

Methods of Economic Investigation II,  
Part I

Oliver Linton

December 2, 1999



# Contents

<b>1</b>	<b>Linear Model</b>	<b>7</b>
1.0.1	The OLS Procedure . . . . .	9
<b>2</b>	<b>Geometry of OLS and Partitioned Regression</b>	<b>15</b>
2.0.2	Goodness of Fit . . . . .	19
<b>3</b>	<b>Statistical Properties of the OLS Estimator</b>	<b>21</b>
3.0.3	Optimality . . . . .	23
<b>4</b>	<b>Hypothesis Testing</b>	<b>27</b>
4.0.4	General Notations . . . . .	28
4.0.5	Examples . . . . .	29
4.0.6	Test of a Single Linear Hypothesis . . . . .	30
4.0.7	Test of a Multiple Linear Hypothesis . . . . .	32
<b>5</b>	<b>Test of Multiple Linear Hypothesis Based on fit</b>	<b>35</b>
<b>6</b>	<b>Examples of <math>F</math>-Tests, <math>t</math> vs. <math>F</math></b>	<b>39</b>
<b>7</b>	<b>Likelihood Based Testing</b>	<b>41</b>
<b>8</b>	<b>Omission of Relevant Variables, Inclusion of Irrelevant Variables, and Model Selection</b>	<b>45</b>

8.0.8	Omission of Relevant Variables . . . . .	45
8.0.9	Inclusion of irrelevant variables . . . . .	47
8.0.10	Model Selection . . . . .	47
<b>9</b>	<b>Data Problems</b>	<b>49</b>
9.0.11	Functional Form . . . . .	49
9.0.12	Multicollinearity . . . . .	51
9.0.13	Influential Observations . . . . .	52
9.0.14	Missing Observations . . . . .	52
<b>10</b>	<b>Asymptotic Theory I</b>	<b>55</b>
<b>11</b>	<b>Asymptotic Theory II</b>	<b>61</b>
11.0.15	The delta method . . . . .	63
<b>12</b>	<b>Errors in Variables</b>	<b>67</b>
12.0.16	Solutions to EIV . . . . .	69
12.0.17	Durbin-Wu-Hausman Test . . . . .	70
<b>13</b>	<b>Heteroskedasticity</b>	<b>73</b>
13.0.18	Effects of Heteroskedasticity . . . . .	73
13.0.19	Plan A: Eicker-White . . . . .	75
13.0.20	Plan B: Model Heteroskedasticity . . . . .	75
13.0.21	Testing for Heteroskedasticity . . . . .	77
<b>14</b>	<b>Nonlinear Regression Models</b>	<b>79</b>
14.0.22	Computation . . . . .	80
<b>15</b>	<b>Asymptotic Properties</b>	<b>83</b>
15.0.23	Consistency of NLLS . . . . .	83
15.0.24	Asymptotic Distribution of NLLS . . . . .	84

<i>CONTENTS</i>	5
15.0.25 Likelihood and Efficiency . . . . .	86
<b>16 Generalized Method of Moments</b>	<b>89</b>
<b>17 Time Series</b>	<b>93</b>
17.0.26 Some Fundamental Properties . . . . .	93
17.0.27 Estimation . . . . .	97
17.0.28 Forecasting . . . . .	99
<b>18 Autocorrelation and Regression</b>	<b>101</b>
18.0.29 Testing for autocorrelation . . . . .	103
<b>19 Dynamic Regression Models</b>	<b>105</b>
19.0.30 Some examples from economics . . . . .	107
19.0.31 Estimation of ADL models . . . . .	108
<b>20 Nonstationarity</b>	<b>111</b>



# Chapter 1

## Linear 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)  $\exists \beta = (\beta_1, \dots, \beta_K)'$  such that  $E(y) = X\beta$ .

$$(A3) \text{Var}(y) = \sigma^2 I_{n \times n}.$$

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 these relations hold with probability one.

An example brings out the difference between these two sampling schemes.

*Example.* A student drops chalk from different heights

$y$  = time elapsed

$x$  = height of chalk.

Each student in the class takes a measurement on  $y, x$ ; in the fixed design we would observe the same  $x$  value for each student. In the random design we might have a range of different  $x$  values.

We stick with fixed design for most of linear regression section. Fixed design is perhaps unconvincing for most economic data sets, especially times series, 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, especially in time series when one regressor might be a lagged value of the dependent variable. However, one can believe in a random design model, but want to conduct inference in the conditional distribution [given  $X$ ]; in this case, we are effectively working in a fixed design model. So the real distinction perhaps is rather whether one evaluates quantities in the conditional or unconditional distribution.

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

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

where

$$E(\varepsilon) = E(\varepsilon|X) = 0$$

$$E(\varepsilon\varepsilon') = \sigma^2 I_n.$$

Frequently, the model is stated like this with statistical assumptions made about the unobservable  $\varepsilon$  rather than directly on the observable  $y$ .

It is worth discussing here some alternative assumptions made about the error terms; for this purpose we shall assume a random design.

1.  $E(\varepsilon_i|x_i) = 0$ , denoted  $\varepsilon_i \perp x_i$
2.  $\varepsilon_i$  are i.i.d. and independent of  $x_i$ , denoted  $\varepsilon_i \perp\!\!\!\perp x_i$ .

The first assumption is quite weak and is really just rewriting the definition of conditional expectation; it is a minimal assumption needed to ‘identify’ the regression. The second assumption is much stronger and is not strictly necessary for estimation purposes although it does have implications about efficiency and choice of estimator. A further assumption, which we will make, is that

3.  $\varepsilon_i \sim N(0, \sigma^2)$ .

This is even stronger than the independence assumption; it is used for hypothesis testing and for establishing optimality of least squares.

### 1.0.1 The OLS Procedure

In practice we don’t know the parameter  $\beta$  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  $\hat{\beta}$  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  $\beta$ , and we shall study its properties at length.

#### COMMENTS

1. Better than  $\sum_{i=1}^n u_i(b)$ , or  $|\sum_{i=1}^n u_i(b)|$ , why?
2. Alternatives are  $\min_b \sum_{i=1}^n |u_i(b)|$ , the Least Absolute Deviation criterion, or other distances say the perpendicular squared distance from the data to the line.
3. Any method needs at least  $n \geq K$ . If  $n = K$ , the solution essentially involves interpolating the data, i.e., the fitted value of  $y$  will be equal to the actual value.
4.  $\hat{\beta}$  is defined for any  $y$  and  $X$  independently of model, provided  $\text{rank}(X) = K$ , so there is no need for assumptions A1-A3 when it comes to computing the estimator.

We briefly mention some alternative estimation methods which actually lead to the same estimator in some cases. These methods are useful in more general situations, which is why we must introduce them.

MAXIMUM LIKELIHOOD. Suppose we also assume that

A4.  $y \sim N(X\beta, \sigma^2 I)$ .

Then the log likelihood function for the observed data is

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

where  $b$  and  $\omega$  are unknown parameters. The maximum likelihood estimator  $\hat{\beta}_{mle}, \hat{\sigma}_{mle}^2$  maximizes  $\ell(b, \omega^2)$  with respect to  $b$  and  $\omega^2$ . It is easy to see that  $\max_{b, \omega} \ell(b, \omega^2)$  is equivalent to  $\min_b S(b)$ , i.e.,

$\widehat{\beta}$  is the MLE of  $\beta$  when A4 holds. However, 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  $\beta$  is defined as the unique parameter that satisfies the  $K$  moment conditions

$$E[x_i(y_k - x_i'\beta)] = 0. \quad (1.1)$$

Note that we need as many moment conditions as parameters. The sample conditions are to find  $b$  such that  $\frac{1}{n} \sum_{i=1}^n x_i(y_k - x_i'b) = 0$ , whose solution is  $\widehat{\beta} = (X'X)^{-1}X'y$  (= OLS). Thus, for the moment conditions (1.1), we are lead to the least squares estimator. However, if we chose some other conditions, say  $E[x_i(y_k - x_i'\beta)^3] = 0$ , we would be lead to a different estimator - any solution of  $\frac{1}{n} \sum_{i=1}^n x_i(y_k - x_i'b)^3 = 0$ . 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.

We now give two derivations of the well-known result that

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

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

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

because  $X'(y - X\hat{\beta}) = X'y - X'X\hat{\beta} = 0$ . But

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

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

**Proof.** (ii) A minimizer of  $S(b)$  must satisfy the vector of first order conditions:

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

Therefore,

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

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

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

as required. To verify that we have found a local minimum rather than maximum it is necessary to calculate the second derivatives<sup>1</sup>

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

■

**CHARACTERIZATION OF THE SOLUTION.** Define the fitted value  $\hat{y} = X\hat{\beta}$  and the OLS residuals  $\hat{u} = y - \hat{y} = y - X\hat{\beta}$ . The OLSE  $\hat{\beta}$  solves the normal equations  $X'\hat{u} = 0$ , i.e.,

$$\sum_{i=1}^n x_{1i} \hat{u}_i = 0$$

---

<sup>1</sup>The vector derivatives follow by straightforward calculus

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

since

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

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

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



## Chapter 2

# Geometry of OLS and Partitioned Regression

We want to give a geometric interpretation to the OLS procedure. The data:  $y, x_1, \dots, x_K$ , are all elements of  $\mathbb{R}^n$ . Define the set

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

otherwise known as the column span of  $X$ . Then,  $\mathcal{C}(X)$  is a linear subspace of  $\mathbb{R}^n$  of dimension  $K$ . The OLS procedure is to find the point in  $\mathcal{C}(X)$  closest to  $y$ . Note that

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

is the Euclidean distance of  $y$  to the point  $Xb \in \mathcal{C}(X)$ . This is an old problem in geometry, which is now given a key role in abstract mathematics. The projection theorem [of Hilbert] says that there is a unique solution to the minimization problem, call it  $\hat{y}$ , which is characterized by the fact that  $y - \hat{y} = \hat{u}$  is orthogonal to  $\mathcal{C}(X)$ . That is, we can write uniquely

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

where  $\hat{y} \in \mathcal{C}(X)$  and  $\hat{u} \in \mathcal{C}^\perp(X)$  [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 with third component zero. What is the closest point for the example above with  $y = (1, 1, 1)'$ ? This is  $(1, 1, 0)' = X\hat{\beta}$ ,  $\hat{u} = (0, 0, 1)'$ . In fact  $\hat{u}$  is orthogonal to  $\mathcal{C}(X)$  (i.e., to  $X$  and any linear combination thereof), i.e.,  $\hat{u} \in \mathcal{C}^\perp(X) = (0, 0, \alpha)'$ .

In general how do we find  $\hat{y}$ ? When  $X$  is of full rank we can give a simple explicit solution;  $\hat{y} = P_X y$ , where the Projector matrix  $P_X = X(X'X)^{-1}X'$  projects onto  $\mathcal{C}(X)$ . Let  $\hat{u} = y - \hat{y} = M_X y$ , 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

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

The matrices  $P_X$  and  $M_X$  are symmetric and idempotent, i.e.,  $P_X = P_X'$  and  $P_X^2 = P_X$ . After applying  $P_X$  once you are ready in  $\mathcal{C}(X)$ . This implies that

$$P_X X = X \tag{2.1}$$

$$M_X X = 0 \tag{2.2}$$

$$P_X M_X y = 0 \text{ for all } y. \tag{2.3}$$

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

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

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

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  $\hat{y}$  and  $\hat{u}$  (but not  $\hat{\beta}$ ); for example, rescaling of the components of  $X$  does not affect the values of  $\hat{y}$  and  $\hat{u}$ .

$$y \text{ on } (x_1, x_2, x_3)(1)$$

$y$  on  $(x_1 + x_2, 2x_2 - x_3, 3x_1 - 2x_2 + 5x_3)$ (2)

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

(1) and (2) yield the same  $\hat{y}$ ,  $\hat{u}$ .

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.

