

UNIVERSITY OF ESSEX
DEPARTMENT OF ECONOMICS

EC996: Econometrics

Lecture Notes 2011/12
(Preliminary and incomplete)

João Santos Silva

jmcss@essex.ac.uk

Room: 5B.204

Tel.: (01206) 872769

Office hours: Thursday 16:00-18:00

- **Why Econometrics**

- Econometrics is the set of techniques that brings together economic theory and data.
- Any economist needs to have some familiarity with at least some econometric tools.
- The recent and significant advances in econometrics are left out of most undergraduate and masters courses.

- **Overview of the course**

- The course is roughly divided into three parts:
 1. Linear models (least squares based);
 2. General estimation methods;
 3. Non-linear models.
- The main objective is to give an overview of the most relevant econometric techniques.
- Empirical illustrations of most of the methods covered will be provided.

- **Reading list**

Greene, W.H. (2012) *Econometric Analysis* (7th edition), Prentice Hall

Cameron, A.C. and P.K. Trivedi (2005) *Microeconometrics: Methods and Applications*, CUP

Cameron, A.C. and P.K. Trivedi (2009) *Microeconometrics using Stata*, Stata press

Davidson, R. and J. MacKinnon (2004) *Econometric Theory and Methods*, OUP

Goldberger, A. S. (1991) *A Course in Econometrics*, Harvard

Lee, M.-j. (2010) *Micro-Econometrics 2nd ed*, Springer

Wooldridge, J.M. (2010) *Econometric Analysis of Cross Section and Panel Data 2nd ed*, MIT

- **Assessment**

- 30 percent Course Work Mark, 70 percent Exam Mark (NO MAX RULE!).

- One-hour mid term exam: 21 November at 5pm.

- Two-hour exam in January.

University week: 2

Regression Equations and Systems of Regression Equations

1. Mean regression and ordinary least squares;
2. Seemingly unrelated regressions;
3. Singular systems.

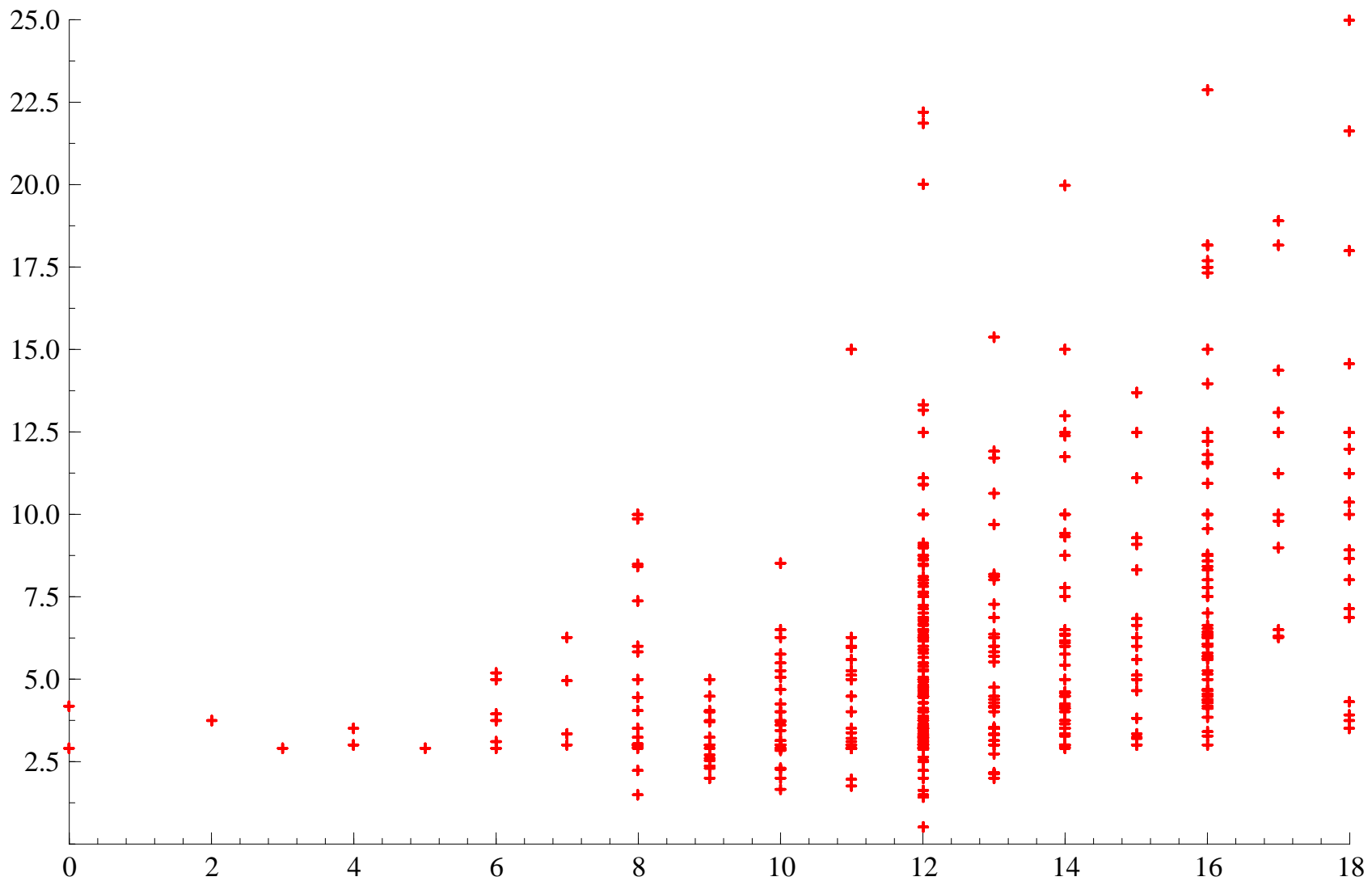
Reading: Greene: 1-4, 10.1, 10.2, 10.5.

1) Mean regression and ordinary least squares

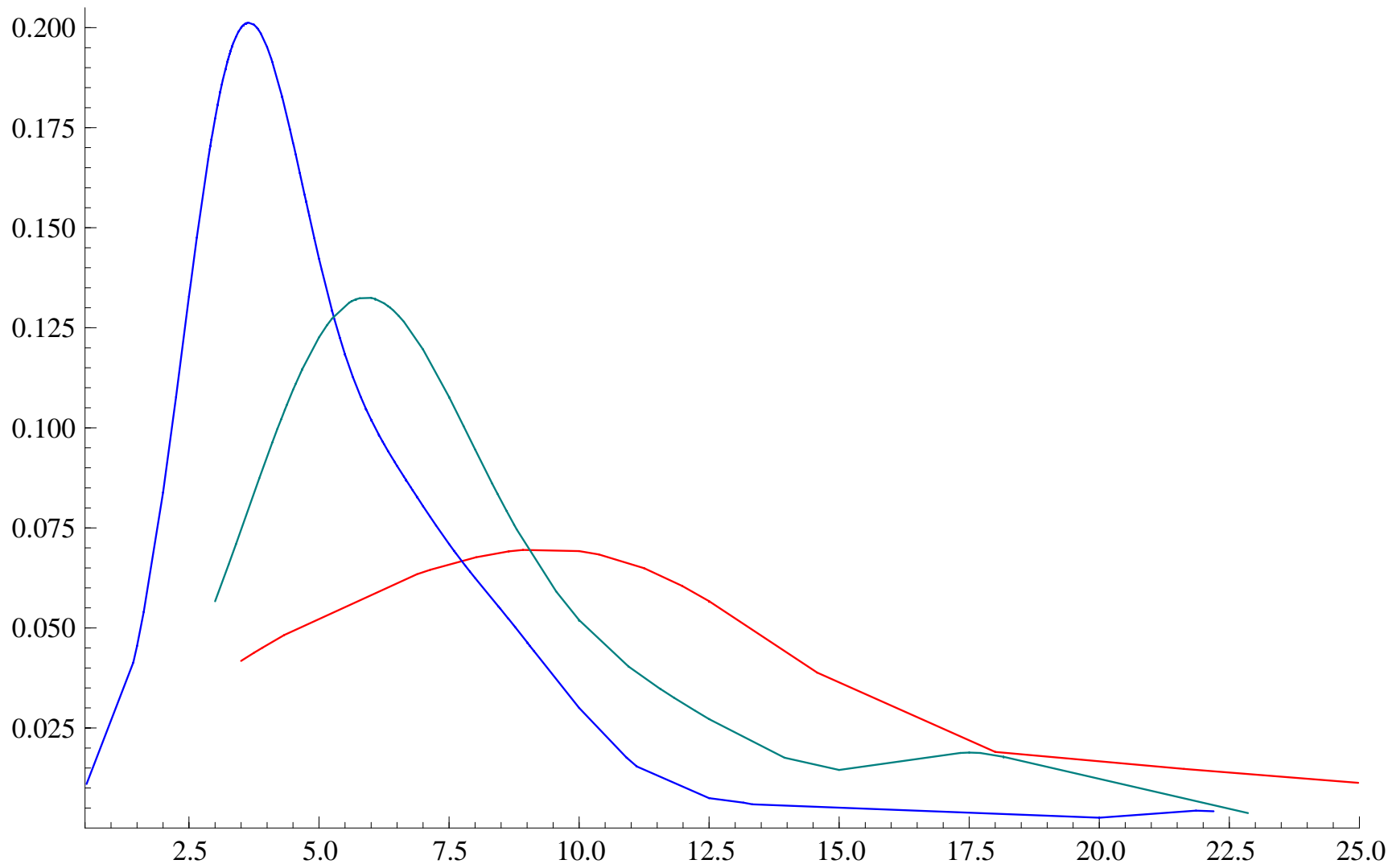
- Economic theory is interested in the relation between certain economic variables.
- Typically, economic theory leads to deterministic functions relating the variables of interest, e.g., $Q_i = AK_i^\alpha L_i^\beta$, $Q_i = f(P_i)$, $C_i = \alpha + \beta Y_i$.
- However, in reality, economic variables are random and are not related by deterministic functions.
- If we want to estimate the unknown parameters of these functions, we need to use data which are obtained by sampling from the joint distribution of the relevant variables.
- We observe data from the joint distribution

$$f(y, x) = f(y|x) g(x)$$

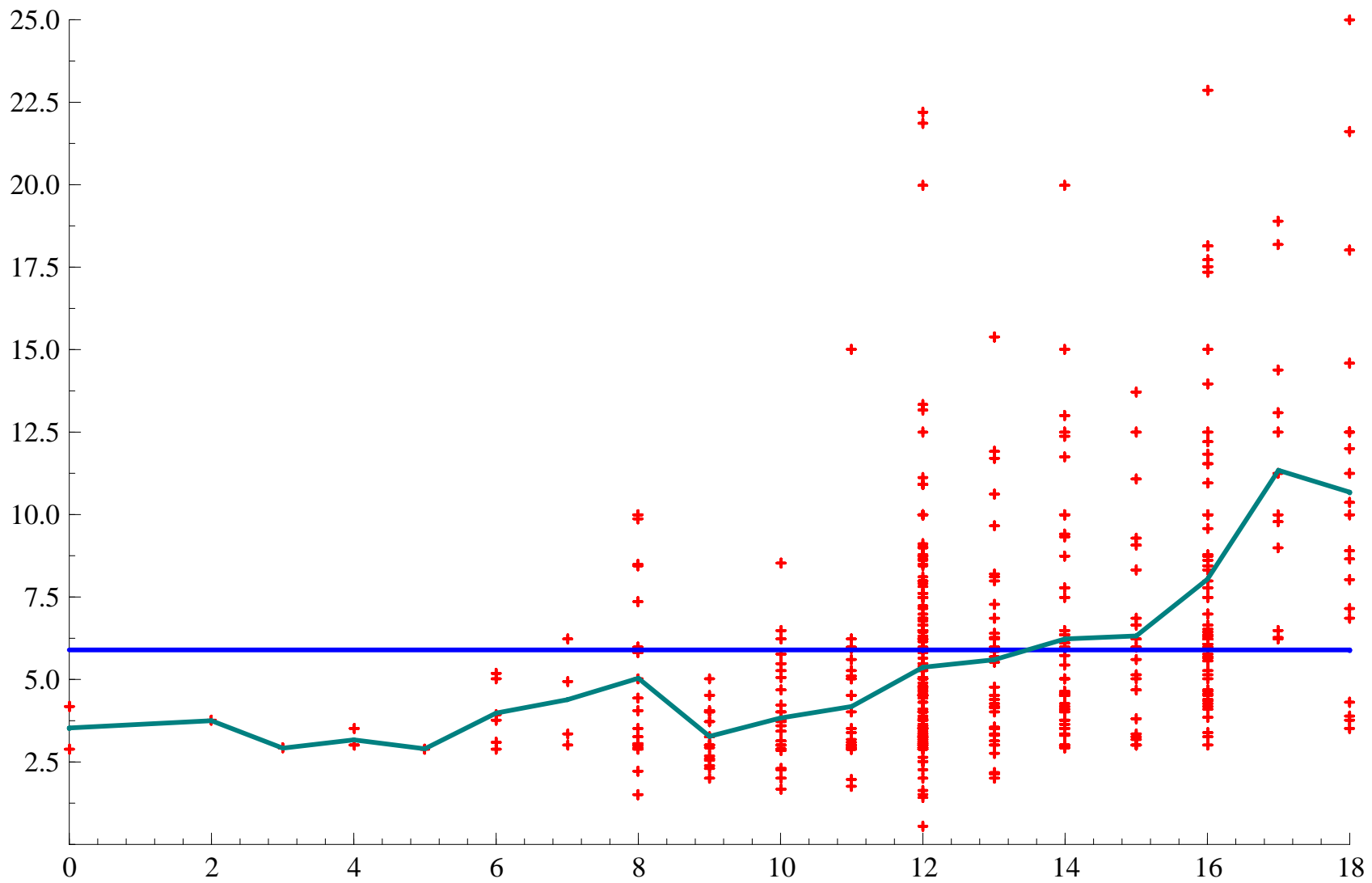
- By conditioning on x we can do *ceteris paribus* analysis because we are able to identify the effect of any regressor on y , keeping all others fixed.
- Defining (and observing) the set of conditioning variables can be tricky.



Hourly wage vs years of education (US data)



Estimated wage densities for 12, 16, and 18 years of education



Hourly wage vs years of education (US data)

- Regression: The regression of y on x is any characteristic of the conditional distribution $f(y|x)$ expressed as a function of x .
- Examples are:
 - Mean regression: $E(y|x)$;
 - Quantile regression: $Q_y(\alpha|x)$;
 - Mode regression: $\text{mode}(y|x)$;
 - Skedastic regression: $\text{Var}(y|x)$.
- Often, interest is focused on the mean regression because $E(y|x)$ is the function of x that minimizes

$$E \left[(y - h(x))^2 \right].$$

- However, $Q_y(0.5|x)$ minimizes $E[|y - h(x)|]$, and other regressions minimize different loss functions.

- Let $\varepsilon = y - h(x)$ denote the errors of a regression.
- The properties of ε depend on the **specific regression** defining $h(x)$.
- For the mean regression, we have $\varepsilon = y - E(Y|x)$, with
 - (a) $E(\varepsilon|x) = 0$
 - (b) $E(\varepsilon) = 0$
 - (c) $\text{Var}(\varepsilon|x) = \sigma_{y|x}^2$
 - (d) $\text{Cov}(\varepsilon, x) = 0$
 - (e) $\text{Cov}(\varepsilon, g(x)) = 0$
- Only the deviations between y and $E(y|x)$ verify (e).
- These properties can be used to define an estimator of $h(x)$ using the *analogy principle*.

- Let x be a scalar random variable and assume that $E(y|x) = \beta_0 + \beta_1 x$.
- Properties (b) and (d) imply

$$E(y - \beta_0 - \beta_1 x) = 0$$

$$\text{Cov}(y - \beta_0 - \beta_1 x, x) = 0,$$

which in turn lead to

$$E[(y - \beta_0 - \beta_1 x)x] = 0.$$

- For a sample of size n , β can be estimated (using the **analogy principle** or the method of moments), by solving the system

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \quad \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0.$$

- This coincides with the least squares estimator, which is defined by

$$\hat{\beta} = \arg \min_{b_0, b_1} \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2,$$

and leads to the first-order conditions

$$-2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \quad -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

which are identical to the ones seen before.

- Using an obvious notation, the estimator can be written in matrix form as

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

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

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

- I assume that you are familiar with the standard assumptions and properties of the OLS estimator.
- I also assume that you are familiar with standard inference tools, such as:
 - generalized least squares and weighted least squares;
 - robust covariance matrix estimation;
 - heteroskedasticity tests (White and Breusch-Pagan);
 - serial-correlation tests (Durbin-Godfrey);
 - linearity tests (RESET);
- See, for example, Greene's book for details.

2) Seemingly unrelated regressions (SUR)

- In many situations we may have multiple dependent variables.
- *Capital Asset Pricing Model (CAPM)*:

$$r_{it} - r_{ft} = \alpha_i + \beta_i(r_{mt} - r_{ft}) + \varepsilon_{it},$$

where r_{it} , r_{ft} and r_{mt} denote the **returns** in period t on **asset i** , on a **risk-free** asset, and on the **market portfolio**, respectively.

- *Grunfeld-Boot and de Witt investment model*:

$$I_{it} = \beta_{1i} + \beta_{2i}F_{it} + \beta_{3i}C_{it} + \varepsilon_{it},$$

where I_{it} , F_{it} and C_{it} denote **gross investment** by firm i in period t , the **market value** of the firm at the end of the previous period, and the value of the **capital stock** of the firm at the end of the previous period, respectively.

- Note that, for the CAPM, the regressor is the same for all equations.

- Joint modelling makes sense because there may be **efficiency gains** from:
 - imposing cross-equation restrictions;
 - exploring error correlation across equations.
- Joint estimation is also useful if we wish to test **cross-equation hypotheses**.
- Notice that this setup does not allow for simultaneity.
- A standard way of writing the general model in matrix form is:

$$y_i = X_i\beta_i + \varepsilon_i, \quad i = 1, \dots, M,$$

with $\varepsilon_i = [\varepsilon_{i1}, \dots, \varepsilon_{iT}]'$, etc.

- Assuming that these are models for $E(y_i|X_1, \dots, X_M)$, we have $E(\varepsilon_i|X_1, \dots, X_M) = 0$.
- Notice that OLS equation-by-equation would estimate $E(y_i|X_i)$ for $i = 1, \dots, M$.
- This strict-exogeneity assumption can be relaxed by modelling $E(y_{it}|X_{1t}, \dots, X_{Mt})$.

- Let $\varepsilon = [\varepsilon'_1, \dots, \varepsilon'_M]'$ and define $E(\varepsilon\varepsilon'|X_1, \dots, X_M) = \Omega$
- It is common to assume

$$E(\varepsilon_i\varepsilon'_j|X_1, \dots, X_M) = \sigma_{ij}I_T \quad \Omega = \begin{bmatrix} \sigma_{11}I_T & \dots & \sigma_{1M}I_T \\ \vdots & \ddots & \vdots \\ \sigma_{M1}I_T & \dots & \sigma_{MM}I_T \end{bmatrix} = \Sigma \otimes I_T,$$

where $\Sigma = [\sigma_{ij}]$, which is assumed non-singular.

- OLS estimation equation-by-equation is the same as OLS estimation of the model

$$Y = X\beta + \varepsilon$$

where $Y = (y'_1, \dots, y'_M)'$, $\beta = (\beta'_1, \dots, \beta'_M)'$ and X is the block-diagonal matrix:

$$X = \begin{bmatrix} X_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & X_M \end{bmatrix}.$$

- OLS, however, ignores the structure of Ω .
- The optimal estimator is the GLS, which is given by

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

- Because $\Omega = (\Sigma \otimes I_T)$, $\Omega^{-1} = (\Sigma^{-1} \otimes I_T)$, and so

$$\begin{aligned} \hat{\beta}_{GLS} &= [X'(\Sigma^{-1} \otimes I_T)X]^{-1} X'(\Sigma^{-1} \otimes I_T)Y \\ &= \begin{bmatrix} \sigma^{11}X'_1X_1 & \dots & \sigma^{1M}X'_1X_M \\ \vdots & \ddots & \vdots \\ \sigma^{M1}X'_MX_1 & \dots & \sigma^{MM}X'_MX_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^M \sigma^{1j}X'_1y_j \\ \vdots \\ \sum_{j=1}^M \sigma^{Mj}X'_My_j \end{bmatrix} \end{aligned}$$

with $\Sigma^{-1} = [\sigma^{ij}]$ and $\text{Var}(\hat{\beta}_{GLS}) = (X'\Omega^{-1}X)^{-1}$.

- Notice that the conditions for consistency of GLS are stronger than those for OLS.

- There are two special situations where GLS is identical to OLS in the SUR:
 - if $\sigma_{ij} = 0$ for all $i \neq j$; because the equations are actually unrelated;
 - all equations have identical explanatory variables (as in the CAPM); because in this case $E(y_i|X_1, \dots, X_M) = E(y_i|X_i)$.
- More generally we have that:
 - The greater the correlation between the disturbances, the greater the gain in efficiency from using GLS.
 - The less correlation there is between the X_i matrices, the greater the gain in efficiency from using GLS.
- Under the stated assumptions, if $\text{plim} \left(\frac{1}{T} X_i' X_j \right) = H_{ij}$ exists for all i and j , and H_{ii} is non-singular for all i , then GLS is consistent (large T asymptotics).
- Under somewhat stronger conditions GLS will be asymptotically normal.

- Except when it is identical to OLS, GLS is infeasible.
- Feasible GLS (FGLS) requires an estimator of Σ .
- A two-step algorithm can be used:
 - (a) Obtain the vectors of OLS residuals for equation i (equation-by-equation) and then estimate σ_{ij} by $s_{ij} = e'_i e_j / T$;
 - (b) Setting $\widehat{\Sigma} = [s_{ij}]$, obtain $\widehat{\beta}_{FGLS} = \left[X' \left(\widehat{\Sigma}^{-1} \otimes I_T \right) X \right]^{-1} X' \left(\widehat{\Sigma}^{-1} \otimes I_T \right) Y$, with $\widehat{\text{Var}} \left(\widehat{\beta}_{FGLS} \right) = \left[X' \left(\widehat{\Sigma}^{-1} \otimes I_T \right) X \right]^{-1}$.
- Under standard regularity conditions, equation-by-equation OLS is consistent and hence $\widehat{\Sigma}$ is consistent for Σ , which leads to $\widehat{\beta}_{FGLS}$ being consistent.
- Furthermore, FGLS is asymptotically equivalent to GLS (and iterating FGLS generates the ML estimator under normality).

- Restrictions on the regression coefficients, including cross-equation restrictions, can be tested using Wald statistics.
- The q restrictions $H_0 : R\beta_0 = r$ can be tested using

$$W = \left(R\hat{\beta}_{FGLS} - r \right)' \left[R \left[\widehat{\text{Var}} \left(\hat{\beta}_{FGLS} \right) \right] R' \right]^{-1} \left(R\hat{\beta}_{FGLS} - r \right) \stackrel{a}{\sim} \chi_{(q)}^2$$

- As usual, the validity of the Wald test depends on homoskedasticity and on absence of serial correlation:
 - Serial correlation should be dealt with by using the appropriate dynamic specification;
 - Heteroskedasticity robust covariance matrices can be used.
- LM and LR tests can also be used but assume joint normality of ε .
- As always, measures of goodness-of-fit are of little use.

3) Singular systems

- So far we have assumed that Σ is non-singular.
- However, in a number of situations, we may have that Σ is singular.
- Consider, for example, the Almost Ideal Demand System of Deaton and Muellbauer (1980)

$$w_i = \alpha_i + \sum_{j=1}^M \gamma_{ij} \ln p_j + \beta_i \ln(X/P) + \varepsilon_i$$

where w_i is the budget share associated with the i -th good, p_j is the price on the j -th good, X is the total expenditure and P is a price index.

- Notice that $\sum_{i=1}^M w_i = 1$ which implies

$$\sum_{i=1}^M \alpha_i = 1, \quad \sum_{i=1}^M \gamma_{ij} = \sum_{i=1}^M \beta_i = 0 \quad \text{and} \quad \sum_{i=1}^M \varepsilon_i = 0.$$

- This means that only $M - 1$ error terms are independent and therefore Σ has rank $M - 1$.
- In this case Σ^{-1} does not exist and GLS cannot be implemented as outline above.
- One possible solution (advocated by Judge, Griffiths, Hill, Lütkepohl and Lee, 1985) is to use Σ^+ in place of Σ^{-1} .
- Alternatively, we can drop one equation and estimate its parameters from the aggregation restrictions.
 - If FGLS is used, finite sample results depend on which equation is dropped.
 - With iterated FGLS (ML) the results are invariant to the choice of equation to drop.

University week: 3

Panel Data

1. Nature of panel data and the basic model;
2. Pooled regression;
3. Random effects;
4. Fixed effects;
5. Hausman test;
6. Dynamic models.

Reading: Greene: 11.1–11.6, 11.8.

1) Nature of panel data and the basic model

- In panel data (also called longitudinal data) we follow a group of individual units over time.
- These units can be micro-level entities such as households, workers and firms, or macro-level entities such as cities, regions or countries.
- Typically, in **micro panel data**, we have a very large number of units relative to the number of time periods per unit; whereas in **macro panel data** we have fewer units and more time periods.
- In general, we have n units each of whom we follow for T_i consecutive time periods (for unit $i = 1, \dots, n$).
- Here we will focus on the case where T_i is small and n is large.
- That is, we will be concerned with the large sample properties of the estimators when T_i is fixed and $n \rightarrow \infty$.

- Another restriction is that we will consider only balanced panels in which T_i is the same for all i , and there are no missing observations.
- Unbalanced panels can be dealt with much in the same way, but we have to take into account that **attrition** may bias the sample.
- Panel data have major benefits over cross-section and time-series data:
 - allows the elimination of certain forms of omitted variable bias;
 - allows the study of dynamics at a micro level.
- For example we can use panel data to distinguish between the effects of unmeasured differences in the propensity to participate in the labour force from the effects of transition into and out of the labour force.
- However, because we have a time dimension, we need to worry about problems not present in cross-sections, like the time-series properties of the data.

