Bavarian Graduate Program in Economics

Prof. Dr. Martin Kukuk

Universität Würzburg

Frontiers in Econometrics

Das vorliegende Manuskript ist für Studierende, die den Graduiertenkurs *Frontiers in Econometrics* besuchen, als Begleitmaterial gedacht, um den Mitschrieb zu erleichtern. Ich kann nicht für die Vollständigkeit und die Fehlerfreiheit des Manuskripts garantieren. Allein die Ausführungen in den Vorlesungen sind relevant.

Juli 2007

Contents

1 The Geometry of Least Squares

1.1 Vector spaces

Representation of a vector

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \in E^n$$

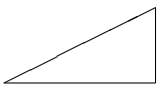
as

Operations

- Addition $a + b = c$
- Scalar product $\alpha \cdot \boldsymbol{a}$ with $\alpha \in \mathbb{R}$
- Inner product $\boldsymbol{a}'\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{a} = \sum_{i=1}^{n} a_i \cdot b_i$
- Length or norm

  $$\|\boldsymbol{x}\| = (\boldsymbol{x}'\boldsymbol{x})^{1/2} = \left(\sum x_i^2\right)^{1/2}$$

  This is inspired by Pythagoras' theorem: $a^2 + b^2 = c^2$



Vectors in E^2 using cartesian coordinates:

Geometric interpretation of innerproduct

General case:

$$\left(\frac{1}{\|\boldsymbol{x}\|}\boldsymbol{x}\right)'\left(\frac{1}{\|\boldsymbol{y}\|}\boldsymbol{y}\right) = \cos\theta \qquad \Leftrightarrow \qquad \boldsymbol{x}'\boldsymbol{y} = \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\| \cdot \cos\theta$$

Orthogonality: $\boldsymbol{x}'\boldsymbol{y} = 0 = \cos(\pi/2)$

Subspaces

Subspace $S(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$ is a set of all **linear combinations**:

$$S(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k) = \left\{ \boldsymbol{z} \in E^n \middle| \boldsymbol{z} = \sum_{i=1}^{k} b_i \cdot \boldsymbol{x}_i \ , \ b_i \in \mathbb{R} \ , \ \boldsymbol{x}_i \in E^n \right\}$$

Denoting $\underset{(n\times k)}{\boldsymbol{X}} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k]$ and $\boldsymbol{b} = (b_1, \ldots, b_k)'$, then

$$S(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k) = S(\boldsymbol{X}) = \left\{ \boldsymbol{z} \in E^n \middle| \boldsymbol{z} = \boldsymbol{X}\boldsymbol{b} \ , \ \boldsymbol{b} \in E^k \ , \ \boldsymbol{x}_i \in E^n \right\}$$

Linear Independence

Columns of $\underset{(n\times k)}{\boldsymbol{X}}$ are linear independent if $\boldsymbol{X}\boldsymbol{b} = \underset{(n\times 1)}{\boldsymbol{0}}$ is only possible with $\boldsymbol{b} = \underset{(k\times 1)}{\boldsymbol{0}}$.

Linear dependent columns of $\boldsymbol{X}$:

$$\begin{aligned} & \boldsymbol{X}\boldsymbol{b} = \boldsymbol{0} \qquad \text{with } \boldsymbol{b} \neq \boldsymbol{0} \\ \Leftrightarrow \quad & \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{0} = \boldsymbol{0} \\ \Leftrightarrow \quad & (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{b} = \boldsymbol{0} \end{aligned}$$

This is a contradiction. Thus if columns of $\boldsymbol{X}$ are linear dependent then $\boldsymbol{X}'\boldsymbol{X}$ is not invertible.

1.2 Geometry of OLS estimation

Considering the model:

$$y_t = \underset{(1\times k)}{\boldsymbol{X}_t} \cdot \underset{(k\times 1)}{\boldsymbol{\beta}} + u_t \qquad ;\ t = 1, \ldots, n$$

In matrix notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

$\boldsymbol{\beta}$ is unknown, $\boldsymbol{y}$ and $\boldsymbol{X}$ are observed, $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{\beta})$ unobservable.

The estimation principle of Least Squares:

$$\min_{\boldsymbol{\beta}} \sum_{t=1}^{n} u_t^2 = \min_{\boldsymbol{\beta}} \boldsymbol{u}(\boldsymbol{\beta})'\boldsymbol{u}(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

In words: Find $\hat{\boldsymbol{u}} = \boldsymbol{u}(\hat{\boldsymbol{\beta}})$ that has least squared length and therefore least length $\|\boldsymbol{u}(\hat{\boldsymbol{\beta}})\|$.

$\boldsymbol{u}(\hat{\boldsymbol{\beta}})$ with least length is found if we drop a perpendicular on $S(\boldsymbol{X})$.

We see that $\hat{\boldsymbol{u}}$ is orthogonal on $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ or more general orthogonal on every vector in $S(\boldsymbol{X})$.

Geometrically, least squares estimation can be decomposed in two steps:

- Find point in $S(\boldsymbol{X})$ closest to $\boldsymbol{y}$: $\boldsymbol{X}\hat{\boldsymbol{\beta}}$
- Knowing $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, determine $\hat{\boldsymbol{\beta}}$

Applying Pythagoras' theorem we obtain

$$\begin{aligned} & \|\boldsymbol{y}\| = \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\| + \|\hat{\boldsymbol{u}}\| \\ \Leftrightarrow \quad & \boldsymbol{y}'\boldsymbol{y} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} \end{aligned}$$

The total sum of squares is equal to the explained sum of squares plus the sum of squared residuals.

OLS estimator

Vector $\hat{\boldsymbol{u}}$ is orthogonal to all vectors in $S(\boldsymbol{X})$ and especially on all the elements of $\boldsymbol{X}$:

$$\begin{aligned} & \boldsymbol{X}'\hat{\boldsymbol{u}} = \underset{(k\times 1)}{\boldsymbol{0}} \qquad \Leftrightarrow \qquad \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0} \\ \Leftrightarrow \quad & \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0} \qquad \Leftrightarrow \qquad \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \end{aligned}$$

If $\boldsymbol{X}$ has linear independent columns then $\boldsymbol{X}'\boldsymbol{X}$ is invertible and we can premultiply the last equation with $(\boldsymbol{X}'\boldsymbol{X})^{-1}$:

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

If $\boldsymbol{X}$ has linear dependent columns we still can project on to $S(\boldsymbol{X})$ but solving for $\hat{\boldsymbol{\beta}}$ is not unique.

Orthogonal projection

First step of OLS is a mapping of $\boldsymbol{y}$ on to $S(\boldsymbol{X})$. We call it a projection of $\boldsymbol{y}$ into a point of subspace $S(\boldsymbol{X})$. This mapping is $\boldsymbol{y} \longrightarrow \boldsymbol{X}\hat{\boldsymbol{\beta}}$ or

$$\underset{(n\times n)}{\boldsymbol{P_X}} \cdot \underset{(n\times 1)}{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

Inserting the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$:

$$\boldsymbol{P_X} \cdot \boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Thus we obtain the orthogonal projector:

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

Since every point (or vector) in subspace $S(\boldsymbol{X})$ is projected on to itself, we have

$$\boldsymbol{P_X} \cdot \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} \qquad \text{or specifically } \boldsymbol{P_X} \cdot \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

Thus we get

$$\boldsymbol{P_X}\boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P_X}\boldsymbol{y}$$

or $\boldsymbol{P_X} \cdot \boldsymbol{P_X} = \boldsymbol{P_X}$. Analytically:

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

A matrix having this property is called **idempotent**.

Projecting $\boldsymbol{y}$ on to $S(\boldsymbol{X})$ yields the fitted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Analogously, projecting on to

$$S^{\perp}(\boldsymbol{X}) = \left\{ \boldsymbol{w} \in E^n \middle| \boldsymbol{w}'\boldsymbol{z} = 0 \ \ \forall \boldsymbol{z} \in S(\boldsymbol{X}) \right\}$$

yields the residuals $\hat{\boldsymbol{u}}$:

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

$$\Rightarrow \boldsymbol{M_X} \cdot \boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P_X}) \cdot \boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P_X} \cdot \boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{u}}$$

$\boldsymbol{M_X}$ annihilates each vector of $S(\boldsymbol{X})$:

$$\boldsymbol{M_X}\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$$

$\boldsymbol{P_X}$ annihilates each vector of $S^{\perp}(\boldsymbol{X})$:

$$\boldsymbol{P_X} \cdot \boldsymbol{w} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}'\boldsymbol{w}}_{=\boldsymbol{0}}$$

$\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are called orthogonal matrices:

$$\boldsymbol{M_X} \cdot \boldsymbol{P_X} = (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{P_X} = \boldsymbol{P_X} - \boldsymbol{P_X}\boldsymbol{P_X} = \boldsymbol{P_X} - \boldsymbol{P_X} = \boldsymbol{O}$$

1.3 Frisch-Waugh-Lovell Theorem

We have seen that $\boldsymbol{P_X} + \boldsymbol{M_X} = \boldsymbol{I}$ and thus

$$\boldsymbol{y} = \boldsymbol{P_X}\boldsymbol{y} + \boldsymbol{M_X}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}$$

Partitioning regressor matrix $\boldsymbol{X}$ and also vector $\boldsymbol{\beta}$

$$\underset{(n\times k)}{\boldsymbol{X}} = [\underset{(n\times k_1)}{\boldsymbol{X}_1} \vdots \underset{(n\times k_2)}{\boldsymbol{X}_2}] \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

we can write our model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u} \tag{1}$$

We now define $\boldsymbol{P}_{\boldsymbol{X}_1} = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ and $\boldsymbol{M}_{\boldsymbol{X}_1} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_1}$ and analogously $\boldsymbol{P}_{\boldsymbol{X}_2}, \boldsymbol{M}_{\boldsymbol{X}_2}$.