Also, note that even 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  $\hat{y}$  and residual  $\hat{u}$  are uniquely defined, but there is no unique coefficient vector  $\hat{\beta}$ . This case will be discussed below under the title 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  $\hat{\beta}_1$  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.,  $X_1'X_2 = 0$ , then  $P_X = P_{X_1} + P_{X_2}$ . This can be verified algebraically, but also should be obvious geometrically. In this case, write

$$\hat{y} = X\hat{\beta} = P_X y = P_{X_1} y + P_{X_2} y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2.$$

This just says that if  $X_1$  and  $X_2$  were orthogonal, then we could get  $\hat{\beta}_1$  by regressing  $y$  on  $X_1$  only, and  $\hat{\beta}_2$  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:

1.  $(X_1, X_2)$  and  $(M_2 X_1, X_2)$  span the same space. This follows because  $X_1 = M_2 X_1 + P_2 X_1$ , where  $\mathcal{C}(P_2 X_1) \subset \mathcal{C}(X_2)$ . Therefore,  $\mathcal{C}(M_2 X_1, X_2) = \mathcal{C}(X_1, X_2)$ .
2.  $M_2 X_1$  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  $\hat{y}$  and  $\hat{u}$ , 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

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

where  $\hat{C} = \hat{\beta}_2 + (X_2'X_2)^{-1}X_2'X_1\hat{\beta}_1$ . So the coefficient on  $M_2X_1$  is the original  $\hat{\beta}_1$ , while that on  $X_2$  is some combination of  $\hat{\beta}_1$  and  $\hat{\beta}_2$ . 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 can get  $\hat{\beta}_1$  by regressing  $y$  [or  $M_2y$  equivalently] on  $M_2X_1$  only, i.e.,

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

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  $X_2 = (1, 1, \dots, 1)'$  =  $i$ , then

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

and

$$M_2x_{1n \times 1} = x_1 - \frac{1}{n} \sum_{i=1}^n x_{1i} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} x_{1i} - \bar{x}_1 \\ \vdots \\ x_{1n} - \bar{x}_1 \end{pmatrix}.$$

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

## 2.0.2 Goodness of Fit

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

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

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,

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

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

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

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

$$R_3^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3).$$

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

**Theorem 1** *When an intercept is included, all three measures are the same.*

**Proof.** [(2) = (3)] *Since an intercept is included, we have  $\sum_{i=1}^n \hat{u}_i = 0$ , which implies that  $\bar{\hat{y}} = \bar{y}$ .*

*Therefore,*

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

*because  $\sum_{i=1}^n \hat{u}_i(\hat{y}_i - \bar{y}) = 0$ . ■*

**Proof.** [(1) = (3)] *Similarly,*

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

■

## COMMENTS

1. 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.
2. If  $y = \alpha + \beta x + u$ , then  $R^2$  is the squared sample correlation between  $y$  and  $x$ .
3. The  $R^2$  is invariant to some changes of units. If  $y \mapsto ay + b$  for any constants  $a, b$ , then  $\hat{y}_i \mapsto a\hat{y}_i + b$  and  $\bar{y} \mapsto a\bar{y} + b$ , so  $R^2$  is the same in this case. Clearly, if  $X \mapsto XA$  for a nonsingular matrix  $A$ , then  $\hat{y}$  is unchanged, as is  $y$  and  $\bar{y}$ .
4.  $R^2$  always increases with addition of variables. With  $K = n$  we can make  $R^2 = 1$ .
5. Theil's adjusted  $R^2$  is defined as follows

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

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

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

It follows that

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

This measure allows some trade-off between fit and parsimony. We will return to this later when we discuss model selection.

## Chapter 3

# 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  $\hat{\beta}$  is that it is linear in  $y$ , i.e., there exists a matrix  $C$  not depending on  $y$  such that

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

This property makes a lot of calculations simple. We want to evaluate how  $\hat{\beta}$  varies across hypothetical repeated samples.

*Fixed Design.* First,

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

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

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

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

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

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

*Random Design.* We first condition on  $X$ ; 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  $\beta$ . Note that this calculation uses the important property called “The Law of Iterated Expectation”, which says (more generally) that  $E(Y|I_1) = E[E(Y|I_2)|I_1]$ , whenever  $I_1 \subseteq I_2$  for two information sets  $I_1, I_2$ .

As for the variance, we use an 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

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

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

The variance for 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 then

$$\text{Var}[\widehat{\beta}_1] = (X_1' M_2 X_1)^{-1} X_1' M_2 E \varepsilon \varepsilon' M_2 X_1 (X_1' M_2 X_1)^{-1} = \sigma^2 (X_1' M_2 X_1)^{-1}.$$

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

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

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.

$$(A4) \quad y \sim N(X\beta, \sigma^2 I) \text{ or}$$

$$(A4r) \quad y|X \sim N(X\beta, \sigma^2 I).$$

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  $\beta$  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.

### 3.0.3 Optimality

There are many estimators of  $\beta$ . 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  $\widetilde{\beta} = \bar{y}/\bar{x}$  and  $\bar{\beta} = \sum_{i=1}^n y_i/x_i$  [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  $\theta$ . We would like a method that does well for all  $\theta$ , 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  $\theta$  moves away, the MSE increases quadratically.

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. 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; but different functions will rank estimators differently [see the example above].

Also note that no estimator can dominate uniformly across  $\theta$  according to MSE because it would have to beat all constant estimators 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  $\max_{\theta} \text{tr}(MSE)$ , 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.

Instead, we reduce the class of allowable estimators. Specifically, we restrict attention to linear unbiased estimates, i.e.,

$$\tilde{\beta} = Ay$$

for some fixed matrix A such that

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

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

**Theorem 2** *Gauss Markov* Assume that A1–A3 hold. The OLS estimator  $\hat{\beta}$  is Best Linear Unbiased (BLUE), i.e.,

$$\text{Var}(\hat{\beta}) \leq \text{Var}(\tilde{\beta})$$

for any other LUE.

**Proof.**  $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ ;  $\text{Var}(\tilde{\beta}) = \sigma^2AA'$

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

■

Comments:

- (1) Makes no assumption about the distribution of the errors; it only assumes 0 mean and  $\sigma^2 I$  variance.
- (2) Result only compares *linear* estimators; it says nothing about for example  $\sum_{i=1}^n |y_i - \beta x_i|$ .
- (3) Result only compares *unbiased* estimators [biased estimators can have 0 variances].
- (4) There are extensions to consider affine estimators  $\tilde{\beta} = a + Ay$  for vectors  $a$ . There are also equivalent results for the invariant quantity  $\hat{y}$ .

Finally, we state without proof the well known result

**theorem 3** Cramèr–Rao Under A1-A4,  $\hat{\beta}$  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.

# Chapter 4

## Hypothesis Testing

We observe certain data  $(y, X)$ . 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  $H_0 \subseteq \mathcal{F}$ . This arises as:

- (1) Prediction of scientific theory, e.g., interest elasticity of demand for money is zero; the gravitational constant is 9.
- (2) Absence of some structure, e.g., independence of error term over time, homoskedasticity etc.
- (3) Pretesting (used as part of model building process).

Our purpose here is to provide a method for interpreting what data say about the hypothesis. 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 (one or more restriction on parameters of  $f$ ) hypotheses. We shall also introduce  $H_A$ , 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.

### Examples

- (a) 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.

(b) 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.

(c) 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  $\gamma = 0$ .

#### 4.0.4 General Notations

Want to test  $H_0$  vs.  $H_A$ ? 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, while 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. Unfortunately, these are usually in conflict. The usual approach is to fix the Type I error and then try to do the best in terms of the Type II error.

We choose  $\alpha \notin [0, 1]$  called the size of the test [magnitude of Type I error]. Let  $T(\text{data})$  be a test statistic, typically scalar valued. Then, find critical region  $C_\alpha$  of size  $\alpha$  such that

$$\Pr[T \notin C_\alpha | H_0] = \alpha. \tag{4.1}$$

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  $C_\alpha$  is easy to find. Define also Power of test:

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

It is desirable, *ceteris paribus*, to have a test which maximizes power for any given size.

An unbiased test satisfies  $\beta \geq \alpha$ , i.e., you are more likely to reject under the alternative hypothesis than the null. This is a desirable property and basically a pretty minimal requirement; all tests we examine have this property.

There are some well known criticisms of the standard hypothesis testing approach:

- It really is very primitive decision-making. The outcome is only one of two things. There is an asymmetric way in which the null hypothesis is treated versus the alternative.
- $\alpha$  is arbitrary. Some argue that  $\alpha$  should be made a function of sample size, because in practice as  $n$  gets large almost any hypothesis is rejected. In reality hypotheses are neighbourhoods not points.

A more informal approach is to report the  $p$ -value of a test  $T$

$$\alpha_{\text{obs}} = \Pr[T \geq T_{\text{obs}} | H_0].$$

In two-sided case, take  $|T| \geq |T_{\text{obs}}|$ . Low  $\alpha_{\text{obs}}$  is evidence against the null hypothesis. More information is conveyed by this approach - the reader can decide on his or her own  $\alpha$  and carry out the test.

## 4.0.5 Examples

Hypothesis Testing in Linear Regression:  $E(y) = X\beta$ ,  $\text{Var}(y) = \sigma^2 I$ ,  $y \sim N(X\beta, \sigma^2 I)$ .

Single (Linear) Hypothesis:  $c'\beta = \gamma_{\text{scalar}}$ , e.g.,  $\beta_2 = 0$  ( $t$ -test).

Multiple (Linear) Hypothesis:  $R_{q \times K}\beta_{K \times 1} = r_{q \times 1}$ ,  $q \leq K$ , e.g.,  $\beta_2 = \beta_3 = \dots = \beta_K = 0$ .

Single Non-linear Hypothesis:  $\beta_1^2 + \beta_2^2 + \dots + \beta_K^2 = 1$ .

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

### 4.0.6 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,

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

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

$$\begin{aligned} \hat{\sigma}_{mle}^2 &= \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n} \\ s^2 &= \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n - K} \end{aligned}$$

The first estimate is the maximum likelihood estimator of  $\sigma^2$ , which can be easily verified. The second estimate is a modification of the MLE, which happens to be unbiased.

Now define the test statistic

$$T = \frac{c'\hat{\beta} - \gamma}{s \sqrt{c'(X'X)^{-1}c}}.$$

**Theorem 4** Under  $H_0$ ,  $T \sim t(n - K)$ .

**Proof.** We show that:

- (1)  $\frac{n-K}{\sigma^2} s^2 \sim \chi_{n-K}^2$
- (2)  $s^2$  and  $c'\hat{\beta} - \gamma$  are independent.

This establishes the theorem by the defining property of a t-random variable, see for example [].

Recall that

$$\frac{\varepsilon'\varepsilon}{\sigma^2} = \sum_{i=1}^n \left(\frac{\varepsilon_i}{\sigma}\right)^2 \sim \chi_n^2.$$

But  $\hat{\varepsilon}$  are residuals that use  $K$  parameter estimates. Furthermore,  $\hat{\varepsilon}'\hat{\varepsilon} = \varepsilon' M_X \varepsilon$  and

$$\begin{aligned} E[\varepsilon' M_X \varepsilon] &= E[\text{tr } M_X \varepsilon \varepsilon'] = \text{tr } M_X E(\varepsilon \varepsilon') \\ &= \sigma^2 \text{tr } M_X = \sigma^2 (n - \text{tr } P_X) \\ &= \sigma^2 (n - K) \end{aligned}$$

$$\text{tr}(X(X'X)^{-1}X') = \text{tr} X'X(X'X)^{-1} = \text{tr} I_K = K.$$

These calculations show that  $E\widehat{\varepsilon}'\widehat{\varepsilon} = n - K$ , which suggests that  $\widehat{\varepsilon}'\widehat{\varepsilon}$  cannot be  $\chi_n^2$  [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  $\Lambda$  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 distribution as  $\varepsilon$  [it has the same mean and variance, which is sufficient to determine the normal distribution]. Therefore,

$$\widehat{\varepsilon}'\widehat{\varepsilon} = \sum_{i=1}^{n-K} \varepsilon_i^{*2} \quad (4.2)$$

for some i.i.d. standard normal random variables  $\varepsilon_i^*$ . Therefore, (4.2) is  $\chi_{n-K}^2$  by the definition of a chi-squared random variable.

Furthermore, under  $H_0$ ,  $c'\widehat{\beta} - \gamma = c'(X'X)^{-1}X'\varepsilon$  is uncorrelated with  $\widehat{\varepsilon} = M_X\varepsilon$ , since

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

Under normality, uncorrelatedness is equivalent to independence. ■

Can now base test of  $H_0$  on

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

using the  $t_{n-k}$  distribution for an exact test under normality. Both one-sided and two-sided alternatives, i.e., reject if  $|T| \geq t_{n-K}(\alpha/2)$  or  $T \geq t_{n-K}(\alpha)$ .

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  $y_i = \beta_1 + \beta_2x_i + \beta_3z_i + u_i$ , 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  $y_i - z_i = \beta_1 + \beta_2(x_i - z_i) + \beta_3z_i + u_i$ .

### 4.0.7 Test of a Multiple Linear Hypothesis

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

$$F = (R\hat{\beta} - r)' [s^2 R(X'X)^{-1} R']^{-1} (R\hat{\beta} - r) / q = \frac{(R\hat{\beta} - r)' [R(X'X)^{-1} R']^{-1} (R\hat{\beta} - r) / q}{(n - K) s^2 / (n - K)}. \quad (4.3)$$

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

$$F = \frac{\chi_q^2 / q}{\chi_{n-K}^2 / (n - K)} \sim F(q, n - K)$$

under  $H_0$ . The rule is that if

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

then reject  $H_0$  at level  $\alpha$ . Note that we can only test against a two-sided alternative  $R\beta \neq r$  because we have squared value in (4.3).

#### Examples

(1) Standard  $F$ -test, which is outputted from computer, is of the hypothesis

$$\beta_2 = 0, \dots, \beta_K = 0,$$

where the intercept  $\beta_1$  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.