- The general form of panel data model we shall consider is

$$y_{it} = x'_{it}\beta + z'_i\phi + \varepsilon_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T, \quad (1)$$

where

y_{it} observable dependent variable

x_{it} $k \times 1$ vector of observable regressors (without an intercept)

β $k \times 1$ vector of unobservable parameters

$z'_i\phi$ **individual effect**

z_i $h \times 1$ vector with individual specific variables:
observable (e.g. gender, location) or unobservable (e.g. tastes)

ϕ $h \times 1$ vector of unobservable parameters

ε_{it} scalar random disturbance with properties to be defined below

n number of cross-sectional units (e.g., individuals)

T number of time periods (e.g., years)

- For the model to have meaning, we need to specify the properties of the error term.
- We start by assuming **strict exogeneity** and interpret our model as

$$E(y_{it} | x_{i1}, \dots, x_{iT}, z_i) = x'_{it}\beta + z'_i\phi \quad \Rightarrow \quad y_{it} = x'_{it}\beta + z'_i\phi + \varepsilon_{it},$$

where ε_{it} satisfies $E(\varepsilon_{it} | x_{i1}, x_{i2}, \dots, z_i) = 0$, and we assume $E(\varepsilon_{it}^2 | x_{i1}, x_{i2}, \dots, z_i) = \sigma^2$ and $E(\varepsilon_{it}\varepsilon_{js} | x_{i1}, x_{i2}, \dots, z_i) = 0$ if $i \neq j$ and/or $t \neq s$.

- Notice that this is stronger than contemporaneous exogeneity, which would imply only $E(\varepsilon_{it} | x_{it}, z_i) = 0$.
- That is, at this stage, we are ruling out dynamic models and models where ε_{is} affects x_{it} , with $t > s$; that is, we rule out feed-back from y_{it} to future observations of the regressors.
- Although not explicitly considered, it is generally wise to include time-dummies in x_{it} .
- **Also, from now on, we include in x_{it} the observable components of z_i .**

2) Pooled regression

- How to estimate the model depends on what is assumed about whether or not (the unobservable part of) z_i is correlated with x_{it} .
- The simplest possible estimator assumes that z_i is **uncorrelated** with x_{it} .
- Under this assumption we can use **Pooled OLS**:

$$y_{it} = x'_{it}\beta + E[z'_i\phi] + \{z'_i\phi - E[z'_i\phi]\} + \varepsilon_{it}$$

$$y_{it} = x'_{it}\beta + \alpha + u_i + \varepsilon_{it}$$

where $\alpha = E[z'_i\phi]$ is fixed, and $u_i = \{z'_i\phi - E[z'_i\phi]\}$ is uncorrelated with x_{it} and ε_{it} .

- OLS estimation of y_{it} on x_{it} and a constant is unbiased and consistent for β and α .
- Also, this estimator has the advantage of requiring only contemporaneous exogeneity.
- However, the estimator is not efficient because it does not account for the serial correlation in the composite errors $\eta_{it} = u_i + \varepsilon_{it}$.

- The serial correlation also needs to be accounted for when estimating the covariance matrix of the estimated parameters.
- From now on, let $\hat{\beta}$ denote the vector of estimated parameters, including the constant.
- Also, let X denote the data matrix, also including the constant.
- An estimator that is robust to heteroskedasticity and serial correlation for $n/T \rightarrow \infty$ is given by (Arellano, 1987)

$$\widehat{\text{Var}}(\hat{\beta}) = \left(\sum_{i=1}^n X_i' X_i \right)^{-1} \left(\sum_{i=1}^n X_i' \hat{\eta}_i \hat{\eta}_i' X_i \right) \left(\sum_{i=1}^n X_i' X_i \right)^{-1}$$

where $\hat{\eta}_i = y_i - X_i \hat{\beta}$ is the vector of OLS residuals for group i .

- This is analogous to White's heteroskedasticity robust estimator, but it also accounts for arbitrary within **cluster** correlation.
- Recall that consistency of this estimator requires $n/T \rightarrow \infty$.

- We can test for the absence of unobserved heterogeneity by testing $H_0 : \sigma_u^2 = 0$.
- Breusch and Pagan (1980) obtained an LM test for this hypothesis under the assumption of normality.
- Recalling that $\hat{\eta}_{it}$ is the residual from the OLS estimation with the pooled data, the test statistic is

$$LM = \frac{nT}{2(T-1)} \left[\frac{\sum_{i=1}^n \left[\sum_{t=1}^T \hat{\eta}_{it} \right]^2}{\sum_{i=1}^n \sum_{t=1}^T \hat{\eta}_{it}^2} - 1 \right]^2 \stackrel{a}{\sim} \chi_{(1)}^2$$

- This is essentially a test for serial correlation.
- An alternative statistic that does not need the normality assumption is presented in Wooldridge (2002, p. 265).
- Serial correlation can also be tested by testing the significance of the added variable $\hat{\eta}_{it-1}$.
- Anyway, these tests are not so relevant because we expect to reject the null in most cases.

3) Random effects

- As before, we are interested in the model

$$y_{it} = x'_{it}\beta + \alpha + \eta_{it}, \quad \eta_{it} = u_i + \varepsilon_{it}.$$

- The efficient estimator in this “error components model” is the GLS. Assuming

$$E[\eta_{it}^2 | X] = \sigma_\varepsilon^2 + \sigma_u^2$$

$$E[\eta_{it}\eta_{is} | X] = \sigma_u^2, \quad t \neq s$$

$$E[\eta_{it}\eta_{js} | X] = 0, \quad i \neq j$$

for the T observations of unit i we have

$$\Sigma_{(T \times T)} = E[\eta_i \eta_i' | X] = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \dots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \dots & \sigma_u^2 \\ & & \vdots & \\ \sigma_u^2 & \sigma_u^2 & \dots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} = \sigma_\varepsilon^2 I_T + \sigma_u^2 i_T i_T'.$$

- Thanks to the independence between groups i and j , the covariance matrix for η is the nT block-diagonal matrix

$$\underset{(nT \times nT)}{\Omega} = \underset{(n \times n)}{I_n} \otimes \underset{(T \times T)}{\Sigma} = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & \Sigma & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & \Sigma \end{bmatrix}.$$

- Then, the Random Effects (GLS) estimator is just

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

where, as before, $\hat{\beta}$ denotes the vector of estimated parameters, including the constant, and X denotes the data matrix, also including the constant.

- It is interesting to notice that $\Omega^{-1} = \Omega^{-\frac{1}{2}} \Omega^{-\frac{1}{2}}$, with $\Omega^{-\frac{1}{2}} = I_n \otimes \Sigma^{-\frac{1}{2}}$ and

$$\Sigma^{-\frac{1}{2}} = \frac{1}{\sigma_\varepsilon} \left[I_T - \frac{\theta}{T} i_T i_T' \right], \quad \text{with:} \quad \theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T \sigma_u^2}}$$

- Therefore, the GLS estimator is equivalent to OLS of the transformed regressand $\Sigma^{-\frac{1}{2}}y_i$ on the transformed regressors $\Sigma^{-\frac{1}{2}}X_i$, where

$$\Sigma^{-\frac{1}{2}}A_i = \frac{1}{\sigma_\varepsilon} \begin{bmatrix} a_{i1} - \theta\bar{a}_i \\ a_{i2} - \theta\bar{a}_i \\ \vdots \\ a_{iT} - \theta\bar{a}_i \end{bmatrix}$$

where we use the group means $\bar{a}_i = T^{-1} \sum_{t=1}^T a_{it}$.

- Notice that this has as special case the pooled OLS for $\theta = 0$, which happens if $\sigma_u^2 = 0$.
- Another special case will be studied subsequently.
- This expression is also interesting because shows clearly why GLS needs **strict exogeneity**.
- From the regression of $\hat{\Sigma}^{-\frac{1}{2}}y_i$ on $\hat{\Sigma}^{-\frac{1}{2}}X_i$, standard (**robust**) inference can be performed.

- Since σ_ε^2 and σ_u^2 are unknown, GLS is **not feasible**.
- We need to consistently estimate σ_ε^2 and σ_u^2 to be able to perform FGLS.
- There are a number of ways to estimate σ_ε^2 and σ_u^2 .
- From the pooled regression we have $\text{plim } s_{Pooled}^2 = \sigma_\varepsilon^2 + \sigma_u^2$.
- We can get further information from the so-called **Between-Groups estimator**.
- The BG estimator is just OLS on

$$\bar{y}_i = \bar{x}_i' \beta + (u_i + \bar{\varepsilon}_i).$$

- From here we can get s_{BG}^2 , with $\text{plim } s_{BG}^2 = \sigma_\varepsilon^2/T + \sigma_u^2$.
- We can now solve the two equations for $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_u^2$.
- Iterated FGLS, with appropriate estimators of σ_ε^2 and σ_u^2 , is ML under normality.

4) Fixed effects

- So far, we have assumed that z_i is uncorrelated with x_{it} . We will now relax this assumption.
- Recall that the basic model is

$$y_{it} = x'_{it}\beta + z'_i\phi + \varepsilon_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}.$$

- β and $\alpha_i, i = 1, \dots, n$ can be estimated directly, including a dummy for each group.
- A **major drawback** of this approach is that it does not permit the identification of the parameters associated with time-invariant observable regressors.
- Moreover, this leads to the so-called **incidental parameters** problem because the number of parameters to be estimated increases with the sample size.
- Consistency requires $T \rightarrow \infty$, but β can be consistently estimated just with $n \rightarrow \infty$.
- Indeed, in the particular case of linear models, we can eliminate the individual effects from the model using a simple transformation.
- Similar transformations are available for some non-linear models, but these are very special cases.

- Using the same notation as before, write $\bar{y}_i = \bar{x}_i' \beta + \alpha_i + \bar{\varepsilon}_i$.
- Then, a model using the deviations from means eliminates α_i (and the intercept):

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)' \beta + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

- The **Fixed Effects** (a.k.a. **LSDV**, or **within-groups**) estimator of β is just OLS of $(y_{it} - \bar{y}_i)$ on $(x_{it} - \bar{x}_i)$.
- To perform inference, we should remember that the errors of the model are likely to be heteroskedastic and that they may also have some serial correlation.
- Again, asymptotically valid inference can be based on the results of Arellano (1987):

$$\widehat{\text{Var}}(\hat{\beta}_{FE}) = (\ddot{X}' \ddot{X})^{-1} \left(\sum_{i=1}^n \ddot{X}_i' \hat{e}_i \hat{e}_i' \ddot{X}_i \right) (\ddot{X}' \ddot{X})^{-1}$$

where \ddot{X} denotes the data in group means deviation form and $\hat{e}_i = \ddot{y}_i - \ddot{X}_i' \hat{\beta}_{FE}$ are the FE residuals for group i .

- The FE estimator is **inefficient** because it ignores the between-groups variation.
- The FE estimator is also a special case of GLS when $\theta = 1$.
- Therefore, the FE estimator also depends on the strict exogeneity assumption (**which can be tested**).
- The hypothesis that all the α_i s are equal (absence of unobserved heterogeneity) can be tested using the standard F test

$$F(n - 1, nT - n - k) = \left[\frac{nT - n - k}{n - 1} \right] \frac{RSS_{POLLS} - RSS_{FE}}{RSS_{FE}}$$

- Orme and Yamagata (2006) show that this test is closely related to the Breusch and Pagan test for random-effects, and that it is robust to non-normality.
- However, the validity of the test depends on the assumption of homoskedasticity.
- Heteroskedasticity robust versions of the test are cumbersome.

5) Hausman test

- If z_i is **uncorrelated** with x_{it} , the RE estimator is consistent and efficient, while the FE estimator is consistent but inefficient.
- If z_i is **correlated** with x_{it} , the RE estimator is inconsistent but the FE estimator is still consistent.
- Under the maintained assumption of strict exogeneity, required both by the RE and FE estimators, these results can be used to test whether or not z_i is correlated with x_{it} .
- The test can be performed using an approach popularized by Hausman (1978).
- The test checks whether the difference $\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)$ is significantly different from zero.
- To proceed, we need a consistent estimate of the asymptotic covariance matrix of $\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)$.
- The Hausman test uses the following key result:

The covariance of an efficient estimator with its difference from an inefficient estimator is zero.

- Under the null that z_i is **uncorrelated** with x_{it} , this implies that

$$\text{Cov} \left[\hat{\beta}_{RE}, \left(\hat{\beta}_{RE} - \hat{\beta}_{FE} \right) \right] = \text{Var} \left(\hat{\beta}_{RE} \right) - \text{Cov} \left(\hat{\beta}_{RE}, \hat{\beta}_{FE} \right) = 0,$$

and hence

$$\text{Cov} \left(\hat{\beta}_{RE}, \hat{\beta}_{FE} \right) = \text{Var} \left(\hat{\beta}_{RE} \right).$$

- Therefore

$$\text{Var} \left(\hat{\beta}_{RE} - \hat{\beta}_{FE} \right) = \text{Var} \left(\hat{\beta}_{FE} \right) - \text{Var} \left(\hat{\beta}_{RE} \right).$$

- This leads to the statistic

$$H = \left(\hat{\beta}_{RE} - \hat{\beta}_{FE} \right)' \left[\widehat{\text{Var}} \left(\hat{\beta}_{FE} \right) - \widehat{\text{Var}} \left(\hat{\beta}_{RE} \right) \right]^{-1} \left(\hat{\beta}_{RE} - \hat{\beta}_{FE} \right) \stackrel{a}{\sim} \chi_{(k)}^2$$

- The test can be made robust to heteroskedasticity and serial-correlation (Wooldridge, 2002, pp. 290-291).

- We have seen that the RE estimator can be obtained by regressing \check{y}_{it} on \check{x}_{it} , where \check{a}_i denotes the quasi-differenced variables $a_{it} - \theta\bar{a}_i$.
- The Hausman test is equivalent to a test for $H_0 : \xi = 0$ in the regression

$$\check{y}_{it} = \check{x}_{it}\beta + \check{\check{x}}_{it}\xi + \zeta_{it}$$

where $\check{\check{x}}_{it} = x_{it} - \bar{x}_i$.

- To make the test robust to heteroskedasticity and serial-correlation just use a "clustered" estimator of the covariance matrix.
- Because of the strict-exogeneity assumption, rejection of the null **does not imply** that the FE estimator is valid. Indeed, the test may reject just because the strict exogeneity assumption is invalid.
- This can be tested by estimating by FE a model including leads of the regressors (see Wooldridge, 2002, p. 285).
- Nobody does this, but it is quite important to consider the result of the strict exogeneity test when interpreting the outcome of the Hausman test.

6) Dynamic models

- One of the main benefits of panel data is that we can use it to model dynamics, while allowing for individual heterogeneity.
- This requires that we use a dynamic panel data model.
- The simplest dynamic model is the first-order auto-regression

$$y_{it} = \gamma y_{it-1} + x'_{it}\beta + \alpha_i + \varepsilon_{it} = w'_{it}\delta + \alpha_i + \varepsilon_{it}.$$

- None of the standard estimators is consistent with $n \rightarrow \infty$:
 - Both Pooled OLS and RE are inconsistent because y_{it-1} is correlated with α_i ;
 - BG is inconsistent because \bar{w}_i is correlated with α_i and with $\bar{\varepsilon}_i$;
 - In the FE, we regress $(y_{it} - \bar{y}_i)$ on $(w_{it} - \bar{w}_i)$ but $(w_{it} - \bar{w}_i)$ is correlated with $(\varepsilon_{it} - \bar{\varepsilon}_i)$ because of the inclusion of y_{it-1} in w_{it} ;
 - However, FE is consistent with $T \rightarrow \infty$.

- A simple consistent IV approach is as follows.
- First-differencing eliminates the individual time-invariant effects and gives

$$y_{it} - y_{it-1} = \gamma(y_{it-1} - y_{it-2}) + (x_{it} - x_{it-1})'\beta + (\varepsilon_{it} - \varepsilon_{it-1}).$$

- OLS fails here because $(y_{it-1} - y_{it-2})$ is correlated with $(\varepsilon_{it} - \varepsilon_{it-1})$.
- IV using suitably lagged levels (y_{it-2}) or differences $(y_{it-2} - y_{it-3})$ as instruments for $(y_{it-1} - y_{it-2})$ is consistent (Anderson and Hsiao, 1982).
- There is a potential **weak-instruments** problem here, especially with lagged differences being used as instruments.
- Also, validity of the estimator depends on ε_{it} having no serial-correlation.
- Stata provides a serial-correlation test for $(\varepsilon_{it} - \varepsilon_{it-1})$: we expect to find first order, but not higher-order serial-correlation.

- Efficiency can be improved by using more identifying information in a GMM procedure.
- The popular Arellano-Bond (1991) estimator uses GMM with all possible lags.
- One should be careful not to use too many instruments because this can affect the performance of the estimator in finite samples.
- There are a number of alternative GMM estimators for dynamic panel data models (Ahn and Schmidt, 1995, Blundell and Bond, 1998), but we will not discuss them here.
- All these estimators can be generalized to include other endogenous regressors.

University week: 4

Simultaneous-Equations Models

1. Fundamental issues in simultaneous-equations models;
2. Limited information methods;
3. System estimation methods;
4. Specification tests;
5. Weak Instruments.

Reading: Greene: 8.1-8.4, 8.7, 10.6.

1) Fundamental issues in simultaneous-equations models

- Simultaneous-equations models are one of the main contributions of econometrics to the general statistics literature.
- Simultaneous-equations models are important in economics because economic theory often deals with systems of equations (e.g., supply and demand determine price and quantity).
- Consider an example with two **behavioural equations** and one **equilibrium condition**:

$$q_{d,t} = \alpha_1 p_t + \alpha_2 x_t + \varepsilon_{d,t}$$

$$q_{s,t} = \beta_1 p_t + \varepsilon_{s,t}$$

$$q_{d,t} = q_{s,t} = q_t$$

- Notice that economic theory is mute about the properties of the error terms.
- Even if we are interested in one equation (e.g., demand), it is important to account for the interactions between the variables.

- It is helpful to introduce some terminology:
 - The equations derived from economic theory and describing any particular feature of the economy are called **structural equations**;
 - Variables determined within the system are the **endogenous** variables (q_t and p_t);
 - Variables determined outside the system are the **exogenous** variables (x_t);
 - Lagged values of the endogenous variables are termed **predetermined variables**;
 - If we assume $E(\varepsilon_t|x_t) = 0$, we have **contemporaneous exogeneity**;
 - If we assume $E(\varepsilon_t|x_1, \dots, x_T) = 0$, we have **strict exogeneity**;
 - The system is **complete** if the numbers of equations and endogenous variables are equal.

- Suppose that we want to estimate the supply equation and write the system as

$$q_t = \alpha_1 p_t + \alpha_2 x_t + \varepsilon_{d,t} \quad (\text{D})$$

$$p_t = \frac{1}{\beta_1} q_t - \frac{1}{\beta_1} \varepsilon_{s,t} \quad (\text{S})$$

- We have **normalized** to 1 the coefficient of p_t in (S).
- Estimation of (S) by OLS is inconsistent due to **two** sources of endogeneity.
 - **Simultaneity** – For $\alpha_1 \neq 0$, q_t is a function of p_t and therefore it is correlated with $\varepsilon_{s,t}$;
 - **Correlation** – If $\varepsilon_{d,t}$ and $\varepsilon_{s,t}$ are correlated, q_t is correlated with $\varepsilon_{s,t}$ even if $\alpha_1 = 0$.
- The fact that the structural equations cannot be estimated by OLS shows that they are not **conditional expectations**.
- Notice that the presence of simultaneity does not imply that the equation is structural.

- Writing the endogenous variables as functions of the exogenous and predetermined variables leads to the **reduced form** of the model

$$p_t = \frac{\alpha_2}{\beta_1 - \alpha_1} x_t + \frac{\varepsilon_{d,t} - \varepsilon_{s,t}}{\beta_1 - \alpha_1} = \pi_1 x_t + \nu_{t1}$$

$$q_t = \frac{\beta_1 \alpha_2}{\beta_1 - \alpha_1} x_t + \frac{\beta_1 \varepsilon_{d,t} - \alpha_1 \varepsilon_{s,t}}{\beta_1 - \alpha_1} = \pi_2 x_t + \nu_{t2}$$

- Under standard assumptions, x is uncorrelated with the error terms and the reduced form can be interpreted as a system of **conditional expectations**.
- Notice that the structural form **imposes restrictions** on the parameters of the reduced form.
- In this particular example, it is clear that the structural form parameters $(\alpha_1, \alpha_2, \beta_1)$ cannot be obtained from π_1 and π_2 .
- This is the so-called **identification problem**: How to match the parameters of the economic model with parameters of a conditional distribution that can be consistently estimated?
- More on the identification problem later...

- Consider now a general linear system of equations with M equations, M endogenous variables and k exogenous variables (including the intercept).
- The structural form for one observation in period t is

$$y_t' \Gamma + x_t' B = \varepsilon_t', \quad t = 1, \dots, T,$$

where $y_t = [y_{t1}, \dots, y_{tM}]'$, $x_t = [x_{t1}, \dots, x_{tk}]'$, $\varepsilon_t = [\varepsilon_{t1}, \dots, \varepsilon_{tM}]'$, and

$$\Gamma = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1M} \\ \vdots & \ddots & \vdots \\ \gamma_{M1} & \cdots & \gamma_{MM} \end{bmatrix}, \quad B = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1M} \\ \vdots & \ddots & \vdots \\ \beta_{k1} & \cdots & \beta_{kM} \end{bmatrix}.$$

- Assume that Γ^{-1} exists, that $E(\varepsilon_t | x_t) = 0$, $E(\varepsilon_t \varepsilon_t' | x_t) = \Sigma$ and $E(\varepsilon_t \varepsilon_s' | x_t, x_s) = 0, \forall t \neq s$.
- Hence, the reduced form is

$$y_t' = -x_t' B \Gamma^{-1} + \varepsilon_t' \Gamma^{-1} = x_t' \Pi + v_t', \quad t = 1, \dots, T,$$

with $E(v_t | x_t) = 0$, $E(v_t v_t' | x_t) = \Omega = (\Gamma^{-1})' \Sigma \Gamma^{-1}$ and $E(v_t v_s' | x_t, x_s) = 0, \forall t \neq s$.

- So far we have arranged the data observation-by-observation.
- We can also stack all the data into a matrix structural form equation

$$Y\Gamma + XB = U$$

where $Y = [y_1, \dots, y_T]'$, $X = [x_1, \dots, x_T]'$ and $U = [\varepsilon_1, \dots, \varepsilon_T]'$, which has the corresponding matrix reduced form

$$Y = X\Pi + V$$

- Alternatively, we can arrange the data equation-by-equation

$$y_{(j)} = X\pi_{(j)} + v_{(j)}, \quad j = 1, \dots, M,$$

where $y_{(j)} = [y_{1j}, \dots, y_{Tj}]'$ is the j -th column of Y , $\pi_{(j)}$ is the j -th column of Π , and $v_{(j)} = [v_{1j}, \dots, v_{Tj}]'$ is the j -th column of V .

- This gives a **SUR-type** arrangement with the same regressors in every equation.
- Hence, if there are no cross-equation restrictions on Π (or Ω), then GLS estimation of the reduced form is the same as OLS.
- However, in the absence of restrictions on (Γ, B, Σ) the structural form is **unidentified**.
- To see this, it is enough to notice that Π has $k \times M$ parameters, whereas Γ and B have $M \times M + k \times M$ parameters (Σ and Ω are both $M \times M$).
- Alternatively, notice that for any non-singular $(M \times M)$ matrix F we have that $y_t' \Gamma F + x_t' B F = \varepsilon_t' F$ leads to the reduced form

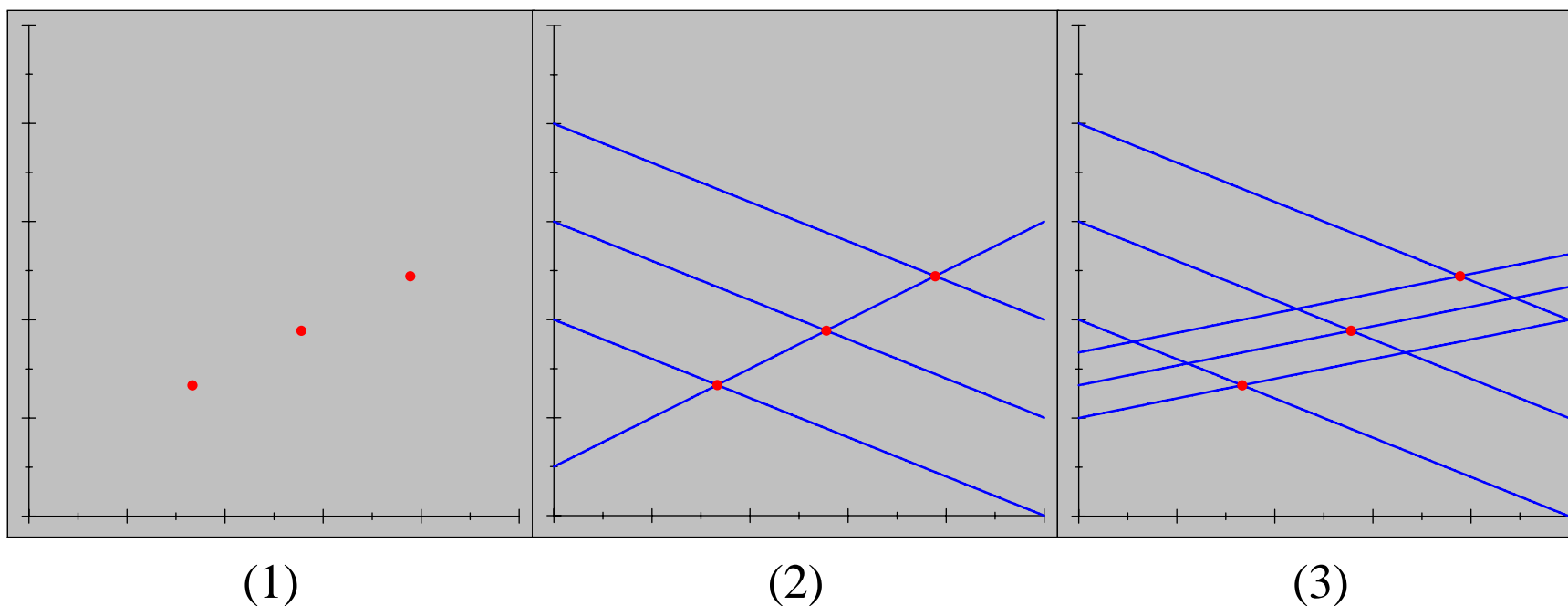
$$y_t' \Gamma F = -x_t' B F + \varepsilon_t' F$$

$$y_t' \Gamma F (F^{-1} \Gamma^{-1}) = -x_t' B F (F^{-1} \Gamma^{-1}) + \varepsilon_t' F (F^{-1} \Gamma^{-1})$$

$$y_t' = -x_t' B \Gamma^{-1} + \varepsilon_t' \Gamma^{-1}$$

- Hence, $y_t' \Gamma F + x_t' B F = \varepsilon_t' F$ is **observationally equivalent** to $y_t' \Gamma + x_t' B = \varepsilon_t'$.

- The identification problem has an interesting graphical illustration due to Working (1926).
- Suppose that observations of the pair (y, x) are collected.
- Without further information, it is not possible to know which structure originated the observed data.
- For us to be able to identify one equation, we must obtain observations when that equation is fixed and the other is moved by some exogenous reason, as in (2).