FWL Theorem

1. OLS estimates for $\boldsymbol{\beta}_2$ from regression (1) and from $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2\boldsymbol{\beta}_2 +$ *residual* are numerically identical.

2. The residuals from from regression (1) and from $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2\boldsymbol{\beta}_2 +$ *residual* are numerically identical.

Proof

1. $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2$ fulfill $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$. Premultiplying with $\boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{X}_1}$ yields:

$$\begin{aligned} & \boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} = \boldsymbol{X}_2'\underbrace{\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_1}\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \underbrace{\boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{M}_{\boldsymbol{X}}}\boldsymbol{y} \\ \Leftrightarrow \quad & \boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} = \boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 \\ \Leftrightarrow \quad & \hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} \end{aligned}$$

   OLS estimator for regression $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2\boldsymbol{\beta}_2 +$ residual :

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{X}_1}'\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{X}_1}'\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y}$$

   We used the result

$$\boldsymbol{M}_{\boldsymbol{X}} \cdot \boldsymbol{M}_{\boldsymbol{X}_1} = \boldsymbol{M}_{\boldsymbol{X}}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_1}) = \boldsymbol{M}_{\boldsymbol{X}} - \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1' = \boldsymbol{M}_{\boldsymbol{X}}$$

   and therefore $\boldsymbol{M}_{\boldsymbol{X}} \cdot \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2 = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{X}_2 = \boldsymbol{O}$.

2. Premultiplying $\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$ with $\boldsymbol{M}_{\boldsymbol{X}_1}$ yields:

$$\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \underbrace{\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{M}_{\boldsymbol{X}}}\boldsymbol{y}$$

   Thus the regression of $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y}$ on $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2$ obtains residuals $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$ just as the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$.

1.4 Goodness of Fit

Due to $\boldsymbol{y} = \boldsymbol{P_X y} + \boldsymbol{M_X y}$ we could construct a goodness of fit measure

$$\frac{\|\boldsymbol{P_X y}\|^2}{\|\boldsymbol{y}\|^2}$$

Suppose $\boldsymbol{X}$ includes a constant, i.e. $\boldsymbol{\iota}$ is a column of $\boldsymbol{X}$, we can increase

$$\frac{\|\boldsymbol{P_X y} + \alpha\boldsymbol{\iota}\|^2}{\|\boldsymbol{y} + \alpha\boldsymbol{\iota}\|^2}$$

close to 1 by increasing α. If we apply FWL and center both sides $\boldsymbol{M_\iota y} = \boldsymbol{P_X M_\iota y} + \boldsymbol{M_X M_\iota y}$ we get

$$R^2 = \frac{\|\boldsymbol{P_X M_\iota y}\|^2}{\|\boldsymbol{M_\iota y}\|^2}$$

It measures the additional fit of a regression model over a benchmark model with just a constant.

1.5 Influential Points and Leverage

In simple regression we can assess high leverage points by inspecting the scatter plot of (x_t, y_t) observations.

To evaluate highly influential points in a multiple regression model we can run a regression in order to see whether observation t is influential

$$\boldsymbol{y} = \boldsymbol{X\beta} + \alpha\boldsymbol{e}_t + \boldsymbol{u}$$

with $\boldsymbol{e}_t$ having 1 as the t^{th} components and zeros elsewhere. Applying FWL yields

$$\boldsymbol{M}_t\boldsymbol{y} = \boldsymbol{M}_t\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}_t\boldsymbol{u}$$

with

$$\boldsymbol{M}_t = \boldsymbol{I} - \boldsymbol{e}_t(\boldsymbol{e}_t'\boldsymbol{e}_t)^{-1}\boldsymbol{e}_t' = \boldsymbol{I} - \boldsymbol{e}_t\boldsymbol{e}_t'$$

$$\Rightarrow \boldsymbol{M}_t\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{e}_t\boldsymbol{e}_t'\boldsymbol{y} = \boldsymbol{y} - y_t\boldsymbol{e}_t$$

Furthermore,

$$\hat{\alpha} = \frac{\boldsymbol{e}_t'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}}{\boldsymbol{e}_t'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{e}_t} = \frac{\hat{u}_t}{1 - \boldsymbol{e}_t'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t} = \frac{\hat{u}_t}{1 - h_t}$$

Thus h_t is the t^{th} diagonal element of $\boldsymbol{P}_{\boldsymbol{X}}$. It holds that $1/n \leq h_t \leq 1$ and $\bar{h}_t = n^{-1}\sum h_t = k/n$.

The effects of the t^{th} observation on the estimates of $\hat{\boldsymbol{\beta}}$ are analyzed by running a regression without t^{th} observation: (Define $\boldsymbol{Z} = [\boldsymbol{X} \vdots \boldsymbol{e}_t]$)

$$\begin{aligned}
& \boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(t)} + \hat{\alpha}\boldsymbol{e}_t + \boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{u} \\
\Leftrightarrow\ & \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(t)} + \hat{\alpha}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t + \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{u} \\
\Leftrightarrow\ & \boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(t)} - \hat{\boldsymbol{\beta}}) = -\hat{\alpha}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t \\
\Leftrightarrow\ & \boldsymbol{e}_t'\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(t)} - \hat{\boldsymbol{\beta}}) = -\hat{\alpha}\boldsymbol{e}_t'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t \\
\Leftrightarrow\ & \boldsymbol{X}_t\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{X}_t\hat{\boldsymbol{\beta}} = -\hat{\alpha}h_t = -\frac{\hat{u}_t}{1 - h_t}h_t = -\frac{h_t}{1 - h_t}\hat{u}_t
\end{aligned}$$

2 Statistical Properties of OLS

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \qquad \boldsymbol{u} \sim IID(\boldsymbol{0}, \sigma^2 \cdot \boldsymbol{I}) \tag{2}$$

Using the LS estimation principle:

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

Assuming $\boldsymbol{X}$ as **non-stochastic** or **fix**:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u}) = \boldsymbol{\beta}$$

- This assumption is obviously appropriate in experimental situations.
- Social sciences seldomly have experimental designs.
- Assumption is mostly used for analytical simplicity.

It can be replaced by the assumption that $\boldsymbol{X}$ is exogenous.

- Randomness of $\boldsymbol{X}$ is *outside* our regression model (2).
- Error term is therefore independent of $\boldsymbol{X}$:

$$\boldsymbol{0} = E(\boldsymbol{u} \cdot \boldsymbol{h}(\boldsymbol{X})') = E_{\boldsymbol{X}}\left[E(\boldsymbol{u} \cdot \boldsymbol{h}(\boldsymbol{X})'|\boldsymbol{X})\right] = E_{\boldsymbol{X}}\left[\underbrace{E(\boldsymbol{u}|\boldsymbol{X})}\cdot\boldsymbol{h}(\boldsymbol{X})'\right]$$

- Thus assumption $E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}$ suffices to ensure tractability:

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

- $E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}$ is especially in time series regression contexts quite restrictive. For instance, if

$$y_t = \mu + \rho \cdot y_{t-1} + u_t \qquad t = 1, \ldots, T$$

In matrix notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

with

$$\boldsymbol{X} = \begin{pmatrix} 1 & y_0 \\ \vdots & \vdots \\ 1 & y_{T-1} \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \rho \end{pmatrix}$$

$E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}$ implies $E(u_t|\underbrace{y_0, \ldots, y_t, y_{t+1}, \ldots, y_{T-1}}_{\text{past and future values of } \boldsymbol{X}_t}) = 0$.

It seems more appropriate in that context to assume $E(u_t|\boldsymbol{X}_t) = 0$ with $\boldsymbol{X}_t$ possibly containing components of $\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots$ (lagged values, past history).

But this does not guarantee $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

- For cross-section analysis

$$E(u_i|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i, \ldots, \boldsymbol{X}_n) = 0$$

is not too restrictive.

2.1 Variance-Covariance matrix of $\hat{\boldsymbol{\beta}}$

$$\begin{aligned} V(\boldsymbol{u}) &= E((\boldsymbol{u} - E(\boldsymbol{u})) \cdot (\boldsymbol{u} - E(\boldsymbol{u}))') \\ &= E(\boldsymbol{u} \cdot \boldsymbol{u}') - E(\boldsymbol{u}) \cdot E(\boldsymbol{u})' \end{aligned}$$

In our context assuming $E(\boldsymbol{u}) = \boldsymbol{0}$ and $V(\boldsymbol{u}) = \sigma^2\boldsymbol{I}$ we have $E(\boldsymbol{u} \cdot \boldsymbol{u}') = \sigma^2\boldsymbol{I}$.