(2) Structural Change. Null hypothesis is  $y = X\beta + u$ . Alternative is

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

where  $n = n_1 + n_2$ . Let

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

Then, we can write the alternative regression as

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

Consider the null hypothesis  $H_0 : \beta_1 = \beta_2$ . Let  $R_{K \times 2K} = [I_K \ : \ -I_K]$ . 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,  $c' \hat{\beta} \pm t_{\alpha/2}(n - K) s \{c'(X'X)^{-1}c\}^{1/2}$  (or  $z_{\alpha/2}$ ). Can also construct multivariate confidence intervals.



## Chapter 5

# Test of Multiple Linear Hypothesis Based on fit

Idea behind  $F$  test is that under  $H_0$ ,  $R\hat{\beta} - r$  should be stochastically small, but under the alternative hypothesis it will not be so. An alternative approach is based on fit. If we estimate  $\beta$  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].

(1) Unrestricted regression:

$$\min_b (y - Xb)'(y - Xb) \mapsto \hat{\beta}, \quad \hat{u} = y - X\hat{\beta}, \quad Q = \hat{u}'\hat{u}.$$

(2) Restricted regression:

$$\min_b (y - Xb)'(y - Xb) \text{ s.t. } Rb = r \mapsto \beta^*, \quad u^* = y - X\beta^*, \quad Q^* = u^{*'}u^*.$$

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 5** Under  $H_0$ ,

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

**Proof.** We show that

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

Then, since  $s^2 = Q/(n - K)$  the result is established.

Now,  $\beta^*$  and  $\lambda^*$  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

$$-X'y + X'X\beta^* + R'\lambda^* = 0(1)$$

$$R\beta^* = r.(2)$$

Now, from (1)

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

which implies that

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

and

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

Therefore,

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

and

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

This gives the restricted least squares estimator in terms of the restrictions and the unrestricted least squares estimator.

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  $\lambda^*$  obtained above and cancelling out, we get

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

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 leads 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:

(a) Zero restrictions

$$\beta_2 = \dots = \beta_K = 0$$

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

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-k)}. \quad (5.1)$$

(b) Structural change. Allow coefficients to be different in two periods. Partition

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

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

Null is of no structural change, i.e.,  $H_0 : \beta_1 = \beta_2$ ,  $R = (I \quad -I)$ .

Consider the more general linear restriction

$$\beta_1 + \beta_2 - 3\beta_4 = 1$$

$$\beta_6 + \beta_1 = 2.$$

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

$$X = \begin{pmatrix} X_1 & X_2 \\ n \times (k-q) & n \times q \end{pmatrix} ; \quad R = \begin{pmatrix} R_1 & R_2 \\ q \times (k-q) & q \times q \end{pmatrix} ; \quad \beta = \begin{pmatrix} \beta_1 \\ (k-q) \times 1 \\ \beta_2 \\ q \times 1 \end{pmatrix},$$

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,

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

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

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  $\beta_1^*$ , and then define  $\beta_2^* = R_2^{-1}(r - R_1\beta_1^*)$ . We then define  $u^* = y - X_1\beta_1^* - X_2\beta_2^*$  and  $Q^*$  accordingly. Check with special cases above.

# Chapter 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.,

$$H_0 : \beta_1 = \beta_2 = \beta \tag{1}.$$

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

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

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.,

$$H_0 : \alpha_1 = \alpha_2 = \alpha \quad (2).$$

Restricted regression  $(\alpha, \beta_1, \beta_2)$

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

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  $n_2 < K$ . Restricted regression is ok. The unrestricted regression runs into problems in second period because  $n_2$  is too small. In fact,  $\hat{u}_2 \equiv 0$ . Degrees of freedom lost are  $n_2$  not  $K$ . Thus

$$F = \frac{(Q^* - Q)/n_2}{Q/(n_1 - K)} \sim F(n_2, n_1 - K).$$

# Chapter 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  $\theta$  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 maximum likelihood estimator is denoted by  $\hat{\theta}$ , while the restricted MLE is denoted by  $\theta^*$ , [this is maximizes  $L$  subject to the restrictions  $R\theta - r = 0$ ]. Now define the following test statistics:

$$\begin{aligned} \text{LR} & : 2 \left[ \log \frac{L(\hat{\theta})}{L(\theta^*)} \right] = 2\{\log L(\hat{\theta}) - \log L(\theta^*)\} \\ \text{Wald} & : (R\hat{\theta} - r)' \left\{ R \left[ -\frac{\partial^2 \log L}{\partial \theta \partial \theta'} \Big|_{\hat{\theta}} \right]^{-1} R' \right\}^{-1} (R\hat{\theta} - r) \\ \text{LM} & : \frac{\partial \log L}{\partial \theta} \Big|_{\theta^*}' \left[ -\frac{\partial^2 \log L}{\partial \theta \partial \theta'} \Big|_{\theta^*} \right]^{-1} \frac{\partial \log L}{\partial \theta} \Big|_{\theta^*}. \end{aligned}$$

The Wald test only requires computation of the unrestricted estimator, while 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. The LR test has certain advantages; but computationally it is the most demanding.

In the linear regression case,  $\theta = (\beta, \sigma^2)$ , and the restrictions only apply to  $\beta$ , so that  $R\beta = r$ . Furthermore, we can replace the derivatives with respect to  $\theta$  by derivatives with respect to  $\beta$  only [this requires an additional justification, which we will not discuss here].

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

$$W = (R\hat{\beta} - r)' [R(X'X)^{-1}R'\hat{\sigma}^2]^{-1} (R\hat{\beta} - r) = \frac{Q^* - Q}{(Q/n)}, \quad (7.1)$$

where  $\hat{\sigma}^2 = Q/n$  is the MLE of  $\sigma^2$ , and so is very similar to the  $F$ -test apart from the use of  $\hat{\sigma}^2$  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

$$LM = \frac{u^{*'} X}{\sigma^{*2}} \left\{ \frac{X' X}{\sigma^{*2}} \right\}^{-1} \frac{X' u^*}{\sigma^{*2}} \quad (7.2)$$

where  $\sigma^{*2} = Q^*/n$ . This can be rewritten in the form

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

where  $\lambda^*$  is the vector of Lagrange Multipliers evaluated at the optimum [Recall that  $X' u^* = R' \lambda^*$ ].

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.

Likelihood Ratio

$$\begin{aligned} \log L(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} \hat{u}' \hat{u} = -\frac{n}{2} \log 2\pi - \frac{1}{2\hat{\sigma}^2} \log \frac{\hat{u}' \hat{u}}{n} - \frac{n}{2} \\ \log L(\beta^*, \sigma^{*2}) &= -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^{*2} - \frac{1}{2\sigma^{*2}} u^{*'} u^* = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \frac{u^{*'} u^*}{n} - \frac{n}{2}. \end{aligned}$$

These two lines follow because  $\hat{\sigma}^2 = \hat{u}' \hat{u}/n$  and  $\sigma^{*2} = u^{*'} u^*/n$ . Therefore,

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

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

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

If we knew the exact distribution of any of them we can obtain the 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.



## Chapter 8

# Omission of Relevant Variables, Inclusion of Irrelevant Variables, and Model Selection

We now work towards a consideration of model selection, i.e., which variables or how many variables to include in a regression. We shall assume that there is a true model, which of course we may or may not know. We have now to consider the effects of including too many variables and of including too few variables.

### 8.0.8 Omission of Relevant Variables

Suppose

$$y = X_1\beta_1 + X_2\beta_2 + u,$$

but we regress  $y$  on  $X_1$  only. Then,

$$\begin{aligned}\hat{\beta}_1 &= (X'X)^{-1}X'y \\ &= (X'_1X_1)^{-1}(X_1\beta_1 + X_2\beta_2 + u) \\ &= \beta_1 + (X'_1X_1)^{-1}X'_1X_2\beta_2 + (X'_1X_1)^{-1}X'_1u,\end{aligned}$$

so that

$$E(\widehat{\beta}_1) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 = \beta_1 + \beta_{12},$$

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  $\beta_2$  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  $X_2$  are 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

$$s^2 = \frac{y'M_1y}{n - K_1} = \frac{(X_2\beta_2 + u)'M_1(X_2\beta_2 + u)}{n - K_1},$$

which has expectation

$$\sigma^2 + \frac{\beta_2'X_2'M_1X_2\beta_2}{n - K_1} \geq \sigma^2,$$

since  $M_1$  is a positive semi-definite matrix. Therefore, the estimated variance of  $\widehat{\beta}_1$  is upwardly biased.

Note that 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

- (1) Seasonality
- (2) Dynamics
- (3) 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.

### 8.0.9 Inclusion of irrelevant variables

Suppose now that

$$y = X_1\beta_1 + u,$$

but we regress  $y$  on both  $X_1$  and  $X_2$ . Then

$$\begin{aligned}\widehat{\beta}_1 &= (X_1' M_2 X_1)^{-1} X_1' M_2 y = \beta_1 + (X_1' M_2 X_1)^{-1} X_1' M_2 u \\ E(\widehat{\beta}_1) &= \beta_1 \text{ all } \beta_1 \\ \text{Var}(\widehat{\beta}_1) &= \sigma^2 (X_1' M_2 X_1)^{-1}.\end{aligned}$$

Compare this with the variance of  $y$  on  $X_1$ , which is only  $\sigma^2 (X_1' X_1)^{-1}$ . Now

$$X_1' X_1 - X_1' M_2 X_1 = X_1' P_2 X_1 \geq 0 \Rightarrow (X_1' X_1)^{-1} - (X_1' M_2 X_1)^{-1} \leq 0.$$

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  $R\beta = r$ . 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.

### 8.0.10 Model Selection

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

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

$$PC_j = \frac{\hat{u}'_j \hat{u}_j}{n - K_j} \left( 1 + \frac{K_j}{n} \right) \quad (2)$$

$$AIC_j = \ln \frac{\hat{u}'_j \hat{u}_j}{n} + \frac{2K_j}{n} \quad (3)$$

$$BIC_j = \ln \frac{\hat{u}'_j \hat{u}_j}{n} + \frac{K_j \log n}{n}. \quad (4)$$

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  $\hat{u}'_j \hat{u}_j / (n - K_j)$ .

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.

#### PROBLEMS

1.  $\mathcal{M}$  may be large and  $2^K$  regressions infeasible.
2. True model may not be in  $\mathcal{M}$ , but procedure is guaranteed to find a best model (data mining).
3. Other criteria are important, especially for nonexperimental data.
  - (a) Consistency with economic theory elasticities the right sign? Demand slopes down?
  - (b) 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.
  - (c) Residuals should be approximately random, i.e., pass diagnostic checks for serial correlation, heteroskedasticity, nonlinearity, etc.
  - (d) How well model performs out-of-sample. (Often used in time series analysis.)
  - (e) 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.

**Theorem 6**  $\overline{R}^2$  falls (rises) when the deleted variable has  $t > (<) 1$

# Chapter 9

## Data Problems

### 9.0.11 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. Specifically, one can provide greater generality by considering transformations of the variables, either one at a time or in pairs.

Dummy variables       $\text{Wages} = \alpha + \beta ed + \gamma UNION + u$

Interaction terms       $\text{Wages} = \alpha + \beta ed + \gamma ab + \delta ed \cdot ab + u$

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

Log transformation     $\ln(\text{wages})$

Logit transformation    $\log \frac{fs}{1-fs} = \alpha + \beta inc + u, fs \in [0, 1], \text{logit}(fs) \in (-\infty, \infty).$

These are all linear in the *parameters* model, i.e., can write  $y = X\beta + u$  for some  $X$ , some  $\beta$ , 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.,  $x_t = t, t = 1, 2, \dots, T$ . 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

$$\begin{aligned} \alpha_1 + \beta_1 t_1^* &= \alpha_1 + \gamma_1 + (\beta_1 + \delta_1) t_1^* \\ \alpha_1 + \beta_1 t_2^* + \gamma_1 + \delta_1 t_2^* &= \alpha_1 + \gamma_1 + (\beta_1 + \delta_1) t_1^* + \gamma_2 + \delta_2 t_2^*, \end{aligned}$$

which implies that  $\gamma_1 = -\delta_1 t_1^*$  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

##### (1) Box-Cox

$$y = \alpha + \beta \frac{x^\lambda - 1}{\lambda} + u \quad \text{as } \lambda \rightarrow 0, \quad \frac{x^\lambda - 1}{\lambda} \rightarrow \ln(x) \quad \text{as } \lambda \rightarrow 1, \quad \frac{x^\lambda - 1}{\lambda} \rightarrow x - 1$$

##### (2) 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.

##### (3) CES production function

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

Methods for treating these models will be considered below.

### 9.0.12 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  $\hat{\beta} = (X'X)^{-1}X'y$ . Solution: Find a minimal (not unique) basis  $X^*$  for  $\mathcal{C}(X)$  and do least squares.

*Example:* Seasonal dummies

$$D1 = 1 \text{ if Quarter 1, } 0 \text{ else}$$

$$D2 = 1 \text{ if Quarter 2, } 0 \text{ else}$$

$$D3 = 1 \text{ if Quarter 3, } 0 \text{ else}$$

$$D4 = 1 \text{ if Quarter 4, } 0 \text{ else.}$$

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,  $C2 + C3 + C4 + C5 = C1$  for all observations. Solution:

(1) Drop  $D4$ , and run  $y = \alpha + \beta_1 D1 + \beta_2 D2 + \beta_3 D3 + u$

(2) Drop intercept, and run  $y = \gamma_1 D2 + \gamma_2 D2 + \gamma_3 D3 + \gamma_4 D4 + u$ .

