Contents

Chapter 1

Ordinary Least Squares

Ordinary Least Squares (OLS) is still the workhorse of econometrics. I develop a quite succinct presentation of the estimator, its assumptions and its properties. It is implemented in Stata through the `regress` command.

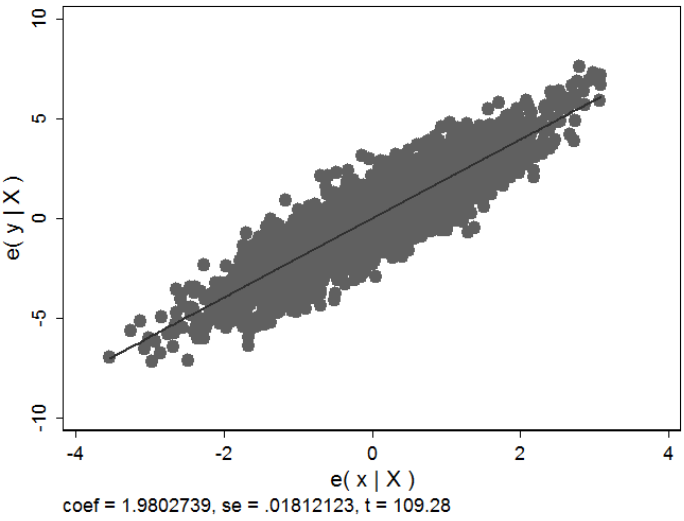

Figure 1.1: OLS regression

1.1 Gauss-Markov Assumptions

The OLS estimator that will briefly derived is based upon the well known Gauss-Markov assumptions. Under these assumptions, the OLS estimator is the Best Linear Unbiased Estimator (BLUE)

1. Linearity: implies that the marginal effect does not depend on the level of regress

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{X}} = \boldsymbol{\beta} \tag{1.1}$$

2. Strict Exogeneity: the conditional mean of the error term is zero

$$E[\varepsilon_i|\boldsymbol{X}] = 0 \tag{1.2}$$

   - unconditional zero mean of the error term

$$E[E[\boldsymbol{\varepsilon}|\boldsymbol{X}]] = E[\boldsymbol{\varepsilon}] \tag{1.3}$$

     (*this is possible by the law of total expectations*)

   - orthogonality of the error term: the basis of the further discussed method of the moments

$$\begin{aligned} E[x_{jk}\varepsilon_i] &= E[E[x_{jk}\varepsilon_i|x_{jk}]] \\ &= E[x_{jk}E[\varepsilon_i|x_{jk}]] \\ &= 0 \end{aligned} \tag{1.4}$$

3. No multicollinearity: The rank of the $n \times K$ data matrix, $\boldsymbol{X}$, is K with probability 1

4. Spherical error variance

   - homoscedasticity

$$E[\boldsymbol{\varepsilon}^2|\boldsymbol{X}] = \sigma^2 \boldsymbol{I}_N > 0 \tag{1.5}$$

   - no correlation between observations

$$E[\varepsilon_i\varepsilon_j|\boldsymbol{X}] = 0 \tag{1.6}$$

1.2 OLS estimator

Assuming the four Gauss-Markov assumptions hold we can now derive our OLS estimator. Let us start with the following multiple regression

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + u_i \tag{1.7}$$

where u_i is a normally distributed zero mean error term. You can rewrite (1.7) in compact matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{1.8}$$

where $\mathbf{X}$ is an $n \times k$ matrix of observations on the explanatory variables

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \mathbf{x_3} \\ ... \\ \mathbf{x_k} \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & ... & x_{1k} \\ 1 & x_{22} & x_{23} & ... & x_{2k} \\ . & & & & \\ . & & & & \\ . & & & & \\ x_{n1} & x_{n2} & x_{n3} & ... & x_{nk} \end{bmatrix}$$

$\boldsymbol{\beta}$ is a $1 \times k$ vector of coefficients and $\boldsymbol{u}$ is an $n \times 1$ vector of unobservable disturbances.

Estimation of $\boldsymbol{\beta}$ proceeds by minimizing the sum of squared residuals (SSR). Define the sum if squared residuals function for any possible $k \times 1$ parameter vector $\boldsymbol{b}$ as

$$\mathbf{b} \equiv \arg\min SSR(\boldsymbol{\beta}) = \arg\min(\boldsymbol{u'u}) \tag{1.9}$$

since $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ we can define (1.9) as

$$(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{y'y} - \boldsymbol{\beta' X' X\beta} + \boldsymbol{2y(X\beta)'} \tag{1.10}$$

notice that the scalar[1] $\boldsymbol{(X\beta)'y = y'X\beta}$. Taking the first order conditions we get

$$\begin{aligned} \frac{\partial SSR\boldsymbol{\beta}}{\partial \beta} &= \mathbf{0} \\ \boldsymbol{-2(X'X)\beta + 2X'y} &= \mathbf{0} \\ \boldsymbol{(X'X)\beta} &= \boldsymbol{X'y} \\ \boldsymbol{(X'X)^{-1}(X'X)\beta} &= \boldsymbol{(X'X)^{-1}X'y} \end{aligned} \tag{1.11}$$

which yields the well-known $\hat{\boldsymbol{\beta}}$ OLS estimator

$$\hat{\boldsymbol{\beta}} = \boldsymbol{(X'X)^{-1}X'y} \tag{1.12}$$

[1]This can easily be verified if one looks at the dimensions of the $\boldsymbol{X}$, $\boldsymbol{y}$ and $\boldsymbol{\beta}$ matrices.

The variance of the OLS estimator is derived as follows

$$\begin{aligned} V[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &= V\left[(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}|\boldsymbol{X}\right] \\ &= V\left[(\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{X\beta}+\boldsymbol{u})|\boldsymbol{X}\right] \\ &= V\left[(\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta}+(\boldsymbol{X'X})^{-1}\boldsymbol{X'u}|\boldsymbol{X}\right] \\ &= V\left[\boldsymbol{\beta}+(\boldsymbol{X'X})^{-1}\boldsymbol{X'u}|\boldsymbol{X}\right] \\ &= V\left[(\boldsymbol{X'X})^{-1}\boldsymbol{X'u}|\boldsymbol{X}\right] \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\left[V[\boldsymbol{u}|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X'X})^{-1}\right] \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\sigma^2\boldsymbol{I_N})\boldsymbol{X}(\boldsymbol{X'X})^{-1} \\ &= \sigma^2(\boldsymbol{X'X})^{-1}\boldsymbol{X'X}(\boldsymbol{X'X})^{-1} \\ &= \sigma^2(\boldsymbol{X'X})^{-1} \end{aligned} \tag{1.13}$$

1.2.1 Small sample properties of the OLS estimator

Unbiasedness

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= [\boldsymbol{X'X}]^{-1}\boldsymbol{X'y} \\ &= [\boldsymbol{X'X}]^{-1}\boldsymbol{X'}[\boldsymbol{X\beta}+\boldsymbol{u}] \\ &= [\boldsymbol{X'X}]^{-1}\boldsymbol{X'X\beta}+[\boldsymbol{X'X}]^{-1}\boldsymbol{X'u} \\ &= \boldsymbol{\beta}+[\boldsymbol{X'X}]^{-1}\boldsymbol{X'u} \end{aligned} \tag{1.14}$$

taking the expectations of $\hat{\boldsymbol{\beta}}$

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &= E[\boldsymbol{\beta}]+E\left([\boldsymbol{X'X}]^{-1}\boldsymbol{X'u}|\boldsymbol{X}\right) \\ &= \boldsymbol{\beta}+[\boldsymbol{X'X}]^{-1}\boldsymbol{X'}E[\boldsymbol{u}|x] \end{aligned} \tag{1.15}$$

since the orthogonality condition imposes $E[\boldsymbol{u}|x]=0$

$$E[\hat{\boldsymbol{\beta}}]=\boldsymbol{\beta} \tag{1.16}$$

1.2.2 Asymptotic distribution of the OLS estimator

The following assumptions are required to derive the asymptotic properties of the OLS estimator:

1. the d.g.p follows $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$
2. data are independent over i with $E[\boldsymbol{u}|\boldsymbol{X}] = \boldsymbol{0}$ and $E[\boldsymbol{uu'}|\boldsymbol{X}] = \boldsymbol{\Omega} = diag[\sigma_i^2]$
3. the matrix $\mathbf{X}$ is of full rank to that $\boldsymbol{X\beta^{(1)}} = \boldsymbol{X\beta^{(2)}}$ iff $\boldsymbol{\beta^{(1)}} = \boldsymbol{X\beta^{(2)}}$
4. the $K \times K$ matrix

$$\boldsymbol{M_{xx}} = p\lim N^{-1}\boldsymbol{X'X} = p\lim N^{-1}\sum_{i=1}^{N}\boldsymbol{x_i x_i'} = \lim N^{-1}\sum_{i=1}^{N}E[\boldsymbol{x_i x_i'}] \tag{1.17}$$

   exists and is finite nonsigular

5. the variance covariance matrix

$$\boldsymbol{M_{x\Omega x}} = p\lim N^{-1}\boldsymbol{X'uu'X} = p\lim N^{-1}\sum u_i^2\boldsymbol{x_i x_i'} = \lim N^{-1}\sum E[u_i^2\boldsymbol{x_i x_i'}] \tag{1.18}$$

Then the OLS estimator $\hat{\boldsymbol{\beta}}$ is **consistent** for $\boldsymbol{\beta}$ and

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to^d \mathcal{N}[\boldsymbol{0}, \boldsymbol{M_{xx}^{-1}M_{x\Omega x}M_{xx}^{-1}}] \tag{1.19}$$

Assumption 1 ensures $E[\boldsymbol{y}|X] = \boldsymbol{X\beta}$ and permits heterostedastic errors with variance σ_i^2, more general than the homoscedastic uncorrelated errors that restrict $\boldsymbol{\Omega} = \sigma^2\boldsymbol{I}$. Assumption 3 rules out perfect collinearity among the regressors. Assumption 4 leads to the rescaling of $\boldsymbol{X'X}$ by N^1. Note that by a law of large numbers $p\lim = \lim E$.

Asymptotic Distribution The asymptotic distribution is interpreted as being applicable in large samples, meaning samples large enough for the limit distribution to be a good approximation but not so large that $\hat{\boldsymbol{\beta}} \to^p \boldsymbol{\beta}$ as then its asymptotic distribution would be degenerate. The asymptotic distribution is obtained from (1.19) by division by $\sqrt{N}$ and addition of $\boldsymbol{\beta}$. This yields the asymptotic distribution

$$\hat{\boldsymbol{\beta}} \sim^a \mathcal{N}[\boldsymbol{\beta}, \boldsymbol{N^{-1}M_{xx}^{-1}M_{x\Omega x}M_{xx}^{-1}}] \tag{1.20}$$

which can also be expressed as

$$\hat{\boldsymbol{\beta}} \sim^a \mathcal{N}[\boldsymbol{\beta}, \boldsymbol{[X'X]^{-1}X'\Omega X[X'X]^{-1}}] \tag{1.21}$$

where $\boldsymbol{[X'X]^{-1}X'\Omega X[X'X]^{-1}}$ is $V[\hat{\boldsymbol{\beta}}]$. With homoscedastic errors (1.22) is simplified to

$$\hat{\boldsymbol{\beta}} \sim^a \mathcal{N}[\boldsymbol{\beta}, \sigma^2\boldsymbol{[X'X]^{-1}}] \tag{1.22}$$

because $\boldsymbol{\Omega} = \sigma^2\boldsymbol{I}$ so that $\boldsymbol{X'\Omega X} = \sigma^2\boldsymbol{X'X}$ and hence $\boldsymbol{M_{x\Omega x}} = \sigma^2\boldsymbol{M_{xx}}$.

1.2.3 Robust Standard Errors

Heteroskedasticity-robust standard errors can be estimated for an OLS regression and are implemented in Stata as the `, robust` option in the `regress` command. In order to do this however we first need to produce an estimate of the OLS variance, which will be given by the sandwich estimate:

$$\hat{V}[\hat{\boldsymbol{\beta}}] = N^{-1}\hat{\boldsymbol{M}}_{\boldsymbol{xx}}^{-\mathbf{1}}\hat{\boldsymbol{M}}_{\boldsymbol{x\Omega x}}\hat{\boldsymbol{M}}_{\boldsymbol{xx}}^{-\mathbf{1}} \tag{1.23}$$

the trouble here being to estimate the filling of the sandwich $\hat{\boldsymbol{M}}_{\boldsymbol{x\Omega x}}$. One solution is to use White's robust standard errors which set $\hat{\boldsymbol{M}}_{\boldsymbol{x\Omega x}} = N^{-1}\sum_i \hat{u}_i^2 \boldsymbol{x_i x_i'}$ which can be seen as a sample transposition from assumption 5.

Chapter 2

Generalized Least Squares

Error term normality condition can be considered unrealistic in several applications of economic theory. Fortunately, this assumption may be relaxed in order to accommodate for both heteroskedasticity and serial correlation. The Generalized Least Squares (GLS) estimator, of which the Feasable GLS and the Weighted Least Squares estimators are particular cases provide us with a robust variance-covariance matrix which is not only consistent but also *efficient*, re-enabling inference over estimated coefficients. Stata has implemented a series of routines to address both heteroskedasticity and serial correlation issues.