We utilize the very important property for linear transformations:

$$\begin{aligned} E(\boldsymbol{Y}) &= \boldsymbol{\mu}; \quad V(\boldsymbol{Y}) = E((\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))') = \boldsymbol{\Sigma} \\ \boldsymbol{Z} &= \boldsymbol{a} + \boldsymbol{B}\boldsymbol{Y} \\ \Rightarrow E(\boldsymbol{Z}) &= \boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu} \\ V(\boldsymbol{Z}) &= V(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{Y}) = \boldsymbol{B} \cdot V(\boldsymbol{Y}) \cdot \boldsymbol{B}' = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}' \end{aligned}$$

- $\boldsymbol{X}$ is fix

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= V(\boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\boldsymbol{u}) = (\boldsymbol{X'X})^{-1}\boldsymbol{X}'V(\boldsymbol{u})\boldsymbol{X}(\boldsymbol{X'X})^{-1} \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X'X})^{-1} = \sigma^2(\boldsymbol{X'X})^{-1}\boldsymbol{X'X}(\boldsymbol{X'X})^{-1} \\ &= \sigma^2(\boldsymbol{X'X})^{-1} \end{aligned}$$

- $\boldsymbol{X}$ is exogenous

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= E((\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})) \cdot (\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))'|\boldsymbol{X}) \\ &= E[(\boldsymbol{X'X})^{-1}\boldsymbol{X}'\boldsymbol{u} \cdot \boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X'X})^{-1}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X'X})^{-1} \end{aligned}$$

We can rewrite this covariance matrix as:

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

Thus the precision of estimator $\hat{\boldsymbol{\beta}}$ in terms of variation around $\boldsymbol{\beta}$ depends on:

- $\sigma^2 = \text{Var}(u_i)$
- n (sample size) (Just like the variance of $\bar{y}$)
- regressor matrix $\boldsymbol{X}$:

  We are, without loss of generality, interested in the variance of $\underbrace{\beta_1}_{1\times 1}$:

$$\boldsymbol{y} = \underbrace{\boldsymbol{X}_1}_{n\times 1}\beta_1 + \underbrace{\boldsymbol{X}_2}_{n\times k-1}\boldsymbol{\beta}_2 + \boldsymbol{u}$$

  Applying FWL-Theorem: (Defining $\boldsymbol{M}_2 = \boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$)

$$\begin{aligned} \boldsymbol{M}_2\boldsymbol{y} &= \boldsymbol{M}_2\boldsymbol{X}_1\beta_1 + \underbrace{\boldsymbol{M}_2\boldsymbol{X}_2}_{=0}\boldsymbol{\beta}_2 + res. \\ \Rightarrow \hat{\beta}_1 &= (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y} = \frac{\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}}{\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1} \\ \Rightarrow \text{Var}(\hat{\beta}_1) &= \sigma^2 \cdot \frac{1}{\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1} \end{aligned}$$

  Since $\boldsymbol{M}_2\boldsymbol{X}_1$ are the regression residuals of the regression of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$:

$$\begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{X}_2 \cdot \boldsymbol{\theta} + res. \\ \boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1 &= \boldsymbol{X}_1'\boldsymbol{M}_2'\boldsymbol{M}_2\boldsymbol{X}_1 = (\boldsymbol{M}_2\boldsymbol{X}_1)'(\boldsymbol{M}_2\boldsymbol{X}_1) \end{aligned}$$

Multicollinearity: $\boldsymbol{X}_1$ is well explained by $\boldsymbol{X}_2$

$\rightarrow$ SSR small

$\rightarrow \|\boldsymbol{M}_2\boldsymbol{X}_1\|$ small

$\rightarrow \text{Var}(\hat{\beta}_1)$ large

2.2 Gauss-Markov-Theorem

$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ is a linear function of $\boldsymbol{y}$ and unbiased.
A class of linear function of $\boldsymbol{y}$ estimators $\tilde{\boldsymbol{\beta}}$ being unbiased look like:

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{a} + \boldsymbol{A}\boldsymbol{y}$$

Since $E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{a} + \boldsymbol{A} \cdot E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}$, unbiasedness requires:

- $\boldsymbol{a} = \boldsymbol{0}$
- $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}$

Gauss-Markov-Theorem states that within that class of linear unbiased estimators the OLS estimator is the most efficient.

Proof

$$V(\tilde{\boldsymbol{\beta}}) = V(\boldsymbol{A}\boldsymbol{y}) = V(\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{u}) = \boldsymbol{A}\sigma^2\boldsymbol{I}\boldsymbol{A}' = \sigma^2\boldsymbol{A}\boldsymbol{A}'$$

Thus

$$\begin{aligned} V(\tilde{\boldsymbol{\beta}}) - V(\hat{\boldsymbol{\beta}}) &= \sigma^2(\boldsymbol{A}\boldsymbol{A}' - (\boldsymbol{X}'\boldsymbol{X})^{-1}) = \sigma^2(\boldsymbol{A}\boldsymbol{A}' - \underbrace{\boldsymbol{A}\boldsymbol{X}}_{=\boldsymbol{I}}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{A}') \\ &= \sigma^2\boldsymbol{A} \cdot (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{A}' = \sigma^2\boldsymbol{A}\boldsymbol{M}\boldsymbol{A}' \\ &= \sigma^2\boldsymbol{A}\boldsymbol{M}\boldsymbol{M}'\boldsymbol{A}' = \sigma^2(\boldsymbol{A}\boldsymbol{M})(\boldsymbol{A}\boldsymbol{M})' \end{aligned}$$

This matrix is positive semidefinit, since

$$\begin{aligned} \boldsymbol{z}'(\boldsymbol{A}\boldsymbol{M})(\boldsymbol{A}\boldsymbol{M})'\boldsymbol{z} &= (\boldsymbol{z}'\boldsymbol{A}\boldsymbol{M}) \cdot (\boldsymbol{M}'\boldsymbol{A}'\boldsymbol{z}) = (\boldsymbol{M}'\boldsymbol{A}'\boldsymbol{z})'\underbrace{(\boldsymbol{M}'\boldsymbol{A}'\boldsymbol{z})}_{\boldsymbol{w}} \\ &= \boldsymbol{w}'\boldsymbol{w} = \sum w_i^2 \geq 0 \qquad\qquad \forall \boldsymbol{z} \neq \boldsymbol{0} \end{aligned}$$

Taking for instance the i-th unit vector $\boldsymbol{e}_i' = (0, \ldots, 0\ , 1,\ 0, \ldots, 0)' = \boldsymbol{z}'$:

$$\boldsymbol{e}_i'(\boldsymbol{AM})(\boldsymbol{AM})'\boldsymbol{e}_i \geq 0 \Rightarrow \mathrm{Var}(\tilde{\boldsymbol{\beta}}_i) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_i) \geq 0$$

The variances of all components of $\hat{\boldsymbol{\beta}}$ are not larger than corresponding components in alternative $\tilde{\boldsymbol{\beta}}$.

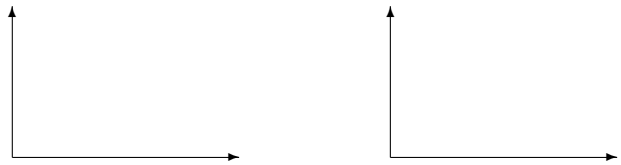
2.3 Some statistical results

- Transformation Theorem

  y has cdf $F(y)$ and pdf $f(y) = F'(y)$

  Transformation $z = h(y)$ with inverse $y = h^{-1}(z)$

  There must be a one-to-one relation $F(y) = G(z)$



  Density $g(z) = G'(z) = \frac{dF(h^{-1}(z))}{dz} = f(h^{-1}(z)) \cdot \left|\frac{dh^{-1}(z)}{dz}\right|$

- Standard normal distribution: $y \sim N(0;1)$

  $\phi(y) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{1}{2}y^2)$

- Linear transformation $z = a + by$

$$\begin{aligned} \Rightarrow y &= \frac{z-a}{b} \quad \Rightarrow \quad \frac{dy}{dz} = \frac{1}{b} \\ f(z) &= \frac{1}{\sqrt{2\pi}}\frac{1}{b}\exp\left(-\frac{1}{2}\left(\frac{z-a}{b}\right)^2\right) = \frac{1}{b}\phi\left(\frac{z-a}{b}\right) \end{aligned}$$

- The sum of two independent $N(0;1)$ random variables is again normally distributed.

- Multivariate normal distribution

  - independent standard normal components: $\boldsymbol{y} \sim N(0; \boldsymbol{I})$

– linear transformation $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{y} \overset{\boldsymbol{B} \text{ full rank}}{\Leftrightarrow} \boldsymbol{y} = \boldsymbol{B}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})$

Transformation theorem (multivariate version) tells us that $\boldsymbol{z}$ is multivariate normal with $E(\boldsymbol{z}) = \boldsymbol{\mu}$ and $V(\boldsymbol{z}) = \boldsymbol{B}\boldsymbol{I}\boldsymbol{B}' = \boldsymbol{B}\boldsymbol{B}' = \boldsymbol{\Sigma}$.

We revert that argument:

$V(\boldsymbol{z}) = \boldsymbol{\Sigma}$ for which we find a decomposition $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}'$

$\Rightarrow \boldsymbol{B}^{-1}(\boldsymbol{z} - \boldsymbol{y})$ is then $N(\boldsymbol{0}; \boldsymbol{I})$

- Chi-Square (χ^2) distribution

  $y_1, \ldots, y_n$ are independent standard normally distributed, i.e. $\boldsymbol{y} \sim N(\boldsymbol{0}; \boldsymbol{I})$

  $z = \sum^n y_i^2 = \|\boldsymbol{y}\|^2 = \boldsymbol{y}'\boldsymbol{y} \sim \chi_n^2$

  It follows that $E(z) = n$ and $\text{Var}(z) = 2 \cdot n$

- If $\boldsymbol{y} \sim N(\boldsymbol{0}; \boldsymbol{\Sigma})$, then

$$\boldsymbol{y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \sim \chi_n^2$$

  Proof

  Assume $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}'$ as above. Then $\boldsymbol{\Sigma}^{-1} = (\boldsymbol{B}')^{-1}\boldsymbol{B}^{-1}$

  $\Rightarrow \boldsymbol{y}'(\boldsymbol{B}')^{-1}\boldsymbol{B}^{-1}\boldsymbol{y} = \boldsymbol{w}'\boldsymbol{w}$

  Since $\boldsymbol{B}^{-1}\boldsymbol{y}$ is $N(.)$ with $V(\boldsymbol{B}^{-1}\boldsymbol{y}) = \boldsymbol{B}^{-1}\boldsymbol{\Sigma}(\boldsymbol{B}^{-1})' = \boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{B}'(\boldsymbol{B}')^{-1} = \boldsymbol{I}$

  $\boldsymbol{w}'\boldsymbol{w} \sim \chi_n^2$

- The orthogonal projection matrix $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ projects orthogonally onto $S(\boldsymbol{X})$ with $\boldsymbol{X}$ a k-dimensional linear subspace