Gives same  $\hat{y}$  and  $\hat{u}$ , 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’.

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

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

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.

### 9.0.14 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: (a) On dependent variable (or on error term); (b) 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  $\geq 5$  years.

We look at ‘ignorable’ case: e.g., time series; e.g., errors in recording add to zero.

(a) 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

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

We can then recompute the least squares estimate of  $\beta$  using ‘all the data’

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

however, some simple algebra reveals that there is no new information in  $\hat{y}_B$ , and in fact  $\hat{\beta} = \hat{\beta}_A$ . Start from

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

and pre-multiply both sides by  $X'X = (X'_A X_A + X'_B X_B)$ , we have

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

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

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

(b) Now suppose that we have some missing  $X$ . For example,  $X = (x, z)$ , and  $x_B$  is missing, i.e., we observe  $(x_A, x_A, y_A)$  and  $(z_B, y_B)$ . Suppose now we predict  $x_b$  by regressing  $x_A$  on  $z_A$

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

Then regress  $y$  on

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

This sometimes improves matters, but sometimes does not! The answer depends on relationship between  $x$  and  $z$ . In any case, the effect of  $x$  is not better estimated; the effect of  $z$  maybe improved.

# Chapter 10

## Asymptotic Theory I

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 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  $x_\infty$ , denoted  $\lim_{n \rightarrow \infty} x_n = x_\infty$ , 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,

$$X_n \xrightarrow{P} X, \tag{10.1}$$

if for all  $\epsilon > 0$ ,

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

$X$  could be a constant or a random variable. We sometimes write  $X = \underset{n \rightarrow \infty}{p} \lim X_n$ .

DEFINITION: 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  $\theta_n \xrightarrow{a.s.} \theta$ , if

$$\Pr[\theta_n \rightarrow \theta] = 1.$$

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

$$X_n \xrightarrow{D} X, \quad (10.2)$$

if for all  $x$ ,

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

Specifically, we often have

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \sigma^2).$$

Note that convergence in probability is stronger than convergence in distribution, but they are equivalent when  $X$  is a constant (i.e., not random).

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

$$\lim_{n \rightarrow \infty} E[|X_n - X|^2] = 0. \quad (10.3)$$

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

$$E[|X_n - X|^2] = E[|X_n - EX_n|^2] + |EX_n - X|^2 = \text{Var}(X_n) + |EX_n - X|^2,$$

and it is necessary and sufficient that  $EX_n \rightarrow X$  and  $\text{Var}(X_n) \rightarrow 0$ .

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

$$\Pr[|X_n - X| > \varepsilon] \leq \frac{E[|X_n - X|^2]}{\varepsilon^2}. \quad (10.4)$$

Almost sure convergence implies convergence in probability, but there is no necessary relationship between almost sure convergence and convergence in mean square.

We now come to the two main theorems, the Law of Large Numbers (LLN), and the Central Limit Theorem (CLT).

**Theorem 7** (Kolmogorov) *Let  $X_1, \dots, X_n$  is 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  $E(|X_i|) < \infty$ .*

**Theorem 8** (*Lindeberg-Levy*) Let  $X_1, \dots, X_n$  be i.i.d. with  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2$ . Then

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

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.

**Theorem 9** (*Lindeberg-Feller*) Let  $X_1, \dots, X_n$  be independent random variables with  $E(X_i) = 0$  and  $\text{Var}(X_i) = \sigma_i^2$ . Suppose also that Lindeberg's condition

$$\frac{1}{\sum_{i=1}^n \sigma_i^2} \sum_{i=1}^n E \left[ X_i^2 1 \left( X_i^2 > \epsilon \sum_{j=1}^i \sigma_j^2 \right) \right] \rightarrow 0 \quad \text{for all } \epsilon > 0 \quad (10.5)$$

holds. Then

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

A sufficient condition for (10.5) is that  $E[|X_i|^3] < \infty$ .

The following results are very useful in conjunction with the LLN and CLT:

**Theorem 10** *Mann-Wald*. 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  $g(X_n) \xrightarrow{P} g(\alpha)$ .

**Theorem 11** *Slutsky*. If  $X_n \xrightarrow{D} X$ ,  $y_n \xrightarrow{P} \alpha$ , then: (i)  $X_n + y_n \xrightarrow{D} X + \alpha$ ; (ii)  $X_n y_n \xrightarrow{D} \alpha X$ ; and (iii)  $X_n / y_n \xrightarrow{D} X / \alpha$ , provided  $\alpha \neq 0$ .

Finally, we point out that when dealing with a vector  $X_n = (X_{n1}, \dots, X_{nk})'$ , we have the result that

$$\|X_n - X\| \xrightarrow{P} 0,$$

where  $\|x\| = (x'x)^{1/2}$  is Euclidean norm, if and only if

$$|X_{nj} - X_j| \xrightarrow{P} 0$$

for all  $j = 1, \dots, k$ . 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  $|X_{nj} - X_j| < \varepsilon$  for each  $j$ . As regards convergence in distribution, we have the Cramers Theorem

**Theorem 12** *A vector  $X_n$  converges in distribution to a normal vector  $X$  if and only if  $c'X_n$  converges in distribution to  $c'X$  for every vector  $c$ .*

We are now able to establish some results about the large sample properties of the least squares estimator. In the i.i.d. case, the result is very simple. If we assume that  $x_i, \varepsilon_i$  are i.i.d. with  $E(\varepsilon_i|x_i) = 0$  with probability one, then provided  $E[x_i x_i'] < \infty$  and  $E[||x_i \varepsilon_i||] < \infty$ , we have  $\widehat{\beta} \xrightarrow{P} \beta$ . These 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 more general sampling schemes including trending variables.

**Theorem 13** *Suppose that A0-A2 hold and that*

$$\lambda_{\min}(X'X) \rightarrow \infty \text{ as } n \rightarrow \infty. \quad (\star)$$

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

**Proof.** First,

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

Since  $\text{Var}(\widehat{\beta}) = \sigma^2(X'X)^{-1}$  but

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

and provided  $(\star)$  is true,  $\text{Var}(\widehat{\beta}) \rightarrow 0$ . ■

If we do have a random design then the conditions and conclusion should be interpreted as holding with probability one in the conditional distribution given  $X$ .

EXAMPLE:

$$\begin{aligned}
 X &= \begin{pmatrix} 1 & 1^{1/2} \\ \vdots & 2^{1/2} \\ 1 & n^{1/2} \end{pmatrix} \\
 X'X &= \begin{pmatrix} n & \sum i^{1/2} \\ \sum i^{1/2} & \sum_{i=1}^n i \end{pmatrix} \approx \begin{bmatrix} n & \frac{2}{3}n^{3/2} \\ \frac{2}{3}n^{3/2} & \frac{n^2}{2} \end{bmatrix} \\
 (X'X)^{-1} &= \begin{pmatrix} n^2/2 & -\frac{2}{3}n^{3/2} \\ -\frac{2}{3}n^{3/2} & n \end{pmatrix} / \left( \frac{n^3}{2} - \frac{4}{9}n^3 \right) \\
 &= 18 \begin{pmatrix} \frac{n^2}{2} & -\frac{2}{3}n^{3/2} \\ -\frac{2}{3}n^{3/2} & n \end{pmatrix} / n^3 \rightarrow 0 \text{ as } n \rightarrow \infty.
 \end{aligned}$$

Note that it is not enough that each element of  $X'X \rightarrow \infty$ , because take

$$X'X = \begin{pmatrix} n & n \\ n & n \end{pmatrix}.$$

We now turn to the limiting distribution of the least squares estimator. In the i.i.d. case, the result is very simple.

**Theorem 14** *If  $x_i, \varepsilon_i$  are i.i.d. with  $\varepsilon_i$  independent of  $x_i$ , then provided  $E(\varepsilon_i^2) = \sigma^2, E[x_i x_i'] < \infty$ , we have*

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

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

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

for some weights depending only on  $X$ . That is, the source of the heterogeneity is the fixed regressors. In this case, a sufficient condition for the 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

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

This is a so-called negligibility requirement, which means that no one of the weights dominates every other term. Therefore,

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

provided  $X'X/n \rightarrow M > 0$  and the following negligibility condition holds:

$$\max_{1 \leq i \leq n} x_i (X'X)^{-1} x_i' \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (10.6)$$

Actually it suffices for the diagonal elements of this matrix to converge to zero. The condition is usually satisfied. Suppose  $k = 1$ , then the condition (10.6) is

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

For example, if  $x_i = i$ ,

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

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.

# Chapter 11

## Asymptotic Theory II

We here consider further applications of the tools of asymptotic theory to standard errors and test statistics. This involves combining the Mann-Wald theorem and Slutsky theorem with the law of large numbers and central limit theorems. In the sequel we shall use the Order notation:

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

$$X_n = O_p(\delta_n) \text{ if for all } K, \lim_{n \rightarrow \infty} \Pr \left[ \left| \frac{X_n}{\delta_n} \right| > K \right] < 1.$$

The latter means that  $X_n$  is of no larger order than  $\delta_n$  [we say that  $X_n$  is *stochastically bounded*], while the first one is stronger and says that  $X_n$  is of smaller order than  $\delta_n$ . 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:

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

We first consider the standard error. We have

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

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

**Theorem 15** Suppose that  $u_i$  are i.i.d. with finite fourth moment, and that the regressors are from a fixed design and satisfy  $(X'X/n) \rightarrow M$ , where  $M$  is a positive definite matrix. Then

$$\sqrt{n}(s^2 - \sigma^2) \xrightarrow{D} N(0, \text{Var}[u^2 - \sigma^2]).$$

PROOF. Note that

$$\text{var}\left(\frac{u'X}{\sqrt{n}}\right) = \sigma^2 \frac{X'X}{n},$$

which stays bounded, so that  $(u'X/\sqrt{n}) = O_p(1)$ . Therefore the second term is  $O_p(n^{-1})$ . Therefore,

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

What about the asymptotic distribution of  $s^2$ ?

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

provided the second moment of  $(u_i^2 - \sigma^2)$  exists, which it does under our assumption. When the errors are normally distributed,  $\text{Var}[u^2 - \sigma^2] = 2\sigma^4$ . ■

Now what about the t statistic:

$$t = \frac{\sqrt{nc'}\hat{\beta}}{s\sqrt{c'\frac{(X'X)^{-1}}{n}c}}.$$

**Theorem 16**

$$\begin{aligned}
t &= \frac{\sqrt{nc'}\hat{\beta}}{\sigma\sqrt{c'M^{-1}c}} + o_p(1) \\
&\xrightarrow{D} \frac{N(0, \sigma^2 c' M^{-1} c)}{\sigma\sqrt{c'M^{-1}c}} \equiv N(0, 1) \quad \text{under } H_0.
\end{aligned}$$

As for the Wald statistic

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

**Theorem 17** 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  $(X'X/n) \rightarrow M$ , where  $M$  is a positive definite matrix. Then,

$$W \xrightarrow{D} N(0, \sigma^2 R M^{-1} R') \cdot [\sigma^2 R M^{-1} R']^{-1} \cdot N(0, \sigma^2 R M^{-1} R') = \chi_q^2.$$

### 11.0.15 The delta method

**Theorem 18** Suppose that  $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \Sigma)$  and that  $f$  is a continuously differentiable function. Then

$$\sqrt{n}(f(\hat{\theta}) - f(\theta)) \xrightarrow{D} N\left(0, \frac{\partial f}{\partial \theta} \Sigma \frac{\partial f}{\partial \theta'}\right).$$

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

$$\begin{aligned} f(\hat{\theta}) &= f(\theta) + (\hat{\theta} - \theta)f'(\theta^*), \text{ i.e.,} \\ \sqrt{n}(f(\hat{\theta}) - f(\theta)) &= f'(\theta^*) \cdot \sqrt{n}(\hat{\theta} - \theta). \end{aligned}$$

But since  $\hat{\theta} \xrightarrow{P} \theta \Rightarrow \theta^* \xrightarrow{P} \theta \Rightarrow f'(\theta^*) \xrightarrow{P} f'(\theta) \neq 0 < \infty$ . Therefore,

$$\sqrt{n}(f(\hat{\theta}) - f(\theta)) = [f'(\theta) + o_p(1)]\sqrt{n}(\hat{\theta} - \theta),$$

and the result now follows. ■

EXAMPLE 1.  $f(\beta) = e^\beta$ , what is the distribution of  $e^{\hat{\beta}}$  (scalar)

$$\sqrt{n}(e^{\hat{\beta}} - e^\beta) = e^\beta \sqrt{n}(\hat{\beta} - \beta) \Rightarrow N(0, e^{2\beta} \sigma^2 M^{-1})$$

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

$$\sqrt{n} \begin{pmatrix} \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{pmatrix} - \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} \xrightarrow{D} N \left( 0, \sigma^2 \frac{\partial f}{\partial \beta} M^{-1} \frac{\partial f}{\partial \beta'} \right),$$