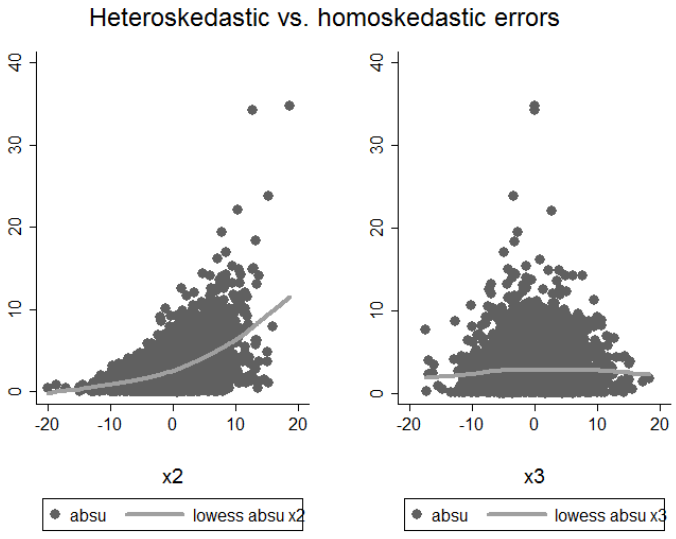

Figure 2.1: Residuals plot

2.1 Generalized Least Squares

If we assume $\boldsymbol{\Omega \neq \sigma^2 I}$ the OLS estimator will be inefficient and therefore needs weighting. That is done with some simple algebra on the d.g.p. Needless is to say that $\boldsymbol{\Omega} = E[\boldsymbol{uu'}|\boldsymbol{X}] = diag[\sigma_i^2]$ needs to be nonsigular

$$\boldsymbol{\Omega^{-1/2} y = \Omega^{-1/2} X\beta + \Omega^{-1/2} u} \tag{2.1}$$

reparametrizing $\boldsymbol{\Omega^{-1/2} y \equiv \tilde{y}}$; $\boldsymbol{\Omega^{-1/2} X \equiv \tilde{X}}$ and $\boldsymbol{\Omega^{-1/2} u \equiv \tilde{u}}$ we get

$$\boldsymbol{\tilde{y} = \tilde{X}\beta + \tilde{u}} \tag{2.2}$$

which produces the $\boldsymbol{\hat{\beta}_{GLS}}$ estimator

$$\begin{aligned} \boldsymbol{\hat{\beta}_{GLS}} &= \boldsymbol{[\tilde{X}'\tilde{X}]^{-1}\tilde{X}'\tilde{y}} \\ &= \boldsymbol{[(\Omega^{-1/2}X)'(\Omega^{-1/2}X)]^{-1}(\Omega^{-1/2}X)'\Omega^{-1/2}y} \\ &= \boldsymbol{[X'\Omega^{-1}X]^{-1}X'\Omega^{-1}y} \end{aligned} \tag{2.3}$$

This would all be fine if $\boldsymbol{\Omega}$, that is to say the shape of heterogeneity in the actual d.g.p, was known. Instead, we specify that $\boldsymbol{\Omega = \Omega(\gamma)}$, where $\boldsymbol{\gamma}$ is a finite-dimensional parameter vector, obtain a consistent estimate $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$, and form $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}}(\boldsymbol{\gamma})$. This is called the feasible generalized least squares (FGLS). FGLS is implemented in Stata for panel data only through the `xtfgls` command. Its estimator is simply

$$\boldsymbol{\hat{\beta}_{FGLS} = [X'\hat{\Omega}^{-1}X]^{-1}X'\hat{\Omega}^{-1}y} \tag{2.4}$$

where $\hat{\boldsymbol{\Omega}} = diag[\hat{u}_i^2]$ and $\hat{u}_i^2 = (y - \boldsymbol{x'\hat{\beta}})^2$. Note that we cannot replace $\boldsymbol{\Omega}$ by $\hat{\boldsymbol{\Omega}} = diag[\hat{u}_i^2]$ as this yields an inconsistent estimator.

Similarly, the variance of the $\hat{\beta}_{FGLS}$ will be

$$\hat{V}[\boldsymbol{\beta_{GLS}}] = \boldsymbol{[X'\hat{\Omega}^{-1}X]^{-1}} \tag{2.5}$$

and finally, the asymptotic distribution given by

$$\sqrt{N}(\boldsymbol{\hat{\beta}_{FGLS} - \beta}) \rightarrow^d \mathcal{N}[\boldsymbol{0}, (p\lim N^{-1}\boldsymbol{X'\Omega^{-1}X})^{-1}] \tag{2.6}$$

2.2 Weighted Least Squares

If $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\gamma})$ is misspecified, the FGLS will still be consistent but will yield the wrong variance in (2.6). Fortunately, a robust estimate of the variance of the GLS estimator can be found even if $\boldsymbol{\Omega}(\boldsymbol{\gamma})$is misspecified. Specifically, define $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\gamma})$ to be a **working variance matrix** that does not necessarily equal the true variance matrix $\boldsymbol{\Omega} = E[\boldsymbol{uu'}|\boldsymbol{X}]$. Form an estimate $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ is an estimate of $\boldsymbol{\gamma}$. Then use weighted least squares with weighting matrix $\hat{\boldsymbol{\Sigma}}^{-1}$. This yields the weighted least-squares (WLS) estimator

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{WLS}} = [\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y} \tag{2.7}$$

Statistical inference is then done without the assumption that $\boldsymbol{\Sigma} = \boldsymbol{\Omega}$, the true variance. The WLS variance is derived in a similar fashion as its OLS counterpart in equation (**??**).

$$\begin{aligned} V[\hat{\boldsymbol{\beta}}_{\boldsymbol{WLS}}|\boldsymbol{X}] &= V\left([\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}[\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}]\right) \\ &= V\left([\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]^{-1}[\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]\boldsymbol{\beta}+[\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{u}\right) \\ &= V[\boldsymbol{\beta}]+[\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}V[u]\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}[\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]^{-1} \\ &= [\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Omega}}\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}[\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]^{-1} \end{aligned} \tag{2.8}$$

nice, isn't it? Notice that the conditional variance of the error $\hat{\boldsymbol{\Omega}}$ is weighted by $\hat{\boldsymbol{\Sigma}}^{-1}$, where $\hat{\boldsymbol{\Omega}}$ is such that

$$p\lim N^{-1}\boldsymbol{X'}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X} = p\lim N^{-1}\boldsymbol{X'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \tag{2.9}$$

In the heteroscedastic case $\hat{\boldsymbol{\Omega}} = diag[\hat{u}_i^{*2}]$, where $\hat{u}_i^{*2} = y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}_{\boldsymbol{WLS}}$

Chapter 3

Model Misspecification

Model misspecification in regression has long been a well-recognized research problem. Depending on the applications, a misidentification of a variable X as a (or even the) cause of Y may result in severe consequences. We present six forms of model misspecification.

3.1 Inconsistency of OLS

Inconsistency is not *per se* a cause of misspecification but rather a symptom of eitheir a poor choice of a model functional form or of model endogeneity. Two key conditions are required to demonstrate the consistency of the OLS estimator

1. the d.g.p. is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$
2. the d.g.p. is such that $p\lim N^{-1}\boldsymbol{X'u} = \boldsymbol{0}$

so that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + [\boldsymbol{X'X}]^{-1}\boldsymbol{X'u} \rightarrow^{p} \boldsymbol{\beta} \tag{3.1}$$

Now, if the $p\lim$ does not converge to the true parameter value $\boldsymbol{\beta}$ we have inconsistency. Notice that the misspecification can raise either from (1), that is to say, assuming an improper shape for the d.g.p. or from (2) if the $p\lim$ of the error term does not converge to zero, meaning endogeneity.

3.2 Functional form misspecification

As mentioned, a poor choice of functional form, say using a linear function to explain an intrinsically nonlinear relationship will inevitably produce biased estimates.

3.3 Endogeneity

Endogeneity is, possibly alongside with selection bias, the boogeyman of empiricists. Endogeneity means that $E[\boldsymbol{u}|\boldsymbol{X}] \neq 0$, i.e. that the error term is correlated with at least one of the regressors. Endogeneity is often curbed with resources to Instrumental Variables (to be exposed in chapter 9) and can be tested via the **Hausman test for endogeneity** which will be presented later on (see section 10.2). In panel models, endogeneity can be removed provided it is time-invariant per unit of analysis through the Fixed Effects[1] estimator.

3.4 Omitted variables

The omission of variables is often pointed out as the first source of biased estimators. The lack of controls, for example, will inflate the effects of the study variable as its value will be intertwined with the contributions of other covariates. The solution for such horrendous issue? Just add variables at you own leisure. Well, not quite, add covariates that make sense in terms of economic theory and then use a Wald test to check whether the extra covariates add value to the model. More specifically, we will be wanting to measure the impact of the new variables in the model. Should they be irrelevant and their coefficient will be very close to zero. In this sense the t-test, which is a squared version of the Wald test, will measure the distance of the new coefficient to zero

$$t^2 = \frac{\hat{\beta}^2_{test} - 0}{V(\hat{\beta}_{test})} \sim \chi^2(h) \tag{3.2}$$

3.5 Pseudo-true value

In the omitted variables example the least-squares estimator is subject to confounding in the sense that it does not estimate β, but instead estimates a function of β, δ, and α. The $p\lim$ of $\hat{\beta}$ of $\beta^* = (\beta + \delta\alpha)$ is referred to as the pseudo-true value. From the pseudo-true value one can obtain the distribution of $\hat{\beta}$ even thought it is inconsistent for β.

[1] The fixed effects estimator is invokable in Stata by the `xtreg yvar xvars, fe`

3.6 Parameter heterogeneity

So far we have assumed error terms could vary across units of observation in what we call idiosyncratic errors, as the i subscript in the regression model has been proof

$$y_i = \boldsymbol{x_i'}\beta + u_i \tag{3.3}$$

Nevertheless it is quite conceivable that the slope of the parameter may not be equal for all individual. If we then allow the marginal effect $E[y_i|\boldsymbol{x_i}] = \beta$ to vary we will be applying a random coefficients model. The random coefficients model specifies β_i to be independently and identically distributed over i with distribution that does not depend on the observables $\boldsymbol{x_i}$. In such case the d.g.p. can be written as

$$y_i = \boldsymbol{x_i'}\beta + (u_i + \boldsymbol{x_i'}(\beta_i - \beta)) \tag{3.4}$$

where we assume the regressors $\boldsymbol{x_i}$ to be uncorrelated with the error term $(u_i + \boldsymbol{x_i'}(\beta_i - \beta))$.

Chapter 4

Panel Data Estimators

In cross-section models data is solely observed at a given point in time which impedes dynamic analyses of reality or the study of behavioural persistency. With longitudinal models this limitation is overcome as the unit of observation (individuals, firms, countries, etc.) is repeated over T periods. Four panel data estimators are presented in this chapter: pooled OLS, between and within estimator, first-differences and random effects. Because of its relevance for practitioners greater relevance will be given to the fixed effects (or within) estimator.

The linear panel data models assume the following specification

$$y_{it} = \alpha_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it} \tag{4.1}$$

where ε_{it} is iid over i and t. The fixed effects (fe) model (4.1) treats α_i as an unobserved random variable that is potentially correlated with the regressors $\boldsymbol{x}_{it}$ If fixed effects are present and correlated with $\boldsymbol{x}_{it}$ then models such as pooled ols will be inconsistent.

4.1 Pooled OLS

The pooled OLS estimator is obtained by stacking the data over i and t into one long regression with $N \times T$ observations. The pooled OLS estimator relies on $Cov[u_{it}, \boldsymbol{x}_{it}] = \mathbf{0}$ to achieve consistency

$$y_{it} = \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + u_{it}, \qquad i = 1, ...N, \qquad t = 1, ..., T. \tag{4.2}$$

In practice however, the $Cov[u_{it}, \boldsymbol{x}_{it}] = \mathbf{0}$ assumption often seems unrealistic. By taking inter-temporal independence given one is implicitly considering contemporaneous individual behaviour to be unrelated from past and future ones.

A more reasonable approach is to expect considerable correlation in y over time, so that $Corr[u_{it}, u_{is}]$ is high. I other words, the (likely) existence of fixed effects in a models specified as pooled OLS will produce inconsistent estimates. This can be verified by disaggregating the u_i error term

$$y_{it} = \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + (\alpha_i - \alpha + \varepsilon_{it}). \tag{4.3}$$

Then the pooled OLS regression of y_{it} on $\boldsymbol{x}_{it}$ and an intercept leads to an inconsistent estimator of $\boldsymbol{\beta}$ if the individual effect α_i is correlated with the regressors $\boldsymbol{x}_{it}$, since such correlation implies that the combined error term $\alpha_i - \alpha + \varepsilon_{it}$ is correlated with the regressors via the constant α.

4.2 Between Estimator

The between estimator is a stupid estimator. It averages data at individual level, uses OLS for estimation.

$$\begin{aligned} \bar{y}_i &= \alpha_i + \bar{\boldsymbol{x}}'_i\boldsymbol{\beta} + \bar{\varepsilon}_i \\ \bar{y}_i &= \alpha + \bar{\boldsymbol{x}}'_i\boldsymbol{\beta} + (\alpha_i - \alpha + \bar{\varepsilon}_i), \qquad i = 1, ...N. \end{aligned} \tag{4.4}$$

where $\bar{y}_i = T^{-1}\sum_t y_{it}$, $\bar{\varepsilon}_i = T^{-1}\sum \varepsilon_{it}$ and $\bar{\boldsymbol{x}}_i = T^{-1}\sum_t \boldsymbol{x}_{it}$.

The between estimator collapses individual-level variability in one data point ignoring changes across time and treating data as cross-sectional. It is consistent if the regressors $\bar{\boldsymbol{x}}_i$ are independent of the composite error $\bar{\varepsilon}_i$ from (4.4). However, as said for the pooled OLS model, such independence assumption is often unrealistic, breaking apart the robustness of the model.

4.3 Within Estimator

The fixed effects or within estimator assumes unobserved individual-level heterogeneity, as worker or team ability, to remain constant over time. Then longitudinal data allows the removal of such unobservables simply by averaging it out both depent variable and covariates over time. Such procedure will wipe out all time-constant unobservables (unfortunately it

will also wipe out constant observables). Derivation is presented bellow

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it} &= \sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{N}\sum_{t=1}^{T}\boldsymbol{x}_{it}\beta + \sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{it} \\
\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y}_i) &= \sum_{i=1}^{N}(\alpha_i-\bar{\alpha}_i) + \sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x_{it}}-\bar{\boldsymbol{x}}_i)'\boldsymbol{\beta} + \sum_{i=1}^{N}\sum_{t=1}^{T}(\varepsilon_{it}-\bar{\varepsilon}_i) \\
\sum_{i=1}^{N}\ddot{y}_i &= \sum_{i=1}^{N}\ddot{\boldsymbol{x}}_i + \sum_{i=1}^{N}\ddot{\varepsilon}_i
\end{aligned} \tag{4.5}$$

which can be written in the following matrix notation

$$\ddot{\boldsymbol{y}} = \ddot{\boldsymbol{X}}\boldsymbol{\beta} + \ddot{\boldsymbol{\varepsilon}} \tag{4.6}$$

solving the problem of the minimisation of the sum of squared residuals as for the OLS estimator we get the fe estimator $\boldsymbol{\beta}_{fe}$

$$\hat{\boldsymbol{\beta}}_{fe} = (\ddot{\boldsymbol{X}}'\ddot{\boldsymbol{X}})^{-1}\ddot{\boldsymbol{X}}'\ddot{\boldsymbol{y}} \tag{4.7}$$

which in a more intuitive index notation equals

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{fe} &= \sum_{i=1}^{N}(\ddot{\boldsymbol{x}}_i\ddot{\boldsymbol{x}}_i')^{-1}\sum_{i=1}^{N}\ddot{\boldsymbol{x}}_i'\ddot{y} \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T}[(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)']^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)(y_{it}-\bar{y}_i).
\end{aligned} \tag{4.8}$$

Only the within estimator and the first-differences estimator (presented in section 4.4) provide consistent estimates for fixed- random- and mixed-effects model specifications. Notice that in spite of this result, relative efficiencies of the within estimator will vary, as discussed in section 4.3.4.