  $\boldsymbol{y} \sim N(\boldsymbol{0}; \boldsymbol{I}) \;\Rightarrow\; \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y} \sim \chi_k^2$

  Proof

$$\underbrace{\boldsymbol{y}'\boldsymbol{X}}_{\boldsymbol{z}'}(\boldsymbol{X}'\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}'\boldsymbol{y}}_{\boldsymbol{z}} = \boldsymbol{z}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{z}$$

  Since $\boldsymbol{y} \sim N(\boldsymbol{0}; \boldsymbol{I})$ it follows that $\boldsymbol{z} = \boldsymbol{X}'\boldsymbol{y} \sim N(\boldsymbol{0}; \boldsymbol{X}'\boldsymbol{I}\boldsymbol{X})$ and therefore $\boldsymbol{z}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{z} \sim \chi_k^2$

- F-distribution

  If $y_1 \sim \chi_{m_1}^2$ and $y_2 \sim \chi_{m_2}^2$ and y_1, y_2 are independent, then

$$\frac{y_1/m_1}{y_2/m_2} \sim F(m_1, m_2)$$

- $\underset{n\times n}{\boldsymbol{A}}$ and $\underset{n\times n}{\boldsymbol{B}}$ are each symmetric and idempotent, $\boldsymbol{z} \sim N(\boldsymbol{0}; \boldsymbol{I})$. Then $\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z}$ and $\boldsymbol{z}'\boldsymbol{B}\boldsymbol{z}$ are independent if and only if $\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{O}$

  Proof

$$\begin{aligned}
\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z} &= \boldsymbol{z}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{z} = \boldsymbol{z}_1'\boldsymbol{z}_1 \quad \text{with } \boldsymbol{z}_1 \sim N(\boldsymbol{0}; \underbrace{\boldsymbol{A}\boldsymbol{I}\boldsymbol{A}'}_{\boldsymbol{A}}) \text{ and} \\
\boldsymbol{z}'\boldsymbol{B}\boldsymbol{z} &= \boldsymbol{z}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{z} = \boldsymbol{z}_2'\boldsymbol{z}_2 \quad \text{with } \boldsymbol{z}_2 \sim N(\boldsymbol{0}; \underbrace{\boldsymbol{B}\boldsymbol{I}\boldsymbol{B}'}_{\boldsymbol{B}}) \\
E(\boldsymbol{z}_1 \cdot \boldsymbol{z}_2') &= E(\boldsymbol{A}\boldsymbol{z}(\boldsymbol{B}\boldsymbol{z})') = E(\boldsymbol{A}\boldsymbol{z}\boldsymbol{z}'\boldsymbol{B}') = \boldsymbol{A}\boldsymbol{I}\boldsymbol{B}'
\end{aligned}$$

2.4 Estimation of σ^2

Since $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is always orthogonal on $S(\boldsymbol{X})$ it is shorter than $\boldsymbol{u}$:

$$\begin{aligned}
\|\hat{\boldsymbol{u}}\| \leq \|\boldsymbol{u}\| &\Leftrightarrow \|\hat{\boldsymbol{u}}\|^2 \leq \|\boldsymbol{u}\|^2 \\
\Rightarrow E(\|\hat{\boldsymbol{u}}\|^2) = E(\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}) &\leq E(\boldsymbol{u}'\boldsymbol{u}) = n \cdot \sigma^2
\end{aligned}$$

Since $\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$ we get
$V(\hat{\boldsymbol{u}}) = \boldsymbol{M}_{\boldsymbol{X}}\sigma^2\boldsymbol{I}\boldsymbol{M}_{\boldsymbol{X}} = \sigma^2\boldsymbol{M}_{\boldsymbol{X}}$

$$\begin{aligned}
E(\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}) &= E(\boldsymbol{u}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}) = \operatorname{tr}(E(\boldsymbol{u}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u})) \\
&= E(\operatorname{tr}(\boldsymbol{u}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u})) = E(\operatorname{tr}(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}\boldsymbol{u}')) = \operatorname{tr}(E(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}\boldsymbol{u}')) \\
&= \operatorname{tr}(\boldsymbol{M}_{\boldsymbol{X}} \cdot \sigma^2\boldsymbol{I}) = \sigma^2(\operatorname{tr}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')) \\
&= \sigma^2(\operatorname{tr}(\boldsymbol{I}) - \operatorname{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) = \sigma^2(n-k)
\end{aligned}$$

We therefore define $s^2 = \frac{1}{n-k}\sum_{t=1}^{n} \hat{u_t}^2 = \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{n-k}$.

2.5 Hypothesis testing

Our model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$; $\boldsymbol{u} \sim IID(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ yields $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$ with $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $V(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
If we assume $\boldsymbol{u} \sim N(\boldsymbol{0}; \sigma^2\boldsymbol{I})$, then

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}; \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Thus

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_k$$

However, it depends on σ^2 which is unknown. Rewriting this yields

$$\begin{aligned}\frac{(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})}{\sigma^2} &= \frac{(\boldsymbol{P}\boldsymbol{y}-\boldsymbol{P}(E(\boldsymbol{y}|\boldsymbol{X})))'(\boldsymbol{P}\boldsymbol{y}-\boldsymbol{P}(E(\boldsymbol{y}|\boldsymbol{X})))}{\sigma^2} \\ &= \frac{(\boldsymbol{P}\boldsymbol{u})'\boldsymbol{P}\boldsymbol{u}}{\sigma^2} = \frac{\boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}}{\sigma^2} \sim \chi^2_k\end{aligned}$$

We just saw in the previous section, that

$$s^2 = \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{n-k} = \frac{\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}}{n-k}$$

Note that $\boldsymbol{u} \sim N(\boldsymbol{0};\sigma^2\boldsymbol{I}) \Rightarrow\ \boldsymbol{u}/\sigma \sim N(\boldsymbol{0};\boldsymbol{I})$ and

$$\frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{\sigma^2} = \frac{\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}}{\sigma^2} \sim \chi^2_{n-k}$$

Since $\boldsymbol{P}_{\boldsymbol{X}} \cdot \boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}}) = \boldsymbol{O}$, $\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}$ is independent of $\boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}$ and therefore

$$\frac{\frac{\boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}}{k}}{\frac{\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}}{n-k}} = \frac{\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{k}}{s^2} \sim F(k,n-k)$$

σ^2 cancels out.

3 Asymptotic Theory

3.1 Probability Limit

We consider an n-vector of random components $\boldsymbol{y}^{(n)}$ and a vector valued function $\boldsymbol{a}(\boldsymbol{y}^{(n)})$.

We call the limiting (random) vector $\boldsymbol{a}_0$ if $\boldsymbol{a}(\boldsymbol{y}^{(n)})$ tends in probability to that $\boldsymbol{a}_0$:

$$\lim_{n\to\infty} P(\|\boldsymbol{a}(\boldsymbol{y}^{(n)}) - \boldsymbol{a}_0\| < \varepsilon) = 1$$

Shortly we write $\text{plim}(\boldsymbol{a}(\boldsymbol{y}^{(n)})) = \boldsymbol{a}_0$ and say $\boldsymbol{a}(\boldsymbol{y}^{(n)})$ is consistent estimator of $\boldsymbol{a}_0$. $\boldsymbol{a}_0$ might be stochastic or nonstochastic.

Example:

$$y_i \sim IID(\mu, \sigma^2) \qquad\qquad i = 1, \ldots, n$$

We consider the arithmetic mean $a(\boldsymbol{y}^{(n)}) = \frac{1}{n}\boldsymbol{y}^{(n)\prime} \cdot \boldsymbol{\iota} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Thus $a(\boldsymbol{y}^{(n)})$ is a scalar function and the Eucledian norm $\|\cdot\|$ reduces to $|\cdot|$.

The first two moments of $a(\boldsymbol{y}^{(n)}) = \bar{y}$:

$$\begin{aligned} E(\bar{y}^{(n)}) &= E\left(\frac{1}{n}\sum y_i\right) = \frac{1}{n}\sum E(y_i) = \mu \\ \text{Var}(\bar{y}^{(n)}) &= \text{Var}\left(\frac{1}{n}\sum y_i\right) \overset{\text{independent}}{=} \left(\frac{1}{n}\right)^2 n\text{Var}(y_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

In order to prove consistency of $\bar{y}$ for μ, we need Chebychev's inequality:

$$P(|Z - E(Z)| > \varepsilon) \leq \frac{\text{Var}(Z)}{\varepsilon^2} \qquad\qquad \forall \varepsilon > 0$$

Applying it to $\bar{y}$:

$$P(|\bar{y}^{(n)} - \mu| > \varepsilon) \leq \frac{\sigma^2}{n \cdot \varepsilon^2}$$

And taking limitis:

$$\lim_{n\to\infty} P(|\bar{y}^{(n)} - \mu| > \varepsilon) \leq 0 \Leftrightarrow \lim P(|\bar{y} - \mu| < \varepsilon) = 1$$

or

$$\text{plim}(\bar{y}^{(n)}) = \mu$$

The example above is called the **Law of Large Numbers(LLN)**. More generally

$$\text{plim}\left(\frac{1}{n}\sum h(y_i)\right) = E(h(y_i))$$

Notes:

- We assumed IID which can be replaced by independence plus the same first moments and bounded variance
- Independence can be relaxed to weak dependence plus weak form of heteroskedasticity:

$$\text{plim}\left(\frac{1}{n}\sum h(y_i)\right) = \lim\left(\frac{1}{n}\sum E(h(y_i))\right)$$