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

$$\sigma^2 \left\{ \begin{pmatrix} \frac{1}{\beta_3} \\ -\frac{\beta_2}{\beta_3^2} \end{pmatrix} \begin{pmatrix} M^{22} & M^{23} \\ M^{32} & M^{33} \end{pmatrix} \begin{pmatrix} \frac{1}{\beta_3} \\ -\frac{\beta_2}{\beta_3^2} \end{pmatrix} \right\}.$$

EXAMPLE 2. Suppose  $y = x\beta + u$ ,  $k = 1$ . Define the reverse regression estimator [from regressing  $x$  on  $y$ ]

$$\widehat{\beta}_r = \frac{y'y}{x'y} = \frac{1}{\widehat{\beta}_{x|y}}.$$

Then

$$\begin{aligned} \widehat{\beta}_r &= \frac{(u + x\beta)'(x\beta + u)}{x'(x\beta + u)} = \frac{\beta^2 x'x + 2\beta x'u + uu'}{\beta x'x + x'u} \\ &= \frac{\beta^2 + 2\beta \frac{x'u}{x'x} + \frac{u'u}{x'x}}{\beta + \frac{x'u}{x'x}} \\ &= \frac{\beta^2 + 2\beta \frac{1}{\sqrt{n}} \frac{x'u/\sqrt{n}}{x'x/\sqrt{n}} + \frac{u'u/n}{x'x/n}}{\beta + \frac{1}{\sqrt{n}} \frac{x'u/\sqrt{n}}{x'x/n}} \\ &= \frac{\beta^2 + O_p(n^{-1/2}) + \sigma^2 M^{-1}}{\beta + O_p(n^{-1/2})} \\ &= \frac{\beta^2 + \sigma^2 M^{-1}}{\beta} + o_p(1) \\ &= \beta + \frac{\sigma^2}{\beta M} + o_p(1) \end{aligned}$$

To derive the asymptotic distribution we use the delta-method. Write

$$\widehat{\beta}_r = \frac{\widehat{N}}{\widehat{D}}.$$

The numerator satisfies

$$\sqrt{n}[\widehat{N} - N] = 2\beta \frac{x'u/\sqrt{n}}{M} + \frac{(u'u - \sigma^2)/\sqrt{n}}{M},$$

where  $N = \beta^2 + \sigma^2 M^{-1}$ . This is asymptotically normal with mean zero and finite variance. Denominator satisfies

$$\sqrt{n}(\widehat{D} - \beta) = \frac{x'u/\sqrt{n}}{M}.$$

Then,

$$\sqrt{n} \begin{pmatrix} \widehat{N} \\ \widehat{D} \end{pmatrix} - \begin{pmatrix} N \\ D \end{pmatrix} = \sqrt{n} \frac{\widehat{N} - N}{D} - \frac{N}{D^2} \sqrt{n}(\widehat{D} - D) + o_p(1)$$

by the delta method. In fact, we need the joint asymptotic distribution, so suppose that

$$\begin{pmatrix} \sqrt{n}(\widehat{N} - N) \\ \sqrt{n}(\widehat{D} - D) \end{pmatrix} \xrightarrow{D} N(0, \Omega), \quad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{21} \\ \Omega_{12} & \Omega_{22} \end{pmatrix}.$$

Then the asymptotic distribution of  $\widehat{\beta}_r$  is

$$N\left(0, \frac{\Omega_{11}}{D^2} + \frac{N^2}{D^4} \Omega_{22} - 2 \frac{N}{D^3} \Omega_{12}\right).$$



# Chapter 12

## Errors in Variables

One interpretation of linear model is that there is some unobservable  $y^*$  satisfying

$$(1a) \quad y^* = X\beta$$

and we observe  $y^*$  subject to error

$$(1b) \quad y = y^* + \varepsilon.$$

Here,  $\varepsilon$  is a mean zero stochastic error term satisfying  $\varepsilon \perp y^*$  [or more fundamentally,  $\varepsilon \perp X$ ]. Therefore,

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

where  $\varepsilon$  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

$$(2) \quad X = X^* + U, \quad U, \varepsilon \perp X^*.$$

Here,  $X$  is stochastic and  $X^*$  is fixed, although one can also consider the case where  $X^*$  is stochastic. Together (1) and (2) imply that

$$y = X\beta + \varepsilon - U\beta = X\beta + \nu, \tag{12.1}$$

where  $\nu = \varepsilon - U\beta$  is correlated with  $X$ . In practice, we have many instances of mismeasured covariates.

The consequences of the specification (12.1) are manifold. First, the least squares estimator has an obvious bias. We have

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

Take expectations

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

It is better to work with asymptotic approximation. The denominator of  $\widehat{\beta}$  satisfies

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

We shall suppose that

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

which would be justified by the Law of Large Numbers. Therefore,

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

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

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

Therefore,

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

In the scalar case,

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

where  $\sigma_\varepsilon^2/q$  is the noise to signal ratio; when  $\frac{\text{noise}}{\text{signal}} = 0$ ,  $\widehat{\beta}$  is unbiased. When  $\frac{\text{noise}}{\text{signal}} \uparrow$ ,  $|\text{bias}|$  increases and  $\beta$  shrinks towards zero.

#### REMARKS

1. In the vector case  $|p \lim \widehat{\beta}| \leq |\beta|$ , but it is not necessarily the case that each element is shrunk towards zero.
2. Suppose that  $K > 1$ , but only one regressor is measured with error, i.e.,  $\Sigma_{\varepsilon\varepsilon} = \begin{bmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & 0 \end{bmatrix}$ . In this case, all  $\widehat{\beta}$  are biased; that particular coefficient estimate is shrunk towards zero.
3. If  $X^*$  is trending, then measurement error may produce no bias; for example, suppose that  $x_t^* = t$  and  $x_t = x_t^* + U_t$ . Now  $X'^*X^* = \sum_{t=1}^T t^2 = O(T^3)$ ,  $U'U = \sum_{t=1}^T U_t^2 = O_p(T)$ . Therefore,

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

Therefore,  $\widehat{\beta} \xrightarrow{P} \beta$ . This is because the signal here is very strong and swamps the noise.

### 12.0.16 Solutions to EIV

(1) Assume knowledge of signal to noise ratio  $q/\sigma_\varepsilon^2$  and adjust  $\widehat{\beta}$  appropriately. This is hard to justify nowadays.

(2) Instrumental variables. Let  $Z_{n \times k}$  be instruments with

$$\frac{Z'X}{n} \xrightarrow{P} Q_{ZX} \tag{12.2}$$

$$\frac{Z'v}{n} \xrightarrow{P} 0, \tag{12.3}$$

i.e.,  $Z'\varepsilon/n \xrightarrow{P} 0$ ,  $Z'U/n \xrightarrow{P} 0$ . 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 assumptions (12.2) and (12.3). Provided also that  $Z'\nu/\sqrt{n}$  satisfies a central limit theorem we can conclude that

$$\sqrt{n}(\widehat{\beta}_{IV} - \beta) = \left( \frac{Z'X}{n} \right)^{-1} \frac{Z'\nu}{\sqrt{n}} \xrightarrow{D} N(0, \sigma_\nu^2 Q_{ZX}^{-1} Q_{ZZ} Q_{ZX}^{-1}).$$

This is the case where  $\nu$  are i.i.d. with mean zero and variance  $\sigma_\nu^2$ .

Where do the instruments come from?

(1) Indicator variables or ranks. Suppose that measurement errors affects cardinal outcome but not ordinality, i.e.,

$$x_i < x_j \Leftrightarrow x_i^* < x_j^*$$

Then

(2a) Take as  $z_i$  the rank of  $x_i$

$$(2b) z_i = \begin{cases} 1 & \text{if } x_i > \text{median } x \\ 0 & \text{if } x_i < \text{median } x \end{cases}$$

Method of grouping Wald (1940). The estimator is  $\bar{y}_1/\bar{x}_1$ .

(3) Time series examples,  $z$  are lagged variables.

(4) Specific examples.

### 12.0.17 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.

$H_0$  : no measurement error.

The test statistic is

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

and one rejects 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?

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

with variance  $V = \sigma_\nu^2 AA'$ . In fact,  $AA'$  simplifies

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

by the Gauss Markov Theorem. So we use

$$\begin{aligned} \widehat{V} &= s_\nu^2 \{(Z'X)^{-1}Z'Z(Z'X)^{-1} - (X'X)^{-1}\} \\ &= s_\nu^2 (Z'X)^{-1}Z'M_X Z(X'Z)^{-1}, \end{aligned}$$

where  $s_\nu^2 = \frac{1}{n-k} \sum_{i=1}^n \widehat{\nu}_i^2$ ,  $\widehat{\nu}_i = y_i - X\widehat{\beta}_{IV}$ . Thus,  $\widehat{V}^{-1} = s_\nu^{-2} X'Z(Z'M_X Z)^{-1}Z'X$ . Under  $H_0$

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

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



# Chapter 13

## Heteroskedasticity

We made the assumption that

$$\text{Var}(y) = \sigma^2 I$$

in the context of the linear regression model. This contains two material parts: (a) off diagonals are zero (independence), and (b) diagonals are the same. Here we extend to the case where

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

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.

### 13.0.18 Effects of Heteroskedasticity

Consider the OLS estimator

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

In the new circumstance, this is unbiased, because

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

but

$$\text{Var}(\widehat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1} \neq \sigma^2(X'X)^{-1}.$$

As sample size increases,  $\text{Var}(\widehat{\beta}) \rightarrow 0$ , so that the OLSE is still consistent. The main problem then is with the variance.

(a) Least squares standard errors are estimating the wrong quantity. We have

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

but

$$\frac{1}{n} \sum_{i=1}^n x_i x_i' \sigma_i^2 \neq \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \cdot \frac{1}{n} \sum_{i=1}^n x_i x_i',$$

so there is no cancellation.

(b) 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,

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

is efficient by Gauss–Markov. So

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

is the efficient estimator here; this is not equal to  $\widehat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y$ , unless  $\Sigma = \sigma^2I$  (or some more complicated conditions are satisfied). Can show directly that

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

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

### 13.0.19 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  $\sigma_i^2$ ,  $i = 1, \dots, n$  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  $\text{Var}(\widehat{\beta})$  by

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

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

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

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

$$n(R\widehat{\beta} - r)' [R\widehat{V}R']^{-1} (R\widehat{\beta} - r) \xrightarrow{D} \chi_J^2.$$

### 13.0.20 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,

$$\text{Var}(u_i) = \frac{1}{n_i} \text{Var}(u_{ij}) = \frac{\sigma^2}{n_i}.$$

Here, the variance is inversely proportional to household size. This is easy case since apart from single constant,  $\sigma^2$ ,  $\sigma_i^2$  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 parameter  $\gamma$ . Suppose we have a normal error and that  $\theta \cap \beta = \phi$ . Then,

$$\begin{aligned} L(\beta, \theta) &= -\frac{1}{2} \ln |\Sigma(\theta)| - \frac{1}{2} (y - X\beta)' \Sigma^{-1}(\theta) (y - X\beta) \\ &= -\frac{1}{2} \sum_{i=1}^n \ln \sigma_i^2(\theta) - \frac{1}{2} \sum_{i=1}^n (y_i - x_i' \beta)^2 \sigma_i^{-2}(\theta). \end{aligned}$$

In this case,

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_{i=1}^n x_i (y_i - x_i' \beta) \sigma_i^{-2}(\theta) \quad (13.1)$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = -\frac{1}{2} \sum_{i=1}^n \frac{\partial \ln \sigma_i^2}{\partial \theta}(\theta) + \frac{1}{2} \sum_{i=1}^n (y_i - x_i' \beta)^2 \sigma_i^{-4}(\theta) \frac{\partial \sigma_i^2}{\partial \theta}. \quad (13.2)$$

The estimators  $(\hat{\beta}_{\text{MLE}}, \hat{\theta}_{\text{MLE}})$  solve (13.1), (13.2), which are nonlinear equations in general. However, the equation for  $\beta$  is conditionally linear, that is suppose that we have a solution  $\hat{\theta}_{\text{MLE}}$ , then

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

Have to use iterations. Start with  $\hat{\beta}_{\text{OLS}}$ , which is consistent, this gives us  $\hat{\theta}$ , 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  $\theta$ . 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

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

where  $\hat{u}_i = y_i - x_i' \hat{\beta}_{\text{MLE}}$ .

## Properties of the Procedure

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

$$\begin{bmatrix} \sqrt{n}(\hat{\beta}_{\text{MLE}} - \beta) \\ \sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \end{bmatrix} \xrightarrow{D} N(0, \Omega) \text{ for some } \Omega$$

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

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

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

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

We say that  $\hat{\beta}_{\text{ML}}$  is asymptotically Gauss–Markov efficient, BLAUE.

Often people use ad hoc estimates of  $\theta$  and construct

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

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

### 13.0.21 Testing for Heteroskedasticity

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

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

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

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,  $\text{Var}\left(\frac{u^2}{2}\right) = 2$ . Therefore,

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

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  $H_0$

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

Reject for large LM.