4.3.1 Estimating the fixed effect

In some cases it may be of interest to compute the estimate from the fixed effect itself. Although such computation may seem at a first glance hard to achieve, after all the α_i term is dropped early in the estimation process, in practice a little algebra is all that is required to obtain $\hat{\alpha}_i$. Specifically, by decomposing $y_i t$ from the estimator equation we get

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{fe} &= \sum_{i=1}^{N}\sum_{t=1}^{T}[(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)']^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)\left[(\alpha_i+\boldsymbol{x}'_{it}\hat{\boldsymbol{\beta}}_{fe})-\bar{y}_i\right] \\
\hat{\boldsymbol{\beta}}_{fe} &= (\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)^{-1}\left[(\alpha_i+\boldsymbol{x}'_{it}\hat{\boldsymbol{\beta}}_{fe})-\bar{y}_i)\right] \\
(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)'\hat{\boldsymbol{\beta}}_{fe} &= \alpha_i+\boldsymbol{x}'_{it}\hat{\boldsymbol{\beta}}_{fe}-\bar{y}_i \\
-\boldsymbol{x}'_i\hat{\boldsymbol{\beta}}_{fe} &= \alpha_i-\bar{y}_i
\end{aligned}$$

which is equal to

$$\hat{\alpha}_i = \bar{y}_i - \boldsymbol{x}'_i\hat{\boldsymbol{\beta}}_{fe}. \tag{4.9}$$

4.3.2 Within estimator Consistency

The within estimator of $\boldsymbol{\beta}$ is consistent if $p\lim(NT)^{-1}\sum_i\sum_t(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)(\varepsilon_{it}-\bar{\varepsilon}_i)=\mathbf{0}$. This should happen in either $T\to\infty$ or (more likelly for panels) $N\to\infty$ and

$$E[\varepsilon_{it}-\bar{\varepsilon}|\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i]=0 \tag{4.10}$$

whose proof is identical to the OLS estimator consistency proof.

4.3.3 Variance of the Within Estimator

We begin with the basic panel model from (4.1)

$$y_{it}=\alpha_i+\boldsymbol{x}_{it}\boldsymbol{\beta}+\varepsilon_{it} \tag{4.11}$$

to which the individual averages are subtracted from, as shown in (4.5)

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y}_i)=\sum_{i=1}^{N}(\alpha_i-\bar{\alpha}_i)+\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x_{it}}-\bar{\boldsymbol{x}}_i)\boldsymbol{\beta}+\sum_{i=1}^{N}\sum_{t=1}^{T}(\varepsilon_{it}-\bar{\varepsilon}_i). \tag{4.12}$$

The previous equation can however be expressed in the following matrix form

$$\boldsymbol{Qy_i}=\boldsymbol{QX_i\beta}-\boldsymbol{Q\varepsilon_i} \tag{4.13}$$

where Q is a $T\times T$ matrix whose permutation with variables creates their deviations for the mean

$$\boldsymbol{QW_i}=\boldsymbol{IW_i}-\boldsymbol{e\bar{w}'_i} \tag{4.14}$$

where $\boldsymbol{I}$ is $T \times T$ identity and $\boldsymbol{e}$ is a $T \times 1$ vector vector of ones. As in the index notation case $\boldsymbol{Q\alpha}_i$ is absorbed by its mean due to lack of intertemporal variance. The estimation of the variance is computed as for the OLS estimator and produces

$$V[\hat{\boldsymbol{\beta}}_{fe}] = \left[\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{X}_i\right]^{-1} V[\boldsymbol{Q\varepsilon}_i|\boldsymbol{X}_i]\boldsymbol{Q}\boldsymbol{X}_i \left[\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{X}_i\right]^{-1} \tag{4.15}$$

4.3.4 Hausman Test

One of the possible applications of the Hausman test (presented in detail in section 10.2 of chapter 10) is to identify the existence of fixed effects in panels. This is done by comparing the consistency of the random and fixed effects (within) estimator. The fe estimator will be consistent in the presence of either fixed or random effects (albeit inefficient in the latter). The RE estimator however will only be consistent (and efficient) in the presence of random effects, making biased estimates for fixed effects. One can therefore test the presence of fixed effects by hypothesising on the statistical significance of the *difference between estimators*.

We begin by assuming that the true model is the random effects model with α_i iid $[0, \sigma_\alpha^2]$ uncorrelated regressors and ε_{it} iid $[0, \sigma_\varepsilon^2]$. Then the estimator $\tilde{\beta}_{RE}$ is fully efficient and the **Hausman test** statistic is

$$H = (\tilde{\beta}_{1,RE} - \hat{\beta}_{1,fe})'[\hat{V}[\hat{\beta}_{1,fe}] - \hat{V}[\tilde{\beta}_{1,RE}]]^{-1}(\tilde{\beta}_{1,RE} - \hat{\beta}_{1,fe}) \tag{4.16}$$

where β_1 denotes the subcomponent of β corresponding to the time-varying regressors since **only those can be estimated by the within estimator**. This test statistic is asymptotically Chi^2 ($\dim[\beta_1]$) distributed under the null hypothesis. Its full derivation is available in chapter 10, section 10.2. Stata provides a somewhat sluggish implementation of the Hausman test. The test is done via a two-step procedure. In the first step the efficient model (here RE) is estimated and coefficients stored. In the second step the inefficient but consistent model is estimated. After this the command `hausman <efficient>` will test the null that the true model is the RE.

4.4 First-Differences Estimator

The first-differences begins by using the individual-specific effects model (4.1) which is then lagged and subtracted

$$y_{it} - y_{i,t-1} = (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1}), \qquad i = 1, ...N, \qquad t = 2, ...T \tag{4.17}$$

as the individual specific α_i term is cancelled out.

4.5 Random Effects estimator

An alternative variant of the model (4.1) is the random effects (RE) model which assumes that the unobservable individual effects α_i are random variables distributed independently of the regressors

$$\begin{aligned} \alpha_i &\sim [\alpha, \sigma_\alpha^2] \\ \varepsilon_{it} &\sim [0, \sigma_\varepsilon^2]. \end{aligned} \tag{4.18}$$

The RE model usually subsumes α_i into a composite error term $u_{it} = \alpha_i + \varepsilon_{it}$. However, α_i is often normalised to zero mean through the addition of a nonrandom scalar intercept coefficient μ so that

$$y_{it} = \mu + \boldsymbol{x}_{it}'\beta + u_{it}. \tag{4.19}$$

Then we can say that

$$Cov\left[(\alpha_i + \varepsilon_{it}), (\alpha_i + \varepsilon_{is})\right] = \begin{cases} \sigma_\alpha^2 \\ \sigma_\alpha^2 + \sigma_\varepsilon^2. \end{cases} \tag{4.20}$$

The RE model therefore imposes the constraint that the composite error term u_{it} is **equicorrelated**, since $Corr[u_{it}, u_{is}] = \sigma_\alpha^2/[\sigma_\alpha^2 + \sigma_\varepsilon^2]$ for $t \neq s$ does not vary over time with the time difference $t - s$.

4.6 Unbalanced Panel Data

Real world data sets often face drop offs of individuals over time as well as missing years. In such cases the fe estimator will remain consistent if the strong exogeneity assumption becomes

$$E[u_i t | \alpha_i, \boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}, d_{i1}, ..., , d_{iT}] = 0 \tag{4.21}$$

where d_{it} is a dummy equal to one if the itth observation is observed and zero otherwise.

At times it may be convenient to convert unbalanced panels into balanced ones by choosing only those individuals with observations in every year. Such procedure will nevertheless lead to loss of efficiency in estimation and potentially drive an attrition bias. Attrition bias occurs when observations are lost in an **non-random** manner. For example, individuals with unusually low incomes may be more likely to miss out some observations in the panel.

Alternatively to deleting individuals with missing data one can also consider data imputation based on existing data.

4.7 Dynamic Models

We now add a dynamic component (one of the regressors is the lagged dependent variable) to the panel model with individual-specific effects. Notice that by adding this covariate **all** previously presented panel estimates will be **inconsistent**. The model is hence defined as

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{x}'_{i,t-1}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}. \tag{4.22}$$

We begin by assuming stationarity, $|\gamma| < 1$, and absence of serial correlation in ε_{it}.

An important result is that even if α_i is a random effect, OLS estimation of (4.22) is inconsistent as $y_{i,t-1}$ is correlated with the unobserved α_1 and hence with the composite error term $\alpha_i + \varepsilon_{it}$.

4.7.1 True State Dependence and Unobserved Heterogeneity

One important issue that needs to be dealt with in a dynamic model is the time-series correlation in y_{it} that will now be induced by $y_{i,t-1}$ in addition to α_i.

Let us begin by assuming $\boldsymbol{\beta} = 0$ in (4.22). Then

$$E[y_{it}|y_{i,t-1}, \alpha_i] = \gamma_i y_{i,t-1} + \alpha_i \tag{4.23}$$

and $Cor[y_{it}, y_{i,t-1}|\alpha_i] = \gamma_i$. The issue though is that α_i is really **unknown** so what is in fact observed is

$$E[y_{it}|y_{i,t-1}] = \gamma_i y_{i,t-1} + E[\alpha_i|y_{i,t-1}] \tag{4.24}$$

and $Cor[y_{it}|y_{i,t-1}] \neq \gamma$. Specifically, if we assume serial uncorrelated errors, $Cor[\varepsilon_{it}, \varepsilon_{i,t-1}] = 0$ and $\boldsymbol{\beta} = 0$ we get

$$\begin{aligned} Cor[y_{it}, y_{i,t-1}] &= Cor[\gamma, y_{i,t-1} + \alpha_i + \varepsilon_{it}, y_{i,t-1} \\ &= \gamma + Cor[\alpha_i, y_{i,t-1}] \\ &= \gamma + \frac{(1-\gamma)}{1 + (1-\gamma)\sigma^2_\varepsilon/(1+\gamma)\sigma^2_\alpha} \end{aligned} \tag{4.25}$$

which makes it clear that are two possible reasons for correlation between y_{it} and $y_{i,t-1}$:

- **true state dependence**: occurs when correlation over time is due to the **causal mechanism** that links the two periods over time ($|\gamma| > 0$. Has economic meaning;
- **unobserved heterogeneity**: arises even when there is no causal mechanism so $\gamma = 0$ but $Cor[y_{it}, y_{i,t-1}] = \sigma^2_\alpha/(\sigma^2_\alpha + \sigma^2_\varepsilon)$;

4.7.2 Inconsistency of Standard Panel Estimators

All previously presented panel estimators are **inconsistent** under a dynamic model setting:

- **OLS**: composite error term $(\alpha_i + \varepsilon)$ is correlated with $y_{i,t-1}$ via α_i, as $y_{i,t-1} = \gamma y_{i,t-2} + \boldsymbol{x}'_{i,t-1}\boldsymbol{\beta} + \alpha_i + \varepsilon_{i,t-1}$;
- **within estimator**: the regressor $(y_{i,t-1}) - \bar{y}_i)$ is correlated with $(\varepsilon_{it} - \bar{\varepsilon}_i)$. Consistency requires $\bar{\varepsilon}_i$ to be very small in relation with ε_{it};
- **random effects**: since it is a combination of the within and the between estimator it is also inconsistent;
- **first differences**: also inconsistent, but one IV approach proposed by Arellano and Bond leads to consistent estimates.

4.8 Arellano-Bond Estimator

The dynamic model presented in (4.22) leads to the following first differences equation

$$y_{it} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1}) \tag{4.26}$$

which is inconsistent because $y_{i,t-1}$ is correlated with $\varepsilon_{i,t-1}$ via its own equation. Therefore the regressors $(y_{i,t-1} - y_{i,t-2})$ and $(\varepsilon_{it} - \varepsilon_{i,t-1})$ will be correlated in (4.26). Anderson and Hsiao (1981) proposed estimating (4.22) using an **IV estimator** where $y_{i,t-2}$ would instrument $(y_{i,t-1} - y_{i,t-2})$. If we assume the error term ε_{it} to be not to be serially correlated, $y_{i,t-2}$ will be a valid instrument. Furthermore, $y_{i,t-2}$ is a good instrument as it will be correlated with $(y_{i,t-1} - y_{i,t-2})$. Anderson and Hsiao also present results suggesting that under the usual $\gamma > 0$ the IV estimator will be more efficient if $\Delta y_{i,t-2}$ is used as instrument instead. Finally, a **more efficient estimation** can be achieved if the model is overidentified, that is to say, if additional lags of the dependent variable are added to the model. In such cases estimation should be computed via 2SLS or panel GMM.

The resulting panel GMM estimator is known as the Arellano-Bond estimator and given by

$$\hat{\boldsymbol{\beta}}_{AB} = \left[\left(\sum_{i=1}^{N} \tilde{\boldsymbol{X}}'_i \boldsymbol{Z}_i \right) \boldsymbol{W}_N \left(\sum_{i=1}^{N} \boldsymbol{Z}'_i \tilde{\boldsymbol{X}}_i \right) \right]^{-1} \left(\sum_{i=1}^{N} \tilde{\boldsymbol{X}}'_i \boldsymbol{Z}_i \right) \boldsymbol{W}_N \left(\sum_{i=1}^{N} \boldsymbol{Z}'_i \tilde{\boldsymbol{y}}_i \right) \tag{4.27}$$

where $\tilde{\boldsymbol{X}}_i$ is a $(T-2)\times(K+1)$ matrix with the ith row $(\Delta y_{i,t-1}, \Delta \boldsymbol{x}'_{it})$, $t = 3, ..., T$, $\tilde{\boldsymbol{y}}_i$ is a $(T-2)\times 1$ vector with the tth row Δy_{it} and $\boldsymbol{Z}_i$ is a $(T-2)\times r$ matrix of instruments

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}'_{i3} & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \boldsymbol{z}'_{i4} & & . \\ . & & & \\ . & & & \\ . & ... & ... & \mathbf{0} \\ \mathbf{0} & ... & \mathbf{0} & \boldsymbol{z}'_{iT} \end{bmatrix}$$

Chapter 5

Maximum-Likelihood

The ML estimator holds special place among estimators. It is the most efficient estimator among consistent asymptotically normal estimators. For a fixed set of data and underlying probability model, maximum likelihood picks the values of the model parameters that make the data "more likely" than any other values of the parameters would make them. The likelihood principle, due to R. A. Fisher, is to choose as estimator of the parameter vector θ_0 that value of θ that maximizes the likelihood of observing the actual sample. In the discrete case this likelihood is the probability obtained from the probability mass function; in the continuous case this is the density.

5.1 Likelihood function

The joint probability mass function or density $f(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\theta})$ is viewed here as a function of θ given the data $(\boldsymbol{y}, \boldsymbol{X})$. This is called the **likelihood function** and is denoted by $L_N(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})$. Maximizing $L_N[\boldsymbol{\theta}]$ is equivalent to maximizing the **log-likelihood function**

$$\mathcal{L}_N(\boldsymbol{\theta}) = \ln L_N(\theta) \tag{5.1}$$

We take the natural logarithm for the sake of simplicity in computing the objective function.