Property

$$\text{plim}\ \boldsymbol{a}(\boldsymbol{y}^{(n)}) = \boldsymbol{a}_0$$

$\boldsymbol{a}_0$ is nonstochastic and $g(.)$ is a smooth function. Then

$$\text{plim}\ g(\boldsymbol{a}(\boldsymbol{y}^{(n)})) = g(\boldsymbol{a}_0)$$

Example

We know that the sequence of random variables $\bar{y}^{(1)}, \bar{y}^{(2)}, \ldots, \bar{y}^{(n)}$ has plim $\bar{y}^{(n)} = \mu$. Thus

$$\text{plim}\ g(\bar{y}^{(n)}) = g(\mu)$$

3.2 Consistency of OLS estimator

In order to focus on $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$ we have to see that

$$\boldsymbol{X}'\boldsymbol{X} = \sum \boldsymbol{X}_t' \cdot \boldsymbol{X}_t \qquad\qquad \boldsymbol{X}_t \text{ is } t^{th} \text{ row of } \boldsymbol{X}$$

Thus ij element of that matrix

$$(\boldsymbol{X}'\boldsymbol{X})_{ij} = \sum^{n} \boldsymbol{X}_{ti} \cdot \boldsymbol{X}_{tj}$$

which is probably growing with increasing n. Therefore, we consider

$$\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)_{ij} = \frac{1}{n}\sum^{n}\boldsymbol{X}_{ti}\cdot\boldsymbol{X}_{tj}$$

for which it is not unreasonable to assume that

$$\text{plim }\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)_{ij} = E(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)$$

as a form of LLN. Specifically, we assume

$$\text{plim }\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right) = \boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}$$

This excludes regressors like a time trend, since e.g.

$$\frac{1}{T}\sum_{t=1}^{T} t = \frac{1}{T}\cdot\frac{T(T+1)}{2}$$

or

$$\frac{1}{T}\sum^{T} t^2$$

are not converging.

Thus we rewrite

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\cdot\frac{1}{n}\boldsymbol{X}'\boldsymbol{u}$$
$$\Rightarrow \text{plim }(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + (\boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}})^{-1}\cdot\text{plim }\frac{1}{n}\boldsymbol{X}'\boldsymbol{u}$$

If we assume that $E(u_t|\boldsymbol{X}_t) = 0$, which is much weaker than $E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}$ implying $E(u_t|\boldsymbol{X}) = 0$, we obtain

$$\boldsymbol{X}_t' E(u_t|\boldsymbol{X}_t) = \boldsymbol{X}_t'\cdot 0 \Leftrightarrow E(\boldsymbol{X}_t' u_t|\boldsymbol{X}_t) = \boldsymbol{0}$$
$$\Leftrightarrow E(\boldsymbol{X}_t'\cdot u_t) = \boldsymbol{0}$$

As a consequence,

$$\text{plim }(\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{X}_t' u_t) = \text{plim }\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{u}\right) \overset{\text{LLN}}{=} E(\boldsymbol{X}_t' u_t) = \boldsymbol{0}$$

As a result,

$$\text{plim }(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

Note that taking expectations and plim are two different concepts.

Examples

- Consider

$$y_t = \beta_1 + \beta_2 \cdot \frac{1}{t} + u_t \qquad u_t \sim IID(0; \sigma^2)$$

OLS is unbiased. But $\hat{\beta}_2$ is not consistent, since as $n \to \infty$ $\lim \frac{1}{t} = 0$. Thus more observations provide less and less information about β_2.

- Estimating population parameter $\mu = E(y_t)$

$$\begin{aligned} \hat{\mu}_1 &= \frac{1}{n-1}\sum_{t=1}^{n} y_t \qquad\qquad \text{plim } \hat{\mu}_1 = \mu \\ \hat{\mu}_2 &= 0,01y_1 + \frac{0,99}{n-1}\sum_{t=2}^{n} y_t \\ E(\hat{\mu}_2) &= 0,01\mu + 0,99 \cdot \frac{1}{n-1}(n-1) \cdot \mu = \mu \\ \text{plim } \hat{\mu}_2 &= 0,99\mu + 0,01y_1 \qquad \text{which is random} \end{aligned}$$

3.3 Asymptotic normality

The second fundamental result is called **central limit theorem (CLT)**. Again, there are various types of CLT's.

Considering $\bar{y}^{(n)}$, for which we know that $\text{Var}(\bar{y}^{(n)}) = \frac{\sigma^2}{n}$. In the limit, this variance is 0. Thus $\bar{y}^{(n)}$ has a degenerate pdf. But note that

$$\sqrt{n}\ \bar{y}^{(n)} \quad \text{has} \quad \text{Var}(\sqrt{n}\ \bar{y}^{(n)}) = n \cdot \text{Var}(\bar{y}^{(n)}) = \sigma^2$$

(Variance stabilizing transformation)

Thus $\sqrt{n}$ is called the rate of convergence

Central Limit Theorem

$$y_t \sim IID(\mu; \sigma^2)$$
$$\lim_{n\to\infty} P\left(\frac{\bar{y}^{(n)} - \mu}{\sigma}\sqrt{n} < z\right) = \Phi(z)$$

We can also write

$$\sqrt{n}\ \left(\frac{\bar{y}^{(n)} - \mu}{\sigma}\right) \xrightarrow{d} N(0;1)$$

or we write

$$\bar{y}^{(n)} \overset{a}{\sim} N\left(\mu; \frac{\sigma^2}{n}\right)$$

Reformulating $\bar{y}^{(n)}$ in the above version of CLT gives the CLT which is most common in econometrics:

CLT

Let y_t be a sequence of random variables, $t = 1, \ldots, \infty$ with $E(y_t) = 0$, then

$$\text{plim } \frac{1}{\sqrt{n}} \sum y_t = y_0 \sim N\left(0; \lim_{n\to\infty} \frac{1}{n} \sum \text{Var}(y_t)\right)$$

In a multivariate version:

Suppose we have a sequence of random m-vectors (m fixed)$\boldsymbol{y}_t$ with $E(\boldsymbol{y}_t) = \boldsymbol{0}$, then

$$\text{plim } \frac{1}{\sqrt{n}} \sum \boldsymbol{y}_t = \boldsymbol{y}_0 \sim N(\boldsymbol{0}; \lim_{n\to\infty} \frac{1}{n} \sum V(\boldsymbol{y}_t))$$

Now we turn to our regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \qquad\qquad \boldsymbol{u} \sim IID(0; \sigma^2 \boldsymbol{I})$$

with $E(u_t|\boldsymbol{X}_t) = 0$ and $E(u_t^2|\boldsymbol{X}_t) = \sigma^2$, plim $(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}) = \boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}$.

Considering

$$\boldsymbol{v} = \frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{u} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t \cdot \boldsymbol{X}_t'$$

By assumption

$$\begin{aligned} E(u_t|\boldsymbol{X}_t) = 0 \quad &\Rightarrow \quad E_X[\boldsymbol{X}_t' E(u_t|\boldsymbol{X}_t)] = E_X E(u_t \boldsymbol{X}_t'|\boldsymbol{X}_t) = \boldsymbol{0} \\ &\Leftrightarrow \quad E(u_t \boldsymbol{X}_t') = \boldsymbol{0} \end{aligned}$$

so that we can apply the multivariate CLT:

$$\boldsymbol{v} \overset{a}{\sim} N\left(\boldsymbol{0}; \lim_{n\to\infty} \frac{1}{n} \sum \text{Var}(u_t \boldsymbol{X}_t')\right) = N\left(\boldsymbol{0}; \lim \frac{1}{n} \sum E(u_t^2 \boldsymbol{X}_t' \boldsymbol{X}_t)\right)$$

The covariance matrix is

(recall that $E(u_t^2|\boldsymbol{X}_t) = \sigma^2 \Leftrightarrow E_X[\boldsymbol{X}_t'\boldsymbol{X}_t E(u_t^2|\boldsymbol{X}_t)] = \sigma^2 E(\boldsymbol{X}_t'\boldsymbol{X}_t)$)

$$\begin{aligned} \lim \frac{1}{n} \sum E(u_t^2 \boldsymbol{X}_t' \boldsymbol{X}_t) &= \lim \sigma^2 \underbrace{\frac{1}{n} \sum E(\boldsymbol{X}_t' \boldsymbol{X}_t)}_{\text{LLN}} \\ &= \text{plim } \sigma^2 \frac{1}{n} \sum \boldsymbol{X}_t' \boldsymbol{X}_t \\ &= \text{plim } \sigma^2 \frac{1}{n} \boldsymbol{X}'\boldsymbol{X} = \sigma^2 \cdot \boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}} \end{aligned}$$

Thus

$$\boldsymbol{v} \overset{a}{\sim} N(\mathbf{0}; \sigma^2 \cdot \boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}})$$

and therefore

$$\begin{aligned} \text{plim } \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \cdot \text{plim } \boldsymbol{v} &= \boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}^{-1} \cdot \text{plim } \boldsymbol{v} \\ &= \boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}^{-1}\boldsymbol{v}_0 \sim N(\mathbf{0}; \sigma^2 \boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}^{-1}\boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}\boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}^{-1'}) = N(\mathbf{0}; \sigma^2 \boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}^{-1}) \end{aligned}$$

Thus we obtain for $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\frac{1}{n}\boldsymbol{X}'\boldsymbol{X})^{-1} \cdot \frac{1}{n}\boldsymbol{X}'\boldsymbol{u}$ and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \underbrace{\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{u}}_{\boldsymbol{v}} \sim N(\mathbf{0}; \sigma^2 \boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}^{-1}) \text{ or } \hat{\boldsymbol{\beta}} \overset{a}{\sim} N\left(\boldsymbol{\beta}, \frac{\sigma^2}{n}\boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}^{-1}\right)$$

4 Generalized Method of Moments

4.1 Method-of-Moments estimation

The basic idea of the method of moments is to replace population means by sample means.