- This graphical analysis suggests that for an equation to be identified it must not depend on all exogenous variables so that changes in the excluded variables displace the other equations, without affecting the equation of interest.
- These **exclusion restrictions** on the structural form are the most common way of obtaining identification.
- As the graphical illustration shows, it is possible that some equations are identified, whereas the others are not.
- Therefore, the identification of a system has to be performed **equation-by-equation**.
- Given that a structural equation cannot be estimated by OLS due to the endogeneity of some variables, identification of the parameters of interest (the ability to estimate them consistently) depends on the existence of **enough instruments** for the endogenous variables in the equation.

- It is now time to have a second look at our supply-and-demand example. Recall that

$$q_t = \alpha_1 p_t + \alpha_2 x_t + \varepsilon_{d,t} \quad q_t = \pi_2 x_t + \nu_{t2} \quad (\text{D})$$

$$p_t = \gamma_1 q_t - \gamma_1 \varepsilon_{s,t} \quad p_t = \pi_1 x_t + \nu_{t1} \quad (\text{S})$$

- It is clear that the supply equation can be estimated by IV, using x_t as an instrument for q_t .
- However, the parameters of the demand equation cannot be identified because there is no valid instrument for p_t in that equation.
- In the case of an equation identified by a normalization restriction and a collection of exclusion restrictions, the necessary order condition for identification of that equation is that:

“The number of exogenous variables excluded from the equation is greater than, or equal to, the number of endogenous explanatory variables included in that equation.”

- A sufficient condition for identification is the **rank condition**.
- Normalize the j -th structural equation by setting $\gamma_{jj} = 1$ and rewrite it as

$$y_{(j)} = Z_{(j)}\delta_{(j)} + \varepsilon_{(j)} = Y_{(j)}\gamma_{(j)} + X_{(j)}\beta_{(j)} + \varepsilon_{(j)}.$$

where $Y_{(j)}$ is the matrix consisting of the non-excluded and non-normalized endogenous variables, and $X_{(j)}$ contains the non-excluded exogenous variables in the j -th equation.

- The reduced form that generates $Y_{(j)}$ is

$$Y_{(j)} = X_{(j)}\Pi_1 + X_{(-j)}\Pi_2 + V_{(j)}.$$

where $X_{(j)}$ is the matrix of non-excluded exogenous variables and $X_{(-j)}$ is the matrix of excluded exogenous variables in the j -th equation.

- The **order condition** requires the number of columns of Π_2 to be at least equal to the number of included endogenous variables.
- The **rank condition** states that if Π_2 has rank equal to the number of included endogenous variables, then the equation is identified.

- Explicit verification of the rank condition impractical.
- Notice that what is relevant is that the rank condition holds **asymptotically**.
- Note also that **this rank condition is not a necessary condition** for identification.
- Besides the normalization and exclusion restrictions, identification can be achieved by imposing other (potentially cross-equation) linear restrictions and restrictions on the covariance matrix of the disturbances.
- In short, what is required for identification is that the **number** of valid and relevant instruments that are available and are **linearly independent** (including those used as explanatory variables) is at **least equal** to the total number of explanatory variables in the equation.
- If the equality holds, the equation is said to be **just-identified**; otherwise it is **over-identified**.
- More on the validity and relevance of the instruments later.

2) Limited information methods

- The estimation of the system can be performed equation-by-equation (**limited information methods**) or using system (**full-information**) methods.
- In general, full information methods may be **more efficient** but can be **less robust**.
- As noted before, **OLS** generally suffers from “simultaneity bias”.
- However, there are situations where it is consistent.
- Consider the following **triangular system** (which does not satisfy the rank condition)

$$y_1 = x' \beta_1 + \varepsilon_1$$

$$y_2 = x' \beta_2 + \gamma y_1 + \varepsilon_2$$

- The first equation can be consistently estimated by OLS under very mild conditions.
- If Σ is diagonal, both equations can be consistently estimated by OLS.
- A triangular system with a diagonal covariance matrix is called a **recursive system**.

- OLS can also obviously be used to estimate the reduced form, giving $\widehat{\Pi}_{OLS}$.
- Under **strict exogeneity**, $\widehat{\Pi}_{OLS}$ is unbiased for Π .
- If we assume that $T^{-1}X'X$ converges to a finite non-singular matrix A , then $\widehat{\Pi}_{OLS}$ is consistent under just **contemporaneous exogeneity**.
- Finally, if an appropriate central limit theorem is assumed to hold

$$T^{1/2} \text{vec}(\widehat{\Pi}_{OLS} - \Pi) \xrightarrow{d} \mathcal{N}(0, \Sigma \otimes A^{-1}).$$

N.B.: $\text{vec}(M)$ is the column vector obtained by stacking the columns of the matrix M .

- In general, estimation of the **structural parameters** requires the use of IV methods.

- Let $W_{(j)}$ be a $T \times (M_j + k_j)$ matrix of instruments and write the j -th structural equation as

$$y_{(j)} = Y_{(j)}\gamma_{(j)} + X_{(j)}\beta_{(j)} + \varepsilon_{(j)}$$

$$y_{(j)} = Z_{(j)}\delta_{(j)} + \varepsilon_{(j)}$$

where it is assumed that

$$\text{plim} \frac{1}{T} Z'_{(j)} \varepsilon_{(j)} \neq 0 \quad \text{plim} \frac{1}{T} W'_{(j)} \varepsilon_{(j)} = 0$$

$$\text{plim} \frac{1}{T} W'_{(j)} Z_{(j)} = \Sigma_{WZ} \quad \text{plim} \frac{1}{T} W'_{(j)} W_{(j)} = \Sigma_{WW}$$

and Σ_{WZ} and Σ_{WW} are finite and Σ_{WW} is non-singular.

- Essentially, $W_{(j)}$ needs to satisfy the following conditions: **Validity** – it is (conditionally) **uncorrelated** with $\varepsilon_{(j)}$; **Relevance** – it is (conditionally) **correlated** with $Z_{(j)}$; **Order condition** – it has at least the dimension of $Z_{(j)}$.

- The **two-stage least squares** (2SLS) is the IV estimator when $W_{(j)} = \hat{Z}_{(j)} = [\hat{Y}_{(j)} \ X_{(j)}]$, where $\hat{Y}_{(j)}$ denotes the fitted values of $Y_{(j)}$ obtained from the reduced form (first stage).

$$\hat{\delta}_{(j),2SLS} = \left(W'_{(j)} Z_{(j)} \right)^{-1} W'_{(j)} y_{(j)} = \left(\hat{Z}'_{(j)} \hat{Z}_{(j)} \right)^{-1} \hat{Z}'_{(j)} y_{(j)}.$$

- Due to the estimation of Π , 2SLS is **biased** in finite samples.
- To see this, notice that the reduced form decomposes $Y_{(j)}$ into a part that is necessarily exogenous ($X\Pi$), and another which is potentially endogenous ($V_{(j)}$).
- Naturally, in finite samples, $X\hat{\Pi}$ is not identical to $X\Pi$ and therefore is not exogenous, leading to bias (more on this later).
- However, under our assumptions, $\hat{\delta}_{(j),2SLS}$ is **consistent** for identified equations.
- Under standard assumptions on the error terms (including homoskedasticity),

$$T^{1/2}(\hat{\delta}_{(j),2SLS} - \delta_{(j)}) \xrightarrow{d} \mathcal{N}(0, \sigma_{jj} \Sigma_{WW}^{-1}).$$

- σ_{jj} can be consistently estimated by (**not** from the residuals of the 2nd stage regression)

$$\hat{\sigma}_{jj} = T^{-1}(y_{(j)} - Z_{(j)}\hat{\delta}_{(j),2SLS})'(y_{(j)} - Z_{(j)}\hat{\delta}_{(j),2SLS}).$$

- Under the additional assumption that the error-terms have a normal distribution, the limited information maximum likelihood estimator (**LIML**) can be used.
- LIML is efficient among the limited information estimators.
- However, LIML is **asymptotically equivalent** to 2SLS.
- Therefore, under normality, 2SLS is fully efficient.
- The main advantage of LIML is its **invariance** to reparameterization.
- It also tends to outperform 2SLS in cases where sample sizes are small and the number of overidentifying restrictions is large (more on this later).
- Finally, an important difference between LIML and 2SLS is that the finite sample distribution of LIML **has no finite moments** (but is **median unbiased**), where the finite sample distribution of 2SLS **may have some finite moments**.

3) System estimation methods

- Equation-by-equation is in general inefficient, unless the structural form is just identified.
- Three-Stage Least Squares (**3SLS**) is an efficient structural form estimator.
- Here we write the system as

$$\begin{bmatrix} y_{(1)} \\ \vdots \\ y_{(M)} \end{bmatrix} = \begin{bmatrix} Z_{(1)} & \dots & 0 \\ \vdots & \ddots & 0 \\ 0 & \dots & Z_{(M)} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_M \end{bmatrix} + \begin{bmatrix} \varepsilon_{(1)} \\ \vdots \\ \varepsilon_{(M)} \end{bmatrix},$$

i.e., as $y = \bar{Z}\delta + \varepsilon$, where $E(\varepsilon|X) = 0$, $E(\varepsilon\varepsilon'|X) = (\Sigma \otimes I_T)$.

- OLS regression of y on \bar{Z} is inconsistent.
- IV regression of y on \bar{Z} using instruments \bar{W} is a way to obtain consistent estimates.
- 2SLS uses $\bar{W} \equiv \widehat{W} = \text{diag}\{\widehat{Z}_{(j)}\}$.

- 3SLS combines IV with GLS, leading to

$$\hat{\delta}_{3SLS} = \left[\widehat{W}' (\widehat{\Sigma}^{-1} \otimes I_T) \widehat{W} \right]^{-1} \widehat{W}' (\widehat{\Sigma}^{-1} \otimes I_T) y.$$

- The result is consistent and asymptotically efficient under standard assumptions.
- Like the SUR, 3SLS can be **iterated**, but this does **not** lead to a ML estimator.
- The **drawback** of the 3SLS, as compared to the 2SLS, is that if any of the equations is misspecified, 3SLS this will typically invalidate all the coefficient estimates, whereas 2SLS will only invalidate the coefficient estimates in the misspecified equation.
- Again, if we assume jointly normality of the errors, we can use maximum likelihood.
- Full Information Maximum Likelihood (**FIML**) uses all the identifying information for the system.
- **Under normality**, FIML is asymptotically equivalent to 3SLS, but it is not robust to non-normality.

4) Specification testing

- Researchers are often interested in testing possible **over-identifying** restrictions (believing to be checking the **validity** of the instruments).
- At the system level, this can be done by estimating the unrestricted reduced form and testing the reduced form restrictions implied by the identification of the system.
- At the single equation level, over-identifying restrictions can be tested using **Sargan's** test.
- Under **homoskedasticity**, Sargan's (1964) test is performed by comparing the test statistic

$$S = T \frac{\hat{\varepsilon}'_{(j)} X (X'X)^{-1} X' \hat{\varepsilon}_{(j)}}{\hat{\varepsilon}'_{(j)} \hat{\varepsilon}_{(j)}},$$

where $\hat{\varepsilon}_{(j)}$ are the 2SLS or LIML residuals for equation j , with critical values from a $\chi^2_{(k-k_j-M_j)}$ distribution (d.f. equal to number of overidentifying restrictions).

- Notice that $S = TR^2$, where R^2 is the the uncentered R^2 in the regression of $\hat{\varepsilon}_{(j)}$ on X .
- The test can be made robust to **heteroskedasticity** (Wooldridge, p. 123).
- The importance of this test is often over-stated...

5) Weak Instruments

- One problem which can arise in practical settings is that of having **weak instruments**.
- Suppose we estimate the equation $y_{(j)} = Z_{(j)}\delta_{(j)} + \varepsilon_{(j)}$ using the following IV estimator

$$\hat{\delta}_{(j),IV} = \left(W'_{(j)}Z_{(j)}\right)^{-1} W'_{(j)}y_{(j)} = \delta_{(j)} + \left(\frac{1}{T}W'_{(j)}Z_{(j)}\right)^{-1} \frac{1}{T}W'_{(j)}\varepsilon_{(j)}.$$

- The weak instrument problem arises when $T^{-1}W'_{(j)}Z_{(j)}$ is not “large” enough:
 - estimators have non-normal distributions and standard inference is invalid;
 - finite sample correlation between $W_{(j)}$ and $\varepsilon_{(j)}$ is amplified, leading to significant bias.
- In the case of the 2SLS estimator, it can be shown that the bias varies inversely with the F -statistic for $H_0 : \Pi_2 = 0$ in the first stage (relevant instruments).
- As the instruments become **weak**, the F -statistic becomes smaller and 2SLS approaches the OLS.
- As a rule-of-thumb, when there is **only one endogenous regressor**, this F -statistic should be at least equal to 10.

- Notice also that the bias increases with the number of instruments.
- To see this, recall that the F -statistic can be written as

$$F = \left[\frac{T - k}{k - k_j} \right] \frac{R_u^2 - R_r^2}{1 - R_u^2}$$

where R_u^2 and R_r^2 are the R-squares for the unrestricted and restricted first-stage models, T is the number of observations, k is the number columns in Π , and $k - k_j$ is the number columns in Π_2 .

- Adding weak instruments will have little impact on R_u^2 but will increase k , thereby tending to reduce the value of the statistic.
- In general, attention should be paid to the first stage regression, which should produce “reasonable” results.
- It is a good idea to look at the results for a just-identified case because in that case 2SLS is median-unbiased.

University week: 5

Cointegrated Processes

1. Introduction;
2. Nonstationary processes and unit roots;
3. Cointegration;
4. Determining cointegrating rank;
5. Estimation;

Recommended reading: Greene: 20.4, 22.1–22.5.

1) Introduction

- Estimation of econometric models using time-series raises a set of (difficult) problems.
- A first issue we have to deal with is the nature of randomness in this context.
- With cross-sectional or panel data, it is clear that different samples from the same population will lead to different results and it is clear that functions of the sample, like the sample mean, are random variables.
- Time series are also random variables: we do not know next year's GDP.
- However, we cannot obtain different samples for GDP over the last 10 years.
- What we observe is just one realization of a sequence of random variables indexed by time.
- A sequence of random variables indexed by time is called a **stochastic process**.
- The population of interest is the set of all possible realizations of a stochastic process.

- There are two additional sources of problems:
 - (a) **The natural ordering of the data:** Because each observation comes from a different period, there is no guarantee that they all have the same distribution, i.e., that the data may not be stationary.
 - (b) **The sampling protocol that has to be used:** Because the past influences the future, in general it is not possible to assume that the sampled observations are independent.
- The sampling method that is used also implies that time series data have other important characteristics: frequency and span are as important as the sample size.
- Properties of the estimators and the estimation strategy to adopt depend on the statistical properties of the time-series (stationarity and independence).
- Therefore, it is important to study the characteristics of the available data before starting the econometric modelling.

Definition: A stochastic process $\{y_t : t = 1, 2, \dots\}$ is (strongly) **stationary** if, for any set of time indexes $1 \leq t_1 < t_2 \dots < t_m$ and for any integer $h \geq 1$, the joint distribution of $(y_{t_1}, \dots, y_{t_m})$ is equal to the joint distribution of $(y_{t_1+h}, \dots, y_{t_m+h})$.

- Often (strong) stationarity is not required for valid inference and it is enough to assume weak or covariance stationarity.

Definition: A stochastic process $\{y_t : t = 1, 2, \dots\}$ is **covariance stationary** if

1. $E(y_t)$ is constant;
 2. $\text{Var}(y_t)$ positive, finite and constant;
 3. $\text{Cov}(y_t, y_{t+h})$ depends only on h , for any t and $h \geq 1$.
- If a stochastic process is (strongly) stationary and has finite second moments, then it must be covariance stationary. The converse is not true.
 - Covariance stationarity is critical for the asymptotic properties of OLS (and other standard estimators) because it allows us to treat a realization of a stochastic process as a sample from the same distribution.

- Covariance stationarity is not enough to ensure meaningful estimation.
- Some form of **weak dependence** is also needed for a Law of Large Numbers and a Central Limit Theorem to be applied.
- In particular, we will assume **ergodicity**.

Definition: A stochastic process $\{y_t : t = 1, 2, \dots\}$ is said to be **ergodic** if, for any two bounded functions $f : \mathbb{R}^k \rightarrow \mathbb{R}$ and $g : \mathbb{R}^\ell \rightarrow \mathbb{R}$,

$$\lim_{T \rightarrow \infty} |\mathbb{E}[f(y_{t_1}, \dots, y_{t_k}) g(y_{t_1+T}, \dots, y_{t_\ell+T})]| = |\mathbb{E}[f(y_{t_1}, \dots, y_{t_k})]| |\mathbb{E}[g(y_{t_1+T}, \dots, y_{t_\ell+T})]|$$

- Heuristically, this implies that random variables sufficiently far apart in time will be almost independently distributed.

Ergodic theorem: Let $\{y_t : t = 1, 2, \dots\}$ be a stationary and ergodic process with $\mathbb{E}[y_t] = \mu$. Then, as $T \rightarrow \infty$,

$$\frac{1}{T} \sum_{t=1}^T y_t \xrightarrow{a.s.} \mu.$$

2) Nonstationary processes and unit roots

- If the time-series are stationary and ergodic, inference can be carried out much in the same way as when cross-sectional data is used.
- However, many time series are clearly non-stationary.
- An important example of a non-stationary process is the **random-walk with drift**

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

where ε_t is a white noise (or at least a zero-mean stationary ergodic process).

- By recursive substitution, we can see that the process is not stationary:

$$y_t = y_0 + t\mu + \sum_{i=0}^{t-1} \varepsilon_{t-i}$$

- Trend-stationary models are also non-stationary:

$$y_t = \mu + \beta t + \varepsilon_t$$

- Inference using non-stationary or non-ergodic series is problematic because the standard asymptotic results do not hold.
- For example, if two independent random-walks $\{x_t\}$ and $\{y_t\}$ are generated, regressing y_t on x_t leads to intriguing results:
 - Granger and Newbold (1974) showed that a t -test for the significance of the parameter associated with x_t often leads to the rejection of the null.
 - For example, for $T = 50$ the rejection frequency for a two-sided test at 5% is 66.2%.
 - With $T = 250$, the rejection frequency goes up to 84.7%.
 - The R^2 converges to a random variable, being often very high.
- These **spurious regressions** arise because, under the null, the model does not satisfy the usual assumptions.
- Regressions using non-stationary variables are only interesting in a particular case to be studied later.

- A model with a deterministic trend is conditionally ergodic and stationary.
- The first difference of the random-walk is ergodic and stationary.

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

- A process y_t such that

$$(1 - L)^d y_t = \mu + \psi(L) \varepsilon_t$$

is stationary and $\psi(1) \neq 0$ (no unit-roots) is said to be **integrated to order d** : $y_t \sim I(d)$.

- The $I(1)$ case is especially interesting. If a process is $I(1)$, it is said to be **difference-stationary** or to be a **unit-root** process.
- The identification of integrated processes is very important from both the **economic** and **econometric** view points.

- Consider first the simple AR(1) model

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

- For $|\gamma| < 1$, c , the least squares estimator of γ is downward biased, but is consistent.
- For $\gamma = 1$, Dickey and Fuller have shown that

$$E[c] < 1, \quad T(c - \gamma) \xrightarrow{d} v$$

where v is a random variable with finite but positive variance (notice the **unusual** rate of convergence).

- Therefore, tests for the presence of unit-roots cannot be performed using standard tools.
- A standard t -ratio can be used, but it has to be referred to a non-standard (Dickey-Fuller) distribution.
- Testing the null of a unit-root in this case is not particularly interesting because it imposes a **zero mean** under the alternative.

- Consider now the following AR(1) model with an **intercept**

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

$$\Delta y_t = \alpha + \theta y_{t-1} + \varepsilon_t$$

with $\rho - 1 = \theta$ and $E(\varepsilon_t | y_{t-1}, y_{t-2}, \dots) = 0$.

- The test of the **null** that the process has a unit-root is performed by comparing

$$DF_\tau = \frac{\hat{\theta}}{s.e.(\hat{\theta})}$$

with the appropriate critical values

1%	5%	10%
-3.43	-2.86	-2.57

- Note that this is a **one-sided left-tailed** test so that one rejects the null of stationarity if the test statistic is smaller (more negative) than the relevant critical value.

- If the data shows a clear **trend**, this test is inappropriate.
- In that case, the test can be based on the model

$$y_t = \alpha + \rho y_{t-1} + \gamma t + \varepsilon_t$$

$$\Delta y_t = \alpha + \theta y_{t-1} + \gamma t + \varepsilon_t$$

- The test of the **null** that the process has a unit-root is performed by comparing

$$DF_\tau = \frac{\hat{\theta}}{s.e.(\hat{\theta})}$$

with the appropriate critical values

1%	5%	10%
-3.96	-3.41	-2.12

- In both with and without the trend, it is assumed that $E(\varepsilon_t | y_{t-1}, y_{t-2}, \dots) = 0$.

- If the auxiliary regression has serial-correlation, these Dickey-Fuller tests are not valid.
- In order for the tests to be valid, the serial-correlation has to be eliminated by including enough lags of the dependent variable ($\Delta y_{t-1}, \Delta y_{t-2}, \dots$).
- The critical values for these **Augmented Dickey-Fuller** tests are as before.
- The problem is how to decide the number of lags used to augment the regression.
- One approach (due to Ng and Perron, 1995) is to start by adding p lags and sequentially drop the last lag if it is not significant at 10%.
- At every step, it is wise to check for possible serial correlation of the errors.
- How to decide on the value of p to start with is an open question but, following Schwert (1989) it is standard to choose $p_{\max} = \left[12 \left(\frac{T}{100} \right)^{0.25} \right]$.
- Alternatively, an information criterion can be used (Akaike, 1974, Schwarz, 1978 or Hannan and Quinn, 1979).

- There are many other flavours of unit-root tests. Two leading examples are:

1. The Phillips-Perron (1988) test

- This test deals with the serial correlation by using an alternative variance estimator (rather than adding lags as in the ADF).
- Critical values are the same used for the ADF test, but we have to choose the kernel and bandwidth for the variance estimator, and the test performs poorly in finite samples.

2. The ADF-GLS of Elliott, Rothenberg and Stock (1996)

- The idea of this test is to estimate the coefficients on the deterministic regressors before doing the test, using a kind of GLS regression.
- Then the test is performed like an ADF test on the new series obtained by subtracting the deterministic part of the model from y_t .
- If a trend is included, the critical values are different (but not much) from the ADF ones.

- Rather than testing the null of non-stationarity, it is possible to test the **null of stationarity** against the alternative that the series is non-stationary.
- This is the idea behind the test proposed by Kwiatkowski, Phillips, Schmidt, & Shin (1992).
- The KPSS is a **one-sided right-tailed** test so that one rejects the null of stationarity if the test statistic is larger than the relevant critical value.
- The denominator of the test statistic is an estimate of the long-run variance of the series, computed from the empirical autocorrelation function.
- Like in the Phillips-Perron test, it is necessary to choose the kernel and the bandwidth used in the estimation of this variance.
- Unfortunately, simulation results by Caner and Kilian (2001) suggest that the KPSS test performs very poorly in finite samples.

- Unit root tests can be extended to a panel data context.
- Perhaps the easiest way to do that is to use the Fisher test due to Maddala and Wu (1999).
- To perform the test, compute the appropriate statistic (e.g., ADF) for each unit in the panel.
- Under the null, the p-value π of a test follows a uniform $(0, 1)$ distribution.
- Hence, by inverting the cdf of the $\chi_{(2)}^2$, we have that $-2 \ln(\pi) \sim \chi_{(2)}^2$.
- Consequently, if we perform n **independent** tests:

$$\lambda = -2 \sum_{i=1}^n \ln(\pi_i) \sim \chi_{(2n)}^2.$$

- This statistic can be used to test the null that **all the series have a unit root** against the alternative that at least one of them is stationary.
- The advantage of the test is that it is quite flexible, for example we can mix DF with ADF tests and the panel does not have to be balanced.
- The disadvantage is that if the tests are not independent, critical values have to be computed by simulation.

3) Cointegration

- The available evidence points towards the existence of many non-stationary macro-economic variables.
- As we have seen, estimating regressions with $I(1)$ variables is likely to lead to erroneous conclusions.
- A possible solution is to work with differenced series (this was the practice adopted by most people after the Granger and Newbold paper).
- However, models in differences are mute about the relation between the levels of the variables in a steady state.
- Moreover, economic theory suggests that there are stable relations between the levels of some of these variables.
- This is possible if some linear combination of non-stationary variable is stationary.
- That is, although the series have random trends, they drift “together”.

- Consider two time series y_{1t} and y_{2t} which are both $I(d)$:
 - In general, any linear combination of y_{1t} and y_{2t} will also be $I(d)$.
 - However, if there exists a vector $\gamma = (1, -\beta)'$, such that the linear combination

$$z_t = y_{1t} - \beta y_{2t} \sim I(d - b),$$

then, y_{1t} and y_{2t} are said to be **cointegrated** of order (d, b) , denoted

$$y_t = [y_{1t}, y_{2t}]' \sim CI(d, b),$$

with γ being the **cointegrating vector**.

- Several points are worth noticing:
 - (a) Cointegration refers to a **linear** combination of nonstationary variables;
 - (b) The cointegrating vector is **not uniquely defined**;
 - (c) Both variables must be integrated of the **same order** to be candidates to form a cointegrating relationship.
 - (d) When the number of variables is **larger than two** it is possible to have cointegration between variables with different orders of integration.
 - (e) Like most of the literature, we will focus on the $CI(1, 1)$ case, since few economic variables prove to be integrated of higher order.
 - (f) If y_{1t} and y_{2t} are $CI(1, 1)$, they must share (up to a scalar) the same stochastic trend, called a **common trend**.
 - (g) Cointegration reflects a **long-run relation** between the levels of the variables.

- Consider now a vector $y_t = [y_{1t}, \dots, y_{Mt}]' \sim CI(1, 1)$ and the **long run** relation

$$y_{1t} = Y_t' \theta + x_t' \delta + \varepsilon_t$$

where $[y_{1t}, \dots, y_{Mt}]' = [y_{1t}, Y_t']'$, ε_t is the stationary and weakly dependent equilibrium error and x may contain a constant, trends, dummies and $I(0)$ exogenous variables.

- The OLS estimator of θ and δ has **non-standard properties** due to the non-stationarity.
- Under cointegration, the OLS estimator of θ is **super-consistent**, i.e., its variance is $O(1/T^2)$ rather than $O(1/T)$.
- This holds even in the presence of **endogeneity**.
- In a system of M variables, there can be up to $M - 1$ independent cointegrating vectors.
- The number of independent cointegrating vectors is the **cointegration rank**.
- When there is more than one cointegrating vector, estimation of the long-run relation identifies the cointegrated vector whose residuals are uncorrelated with any other $I(0)$ linear combination of y_{2t}, \dots, y_{Mt} .