5.2 Objective Function

For cross-section data the observations $(y_i, \boldsymbol{x_i})$ are independent over i with conditional density function $f(yi|\boldsymbol{x_i}, \boldsymbol{\theta})$. Then by independence the joint conditional density $f(\boldsymbol{y}|\boldsymbol{X},) =$

$\prod_{i=1}^{N} f(y_i|\boldsymbol{x_i},)$, leading to the (conditional) log-likelihood function

Notice that the only difference between the likelihood and the objective function is the N^{-1} normalization present in the objective function.

$$Q_N(\boldsymbol{\theta}) = N^{-1}\mathcal{L}_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln f(y_i|\boldsymbol{x_i}, \boldsymbol{\theta}) = \mathbf{0} \tag{5.2}$$

5.3 Maximum Likelihood Estimator

The maximum likelihood estimator (MLE) is the estimator that maximizes the (conditional) log-likelihood function and is clearly an extremum estimator[1]. Usually the MLE is the local maximum that solves the first-order conditions

$$\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{N} \frac{\partial \ln f(y_i|\boldsymbol{x_i}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{5.3}$$

The gradient vector $\partial \mathcal{L}_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ **score vector**, as it sums the first derivatives of the log density, and when evaluated at θ_0 it is called the *efficient score*.

5.4 Information Matrix Equality

The ML regularity conditions are the following

$$E_f\left(\frac{\partial \ln f(y|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = \mathbf{0} \tag{5.4}$$

$$-E_f\left(\frac{\partial^2 \ln f(y|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) = E_f\left(\frac{\partial \ln f(y|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(y|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right) \tag{5.5}$$

where (5.5) is a quite relevant practical result as it facilitates the computation of the **information matrix**. The **information matrix** is the expectation of the *outer product of the score vector*,

$$\mathcal{I} = E\left(\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right) \tag{5.6}$$

For log-likelihood function (5.2) the regularity condition (5.5) implies that

[1] extremum estimators are those that are calculated through maximization (or minimization) of a certain objective function, which depends on the data

$$- E_f \left(\frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \bigg|_{\theta_0} \right) = E_f \left(\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \bigg|_{\theta_0} \right) \tag{5.7}$$

which is known as the **information matrix (IM) equality**.

5.5 Distribution of the ML Estimator

The distribution of the ML estimator will be consistent under the following assumptions

1. the d.g.p is the conditional density $f(y_i|\boldsymbol{x_i}, \boldsymbol{\theta_0})$ used to define the likelihood function
2. the density function $f(.)$ satisfies $f(y, \boldsymbol{\theta}^{(1)}) = (y, \boldsymbol{\theta}^{(2)})$ iff $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(2)}$
3. the matrix
$$\boldsymbol{A_0} = p \lim N^{-1} \frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \bigg|_{\theta_0} \tag{5.8}$$
exists and is finite nonsingular (think of it as analogous to the $p \lim N^{-1} \boldsymbol{X'X}$ in OLS)
4. the order of differentiation and integration of the log-likelihood can be reversed

Then the ML estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$, defined to be a solution of the F.O.C.

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}} - \boldsymbol{\theta_0}) \rightarrow^d \mathcal{N}[\mathbf{0}, -\boldsymbol{A}_\mathbf{0}^{-\mathbf{1}}] \tag{5.9}$$

The resulting *asymptotic distribution* of the MLE is often expressed as

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}} \sim^a \mathcal{N} \left\{ \boldsymbol{\theta}, \left(-E \left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \right)^{-1} \right\} \tag{5.10}$$

which is again quite similar to the OLS asymptotic distribution. The $p \lim$ operator used for defining $\boldsymbol{A_0}$ in equation (5.8) is replaced by $\lim E$ and then drop the limit. The right-hand side of (5.10) is the Cramer-Rao lower bound (CRLB), which from basic statistics courses is the lower bound of the variance of unbiased estimators in small samples. The MLE has the strong attraction of having the smallest asymptotic variance among root-N consistent estimators. This result requires the strong assumption of correct specification of the conditional density.

5.5.1 Poisson regression Example

We shall begin with the example of what is perhaps the simplest m-estimator, the Poisson regression model. Recall that the Poisson distribution is a discrete distribution characterized by an expected value λ that equals its variance. The $f(y|\lambda)$ density function equals

$$f(y|\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} \tag{5.11}$$

The log-likelihood density will be given by

$$\ln f(y_i) = -\lambda + y_i \ln \lambda - \ln y_i! \tag{5.12}$$

and the $Q_N(\lambda)$ objective function

$$\mathcal{L}_N(\lambda) = \sum_{i=1}^{N} [-\lambda + y_i \ln \lambda - \ln y_i!] \tag{5.13}$$

the first order conditions are taken in relation to $\boldsymbol{\beta}$

$$\frac{\partial \mathcal{L}_N(\lambda)}{\partial \lambda} \equiv \text{Score} = 0 \tag{5.14}$$

the Score matrix being the partial derivative in terms of λ. In this case we will have

$$\begin{aligned} \frac{\partial \mathcal{L}_N(\lambda)}{\partial \lambda} &= \sum_{i=1}^{N} \left[-1 + \frac{y_i}{\lambda} - 0\right] = 0 \\ N &= \sum_{i=1}^{N} \frac{y_i}{\lambda} \\ \lambda &= \sum_{i=1}^{N} \frac{y_i}{N} \end{aligned} \tag{5.15}$$

5.6 Quasi-Maximum Likelihood

The **quasi-MLE $\hat{\boldsymbol{\theta}}_{\boldsymbol{QML}}$** is defined to be the estimator that maximizes a log-likelihood function that is *misspecified*, as the result of specification of the wrong density. Generally such misspecification leads to inconsistent estimation. In terms of intuition, the QML concept is closely related to the *pseudo-tru value* misspecification presented in section 3.5.

5.6.1 Pseudo-True Value

The quasi-MLE $\hat{\boldsymbol{\theta}}_{\boldsymbol{QML}}$ converges in probability to the pseudo true value $\boldsymbol{\theta}^*$ defined as

$$\boldsymbol{\theta}^* = \arg\max_{\theta\in\Theta}[p\lim N^{-1}\mathcal{L}_N(\boldsymbol{\theta})] \tag{5.16}$$

The probability limit is taken with respect to the true d.g.p. If a misspecification of the density occurs then the true d.g.p differs from the assumed density $f(y|\boldsymbol{x},\boldsymbol{\theta})$ used to form $\mathcal{L}_N(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^* \neq \boldsymbol{\theta_0}$ will imply that the quasi-MLE is inconsistent.

5.6.2 Linear Exponential Family

Exponential family is a class of probability distributions sharing the form

$$f(y|\mu) = \exp\alpha(\mu) + b(y) + c(\mu)y \tag{5.17}$$

It is said that such distributions belong to the linear exponential family (LEF) of density functions. Its big advantage is to provide consistency even when the density is partially misspecified. Asides from that, LEF form is chosen for mathematical convenience, on account of some useful algebraic properties, as well as for generality, as exponential families are in a sense very natural distributions to consider. The mean parametrization for the LEF is such that $\mu = E[y]$.

Distribution	$f(y) = \exp\alpha(\mu) + b(y) + c(\mu)y$	$E[y]$	$V[y] = [c'(\mu)]^{-1}$
Normal (σ^2 is known)	$\exp\frac{-\mu^2}{e\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{\mu}{\sigma^2}y$	μ	σ^2
Bernoulli	$\exp\ln(1-p) + \ln[p/(1-p)]y$	$\mu = p$	$\mu(1-\mu)$
Exponential	$\exp\ln\lambda - \lambda y$	$\mu = \lambda^{-1}$	μ^2
Poisson	$\exp -\lambda - \ln y! + y\ln\lambda$	$\mu = \lambda$	μ

Table 5.1: LEF densities

As can be seen on 5.1, LEFs are very special cases. In general, misspecification of any aspect of the density leads to inconsistency of the MLE. Even in the LEF case the quasi-MLE can be used only to predict the conditional mean whereas with a correctly specified density one can predict the conditional distribution.

5.7 Numerical Maximization

The integrals encountered in a basic calculus course are deliberately chosen for simplicity; those found in most m-estimators however are not always so accommodating. In such cases, where no closed-form antiderivatives can be found one needs to resort to numerical methods in order to proceed with the estimation of the desired coefficients. We present here the four algorithms implemented in Stata via the `, technique(''algorithm'')` option in m-estimators: Newton-Raphson (NR); Berndt-Hall-Hall-Hausman (BHHH); Davidon-Fletcher-Powell (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS).

5.7.1 Newton-Raphson

The NR algorithm is the standard numerical maximization algorithm. It simply maximizes the second order of a Taylor approximation to the log-likelihood objective function $Q_N(\boldsymbol{\beta}_{i+1})$

$$Q_N(\beta_{i+1}) = Q_N(\boldsymbol{\beta}_i) + (\beta_{i+1} - \beta_i)' \frac{\partial Q_N(\beta)}{\partial \beta} + \frac{1}{2}(\beta_{i+1} - \beta_i)' \frac{\partial^2 Q_N(\beta)}{\partial \beta \partial \beta'} (\beta_{i+1} - \beta_i) \tag{5.18}$$

The method then finds the value of β_{i+1} that maximizes this approximation to $Q_N(\beta_{i+1})$

$$\begin{aligned}
\frac{\partial Q_N(\beta_{i+1})}{\partial \beta_{i+1}} &= \frac{\partial Q_N(\beta)}{\partial \beta} + \frac{\partial^2 Q_N(\beta)}{\partial \beta \partial \beta'} (\beta_{i+1} - \beta_i) = 0 \\
\frac{\partial^2 Q_N(\beta)}{\partial \beta \partial \beta'} (\beta_{i+1} - \beta_i) &= -\frac{\partial Q_N(\beta)}{\partial \beta} \\
(\beta_{i+1} - \beta_i) &= -\left(\frac{\partial^2 Q_N(\beta)}{\partial \beta \partial \beta'} \right)^{-1} \frac{\partial Q_N(\beta)}{\partial \beta} \\
\beta_{i+1} &= \beta_i - \left(\frac{\partial^2 Q_N(\beta)}{\partial \beta \partial \beta'} \right)^{-1} \frac{\partial Q_N(\beta)}{\partial \beta}
\end{aligned} \tag{5.19}$$

In order to address the case where NR algorithm steps past the maximum log-likelihood value we can introduce a scalar λ which accounts for the step size. This way we modify (5.19) by adding λ

$$\beta_{i+1} = \beta_i + \lambda \left(-\frac{\partial^2 Q_N(\beta)}{\partial \beta \partial \beta'} \right)^{-1} \frac{\partial Q_N(\beta)}{\partial \beta} \tag{5.20}$$

The NR procedure has two drawbacks. First, calculation of the Hessian is usually computation-intensive. Procedures that avoid calculating the Hessian at every iteration can be much faster. Second, the NR procedure does not guarantee an increase in each step if the log-likelihood function is not globally concave. When $-H^{-1} = -\left(\frac{\partial^2 Q_N(\beta)}{\partial \beta \partial \beta'} \right)^{-1}$ is not positive definite, an increase is not guaranteed.

5.7.2 Berndt-Hall-Hall-Hausman

BHHH (and commonly pronounced B-triple H), proposed using this relationship in the numerical search for the maximum of the log-likelihood function. In particular, the BHHH procedure uses B_i in the optimization routine in place of $-H^{-1}$. Each iteration is defined by

$$\beta_{i+1} = \beta_i + \lambda B_i^{-1} \frac{\partial Q_N(\beta)}{\partial \beta} \tag{5.21}$$

which is the same as in NR except $-H^{-1}$ is substituted by B^{-1}, which equals

$$B_i = \sum_{i=1}^{n} N^{-1} s_N(\beta_i) s_N(\beta_i)' \tag{5.22}$$

and $s_N(\beta_i)s_N(\beta_i)'$ equal to

$$s_N(\beta_i)s_N(\beta_i)' = \begin{bmatrix} s_n^1 s_n^1 & s_n^1 s_n^2 & ... & s_n^1 s_n^k \\ s_n^1 s_n^2 & s_n^2 s_n^2 & ... & s_n^2 s_n^k \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ s_n^1 s_n^k & s_n^2 s_n^k & ... & s_n^k s_n^k \end{bmatrix}$$

In other words the substitution of $-H^{-1}$ by B^{-1} means that instead of the Hessian we are now maximizing the outer product of the scores.

5.7.3 Davidon-Fletcher-Powell and Broyden-Fletcher-Goldfarb-Shanno

The DFP and BFGS methods calculate the approximate Hessian in a way that uses information at more than one point on the likelihood function. If the function is quadratic, then information at one point on the function provides all the information that is needed about the shape of the function. These methods work well, therefore, when the log-likelihood function is close to quadratic. In contrast, the DFP and BFGS procedures use information at several points to obtain a sense of the curvature of the log-likelihood function. The DFP and BFGS procedures use these concepts to approximate the Hessian. The Hessian is the matrix of second derivatives. As such, it gives the amount by which the slope of the curve changes as one moves along the curve. The Hessian is defined for infinitesimally small movements. Since we are interested in making large steps, understanding how the slope changes for non-infinitesimal movements is useful. An *arc Hessian* can be defined on the basis of how the gradient changes from one point to another.

Chapter 6

Binary Response Models

Discrete outcome or qualitative response models are models for a dependent variable that indicates in which one of m mutually exclusive categories the outcome of interest falls. Binary outcomes are simple to model and estimation is usually by maximum likelihood because the distribution of the data is necessarily defined by the Bernoulli model. If the probability of one outcome equals p, then the probability of the other outcome must be $(1p)$. The two standard binary outcome models, the **logit** and the **probit** models, specify different functional forms for this probability as a function of regressors and can be seen on figure 11. The difference between these estimators is qualitatively similar to use of different functional forms for the conditional mean in least-squares regression.

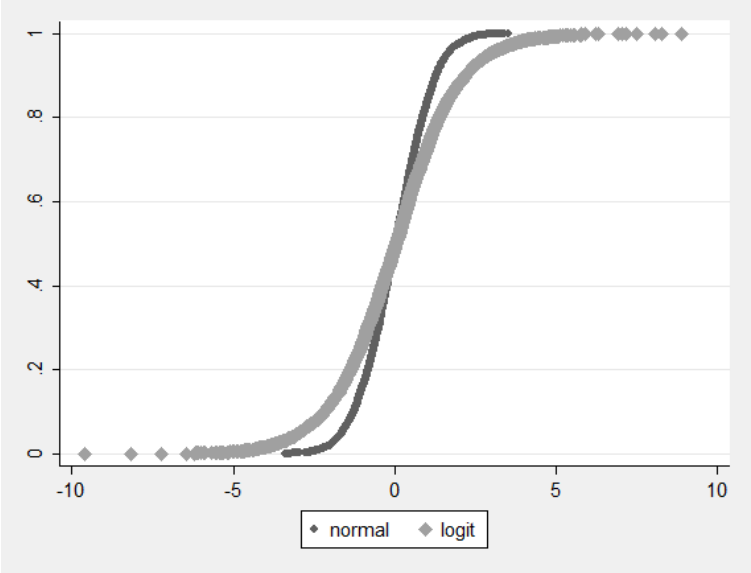

Figure 6.1: The Logistic and Normal c.d.f.

6.1 General Binary Model

We begin by presenting a generalized version of the estimation process for binary models that is valid for both the logit and the probit models. We then proceed with the derivations of those models which are of particular interest for empirical researchers.