For our model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \qquad\qquad \boldsymbol{u} \sim IID$$

we assume $E(u_t|\boldsymbol{X}_t) = 0$ which implies

$$\begin{aligned} \boldsymbol{X}_t' \cdot E(u_t|\boldsymbol{X}_t) = \boldsymbol{X}_t' \cdot 0 = \boldsymbol{0} \Leftrightarrow E(\boldsymbol{X}_t' u_t|\boldsymbol{X}_t) = \boldsymbol{0} \\ \Leftrightarrow E_{X_t} E(\boldsymbol{X}_t' u_t|\boldsymbol{X}_t) = E(\boldsymbol{X}_t' u_t) = E_{X_t}(\boldsymbol{0}) = \boldsymbol{0} \end{aligned}$$

Thus replacing $E(\boldsymbol{X}_t' u_t)$ by its sample counterpart

$$\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{X}_t' u_t = \frac{1}{n}\boldsymbol{X}'\boldsymbol{u} = \frac{1}{n}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \stackrel{!}{=} \boldsymbol{0}$$

Multiplication with n yields the LS normal equations and therefore we obtain

$$\hat{\boldsymbol{\beta}}_{MM} = \hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

4.2 Errors-in-Variables

Suppose the true model is

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{u}^* \qquad\qquad \boldsymbol{u}^* \sim IID(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{3}$$

A problem arises whenever $\boldsymbol{X}^*$ is not observed directly. Instead we observe $\boldsymbol{X}^*$ with some measurement error:

$$\boldsymbol{X} = \boldsymbol{X}^* + \boldsymbol{V}$$

Inserting this into model (3) yields

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u}^*$$

Thus the new error term is $\boldsymbol{u} = \boldsymbol{u}^* - \boldsymbol{V}\boldsymbol{\beta}$ and it is quite obvious that $u_t = u_t^* - \boldsymbol{V}_t\boldsymbol{\beta}$ is correlated with $\boldsymbol{X}_t = \boldsymbol{X}_t^* + \boldsymbol{V}_t$. Thus $E(\boldsymbol{X}_t'\boldsymbol{u}_t) \neq \boldsymbol{0}$.

However, if we find instruments $\boldsymbol{W}_t$ that are uncorrelated with u_t implying that they are uncorrelated with $\boldsymbol{V}_t$ but highly correlated with $\boldsymbol{X}_t$ we obtain the MM estimator:

$$\begin{aligned} \frac{1}{n}\sum \boldsymbol{W}_t' \cdot u_t &= \frac{1}{n}\boldsymbol{W}'\boldsymbol{u} = \frac{1}{n}\boldsymbol{W}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \stackrel{!}{=} 0 \\ \Leftrightarrow \boldsymbol{W}'\boldsymbol{y} &= \boldsymbol{W}'\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

Having exactly as many instruments as regressors and $\boldsymbol{W}'\boldsymbol{X}$ having full rank, we get

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{IV} &= (\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{W}'\boldsymbol{y} \\ &= (\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{W}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{W}'\boldsymbol{u} \\ &= \boldsymbol{\beta} + (\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{W}'\boldsymbol{u} \end{aligned}$$

This is a linear transformation in $\boldsymbol{y}$ and $\boldsymbol{y}$ is projected onto $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{W}'\boldsymbol{y}$ so that the projector now is

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{W}'$$

For consistency we write $\hat{\boldsymbol{\beta}}_{IV} = \boldsymbol{\beta} + (n^{-1}\boldsymbol{W}'\boldsymbol{X})^{-1}(n^{-1}\boldsymbol{W}'\boldsymbol{u})$ and see that if

$$\operatorname{plim} \frac{1}{n}\boldsymbol{W}'\boldsymbol{X} = \boldsymbol{S}_{\boldsymbol{W}'\boldsymbol{X}}$$

we obtain consistency of $\hat{\boldsymbol{\beta}}_{IV}$ since

$$\operatorname{plim} \frac{1}{n}\boldsymbol{W}'\boldsymbol{u} = \operatorname{plim} \frac{1}{n}\sum \boldsymbol{W}_t' u_t \stackrel{\text{LLN}}{=} \lim \frac{1}{n}\sum E(\boldsymbol{W}_t\boldsymbol{u}_t) \stackrel{\text{assumption } E(u_t|\boldsymbol{W}_t)=0}{=} \boldsymbol{0}$$

Examining $\sqrt{n}(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) = n^{-1}(\boldsymbol{W}'\boldsymbol{X})^{-1}n^{-\frac{1}{2}}\boldsymbol{W}'\boldsymbol{u}$ we can apply a multivariate CLT as above and obtain

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) \stackrel{a}{\sim} N(\boldsymbol{0}; \sigma^2 \boldsymbol{S}_{\boldsymbol{W}'\boldsymbol{X}}^{-1}\boldsymbol{S}_{\boldsymbol{W}'\boldsymbol{W}}(\boldsymbol{S}_{\boldsymbol{W}'\boldsymbol{X}}^{-1})')$$

where $\operatorname{plim} \frac{1}{n}\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{S}_{\boldsymbol{W}'\boldsymbol{W}}$. Thus the asymptotic variance is

$$\begin{aligned} \sigma^2 \boldsymbol{S}_{\boldsymbol{W}'\boldsymbol{X}}^{-1}\boldsymbol{S}_{\boldsymbol{W}'\boldsymbol{W}}(\boldsymbol{S}_{\boldsymbol{W}'\boldsymbol{X}}^{-1})' &= \sigma^2 \operatorname{plim} \left(\frac{1}{n}\boldsymbol{W}'\boldsymbol{X}\right)^{-1} \frac{1}{n}\boldsymbol{W}'\boldsymbol{W} \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{W}\right)^{-1} \\ &= \sigma^2 \operatorname{plim} \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{W}\left(\frac{1}{n}\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\frac{1}{n}\boldsymbol{W}'\boldsymbol{X}\right)^{-1} \end{aligned}$$

$$= \sigma^2 \text{plim} \left(\frac{1}{n} \boldsymbol{X}' \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} \right)^{-1} \tag{4}$$

If we could choose instruments $\boldsymbol{W}$ we could use those ones implying the least $V(\sqrt{n}\,(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}))$.

If we write for the explanatory variables

$$\boldsymbol{X} = \bar{\boldsymbol{X}} + \boldsymbol{z} \qquad\qquad E(z_t|\Omega_t) = 0$$

where Ω_t is the information set in t and $\bar{\boldsymbol{X}}_t$ denoting the t-th row of $\bar{\boldsymbol{X}}$.

These conditional expectations of $\boldsymbol{X}_t$ are the optimal instruments. To see that we first note that

$$\begin{aligned} \text{plim}\, \frac{1}{n} \boldsymbol{X}'\boldsymbol{W} &= \lim \frac{1}{n} E(\boldsymbol{X}'\boldsymbol{W}) = \lim \frac{1}{n} E((\bar{\boldsymbol{X}} + \boldsymbol{z})'\boldsymbol{W}) \\ &= \lim \frac{1}{n} E(\bar{\boldsymbol{X}}'\boldsymbol{W}) = \text{plim} \frac{1}{n} \bar{\boldsymbol{X}}'\boldsymbol{W} \end{aligned}$$

Thus the covariance matrix becomes to $\sigma^2 \text{plim}\, (\frac{1}{n} \bar{\boldsymbol{X}}' \boldsymbol{P}_W \bar{\boldsymbol{X}})^{-1}$.

Now setting $\boldsymbol{W} = \bar{\boldsymbol{X}}$ we get

$$\sigma^2 \text{plim} \left(\frac{1}{n} \bar{\boldsymbol{X}}' \boldsymbol{W} (\boldsymbol{W}'\boldsymbol{W})^{-1} \boldsymbol{W}' \bar{\boldsymbol{X}} \right)^{-1} = \sigma^2 \text{plim} \left(\frac{1}{n} \bar{\boldsymbol{X}}' \bar{\boldsymbol{X}} \right)^{-1} \tag{5}$$

Considering the differences between (4) and (5) or rather their inverses (to show pos. semi-definiteness of $\boldsymbol{A} - \boldsymbol{B}$ it is equivalent to prove pos. semi-definiteness of $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}$)

$$\sigma^2 \text{plim} \left[\frac{1}{n} (\bar{\boldsymbol{X}}' \bar{\boldsymbol{X}} - \bar{\boldsymbol{X}}' \boldsymbol{P}_W \bar{\boldsymbol{X}}) \right] = \sigma^2 \text{plim} \frac{1}{n} \bar{\boldsymbol{X}}' \boldsymbol{M}_W \bar{\boldsymbol{X}}$$

which is a positiv semi-definit matrix showing the optimality of using $\bar{\boldsymbol{X}}$ as instruments.

So far we assumed that $\boldsymbol{W}$ has the same number of columns as $\boldsymbol{X}$. What if we have more instruments than regressors?

The basic idea is to use k linear combinations of $\boldsymbol{W}$ (e.g. $\boldsymbol{W} \cdot \boldsymbol{J}$) to obtain a just identified problem:

$$(\boldsymbol{W}\boldsymbol{J})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \stackrel{!}{=} \boldsymbol{0} \Leftrightarrow \boldsymbol{J}'\boldsymbol{W}'\boldsymbol{y} = \boldsymbol{J}'\boldsymbol{W}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \Leftrightarrow \hat{\boldsymbol{\beta}}_{GIVE} = (\boldsymbol{J}'\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{J}'\boldsymbol{W}'\boldsymbol{y}$$

This generalized estimator $\sqrt{n}(\hat{\boldsymbol{\beta}}_{GIVE} - \hat{\boldsymbol{\beta}})$ has asymptotic covariance matrix

$$\sigma^2 \text{plim}\, n(\boldsymbol{J}'\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{J}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{J}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{J})^{-1}$$

The sandwich form tells us that it is not the most efficient one. If we choose $\boldsymbol{J} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{X}$ the asymptotic covariance becomes:

$$\begin{aligned}\sigma^2\text{plim} \quad & \left(\frac{1}{n}(\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\right. \\ & \cdot \quad \left.\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{X})^{-1}\right) \\ = \quad & \sigma^2\text{plim}\, \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{X}\right)^{-1} = \sigma^2\text{plim}\, \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\end{aligned}$$

Thus our generalized IV estimator is

$$\hat{\boldsymbol{\beta}}_{GIVE} = (\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y} \tag{6}$$

The form of the optimal $\boldsymbol{J}$ is not surprising. $\boldsymbol{W}\boldsymbol{J}$ are the orthogonal projections of $\boldsymbol{X}$ onto $S(\boldsymbol{W})$. Thus each column of $\boldsymbol{W}\boldsymbol{J}$ has shortest distance to the corresponding column of $\boldsymbol{X}$.