- If the variables in y_t are cointegrated, ε_t is stationary.
- This suggests that the null of no cointegration can be tested by checking the stationarity of ε_t using DF tests.
- Because ε_t is not observed, the test is based on $\hat{\varepsilon}_t$ and the critical values have to be modified.
- The test equation is:

$$\Delta \hat{\varepsilon}_t = (\rho - 1) \hat{\varepsilon}_{t-1} + u_t$$

Critical values for the Engle-Granger test (5%)					
M	2	3	4	5	6
Without trend	-3.34	-3.74	-4.10	-4.42	-4.71
With trend	-3.78	-4.12	-4.43	-4.72	-4.98

- As usual, the test equation can be “**augmented**” to account for serial correlation.
- The test is **not invariant** to the choice of dependent variable in the long-run regression.

- The long-run equation identifies a cointegrating vector, but, because the series are $I(1)$, **standard inference is invalid**.
- Saikkonen (1991) showed that valid inference about the long-run parameters can be performed using standard tools in the following dynamic model (**DOLS**):

$$y_{1t} = Y_t' \theta + x_t' \delta + \sum_{j=-p}^p \Delta Y_{t+j}' \delta_j + e_t$$

where p is allowed to increase with T at an appropriate rate.

- In practice, the **choice** of p is as in the ADF tests.
- The leads and lags may be enough to ensure that the errors are serially uncorrelated; if that is not the case, robust estimators of the covariance matrix can be used.
- Cointegration also has implications for the **short-run dynamics** of the series.
- If the series are cointegrated, the equilibrium error ε_t contains information about Δy_{1t} .
- Models incorporating this **error correction mechanism** (ECM) were introduced by Sargan and popularized by Davidson, Hendry, Srba and Yeo (1978).

- The ECM can be incorporated in the model in different ways.
- For example, one can estimate models of the form

$$\Delta y_{1t} = \alpha + \rho \Delta y_{1t-1} + \Delta Y_t' \varphi + \Delta Y_{t-1}' \phi + \lambda \hat{\varepsilon}_{t-1} + \eta_t$$

$$\Delta y_{1t} = \alpha + \rho \Delta y_{1t-1} + \Delta Y_t' \varphi + \Delta Y_{t-1}' \phi + \lambda \left(y_{1t-1} - Y_{t-1}' \hat{\theta} - x_{t-1}' \hat{\delta} \right) + \eta_t$$

- Standard inference is valid because of the super-consistency of the first step.
- Alternatively, we may use the model

$$\Delta y_{1t} = \alpha + \rho \Delta y_{1t-1} + \Delta Y_t' \varphi + \Delta Y_{t-1}' \phi + \lambda_1 y_{1t-1} + Y_{t-1}' \lambda_2 + x_{t-1}' \lambda_3 + \eta_t$$

- Although this is an unbalanced regression, inference on individual parameters is valid because the model can be written in such way that all parameters are associated with $I(0)$ regressors (Sims, Stock and Watson, 1990).
- However, F -tests and tests involving more than one parameter are not valid.

4) Determining cointegrating rank

- How to determine the cointegration rank and identify the different cointegrating vectors?
- To answer this question, we need to look at a VAR model of the form (Johansen, 1988)

$$y_t = Bx_t + \sum_{i=1}^{p+1} \Phi_i y_{t-i} + \varepsilon_t$$

- This model can be rewritten in the **vector error correction model (VECM)** form

$$\Delta y_t = Bx_t + \Pi y_{t-1} + \sum_{i=1}^p \Gamma_i \Delta y_{t-i} + \varepsilon_t$$

with $\Gamma_p = -\Phi_{p+1}$, $\Gamma_i = \Gamma_{i+1} - \Phi_{i+1}$ for $i = 1, \dots, p-1$, and Π is the $M \times M$ **impact matrix**

$$\Pi = \sum_{i=1}^{p+1} \Phi_i - I.$$

- The **Granger representation theorem** implies that if the series are cointegrated, then they form a VECM.

- Under the **usual assumptions**, only the term Πy_{t-1} may not be $I(0)$.
- However, **for the equation to hold**, Πy_{t-1} must be $I(0)$.
- Different cases are possible:
 - (a) If $\Pi = 0$ **there is no cointegration**;
 - (b) If Π has full rank, there are M stationary linear combinations of the series and that is only possible if they are all $I(0)$.
 - (c) If the rank of Π is $0 < r < M$, and Πy_{t-1} is $I(0)$, there are r cointegrating vectors.
- To see why, write $\Pi = \alpha\eta'$, where η and α are $M \times r$ matrices (with rank r).
- Hence, the r columns of η are the r (independent) cointegrating vectors.
- α is the **loading matrix** which gives the coefficients of each ECM in each equation.
- Then, $\alpha\eta'y_{t-1}$ can only be stationary if the r columns of $\eta'y_{t-1}$ are $I(0)$.

- LR tests for $H_0 : r = r_1$ vs. $H_1 : r = r_2$ with $r_1 < r_2 \leq M$ can be performed by comparing models imposing $\Pi = \alpha\eta'$ for different values of r .
- This is the **trace test** and it has a non-standard distribution (but appropriate critical values are available).
- The distribution depends on $r_2 - r_1$, on the composition of x_t and on the possible restrictions on B .
- An important advantage of this test is that it does not depend on the chosen normalization of the cointegrating vector.

5) Estimation

- Estimating the system imposing $\Pi = \alpha\eta'$ yields (super-consistent) estimates of the cointegrating vectors.
- However, because $\alpha\eta' = \alpha\Theta\Theta^{-1}\eta'$, additional restrictions are needed to identify η .
- These restrictions have to be imposed on a case-by-case basis and can lead to non-linear models.
- Estimation can be performed by (quasi-) maximum likelihood under normality.

University week: 6

Maximum Likelihood Estimation

1. Likelihood function and the ML principle;
2. Properties of ML estimators;
3. Notes on the maximization of $\ln L(\theta)$;
4. Robust covariance matrix estimation;
5. Hypothesis testing.

Reading: Greene: 14.1–14.6, 14.8, E3.

1) Likelihood function and the ML principle

- Let $f(y; \theta)$ denote the probability density function of the random variable y , given θ .
- The **joint density** of n iid observations of y is

$$f(y_1, \dots, y_n | \theta) = \prod_{i=1}^n f(y_i; \theta).$$

- If y is a discrete random variable, $f(y_1, \dots, y_n | \theta)$ gives the probability of observing a particular sample, given θ .
- Let us now take $f(y_i; \theta)$ as a function of θ given y and write

$$L(\theta | y_1, \dots, y_n) = \prod_{i=1}^n f(y_i; \theta).$$

- This is the **likelihood function**, which gives the likelihood that the population parameter is θ , given the observed sample.
- Note: $L(\theta | y_1, \dots, y_n)$ is often abbreviated to $L(\theta)$.

- The **Maximum Likelihood (ML) principle** suggests that estimators of the unknown parameters are obtained by maximizing $L(\theta)$ with respect to θ .

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta).$$

- It is often convenient to work with the natural logarithm of the likelihood function $\ln L(\theta)$. For example, in the iid case:

$$\ln L(\theta|y_1, \dots, y_n) = \sum_{i=1}^n \ln f(y_i|\theta);$$

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta) = \arg \max_{\theta \in \Theta} \ln L(\theta)$$

- Let θ_0 denote the true parameter value. In the context of ML estimation, θ_0 is **identified** (estimable) if, for any other parameter vector $\theta^* \neq \theta_0$, there exists some set of outcomes with positive probability under θ_0 for which $L(\theta^*) \neq L(\theta_0)$.
- More generally, identification depends on the model, on the data and on the parameter.

- In the so-called **regular problems**, $\hat{\theta}$ can be obtained by solving the **likelihood equation**

$$\left. \frac{\partial \ln L(\theta)}{\partial \theta} \right|_{\hat{\theta}} = 0$$

- Occasionally the ML estimator is **not unique**.
- Also, $\ln L(\theta)$ may have only one global maximum, but multiple **local maxima**.
- The main **regularity conditions** (now assumed to hold) are as follows:
 - (a) The first three derivatives of $\ln f(y|\theta)$ with respect to θ are continuous and finite for almost all y and for all θ ;
 - (b) For all values of θ , $|\partial^3 \ln L(\theta) / \partial \theta_j \partial \theta_k \partial \theta_l|$ is limited by a function that has finite expectation;
 - (c) The domain of y does not depend on θ ;
 - (d) θ is an interior point to the convex parameter space Θ .

- In order to proceed, it is interesting to look at some important results (Bartlett identities).

- Define the **score** vector $g(\theta)$ and the **Hessian** matrix $H(\theta)$ as

$$g(\theta) = \frac{\partial \ln L(\theta)}{\partial \theta} = \sum_{i=1}^n \frac{\partial \ln f(y_i|\theta)}{\partial \theta} = \sum_{i=1}^n g_i, \quad H(\theta) = \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} = \sum_{i=1}^n \frac{\partial^2 \ln f(y_i|\theta)}{\partial \theta \partial \theta'} = \sum_{i=1}^n H_i.$$

- Notice that $\int f(y|\theta) dy = 1$ and therefore $\int \frac{\partial f(y|\theta)}{\partial \theta} dy = 0$.

- Noting that $\frac{\partial \ln h(a)}{\partial a} = \frac{\partial h(a)}{\partial a} / h(a) \Leftrightarrow \frac{\partial \ln h(a)}{\partial a} h(a) = \frac{\partial h(a)}{\partial a}$

$$\int \frac{\partial \ln f(y|\theta)}{\partial \theta} f(y|\theta) dy = 0$$

- Hence, for correctly specified models, $E_0 \left[\frac{\partial \ln f(y|\theta)}{\partial \theta} \right] = E_0 [g_i] = 0$ and $E_0 [g(\theta)] = 0$.

- Taking derivatives again and recalling that $\frac{\partial \ln h(a)}{\partial a} h(a) = \frac{\partial h(a)}{\partial a}$,

$$\int \left[\frac{\partial^2 \ln f(y|\theta)}{\partial \theta \partial \theta'} f(y|\theta) + \frac{\partial \ln f(y|\theta)}{\partial \theta} \frac{\partial f(y|\theta)}{\partial \theta'} \right] dy = 0$$

$$\int \left[\frac{\partial^2 \ln f(y|\theta)}{\partial \theta \partial \theta'} f(y|\theta) + \frac{\partial \ln f(y|\theta)}{\partial \theta} \frac{\partial \ln f(y|\theta)}{\partial \theta'} f(y|\theta) \right] dy = 0$$

- Hence,

$$\int \left[\frac{\partial \ln f(y|\theta)}{\partial \theta} \frac{\partial \ln f(y|\theta)}{\partial \theta'} f(y|\theta) \right] dy = - \int \left[\frac{\partial^2 \ln f(y|\theta)}{\partial \theta \partial \theta'} f(y|\theta) \right] dy$$

and we have that, for correctly specified models,

$$\mathbb{E}_0 \left[\frac{\partial \ln f(y|\theta)}{\partial \theta} \frac{\partial \ln f(y|\theta)}{\partial \theta'} \right] = \text{Var}_0 [g_i] = \mathbb{E}_0 \left[-\frac{\partial^2 \ln f(y|\theta)}{\partial \theta \partial \theta'} \right] = \mathbb{E}_0 [-H_i]$$

- $\text{Var}_0 [g_i] = E_0 [g_i g_i']$ defines Fisher's **information matrix**, denoted $\mathcal{I}(\theta)$.
- Hence, the result

$$\text{Var}_0 [g_i] = E_0 [g_i g_i'] = E_0 [-H_i]$$

is called the **information matrix identity**.

- Obviously, it implies that, for correctly specified models,

$$E_0 \left[\sum_{i=1}^n g_i g_i' \right] = E_0 [-H(\theta)] = n\mathcal{I}(\theta)$$

which is a result that will be used below.

2) Properties of MLE

- Under the assumed regularity conditions the MLE possesses the following properties:

(a) **Consistency:** $\text{plim } \hat{\theta} = \theta_0$;

(b) **Asymptotic normality:** $\hat{\theta} \stackrel{a}{\sim} \mathcal{N}(\theta_0, [n\mathcal{I}(\theta_0)]^{-1})$;

(c) **Asymptotic efficiency:** if $\tilde{\theta}$ is a regular consistent asymptotically normal estimator such that $\tilde{\theta} \stackrel{a}{\sim} \mathcal{N}(\theta_0, \Omega_0)$, then $\Omega_0 \geq [n\mathcal{I}(\theta_0)]^{-1}$;

* i.e., under these RC, the MLE asymptotically achieves the **Cramer-Rao** lower bound;

(d) **Invariance:** If $c(\theta)$ is a continuous and continuously differentiable one-to-one function, the MLE of $\gamma_0 = c(\theta_0)$ is $c(\hat{\theta})$.

- It is interesting to sketch the proof of the asymptotic normality.
- Under the assumed regularity conditions, we have $g(\hat{\theta}) = 0$.
- Expanding this result in a 1st order Taylor series around θ_0 we have

$$g(\hat{\theta}) = g(\theta_0) + H(\bar{\theta})(\hat{\theta} - \theta_0) = 0$$

where $\bar{\theta} = w\hat{\theta} + (1 - w)\theta_0$ for $0 < w < 1$ (mean-value theorem). Then,

$$(\hat{\theta} - \theta_0) = [-H(\bar{\theta})]^{-1} g(\theta_0)$$

$$\sqrt{n}(\hat{\theta} - \theta_0) = [-H(\bar{\theta})]^{-1} [\sqrt{n}g(\theta_0)]$$

- Notice that, because $\text{plim}(\hat{\theta} - \theta_0) = 0$, we have that $\text{plim}(\bar{\theta} - \theta_0) = 0$ and therefore

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} [-H(\theta_0)]^{-1} [\sqrt{n}g(\theta_0)]$$

- We now divide $H(\theta_0)$ and $g(\theta_0)$ by n to obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} [-\bar{H}(\theta_0)]^{-1} [\sqrt{n}\bar{g}(\theta_0)].$$

- Because $\bar{g}(\theta_0)$ is the mean of n zero-mean random variables, we can apply an appropriate Central Limit Theorem to obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, A^{-1}BA^{-1})$$

$$A = \lim_{n \rightarrow \infty} E_0[-\bar{H}(\theta_0)] \quad B = \lim_{n \rightarrow \infty} n[E_0(\bar{g}(\theta_0))(\bar{g}(\theta_0)')]$$

- Notice that $(\bar{g}(\theta_0))(\bar{g}(\theta_0)') = \frac{1}{n^2}(\sum_{i=1}^n g(\theta_0)_i)(\sum_{i=1}^n g(\theta_0)_i)'$ and because the expectation of the cross-products is zero under independence

$$B = \lim_{n \rightarrow \infty} E_0 \left[\frac{1}{n} \sum_{i=1}^n g(\theta_0)_i g(\theta_0)_i' \right].$$

- Hence, for correctly specified models, $B = \lim_{n \rightarrow \infty} E_0 [-\bar{H}(\theta)] = A = \mathcal{I}(\theta_0)$ and

$$\sqrt{n} (\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, A^{-1});$$

$$\hat{\theta} \overset{a}{\approx} \mathcal{N}(\theta_0, [n\mathcal{I}(\theta_0)]^{-1}).$$

- There are three commonly used estimators of $[n\mathcal{I}(\theta_0)]^{-1}$.
 - (a) **Expected Information:** If the form of the expected values of the second derivatives of the log-likelihood function is known, then we can evaluate $[n\mathcal{I}(\theta_0)]^{-1}$ at $\hat{\theta}$.
 - (b) **Observed Information:** Simply use $[-H(\hat{\theta})]^{-1}$.
 - (c) **Outer Product of the Gradient (OPG or BHHH):** Because of the information matrix identity, we can also use $\left[\sum_{i=1}^n g_i(\hat{\theta}) g_i(\hat{\theta})' \right]^{-1}$.
- None of these estimators is dominant, but the OPG is notorious for its poor finite sample performance.

- The previous results are easy to extend to accommodate the presence of covariates.
- Suppose the joint distribution of y and x depends on α , giving $f(y, x|\alpha) = f(y|x, \alpha)g(x|\alpha)$.
- Next, suppose that α can be divided into θ and δ , so that (exogeneity of x) $f(y, x|\alpha) = f(y_i|x_i, \theta)g(x_i|\delta)$.
- For an iid sample $\{(y_i, x_i)\}_{i=1}^n$ then

$$\ln L(\theta, \delta|y_i, x_i) = \sum_{i=1}^n \ln f(y_i, x_i|\alpha) = \sum_{i=1}^n \ln f(y_i|x_i, \theta) + \sum_{i=1}^n \ln g(x_i|\delta).$$

- $\hat{\theta}$ can then be obtained by maximizing just $\sum_{i=1}^n \ln f(y_i|x_i, \theta)$ with respect to θ . Therefore, frequently we will work directly with the conditional log-likelihood

$$\ln L(\theta|y_i, x_i) = \sum_{i=1}^n \ln f(y_i|x_i, \theta),$$

and this (under appropriate regularity conditions) will behave to a large extent like a standard log-likelihood.

- However, now $E_0[-H(\theta)] = \sum_{i=1}^n E_0[-H_i] = \sum_{i=1}^n \mathcal{I}_i(\theta) \neq n\mathcal{I}(\theta)$, and so on.

3) Notes on the maximization of $\ln L(\theta)$

- The objective function $\ln L(\theta)$ can be maximized using the Gauss-Newton algorithm:

$$g(\hat{\theta}) = 0 \simeq g(\theta^1) + H(\theta^1)(\hat{\theta} - \theta^1)$$

$$\hat{\theta} \simeq \theta^1 - [H(\theta^1)]^{-1} g(\theta^1)$$

- The step depends on gradient (direction) and on the Hessian (curvature).
- Except when $\ln L(\theta)$ is quadratic, this procedure has to be iterated.
- Convergence is achieved when $g(\theta^1)' [H(\theta^1)]^{-1} g(\theta^1)$ is small enough.
- In practice it is generally important to adjust the step-size to ensure that the objective function increases at each iteration. That is, for $\lambda_1 > 0$, we use

$$\hat{\theta} \simeq \theta^1 - \lambda_1 [H(\theta^1)]^{-1} g(\theta^1).$$

- For regions where $\ln L(\theta)$ is not concave, one can reverse the sign of the Hessian, but there is no particular justification to use a Hessian based step.
- Other algorithms are obtained by replacing $H(\theta^1)$ with other matrices:
 - ◇ $-\sum_{i=1}^n \mathcal{I}_i(\theta^1)$ (Fisher's scoring algorithm): This is rarely used;
 - ◇ $-\sum_{i=1}^n g_i(\theta^1) g_i(\theta^1)'$ (BHHH): Easy to compute but rather slow to converge;
 - ◇ $-I_k$ (steepest ascent): Even easier and slower.
 - ◇ An *arc* Hessian can be used in non-quadratic problems, as in the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm: this tends to work well.
- In many problems multiple maxima are possible. In these cases we should either try many (random) starting values or use other algorithms, like generalized simulated annealing.

4) Robust covariance matrix estimation

- If the likelihood function is misspecified, the MLE is generally inconsistent for the parameters of interest.
- However, under very general conditions, $\text{plim } \hat{\theta} = \theta^*$, where the **pseudo-true value** θ^* minimizes the Kullback-Leibler index defined by

$$E_0 [\ln f_0 - \ln f(\theta)].$$

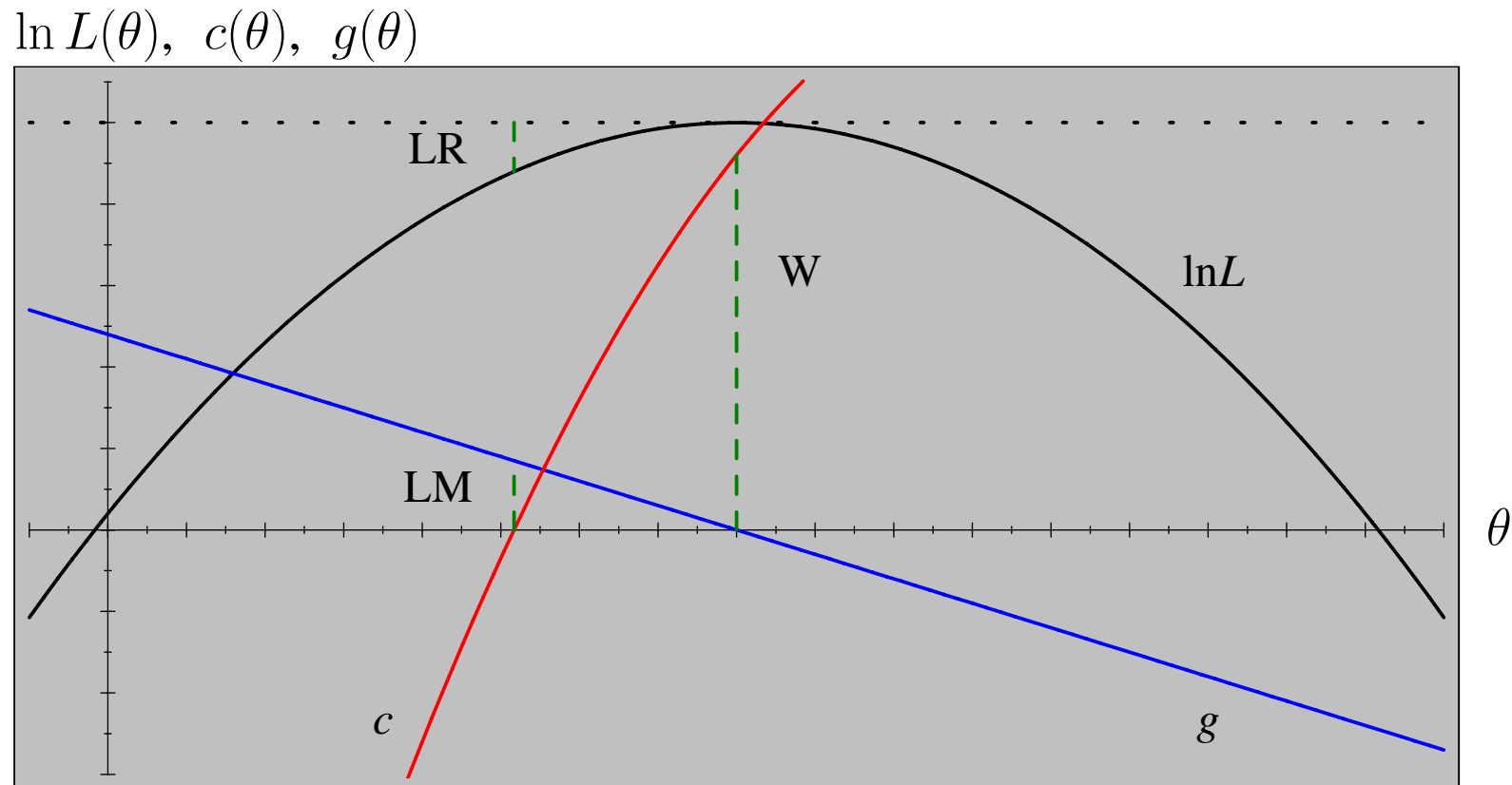
- That is, the MLE leads to the **best approximation**, in the Kullback-Leibler sense, to f_0 , the true density.
- However, because the IM identity does not hold, the asymptotic covariance matrix is given by the so-called **sandwich** estimator:

$$A^{-1}BA^{-1}, \quad A = \lim_{n \rightarrow \infty} E_0 [-\bar{H}(\theta_0)] \quad \text{and} \quad B = \lim_{n \rightarrow \infty} \left[n^{-1} E_0 \left(\sum_{i=1}^n g(\theta_0)_i g(\theta_0)_i' \right) \right].$$

- Notice that in some special cases (when the score only depends on correctly specified low order moments of y) the MLE is **robust** to some forms of misspecification (i.e., it is still consistent).

5) Hypothesis Testing

- Suppose now that we want to test the null hypothesis $H_0 : c(\theta) = 0$ against $H_A : c(\theta) \neq 0$, where $c(\theta)$ is a $m \times 1$ vector.
- There are three likelihood-based procedures for testing: likelihood ratio (LR), Wald (W) and score or Lagrange multiplier (LM).



- The main result to be used is that if $x \sim \mathcal{N}_J(\mu, \Sigma)$, then $(x - \mu)' \Sigma^{-1} (x - \mu) \sim \chi_{(J)}^2$
- **Wald test:** Expand $\sqrt{n}c(\hat{\theta})$ around θ_0 to obtain

$$\sqrt{n}c(\hat{\theta}) = \sqrt{n}c(\theta_0) + \sqrt{n}C(\bar{\theta})(\hat{\theta} - \theta_0)$$

– Therefore, under the null

$$\sqrt{n}c(\hat{\theta}) \xrightarrow{d} \mathcal{N}_m\left(0, C(\theta_0)\text{Var}_0\left(\sqrt{n}(\hat{\theta} - \theta_0)\right)C(\theta_0)'\right)$$

$$W = c(\hat{\theta})' \left[C(\hat{\theta})\text{Var}_0(\hat{\theta})C(\hat{\theta})' \right]^{-1} c(\hat{\theta}) \stackrel{a}{\sim} \chi_{(m)}^2$$

- Notice that all that is needed for the validity of the test is a consistent estimate of $\text{Var}_0(\hat{\theta})$.
- The test is not invariant to reparameterizations of $c(\theta)$.

- **Likelihood Ratio:** Let $\hat{\theta}_R$ be

$$\hat{\theta}_R = \arg \max_{\theta \in \Theta, c(\theta)=0} \ln L(\theta).$$

- Now, expand $\ln L(\hat{\theta}_R)$ around $\hat{\theta}$ to obtain

$$\ln L(\hat{\theta}_R) = \ln L(\hat{\theta}) + (\hat{\theta} - \hat{\theta}_R)' g(\hat{\theta}) + \frac{1}{2} (\hat{\theta} - \hat{\theta}_R)' H(\ddot{\theta}) (\hat{\theta} - \hat{\theta}_R)$$

$$2 \left[\ln L(\hat{\theta}) - \ln L(\hat{\theta}_R) \right] = (\hat{\theta} - \hat{\theta}_R)' \left[-H(\ddot{\theta}) \right] (\hat{\theta} - \hat{\theta}_R)$$

- It is possible to show that, under the null,

$$\text{LR} = 2 \left[\ln L(\hat{\theta}) - \ln L(\hat{\theta}_R) \right] \stackrel{a}{\sim} \chi_{(m)}^2.$$

- Notice that the observed information is implicitly used as an estimate of the information matrix.

- **Lagrange multiplier (score):** It is also possible to show that, under the null,

$$\text{LM} = \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R} \right)' \left[\sum_{i=1}^n \hat{\mathcal{I}}_i(\hat{\theta}_R) \right]^{-1} \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R} \right) \stackrel{a}{\sim} \chi_{(m)}^2.$$

where $\sum_{i=1}^n \hat{\mathcal{I}}_i(\hat{\theta}_R)$ is an estimator of the information matrix based on the restricted MLE.