# Chapter 14

## Nonlinear Regression Models

We now consider a more general class of regression models; the Nonlinear regression model. We have noted that the linear in parameters but nonlinear in variables can really be accommodated inside the classical linear framework, so the real difficulty is in accommodating nonlinear in parameters models. We suppose that

$$E[y|X] \text{ or } E[y] = \begin{bmatrix} g_1(x_1, \beta) \\ \vdots \\ g_n(x_n, \beta) \end{bmatrix},$$

where for example  $g(x, \beta) = x^\beta$ . Can also specify  $\text{Var}[y] = \sigma^2 I$ , to obtain

$$y_i = g(x_i, \beta) + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

where  $\varepsilon_i$  are i.i.d. mean zero with variance  $\sigma^2$ .

In this case, how do we estimate  $\beta$ ? 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  $\beta$  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

$$\hat{\beta}_{NLLS} = \arg \min_{\beta \in B} S_n(\beta).$$

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. In any event, computation of ‘the estimator’ is a big issue. 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.

### 14.0.22 Computation

We make some remarks

- 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.
- 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  $\lambda$  were known, would estimate  $\beta$  by

$$\widehat{\beta} = [X(\lambda)'X(\lambda)]^{-1}X(\lambda)'y, \quad X(\lambda) = \begin{bmatrix} \frac{x_1^\lambda - 1}{\lambda} \\ \vdots \\ \frac{x_n^\lambda - 1}{\lambda} \end{bmatrix}.$$

Then write

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

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

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

$$\frac{\partial S_n}{\partial \beta}(\hat{\beta}_{NNLS}) = 0$$

Can evaluate  $S_n$ ,  $\partial S_n/\partial \beta$ ,  $\partial^2 S_n/\partial \beta \partial \beta'$ , for any  $\beta$ . Suppose we take an initial guess  $\beta_1$  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  $\beta_1$  and find where that line intersects the zero. The tangent at  $\beta_1$  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. In practice the following criteria are used

$$|\beta_{r+1} - \beta_r| < \tau, \quad |S_n(\beta_{r+1}) - S_n(\beta_r)| < \tau.$$

This is Newton's method. 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 product (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  $\lambda$

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

and choose  $\lambda$  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 the 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.

# Chapter 15

## Asymptotic Properties

### 15.0.23 Consistency of NLLS

We state the following theorem about the consistency of the nonlinear least squares estimator. We shall just make high level conditions that could really apply to much more general criterion function and then investigate these conditions in the special case of interest.

**Theorem 19** *Suppose that (1) The parameter space  $B$  is a compact subset of  $\mathbb{R}^K$ ; (2)  $S_n(\beta)$  is continuous in  $\beta$  for all possible data; (3)  $S_n(\beta)$  converges in probability to a non-random function  $S(\beta)$  uniformly in  $\beta \in B$ ; (4) The function  $S(\beta)$  is uniquely minimized at  $\beta = \beta_0$ . Then*

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

**Proof.** See Amemiya (1986, Theorem 4.1.1). ■

We just show why (3) and (4) are plausible. Substituting in for  $y_i$ , we have

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

By the Law of Large numbers

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

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

where the latter result is true for any given  $\beta$ . Therefore,

$$S_n \xrightarrow{P} \sigma^2 + p \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n [g_i(\beta) - g_i(\beta_0)]^2 \equiv S(\beta).$$

Need convergence in probability to hold uniformly in a compact set containing  $\beta_0$  (or over  $B$ ), which requires subtle arguments. Now  $S(\beta_0) = \sigma^2$ , but  $S(\beta) \geq \sigma^2$  for all  $\beta$ . So, in the limit,  $\beta_0$  minimizes  $S$ . Need  $S(\beta)$  to be uniquely minimized at  $\beta_0$  (identification condition).

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

$$S(\beta) = \sigma^2 + (\beta - \beta_0)' X' X (\beta - \beta_0),$$

which is a quadratic function of  $\beta$ . (3) also holds in this case under mild conditions on  $X$ .

### 15.0.24 Asymptotic Distribution of NLLS

We state the following theorem about the asymptotic distribution of the nonlinear least squares estimator.

**Theorem 20** *Suppose that: (1)  $\widehat{\beta}$  is such that  $\partial S_n(\widehat{\beta})/\partial \beta = 0$  and satisfies  $\widehat{\beta} \xrightarrow{P} \beta_0$ , where  $\beta_0$  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  $\beta_0$ ; (3)  $\partial^2 S_n(\beta)/\partial \beta \partial \beta'$  converges in probability to a finite nonsingular matrix  $A(\beta)$  uniformly in  $\beta$  over any shrinking neighbourhood of  $\beta_0$ ; (4)  $\sqrt{n} \partial S_n(\beta_0)/\partial \beta \xrightarrow{D} N(0, B(\beta_0))$  for some finite matrix  $B(\beta_0)$ . Then,*

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{D} N(0, A^{-1}(\beta_0)B(\beta_0)A^{-1}(\beta_0)).$$

**Proof.** We have

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

where  $\beta^*$  lies between  $\beta_0$  and  $\hat{\beta}$  by the multivariate mean value theorem. Applying assumptions (1)-(3) we get

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

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

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

Suppose that  $(x_i, \varepsilon_i)$  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

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

What about (3)?

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

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

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

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

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \frac{\partial^2 g}{\partial \beta \partial \beta'}(x_i, \beta_0) &\xrightarrow{P} 0 \\ \frac{1}{n} \sum_{i=1}^n \frac{\partial g}{\partial \beta} \frac{\partial g}{\partial \beta'}(x_i, \beta_0) &\xrightarrow{P} E \left[ \frac{\partial g}{\partial \beta} \frac{\partial g}{\partial \beta'}(x_i, \beta_0) \right]. \end{aligned}$$

These conditions need to be strengthened a little to obtain uniformity over the neighbourhood of  $\beta_0$ .

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  $\beta^{**}$ . Then, provided

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

for some function  $D$  for which  $ED^2(x) < \infty$ , 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  $\varepsilon_i$  are i.i.d. and independent of  $x_i$ , we have

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

and the asymptotic distribution is

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, \sigma^2 A^{-1}(\beta_0)).$$

### 15.0.25 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  $\hat{\theta}$  maximize  $\ell(\text{data}, \theta)$

$$\hat{\theta} \xrightarrow{D} \theta_0$$

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N \left( 0, \left[ \frac{1}{n} \sum_{i=1}^n \frac{\partial \ell_i}{\partial \theta} \frac{\partial \ell_i}{\partial \theta'} \right]^{-1} \right) \equiv I \quad [\text{information matrix}]$$

under variety of conditions on  $\ell$ : i.i.d., smoothness, moments (normal linear regression is a special case of this).

**Theorem 21** Cramèr–Rao. Any unbiased estimator  $\tilde{\theta}$  of  $\theta$  has

$$\text{Var}(\tilde{\theta}) \geq I^{-1}/n.$$

**Corollary 22** The MLE is asymptotically “efficient” amongst the class of all asymptotically normal estimates (stronger than Gauss–Markov).

In the normal linear regression case  $\theta = (\beta, \sigma^2)$

$$I = \begin{bmatrix} \frac{X'X}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix},$$

which is a diagonal information matrix. This provides a quick way of finding asymptotic distribution of  $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$ , i.e.,

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{D} N(0, 2\sigma^4).$$

Note that  $\sqrt{n}(s^2 - \sigma^2) \xrightarrow{D} N(0, 2\sigma^4)$ , where  $s^2 = \hat{u}'\hat{u}/(n - K)$ .

Both *asymptotically* efficient estimators of  $\sigma^2$ . In finite samples

- (1)  $\text{Var}[\sqrt{n}(\hat{\beta} - \beta)] = \sigma^2(X'X/n)^{-1}$
  - (2)  $\text{Var}[\sqrt{n}(s^2 - \sigma^2)] = 2\sigma^4n/(n - K)$  strictly greater than limit.
- (1) is case where efficiency is achieved in finite samples.  
 (2) is case where efficiency is only achieved in infinite sample.



# Chapter 16

## Generalized Method of Moments

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  $u_c > 0$ ,  $u_{cc} < 0$ ,  $1 + r_{i,t+1}$ ,  $i = 1, \dots, m$  is gross return on asset  $i$  at time  $t + 1$ . The representative agent solves the following optimization problem

$$\max_{\{c_t, w_t\}_{t=0}^{\infty}} \sum_{\tau=0}^{\infty} \beta^{\tau} E[u(c_{t+\tau})|I_t],$$

where  $w_t$  is a vector of portfolio weights. This is a dynamic programming problem. We assume that there is a unique interior solution; this is characterized by the following condition

$$u'(c_t) = \beta E[(1 + r_{i,t+1})u'(c_{t+1})|I_t], \quad i = 1, \dots, m.$$

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,  $\gamma$  is the coefficient of relative risk aversion. Then

$$c_t^{-\gamma} = \beta E[(1 + r_{i,t+1})c_{t+1}^{-\gamma}|I_t],$$

and rearranging we get

$$E \left[ 1 - \beta \left\{ (1 + r_{i,t+1}) \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} \right\} \middle| I_t^* \right] = 0, \quad i = 1, \dots, m$$

where  $I_t^* \subset I_t$  and  $I_t^*$  is the econometrician's information set.

We want to estimate the parameters and test the theory given a dataset consisting of  $c_t, r_{i,t+1}, I_t^*$ . Let  $\theta_{p \times 1} = (\beta, \gamma)$  and define the vector

$$g(\theta, x_t) = \begin{bmatrix} \vdots \\ \left[ 1 - \beta \left\{ (1 + r_{i,t+1}) \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} \right\} \right] z_t \\ \vdots \end{bmatrix}_{q \times 1},$$

where  $z_t \in I_t^*$ , and  $x_t = (z_t, c_t, c_{t+1}, r_{1,t+1}, \dots, r_{m,t+1})'$ . We shall suppose that  $q \geq p$ . When  $q = p$  we say that we are exactly identified, while when  $q > p$  we are overidentified. In any case, we shall suppose that  $E[g(\theta_0, x_t)] = 0$  for some unique  $\theta_0$ . We also suppose that  $x_t$  is i.i.d., but this is not necessary.

Estimation strategy is based on the consequence that

$$G_T(\theta) = \frac{1}{T} \sum_{t=1}^T g(\theta, x_t) \approx 0$$

when  $\theta$  is the true parameter value. Let  $Q_T(\theta) = G_T(\theta)' W_T G_T(\theta)$ , where  $W_T$  is a weighting matrix. Then let

$$\hat{\theta}_{GMM} = \arg \min_{\theta} Q_T(\theta).$$

This defines a large class of estimators, one for each weighting matrix  $W_T$ . It is generally a nonlinear optimization problem like nonlinear least squares; various techniques are available for finding the minimizer.

We now turn to the asymptotic properties. Provided  $W_T \xrightarrow{P} W > 0$ ,  $q \geq p$ , and  $\theta_0$  is the unique value, we have

$$\hat{\theta}_{GMM} \xrightarrow{P} \theta_0.$$

Under further conditions, we can establish

$$\sqrt{T}(\widehat{\theta}_{GMM} - \theta_0) \xrightarrow{D} N(0, V(W))$$

where  $V$  is a complicated function of  $W$ . Specifically,

$$V(W) = (\Gamma'W\Gamma)^{-1}\Gamma'W\Omega W\Gamma(\Gamma'W\Gamma)^{-1},$$

where (in the i.i.d. case)

$$\Omega(\theta_0) = \text{Var}\sqrt{T}G_T(\theta_0) = \frac{1}{T} \sum_{t=1}^T E g_t(\theta_0)g_t(\theta_0)',$$

while  $\Gamma = p \lim \partial G_T(\theta_0)/\partial \theta$ . Hansen went step further and asked what is the optimal choice of  $W$ . In fact  $W_T$  should be an estimate of  $\Omega^{-1}$ . We take

$$\Omega(\tilde{\theta}) = \frac{1}{T} \sum_{t=1}^T g_t(\tilde{\theta})g_t(\tilde{\theta})', \quad q \times q$$

where  $\tilde{\theta}$  is a preliminary estimate of  $\theta_0$  obtained using some arbitrary weighting matrix, e.g.,  $I_q$ . In sum, then, the full procedure is

- (1)  $\tilde{\theta} = \arg \min G_T(\theta)'G_T(\theta)$
- (2)  $\widehat{\theta} = \arg \min G_T(\theta)'\widehat{\Omega}^{-1}G_T(\theta)$ .

The asymptotic distribution is now normal with mean zero and variance  $(\Gamma'S^{-1}\Gamma)^{-1}$ . This is minimum among all  $\Gamma$  is efficient in these terms. Can estimate the asymptotic variance by  $[\widehat{\Gamma}'\widehat{\Omega}^{-1}\widehat{\Gamma}]^{-1}$ . Can now do inference about  $\theta$ .

Trivial example, linear regression  $u = y - X\beta$ . Moment conditions

$$E[X'u(\beta_0)] = 0.$$

Here there are  $K$  conditions and  $K$  parameters exactly identified case. In this case, there exists a unique  $\widehat{\beta}$  that satisfies the empirical conditions.