From equation (6) it is quite obvious that $\hat{\boldsymbol{\beta}}_{GIVE}$ is also called the 2SLS estimator.

It is important to note that $\hat{\boldsymbol{\beta}}_{GIVE}$ is also the solution to

$\min Q(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$.

4.3 Generalized Method of Moments

In the linear regression models we have the unconditional expectation

$$E(u_t) = 0$$

which we replace by the sample counterpart

$$\frac{1}{n}\sum_{t=1}^{n} u_t = \frac{1}{n}\sum_{t=1}^{n}(y_t - \boldsymbol{X}_t\boldsymbol{\beta}) = 0$$

We have one equation with k unknown parameters.

Introducing k instruments fulfilling the predetermined condition we get

$$\frac{1}{n}\sum \boldsymbol{W}_t'(y_t - \boldsymbol{X}_t\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{W}'\boldsymbol{y} - \boldsymbol{W}'\boldsymbol{X}\boldsymbol{\beta} \stackrel{!}{=} \boldsymbol{0}$$

In this case, the moment condition is linear in $\boldsymbol{\beta}$. The GMM allows $u_t = u_t(\boldsymbol{\beta})$ to be nonlinear.

By the same token, we need at least k instruments in $\boldsymbol{W}_t$ to solve

$$\frac{1}{n}\sum \boldsymbol{W}_t \cdot \boldsymbol{u}_t(\boldsymbol{\beta}) \stackrel{!}{=} \mathbf{0}$$

Or more generally we allow for the just identified case

$$\boldsymbol{g}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum \boldsymbol{f}(y_t, \boldsymbol{X}_t, \boldsymbol{W}_t, \boldsymbol{\beta}) \stackrel{!}{=} \mathbf{0}$$

If we have more instruments than parameters the GMM estimator minimizes the function

$$Q(\boldsymbol{\beta}) = \boldsymbol{g}_n(\boldsymbol{\beta})' \boldsymbol{\Psi} \boldsymbol{g}_n(\boldsymbol{\beta}) \tag{7}$$

with weighting matrix $\boldsymbol{\Psi}$.

The first order conditions of minimizing (7) turn out to be

$$\boldsymbol{G}'\boldsymbol{\Psi}\boldsymbol{g} \stackrel{!}{=} \mathbf{0} \qquad \text{with } \boldsymbol{G} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\beta}'}$$

Since $E(\boldsymbol{f}) = \mathbf{0}$ the covariance matrix $V(\boldsymbol{f}) = E(\boldsymbol{f}\boldsymbol{f}')$. Thus the GMM estimator has asymptotic covariance matrix

$$\boldsymbol{G}'\boldsymbol{\Psi}V(\boldsymbol{f})\boldsymbol{\Psi}'\boldsymbol{G}$$

Again, the sandwich form disappears if we choose $\boldsymbol{\Psi} = \boldsymbol{\Psi}'$ optimally as $\boldsymbol{\Psi}_{opt} = V(\boldsymbol{f})^{-1}$.

5 Maximum Likelihood

We consider now fully specified parametric models, i.e. we know the joint pdf

$$f(\boldsymbol{y}, \boldsymbol{\theta})$$

For given $\boldsymbol{\theta}$ we obtain the pdf of vector $\boldsymbol{y}$. However, if we interpret $f(\boldsymbol{y}, \boldsymbol{\theta})$ for a given vector $\boldsymbol{y}$ it is now called a likelihood function for a given data set.
Stochastically independent y_t thus yield:

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = \prod_{t=1}^{n} f(y_t, \boldsymbol{\theta})$$

The parameter vector $\hat{\boldsymbol{\theta}}$ that maximizes $f(\boldsymbol{y}, \boldsymbol{\theta})$ ist called MLE.
For computational and analytical ease a strictly monotonic transformation of $f(\boldsymbol{y}, \boldsymbol{\theta})$ is considered:

$$\ell(\boldsymbol{y}, \boldsymbol{\theta}) \equiv \log f(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(y_t, \boldsymbol{\theta})$$

$\ell_t(y_t, \boldsymbol{\theta})$ is called the t^{th} contribution to loglikelihood function.

Example 1 Exponential Distribution

$$f(y_t, \theta) = \theta \exp(-\theta y_t) \qquad y_t > 0; \theta > 0$$
$$\ell(\boldsymbol{y}, \theta) = \sum_{t=1}^{n} (\log \theta - \theta y_t) = n \log \theta - \theta \sum y_t$$

FOC:

$$\begin{aligned} \frac{d\ell}{d\theta} &= \frac{n}{\theta} - \sum y_t = 0 \\ \Leftrightarrow \hat{\theta} &= \frac{n}{\sum y_t} = \frac{1}{\bar{y}} \end{aligned}$$

Example 2 Regression with normal errors

$$\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

The elements u_t are independently distributed as $N(0;\sigma^2)$. Thus y_t conditional on $\boldsymbol{X}_t$ has the same distribution law:

$$\begin{aligned} f_t(y_t,\boldsymbol{\beta},\sigma|\boldsymbol{X}_t) &= \frac{1}{\sigma}\phi\left(\frac{y_t-\boldsymbol{X}_t\boldsymbol{\beta}}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y_t-\boldsymbol{X}_t\boldsymbol{\beta}}{\sigma}\right)^2\right) \\ \ell_t(y_t,\boldsymbol{\beta},\sigma|\boldsymbol{X}_t) &= -\frac{1}{2}\log(2\pi)-\frac{1}{2}\log\sigma^2-\frac{1}{2\sigma^2}(y_t-\boldsymbol{X}_t\boldsymbol{\beta})^2 \\ \ell(\boldsymbol{y},\boldsymbol{\beta},\sigma|\boldsymbol{X}) &= -\frac{n}{2}\log(2\pi)-\frac{1}{2}\log\sigma^2-\frac{n}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

FOC:

$$\begin{aligned} \frac{\partial\ell}{\partial\boldsymbol{\beta}} &= -\frac{1}{2\sigma^2}(-2\boldsymbol{X}'\boldsymbol{y}+2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}) \stackrel{!}{=} \mathbf{0} \\ \frac{\partial\ell}{\partial\sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2(\sigma^2)^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \stackrel{!}{=} 0 \end{aligned}$$

The first set of equations obviously yields $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{OLS}$. Inserting this into the second equation obtains

$$\hat{\sigma}^2_{ML} = \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{n}$$

As the examples showed most MLE can be characterized by

$$\left.\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}} = \mathbf{0}$$

Since $\ell(\boldsymbol{\theta}) = \sum\ell_t(\boldsymbol{\theta})$, it follows that we can write the FOC:

$$\boldsymbol{g}(\boldsymbol{\theta}) = \frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum\frac{\partial\ell_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum\boldsymbol{g}_t(\boldsymbol{\theta}) \stackrel{!}{=} \mathbf{0}$$

$\boldsymbol{g}(\boldsymbol{\theta})$ is called gradient vector or score vector. The true parameter vector is denoted as $\boldsymbol{\theta}_0$ which generated the data set $\boldsymbol{y}$. Thus

$$\begin{aligned} E(\boldsymbol{g}(\boldsymbol{\theta}_0)) &= E\left(\frac{\partial\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right) = E\left(\frac{\partial\log f(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right) = E\left(\frac{1}{f(\boldsymbol{\theta}_0)}\cdot\frac{\partial f(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right) \\ &= \int\frac{1}{f(\boldsymbol{\theta}_0)}\cdot\frac{\partial f(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\cdot f(\boldsymbol{\theta}_0)d\boldsymbol{y} = \int\frac{\partial f(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}d\boldsymbol{y} = \mathbf{0} \end{aligned}$$

This last equality follows from the fact that f is a pdf and thus $\int f(\boldsymbol{y},\boldsymbol{\theta})d\boldsymbol{y} = 1$. Differentiating both sides yields

$$\frac{\partial}{\partial\boldsymbol{\theta}}\int f(\boldsymbol{y},\boldsymbol{\theta})d\boldsymbol{y} = \int\frac{\partial f(\boldsymbol{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}d\boldsymbol{y} = \mathbf{0}$$

Differentiating both sides of $E(\boldsymbol{g}(\boldsymbol{\theta}_0)) = \mathbf{0}$ once again and using the fact that

$$\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \frac{\partial \log f(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \frac{1}{f(\boldsymbol{\theta}_0)} \cdot \frac{\partial f(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$$

we get

$$\int \frac{\partial^2 f(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} d\boldsymbol{y} = \mathbf{0} \quad \Leftrightarrow \quad \int \frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \cdot f(\boldsymbol{\theta}_0) d\boldsymbol{y} + \int \frac{\partial \ell}{\partial \boldsymbol{\theta}} \cdot \frac{\partial f}{\partial \boldsymbol{\theta}} d\boldsymbol{y} = \mathbf{0}$$

$$\Leftrightarrow \quad -E\left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) = E\left(\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right)$$

The left hand side of this equation is called information matrix which is usually denoted as $\boldsymbol{I}(\boldsymbol{\theta})$ and it is the covariance matrix of the score vector $\boldsymbol{g}(\boldsymbol{\theta}_0)$.

