'.

- We next discuss two alternative models of trending data: trend stationary and difference stationary.
- Trend stationary. Consider the following process

$$y_t = \mu + \beta t + u_t,$$

where $\{u_t\}$ is a stationary mean zero process e.g.,

$$A(L)u_t = B(L)\varepsilon_t$$

with the polynomials A, B satisfying the usual conditions required for stationarity and invertibility. This is the trend+stationary decomposition.

  - We have

$$Ey_t = \mu + \beta t; \qquad \text{var}(y_t) = \sigma^2$$

for all t. The lack of stationarity comes only through the mean.

  - The shocks (u_t) are transitory - they last for some period of time and then are forgotten as y_t returns to trend.
  - Example. GNP grows at 3% per year (on average) for ever after.

- Difference stationary $I(1)$

$$y_t = \mu + y_{t-1} + u_t,$$

  where $\{u_t\}$ is a stationary process. This is called the random walk plus drift. When $\mu = 0$, we have the plain vanilla random walk.

  - We can't now suppose that the process has been going on for an infinite amount of time, and the starting condition is of some significance.
  - We can make two assumptions about the initial conditions:

$$y_0 = \begin{cases} \text{fixed} \\ \text{random Variable } N(0, v) \end{cases}$$

    for some variance v.
  - Any shocks have permanent affects

$$y_t = y_0 + t\mu + \sum_{s=1}^{t} u_s.$$

    The differenced series is than

$$\Delta y_t = y_t - y_{t-1} = \mu + u_t.$$

- Both the mean and the variance of this process are generally explosive.

$$Ey_t = y_0 + t\mu; \qquad \text{var } y_t = \sigma^2 t.$$

  If $\mu = 0$, the mean does not increase over time but the variance does.

- Note that differencing in the trend stationary case gives

$$\Delta y_t = \beta + u_t + u_{t-1},$$

  which is a unit root MA. So although differencing apparently eliminates stationarity it induces non-invertibility. Likewise detrending the difference stationary case is not perfect.

- A model that nests both trend stationary and difference stationary is

$$y_t = \mu + \beta t + u_t,\ u_t = \rho u_{t-1} + \eta_t,$$

where η_t is a stationary ARMA process. We have

$$y_t = \mu + \beta t + \rho(y_{t-1} - \mu + \beta(t-1)) + \eta_t,$$

When $\rho = 1$ and $\beta = 0$ we get the random walk plus drift.

10.12 Estimation

- Effects of time trend on estimation:
  - you get superconsistent $T^{3/2}$ estimates of β, but still Gaussian t-tests still valid.
- Effects of unit root:
  - superconsistent estimates, but with nonstandard distributions: t-tests not valid!
- Suppose that

$$y_t = \rho y_{t-1} + u_t, \quad \text{where } u_t \sim 0,\ \sigma^2.$$

Then,

$$\widehat{\rho}_{OLS} = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} \xrightarrow{P} \rho, \quad \forall \rho.$$

  - If $|\rho| < 1$

$$T^{\frac{1}{2}}(\widehat{\rho} - \rho) \to N(0, 1 - \rho^2).$$

  - If $\rho = 1$, $1 - \rho^2 = 0$, so the implied variance above is zero. So what happens in this case? If $\rho = 1$,

$$T(\widehat{\rho} - \rho) \xrightarrow{D} X,$$

where X is not Gaussian; it is asymmetric and in fact $E(\widehat{\rho}) < 1$, $\forall T$. The rate of convergence is faster but the asymptotic distribution is non standard.

  - Dickey-Fuller (1981) derived the distribution of $\widehat{\rho}$ and the distribution of the corresponding t-statistic, t_ρ, when $\rho = 1$, and they tabulated it.

10.13 Testing for Unit Roots

- Suppose that

$$y_t = \mu + \rho y_{t-1} + u_t,$$

where the process u_t is i.i.d. By taking differences we obtain

$$\Delta y_t = \mu + \gamma y_{t-1} + u_t$$

with $\gamma = \rho - 1$.