Suppose now  $E[X'u(\beta_0)] \neq 0$  but

$$E[Z'u(\beta_0)] = 0$$

for  $Z_{J \times 1}$  instruments,  $J > K$ . 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. Suppose also that  $u$  has variance matrix  $\sigma^2 I$  independent of  $Z$ , then

$$\text{Var} \frac{1}{\sqrt{T}} Z' u = \sigma^2 \frac{Z' Z}{T}.$$

Therefore, we take

$$Q_T(\theta) = (y - X\beta)' Z (Z' Z)^{-1} Z' (y - X\beta).$$

This has a closed form solution. Let  $P_Z = Z(Z' Z)^{-1} Z'$ . Then, the solution is the same as

$$\min(y^* - X^* \beta)' (y^* - X^* \beta),$$

where  $y^* = P_Z y$  and  $X^* = P_Z X$ , which implies that

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

i.e., it is an instrumental variables estimator with instruments  $X^*$ .

# Chapter 17

## Time Series

We now consider time series data, i.e., data that are produced in sequence. There are some special features of these models.

We start with univariate time series  $\{y_t\}_{t=1}^T$ .

### 17.0.26 Some Fundamental Properties

There are two main features:

- (a) stationarity/nonstationarity
- (b) dependence.

We first define stationarity.

Strong Stationarity. The stochastic process  $y$  is said to be strongly stationary if the vectors  $(y_t, \dots, y_{t+r})$  and  $(y_{t+s}, \dots, y_{t+s+r})$  have the same distribution for all  $t, s, r$ .

Weak Stationarity. The stochastic process  $y$  is said to be weakly stationary if the vectors  $(y_t, \dots, y_{t+r})$  and  $(y_{t+s}, \dots, y_{t+s+r})$  have the same mean and variance for all  $t, s, r$ .

Most of what we say is restricted to [weakly] stationary series, but in the last 20 years there have been major advances in the theory of nonstationary time series. There has also been some exploration of the difference between weak and strong stationarity.

Dependence: One measure of dependence is given by the covariogram [or correlogram]

$$\text{Cov}(y_t, y_{t-s}) = \gamma_s ; \quad \rho_s = \frac{\gamma_s}{\gamma_0}.$$

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,  $\gamma_s = 0$  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  $\gamma_s \rightarrow 0$  as  $s \rightarrow \infty$ .

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( $p, q$ ) :

$$y_t = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q},$$

where  $\varepsilon_t$  is i.i.d., mean zero and variance  $\sigma^2$ . We shall for convenience usually assume that  $\mu = 0$ . We also assume for convenience that this model holds for  $t = 0, \pm 1, \dots$ . It is convenient to write this model using lag polynomial notation

$$A(L)y_t = B(L)\varepsilon_t,$$

where the lag polynomials  $A(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$  and  $B(L) = 1 - \theta_1 L - \cdots - \theta_q L^q$ . Here,  $Ly_t = y_{t-1}$ . 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

$$y_{t-1} = \phi y_{t-2} + \varepsilon_{t-1}.$$

This implies that

$$y_t = \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 y_{t-2}$$

$$\begin{aligned}
&= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots \\
&= \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j},
\end{aligned}$$

which is called the MA( $\infty$ ) representation of the time series; this shows that  $y_t$  depends on all the past shocks. Note that we need  $|\phi| < 1$  for the above sum to ‘exist’.

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,

$$\text{var}(y_t) = \phi^2 \text{var}(y_{t-1}) + \sigma^2,$$

which implies that

$$\gamma_0 = \frac{\sigma^2}{1 - \phi^2},$$

where  $\gamma_0 = \text{var}(y_t) = \text{var}(y_{t-1})$ . This last calculation of course requires that  $|\phi| < 1$ , which we are assuming for stationarity. Finally,

$$\text{cov}(y_t, y_{t-1}) = E(y_t y_{t-1}) = \phi E(y_{t-1}^2) + 0,$$

which implies that

$$\gamma_1 = \phi \frac{\sigma^2}{1 - \phi^2},$$

while

$$\text{Cov}(y_t, y_{t-2}) = E(y_t y_{t-2}) = \phi E(y_{t-1} y_{t-2}) = \phi^2 \frac{\sigma^2}{1 - \phi^2}.$$

In general

$$\gamma_s = \sigma^2 \frac{\phi^{2s}}{1 - \phi^2}; \quad \rho_s = \frac{\phi^s}{1 - \phi^2}.$$

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  $\varepsilon_t$  are i.i.d. mean zero with variance  $\sigma^2$ . In this case,

$$E(y_t) = 0,$$

and

$$\text{var}(y_t) = \sigma^2(1 + \theta^2).$$

Furthermore,

$$\text{cov}(y_t, y_{t-1}) = E\{(\varepsilon_t - \theta\varepsilon_{t-1})(\varepsilon_{t-1} - \theta\varepsilon_{t-2})\} = -\theta E(\varepsilon_{t-1}^2) = -\theta\sigma^2.$$

Therefore,

$$\rho_1 = \frac{-\theta}{1 + \theta^2}, \quad \rho_j = 0, \quad j = 2, \dots$$

This is a 1-dependent series. MA( $q$ ) is a  $q$ -dependent series. Note that the process is automatically stationary for any value of  $\theta$ . 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

$$A(L)y_t = B(L)\varepsilon_t,$$

where  $A(L) = 1 - \phi_1 L - \dots - \phi_p L^p$  and  $B(L) = 1 - \theta_1 L - \dots - \theta_q L^q$ , where  $L$  is the lag operator  $Ly_t = y_{t-1}$ ,  $L^2 y_t = L(Ly_t) = y_{t-2}$ , while  $A(L)$  and  $B(L)$  are polynomials in the lag operator. The stationarity condition for an ARMA( $p, q$ ) process is just that we need the roots of the autoregressive polynomial  $1 - \phi_1 z - \dots - \phi_p z^p$  to be outside unit circle. Likewise the condition for invertibility

is that the roots of the moving average polynomial  $1 - \theta_1 L - \dots - \theta_q L^q$  lie outside the unit circle. Assuming these conditions are satisfied we can write this process in two different ways

$$\begin{aligned}\frac{A(L)}{B(L)}y_t &= \sum_{j=0}^{\infty} \gamma_j y_{t-j} = \varepsilon_t \\ y_t &= \frac{B(L)}{A(L)}\varepsilon_t = \sum_{j=0}^{\infty} \delta_j \varepsilon_{t-j}.\end{aligned}$$

The first line is called the  $AR(\infty)$  representation, and expresses  $y$  in terms of its own past. The second line is called the  $MA(\infty)$  representation, and expresses  $y$  in terms of the past history of the random shocks.

## 17.0.27 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

$$\hat{\gamma}_s = \frac{1}{T-s} \sum_{t=s+1}^T (y_t - \bar{y})(y_{t-s} - \bar{y})$$

$$\hat{\rho}_s = \hat{\gamma}_s / \hat{\gamma}_0.$$

These sample quantities are often used to describe the actual series properties. 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

A popular estimation criterion is the Likelihood under normality. Suppose that

$$\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{pmatrix} \sim N(0, \sigma^2 I), \quad \text{then} \quad \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix} \sim N(0, \Sigma),$$

where for an AR(1) process

$$\Sigma = \frac{\sigma^2}{1 - \gamma^2} \begin{bmatrix} 1 & \gamma & \gamma^2 & \cdots & \gamma^{T-1} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix},$$

while for an MA(1) process

$$\Sigma = \sigma^2(1 + \theta^2) \begin{bmatrix} 1 & \frac{-\theta}{1+\theta^2} & & 0 \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}.$$

In either case,

$$\ell = \frac{-T}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} y' \Sigma^{-1} y.$$

Maximize with respect to all the parameters. In practice,  $|\Sigma|$  and  $\Sigma^{-1}$  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,

$$f(x, y) = f(x|z)f(z).$$

We use this repeatedly and take logs to give

$$\ell(y_1, \dots, y_T; \theta) = \sum_{t=p+1}^T \ell(y_t | y_{t-1}, \dots, y_1) + \ell(y_1, \dots, y_p).$$

This writes the log likelihood in terms of conditional distributions and a single marginal distribution. In AR cases  $\ell(y_t|y_{t-1}, \dots, y_1)$  is easy to find:  $y_t|y_{t-1}, \dots, y_1$  is  $N(\phi_1 y_{t-1} + \dots + \phi_p y_{t-p}, \sigma^2)$ . In the AR(1) case

$$\ell_{t|t-1} \sim -\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y_t - \phi_1 y_{t-1})^2.$$

The marginal distribution of  $y_1$  is  $N(0, \sigma^2/(1 - \phi^2))$ , which means that

$$\ell(y_1) = -\frac{1}{2} \log \frac{\sigma^2}{1 - \phi^2} - \frac{(1 - \phi^2)}{2\sigma^2} y_1^2.$$

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  $\sum_{t=2}^T \ell(y_t|y_{t-1}, \dots, y_1)$ , in which case we use

$$-\frac{T-1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=2}^T (y_t - \phi y_{t-1})^2.$$

This criterion is equivalent to the least squares criterion, and has unique maximum

$$\hat{\phi} = \frac{\sum_{t=2}^T y_t y_{t-1}}{\sum_{t=2}^T y_{t-1}^2}.$$

This estimator is just the OLS of  $y_t$  on  $y_{t-1}$  [but using the reduced sample]. 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. For example, the MLE imposes that  $\hat{\phi}$  be less than one - as  $\phi \rightarrow \pm 1$ ,  $\ell \rightarrow -\infty$ . The OLS estimate however can be either side of the unit circle.

## 17.0.28 Forecasting

Let the sample be  $\{y_1, \dots, y_T\}$ . Suppose that

$$y_t = \gamma y_{t-1} + \varepsilon_t, \quad |\gamma| < 1,$$

where we first assume that  $\gamma$  is known. Want to forecast  $y_{T+1}, y_{T+2}, \dots, y_{T+r}$  given the sample information. We have

$$y_{T+1} = \gamma y_T + \varepsilon_{T+1}.$$

Therefore, forecast  $y_{T+1}$  by

$$\hat{y}_{T+1|T} = E[y_{T+1}|\text{sample}] = \gamma y_T.$$

The forecast error is  $\varepsilon_{T+1}$ , which is mean zero and has variance  $\sigma^2$ .

What about forecasting  $r$  periods ahead?

$$y_{T+r} = \gamma^r y_T + \gamma^{r-1} \varepsilon_{T+1} + \dots + \varepsilon_{T+r}.$$

Therefore, let

$$\hat{y}_{T+r|T} = \gamma^r y_T$$

be our forecast. The forecast error  $\hat{y}_{T+r|T} - y_{T+r}$  has mean zero and variance  $\sigma^2(1 + \gamma^2 + \dots + \gamma^{2r-2})$ .

In practice, we must use an estimate of  $\gamma$ , so that

$$\hat{y}_{T+r|T} = \hat{\gamma}^r y_T,$$

where  $\hat{\gamma}$  is estimated from sample data. If  $\gamma$  is estimated well, then this will not make much difference.

Forecast interval

$$\hat{y}_{T+r|t} \pm 1.96 \cdot SD, \quad SD = \sigma^2(1 + \gamma^2 + \dots + \gamma^{2r-2}).$$

This is to be interpreted like a confidence interval. Again we must replace the unknown parameters by consistent estimates.

# Chapter 18

## 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 like fixed regressors. 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

$$\ln \text{GNP} = \beta_1 + \beta_2 \text{time} + u_t.$$

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

$$y = X\beta + u \quad , \quad E(uu') = \Sigma = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{T-1} \\ & & & \ddots & \gamma_2 \\ & & & & \ddots & \gamma_0 \end{bmatrix} .$$

The consequences for estimation and testing of  $\beta$  are the same as with heteroskedasticity: OLS is consistent and unbiased, but inefficient, while the SE's are wrong. Specifically,

$$\text{var}(\widehat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1},$$

where

$$\Psi_T = X'\Sigma X = \sum_t \sum_s X_t X'_s \gamma_{|t-s|} = \sum_{t=1}^T X_t X'_t \gamma_0 + \sum_{j=1}^{T-1} \gamma_j \sum_{|t-s|=j} X_t X'_{t+j}. \quad (18.1)$$

A naive implementation of the Eicker-White strategy is going to fail here, i.e., if we replace  $\gamma_j$  by  $\widehat{u}_t \widehat{u}_{t+j}$  in (18.1), then we get inconsistent estimates. This is basically because there are two 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 the terms. Bartlett/White/Newey/West SE's: Replace by sample equivalents and use weights

$$w(j) = 1 - \frac{j}{n+1},$$

so that

$$\widehat{\Psi}_T = \sum_{t,s:|t-s|\leq n(T)} X_t X'_s w(|t-s|) \widehat{u}_t \widehat{u}_s.$$

This also ensures a positive definite covariance matrix estimate. Provides consistent standard errors [even when  $u_t$  is also heterogeneous, this is still true].