6.1.1 General Binary Outcome Model

For binary outcome data the dependent variable y takes one of two values. We let

$$y = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}$$

A regression model is formed by parameterizing the probability p to depend on a regressor vector x and a $K \times 1$ parameter vector $\boldsymbol{\beta}$. The commonly used models are of single-index form with conditional probability given by

$$p_i \equiv \Pr[y_i = 1|\boldsymbol{x}] = F(\boldsymbol{x_i'\beta}) \tag{6.1}$$

where $F(.)$ is a specified c.d.f. function. Usually we will work with $p_i = F(\boldsymbol{x_i'\beta})$.

6.1.2 ML Estimation

The Bernoulli distributed independent sample $(y_i, \boldsymbol{x_i}), i = 1, ..., N$ is characterized by the following p.d.f.

$$f(y_i|\boldsymbol{x_i}) = p_i^{y_i}(1-p_i)^{1-y_i}, \qquad y = 0, 1 \tag{6.2}$$

where $p_i = F(\boldsymbol{x_i'\beta})$. Notice that this the yields probabilities p_i and $(1-p_i)$ since $f(1) = p^1(1-p)^0$ and $f(0) = p^0(1-p)^1$. The log-density is given by

$$\begin{aligned} \ln f(y_i|\boldsymbol{x_i}) &= y_i \ln p_i + (1-y_i)\ln(1-p_i) \\ &= y_i \ln F(\boldsymbol{x_i'\beta}) + (1-y_i)\ln F(\boldsymbol{x_i'\beta}) \end{aligned} \tag{6.3}$$

and consequently the log-likelihood by

$$\mathcal{L}_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i \ln F(\boldsymbol{x_i'\beta}) + (1-y_i)\ln(1-F(\boldsymbol{x_i'\beta})) \tag{6.4}$$

Now, according to the used model, different c.d.f.'s will fill in the $F(\boldsymbol{x_i'\beta})$ places. Next comes the first order conditions

$$\begin{aligned}
\frac{\partial \mathcal{L}_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= 0 \\
\sum_{i=1}^{N} \frac{y_i}{F(\boldsymbol{x_i'\beta})} F'(\boldsymbol{x_i'\beta}) - \frac{1-y_i}{(1-F(\boldsymbol{x_i'\beta}))} F'(\boldsymbol{x_i'\beta}) &= 0 \\
\sum_{i=1}^{N} \frac{[1-F(\boldsymbol{x_i'\beta})]y_i + (1-y_i)F(\boldsymbol{x_i'\beta})}{F(\boldsymbol{x_i'\beta})[1-F(\boldsymbol{x_i'\beta})]} F'(\boldsymbol{x_i'\beta}) &= 0 \\
\sum_{i=1}^{N} \frac{y_i - y_i F(\boldsymbol{x_i'\beta}) + F(\boldsymbol{x_i'\beta}) - y_i F(\boldsymbol{x_i'\beta})}{F(\boldsymbol{x_i'\beta})[1-F(\boldsymbol{x_i'\beta})]} F'(\boldsymbol{x_i'\beta}) &= 0 \\
\sum_{i=1}^{N} \frac{y_i - F(\boldsymbol{x_i'\beta})}{F(\boldsymbol{x_i'\beta})[1-F(\boldsymbol{x_i'\beta})]} F'(\boldsymbol{x_i'\beta}) &= 0 \qquad (6.5)
\end{aligned}$$

There is no explicit solution for $\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$ but one can apply numerical methods as the Newton-Raphson algorithm (see section 5.19).

6.1.3 Consistency of the MLE

The MLE is consistent if the conditional density of y given x is correctly specified. Since the density here must be the Bernoulli, the only possible misspecification is that the Bernoulli probability is misspecified. *So the MLE is consistent if* $p_i \equiv F(\boldsymbol{x_i\beta})$ *and is inconsistent otherwise*

$$E[y] = 1 \times p + 0 \times (1-p) = p \qquad (6.6)$$

which implies

$$E[y_i|\boldsymbol{x_i}] = F(\boldsymbol{x_i'\beta}) \qquad (6.7)$$

Given correct density specification we get the ML asymptotic distribution

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{ML}} \sim^a \mathcal{N}\left\{\boldsymbol{\beta}, \left(-E\left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]\right)^{-1}\right\} \qquad (6.8)$$

6.2 Logit

We begin with c.d.f. of the logistic distribution

$$p = \Lambda(\boldsymbol{x'\beta}) = \frac{\exp(\boldsymbol{x'\beta})}{1+\exp(\boldsymbol{x'}\beta)} \qquad (6.9)$$

which specifies the following model density:

$$f(y_i|\boldsymbol{x_i'\beta}) = \Lambda(\boldsymbol{x_i'\beta})^{y_i}[1-\Lambda(\boldsymbol{x_i'\beta})]^{1-y_i} \tag{6.10}$$

and hence the log-density

$$\ln f(y_i|\boldsymbol{x_i'\beta}) = y_i \ln \Lambda(\boldsymbol{x_i'\beta}) + (1-y_i)[1-\Lambda(\boldsymbol{x_i'\beta})] \tag{6.11}$$

Then the objective function will be

$$\begin{aligned}\mathcal{L}_N(\boldsymbol{\beta}) &= \sum_{i=1}^{N} y_i \ln \Lambda(\boldsymbol{x_i'\beta}) + (1-y_i)[1-\Lambda(\boldsymbol{x_i'\beta})] \\ &= \sum_{i=1}^{N} y_i \ln \Lambda(\boldsymbol{x_i'\beta}) + \sum_{i=1}^{N}[1-\Lambda(\boldsymbol{x_i'\beta})] - \sum_{i=1}^{N} y_i[1-\Lambda(\boldsymbol{x_i'\beta})] \end{aligned} \tag{6.12}$$

The Score vector will be given by the F.O.C.[1]

$$\begin{aligned}\frac{\partial \mathcal{L}_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \mathcal{L}_N(\boldsymbol{\beta})}{\partial \Lambda(\boldsymbol{x_i\beta})} \times \frac{\partial \Lambda(\boldsymbol{x_i\beta})}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{N} y_i \frac{\Lambda'(\boldsymbol{x_i'\beta})}{\Lambda(\boldsymbol{x_i'\beta})} + \sum_{i=1}^{N} \frac{[1-\Lambda(\boldsymbol{x_i'\beta})]'}{1-\Lambda(\boldsymbol{x_i'\beta})} - \sum_{i=1}^{N} y_i \frac{[1-\Lambda(\boldsymbol{x_i'\beta})]'}{1-\Lambda(\boldsymbol{x_i'\beta})} \\ &= \sum_{i=1}^{N} y_i \frac{\Lambda(\boldsymbol{x_i'\beta})[1-\Lambda(\boldsymbol{x_i'\beta})]}{\Lambda(\boldsymbol{x_i'\beta})}\boldsymbol{x_i} - \sum_{i=1}^{N} \frac{\Lambda(\boldsymbol{x_i'\beta})[1-\Lambda(\boldsymbol{x_i'\beta})]}{1-\Lambda(\boldsymbol{x_i'\beta})}\boldsymbol{x_i} + \\ \sum_{i=1}^{N} y_i \frac{\Lambda(\boldsymbol{x_i'\beta})[1-\Lambda(\boldsymbol{x_i'\beta})]}{1-\Lambda(\boldsymbol{x_i'\beta})}\boldsymbol{x_i} & \\ &= \sum_{i=1}^{N} y_i[1-\Lambda(\boldsymbol{x_i'\beta})]\boldsymbol{x_i} - \sum_{i=1}^{N} \Lambda(\boldsymbol{x_i'\beta})\boldsymbol{x_i} + \sum_{i=1}^{N} y_i \Lambda(\boldsymbol{x_i'\beta})\boldsymbol{x_i} \\ &= \sum_{i=1}^{N} y_i - \Lambda(\boldsymbol{x_i'\beta})\boldsymbol{x_i} \end{aligned} \tag{6.13}$$

So we have the score vector. The $\hat{\boldsymbol{\beta}}$ estimate however will need to resort to numerical procedures in order to be found. The hessian matrix given by the S.O.C. will be

[1] notice that the logistic c.d.f. has the convenient property that $\Lambda'(.) = \Lambda(.)[1-\Lambda(.)]$ and $\Lambda''(.) = [1-2\Lambda(.)]\Lambda(.)[1-\Lambda()]$.

$$\begin{aligned}\frac{\partial^2 \mathcal{L}_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= \frac{\partial \text{Score}}{\partial \Lambda(\boldsymbol{x_i}\boldsymbol{\beta})} \times \frac{\Lambda(\boldsymbol{x_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= -N^{-1} \sum_{i=1}^{N} \Lambda'(\boldsymbol{x_i'}\boldsymbol{\beta}) \boldsymbol{x_i} \boldsymbol{x_i'} \\ &= -N^{-1} \sum_{i=1}^{N} \Lambda(\boldsymbol{x_i'}\boldsymbol{\beta})[1 - \Lambda(\boldsymbol{x_i'}\boldsymbol{\beta})] \boldsymbol{x_i} \boldsymbol{x_i'}\end{aligned} \tag{6.14}$$

6.3 Probit

Again, we begin with the well known c.d.f. of the normal distribution

$$p = \Phi(\boldsymbol{x_i'}\boldsymbol{\beta}) = \int_{-1}^{\boldsymbol{x_i'}\boldsymbol{\beta}} \phi(z) dz \tag{6.15}$$

where the p.d.f. $\phi(z)$ is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) \tag{6.16}$$

the conditional density of the model is

$$f(y_i | \boldsymbol{x_i'}\boldsymbol{\beta} = \Phi(\boldsymbol{x_i'}\boldsymbol{\beta})^{y_i} [1 - \Phi(\boldsymbol{x_i'}\boldsymbol{\beta})]^{1-y_i} \tag{6.17}$$

which yields the log-density

$$\begin{aligned}\ln f(y_i | \boldsymbol{x_i'}\boldsymbol{\beta}) &= y_i \ln \Phi(\boldsymbol{x_i'}\boldsymbol{\beta}) + (1 - y_i) \ln[1 - \Phi(\boldsymbol{x_i'}\boldsymbol{\beta})] \\ &= y_i \ln \Phi(\boldsymbol{x_i'}\boldsymbol{\beta}) + \ln[1 - \Phi(\boldsymbol{x_i'}\boldsymbol{\beta})] - y_i \ln[1 - \Phi(\boldsymbol{x_i'}\boldsymbol{\beta})]\end{aligned} \tag{6.18}$$

The objective function $Q_N(\boldsymbol{\beta})$ will be

$$\mathcal{L}_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i \ln \Phi(\boldsymbol{x_i'}\boldsymbol{\beta}) + \ln[1 - \Phi(\boldsymbol{x_i'}\boldsymbol{\beta})] - y_i \ln[1 - \Phi(\boldsymbol{x_i'}\boldsymbol{\beta})] \tag{6.19}$$

and the F.O.C.

$$\begin{aligned}
\frac{\partial \mathcal{L}_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \mathcal{L}_N(\boldsymbol{\beta})}{\partial \Phi(\boldsymbol{x_i'\beta})} \times \frac{\partial \Phi(\boldsymbol{x_i'\beta})}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{N} y_i \frac{\phi(\boldsymbol{x_i'\beta})}{\Phi(\boldsymbol{x_i'\beta})}\boldsymbol{x_i} - \frac{\phi(\boldsymbol{x_i'\beta})}{[1-\Phi(\boldsymbol{x_i'\beta})]}\boldsymbol{x_i} + y_i \frac{\phi(\boldsymbol{x_i'\beta})}{[1-\Phi(\boldsymbol{x_i'\beta})]}\boldsymbol{x_i} \\
&= \sum_{i=1}^{N} y_i \frac{\phi(\boldsymbol{x_i'\beta})}{\Phi(\boldsymbol{x_i'\beta})}\boldsymbol{x_i} + \frac{\phi(\boldsymbol{x_i'\beta})}{[1-\Phi(\boldsymbol{x_i'\beta})]}\boldsymbol{x_i}(1-y_i) \\
&= \sum_{i=1}^{N} \frac{[1-\Phi(\boldsymbol{x_i'\beta})][y_i - \phi(\boldsymbol{x_i'\beta})\boldsymbol{x_i}] - \Phi(\boldsymbol{x_i'\beta})[\phi(\boldsymbol{x_i'\beta})\boldsymbol{x_i} - y_i\phi(\boldsymbol{x_i'\beta})\boldsymbol{x_i}]}{\Phi(\boldsymbol{x_i'\beta})[1-\Phi(\boldsymbol{x_i'\beta})]} \\
&= \sum_{i=1}^{N} \frac{y_i\phi(\boldsymbol{x_i'\beta})\boldsymbol{x_i} - \Phi(\boldsymbol{x_i'\beta})y_i\phi(\boldsymbol{x_i'\beta})\boldsymbol{x_i} - \Phi(\boldsymbol{x_i'\beta})\phi(\boldsymbol{x_i'\beta})\boldsymbol{x_i} + \Phi(\boldsymbol{x_i'\beta})y_i\phi(\boldsymbol{x_i'\beta})\boldsymbol{x_i}}{\Phi(\boldsymbol{x_i'\beta})[1-\Phi(\boldsymbol{x_i'\beta})]} \\
&= \sum_{i=1}^{N} \frac{\phi(\boldsymbol{x_i'\beta})\boldsymbol{x_i'}[y_i - \Phi(\boldsymbol{x_i'\beta})]}{\Phi(\boldsymbol{x_i'\beta})[1-\Phi(\boldsymbol{x_i'\beta})]} \qquad (6.20)
\end{aligned}$$

yeah,so I guess Econometrics *does* beat Physics.

Chapter 7

Duration Analysis

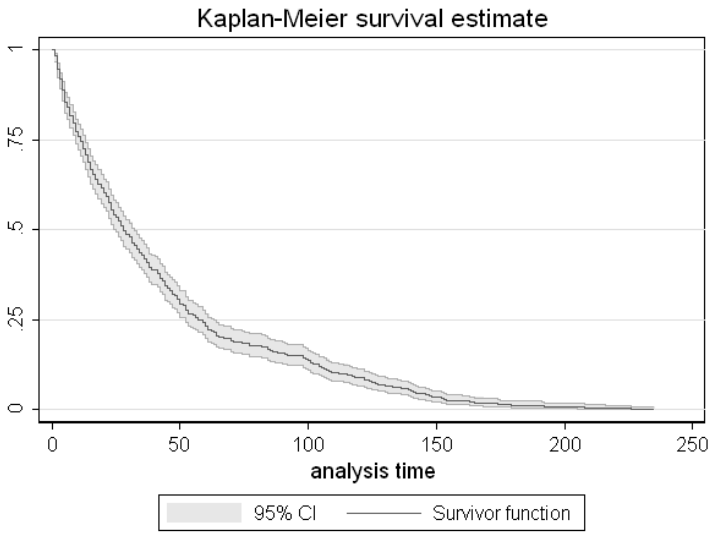

Figure 7.1: Kaplan-Meier estimate of US strike duration

Duration analysis models the length of time individuals remain in a given state before transition to another state occurs. Originally developed by biostatisticians and epidemiologists the field has several applications in economic sciences, particularly in labour and industrial economics. Typical examples of applications include strike (as in figure7) and unemployment duration, industry evolution or self-employment spell maturity.