- To test whether $\rho = 1$ is equivalent to testing $\gamma = 0$ in the model

$$\Delta y_t = \mu + \gamma y_{t-1} + u_t.$$

We do a one-sided test

$$H_0 : \gamma = 0 \text{ vs } \gamma < 0$$

because the explosive alternatives are not interesting.

- Dickey and Fuller (1979) tabulated the distribution of the least squares estimator $\widehat{\gamma}$ and its associated t-test in the case that $\rho = 1$ i.e., $\gamma = 0$. This is exactly the null case. Their critical values can be used to do the test. Large negative values of the test statistic are evidence against the null hypothesis.

- The critical values are -3.96 and -3.41 at the 1% and 5% levels respectively.

- If you do it without the intercept, i.e., run the regression

$$\Delta y_t = \gamma y_{t-1} + u_t.$$

the critical values are -3.43 and -2.86 at the 1% and 5% levels respectively. This assumes that the null hypothesis is the driftless random walk.

- Can also do a test based on the raw estimates.

- The DF test is only valid if the error term u_t is i.i.d. Have to adjust for the serial correlation in the error terms to get a valid test. The Augmented D-F allows the error term to be correlated over time upto a certain order. Their test is based on estimating the regression

$$\Delta y_t = \mu + \gamma y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + \eta_t$$

  by least squares and using the ADF critical values for $\widehat{\gamma}$ or rather the t-ratio.

- Can also add trend terms in the regression. Phillips-Perron test (PP) is an alternative way of correcting for serial correlation in u_t.

- Applications

10.14 Cointegration

- Suppose y_t and x_t are $I(1)$ but there is a β such that

$$y_t - \beta x_t$$

  is $I(0)$, then we say that y_t, x_t are cointegrated.

  - For example, aggregate consumption and income appear to be nonstationary processes, but appear to deviate from each other in only a stationary fashion, i.e., there exists a long-run equilibrium relationship about which there are only stationary deviations.
  - Note that β is not necessarily unique.

- Can estimate the cointegrating parameter β by an OLS regression of y_t on x_t but although the estimator is consistent, the distribution theory is again non-standard, but has been tabulated.

- More general system. Suppose that $y_t = (y_{1t}, y_{2t})' \in \mathbb{R}^{k_1+k_2}$ and that

$$\begin{aligned} y_{1t} &= \beta' y_{2t} + u_t \\ y_{2t} &= y_{2t-1} + \eta_t, \end{aligned}$$

If u_t and η_t are mutually uncorrelated, then we call the system triangular. Special results apply in this case. This model assumes knowledge about the number of cointegrating relations, i.e., k_1, and it makes a particular normalization. Can

- Johansen test for the presence of cointegration and the number of cointegrating relations. If we have a k-vector unit root series y_t there can be no cointegrating relations, one,..., $k-1$ cointegrating relations.. Johansen tests these restrictions sequentially to find the right number of cointegrating relations in the data.

10.15 Martingales

- We say that the process y_t is a martingale if

$$E\left[y_t \left| I_{t-1}\right.\right] = y_{t-1} \text{ a.s.},$$

where I_{t-1} is information available at time t, for example $I_{t-1} = \{y_{t-1}, \ldots\}$, i.e.,

$$y_t = y_{t-1} + u_t,$$

where u_t is a martingale difference sequence and satisfies

$$E\left[u_t \left| I_{t-1}\right.\right] = 0 \text{ a.s.}$$

The process u_t may be heterogeneous but is uncorrelated.

  - Hall (1978): Consumption is a martingale.
  - Fama: Stock prices are martingales.

$$E\left(P_{t+1} \left| P_t, \ldots\right.\right) = P_t.$$

    This is a bit too strong and is unsupported by the data.