An alternative strategy is to parameterize  $u_t$  by say an ARMA process and do maximum likelihood

$$\ell = -\frac{1}{2} \ln |\Sigma(\theta)| - \frac{1}{2} (y - X\beta)' \Sigma(\theta)^{-1} (y - X\beta).$$

Same as before, efficient estimate of  $\beta$  (under Gaussianity) is

$$\widehat{\beta}_{\text{ML}} = (X'\Sigma(\widehat{\theta})^{-1}X)^{-1}X'\Sigma(\widehat{\theta})^{-1}y,$$

where  $\widehat{\theta}$  is the MLE of  $\theta$ . This will be asymptotically efficient when the chosen parametric model is correct.

### 18.0.29 Testing for autocorrelation

Suppose that we observe  $u_t$ , which is generated from an  $AR(1)$  process

$$u_t = \rho y_{t-1} + \varepsilon_t,$$

where  $\varepsilon_t$  are i.i.d. The null hypothesis is that  $u_t$  is i.i.d., i.e.,

$$H_0 : \rho = 0$$

$$\text{vs. } H_A : \rho \neq 0, \rho > 0.$$

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

$$LM = T \left( \frac{\sum_t \hat{u}_t \hat{u}_{t-1}}{\sum_t \hat{u}_{t-1}^2} \right)^2 = Tr_1^2 \xrightarrow{D} \chi_1^2,$$

where  $\hat{u}_t$  are the OLS residuals. Therefore, we reject the null hypothesis when  $LM$  is large relative to the critical value from  $\chi_1^2$ . This approach is limited to two sided-alternatives. We can however also use the signed version,  $\sqrt{T}r_1$ , which satisfies

$$\sqrt{T}r_1 \xrightarrow{D} N(0, 1)$$

under the null hypothesis. We could also just use  $\partial \ell / \partial \rho = \sum_t (u_t - \rho u_{t-1}) u_{t-1}$ , and do a similar test.

#### OTHER TESTS

The Durbin–Watson  $d$  is

$$d = \frac{\sum_{t=2}^T (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^T \hat{u}_t^2}.$$

This is always printed out by many regression packages. We have

$$d = 2 \Leftrightarrow \rho = 0$$

$$d > 2 \Leftrightarrow \rho < 0$$

$$d < 2 \Leftrightarrow \rho > 0.$$

Using the approximation  $d \approx 2(1 - r_1)$ , we have

$$\sqrt{T} \left( 1 - \frac{d}{2} \right) \rightarrow N(0, 1).$$

Don't bother with the Bounds test, just use asymptotics.

Generalization (test against AR( $p$ )). Box-Pierce  $Q$

$$Q = T \sum_{j=1}^P r_j^2 \xrightarrow{D} \chi_P^2.$$

The finite sample adjustment

$$Q^* = T(T+2) \sum_{j=1}^P \frac{r_j^2}{T-j} \xrightarrow{D} \chi_P^2,$$

gives a better approximation.

# Chapter 19

## Dynamic Regression Models

Distributed lag

$$y_t = \alpha + \sum_{j=0}^q \beta_j X_{t-j} + u_t,$$

where for now  $u_t \stackrel{\text{iid}}{\sim} 0, \gamma^2$ . [Could have  $q = \infty$ ]. Captures the idea of dynamic response: affect on  $y$  of change in  $x$  may take several periods to work through.

TEMPORARY CHANGE. Suppose that  $x_t \rightarrow x_t + \Delta$  but that future  $x_s$  are unaffected, then

$$y_t \rightarrow y'_t + \beta_0 \Delta$$

$$y_{t+1} \rightarrow y_t + \beta_1 \Delta.$$

PERMANENT CHANGE. Suppose that  $x_s \rightarrow x_s + \Delta$ ,  $\forall s \geq t$ . Then

$$y_t \rightarrow y_t + \beta_0 \Delta$$

$$y_{t+1} \rightarrow y_t + (\beta_0 + \beta_1) \Delta$$

The impact effect is  $\beta_0 \Delta$ . Long run effect is  $\lim_{s \rightarrow \infty} \beta_s / \Delta = 0$ .

When  $q$  is large (infinite) there are too many free parameters  $\beta_j$ , which makes estimation difficult and imprecise. To reduce the dimensionality it is appropriate to make restrictions on  $\beta_j$ . For example,

the polynomial lag

$$\beta_j = \begin{cases} a_0 + a_1j + \cdots + a_pj^p & \text{if } j \leq p \\ 0 & \text{else.} \end{cases}$$

The Geometric lag

$$\beta_j = \beta\lambda^j, \quad j = 0, 1, \dots \quad (0 < \lambda < 1),$$

which 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,

$$(1 - \lambda L)y_t = \alpha(1 - \lambda L) + \beta x_t + (1 - \lambda L)u_t.$$

Therefore,

$$y_t = \alpha(1 - \lambda) + \lambda y_{t-1} + \beta x_t + u_t - \lambda u_{t-1}.$$

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.

### 19.0.30 Some examples from economics

#### Adaptive expectations

Suppose that

$$\underbrace{y_t}_{\text{demand}} = \alpha + \beta \underbrace{x_{t+1,t}^*}_{\text{expected Price}} + \varepsilon_t,$$

but that the expected price is unobserved by the econometrician. We do however observe  $x_t$ , where

$$\underbrace{x_{t,t+1}^* - x_{t-1,t}^*}_{\text{revised expectations}} = (1 - \lambda) \underbrace{(x_t - x_{t-1,t}^*)}_{\text{forecast error}},$$

i.e.,

$$x_{t,t+1}^* = \underbrace{\lambda x_{t-1,t}^*}_{\text{old forecast}} + \underbrace{(1 - \lambda)x_t}_{\text{news}}.$$

Write

$$(1 - \lambda L)x_t^* = (1 - \lambda)x_t,$$

which implies that

$$x_t^* = \frac{(1 - \lambda)}{1 - \lambda L} x_t = (1 - \lambda)[x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \dots],$$

and 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.

#### 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.

### Error correction

Suppose long run equilibrium is

$$y = \lambda x,$$

and disequilibria are corrected according to

$$\Delta y_t = \beta(y_{t-1} - \lambda x_{t-1}) + \lambda \Delta x_{t-1} + \varepsilon_t,$$

where  $\beta < 0$ . This implies that

$$y_t = y_{t-1}(1 + \beta) + \lambda(1 - \beta)x_{t-1} - \lambda x_{t-2} + \varepsilon_t.$$

### 19.0.31 Estimation of ADL models

Suppose that

$$y_t = \alpha_1 + \gamma_2 y_{t-1} + \beta_3 x_t + \varepsilon_t,$$

where we have two general cases regarding the error term:

- (1)  $\varepsilon_t$  is i.i.d. 0,  $\gamma^2$
- (2)  $\varepsilon_t$  is autocorrelated.

In case (1), we can use OLS regression to get consistent estimates of  $\alpha$ ,  $\gamma$ , and  $\beta$ . If there are constraints, e.g.,

$$\left. \begin{aligned} \theta_1 &= \alpha(1 - \lambda) \\ \theta_2 &= \lambda \\ \theta_3 &= \beta(1 - \lambda) \end{aligned} \right\},$$

then we estimate the parameters by

$$\begin{aligned} \widehat{\lambda} &= \widehat{\theta}_2 \\ \widehat{\alpha} &= \frac{\widehat{\theta}_1}{1 - \widehat{\theta}_2} \\ \widehat{\beta} &= \frac{\widehat{\theta}_3}{1 - \widehat{\theta}_2}. \end{aligned}$$

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  $\varepsilon_t$  through  $\eta_{t-1}$ . In this case there are many instruments: (1) All lagged  $x_t$ ; (2)  $y_{t-2}, \dots$ . However, when  $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$ ,  $\eta_t$  i.i.d. lagged  $y$  are no longer valid instruments.

REMARKS.

1. IV are not generally as efficient as ML when the error terms are normally distributed.
2. Also note that when there is a lagged dependent variable, the DW test is no longer valid. In this case, the corresponding test is based on the statistic

$$h = \left(1 - \frac{1}{2}d\right) \sqrt{\frac{T}{1 - T \cdot \alpha \text{Var}(\widehat{\alpha})}}.$$



# Chapter 20

## Nonstationarity

### NONSTATIONARY TIME SERIES MODELS

#### (1) Trend stationary

$$y_t = \mu + \beta t + u_t,$$

where  $\{u_t\}$  is a stationary process, possibly  $A(L)u_t = B(L)\varepsilon_t$ . This is the trend+stationary decomposition.

GNP grows at 3% per year (on average) for ever after. Any shocks ( $u_t$ ) are transitory - they last for some period of time and then are forgotten as  $y$  returns to trend.

#### (2) Difference stationary $I(1)$

$$y_t = \mu + y_{t-1} + u_t,$$

where  $\{u_t\}$  is stationary process. 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 condition:

$$y_0 \begin{cases} \text{fixed} \\ \text{random Variable } N(0, \cdot). \end{cases}$$

This completes the specification of the process. The differenced series is then  $\Delta y_t = \mu + u_t$ . Any

shocks have permanent affects

$$y_t = y_0 + t\mu + \sum_{s=1}^t u_s.$$

Another important differences between (1) and (2) is that in case (1),  $\text{var } y_t = \sigma^2$ ,  $\forall t$ , while in case (2),

$$\text{var } y_t = \sigma^2 \cdot t \rightarrow \infty \text{ as } t \rightarrow \infty \quad (20.1)$$

In (2) both mean and variance explode, in (1) only the mean.

Martingales

$$E[y_t | y_{t-1}, \dots] = y_{t-1} \text{ a.s.},$$

i.e.,  $y_t = y_{t-1} + u_t$ , where  $u_t$  may be heterogeneous but uncorrelated. Hall (1978): Consumption is a martingale. Fama: Stock prices are martingales.

Note that differencing in (1) 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 (2) is not perfect. In any case, want to know which one is the case.

Effects of time trend: you get superconsistent  $T^{3/2}$  estimates of  $\beta$ , but still Gaussian  $t$ -tests still valid.

Effects of unit root: superconsistent estimates, but with nonstandard distributions:  $t$ -tests not valid!

Suppose  $y_t = \rho y_{t-1} + u_t$ ,  $u_t \sim 0, \sigma^2$ . Then,

$$\hat{\rho}_{\text{OLS}} = \frac{\sum_{t=2}^T y_t y_{t-1}}{\sum_{t=2}^T y_{t-1}^2} \rightarrow \rho, \quad \forall \rho.$$

If  $\rho < 1$

$$\sqrt{T}(\hat{\rho} - \rho) \rightarrow N(0, 1 - \rho^2),$$

and note that the scale parameter  $\sigma^2$  doesn't appear on the limiting distributions. This is one difference from the regression case.

If  $\rho = 1$ ,  $1 - \rho^2 = 0$ , so the implied variance above is zero. So what happens in this case? If  $\rho = 1$ ,

$$T(\hat{\rho} - \rho) \xrightarrow{D} X,$$

where  $X$  is not Gaussian; it is asymmetric and in fact  $E(\hat{\rho}) < 1$ ,  $\forall T$ . The rate of convergence is faster but the asymptotic distribution is non standard.

### Testing for unit roots

Dickey–Fuller (1981) derived the distribution of  $\hat{\rho}$  and  $t_\rho$  when  $\rho = 1$ , and they tabulated it. Critical values are quite different from the usual  $t$ -tables. Critical values can be large negative numbers  $\ll -1.96$ .

Can also rewrite as test of  $\gamma = 0$  in the model

$$\Delta y_t = \gamma y_{t-1} + u_t.$$

We do a one-sided test vs.  $\gamma < 0$ . Augmented D–F

$$\Delta y_t = \mu + \gamma y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + \eta_t.$$

This allows for stationary dynamics in  $u_t$ .

### Cointegration

Suppose  $y_t$  and  $x_t$  are  $I(1)$  but there is a  $\beta$  such that  $y_t - \beta x_t$  is  $I(0)$ , then  $y_t$ ,  $x_t$  are cointegrated. Note that  $\beta$  is not necessarily unique. For example, consumption and income are both generally  $I(1)$  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.

### GARCH Models

For financial data (specifically, stock returns), the scale of the process appears to have some structure. Engle (1982) introduced the following class of models

$$y_t = \beta' x_t + \overbrace{\varepsilon_t \sigma_t}^{u_t},$$

where  $\varepsilon_t$  is i.i.d.  $(0,1)$ , while  $\sigma_t^2 = \text{Var}[y_t | \mathcal{F}_{t-1}]$  is the (time-varying) conditional variance.

GARCH( $p, q$ )

$$A(L)\sigma_t^2 = \alpha_0 + C(L)u_{t-1}^2,$$

where  $A$  and  $B$  are lag polynomials. For example,

$$\sigma_t^2 = \alpha_0 + \gamma u_{t-1}^2,$$

which is the ARCH(1).

- Nonlinear process
- Generates volatility clustering, i.e., some persistence in volatility
- The unconditional distribution of  $u_t$  is thick-tailed
- Estimation by ML: nonlinear. More efficient than OLS for  $\beta$ .