- Note that we can use any of a variety of information matrix estimators.
- The notorious OPG version of the test can be computed as n times the R^2 of an auxiliary regression.
- This is the only test whose distribution is the same, whether or not the null is on the boundary of the parameter space.
- Although the 3 tests are asymptotically equivalent, their finite sample performance can be very distinct.
- Under the alternative, all tests have non-central $\chi_{(m)}^2$ distributions.

- These tests are asymptotically equivalent under the null and are only asymptotically valid.
- However, different inferences can be drawn from testing the same hypothesis with the different procedures.
- Which test should be used, and when?
- The Wald test is not invariant to reparameterizations.
- On the other hand, it is easy to modify (by using the sandwich covariance matrix) so that it is valid even if some assumptions of the model do not hold.
- The LM test can be modified in a similar way, but this is more cumbersome.
- The LM test is often interesting when we want to check for departures from the maintained assumptions as we do not have to fully specify the alternative.
- The LR test tends to have good finite sample behaviour but is dependent on the distributional assumptions.

University week: 7

Generalized Method of Moments (GMM)

1. Method of moments estimation;
2. GMM; Optimal GMM;
3. Hypothesis testing;
4. Testing overidentifying restrictions.

Reading: Greene: 13.1, 13.2, 13.4–13.7.

1) Method of moments estimation

- The method of moments provides an alternative to ML estimation.
- The method of moments estimators are often less efficient than ML, but require less demanding assumptions.
- Suppose we wish to estimate a $(k \times 1)$ parameter vector $\theta_0 = (\theta_{01}, \dots, \theta_{0k})'$ using $\{y_i\}_{i=1}^n$, an iid sample of observations on a scalar random variable y with pdf $f(y; \theta_0)$, whose moments exist up to order $2k$.
- Recall that the j -th uncentered population moment of y is defined by

$$\mu_j \equiv \mathbb{E}[y^j] = \int_{-\infty}^{\infty} y^j f(y; \theta_0) dy,$$

and suppose that $\mu_j = h_j(\theta_0)$, where $h_j(\cdot)$ is a known function.

- With an iid sample, we can estimate μ_j by $\hat{\mu}_j = n^{-1} \sum_{i=1}^n y_i^j$ for $j = 1, \dots, 2k$.

- Let $\mu = (\mu_1, \dots, \mu_k)'$ and $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_k)'$. By the weak law of large numbers

$$\hat{\mu} \xrightarrow{p} \mu$$

and by the central limit theorem:

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \Omega),$$

where the element (s, t) of Ω is $\omega_{st} = \mu_{s+t} - \mu_s \mu_t$.

- With $h(\theta) = (h_1(\theta), \dots, h_k(\theta))'$, the **method of moments** estimator $\hat{\theta}$ of θ_0 is defined by the following set of k equations in k unknowns:

$$h(\hat{\theta}) = \hat{\mu}.$$

- If $h(\cdot)$ is invertible then $\hat{\theta} = h^{-1}(\hat{\mu})$. Furthermore, if $h^{-1}(\cdot)$ is continuous at μ then:
 - by the Slutsky theorem, $\hat{\theta} \xrightarrow{p} \theta_0$;
 - by the delta method, $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, H_0^{-1} \Omega (H_0^{-1})'\right)$, where $H_0 = \partial h(\theta) / \partial \theta'$ at $\theta = \theta_0$.

- We can estimate ω_{st} and H_0 by

$$\hat{\omega}_{st} = \hat{\mu}_{s+t} - \hat{\mu}_s \hat{\mu}_t$$

$$\hat{H} = \left. \frac{\partial h(\theta)}{\partial \theta'} \right|_{\theta=\hat{\theta}}.$$

- Hence, we can estimate the asymptotic variance of $\hat{\theta}$ and construct Wald-type tests of hypotheses about θ_0 using

$$\hat{\theta} \stackrel{a}{\sim} \mathcal{N} \left(\theta_0, n^{-1} \hat{H}^{-1} \hat{\Omega} \left(\hat{H}^{-1} \right)' \right).$$

- Which moments to match?
- Why not use more than k sample moments to estimate θ_0 ?

2) GMM; Optimal GMM

- Suppose that an economic or statistical model leads to moment conditions of the form

$$E_0[m(y; \theta)] = 0,$$

where $m(y; \theta)$ is a $(L \times 1)$ vector function of y , **which hold if and only if** $\theta = \theta_0$.

- An example is $E_0[g_i] = 0$, in regular ML estimation.
- The method of moments estimator is defined by moment conditions of the form

$$E_0[y^j - h_j(\theta_0)] = 0, \quad j = 1, \dots, k.$$

- First order conditions in stochastic optimization problems also lead to moment conditions of this form.
- For $L \geq k$, the Generalized Method of Moments (GMM) provides a way of estimating θ using only these moment conditions.

- Let W_n be a $(L \times L)$ symmetric positive definite (pd) matrix, which may depend on the sample data, and which converges in probability to a $(L \times L)$ symmetric pd matrix W_0 .
- Then, if $\{y_i\}_{i=1}^n$ is an iid sample of y and $\bar{m}_n(\theta) = n^{-1} \sum_{i=1}^n m(y_i; \theta)$, the GMM estimator of $\hat{\theta}$ is defined by

$$\hat{\theta} = \arg \min_{\theta \in \Theta} [\bar{m}_n(\theta)]' W_n [\bar{m}_n(\theta)].$$

- Under fairly weak regularity conditions (including $\Sigma(\theta) = \text{Var}[\sqrt{n}\bar{m}_n(y; \theta)]$ being finite non-singular for all θ),

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

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, A_0^{-1} B_0 A_0^{-1})$$

$$\hat{\theta} \stackrel{a}{\sim} \mathcal{N}(\theta_0, n^{-1} A_0^{-1} B_0 A_0^{-1})$$

with $A_0 = \bar{M}'_0 W_0 \bar{M}_0$, $B_0 = \bar{M}'_0 W_0 \Sigma_0 W_0 \bar{M}_0$ and $\Sigma_0 = \Sigma(\theta_0)$, $\bar{M}_0 = \left. \frac{\partial \bar{m}_n(y; \theta)}{\partial \theta'} \right|_{\theta = \theta_0}$.

- Under mild conditions, we can consistently estimate M_0 and Σ_0 by

$$\hat{M}_n = \bar{M}_n(\hat{\theta}) \equiv \frac{1}{n} \sum_{i=1}^n \left[\frac{\partial m(y; \theta)}{\partial \theta'} \Big|_{\theta=\hat{\theta}} \right], \quad \hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n m(y_i; \hat{\theta}) m(y_i; \hat{\theta})',$$

and hence construct consistent standard errors and Wald-type tests.

- If $L = k$ we can find $\hat{\theta}$ such that $\bar{m}_n(\hat{\theta}) = 0$, in which case the choice of weighting matrix W_n is unimportant:
 - $\hat{\theta}$ does not depend on W_0 ;
 - Notice that in this case \bar{M}_0 is $k \times k$ and if the rank condition is met we have that $A_0^{-1} B_0 A_0^{-1} = M_0^{-1} \Sigma_0 M_0^{-1}$.
- For the **over-identified** case, typically there is no solution to $\bar{m}_n(\hat{\theta}) = 0$ and the choice of weighting matrix W_n is important.
- Because W_n is only required to be symmetric pd, we can set $W_n = I_L$.
- However, we can **gain efficiency** by using a matrix that gives less weight to the noisier moment conditions.

- The asymptotic variance of $\hat{\theta}$ is **minimised** when $W_0 = \Sigma_0^{-1}$.
- In that case, $A_0^{-1}B_0A_0^{-1} = (M_0'\Sigma_0^{-1}M_0)^{-1}$.
- **Optimal GMM** can be implemented in several ways.
- A popular choice is to use a two-step procedure (which can be iterated):
 - **Step I:** Set W_n to be some known symmetric positive matrix function of the data (e.g., $W_n = I_L$) and implement GMM to obtain an initial consistent estimator $\hat{\theta}$ of θ_0 ;
 - **Step II:** Construct $\hat{\Sigma}_n = n^{-1} \sum_{i=1}^n m(y_i; \hat{\theta})m(y_i; \hat{\theta})'$ and then perform a second round of GMM with W_n set equal to $\hat{\Sigma}_n^{-1}$ to obtain an optimal GMM estimator $\tilde{\theta}$ of θ_0 .
- There is now quite a lot of evidence that, unfortunately, the 2-step optimal GMM performs badly in finite samples.
- The Windmeijer (2005) correction can be used to improve inference.
- An alternative is to use the continuous updating estimator (CUE).

- Up to now, we have only considered **unconditional moment conditions** of the form $E_0[m(y; \theta)] = 0$.

- However, in many econometric models the parameters are characterized by **conditional moment** conditions like

$$E_0[u(y; x, \theta) | Z] = 0.$$

- Even if there is only one **parameter-defining** equation $u(y; x, \theta)$, as we now assume to be the case, estimation is possible because this moment condition implies that, for a very wide range of vector functions $g(Z)$,

$$E_0[g(Z)u(y; x, \theta)] = 0.$$

- These new moment conditions can be seen as the **orthogonality** between a “residual” and an “instrument”.
- Because there is a very large set of “instruments”, there is a wide collection of unconditional moment restrictions (in fact the collection is infinite).
- This leads to the issue of selecting the instruments to use.

- Several results are useful here:

(a) The **asymptotic efficiency** of the estimator cannot be reduced by adding instruments.

* However, it is not wise to use too many instruments because there is a **trade-off** between asymptotic efficiency and finite sample bias.

* Also, as we will see, there are **diminishing returns** to adding instruments.

(b) The asymptotic covariance of GMM estimators constructed using instruments belonging to the information set are bounded from below (**GMM bound**).

* This explains the diminishing returns mentioned above.

(c) The GMM bound is attained when the instruments for observation i are given by the i -th column of

$$E_0 [U' | Z] \Omega^{-1}$$

where Ω denotes the $n \times n$ variance matrix of the “residuals” $u(y; x, \theta)$ and U is a $n \times k$ matrix with typical element $\partial u(y_i; x_i, \theta) / \partial \theta_j$

* In general, **this result cannot be directly applied** because Ω is unknown and the expectation may be impossible to compute without further assumptions.

- So far we have assumed that the observations y_i come from an iid sample.
- However, in dealing with time series data, it is often the case that there is serial correlation in $m(y_t; \theta)$.
- Hence, in that case, $\text{Var}[\bar{m}_T(\theta)] \neq \text{Var}[m(y_t; \theta)]/T$.
- This will cause problems for the properties of the GMM standard errors and the two-step optimal GMM estimator outlined earlier.
- Under fairly weak regularity conditions, $\Sigma_0 = \text{Var}[\sqrt{n}\bar{m}_T(\theta)]$ can be estimated using the kernel-based heteroskedasticity-autocorrelation consistent (HAC) variance matrix estimator (Newey and West, 1987).

3) Hypothesis testing

- Suppose that we wish to test $H_0 : c(\theta_0) = 0$, where $c(\theta)$ is $(l \times 1)$ and $C(\theta) = \partial c / \partial \theta'$ is continuously differentiable with rank l .
- It is possible to define **optimal GMM** counterparts to the LR, Wald and score tests.
- The optimal GMM counterpart to the LR test is based on the following result.
- Let $\hat{q}_n(\tilde{\theta}) = \bar{m}_n(\theta)' \hat{\Sigma}^{-1} \bar{m}_n(\theta)$ denote the GMM objective function from the second stage of optimal GMM. Provided that $E_0[m(y; \theta)] = 0$ and $c(\theta_0) = 0$ hold,

$$n \times [\hat{q}_n(\tilde{\theta}_R) - \hat{q}_n(\tilde{\theta})] \xrightarrow{d} \chi_{(l)}^2,$$

where

$$\tilde{\theta}_R = \arg \min_{\theta \in \Theta, c(\theta)=0} \hat{q}_n(\theta)$$

- **Notice** that the same matrix W_n and the same “instruments” have to be used for the restricted and unrestricted estimation.

- Naturally, the Wald test statistic is given by

$$\mathbf{W} = n \times c(\tilde{\theta})' \left[C(\tilde{\theta}) \left(\hat{M}'_n \hat{\Sigma}_n^{-1} \hat{M}_n \right)^{-1} C(\tilde{\theta})' \right]^{-1} c(\tilde{\theta}) \stackrel{a}{\sim} \chi^2_{(l)}.$$

- Finally, the analog of the LM test statistic is

$$\mathbf{LM} = n \times \left[\bar{m}_n(\tilde{\theta}_R) \hat{\Sigma}_{\tilde{\theta}_R}^{-1} \bar{M}_{\tilde{\theta}_R} \right] \left[\bar{M}'_{\tilde{\theta}_R} \hat{\Sigma}_{\tilde{\theta}_R}^{-1} \bar{M}_{\tilde{\theta}_R} \right]^{-1} \left[\bar{M}'_{\tilde{\theta}_R} \hat{\Sigma}_{\tilde{\theta}_R}^{-1} \bar{m}_n(\tilde{\theta}_R) \right] \stackrel{a}{\sim} \chi^2_{(l)}.$$

- This is just a Wald test for the significance of the “score” vector of the unrestricted models, evaluated at the restricted optimal GMM estimator.

4) Testing overidentifying restrictions

- We can use two-step optimal GMM to perform a different type of hypothesis test.
- Let again $\hat{q}_n(\tilde{\theta}) = \bar{m}_n(\theta)' \hat{\Sigma}^{-1} \bar{m}_n(\theta)$ denote the GMM objective function from the second stage of optimal GMM.
- If there exists a value $\theta_* = \text{plim } \tilde{\theta}$ of θ such that $E_*[m(y; \theta)] = 0$ then

$$n \times \hat{q}_n(\tilde{\theta}_n) \xrightarrow{d} \chi_{(L-k)}^2.$$

- This is Hansen's J-test (a robust version of Sargan's test) for overidentifying restrictions.
- This test checks whether $E_*[m(y_i; \theta)] = 0$, not whether $E_0[m(y_i; \theta)] = 0$.
- This point is often ignored and therefore the J-test is often misinterpreted as checking the validity of the moment conditions (or the validity of the instruments).
- Notice also that the LR-type test is just an over-identification test when θ is exactly identified in the unrestricted model.
- Specification tests based on additional moment conditions may also be performed this way.

- Suppose that under correct specification, the model implies l additional moment restrictions.
- These can be used in the estimation and the J-test can be used to check their validity.
- Alternatively, a score test approach can be used to construct these **conditional moment** tests.
- Let $m^*(y; \tilde{\theta})$ denote the extended set of moment conditions, evaluated at $\tilde{\theta}$ (obtained using only $m(y; \theta)$).
- Then, a valid test for the additional moment conditions can be performed by comparing

$$CM = \left[\frac{1}{n} \sum_{i=1}^n m^*(y_i; \tilde{\theta}) \right]' \left[\frac{1}{n^2} \sum_{i=1}^n m^*(y_i; \tilde{\theta}) m^*(y_i; \tilde{\theta})' \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^n m^*(y_i; \tilde{\theta}) \right]$$

to the appropriate critical values from the $\chi_{(l)}^2$ distribution.

- This is just the ESS (or nR^2) of a regression of 1 on $m^*(y_i; \tilde{\theta})$.

University week: 8

Semiparametric and Nonparametric Estimation

1. Nonparametric estimation;
2. Semiparametric estimators;
3. Quantile Regression;
4. Bootstrapping.

Reading: Greene: 7.3–7.5, 12.4, 15.2, 15.4.

1) Nonparametric estimation

- A **fully parametric** model is one where the parameter values uniquely identify the conditional probability distribution of the dependent variables.
- A **nonparametric** model is one in which there are no simple parameters as such.
- A **semiparametric** model is one in which there are parameters but they do not uniquely identify the conditional probability distribution of the dependent variables.
- Sometimes, semiparametric models are called seminonparametric, but this term should be reserved for the case in which a semiparametric model involves some form of nonparametric estimation.
- The basic building block of most semi- and nonparametric estimators is the kernel density estimator.

- Suppose $\{y_i\}_{i=1}^n$ are iid, where y_i is a random variable with a continuous distribution.
- A crude estimator of the density of y is the histogram.
- The shape of the histogram depends on the **bin width**, on the **number of bins** and on the chosen **origin**.
- Fixing these parameters, the density of y at $y = y_0$ given by the histogram is

$$\hat{f}(y_0) = \frac{1 \text{ \# of observations in the same bin as } y_0}{n \text{ width of the bin containing } y_0}$$

- If we define h as the bin width and let y^* denote the midpoint of the interval containing y_0 , the histogram estimate is given by

$$\hat{f}(y_0) = \frac{1}{nh} \sum_{j=1}^n \mathbf{1} \left(-\frac{1}{2} \leq \frac{y_j - y^*}{h} < \frac{1}{2} \right)$$

- This estimator is not continuous and depends on 3 “parameters”.

- An improvement over the histogram is obtained as follows.
- From the definition of a probability density, we have that

$$f(y_0) = \lim_{h \rightarrow 0} \frac{F(y_0 + h/2) - F(y_0 - h/2)}{h} = \lim_{h \rightarrow 0} \frac{1}{h} \Pr(y_0 - h/2 < y < y_0 + h/2)$$

- Estimating this probability by the proportion of the sample in the interval $(y_0 - h/2, y_0 + h/2)$, leads to the naïve density estimator:

$$\hat{f}(y_0) = \frac{1}{nh} \sum_{j=1}^n \mathbf{1} \left(y_0 - \frac{h}{2} \leq y_j < y_0 + \frac{h}{2} \right) = \frac{1}{nh} \sum_{j=1}^n \mathbf{1} \left(-\frac{1}{2} \leq \frac{y_j - y_0}{h} < \frac{1}{2} \right)$$

- Like the histogram, this estimator is not continuous, but it does not depend on the origin nor on the number of bins.
- A continuous and smooth estimator can be obtained by replacing the function $K[z] = \mathbf{1}(-1/2 \leq z < 1/2)$ by another **kernel** function such that $K(\cdot) \geq 0$ and $\int K(u) du = 1$.

- Many kernels are available, but a popular choice is to use $K [z] = \phi (z)$.
- In general, for a **bandwidth** $h > 0$, the kernel density estimator is given by

$$\hat{f}(y_0) = \frac{1}{nh} \sum_{j=1}^n K \left(\frac{y_j - y_0}{h} \right).$$

- This can be generalized to d -dimensional data.
- Let $\{Y_i\}_{i=1}^n$ be iid, where Y_i is a d -dimensional random vector with a continuous distribution, and let $K_d(\cdot)$ be a multivariate kernel function such that $\int_{\mathbb{R}^d} K_d(u) du = 1$.
- Then, the kernel estimator of $f(Y_0)$ is

$$\hat{f}(Y_0) = \frac{1}{nh^d} \sum_{j=1}^n K_d \left(\frac{Y_j - Y_0}{h} \right).$$

- Because a single smoothing parameter is used, it is convenient to **pre-scale** the data.

- To ensure **consistency**, among other things, we need to assume
 - $f(\cdot)$ is continuous at $Y = Y_0$;
 - $\lim_{n \rightarrow \infty} h = 0$;
 - $\lim_{n \rightarrow \infty} nh^d = \infty$.
- Notice the “**curse of dimensionality**”: the larger the value of d , the slower the rate of asymptotic convergence in distribution.
- Under additional conditions on $f(\cdot)$, $K_d(\cdot)$ and h , it is also possible to ensure asymptotic **normality**.
- However, convergence is slower than rate $n^{-1/2}$ and the estimator is biased.
- The rate of convergence depends on $K_d(\cdot)$ but in practice this choice is relatively unimportant.
- The **bandwidth** h determines the bias and variance of the estimator.
- As h decreases, so does the bias; however, the variance increases.

- The choice of bandwidth is critical for the results.
- Too large a bandwidth, means over-smoothing and eliminates important features of the function being estimated.
- Too small a bandwidth, means under-smoothing and the important features of the function become obscured by noise.
- Choices usually implemented by:
 - Automatic rules;
 - Graphically analysing the effects of varying the bandwidth;
 - Minimizing selection criterion, e.g. using cross-validation.
- Silverman's rule-of-thumb, for the Gaussian kernel, is a popular choice

$$h = \min \left\{ \sigma, \frac{IQR}{1.34} \right\} \frac{0.9}{n^{1/5}}.$$

- The kernel density estimator can be used for **Kernel Regression**.
- Particularly in the **exploratory analysis** of the data, it may be convenient not to assume a functional form for the relation between two variables.
- Let $\{(y_i, x_i)\}_{i=1}^n$ be iid, where y_i and x_i are random variables with continuous distributions and suppose we want to estimate $E(y|x) = m(x_i)$.
- Consider the model

$$y_i = m(x_i) + \epsilon_i; \quad i = 1, \dots, n$$

where ϵ_t is a zero mean iid error term and $m(x_i)$ is an **unknown but smooth** function.

- If we were interested in estimating $m(x_i)$ at a particular value of x and **if multiple observations** of y were available for that particular value of x , we could use

$$\hat{m}(x_0) = \frac{\sum_{j=1}^n \mathbf{1}(x_j = x_0) y_j}{\sum_{j=1}^n \mathbf{1}(x_j = x_0)} = \sum_{j=1}^n \frac{\mathbf{1}(x_j = x_0)}{\sum_{j=1}^n \mathbf{1}(x_j = x_0)} y_j = m(x_0) + \frac{\sum_{j=1}^n \mathbf{1}(x_j = x_0) \epsilon_j}{\sum_{j=1}^n \mathbf{1}(x_j = x_0)}$$

which, under mild assumptions, is consistent if $\sum_{j=1}^n \mathbf{1}(x_j = x_0) \rightarrow \infty$.

- In general, repeated observations are not available, but if $m(\cdot)$ is sufficiently smooth, we can estimate $m(x_0)$ using observations in a **neighbourhood** of x_0 .
- A **smoothing estimator** of $m(x_0)$ can be written as

$$\hat{m}(x_0) = \sum_{j=1}^n \omega_h(x_0 - x_j) y_j$$

where the weights $\omega_h(x_0 - x_j)$ decrease with the distance between x_j and x_0 .

- Frequently, the weights are defined as

$$\omega_h(x_0 - x_j) = \frac{K\left(\frac{x_0 - x_j}{h}\right)}{\sum_{k=1}^n K\left(\frac{x_0 - x_k}{h}\right)}.$$

- This defines the **Nadaraya-Watson** estimator of $m(x_0)$.
- Under appropriate regularity conditions, $\hat{m}(x_0)$ is consistent when $\lim_{n \rightarrow \infty} h = 0$, but the **convergence is slow** (not $n^{1/2}$).

- Again, the properties of the estimator depend on the choice of $K(\cdot)$ and h .
 - The choice of $K(\cdot)$ is not that important, and often $K(\cdot) = \phi(\cdot)$.
 - The choice of h is critical: If h is too small, the estimates become too irregular; If h is too large, the estimates approach \bar{y} .
- A popular method to choose h is to use leave-one-out cross-validation, i.e.,

$$h_{CV} = \arg \min_h \frac{1}{n} \sum_{i=1}^n \delta(x_i) \left[y_i - \sum_{j \neq i} \omega_h(x_i - x_j) y_j \right]^2$$

where $\delta(x)$ is a trimming function to reduce boundary effects.

- This can be generalized for the case where we condition on a **vector of regressors**.
- However, for the multivariate case, not only do we have to face the “**Curse of Dimensionality**”, but we also lose the ability to display the results **graphically**.