- The assumption of unforecastability rules out serial correlation in ε_t and hence r_t, but it does not by itself say anything more about the distribution of ε_t. That is, ε_t could be heterogeneous and be non-normal. It could itself be non-stationary - for example ε_t independent over time with

$$\varepsilon_t \sim N\left(0, f(t)\right)$$

is consistent with the efficient markets hypothesis. However, it is frequently assumed that the error term is itself stationary process.

10.16 GARCH Models

- Engle (1982) introduced the following class of models

$$r_t = \varepsilon_t \sigma_t,$$

where ε_t is i.i.d. $(0, 1)$, while

$$\sigma_t^2 = var\left(r_t \,|\mathcal{F}_{t-1}\right)$$

is the (time-varying) conditional variance.

- For example,

$$\sigma_t^2 = \alpha + \gamma r_{t-1}^2,$$

which is the $ARCH(1)$ model. Provided $\gamma < 1$, the process r_t is weakly stationary and has finite unconditional variance σ^2 given by

$$\sigma^2 = E(\sigma_t^2) < \infty,$$

where

$$\sigma^2 = \alpha + \gamma\sigma^2 = \frac{\alpha}{1-\gamma}.$$

- This uses the law of iterated expectations $E(Y) = E\left(E(Y\,|I)\right)$ to argue

$$\begin{aligned} E\left(r_{t-1}^2\right) &= E\left(E\left(\varepsilon_{t-1}^2\,|I_{t-1}\right)\sigma_{t-1}^2\right) \\ &= E\left(\sigma_{t-1}^2\right) = \sigma^2. \end{aligned}$$

- The unconditional distribution of r_t is thick-tailed; that is, even if ε_t is normally distributed, r_t is going to have an unconditional distribution that is a mixture of normals and is more leptokurtic. Suppose ε_t is standard normal, then $E(\varepsilon_t^4) = 3$ and

$$\begin{aligned} \mu_4 &= E(r_t^4) = E\left(\varepsilon_t^4\sigma_t^4\right) \\ &= 3E(\sigma_t^4), \end{aligned}$$

where

$$\begin{aligned} E(\sigma_t^4) &= E\left[\left(\alpha^2 + \gamma^2 r_{t-1}^4 + 2\alpha\gamma r_{t-1}^2\right)\right] \\ &= \alpha^2 + \gamma^2\mu_4 + 2\alpha\gamma\sigma^2. \end{aligned}$$

Therefore,

$$\begin{aligned}\mu_4 &= 3\left(\alpha^3 + \gamma^2\mu_4 + 2\alpha\gamma\sigma^2\right) \\ &= \frac{3\left(\alpha^2 + 2\alpha\gamma\sigma^2\right)}{1 - 3\gamma^2} \\ &\geq 3\sigma^4 = \frac{3\alpha^2}{(1-\gamma)^2}.\end{aligned}$$

- The process r_t is uncorrelated, i.e.,

$$cov(r_t, r_{t-s}) = 0$$

for all $s \neq 0$. However, the process r_t is dependent so that

$$E\left(g(r_t)g(r_{t-s})\right) \neq E\left(g(r_t)\right)E(h(r_{t-s}))$$

for arbitrary functions g, h, certainly for $g(r) = h(r) = r^2$ this is not true.

- Can write the process as an $AR(1)$ process in u_t^2, i.e.,

$$r_t^2 = \alpha + \gamma r_{t-1}^2 + \eta_t,$$

where $\eta_t = r_t^2 - \sigma_t^2$ is a mean zero innovation that is uncorrelated with its past.

- Therefore, since $\gamma > 0$, the volatility process is positively autocorrelated, i.e.,

$$cov\left(\sigma_t^2, \sigma_{t-j}^2\right) > 0.$$

Hence we get volatility clustering.