Duration analysis models need to deal with a relatively large number of issues. First, both duration as well as the probability of transition need to be modelled via a wide set of distributional functions. Second, there are two sampling method impacts analysis: **stock**

sampling refers to sampling in the survey period from the stock of individuals who are then in a given state. **Flow sampling** means that we sample those who enter the state during a particular interval. Third, the data on the duration of a spell is often censored. This is a major reason for modelling transitions rather than mean duration. Forth, several states and destinations of transitions can occur.

7.1 Key Concepts

Duration is a state in a nonnegative random variable, denoted T, which in economic data is often a discrete random variable. The **cumulative distribution function** (c.d.f.) of T is denoted $F(t)$ and the **density function** is $f(t) = dF(t)/dt$. Then the probability that the duration or spell length is less than t is

$$\begin{aligned} F(t) &= Pr[T \leq t] \\ &= \int_0^t f(s)ds. \end{aligned} \tag{7.1}$$

A complementary concept to the c.d.f. is the probability that the duration equals or exceeds t, called the **survival function**, which is defined by

$$\begin{aligned} S(t) &= PrT > t \\ &= 1 - F(t) \end{aligned} \tag{7.2}$$

The survivor function is monotonically declining (oh really?) from one to zero since the c.d.f. is monotonically increasing from zero. If all individuals at risk eventually transit from one state to another $S(\infty) = 0$. Otherwise $S(\infty) > 0$ and the duration distribution is called **defective**. The sample mean of a completed spell length is the integral $\int_0^\infty uf(u)du$. In other words, *the mean duration equals area under the survival curve*

$$E[T] = \int_0^\infty (1 - F(u))du = \int_0^\infty S(u)du. \tag{7.3}$$

Another key concept is the **hazard function**, which is the *instantaneous probability of leaving a state conditional on survival to time t.* This is defined as

$$\begin{aligned} \lambda(t) &= \lim_{\Delta \to 0} \frac{Pr[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{7.4}$$

It is easily verifiable that the **hazard function** equals the change in the log-survivor function,

$$\lambda(t) = \frac{d\ln(S(t))}{dt}. \tag{7.5}$$

Finally the **cumulative hazard function**

$$\begin{aligned} \Lambda(t) &= \int_0^t \lambda(s)ds \\ &= -\ln S(t) \end{aligned} \tag{7.6}$$

Function	**Symbol**	**Definition**	**Relationships**
Density	f(t)		f(t)=dF(t)/dt
Distribution	F(t)	Pr[T≤t]	F(t)=$\int_0^\infty$f(s)ds
Survivor	S(t)	Pr[T≥t]	S(t)=1-F(t)
Hazard	λ(t)	$\lim_{h\to 0} \frac{Pr[t\leq T<t+h\|T\geq t]}{h}$	λ(t)=f(t)/S(t)
Cumulative hazard	Λ(t)	$\int_0^\infty \lambda$(s)ds	Λ(t)= − ln S(t)

Table 7.1: Duration Analysis: Definitions and Key concepts

7.2 Censoring

Survival data are usually censored, as some spells are incompletelly observed. Data may be right-, left- or interval censored. **Right-censored** data spells are observed from time 0 up to a censoring time c.

Chapter 8

Generalized Method of Moments

The most important assumption made for the OLS is the orthogonality between the error term and regressors. Without it, the OLS estimator is not even consistent. Since in many important applications the orthogonality condition is not satisfied, it is imperative to be able to deal with endogenous regressors. The estimation method called the **Generalized Method of Moments** (GMM), which includes OLS as a special case, provides a solution.

8.1 MM Estimator

We begin by defining the orthogonality condition on the assumption of the existence of r moment conditions for q parameters

$$E[\boldsymbol{h}(\boldsymbol{w_i}, \boldsymbol{\theta_0})] = \mathbf{0} \tag{8.1}$$

where $\boldsymbol{\theta}$ is a $q \times 1$ vector, $\boldsymbol{h}(.)$ an $r \times 1$ vector with $r \geq q$ and $\boldsymbol{\theta_0}$ denotes the value of $\boldsymbol{\theta}$ in the d.g.p. Notice that the vector $\boldsymbol{W}$ is an aggregate vector of one or more dependent variables $\boldsymbol{y}$, the independent variables $\boldsymbol{X}$ including some potentially endogenous covariates as well as a vector of their instruments $\boldsymbol{Z}$. The choice of $\boldsymbol{h(.)}$ is analogous to the choice of model specification in say, m-estimation methods. Due to the immense versatility of the GMM model, any econometric specification can fit to $\boldsymbol{h(.)}$, from OLS to maximum likelihood.

From the population condition stated in (8.1) we derive its sample counterpart

$$N^{-1} \sum_{i=1}^{N} \boldsymbol{h}(\boldsymbol{w_i}, \hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{8.2}$$

If we have a *just identified* model with $r = q$ then (8.2) will produce a $\hat{\theta}$ estimate by minimizing the following objective function

Which is equivalently to minimize

$$\arg\min Q_N(\boldsymbol{\theta}) = \left(N^{-1} \sum_{i=1}^{N} (\boldsymbol{w_i}, \boldsymbol{\theta}) \right)' \left(N^{-1} \sum_{i=1}^{N} (\boldsymbol{w_i}, \boldsymbol{\theta}) \right) \tag{8.3}$$

If however $r > q$ then we have an *overidentified* model and (8.2) has no solution for $\hat{\theta}$ as there are more r equations then q unknowns. Therefore we need to introduce a weighting matrix $\boldsymbol{W}_N$ in the (8.3) objective function

$$\arg\min Q_N(\boldsymbol{\theta}) = \left(N^{-1} \sum_{i=1}^{N} (\boldsymbol{w_i}, \boldsymbol{\theta}) \right)' \boldsymbol{W}_N \left(N^{-1} \sum_{i=1}^{N} (\boldsymbol{w_i}, \boldsymbol{\theta}) \right) \tag{8.4}$$

where $\boldsymbol{W}$ is an $r \times r$ symmetric positive definite, possibly stochastic with finite probability limit weighting matrix. Differentiating $Q_N(\boldsymbol{\theta})$ in respect to (8.3) we get the GMM F.O.C.

$$\begin{aligned} \left. \frac{\partial Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\hat{\theta}_0} &= \mathbf{0} \\ \left(N^{-1} \sum_{i=1}^{N} \left. \frac{\partial (\boldsymbol{w_i}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right|_{\hat{\theta}} \right) \boldsymbol{W}_N \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{h_i}(\hat{\boldsymbol{\theta}}) \right) &= \mathbf{0} \end{aligned} \tag{8.5}$$

which will typically be solved with the numerical solutions presented in section 5.7.

8.1.1 Distribution of the GMM estimator

The following propositions are required for the deriving the GMM estimator distribution

1. the d.g.p. imposes the moment condition 8.1, that is, $E[\boldsymbol{h}(\boldsymbol{w_i}, \boldsymbol{\theta_0})] = \mathbf{0}$
2. the $r \times 1$ vector function $\boldsymbol{h}(.)$ satisfies $\boldsymbol{h}(\boldsymbol{w}, \boldsymbol{\theta^{(1)}}) = \boldsymbol{h}(\boldsymbol{w}, \boldsymbol{\theta^{(2)}})$ iff $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(1)}$
3. the following $r \times q$ matrix exists and is finite with rank q

$$\boldsymbol{G_0} = p \lim N^{-1} \sum_{i=1}^{N} \left(\left. \frac{\partial \boldsymbol{h}_i}{\partial \boldsymbol{\theta}'} \right|_{\theta_0} \right) \tag{8.6}$$

4. $\boldsymbol{W}_N \to^p \boldsymbol{W}_0$ where $\boldsymbol{W}_0$ is finite symmetric positive definite
5. $N^{-1/} \sum_{i=1}^{n} \boldsymbol{h}_i|_{\theta_0} \to^d \mathcal{N}[\mathbf{0}, \boldsymbol{S}(\boldsymbol{\theta_0})]$ where

$$\boldsymbol{S}_0 = p \lim N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} [\boldsymbol{h}_i \boldsymbol{h}_j'|_{\theta_0}] \tag{8.7}$$

Then the **GMM estimator** $\hat{\boldsymbol{\theta}}_{GMM}$ defined to be the root of the f.o.c. $\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is consistent for $\boldsymbol{\theta}_0$ and

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) \to^d \mathcal{N}[\mathbf{0}, (\boldsymbol{G}_\mathbf{0}'\boldsymbol{W}_\mathbf{0}\boldsymbol{G}_\mathbf{0})^{-\mathbf{1}}(\boldsymbol{G}_\mathbf{0}'\boldsymbol{W}_\mathbf{0}\boldsymbol{S}_\mathbf{0}\boldsymbol{W}_\mathbf{0}\boldsymbol{G}_\mathbf{0})(\boldsymbol{G}_\mathbf{0}'\boldsymbol{W}_\mathbf{0}\boldsymbol{G}_\mathbf{0})^{-\mathbf{1}}] \tag{8.8}$$

8.2 Optimal GMM

The optimal GMM will be given by the $\boldsymbol{W}_N$ that achieves the smallest asymptotic variance given the chosen $\boldsymbol{h(.)}$ model specification. For just-identified models the MM estimator is obtained for *any* full rank weighting matrix, so $\boldsymbol{W}_N$ is usually set as $\boldsymbol{W}_N = \boldsymbol{I}_q$. This is visible in equation (8.3). For overidentified models $r > q$ and $\boldsymbol{S}_0$ known, the most efficient GMM estimator is obtained by setting $\boldsymbol{W}_N = \boldsymbol{S}_0^{-1}$. The $Q_N(\boldsymbol{\theta})$ objective function will then be given by

$$\arg\min Q_N(\boldsymbol{\theta}) = \left(N^{-1}\sum_{i=1}^{N}(\boldsymbol{w_i}, \boldsymbol{\theta})\right)' \hat{\boldsymbol{S}}^{-1}\left(N^{-1}\sum_{i=1}^{N}(\boldsymbol{w_i}, \boldsymbol{\theta})\right) \tag{8.9}$$

where $\boldsymbol{G_0}$ is defined in equation (8.6) and $\hat{\boldsymbol{S}}^{-1}$ is the inverted (assuming $\hat{\boldsymbol{S}}$ invertible) matrix

$$\hat{\boldsymbol{S}}^{-1} = \sum_{i=1}^{N}\boldsymbol{h_i}(\hat{\boldsymbol{\theta}})\boldsymbol{h_i}(\hat{\boldsymbol{\theta}})' \tag{8.10}$$

The optimal GMM estimator will then follow the limit distribution

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) \to^d \mathcal{N}[\mathbf{0}, (\boldsymbol{G}_{\boldsymbol{O}}'\boldsymbol{S}_{\boldsymbol{O}}^{-\mathbf{1}}\boldsymbol{G}_\mathbf{0})^{-\mathbf{1}}] \tag{8.11}$$

which, if you notice is actually remarkably similar to the ML limit distribution with a weighted outer product of the F.O.C.'s.

8.2.1 Number of Moment Restrictions

In general adding further moment restrictions **improves asymptotic efficiency**, as it reduces the limit variance $(\boldsymbol{G}_\mathbf{0}'\boldsymbol{S}_\mathbf{0}^{-\mathbf{1}}\boldsymbol{G}_\mathbf{0})^{-1}$ of the optimal GMM estimator or at worst leaves it unchanged. The benefits of adding further moment conditions vary with the application. For example, if the estimator is the MLE then there is no gain since the MLE is already fully efficient. The literature has focused on IV estimation where gains may be considerable because the variable being instrumented may be much more highly correlated with a combination of many instruments than with a single instrument. There is a limit, however, as the

number of moment restrictions cannot exceed the number of observations. Moreover, adding more moment conditions increases the likelihood of finite-sample bias and related problems similar to those of weak instruments in linear models.

Chapter 9

Instrumental Variables

The instrumental variables (IV) estimator provides a way to address endogeneity issues ($corr(X, u) \neq 0$). An obvious alternative to solve this issue would be through a randomized experiment, but for most economics applications such experiments are too expensive or even infeasible. The IV approach consists in using an **instrument** z which will be able to explain changes in x without the covariance effect upon u.

9.1 IV Estimator

The most simple case for the IV estimator is to assume the existence of a single scalar instrument z. We begin the derivation from a generic model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \tag{9.1}$$

$$\mathbf{X'Y} = \mathbf{X'X}\beta + \mathbf{X'u} \tag{9.2}$$

In the endogeneity case we know that $\mathbf{X'u} \neq 0$. Therefore, we need to apply a vector instrument $\mathbf{Z}$ in order to observe that orthogonality condition and hence obtain a consistent estimate. Therefore

$$\mathbf{Z'Y} = \mathbf{Z'X}\beta + \mathbf{Z'u} \tag{9.3}$$

which under the LS minimizing condition turns out to yield

$$\beta_{IV} = [\mathbf{Z'X}]^{-1}\mathbf{Z'Y} \tag{9.4}$$

The corresponding sample estimator is given by

$$\hat{\beta}_{IV} = [\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{Z}'(\mathbf{X}\beta + \mathbf{u}) \tag{9.5}$$

$$\hat{\beta}_{IV} = [\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{Z}'\mathbf{X}]\beta + [\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{Z}'\mathbf{u} \tag{9.6}$$

$$\hat{\beta}_{IV} = \beta + [N^{-1}\mathbf{Z}'\mathbf{X}]^{-1}N^{-1}\mathbf{Z}'\mathbf{u} \tag{9.7}$$

Instrument Quality The IV estimator will be consistent if the following two conditions are verified

$$p\lim N^{-1}\mathbf{Z}'\mathbf{u} = 0 \tag{9.8}$$

$$p\lim N^{-1}\mathbf{Z}'\mathbf{X} \neq 0 \tag{9.9}$$

Both conditions are necessary for consistency. However, a *good* instrument will also be characterized by a strong correlation with the $\mathbf{X}$ vector it is instrumenting. That is the hidden imposition from equation (9.9). Now, with heteroscedastic errors, the IV estimator is asymptotically normal with mean β and variance matrix consistently estimated by the following sandwich type estimator

$$\hat{V}[\hat{\beta}_{IV}] = [\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{Z}'\hat{\Omega}\mathbf{Z}[\mathbf{Z}'\mathbf{X}]^{-1} \tag{9.10}$$

Identification Identification issues arise when the number of unknown parameters is greater than the number of known parameters. In $y = 2x + z$, we have three unknowns (x, y and z) for only one equation, making it *unidentified*. The **order condition** requires that the number of instruments must at least equal the number of independent endogenous components, so that $r \geq K$. The model is said to be *just-identified* if $r = K$ and *overidentified* if $r > K$. The previous IV estimator is an example of a just-identified model, where the number endogeneous regressors is equal to the number of instruments. The two-stage least squares model, that will be presented next is an example of an overidentified model. In practice however, as good instruments may be extremely hard to find, the just-identified IV estimator tends to be widely used.