2) Semiparametric estimators

- Most non-ML estimators are semiparametric in the sense that they do not require the full specification of the conditional probability distribution of the dependent variable.
- Later we will see examples of semiparametric estimators in the context of some specific microeconomic models.
- Here, we will look at a useful semi(non)parametric estimator: The partially linear model.
- Suppose that $E(y_i|z, x) = z_i\beta + g(x_i)$, where $g(\cdot)$ is left unspecified, and

$$y_i = z_i\beta + g(x_i) + \epsilon_i \quad ; \quad i = 1, \dots, n.$$

- Robinson (1988) and Speckman (1988) have shown that, under certain conditions, it is possible to obtain \sqrt{n} -consistent estimators for β .
- Notice that we can write

$$E(y_i|x) = E(z_i|x)\beta + g(x_i)$$

$$y_i - E(y_i|x) = [z_i - E(z_i|x)]\beta + \epsilon$$

- If $E(y_i|x)$ and $E(z_i|x)$ where known, β could be estimated by OLS.
- However, if $E(y_i|x)$ and $E(z_i|x)$ are replaced by kernel estimators that converge sufficiently quickly, their use will not affect the rate of convergence of the OLS estimator of β , which, using an obvious notation, can be estimated by (Yatchew, 2003)

$$\hat{\beta} = \left[\left(Z - \widehat{E}(Z|x) \right)' \left(Z - \widehat{E}(Z|x) \right) \right]^{-1} \left(Z - \widehat{E}(Z|x) \right)' \left(Y - \widehat{E}(Y|x) \right)$$

with

$$\sqrt{n} \left(\hat{\beta} - \beta \right) \xrightarrow{d} \mathcal{N} \left(0, A^{-1} B A^{-1} \right).$$

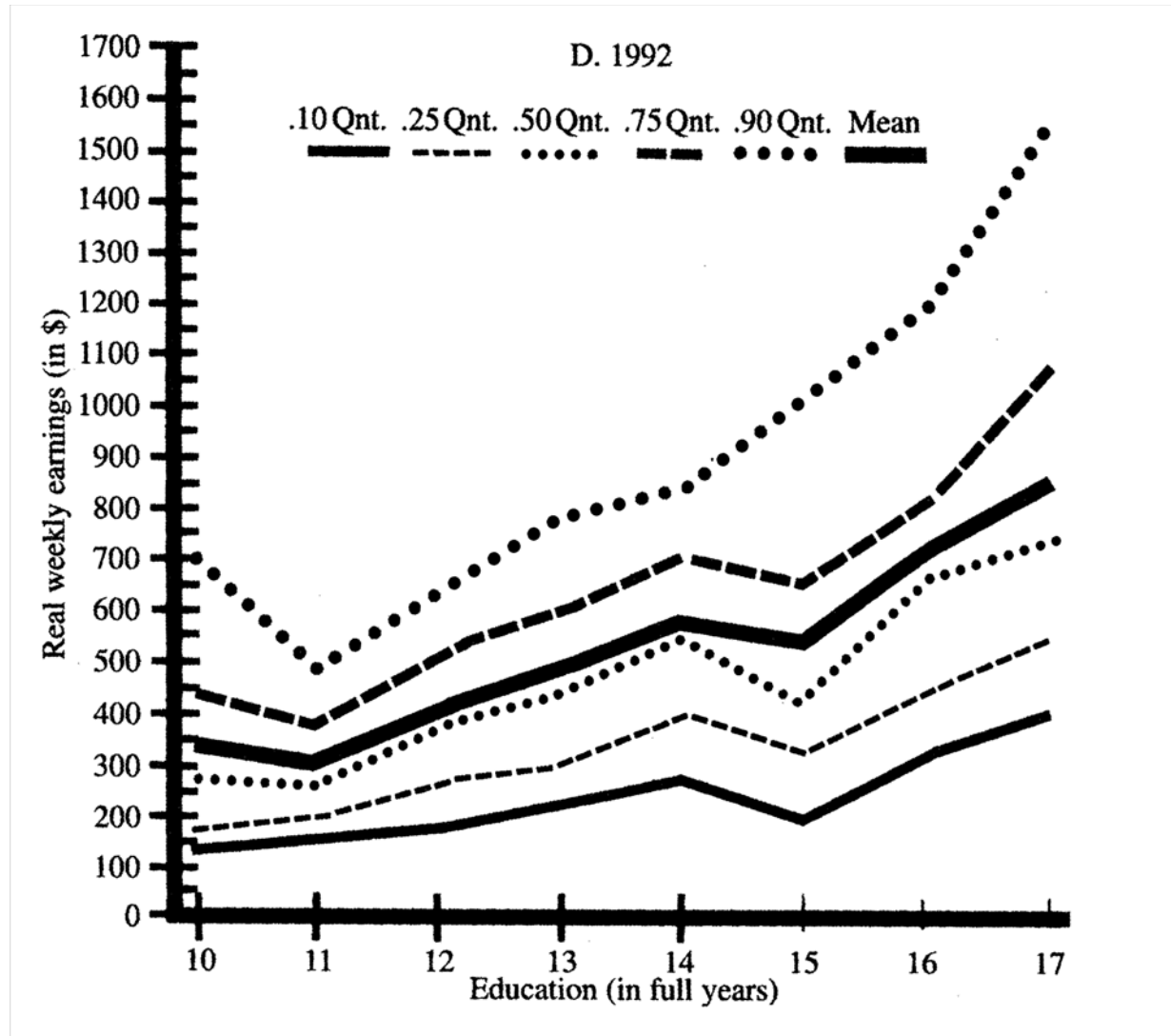
- If the errors are iid,

$$A = \lim_{n \rightarrow \infty} \frac{1}{n} \left[\left(Z - \widehat{E}(Z|x) \right)' \left(Z - \widehat{E}(Z|x) \right) \right]; \quad B = \sigma_{\epsilon}^2 A.$$

- If the errors are independent but heteroskedastic, the usual “sandwich” can be used with

$$B = \lim_{n \rightarrow \infty} \frac{1}{n} \left[\left(Z - \widehat{E}(Z|x) \right)' \Omega \left(Z - \widehat{E}(Z|x) \right) \right].$$

3) Quantile Regression



- For $0 < \alpha < 1$, the α -th quantile of y given x is defined by

$$Q_y(\alpha|x) = \min\{\eta | P(y \leq \eta|x) \geq \alpha\}.$$

- Assume that $Q_y(\alpha|x)$ is linear, so that $Q_y(\alpha|x_i) = x_i' \beta(\alpha)$, which is equivalent to

$$y_i = x_i' \beta(\alpha) + u(\alpha)_i; \quad Q_{u(\alpha)}(\alpha|x_i) = 0$$

- For $\alpha = 0.5$ we have the median regression: $\text{med}(y|x) = x' \beta(0.5)$.
- It is interesting to notice that quantiles are equivariant to monotonic transformations so that if $T(\cdot)$ is a monotonic function,

$$Q_y(\alpha|x_i) = T(x_i' \beta(\alpha))$$

implies

$$Q_{T^{-1}(y)}(\alpha|x_i) = x_i' \beta(\alpha).$$

- The estimator of $\beta(\alpha)$ is defined by (which is LAD for $\alpha = 0.5$)

$$\hat{\beta}(\alpha) = \arg \min_b \frac{1}{n} \left\{ \sum_{i:y_i \geq x'_i b} \alpha |y_i - x'_i b| + \sum_{i:y_i < x'_i b} (1 - \alpha) |y_i - x'_i b| \right\}$$

- This is equivalent to

$$\hat{\beta}(\alpha) = \arg \min_b \frac{1}{n} \sum_{i=1}^n \rho_\alpha(u(\alpha)_i),$$

where $\rho_\alpha(a) = a(\alpha - I(a < 0))$ is the so-called *check function* (\checkmark), and to

$$\hat{\beta}(\alpha) = \arg \min_b \frac{1}{n} \sum_{i=1}^n (\alpha - 0.5 + 0.5 \operatorname{sgn}(y_i - x'_i b)) (y_i - x'_i b)$$

- Notice that the F.O.C. are given by

$$\frac{1}{n} \sum_{i=1}^n (\alpha - 0.5 + 0.5 \operatorname{sgn}(y_i - x'_i b)) x_i = 0$$

- So, the estimator is invariant to perturbations of y that do not change the sign of $(y_i - x'_i b)$.

- $\hat{\beta}(\alpha)$ is usually estimated by **linear programming** methods.
- Asymptotic theory is **not standard** because the objective function is not differentiable.
- It is possible to show that, under certain regularity conditions, $\hat{\beta}(\alpha)$ is consistent with

$$\sqrt{n} \left(\hat{\beta}(\alpha) - \beta(\alpha) \right) \xrightarrow{d} \mathcal{N} \left(0, D^{-1} A D^{-1} \right)$$

$$D = E \left[f_{u(\alpha)}(0|x_i) x_i x_i' \right] \quad A = \alpha(1 - \alpha) E \left[x_i x_i' \right]$$

- Notice that A is always easy to estimate in one of the following ways

$$\hat{A} = \alpha(1 - \alpha) \frac{1}{n} \sum_{i=1}^n x_i x_i', \quad \tilde{A}_q = \frac{1}{n} \sum_{i=1}^n (\alpha - \mathbf{1}(\hat{u}(\alpha)_i \leq 0))^2 x_i x_i'.$$

- For D , Powell (1984) suggested (other kernels can be used in place of the uniform kernel)

$$\hat{D} = \frac{1}{nh} \sum_{i=1}^n \mathbf{1} \left(-0.5 \leq \frac{\hat{u}(\alpha)_i}{h} \leq 0.5 \right) x_i x_i',$$

- If $u(\alpha)_i$ and x_i are independent then $f_{u(\alpha)}(0|x_i) = f_{u(\alpha)}(0)$ for all x in which case and the covariance matrix simplifies to

$$\left(\frac{\alpha(1-\alpha)}{[f_{u(\alpha)}(0)]^2} \right) [E(xx')]^{-1}.$$

- In this case, $f_{u(\alpha)}(0)$ may be estimated by kernel density estimation.
- Given an estimate of the covariance matrix, standard Wald tests on quantile coefficients can be performed.
- As usual, Wald test are easy to implement but are not invariant to reformulation of non-linear hypotheses.
- Alternative inference procedures are available, but are more cumbersome.
- Quantile regression also allows us to test hypotheses about the **characteristics of the conditional distribution.**

- For instance, independence between errors and regressors can be tested by checking whether the slopes are the same for all quantiles (Koenker and Bassett, 1982).
- A comparable test based on **expectiles** was proposed by Newey and Powell (1987).
- A simpler test for a closely related hypothesis can be performed as follows (Machado and Santos Silva, 2000):
 - regress $\rho_\alpha(\hat{u}(\alpha)_i)$ on a constant and functions of x ;
 - compute the test statistic as n times the R^2 from the auxiliary regression;
 - Compare the test statistic to the appropriate critical value from the $\chi^2_{(J)}$, where J is the number of functions of x included in the auxiliary regression.
- For $\alpha = 0.5$, this is the well known Glejser (1969) test for heteroskedasticity.

4) Bootstrap

- In most cases, statistical tests have to be performed by comparing the relevant test statistic with a critical value obtained from its asymptotic distribution.
- For finite samples, even large ones, tests performed in this way may have a size which is very different from the nominal significance level.
- In some important cases, the bootstrap may be used to obtain tests procedures where the empirical and nominal significance levels are much closer that when using critical values from the asymptotic distribution.
- Bootstrap is a resampling method that uses the original sample to generate new (bootstrap) samples, from which quantities of interest can be computed.
- The properties of the distribution of interest can then be studied by studying the corresponding bootstrap distribution.
- As we will see later, bootstrap samples can be **generated in different ways**, depending on the problem in hand.
- Bootstrap is often used to estimate the covariance matrices (although this is not a good use of the bootstrap!).

- Suppose that we have estimated $\hat{\beta}$ and, by using one of the methods to be presented later, B bootstrap samples of size n are generated. Then:
 - for sample b , $1 \leq b \leq B$, the parameters $\hat{\beta}_b$ can be estimated;
 - the covariance of $\hat{\beta}$ is estimated by

$$\frac{1}{B} \sum_{b=1}^B \left(\hat{\beta}_b - \hat{\beta} \right) \left(\hat{\beta}_b - \hat{\beta} \right)' .$$

- **Bootstrap is better used** to perform statistical tests or to construct confidence intervals.
- The advantages of this are:
 - (a) It is possible to perform inference about quantities whose distribution **is unknown** (even asymptotically);
 - (b) If the quantities involved are asymptotically pivotal, the bootstrap procedures have **improved finite sample properties**.
 - * A **pivotal** statistic is one whose distribution does not depend on unknown quantities;
 - * If a statistic is exactly pivotal, exact tests can be performed by simulation.

- When **asymptotic critical** values are used, the empirical size of the test with nominal significance level α is $\alpha + O(n^{-1})$.
- For asymptotically **pivotal** test statistics, when **bootstrap** critical values are used, the empirical size of the test with nominal significance level α is $\alpha + O(n^{-2})$.
- To perform a bootstrap test:
 - (a) A collection of bootstrap tests statistics τ_1, \dots, τ_B is generated **under the null**;
 - (b) For one-sided tests, the bootstrap p-value is computed by comparing the original test statistic, τ_0 , with the ordered bootstrap statistics

$$P_B = \frac{\sum_{b=1}^B \mathbf{1}(\tau_b \geq \tau_0) + 1}{B + 1}.$$

- The value of B should be chosen so that $\alpha(B + 1)$ is an integer. That is, for a given α , B has to be at least $\alpha(B + 1) \geq 1 \Rightarrow \frac{1}{\alpha} - 1$.

- How to generate bootstrap samples?
- For simplicity, consider the model $y_i = x_i'\beta + u_i$ and suppose we want to test an hypothesis about β .
- Let $\bar{\beta}$ denote the estimate of β under the null and set $\bar{u}_i = y_i - x_i'\bar{\beta}$.
- **Residual bootstrap:** denoting by u_i^* the bootstrap errors, generate the dependent variable as

$$y_i^* = x_i'\bar{\beta} + u_i^*$$

- Different ways of generating the bootstrap errors are available:
 - (a) **Parametric residual bootstrap:** if the distribution of the errors is known, bootstrap errors can be generated from it.
 - (b) **Semiparametric residual bootstrap:** if the errors are IID, bootstrap errors can be obtained as draws with replacement from the vector of estimation residuals \bar{u}_i .
 - * N.B.: A smoothed version can be used. This is done by jittering the residuals in \bar{u}_i .

(c) **Block-bootstrap** (a variant of semiparametric residual bootstrap): in case the errors are serially correlated, we draw blocks of observations of the vector of estimation residuals to preserve the dynamic dependence (at least asymptotically).

(d) **Wild-bootstrap** (a variant of semiparametric residual bootstrap): in presence of heteroskedasticity, to preserve the link between the errors and regressors, the bootstrap error for observation i can be obtained as $u_i^* = f(\bar{u}_i) \varepsilon_i^b$, where

* $f(\bar{u}_i)$ is an appropriate function of the residuals, e.g., the identity function;

* ε_i^b is a random variable independent of u and x , such that $E(\varepsilon_i^b) = 0$ and $Var(\varepsilon_i^b) = 1$;

* a possibility is to use the Rademacher distribution for ε_i^b :

$$F_1 : \varepsilon_i^b = \begin{cases} -1 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2} \end{cases}$$

- **Pairs (case-resampling) bootstrap:** Denoting by \hat{u}_i the residuals from the unrestricted regression, resample the pairs (x_i, \hat{u}_i) and generate the dependent variable as

$$y_i^* = x_i^{*'} \bar{\beta} + \hat{u}_i^*$$

- Notice that u_i is not required to be independent of x_i because their link is maintained in the bootstrap sample.
- Inference with the pairs-bootstrap **is marginal** in x .
- An alternative version of the pairs-bootstrap can be obtained by resampling the (y_i, x_i) pairs themselves.
- In this case, it is not possible to impose the null on the bootstrap data generating process and therefore we have to replace the test that is performed.
- Rather than testing $H_0 : R\beta = r$, which is invalid in the bootstrap data, one can test $H_0^* : R\beta = R\hat{\beta}$ (where $\hat{\beta}$ is the unrestricted estimator of β), which is valid in the bootstrap data.

University week: 9

Discrete Choice Models

1. Binary choice models, logit and probit;
2. Extensions of binary choice models;
3. Multinomial logit and probit.

Recommended reading: Greene: 15.2, 15.6, 17.1–17.4, 18.1.

1) Binary choice models, logit and probit

- In many applications, the variate of interest is binary, i.e., takes only the values 0 and 1.
- Examples include labour force participation, house ownership and passing an exam.
- We can consider three different frameworks to interpret models for this sort of data.

(a) The (nonlinear) regression approach:

- * Notice that $E(y|x) = \Pr(y = 1|x)$, which completely characterizes $f(y|x)$.
- * In this case, linear regression is not appropriate since it is unlikely that the marginal effect of x on y is constant.

(b) Partial observability:

- * Suppose that

$$y_i^* = x_i' \beta + u_i, \quad y_i = \begin{cases} 0 & y_i^* \leq 0 \\ 1 & y_i^* > 0 \end{cases}$$

where y_i^* is unobservable. In this case, **up to scale**, β can be estimated from

$$\Pr(y_i = 1|x) = \Pr(u_i > -x_i' \beta | x).$$

(c) Random utility models

- * Suppose that an individual has to choose between alternatives a and b , with utilities U^a and U^b .
- * The researcher does not observe the utilities, but observes some characteristics of the alternatives, and writes

$$U^a = x'_a \beta + u_a,$$

$$U^b = x'_b \beta + u_b.$$

- * The researcher observes the chosen alternative, say a , which is indicated by $y = 1$.
- * Then, we know that

$$\begin{aligned} \Pr(y_i = 1|x) &= \Pr(U^a > U^b|x) = \Pr(x'_a \beta + u_a > x'_b \beta + u_b|x) \\ &= \Pr(u_a - u_b > -(x'_a - x'_b) \beta|x). \end{aligned}$$

- Whatever the interpretation, we have to make inference about $\Pr(y_i = 1|x)$.

- Suppose that we have a sample $\{y_i, x_i\}_{i=1}^n$ of n independent observations, and that for each i the conditional probability of y_i given x_i has the form $\Pr(y_i = 1|x_i) = p(x_i; \theta) = p_i$.
- Often, the form of $p(x_i; \theta)$ is assumed to be known. Examples include,
 - the linear probability model: $p_i = x_i'\theta$;
 - the logit model: $p_i = \Lambda(x_i'\theta)$, where $\Lambda(z) = e^z / (1 + e^z)$;
 - the probit model: $p_i = \Phi(x_i'\theta)$, where $\Phi(\cdot)$ is the standard normal cdf.
- Because the $E(y_i|x_i)$ defines $\Pr(y_i = 1|x_i)$, estimation is often likelihood based.
- Since, given x_i , the y_i 's are independent Bernoulli random variables, the likelihood function has the form

$$L(\theta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{(1-y_i)} .$$

- Hence, the log-likelihood is

$$\ln L(\theta) = \sum_{i=1}^n [y_i \ln p_i + (1 - y_i) \ln (1 - p_i)].$$

- It is useful to obtain the score, the Hessian and the (sum of the individual) Information Matrix, which are given by

$$g(\theta) = \sum_{i=1}^n \left[\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right] \frac{\partial p_i}{\partial \theta} = \sum_{i=1}^n \left[\frac{y_i - p_i}{p_i (1 - p_i)} \right] \frac{\partial p_i}{\partial \theta}$$

$$H(\theta) = \sum_{i=1}^n \left[\frac{y_i - p_i}{p_i (1 - p_i)} \right] \left[\frac{\partial^2 p_i}{\partial \theta \partial \theta'} \right] - \sum_{i=1}^n \left[\frac{y_i - p_i}{p_i (1 - p_i)} \right]^2 \left[\frac{\partial p_i}{\partial \theta} \right] \left[\frac{\partial p_i}{\partial \theta} \right]'$$

$$\sum_{i=1}^n \mathcal{I}_i(\theta) = -E(H(\theta)) = E \left(\sum_{i=1}^n g_i(\theta) g_i(\theta)' \right) = \sum_{i=1}^n \frac{1}{p_i (1 - p_i)} \left[\frac{\partial p_i}{\partial \theta} \right] \left[\frac{\partial p_i}{\partial \theta} \right]'$$

- It is worth noting that $p_i (1 - p_i) = \text{Var}(y_i | x_i)$, which implies that ML is GLS.

- **Logit and Probit**

- It can be shown that log-likelihood function for the logit and probit models is globally **concave**.
- With appropriate distributional assumptions about the error terms, the both models fit into the partial observability and random utility frameworks.
- The logit and probit are very similar: both $\Lambda(x'_i\beta)$ and $\Phi(x'_i\beta)$ are symmetric and the two functions are almost indistinguishable.
- Notice, however, that the parameters are measured in **different scales**.

- Here are some examples of other parametric models:

(i) $p_i = \exp(-\exp(x'_i\beta))$

(ii) $p_i = 1 - \exp(-\exp(x'_i\beta))$

(iii) $p_i = \Phi(x'_i\beta)^\tau, \quad \tau > 0$

(iv) $p_i = 1 - (1 + \omega \exp(x'_i\beta))^{-1/\omega}, \quad \omega > 0$ (for $\omega = 1, p_i = \Lambda(x'_i\beta)$ and $\lim_{\omega \rightarrow 0} p_i = 1 - \exp(-\exp(x'_i\beta))$)

- Generally, in nonlinear models, parameters are not directly interpretable.
- This difficulty can be overcome by computing partial effects.
- In binary choice models, partial effects are defined as:

(a) If x_{ij} is a continuous regressor

$$pe = \frac{\partial p_i}{\partial x_{ij}} = \beta_j \frac{\partial p}{\partial x' \beta}.$$

(b) If x_{ij} is a discrete regressor

$$pe = p_i(x_{ij} = 1) - p_i(x_{ij} = 0).$$

- As usual, standard errors can be computed using the delta-method. If $F(\theta)$ is a continuous function, then

$$\text{Var} F(\hat{\theta}) \approx \left[\frac{\partial F(\theta)}{\partial \theta} \right]' \text{Var}(\hat{\theta}) \left[\frac{\partial F(\theta)}{\partial \theta} \right]$$

- If p_i is misspecified, the ML estimator has the usual interpretation.
- However, misspecification may have severe consequences.
- Some simple specification tests are available:
 - (a) A **RESET**-type test can be performed by checking the significance of δ_1 and δ_2 in the model

$$p \left(x'_i \beta + \delta_1 \left(x'_i \hat{\beta} \right)^2 + \delta_2 \left(x'_i \hat{\beta} \right)^3 \right)$$

* This is actually a normality test in the probit.

- (b) The model can be tested against **more general** parametric specifications, which include additional shape parameters (see (iii) and (iv) above).
- (c) Tests against **non-nested** alternatives can be performed by testing $\alpha = 0$ or $\alpha = 1$ in the model

$$p_i = (1 - \alpha) p_i^1 + \alpha p_i^2$$

- Finally, as usual, measures of **goodness-of-fit** are not so useful.

2) Extensions of binary choice models

(a) Klein and Spady's nonparametric maximum likelihood estimator

- Recall that the log-likelihood for a binary choice model has the form

$$\ln L(\theta) = \sum_{i=1}^n [y_i \ln p_i + (1 - y_i) \ln (1 - p_i)].$$

- Note also that $p_i = E(y_i|x)$.
- Assuming that p_i depends on x only through $x'_i\beta$ (**single index** assumption), the likelihood function can be maximized without specifying the functional form of p_i .
- This is because p_i can be estimated nonparametrically as a function of the index $x'_i\beta$.

- In particular, Klein and Spady (1993) suggested that β can be estimated by maximizing

$$\sum_{i=1}^n \zeta_i [y_i \ln \hat{p}_i + (1 - y_i) \ln (1 - \hat{p}_i)].$$

where \hat{p}_i is a non parametric estimate of $E(y_i|x'_i\beta)$ and ζ_i is a trimming function to reduce boundary effects.

- Notice that the slope parameters are identified up to **scale**.
- Also, the **intercept** is embedded in the form of \hat{p}_i , and therefore is not identified.
- The standard **normalization** is to set to 1 the coefficient of a regressor known to have a positive slope parameter.
- Identification requires that at least one **continuous regressor** has non-zero coefficient.
- The KS estimator is \sqrt{n} -consistent, asymptotically normal, and attains the semiparametric **efficiency bound**.
- The single index assumption is, however, somewhat restrictive; for instance, it is generally incompatible with heteroskedasticity.

(b) Manski's (1975) maximum score estimator

- This method estimates the conditional median of y_i^* rather than the conditional mean.
- Under conditional symmetry the mean and median coincide.
- If $Q_{y_i^*}(0.5|x_i) = x_i'\beta$, β can be estimated as

$$\hat{\beta} = \arg \max_{\beta: \beta'\beta=1} S_n(\beta),$$

$$S_n(\beta) = \sum_{i=1}^n [y_i I(x_i'\beta > 0) + (1 - y_i)(1 - I(x_i'\beta > 0))].$$

- This can be thought of as maximizing the number of observations for which the sign of $x_i'\beta$ correctly indicates whether $y_i = 1$ or $y_i = 0$.
- $S_n(\beta)$ is a step function, leading to multiple maxima and making it hard to maximize.
- The estimator is consistent but, because y_i is discrete and the objective function is non-differentiable, $\hat{\beta}$ has a non-normal limiting distribution with a $n^{1/3}$ rate of convergence.

- In 1992, Horowitz proposed an alternative method, which overcomes some of the problems with Manski's estimator.

- Notice that

$$S_n(\beta) = \sum_{i=1}^n [2y_i I(x'_i \beta > 0) - I(x'_i \beta > 0) + 1 - y_i]$$

- Therefore, the estimator can be defined as

$$\hat{\beta} = \arg \max_{\beta: \beta' \beta = 1} \sum_{i=1}^n (2y_i - 1) I(x'_i \beta > 0)$$

- Horowitz's idea was to replace $I(x'_i \beta > 0)$ by a smooth function

$$K\left(\frac{x'_i \beta}{\delta_n}\right)$$

such that, $\lim_{\nu \rightarrow -\infty} K(\nu) = 0$; $\lim_{\nu \rightarrow \infty} K(\nu) = 1$; $\lim_{n \rightarrow \infty} \delta_n = 0$.

- Notice that $K(\nu)$ is not a kernel function as before, but an integrated kernel.

- This leads to

$$\tilde{\beta} = \arg \max_{\beta: \beta' \beta = 1} \sum_{i=1}^n (2y_i - 1) K \left(\frac{x_i' \beta}{\delta_n} \right)$$

- Clearly, $\tilde{\beta}$ and $\hat{\beta}$ converge to the same limit, but $\tilde{\beta}$ is asymptotically **normal** and its convergence rate is at least $n^{2/5}$ and can be made **arbitrarily close** to $n^{1/2}$, provided that the problem is sufficiently smooth.
- The choice of δ_n is determinant for the performance of the method.
- The choice of $K(\cdot)$ is important, but not as much.
- The performance is only reasonable in (very) large samples.
- Both the maximum score estimator and its smooth version can be adapted to the estimation of other quantiles.
- The **advantage** of the (smooth) maximum score estimator is that it does not require the single-index assumption. However, it has the **drawback** of not permitting the estimation of probabilities or marginal effects.

(c) Choice based sampling

- In some cases it is convenient to work with a sample in which the percentage of observations with $y_i = 1$ is defined by the researcher.
- In these cases, the models have to be adapted in order to perform valid inference.
- As usual, let $\Pr(y_i = 1|x) = p_i$ and define $p = \Pr(y_i = 1) = E_x[p_i(x)]$.
- Therefore the marginal probability density function of y is given by

$$f(y) = p^{y_i} (1 - p)^{1-y_i}$$

- Suppose that individuals are sampled from the population in such a way that there is a probability π that an individual with $y_i = 1$ will enter the sample.
- Then, the sample can be viewed as a random draw from a population with

$$f_s(y) = \pi^{y_i} (1 - \pi)^{1-y_i}$$

- It is possible to show that

$$P_s(y = 1|x) = \frac{p_i^{\frac{\pi}{p}}}{\frac{(1-\pi)}{(1-p)} (1 - p_i) + p_i^{\frac{\pi}{p}}}$$

from where the parameters of p_i can be estimated by maximum (conditional) likelihood.

- In the case of the logit the result is very convenient:

$$P_s(y = 1|x) = \frac{\exp(x'_i\beta)^{\frac{\pi}{p}}}{\frac{(1-\pi)}{(1-p)} 1 + \exp(x'_i\beta)^{\frac{\pi}{p}}} = \frac{\exp(x'_i\beta + \gamma)}{1 + \exp(x'_i\beta + \gamma)}$$

$$\gamma = \log \left(\frac{(1 - \pi) p}{(1 - p) \pi} \right)$$

- Therefore, in the case of the logit, only the intercept is biased under choice based sampling.

(d) Estimation with panel data

(d.1) If we are willing to specify $\Pr(y_{it} = 1|x_{it}) = F(x'_{it}\beta)$, β can be consistently estimated without specifying the distribution of $y_i \equiv (y_{i1}, \dots, y_{iT})$ given $x_i \equiv (x_{i1}, \dots, x_{iT})$ using **pooled maximum-likelihood estimation**.

- **Robust** covariance matrix has to be used to account for correlation over time.

(d.2) Under **strict-exogeneity**, efficiency can be gained by using **GLS** (population average).

(d.3) A popular alternative is the **random-effects** probit model. Using the usual panel data notation, let y_{it} be defined by

$$y_{it} = \mathbf{1}(x'_{it}\beta + \alpha + c_i + u_{it} > 0) \quad c_i|x_i \sim \mathcal{N}(0, \sigma_c^2).$$

- Defining $\beta_c \equiv \beta / \sqrt{1 + \sigma_c^2}$, the random-effects probit model is obtained by assuming also:

Strict-exogeneity: $\Pr(y_{it} = 1|x_i, c_i) = \Pr(y_{it} = 1|x_{it}, c_i) = \Phi(x'_{it}\beta + \alpha + c_i)$,

Independence: $y_{i1} \dots y_{iT}$ are independent conditional on x_i, c_i .