- We can rewrite the process as

$$\sigma_t^2 - \sigma^2 = \gamma\left(r_{t-1}^2 - \sigma^2\right).$$

Suppose that $\sigma_{t-1}^2 = \sigma^2$. When we get a large shock, i.e., $\varepsilon_{t-1}^2 > 1$, we get $\sigma_t^2 > \sigma^2$ but the process decays rapidly to σ^2 unless we get a sequence of large shocks $\varepsilon_{t-1+s}^2 > 1$, $s = 0, 1, 2, \ldots$. In fact, for a normal distribution the probability of having $\varepsilon^2 > 1$ is only about 0.32 so we generally see little persistence.

- Although the ARCH model implies volatility clustering, it does not in practice generate enough.

- Generalize to $ARCH(p)$, write

$$\sigma_t^2 = \alpha + \sum_{j=1}^{p} \gamma_j r_{t-j}^2,$$

where p is some positive integer and γ_j are positive coefficients.

- This model is fine, but estimation is difficult. When p is large one finds that the coefficients are imprecisely estimated and can be negative. Have to impose some restrictions on the coefficients.

- Instead $GARCH(1,1)$

$$\sigma_t^2 = \alpha + \beta\sigma_{t-1}^2 + \gamma r_{t-1}^2,$$

where α, β, γ are positive.

  - We have

$$\sigma_t^2 = \frac{\alpha}{1-\beta} + \gamma \sum_{j=1}^{\infty} \beta^{j-1} r_{t-j}^2,$$

    so that it is an infinite order ARCH model with geometric decline in the coefficients.

  - If $\gamma + \beta < 1$, then the process r_t is weakly stationary, i.e., the unconditional variance exists, and

$$\sigma^2 = E(\sigma_t^2) < \infty,$$

    where

$$\sigma^2 = \alpha + \beta\sigma^2 + \gamma\sigma^2 = \frac{\alpha}{1-(\beta+\gamma)}.$$

  - Surprisingly, even for some values of β, γ with $\gamma + \beta \geq 1$, the process σ_t^2 is strongly stationary although the unconditional variance does not exist in this case.

- More general class of models $GARCH(p,q)$

$$B(L)\sigma_t^2 = \alpha + C(L)r_{t-1}^2,$$

where A and B are lag polynomials. Usually assume that the parameters in α, B, $C > 0$ to ensure that the variance is positive.

- Other models. For example, one can write the model for log of variance, i.e.,

$$\log \sigma_t^2 + \alpha + \beta \log \sigma_{t-1}^2 + \gamma r_{t-1}^2.$$

This automatically imposes the restriction that $\sigma_t^2 \geq 0$ so there is no need to impose restrictions on the parameters.

- Nelsons $EGARCH$

$$\log \sigma_t^2 = \alpha + \beta \log \sigma_{t-1}^2 + \gamma \varepsilon_t + \delta \left(|\varepsilon_t| - E\left(|\varepsilon_t| \right) \right).$$

- $TARCH$, $SGARCH$, $CGARCH$ etc.

10.17 Estimation

- More general model

$$\begin{aligned} y_t &= b'x_t + \varepsilon_t \sigma_t \\ B(L)\sigma_t^2 &= \alpha + C(L)\left(y_{t-1} - b'x_{t-1} \right)^2. \end{aligned}$$

- If ARCH effects are present, then we need to use robust estimates of the standard errors for the parameters b of the mean model.

- Also, the variance process itself is of interest. Want to estimate the parameters of σ_t^2 too.

- Let $\theta = \left(b, \alpha, \beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q \right)$. Estimation by ML suggested by ε_t being standard normal. In this case

$$\ell_T(\theta) = -\frac{1}{2}\sum_{t=1}^{T} \log \sigma_t^2(\theta) - \frac{1}{2}\sum_{t=1}^{T} \frac{\left(y_t - b'x_t \right)^2}{\sigma_t^2(\theta)}.$$

The ML estimator of b, θ can be obtained from this criterion. This involves nonlinear optimization. Have to impose the inequality restrictions on the parameters which can be tricky.