9.2 Two-Stage Least Squares

The Two-Stage Least Squares (2SLS) estimator is used in overidentified models, where there are more instruments than endogenous regressors. It is implemented in Stata under the code `ivregress 2sls` and is computed as the name describes, that is, in two steps. In the first step, the endogenous parameter is regressed on its instruments and on the remaining exogenous variables in the model. In the second step, the fitted values from the estimated endogenous regressor are plugged into the original regression and regressed. In equation form we will have after having estimated the endogenous regressors in the first step

$$\boldsymbol{y} = \hat{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{u} \tag{9.11}$$

where $\hat{\boldsymbol{X}}$ is the vector estimated with the instrumental variables vector $\boldsymbol{Z}$ in step one

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} \tag{9.12}$$

substituting in equation (9.11) we simply have

$$\boldsymbol{y} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{9.13}$$

which, after the minimization of the sum of squared residuals yields the 2SLS estimator $\boldsymbol{\beta_{2SLS}}$

$$\begin{aligned}
\boldsymbol{\beta_{2SLS}} &= [\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}]^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y} \\
&= [\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1}\boldsymbol{X}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} \\
&= [\hat{\boldsymbol{X}}'\boldsymbol{X}]^{-1}\hat{\boldsymbol{X}}\boldsymbol{y} \\
&= [\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}
\end{aligned} \tag{9.14}$$

or in a compact form by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{2SLS}} = [\boldsymbol{X}'\boldsymbol{P}_z\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{P}_z\boldsymbol{y}] \tag{9.15}$$

where $\boldsymbol{P_z} = \boldsymbol{Z}(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}'$. The 2SLS homoscedastic variance will be equal to

$$\hat{V}[\hat{\boldsymbol{\beta}}] = N[\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\hat{\boldsymbol{S}}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}][\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1} \tag{9.16}$$

If you think this formula is big then you should take a look at the heteroscedastic one. Luckily Stata does all of that for you, with the `, vce(robust)` option for heteroscedastic error term.

9.3 GMM IV's

The Generalized Method of Moments (GMM) consists of a pragmatic approach to econometrics. Its tremendous versatility is derived from the fact that it can be derived from the moment conditions imposed upon the error term. Take the following model

$$\mathbf{y_i} = \mathbf{x_i'}\beta + \mathbf{u_i} \tag{9.17}$$

where each component of x is viewed as being an exogenous regressor if it is uncorrelated with the error or an endogenous regressor if it is correlated. If all regressors are exogenous then LS estimators can be used, but if any components of x are endogenous then LS estimators are inconsistent for β. In the latter case, we will be using an instrument vector $\mathbf{z}$ in order to attain the orthogonality condition

$$E[\mathbf{z_i u_i}] = \mathbf{0} \tag{9.18}$$

as $\mathbf{u_i} = \mathbf{y_i} - \mathbf{x_i}\beta$ we can rewrite (9.18) as

$$E[\mathbf{z_i}(\mathbf{y_i} - \mathbf{x_i'}\beta)] = \mathbf{0} \tag{9.19}$$

Exogenous regressors can be instrumented by themselves. As there must be at least as many instruments as regressors, the challenge is to find additional instruments that at least equal the number of endogenous variables in the model.

9.3.1 GMM Estimator

The GMM estimator[1] uses the population condition defined in (9.19) to build its objective function $Q_N(\beta)$. The sample correspondent of the expectation will simply be $N^{-1}\sum_i \mathbf{z_i}(\mathbf{y_i} - \mathbf{x_i'}\beta)$ and hence the objective function

$$\begin{aligned} Q_N(\beta) &= [\boldsymbol{Z'(y - X\beta)}]'\boldsymbol{W_N}[\boldsymbol{Z'(y - X\beta)}] \\ &= [\boldsymbol{Z'y - Z'X\beta}]'\boldsymbol{W_N}[\boldsymbol{Z'y - Z'X\beta}] \end{aligned} \tag{9.20}$$

[1] for IV uses of the GMM estimator in Stata use `ivregress gmm`

where $\mathbf{W_N}$ is an $r \times r$ full-rank symmetric weighting matrix. The first order conditions for $Q_N(\beta)$ will be

$$\begin{aligned} \frac{\partial Q_N(\beta)}{\partial \beta} &= \mathbf{0} \\ -2[\boldsymbol{X'Z}]\boldsymbol{W_N}[\boldsymbol{Z'(y - X\beta)}] &= \mathbf{0} \\ \boldsymbol{X'ZW_NZ'y - X'ZW_NZ'X\beta} &= \mathbf{0} \\ \boldsymbol{X'ZW_NZ'X\beta} &= \boldsymbol{X'ZW_NZ'y} \\ \boldsymbol{\hat{\beta}_{GMM}} &= \boldsymbol{[X'ZW_NZ'X]^{-1}X'ZW_NZ'y} \end{aligned} \tag{9.21}$$

which is exactly the same as the 2SLS estimator only with W_N instead of P_z.

Similarly, the GMM variance estimator will share many similarities with the 2SLS analogue. Notice however the presence of an $\boldsymbol{\hat{S}}$ optimal weighting matrix which will be further discussed

$$\hat{V}[\hat{\beta}] = N[\mathbf{X'ZW_NZ'X}]^{-1}[\mathbf{X'ZW_N\hat{S}W_NZ'X}][\mathbf{X'ZW_NZ'X}]^{-1} \tag{9.22}$$

9.3.2 Optimal GMM

The optimal $\boldsymbol{W} = \boldsymbol{S_0^{-1}}$ where

$$\mathbf{S_0} = \lim N^{-1} \sum_{i=1}^{N} E[u_i^2 \mathbf{z_i z_i'}] \tag{9.23}$$

however, since in practice $\boldsymbol{S_0}$ is unknown we set $\boldsymbol{W}_N = \boldsymbol{\hat{S}^{-1}}$, where $\boldsymbol{\hat{S}_0}$ is consistent for $\boldsymbol{S_0}$ and estimated as follows

$$\hat{\mathbf{S}} = N^{-1} \sum_{i=1}^{N} \hat{u}_i^2 \mathbf{z_i z_i'} = \frac{\mathbf{Z'DZ}}{N} \tag{9.24}$$

where $\hat{u}_i = y_i - \mathbf{x_i'}\hat{\beta}_{GMM}$ is the GMM residual and $\mathbf{D}$ is an $N \times N$ diagonal matrix with entries $\hat{u}_i^2$.

The optimal GMM is then estimated through a two-step procedure. In the first step, a GMM estimator is obtained using a suboptimal choice of $\boldsymbol{W}_N$, usually $\boldsymbol{W}_N = \boldsymbol{I}_N$. From this step we gather the estimated squared residuals $\hat{u}_i^2$ and form estimate $\boldsymbol{\hat{S}}$. In the second step we use an optimal GMM estimator now with the newly estimated weighting matrix so that $\boldsymbol{W}_N = \boldsymbol{\hat{S}^{-1}}$. Then the optimal GMM estimator will be given by the following objective function

$$Q_N(\beta) = \left[N^{-1} \sum_{i=1}^{N} (u_i \boldsymbol{z_i}) \right]' \boldsymbol{\hat{S}^{-1}} \left[N^{-1} \sum_{i=1}^{N} (u_i \boldsymbol{z_i}) \right] \tag{9.25}$$

9.3.3 GMM vs. 2SLS

Both the optimal GMM and the 2SLS estimator lead to efficiency gains in overidentified models. Optimal GMM has the advantage of being more efficient than 2SLS, if errors are heteroscedastic, though the efficiency gain need not be great. Optimal GMM has the disadvantage of requiring additional computation compared to 2SLS. In cross-section applications it is common to use the less efficient 2SLS, though with inference based on heteroscedastic robust standard errors.

9.4 Three-Stage Least Squares

Three Stage Least Squares (3SLS) estimates a system of structural equations, where some equations contain *endogenous* and potentially *serially correlated* variables. The Three-Stage Least Squares (3SLS) method generalizes the 2SLS approach to take account of the correlations between equations in the same way that SUR generalizes OLS. 3SLS is implemented in Stata under the `reg3 (depvar1 varlist1) (depvar2 varlist2) ... (depvarN varlistN)` command. 3SLS requires three steps:

1. use OLS regressions to estimate the fitted values for all the endogeneous variables. Each of these regressions shall include *all* exogenous variables of the system as right-hand side variables[2]

2. obtain a consistent estimate of the covariance matrix of the equation disturbances. These estimates are based on the residuals from the 2SLS estimation of each structural equation

3. perform a GLS-type estimation using the covariance matrix $\hat{\boldsymbol{W}}$ estimated in the second stage with the predicted values in place of the right-hand side endogenous variables

9.4.1 3SLS Estimator

The 3SLS is a GMM estimator that uses a particular weighting matrix. To define the 3SLS estimator let $\hat{\boldsymbol{u}}_{\boldsymbol{i}} = \boldsymbol{y_i} - \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}}$ be the residuals from step 1. Define the $G \times G$ matrix

$$\hat{\boldsymbol{\Omega}} \equiv N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{u}}_{\boldsymbol{i}} \hat{\boldsymbol{u}}_{\boldsymbol{i}}' \tag{9.26}$$

[2]notice that this step is identical to the first step of the 2SLS estimation

since $\hat{\boldsymbol{\Omega}} \to^p \boldsymbol{\Omega} = E(\boldsymbol{u_i u_i'})$ the weighting matrix used in 3SLS is

$$\hat{\boldsymbol{W}} = \left[N^{-1} \sum_{i=1}^{N} \boldsymbol{Z_i' \Omega Z_i} \right)^{-1} = \left[\frac{\boldsymbol{Z'}(\boldsymbol{I_N} \otimes \hat{\boldsymbol{\Omega}})\boldsymbol{Z}}{N} \right]^{-1} \tag{9.27}$$

where $\boldsymbol{I_N}$ is the $N \times N$ identity matrix. The $\hat{\boldsymbol{\beta}}_{\boldsymbol{3SLS}}$ estimator is then given by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{3SLS}} = \left[\boldsymbol{X'Z}\{\boldsymbol{Z'}(\boldsymbol{I_N} \otimes \hat{\boldsymbol{\Omega}})\boldsymbol{Z}\}^{-1}\boldsymbol{Z'X} \right]^{-1} \boldsymbol{X'Z} \left[\boldsymbol{Z'}(\boldsymbol{I_N} \otimes \hat{\boldsymbol{\Omega}})\boldsymbol{Z} \right]^{-1} \boldsymbol{Z'Y} \tag{9.28}$$

$\hat{\boldsymbol{\beta}}_{\boldsymbol{3SLS}}$ is consistent if $E[\boldsymbol{u_i}|\boldsymbol{z_i}] = \boldsymbol{0}$ and asymptotically normal under the usual assumptions.

Chapter 10

Specification Tests Revisited

Having presented the potential sources of misspecification in linear models we now turn to the tools used for testing model fit to data. In the present chapter we present two testing philosophies for model specifications (m-tests and Hausman tests) plus a set of some other tests for some common model misspecifications.

10.1 M-tests

M-tests use moment conditions in a similar fashion to the GMM approach with the distinction that moment conditions are not imposed in the estimation but rather used for testing. M-tests are usually implemented using auxiliary regressions and estimated via ML. You can find them in Stata under the `_mtest` command. Unsurprisingly, the null is the population moment orthogonality condition:

$$H_0 : E[\boldsymbol{m_i}(\boldsymbol{w_i}, \boldsymbol{\theta})] = \mathbf{0} \tag{10.1}$$

The sample moment correspondent to (10.1) will be our **m-test**, which will test of the closeness to zero of the null

$$\hat{\boldsymbol{m}}_{\boldsymbol{N}}(\hat{\boldsymbol{\theta}}) = N^{-1} \sum_{i=1}^{N} \boldsymbol{m_i}(\boldsymbol{w_i}, \hat{\boldsymbol{\theta}}) \tag{10.2}$$

where $\boldsymbol{w}$ is a vector of observables, usually the dependent variable y and the regressors $\boldsymbol{x}$ and sometimes additional variables $\boldsymbol{z}$, $\boldsymbol{\theta}$ is a $q \times 1$ vector of parameter and $\boldsymbol{m}_i(.)$ is an $h \times 1$ vector. This approach is similar to that for the Wald test, where $\boldsymbol{h}(\boldsymbol{\theta}) = 0$ is tested by testing the closeness to zero of $\boldsymbol{h}(\hat{\boldsymbol{\theta}})$. A chi-square test statistic can then be obtained by

taking the corresponding quadratic form. Thus the **m-test statistic** is

$$M = N\hat{\boldsymbol{m}}_{\boldsymbol{N}}(\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{V}}_{\boldsymbol{m}}^{-\mathbf{1}}\hat{\boldsymbol{m}}_{\boldsymbol{N}}(\hat{\boldsymbol{\theta}}) \sim \chi^2(\text{rank}[\boldsymbol{V}_m]) \tag{10.3}$$

The m-test rejects the moment conditions (10.1) at significance level α $M > \chi^2_\alpha$ and does not reject otherwise.The m-test approach is conceptually very simple. The **moment** restriction (10.1) is rejected if a quadratic form in the **sample** estimate (10.2) is far enough from zero. The true practical complication lies in the estimation of the inverse matrix variance[1] $\boldsymbol{V}_{\boldsymbol{m}}^{-\mathbf{1}}$.

10.1.1 CM test

The **conditional moment test** (cm) is an m-test based on the implied unconditional moment restrictions

$$E[r(y, \boldsymbol{x}, \boldsymbol{\theta})|\boldsymbol{x}] = 0 \tag{10.4}$$

for some scalar function $r(.)$. The conditional moment test will take use of this condition and redefine it as an *unconditional* moment restriction

$$E[\boldsymbol{g}(\boldsymbol{x})r(y, \boldsymbol{x}, \boldsymbol{\theta})] = 0 \tag{10.5}$$

where $\boldsymbol{g}(\boldsymbol{x})$ and/or $r(y, \boldsymbol{x}, \boldsymbol{\theta})$ are chosen so that these restrictions **are not already used** in estimation.

10.1.2 Test of Overidentifying Restrictions

Tests for overidentifying restrictions are m-tests. In an overidentified model, there is an excess of $r - q$ unused orthogonality conditions which can be used to form an m-test which follows a chi-square distribution with $r - q$ degress of freedom.

10.2 Hausman Test

The Hausman test is based on the comparison between two the properties of two competing estimators. Hypotheses are

$$\boldsymbol{H_0} \quad : \quad p\lim(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) = 0 \tag{10.6}$$

$$\boldsymbol{H_a} \quad : \quad p\lim(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \neq 0 \tag{10.7}$$

More intuitively we have that under the null

[1] In fact many practitioners do it via bootstrapping techniques discussed in chapter 11.