- As before, estimation is by (pseudo) ML and robust covariance matrix has to be used.
- To allow the unobservables to be correlated with the regressors, we may use the **Mundlak-Chamberlain** formulation

$$c_i | x_i \sim \mathcal{N}(\bar{x}_i' \xi, \sigma_a^2).$$

- Naturally, coefficients associated to regressors that do not vary in time are not identified.

(d.4) If, conditional on c_i and x_i , the elements of y_i are assumed to be **independent**, estimates of β and σ_c^2 (or σ_a^2) can be obtained by explicitly integrating-out the unobservables from the likelihood function, e.g.,

$$L(y_i) = \int \left[\prod_{t=1}^T \Phi(x'_{it}\beta + \alpha + c_i)^{y_{it}} (1 - \Phi(x'_{it}\beta + \alpha + c_i))^{(1-y_{it})} \right] \frac{1}{\sigma_c} \phi\left(\frac{c_i}{\sigma_c}\right) dc$$

- The estimator can be adapted to be used with different distributional assumptions.

(d.5) Treating c_i as parameters leads to an inconsistent estimator of β due to the **incidental parameters** problem (except in the logit).

(d.6) Maintaining strict-exogeneity and conditional independence of the elements of y_i it is possible to allow c_i and x_i to be arbitrarily related in a **logit model**.

- This is done by **conditioning** on $\sum_{t=1}^T y_{it}$, which eliminates the individual effects from the objective function.
- However, because c_i is conditioned out, it is **impossible** to estimate probabilities and compute marginal effects.
- This is a very serious limitation in practice.
- Finally, recall that this “fixed-effects” logit estimator requires **conditional independence** and depends on the **logit distribution**.

3) Multinomial logit and probit

- Consider now the case of **multinomial choice models**, i.e., models with more than two (unordered) discrete outcomes.
- Suppose that y can take the values $1, \dots, J$, where J is some predefined fixed integer, and x^* is a vector of explanatory variables. Then we model $\{\Pr(y_i = j|x_i^*, \theta)\}_{j=1}^J$.
- The log-likelihood function for the multinomial choice model is given by

$$\ln L(\theta) = \sum_{i=1}^n \sum_{j=1}^J \mathbf{1}(y_i = j) \ln \Pr(y_i = j|x_i^*, \theta).$$

- The simplest specification is the **multinomial logit**, with

$$\Pr(y_i = j|x_i^*, \theta) = \frac{\exp(x'_{ij}\beta)}{\sum_{l=1}^J \exp(x'_{il}\beta)}, \quad j = 1, \dots, J.$$

- For identification, we set $x'_{ij}\beta = 0$, for some alternative j (usually the first or the last).

- The multinomial logit can be obtained from a random utility model of the form

$$U_i^j = x'_{ij}\beta + \epsilon_{ij},$$

where the ϵ_{ij} have independent **Type-I extreme value** distributions with common cdf given by (up to a common scaling factor)

$$f(\epsilon) = \exp[-\exp(-\epsilon)].$$

- This model has the **Independence of Irrelevant Alternatives (IIA)** property which means that for any pair (j, l) the odds ratio

$$\frac{\Pr(y_i = j | x_i^*, \theta)}{\Pr(y_i = l | x_i^*, \theta)} = \frac{\exp(x'_{ij}\beta)}{\exp(x'_{il}\beta)}$$

does not depend on the characteristics or availability of any other options.

- This implies that the elasticity of the probability of choosing j with respect to changes in the value of a characteristic of i is the same for any $j \neq i$.
- This property characterizes the logit and is a consequence of the distributional assumptions.

- A specification that does not impose the IIA is the **multinomial probit** model defined by a random utility model

$$U_i^j = x'_{ij}\beta + \epsilon_{ij},$$

where the disturbances $\{(\epsilon_{i1}, \dots, \epsilon_{iJ})'\}_{i=1}^n$ are iid $\mathcal{N}_J(0, \Sigma_\epsilon)$.

- Location and scale normalization implies that Σ_ϵ has rank $J - 1$ and is estimated up to scale.
- Normalizing with respect to the first alternative, only the covariance of $\epsilon_{ij}^* = \epsilon_{ij} - \epsilon_{i1}$, $j = 2, \dots, J$, is estimated.
- Moreover, one of the parameters has to be fixed to impose scale normalization.
- That is, only $J(J - 1)/2 - 1$ covariance parameters can be identified.
- Without exclusion restrictions, identification is often difficult.
- Estimation requires evaluation of multivariate normal probabilities, which can be done by simulation, but even that is expensive for moderate to large J .

- Currently, a very popular specification for multinomial choice models is the kernel, mixed or random parameter logit, where

$$\Pr(y_i = j | x_i^*, \beta_i) = \frac{\exp(x'_{ij}\beta_i)}{\sum_{l=1}^J \exp(x'_{il}\beta_i)}, \quad j = 1, \dots, J.$$

- Because β_i is not observable, the probability of interest is

$$\Pr(y_i = j | x_i^*) = \int \frac{\exp(x'_{ij}\beta_i)}{\sum_{l=1}^J \exp(x'_{il}\beta_i)} f(\beta_i) d\beta_i$$

- Under standard assumptions, estimation can be performed by maximizing

$$\ln L(\theta) = \sum_{i=1}^n \sum_{j=1}^J \mathbf{1}(y_i = j) \ln \left[\int \frac{\exp(x'_{ij}\beta_i)}{\sum_{l=1}^J \exp(x'_{il}\beta_i)} f(\beta_i) d\beta_i \right].$$

- In practice, $L(\theta)$ is evaluated by simulation, replacing the integral with

$$P(y_i = j | x_i^*) = \frac{1}{D} \sum_{r=1}^D \frac{\exp(x_{ij}^* b_r)}{\sum_{l=1}^J \exp(x_{il}^* b_r)},$$

where b_r is a random draw from $f(\beta_i)$, the assumed distribution of β_i .

- $P(y_i = j | x_i^*, \beta_i)$ is an unbiased estimator of $\Pr(y_i = j | x_i^*, \beta_i)$.
- Notice that we do not estimate β_i , but the parameters of $f(\beta_i)$.
- Notice also that consistency depends on the correct specification of $f(\beta_i)$.
- Estimation is complicated by the fact that the likelihood function is not strictly concave and therefore multiple maxima are possible.
- The choice of D depends on the nature of the problem and on how the draws are obtained: pseudo-random or quasi-random (Halton). In any case, it is vital to ensure that the parameter estimates are stable when the value of D is increased.

University week: 10

Ordered Data and Count Data Models

1. Ordered data
2. The Poisson Regression Model
3. Testing for Overdispersion
4. Heterogeneity and the Negative Binomial Regression Model
5. Hurdle and Zero-Altered Poisson Models
6. Models for Panel Data

Recommended reading: Greene: 18.3, 18.4.

1) Ordered data

- In some problems, the variate of interest assumes more than two discrete outcomes, but these are inherently ordered.
- An example are the results of surveys on the degree of satisfaction with some service.
- This kind of data can be modelled using the following latent variable framework

$$y_i^* = x_i' \beta + u_i, \quad y_i = \begin{cases} 1 & y_i^* \leq \mu_1 \\ 2 & \mu_1 < y_i^* \leq \mu_2 \\ 3 & \mu_2 < y_i^* \leq \mu_3 \\ \vdots & \vdots \\ J & \mu_{J-1} < y_i^* \end{cases}$$

where the threshold parameters are such that $0 = \mu_1 < \mu_2 < \dots < \mu_{J-1}$.

- If the distribution of u_i is specified, the unknown parameters β and μ_2, \dots, μ_{J-1} can be estimated by maximum likelihood.

- Notice that

$$\Pr(y_i = 1|x_i) = \Pr(x_i'\beta + u_i \leq 0) = \Pr(u_i \leq -x_i'\beta)$$

$$\Pr(y_i = 2|x_i) = \Pr(0 < x_i'\beta + u_i \leq \mu_2) = \Pr(u_i \leq \mu_2 - x_i'\beta) - \Pr(u_i < -x_i'\beta)$$

⋮

$$\Pr(y_i = J|x_i) = \Pr(\mu_{J-1} - x_i'\beta < u_i) = 1 - \Pr(u_i < \mu_{J-1} - x_i'\beta)$$

- Therefore, the log-likelihood function is simply

$$\ln L(\theta) = \sum_{i=1}^n \sum_{j=1}^J \mathbf{1}(y_i = j) \ln [\Pr(u_i \leq \mu_j - x_i'\beta) - \Pr(u_i < \mu_{j-1} - x_i'\beta)]$$

- As in all discrete choice models, the variance of u_i is not identified.
- The ordered-probit and ordered-logit are the most used special cases of this model.

- For the ordered-probit

$$\Pr(u_i \leq \mu_j - x_i'\beta) = \Phi(\mu_j - x_i'\beta)$$

- For the ordered-logit

$$\Pr(u_i \leq \mu_j - x_i'\beta) = \frac{\exp(\mu_j - x_i'\beta)}{1 + \exp(\mu_j - x_i'\beta)}$$

- Interpretation of the coefficients is not obvious, except within the latent variable formulation.
- A closely related model can be used for **grouped data**.
 - In this case, the threshold parameters are the limits of the intervals.
 - The main difference is that, for $J > 1$, the variance of u_i is identified because the thresholds give information on the scale of u_i .
 - **Additional flexibility** (at a cost) is obtained if the limits of the intervals are estimated.
- Alternatively, use **sequential binary models** for $\Pr(y_i = j | x_i, y_i \geq j), j = 1, \dots, J - 1$.

2) The Poisson Regression Model

- In many relevant applications, the variate of interest is the count of the number of occurrences of some event in a given period of time.
- Examples include: accidents, patents, takeovers, purchases, doctor visits, jobs and trips.
- These data have some very specific characteristics:
 - Discreteness;
 - Bounded support;
 - Many zeros and a long right-hand tail.
- In this context, standard linear models are not appealing because:
 - The conditional expectation is necessarily non-linear and positive;
 - The data is intrinsically heteroskedastic;
 - Do not allow the computation of the probability of events of interest.

- The basic model for count data is the Poisson regression, defined by

$$\Pr(y_i = j|x_i) = \frac{\exp(-\lambda(x_i, \beta)) \lambda(x_i, \beta)^j}{j!}, \quad j = 0, 1, 2, \dots$$

$$E(y_i|x_i) = \text{Var}(y_i|x_i) = \lambda(x_i, \beta)$$

- Notice, however, that

$$\text{Var}(y_i) = E_x [\lambda(x_i, \beta)] + \text{Var}_x [\lambda(x_i, \beta)] \geq E_x [\lambda(x_i, \beta)] = E(y_i).$$

- In general, the following specification is adopted: $\lambda(x_i, \beta) = \exp(x_i' \beta)$.

- Therefore,

$$\frac{\partial E(y_i|x_i)}{\partial x_i} = \lambda(x_i' \beta) \beta$$

- ML estimation of β is straightforward.

- The log-likelihood function, likelihood equations and the Hessian are given by

$$\ln L(\beta) = \sum_{i=1}^n [-\exp(x'_i\beta) + (x'_i\beta)y_i - \ln(y_i!)]$$

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \sum_{i=1}^n [y_i - \exp(x'_i\hat{\beta})] x_i = 0$$

$$\frac{\partial^2 \ln L(\beta)}{\partial \beta \partial \beta'} = - \sum_{i=1}^n \exp(x'_i\beta) x_i x'_i$$

- Notice that the Hessian is **negative definite** for all x and β , which facilitates the estimation and ensures the uniqueness of the maximum, **if it exists**.
- The MLE has the usual properties. In particular

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \text{plim}\left(n^{-1} \sum \exp(x'_i\beta) x_i x'_i\right)^{-1}\right)$$

- As usual, inference can be performed using the LR, W and LM tests, and goodness-of-fit is uninteresting.

3) Testing for Overdispersion

- The Poisson model imposes (conditional) equidispersion, which is very restrictive.
- There are many possible causes for overdispersion:
 - Contagion;
 - Measurement error;
 - Misspecification of the conditional mean;
 - Neglected heterogeneity (random parameter variation).
- Economists tend to focus on the neglected heterogeneity issue, assuming

$$\lambda_i = \exp(x_i' \beta + \varepsilon_i)$$

$$E(\exp(\varepsilon_i) | x_i) = 1, \quad \text{Var}(\exp(\varepsilon_i) | x_i) = \sigma^2$$

- In this particular case

$$E(y_i|x_i) = E_\varepsilon [\exp(x_i'\beta + \varepsilon_i)] = \exp(x_i'\beta)$$

$$\begin{aligned} \text{Var}(y_i|x_i) &= E_\varepsilon [\exp(x_i'\beta + \varepsilon_i)] + \text{Var}_\varepsilon [\exp(x_i'\beta + \varepsilon_i)] \\ &= \exp(x_i'\beta) + \sigma^2 \exp(2x_i'\beta). \end{aligned}$$

- Therefore, this sort of neglected heterogeneity does not change the form of the conditional expectation of y_i , and the (pseudo) MLE is still consistent.
- The presence of overdispersion can be tested by testing $H_0 : \sigma^2 = 0$.
- This can be done using the following LM (IM) test statistic (Cox, 1983, and Chesher, 1984)

$$T = \sum_{i=1}^n \frac{\left(y_i - \exp(x_i'\hat{\beta})\right)^2 - y_i}{\sqrt{2 \sum_{i=1}^n \exp(2x_i'\hat{\beta})}} \stackrel{a}{\sim} \mathcal{N}(0, 1)$$

- Alternatively, we can regress $\left[\left(y_i - \exp(x'_i \hat{\beta}) \right)^2 - y_i \right] \exp(-x'_i \hat{\beta})$ on $\exp(x'_i \hat{\beta})$ (or on a constant or other functions of $\exp(x'_i \hat{\beta})$) and test the significance of the regressor (Cameron & Trivedi, 1986).
- All these tests can also detect underdispersion.
- Overdispersion tests are overplayed in the literature:
 - (a) in practice, the null is almost always rejected;
 - (b) if this is the only source of misspecification, the Poisson (pseudo) MLE is still consistent.
- Other specification tests are available, like the RESET test that checks the moment condition

$$E \left[(y_i - \exp(x'_i \beta)) (x'_i \beta)^2 \right] = 0$$

- In practice, the test can be performed by checking the significance of the additional regressor $\left(x'_i \hat{\beta} \right)^2$.

4) Heterogeneity and the Negative Binomial Regression Model

- The assumption that y_i has a Poisson distribution with $\lambda_i = \exp(x_i'\beta + \varepsilon_i)$, leads to the compound Poisson regression model

$$\Pr(y_i = j | x_i, \varepsilon_i) = \frac{\exp[-\exp(x_i'\beta + \varepsilon_i)] \exp(x_i'\beta + \varepsilon_i)^j}{j!}$$

$$\Pr(y_i = j | x_i) = \int \left[\frac{\exp[-\exp(x_i'\beta + \varepsilon_i)] \exp(x_i'\beta + \varepsilon_i)^j}{j!} \right] g(\varepsilon_i) d\varepsilon_i.$$

- This model can be made operational in different ways:
 - (a) Pseudo maximum likelihood estimation;
 - (b) Parametric estimation for specified $g(\varepsilon_i)$;
 - (c) Semiparametric estimation of β and $g(\varepsilon_i)$.

- Since the Poisson (pseudo) MLE is consistent in presence of this sort of misspecification, valid inference can be based on

$$\sqrt{n} \left(\hat{\beta}_{PML} - \beta \right) \xrightarrow{d} \mathcal{N} \left(0, \text{plim } A^{-1} B A^{-1} \right)$$

$$A = \frac{1}{N} \sum \exp(x'_i \hat{\beta}_{PML}) x_i x'_i \quad B = \frac{1}{N} \sum \left(y_i - \exp(x'_i \hat{\beta}_{PML}) \right)^2 x_i x'_i$$

- If $g(\varepsilon_i)$ is specified, the MLE can be obtained, but the estimator may not be robust to departures from the additional distributional assumptions.
- Assuming that $\exp(\varepsilon_i) \sim \Gamma(\sigma^{-2}, \sigma^2)$, $\Pr(y_i = j | x_i)$ is given by the **negative-binomial (NegBinII)** model:

$$\Pr(y_i = j | x_i) = \frac{\Gamma(j + \sigma^{-2}) [1 + \sigma^{-2} \lambda(x'_i \beta)^{-1}]^{-j}}{\Gamma(\sigma^{-2}) \Gamma(j + 1) (1 + \sigma^2 \lambda(x'_i \beta))^{\sigma^{-2}}}$$

- The Poisson model is obtained as a limiting case when $\sigma^2 \rightarrow 0$, but $H_0 : \sigma^2 = 0$ **cannot** be tested with a standard LR or W test.
- If $E(y_i | x_i)$ is correctly specified, the NegBinII estimator is consistent for β , but not for σ .

- The score test for $H_0 : \sigma^2 = 0$ is the overdispersion test studied before.
- Other parametric alternatives to the Poisson regression are available:
 - different parameterizations of the negative-binomial model;
 - generalized-Poisson regression;
 - beta-binomial regression (if the support of y has an upper bound);
 - Poisson-normal model (for which there is no closed form).
- Tests for nonnested hypotheses can be used to compare the chosen specification with alternative models.
- A semiparametric alternative is to assume that ε has a discrete distribution with Q support points $\alpha_1, \dots, \alpha_Q$ and corresponding probabilities π_1, \dots, π_Q , leading to

$$\Pr(y_i = j | x_i) = \sum_{q=1}^Q \frac{\exp[-\exp(x_i' \beta + \alpha_q)] \exp(x_i' \beta + \alpha_q)^j}{j!} \pi_q,$$

- **For a given Q** , estimation of $\beta, \alpha_1, \dots, \alpha_Q$ and π_1, \dots, π_{Q-1} can be performed by ML.
- Estimation by the EM algorithm is also popular.
- This model can be interpreted as:
 - (a) Finite-mixture (compound) Poisson model;
 - (b) Semiparametric approximation to a compound Poisson model with unspecified mixing distribution.
 - * This leads to a consistent estimator if Q is allowed to increase at an appropriate rate;
 - * In practice, the value of Q has to be chosen (for example using an information criterion);
 - * Inference is complicated by the fact that the number of parameters is not fixed.
 - * In general, inference is performed conditional on the value of Q , which underestimates the true variance of the parameters.

5) Hurdle and Zero-Altered Poisson Models

- In some cases, the population may be contaminated by individuals for which $y_i \equiv 0$.
- Let z_i be a dummy variable such that

$$z_i = \begin{cases} 1 & \text{if } \Pr(y_i = 0|x_i) = 1 \\ 0 & \text{if } \Pr(y_i = 0|x_i) < 1 \end{cases}$$

- Furthermore, let $\Pr(z_i = 1|x_i) = p_i$ and $\Pr(y_i = j|x_i, z_i = 0) = \pi_i(j)$.
- Then, the log-likelihood function for this **zero-inflated** (Mullahy, 1986) model can be written as

$$\ln L(\theta) = \sum_{i=1}^n \ln \left\{ [p_i + (1 - p_i) \pi_i(0)]^{\mathbf{1}(y_i=0)} [(1 - p_i) \pi_i(j)]^{\mathbf{1}(y_i>0)} \right\}$$

- Notice that $E(y_i|x_i) = (1 - p_i) E(y_i|x_i, z_i = 0)$ and that the pseudo maximum likelihood result does not hold here, except if p_i is constant.

- A different extension of the basic count data model is obtained by letting the zero and positive observations be generated by different mechanisms.
- This is what is called a **hurdle model** (Mullahy, 1986).
- In this case we have

$$\Pr(y_i = j|x_i) = \begin{cases} p_i(x_i\gamma) & \text{if } j = 0 \\ (1 - p_i(x_i\gamma)) \Pr(y_i = j|x_i, \beta) & \text{if } j > 0 \end{cases}$$

- Then, the likelihood function has the form

$$\ln L(\gamma, \beta) = \sum_{i=1}^n \{ \mathbf{1}(y_i = 0) (\ln p_i(x_i\gamma)) + \mathbf{1}(y_i > 0) \ln(1 - p_i(x_i\gamma)) + \mathbf{1}(y_i > 0) \ln[\Pr(y_i = j|x_i, \beta)] \}$$

- Notice that this function is separable.
- Correlated unobserved heterogeneity can be allowed for and integrated-out numerically.

- Usually, $\Pr (y_i = j|x_i, \beta)$ is specified as a truncated Poisson of the form

$$\Pr (y_i = j|x_i, \gamma) = \frac{\exp (-\lambda_i) \lambda_i^j}{(1 - \exp (-\lambda_i)) j!}, \quad j > 0,$$

with $\lambda_i = \exp (x_i' \beta)$.

- However, in this model **there is no real truncation** and therefore an equally valid specification would be

$$\Pr (y_i = j|x_i, \gamma) = \frac{\exp (-\lambda_i) \lambda_i^{j-1}}{(j-1)!}, \quad j > 0,$$

- When the truncated Poisson specification is used and $p_i (x_i \gamma)$ is specified as

$$p_i (x_i \gamma) = \exp (-\exp (x_i' \gamma)),$$

the null of no hurdle can be tested by testing $H_0 : \beta = \gamma$.

- In any case, consistency depends on the distributional assumptions.

6) Models for Panel Data

- Define $y_i = (y_{i1}, \dots, y_{iT})$ and $j_i = (j_{i1}, \dots, j_{iT})$, and let

$$\Pr(y_{it} = j_{it} | x_{it}, \varepsilon_i) = \frac{\exp(-\lambda_{it}) \lambda_{it}^{j_{it}}}{j_{it}!}$$

$$\lambda_{it} = \exp(x'_{it}\beta + \varepsilon_i) = \exp(x'_{it}\beta)\alpha_i.$$

6.1 Pooled Poisson regression based on the assumption that $E(y_{it} | x_{it}) = \exp(x'_{it}\beta)$ is consistent under mild assumptions.

- In particular, the **Poisson assumption**, **strict-exogeneity** and **conditional independence** between the elements of y_i , are not needed.
- However, we must assume that $E(\alpha_i | x_{it})$ is a constant (normalized to 1).
- Naturally, inference must be based on a robust (clustered) covariance estimator.
- Inclusion of time dummies in the model is generally recommended.

6.2 Under additional assumptions, a more efficient (random effects) estimator can be obtained.

- If strict-exogeneity of the regressors $E(y_{it}|x_i, \alpha_i) = E(y_{it}|x_{it}, \alpha_i)$ is assumed, a GLS estimator accounting for some overdispersion and serial-correlation can be used.
- This is often called the “**Population Averaged**” estimator.

6.3 With even more assumptions, further efficiency gains are possible.

- For that, we require:
 - (a) independence of the elements of $y_i = (y_{i1}, \dots, y_{iT})$, conditional on α_i and x_i ;
 - (b) strict-exogeneity of the regressors $E(y_{it}|x_i, \alpha_i) = E(y_{it}|x_{it}, \alpha_i)$;
 - (c) the following distributional assumptions
 - (i) $\Pr(y_{it} = j_{it}|x_{it}, \varepsilon_i)$ is given by the Poisson model;
 - (ii) distribution of ε_i is known and independent of x_{it} .

- In this case,

$$L(y_i) = \int \left[\prod_{t=1}^T \frac{\exp(-\exp(x'_{it}\beta)\alpha_i) (\exp(x'_{it}\beta)\alpha_i)^{j_{it}}}{j_{it}!} \right] g(\alpha_i) d\alpha$$

- If $\alpha_i = \exp(\varepsilon_i)$ is assumed to have a gamma distribution, the model has a closed form based on the negative-binomial distribution.
 - Often, it is assumed that α_i has a log-normal distribution (no closed form).
 - Consistency depends, of course, on the validity of the distributional assumptions.
 - A Mundlak-Chamberlain formulation can be used to relax the assumption of independence between α_i and the regressors.
- 6.4** As for the logit, there is a consistent fixed-effects estimator for the Poisson model, that does not require independence between α_i and the regressors.
- As before, this estimator requires strict-exogeneity.

- By the additivity property of the Poisson distribution, we have that

$$\sum_{t=1}^T y_{it} \sim \text{Poisson} \left(\sum_{t=1}^T \lambda_{it} \right).$$

- It turns out that $\sum_{t=1}^T y_{it}$ is a sufficient statistic for α_i and, therefore, the distribution of y_{it} conditional on x_{it} , α_i and $\sum_{t=1}^T y_{it}$ does not depend on α_i .
- Indeed, assuming independence of the elements of $y_i = (y_1, \dots, y_T)$ conditional on α_i and x_i , we can write

$$\begin{aligned} \Pr \left(y_i = j_i \mid x_{it}, \varepsilon_i, \sum_{t=1}^T y_{it} \right) &= \frac{\prod_{t=1}^T \frac{\exp(-\lambda_{it}) \lambda_{it}^{j_{it}}}{j_{it}!}}{\frac{\exp(-\sum_{t=1}^T \lambda_{it}) (\sum_{t=1}^T \lambda_{it})^{\sum_{t=1}^T j_{it}}}{(\sum_{t=1}^T j_{it})!}} \\ &= \frac{(\sum_{t=1}^T j_{it})!}{\prod_{t=1}^T j_{it}!} \prod_{t=1}^T \left(\frac{\lambda_{it}}{\sum_{t=1}^T \lambda_{it}} \right)^{j_{it}} = \frac{(\sum_{t=1}^T j_{it})!}{\prod_{t=1}^T j_{it}!} \prod_{t=1}^T \left(\frac{\exp(x'_{it}\beta)}{\sum_{t=1}^T \exp(x'_{it}\beta)} \right)^{j_{it}} \end{aligned}$$

- Estimation is simple due to the MNL structure of the likelihood function.
- Notice that the estimator **is consistent** even if:
 - (a) y_{it} is not Poisson.
 - (b) the elements of $y_i = (y_{i1}, \dots, y_{iT})$ are not independent, conditional on α_i and x_i .
- Naturally, if these assumptions do not hold, inference must be based on a robust (clustered) covariance matrix.
- As for the logit, the appeal of this estimator is limited by the fact that probabilities and marginal effects cannot be computed.

University week: 11

Limited Dependent Variable Models

1. Introduction;
2. Truncated data;
3. Censored data;
4. Sample selection;
5. Corner solutions.

Recommended reading: Greene: 19.1–19.3, 19.5.