- estimator $\hat{\boldsymbol{\theta}}$ will be consistent but inefficient
- estimator $\tilde{\boldsymbol{\theta}}$ will be consistent and efficient

and that under the alternative

- estimator $\hat{\boldsymbol{\theta}}$ will *still* be consistent but inefficient
- estimator $\tilde{\boldsymbol{\theta}}$ will be become inconsistent

The Hausman test is therefore a test on both the value and the variance of the estimator. It will take the following form

$$H = [\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}]'[V(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})]^{-1}[\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}] \tag{10.8}$$

The obvious trouble here is how to disentangle the two variances. Fortunately Hausman figured that out for us

$$V_H = V[(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})] = V[\hat{\boldsymbol{\theta}}] + V[\tilde{\boldsymbol{\theta}}] - 2Cov[\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}] \tag{10.9}$$

if $\hat{\theta}$ is the efficient estimator in the null hypothesis model, then $Cov[\hat{\theta}, \tilde{\theta}]$ is simply the variance of $\hat{\theta}$, so that

$$\begin{aligned} V[(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})] &= V[\hat{\boldsymbol{\theta}}] + V[\tilde{\boldsymbol{\theta}}] - 2V[\hat{\boldsymbol{\theta}}] \\ &= V[\tilde{\boldsymbol{\theta}}] - V[\hat{\boldsymbol{\theta}}] \end{aligned} \tag{10.10}$$

which, plugged back into (10.8) turns out to be

$$H = [\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}]'[V(\tilde{\boldsymbol{\theta}}) - V(\hat{\boldsymbol{\theta}})]^{-1}[\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}] \sim \chi^2(k) \tag{10.11}$$

where $\dim[\hat{\boldsymbol{\theta}}] = \dim[\tilde{\boldsymbol{\theta}}] = k \times 1$

The Hausman test. Typical applications of this test are the the panel data comparison of fixed-effects vs. random-effects estimator or the endogeneity test for the 2SLS vs. OLS estimators. In the latter case we test the null that the 2SLS estimator is consistent but inefficient under the null but still consistent under the alternative (provided the $\mathbf{z}$ vector of instruments is appropriate).

10.3 Common Misspecifications

Model misspecifications have previously been discussed in chapter 3. Here we provide some further checks for heteroscedasticity, exogeneity of instruments and nonlinearity of the regressors.

10.3.1 Heteroscedasticity

Heteroscedasticity occurs whenever a random variable has differing variances across its distribution. Even though heteroscedasticity *per se* does not affect the consistency and can be easily corrected with the usage weights in the variance-covariance matrix, it can quite often be a symptom of of more troubling econometric issues as endogeneity (as is the case in figure 10.3.1). Heteroscedasticity is usually dealt with by using the GLS model presented in chapter 2.

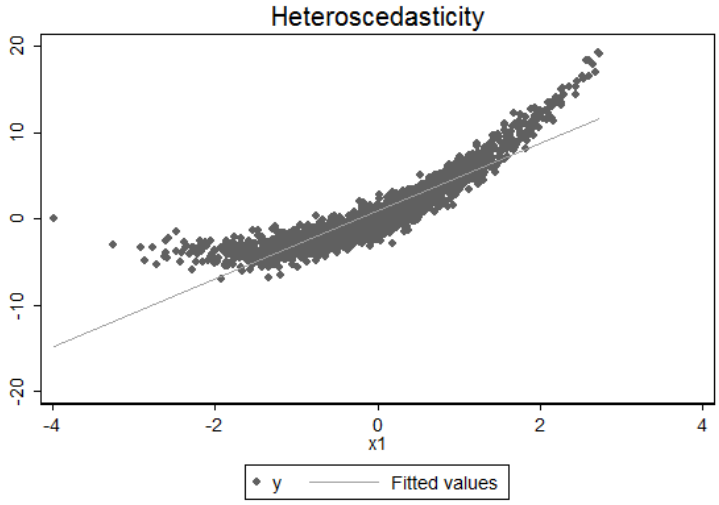

Figure 10.1: Example of Heteroscedasticity

Breusch-Pagan test The Breusch Pagan (BP) test, implemented in Stata as a postestimaton command for linear regressions `estat hettest` is an chi-square test with $n\chi^2(k)$ degrees of freedom. The BP tests the null of constant variance in the model (no heteroscedasticity) and can easily computed as follows

1. run the OLS regression

$$y = \beta_0 + \boldsymbol{x}_1'\beta_1 + ... + \boldsymbol{x}_n'\beta_n + \boldsymbol{z}_1'\beta_1 + ... + \boldsymbol{z}_n'\beta_n + \boldsymbol{u} \tag{10.12}$$

2. obtain the $\hat{u}^2$ residuals of the estimated OLS regression equation

3. use the squared residuals $\hat{u}^2$ as the dependent variable in a secondary equation that includes the independent variables **suspected** of being related to error term

$$\hat{u}^2 = \beta_0 + \boldsymbol{z}_1'\beta_1 + ... + \boldsymbol{z}_n'\beta_n + \boldsymbol{u} \tag{10.13}$$

4. perform a chi-square test on the mode estimated in step 3. If the test confirms that the independent variables are jointly significant then we can *reject the hypothesis of no heteroscedasticity*. Notice however that under the null, the BP test assumes normally distributed errors as it uses the restriction $E[u^4|\boldsymbol{x}^4] = 3\sigma^4$.

White test Unlike the BP test the White test does not require modeling heteroscedasticity, as heteroscedastic-robust standard errors can be computed under minimal distributional assumptions. Asides from that, the White test is quite similar to the BP test. It is computed as follows

1. run the OLS regression
$$y = \beta_0 + \boldsymbol{x}_1'\beta_1 + ... + \boldsymbol{x}_n'\beta_n + \boldsymbol{u} \tag{10.14}$$
2. obtain the $\hat{u}^2$ residuals of the estimated OLS regression equation
3. use the squared residuals $\hat{u}^2$ as the dependent variable and estimate the following equation where $\boldsymbol{X}$'s are **all** explanatory variables from the original equation
$$\hat{u}^2 = \beta_0 + \boldsymbol{x}_1'\beta_1 + ... + \boldsymbol{x}_n'\beta_n + \boldsymbol{u} \tag{10.15}$$
4. test the joint hypothesis that all the coefficients are zero (Chi-square test)

The White test is equivalent to the LM test $LM = NR^2$. As mentioned the advantage of the White test is that it requires no induction on which regressors are causing heteroscedasticity.

10.3.2 OIR Tests

If an IV estimator is used then the instruments must be exogenous for the IV estimator to be consistent. For just-identified models it is not possible to test for instrument exogeneity. Instead, a priori arguments need to be used to justify instrument validity. The Sargan (which by the way is an m-test) is the standard over-identifying restrictions test. It sets the null at the moment condition level $H_0 : E[\boldsymbol{z}u] = \boldsymbol{0}$ which corresponds to the sample equivalent

$$N^{-1}\sum_{i=1}^{N} u_i^2 \boldsymbol{z_i}\boldsymbol{z_i'} \sim \chi^2(r-q) \tag{10.16}$$

leading to the Sargan test OIR

$$\text{OIR} = \hat{\boldsymbol{u}}'\boldsymbol{Z}\hat{\boldsymbol{S}}^{-1}\boldsymbol{Z}'\hat{\boldsymbol{u}} \tag{10.17}$$

where $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{S}} = \hat{\sigma}^2\boldsymbol{Z}'\boldsymbol{Z}$ is consistent for $p\lim N^{-1}\sum_i u_i^2\boldsymbol{z_i}\boldsymbol{z_i'}$. If OIR is large then the moment conditions are rejected and the IV estimator is inconsistent. Rejection of H_0 is usually interpreted as evidence that the instruments $\boldsymbol{z}$ are endogenous, but it could also be evidence of model misspecification so that in fact $y \neq \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{u}$. In either case rejection indicates problems for the IV estimator.

Chapter 11

Bootstrapping

Bootstrapping is a computer intensive method based in Monte Carlo simulation and re-sampling techniques which allow for an approximation asymptotic results using finite-samples. The wide range of bootstrap methods can be classified into (1) statistical inference when conventional methods such as standard error computation are difficult to implement and (2) provide asymptotic refinements that can lead to a better approximation in finite samples.

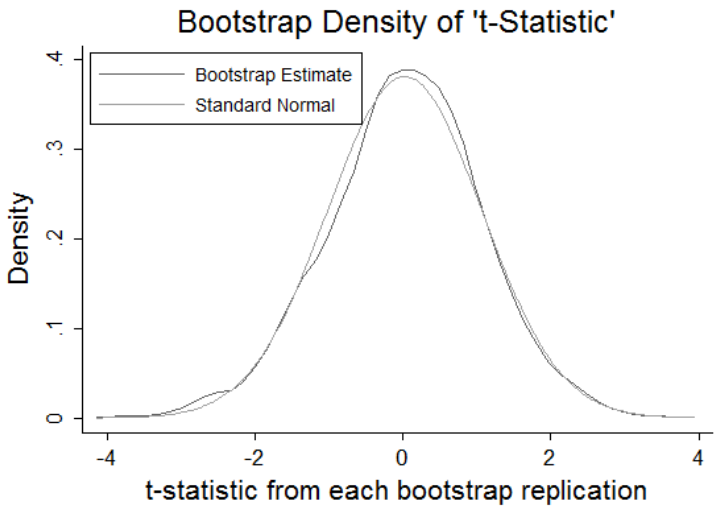

Figure 11.1: Bootstrap density of t-test statistic

11.1 Asymptotic Pivotal Statistic

For asymptotic refinement to occur, the statistic being bootstrapped must be an asymptotically pivotal statistic, meaning a statistic whose limit distribution does not depend on unknown parame-

ters. As an example. consider sampling from) $y_i \sim [\mu, \sigma^2]$. Then the estimate $\hat{\mu} = \bar{y} \sim^a \mathcal{N}[\mu, \sigma^2]$ is not asymptotically pivotal even given a null hypothesis value $\mu = \mu_0$ since its distribution depends on the unknown parameter σ^2. However, the studentized statistic, $t = (\hat{\mu} - \mu_0)/s_{\hat{\mu}} \sim^a \mathcal{N}[0, 1]$ is asymptotically pivotal.

11.2 The Bootstrap Procedure

11.2.1 Boostrap Algorithm

A general **bootstrap algorithm** works as follows:

1. Given data $\boldsymbol{w_i}, ..., \boldsymbol{w_N}$, draw a bootstrap sample of size N using a method given in 11.2.2 and denote this new samples as $\boldsymbol{w_i^*}, ..., \boldsymbol{w_N^*}$
2. Calculate a desired statistic using the bootstrap sample. Examples include
   - the estimate $\hat{\theta}^*$ of θ
   - the standard error $s_{\hat{\theta}}$ of the estimate $\hat{\theta}^*$
   - a t-statistic $t^* = (\hat{\theta}^* - \hat{\theta})/s_{\hat{\theta}}$ centered at the original estimate $\hat{\theta}$
3. Repeat steps 1.. and 2. B independent times where B is a larger number of bootstrap replications of the statistic of interest
4. Use there B bootstrap replications ti obtain a bootstrapped version of the statistic

11.2.2 Bootstrap Sampling

The bootstrap d.g.p. in step 1. is used to approximate the true unknown d.g.p. We present three sampling techniques for bootstrap sampling

Empirical distribution function use the empirical distribution of the data and perform a sampling with replacement on that data B times. Remember that since the sampling is done with replacement observations can be repeated or not show up within each new sample.

Parametric bootstrap if the conditional distribution $y|\boldsymbol{x} \sim F(\boldsymbol{x}, \boldsymbol{\theta_0})$ is known, then the bootstrapped estimates will follow the same $F(\boldsymbol{x}, \boldsymbol{\theta_0})$ parametric distribution.

Residual bootstrap Each residual is randomly multiplied by a random variable with mean 0 and variance 1. This method assumes that the 'true' residual distribution is symmetric and can offer advantages over simple residual sampling for smaller sample sizes.

11.2.3 Number of Bootstraps

The bootstrap asymptotics rely on $N \to \infty$ and so the bootstrap can be asymptotically valid even for low B. However, clearly the bootstrap is more accurate as $B \to \infty$. Pedro Portugal's rule of thumb is 999. Some other authors propose 399 for tests at level 0.05 and $1,499$ for tests at 0.01.

11.2.4 Standard Error Estimation

The bootstrap estimate of variance of an estimator is given by

$$s^2_{\hat{\theta},boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_b - \bar{\hat{\theta}}^*)^2 \tag{11.1}$$

where

$$\bar{\hat{\theta}}^* = B^{-1}\sum_{b=1}^{B}\hat{\theta}^*_b \tag{11.2}$$

11.2.5 Hypothesis Testing

Tests with asymptotic refinement Asymptotic refinement is achieved through the repetition of the bootstrapping procedure B times. As mentioned as $B \to \infty$ the more and more accurate the bootstrapped estimates become.

1. Consider a t-test: compute B bootstrap replications of B test statistics, $t*_1, ..., t*_B$ where

$$t^*_b = \frac{(\hat{\theta}^*_b - \hat{\theta})}{s_{\theta^*_b}} \tag{11.3}$$

2. order the empirical distribution of $t*_1, ..., t*_B$ from smallest to largest

3. get the bootstrap critical value for the defined α level of confidence

Tests without asymptotic refinement Alternative bootstrap methods can be used that although asymptotically valid do not provide an asymptotic refinement.

1. compute t

$$t = \frac{(\hat{\theta} - \theta_0)}{s_{\hat{\theta},boot}} \tag{11.4}$$

   where θ_0 is the null hypothesis testing parameter

2. compare the single test statistic to critical values from the standard normal distribution

11.2.6 Sampling Bias Reduction

The bootstrap estimate of the sampling bias is

$$\text{Bias}_{\hat{\theta}} = (\hat{\bar{\theta}}^* - \hat{\theta}) \tag{11.5}$$

Suppose, for example that $\hat{\theta} = 4$ and $\hat{\bar{\theta}}^* = 5$. Then the estimated bias is $(5 - 4) = 1$, an upward bias of 1 that needs to be subtracted from $\hat{\bar{\theta}}^*$. More generally the **bootstrap bias-corrected estimator** of θ is

$$\begin{aligned} \hat{\theta}_{Boot} &= \hat{\theta} - (\hat{\bar{\theta}}^* - \hat{\theta}) \\ &= 2\hat{\theta} - \hat{\bar{\theta}}^* \end{aligned} \tag{11.6}$$

In practice however bias correction is seldom used for $\sqrt{N}$-consistent estimators, as the bootstrap estimate can be more variable than the original estimate $\hat{\theta}$ and the bias is often small relative to the standard error of the estimate. Bootstrap bias correction is used for estimators that converge at rate less than $\sqrt{N}$, notably nonparametric regression and density estimators.

Bibliography

[1] A. C. Cameron and P. K. Trivedi, *Microeconometrics: Methods and Applications*, Cambridge University Press, 2005.

Typeset in LaTeX

This Version: JUN. 2013