1) Introduction

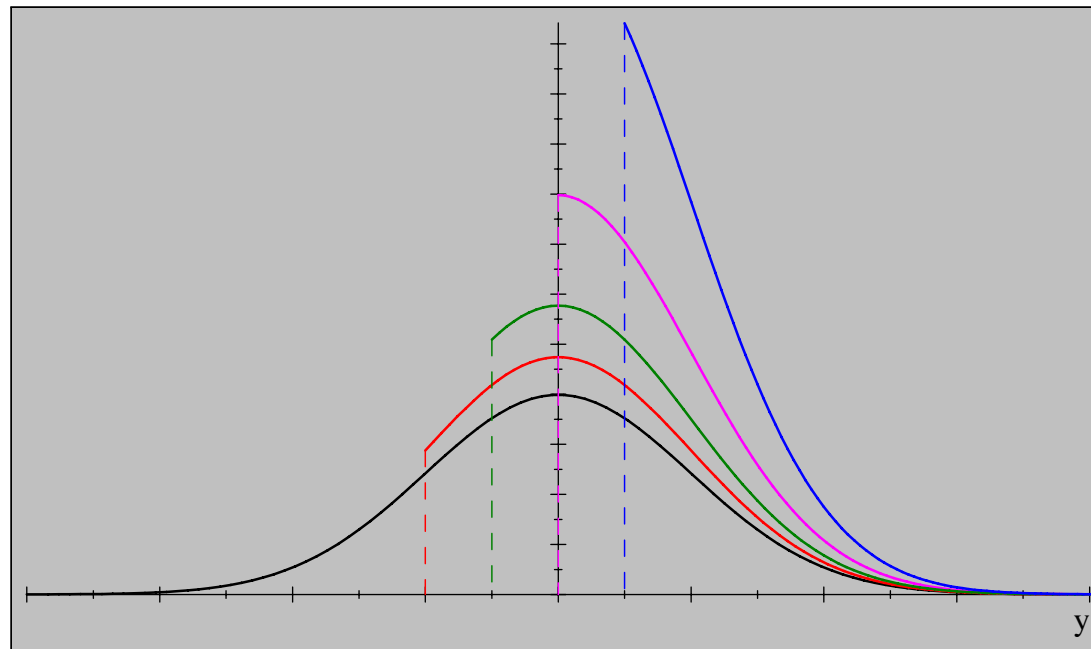
- In most applications, the domain of the variate of interest is limited.
- Although in many cases those bounds are irrelevant, in microeconomic applications the bounds on the domain of y often have to be taken into account.
- Traditionally, textbooks consider three forms of limited dependent variable models:
 - (a) Truncated data;
 - (b) Censored data;
 - (c) Sample selection (incidental truncation).
- All these cases treat the bounds on the support of the dependent variable as a result of a **limited observability** problem.
- However, in a large number of cases, the bounds are not the result of an observability problem but are intrinsic to the nature of the data (“**corner solutions**”).

2) Truncated data

- A sample of y is said to come from a truncated distribution when it is not possible to obtain observations from part of the domain of y .
- For example, if y is a continuous random variable with pdf $f(y)$ and a is a constant, we have left-truncation at a if the sample is drawn from the conditional distribution

$$f(y|y > a) = \frac{f(y)}{\Pr(y > a)}.$$

- In a regression, we have truncation if **the regressors are also not observed** when $y < a$.



- For a discrete random variable with support on $0, 1, \dots, \infty$, we have left-truncation at a if the sample is drawn from

$$\Pr(y = j | y > a) = \frac{\Pr(y = j)}{1 - \sum_{k=0}^{\lfloor a \rfloor} \Pr(y = k)},$$

where $\lfloor a \rfloor$ denotes the integer part of a .

- For example, we have already seen the zero-truncated Poisson defined by

$$\Pr(y = j | y > 0) = \frac{\exp(-\lambda_i) \lambda_i^j}{(1 - \exp(-\lambda_i)) j!}.$$

- **On-site samples** of count data are sometimes **mistaken for** truncated samples.
- **Notice that** truncation is only problematic if it depends on y or on another variable that is not independent of y and that is not used as a regressor.
- For example, in general, standard inference can be performed if the sample is obtained from $\Pr(y = j | x, x > a)$.

- Under certain conditions, it is possible to perform inference about the entire population using truncated samples.
- The problem is that this inference tends to be sensitive to **distributional assumptions**.
- For example, for the normal distribution with mean μ and variance σ^2 ,

$$E(y_i | y_i > a) = \mu + \sigma \frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)}.$$

- For the case of count data with $E(y_i | x_i) = \exp(x_i' \beta)$,

$$E(y_i | x_i, y_i > a) = \frac{\exp(x_i' \beta)}{1 - \sum_{k=0}^{\lfloor a \rfloor} \Pr(y = k | x_i)}.$$

- In both cases, the expectation for the truncated data depends on the shape of the conditional distribution.
- Because of this, estimation is often performed by maximum likelihood.

- An example is the truncated normal regression model defined by

$$y_i^* = x_i' \beta + u_i^*, \quad u_i^* \sim \mathcal{N}(0, \sigma^2),$$

$$(y_i, x_i) = \begin{cases} (y_i^*, x_i) & \text{if } y_i^* > 0 \\ \text{non-observable} & \text{if } y_i^* \leq 0 \end{cases}.$$

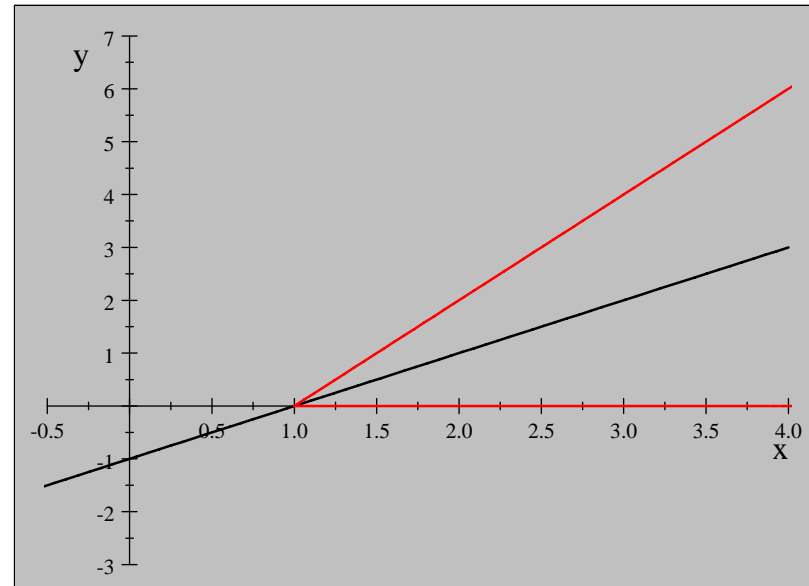
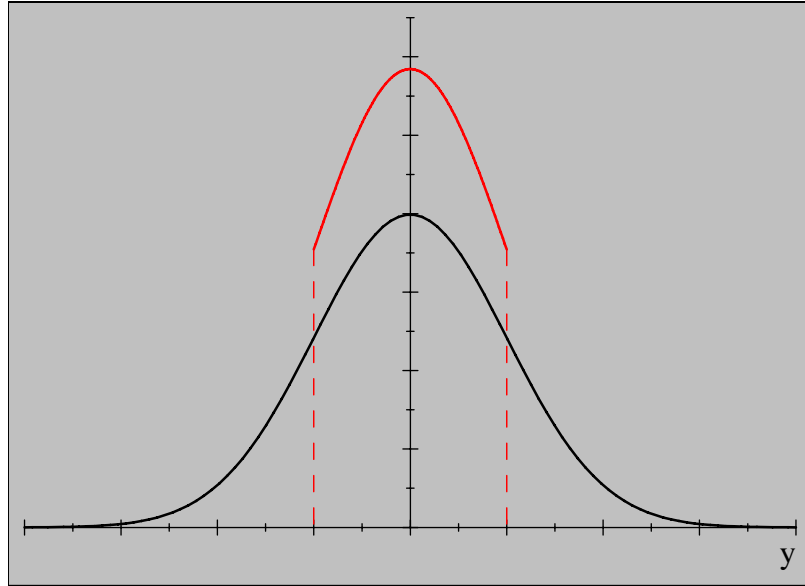
- This implies that

$$y_i = x_i' \beta + u_i, \quad E(u_i | x_i) = \sigma \frac{\phi(-x_i' \beta / \sigma)}{1 - \Phi(-x_i' \beta / \sigma)}.$$

- The parameters of interest can be estimated from the likelihood function

$$L(\theta) = \prod_{i=1}^n \frac{1}{\sigma} \frac{\phi\left(\frac{y_i - x_i' \beta}{\sigma}\right)}{1 - \Phi(-x_i' \beta / \sigma)}.$$

- A more robust estimator is Powell's (1986) Symmetrically Trimmed Least Squares (STLS).
- For symmetrical unimodal distributions, the mean is unchanged by symmetric truncation.



- Therefore, for symmetrical unimodal distributions, β can be estimated by

$$\hat{\beta} = \arg \min_b \sum_{i=1}^n (y_i - \max \{0.5y_i, x_i' b\})^2 .$$

- Under mild conditions, $\hat{\beta}$ is \sqrt{n} -consistent and asymptotically normal.
- For symmetrical unimodal distributions, an alternative estimator can be based on the mode.

3) Censored data

- Censoring is also a problem of partial observability. For example, if y^* is a random variable and a is a constant, we have left-censoring at a if we can only observe

$$y = \begin{cases} a & \text{if } y^* \leq a \\ y^* & \text{if } y^* > a \end{cases} .$$

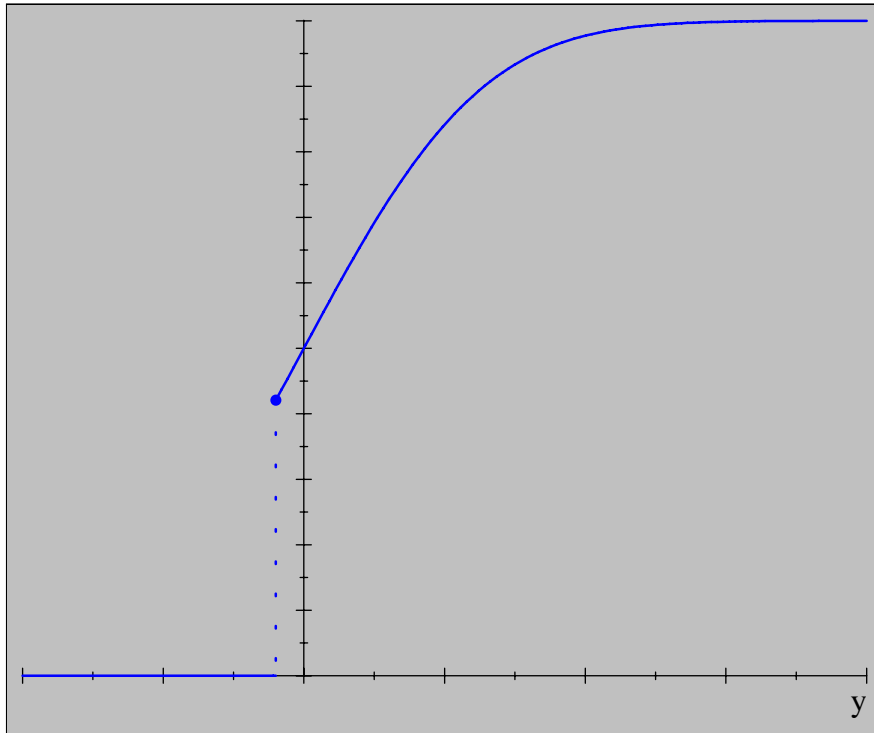
- If y^* is a discrete random variable with support on $0, 1, \dots, \infty$, the probability function of y is given by

$$\Pr(y = j) = \begin{cases} 0 & \text{if } j < a \\ \Pr(y^* \leq j) & \text{if } j = a \\ \Pr(y^* = j) & \text{if } j > a \end{cases} .$$

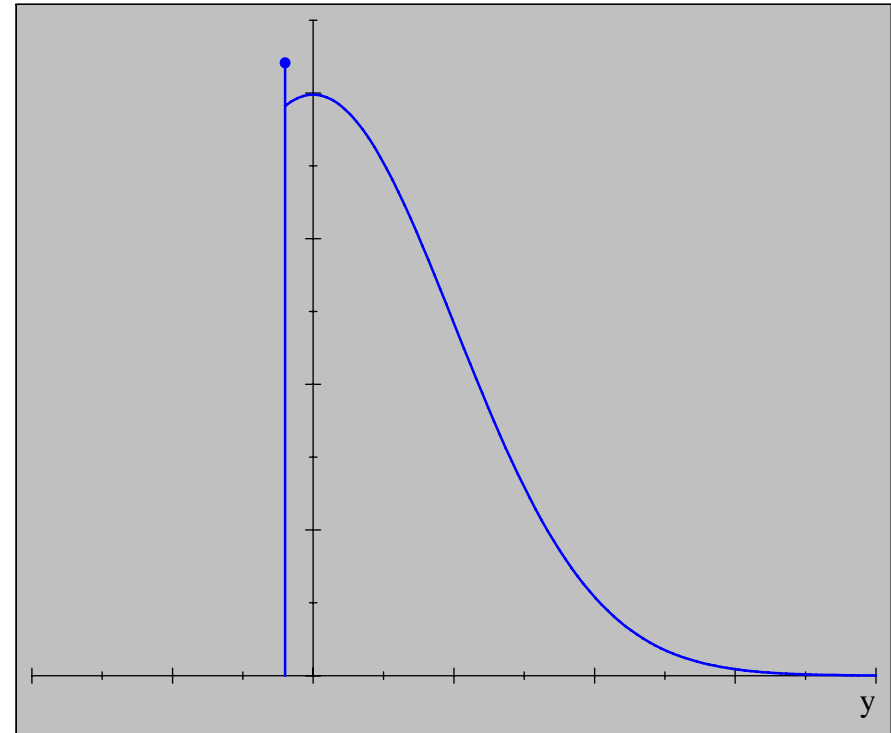
- If y^* is a continuous random variable with pdf $f(y^*)$, y has the **mixed distribution**

$$\Pr(y \leq k) = \begin{cases} 0 & \text{if } k < a \\ \Pr(y^* \leq k) & \text{if } k \geq a \end{cases} .$$

- It is assumed that the **regressors are fully observed**, even for cases with censored y^* .



Censored distribution



Censored density

- **Mixed distributions** also arise due to “**corner solutions**”.
- Often, data with mixed distributions are incorrectly treated as censored.
- The standard regression model for continuous censored data is the **Tobit** (normal censored regression model).

- The Tobit for censoring at zero is defined by

$$y_i^* = x_i' \beta + u_i^*, \quad u_i^* \sim \mathcal{N}(0, \sigma^2).$$

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ y_i^* & \text{if } y_i^* > 0 \end{cases}.$$

- The model is appropriate when we are interested in making inference about y_i^* , not y_i .
- ML inference is based on the likelihood function

$$L(\theta) = \prod_{i=1}^n \left[1 - \Phi\left(\frac{x_i' \beta}{\sigma}\right) \right]^{\mathbf{1}(y_i=0)} \left[\frac{1}{\sigma} \phi\left(\frac{y_i - x_i' \beta}{\sigma}\right) \right]^{\mathbf{1}(y_i>0)}.$$

- Reparameterizing $L(\theta)$ with $\gamma = \beta/\sigma$ and $\omega = 1/\sigma$, the Hessian is **negative definite** and estimation is easy, but the estimates may not exist (like in binary or count data models).
- Consistency depends on the **distributional assumptions** (but independence is not required) and appropriate specification tests are available.

- Powell's symmetric trimming idea can also be used to estimate β under milder assumptions.
- In this case, for symmetrical unimodal distributions, $\hat{\beta}$ can be estimated as

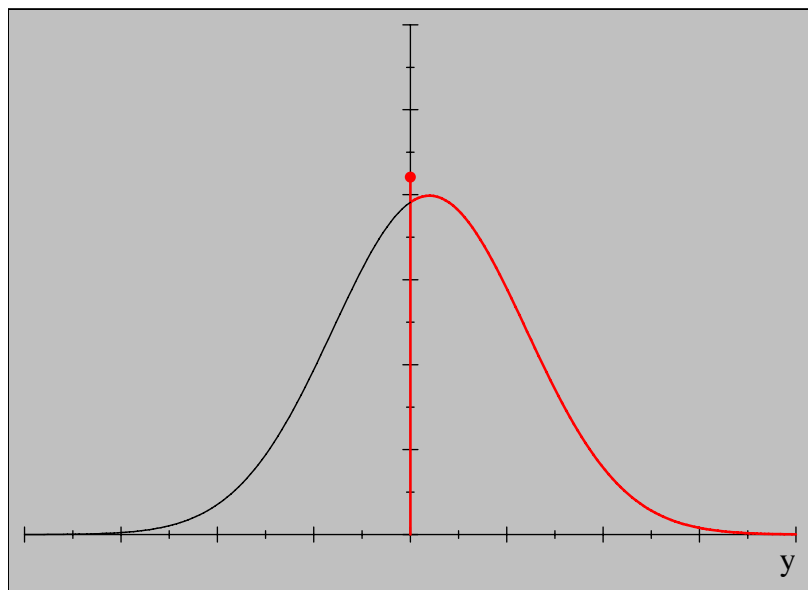
$$\hat{\beta} = \left[\sum_{i=1}^n \mathbf{1} \left(x_i' \hat{\beta} > 0 \right) x_i x_i' \right]^{-1} \sum_{i=1}^n \mathbf{1} \left(x_i' \hat{\beta} > 0 \right) x_i \min \left\{ y_i, 2x_i' \hat{\beta} \right\}.$$

- Under mild conditions, $\hat{\beta}$ is \sqrt{n} -consistent and asymptotically normal, with a covariance matrix that can be estimated by $\widehat{\text{Var}} \left(\hat{\beta} \right) = A^{-1} B A^{-1}$, with

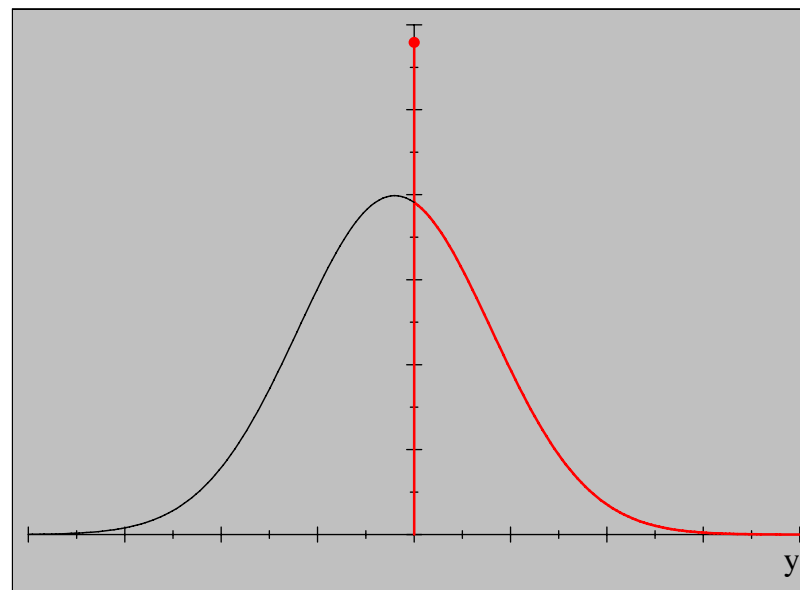
$$A = \left[\sum_{i=1}^n \mathbf{1} \left(0 < y_i < 2x_i' \hat{\beta} \right) x_i x_i' \right], \quad B = \left[\sum_{i=1}^n \mathbf{1} \left(x_i' \hat{\beta} > 0 \right) x_i x_i' \hat{r}^2 \right],$$

$$\hat{r} = \min \left\{ y_i, 2x_i' \hat{\beta} \right\} - x_i' \hat{\beta}.$$

- Also under symmetry, an alternative consistent estimator is available.
- Under symmetry, the median and the mean coincide and therefore median regression identifies the conditional mean.
- Moreover, if the median of y^* is positive, it coincides with the median of y .
- If the median of y^* is negative, the median of y is 0.



Density censored below the median



Density censored above the median

- Therefore,

$$Q_{y_i}(0.5|x_i) = \max\{0, x_i'\beta\}.$$

- Even if the distribution is not symmetric, this estimator has an interesting interpretation.
- The censored least absolute deviations (CLAD) is also due to Powell (1984).
- Like the STLS, the CLAD is robust to non-normality and to heteroskedasticity.
- Several algorithms to estimate β are available.
- In general, estimation can be performed as a sequence of uncensored linear median regressions in selected subsamples.
- As usual, estimation of the covariance matrix requires kernel density estimation.
- Naturally, other quantiles can be estimated using the same idea.

4) Sample selection

- In many cases, the sample available depends on individual decisions (self-selection).
- The leading example is the wage equation, which can only be estimated for individuals that participate in the labour market.
- Suppose that we are interested in the regression

$$y_i = x_i' \beta + \varepsilon_i, \quad E(\varepsilon_i | x_i) = 0,$$

but y_i is observable only when

$$z_i^* = x_i' \gamma + u_i > 0.$$

- If ε_i and u_i are not independent, OLS estimation of β is inconsistent because $E(\varepsilon_i | x_i, u_i > -x_i' \gamma) \neq 0$.
- Heckman (1976) popularized an estimator of β which is consistent under the assumption

$$\begin{bmatrix} u_i \\ \varepsilon_i \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix} \right).$$

- Under this assumption, β , γ , ρ and σ can be estimated by maximum likelihood.
- Alternatively, Heckman (1976) used a two-step estimator (Heckit) based on the result

$$E(\varepsilon_i | x_i, u_i > -x_i' \gamma) = \rho \sigma \frac{\phi(x_i' \gamma)}{\Phi(x_i' \gamma)}.$$

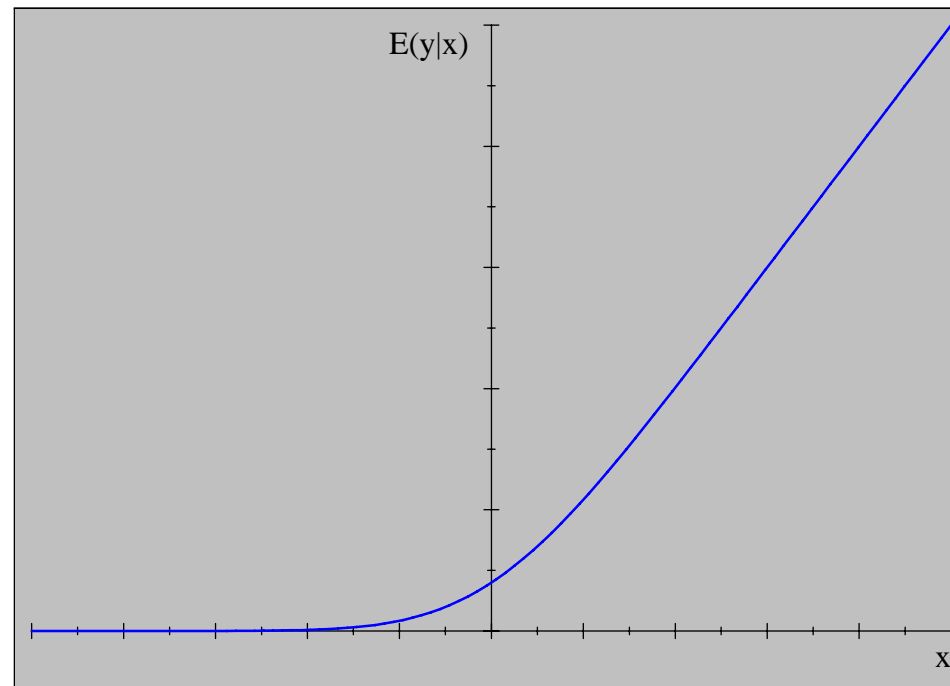
- Therefore, in the **first step**, $x_i' \gamma$ can be estimated by a probit.
- In the **second step**, β and $\rho \sigma$ are estimated in the OLS regression of y_i on x_i and $\frac{\phi(x_i' \gamma)}{\Phi(x_i' \gamma)}$.
- A standard t -test for the significance of $\rho \sigma$ is a **valid test** for correlation between u_i and ε_i .
- In case u_i and ε_i are correlated, the covariance matrix of the second step has to be corrected to account for the **estimated regressor**.
- Identification is easier if there are **exclusion restrictions**.
- **Semiparametric** versions of this estimator have been proposed. However, these are impractical and are rarely used.

5) Corner solutions

- Corner solutions data are data with mixed distributions where the mass-points do not result from censoring or sample selection.
- In this case we are interested in modeling how the regressors affect the distribution of the observed y , not some latent unobservable y^* .
- As noted before, because they also have mixed distributions, corner solutions data are often incorrectly treated as censored.
- In particular, the Tobit and Heckman's sample-selection estimator are often used to model corner solutions data.

- When the Tobit is used to model uncensored data with mixed distributions, that is equivalent to assuming the following functional form

$$E(y_i|x_i) = x_i'\beta\Phi\left(\frac{x_i'\beta}{\sigma}\right) + \sigma\phi\left(\frac{x_i'\beta}{\sigma}\right).$$



- Notice that, in this case, the **marginal effect** is given by $\frac{\partial E(y_i|x_i)}{\partial x_i} = \beta\Phi\left(\frac{x_i'\beta}{\sigma}\right)$.

- The Heckit estimator is also often used to model corner solutions.
- This is equivalent to assuming the following functional form

$$E(y_i|x_i) = x_i'\beta\Phi(x_i'\gamma) + \rho\sigma\phi(x_i'\gamma),$$

which is a generalization of the Tobit specification.

- In practice, the dependent variable of the second step is often $\ln(y_i)$ and in that case

$$E(y_i|x_i) = \exp\left(x_i'\beta + \frac{\sigma^2}{2}\right)\Phi(x_i'\gamma + \rho\sigma).$$

- These estimators impose specific functional forms that may be inadequate or undesirable.
- An alternative to this approach is to directly specify the form of $E(y_i|x_i)$ and estimate the parameters of interest using a robust estimator.

- For example, if y is a fractional variate with domain $[0, 1]$, with possible mass-points at 0 or 1, we can specify

$$E(y_i|x_i) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}.$$

- If y is a non-negative variate, with a possible mass-point at 0, it is often appropriate to specify

$$E(y_i|x_i) = \exp(x_i'\beta).$$

- Estimation can be performed by (non-linear) least squares.
- However, it is generally important to (at least) partially account for heteroskedasticity.
- An attractive way of doing that is to use pseudo maximum likelihood (PML) estimators.

- For the case of fractional data, Papke and Wooldrige (1996) have suggested the use of the Bernoulli PML estimator based on

$$\ln L(\beta) = \sum_{i=1}^n \left[y_i \ln \left(\frac{\exp(x'_i \beta)}{1 + \exp(x'_i \beta)} \right) + (1 - y_i) \ln \left(\frac{1}{1 + \exp(x'_i \beta)} \right) \right].$$

- For non-negative data, Santos Silva and Tenreyro (2006) suggested the use of the Poisson PML based on

$$\ln L(\beta) = \sum_{i=1}^n [-\exp(x'_i \beta) + y_i (x'_i \beta) - \ln \Gamma(y_i + 1)]$$

- The estimators are consistent as long as $E(y_i | x_i)$ is correctly specified.
- In both cases, robust covariance matrices should be used for inference.
- In both cases, the estimators can be adapted to deal with panel data and endogeneity.