Consistency

Due to Jensens's inequality we know that

$$E \log\left(\frac{f(\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}_0)}\right) < \log E\left(\frac{f(\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}_0)}\right)$$

As we have just seen

$$E\left(\frac{f(\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}_0)}\right) = \int \frac{f(\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}_0)} \cdot f(\boldsymbol{\theta}_0) d\boldsymbol{y} = \int f(\boldsymbol{\theta}^*) d\boldsymbol{y} = 1$$

Thus we get

$$E(\ell(\boldsymbol{\theta}^*)) - E(\ell(\boldsymbol{\theta}_0)) < 0 \Leftrightarrow E(\ell(\boldsymbol{\theta}^*)) < E(\ell(\boldsymbol{\theta}_0))$$

Applying a LLN to the contributions ℓ_t we obtain $\operatorname{plim} n^{-1}\ell(\boldsymbol{\theta}) = \lim n^{-1} E(\ell(\boldsymbol{\theta}))$:

$$\operatorname{plim} \frac{1}{n}\ell(\boldsymbol{\theta}^*) \leq \operatorname{plim} \ell(\boldsymbol{\theta}_0)$$

This holds for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$. The strict inequality disappears due to taking limits. Since the MLE $\hat{\boldsymbol{\theta}}$ maximizes $\ell(\boldsymbol{\theta})$ it must hold that

$$\operatorname{plim} \frac{1}{n}\ell(\hat{\boldsymbol{\theta}}) \geq \operatorname{plim} \ell(\boldsymbol{\theta}_0)$$

Asymptotic normality of MLE

Performing a Taylor expansion of the likelihood equation $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ around $\boldsymbol{\theta}_0$ yields

$$\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \boldsymbol{g}(\boldsymbol{\theta}_0) + \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0} \tag{8}$$

where $\boldsymbol{H}(\boldsymbol{\theta}) = \partial^2 \ell / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$ is the Hessian of the loglikelihood function. Since $\bar{\boldsymbol{\theta}}$ must lie somewhere between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ we may write

$$\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$$

Solving equation (8) and inserting appropriate factors of n we get:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\left(\frac{1}{n}\boldsymbol{H}(\bar{\boldsymbol{\theta}})\right)^{-1} \frac{1}{\sqrt{n}}\boldsymbol{g}(\boldsymbol{\theta}_0)$$

Components of the score vector $\boldsymbol{g}_i(\boldsymbol{\theta}_0) = \sum_{t=1}^{n} \frac{\partial \ell_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_i}$ having mean of 0 multiplied with $n^{-1/2}$ is asymptotically normally distributed according to CLT.
Since $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$, vector $\bar{\boldsymbol{\theta}}$ is also consistent.
Thus plim $n^{-1}\boldsymbol{H}(\bar{\boldsymbol{\theta}})$ = plim $n^{-1}\boldsymbol{H}(\hat{\boldsymbol{\theta}})$ = plim $n^{-1}\boldsymbol{H}(\boldsymbol{\theta}_0) = \mathcal{H}(\boldsymbol{\theta}_0)$. Therefore, it follows with the help of $-\mathcal{H}(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0) =$ plim $n^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)$

$$\frac{1}{\sqrt{n}}\,\boldsymbol{g}(\boldsymbol{\theta}_0) \overset{a}{\sim} N(\mathbf{0}; \mathcal{I}(\boldsymbol{\theta}_0))$$

Thus

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} N(\mathbf{0}; \underbrace{\mathcal{H}(\boldsymbol{\theta}_0)^{-1}\mathcal{I}(\boldsymbol{\theta}_0)\mathcal{H}(\boldsymbol{\theta}_0)^{-1}}_{=\mathcal{I}(\boldsymbol{\theta}_0)^{-1}})$$

Computational Issues

Solving a system of nonlinear equations

$$\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$$

involves iterative maximization procedure like Newton-Raphson

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \boldsymbol{H}(\boldsymbol{\theta}^{(j)})^{-1} \cdot \boldsymbol{g}(\boldsymbol{\theta}^{(j)})$$

It involves the Hessian which can be replaced by approximations to it. One approach is to use the information equality and use $\boldsymbol{D}$ instead of $\boldsymbol{H}$ with

$$\boldsymbol{D} = \sum \frac{\partial \ell_t}{\partial \boldsymbol{\theta}} \cdot \left(\frac{\partial \ell_t}{\partial \boldsymbol{\theta}} \right)'$$

The inverse of this matrix or of $-\boldsymbol{H}(\hat{\boldsymbol{\theta}})$ or of $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ are all common estimates for the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$.

Delta Theorem

Suppose a k-vector $\boldsymbol{\theta}$ with $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} N(\mathbf{0}; V(\hat{\boldsymbol{\theta}}))$ and the l-vector $(l < k)\boldsymbol{\gamma} = \boldsymbol{g}(\boldsymbol{\theta})$ with $\boldsymbol{g}(.)$ having l monotonic and differentiable functions g_i, then

$$\sqrt{n}\,(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \overset{a}{\sim} N(\mathbf{0}; \boldsymbol{G}_0 V(\hat{\boldsymbol{\theta}}) \boldsymbol{G}_0')$$

with $\boldsymbol{G}_0$ a $l \times k$ matrix with typical element $\partial g_i / \partial \theta_j$ evaluated at $\boldsymbol{\theta}_0$. In practice we estimate this covariance matrix by $\boldsymbol{G}(\hat{\boldsymbol{\theta}}) \widehat{V(\hat{\boldsymbol{\theta}})} \boldsymbol{G}(\hat{\boldsymbol{\theta}})'$.

5.1 Hypothesis tests

Testing hypothesis of type $\boldsymbol{r}(\boldsymbol{\theta}) = \mathbf{0}$ having r components we have three asymptotically identical tests:

Likelihood ratio test (LR)

$$LR = 2(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}}))$$

where $\tilde{\boldsymbol{\theta}}$ denotes the restricted MLE fulfilling $\boldsymbol{r}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$. Taylor expansion of $\ell(\tilde{\boldsymbol{\theta}})$ around $\hat{\boldsymbol{\theta}}$ obtains

$$\begin{aligned} \ell(\tilde{\boldsymbol{\theta}}) &= \ell(\hat{\boldsymbol{\theta}}) + \boldsymbol{g}(\hat{\boldsymbol{\theta}})'(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})' \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \\ \Leftrightarrow \; & 2(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})) = -(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})' \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \end{aligned}$$

Intuitively, it shows that LR is approximatly χ^2_r distributed.

Wald test

$$W = \boldsymbol{r}(\hat{\boldsymbol{\theta}})'[\boldsymbol{R}\boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{R}']^{-1}\boldsymbol{r}(\hat{\boldsymbol{\theta}})$$

with $r \times k$ matrix $\boldsymbol{R}$ having typical element $\partial \boldsymbol{r}_i / \partial \theta_j$. Since $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed and under $H_0 : \boldsymbol{r}(\boldsymbol{\theta}) = \mathbf{0}$ we can apply the Delta Theorem and get

$$\boldsymbol{r}(\hat{\boldsymbol{\theta}}) \overset{a}{\sim} N(\mathbf{0}; \boldsymbol{R}\boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{R}')$$

The asymptotic distribution of the quadratic form is thus χ^2_r.

Lagrange multiplier test (LM)

$$LM = \boldsymbol{g}(\tilde{\boldsymbol{\theta}})'\boldsymbol{I}(\tilde{\boldsymbol{\theta}})^{-1}\boldsymbol{g}(\tilde{\boldsymbol{\theta}})$$

The FOC of the restricted otpimization problem are

$$\begin{aligned} \boldsymbol{g}(\tilde{\boldsymbol{\theta}}) + \boldsymbol{R}(\tilde{\boldsymbol{\theta}})'\boldsymbol{\lambda} &= \mathbf{0} \\ \boldsymbol{r}(\tilde{\boldsymbol{\theta}}) &= \mathbf{0} \end{aligned}$$

Under the null hypothesis $\boldsymbol{r}(\boldsymbol{\theta}_0) = \mathbf{0}$ we would get $\boldsymbol{\lambda} = \mathbf{0}$ implying $\boldsymbol{g}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$. Thus the quadratic form $\boldsymbol{g}(\tilde{\boldsymbol{\theta}})'\boldsymbol{I}(\tilde{\boldsymbol{\theta}})^{-1}\boldsymbol{g}(\tilde{\boldsymbol{\theta}})$ should be asymptotically χ^2_r distributed.

6 Generalized Least Squares

We now assume a general linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \qquad ; E(\boldsymbol{u}) = \boldsymbol{0} \quad , \quad E(\boldsymbol{u}\boldsymbol{u}') = \boldsymbol{\Omega} \tag{9}$$

Clearly, if $\boldsymbol{\Omega} = \sigma^2 \boldsymbol{I}$ we are back to the classical model.

First we can note that $\hat{\boldsymbol{\beta}}_{OLS}$ is still unbiased given that either $\boldsymbol{X}$ is fix or uncorrelated wit $\boldsymbol{u}$:

$$E(\hat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u})$$

However, its covariance matrix now is

$$V(\hat{\boldsymbol{\beta}}_{OLS}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u}\boldsymbol{u}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

As a consequence, inference based on the covariance matrix $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is not valid anymore.

6.1 GLS estimator

The covariance matrix $\boldsymbol{\Omega}$ is a positive definite matrix implying that its inverse can be decomposed

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}' \qquad \Rightarrow \qquad \boldsymbol{\Omega} = (\boldsymbol{\Psi}')^{-1}\boldsymbol{\Psi}^{-1}$$

Premultiplying equation (9) with $\boldsymbol{\Psi}'$ yields

$$\boldsymbol{\Psi}'\boldsymbol{y} = \boldsymbol{\Psi}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Psi}'\boldsymbol{u}$$

Analyzing the transformed error terms we see

$$V(\boldsymbol{\Psi}'\boldsymbol{u}) = \boldsymbol{\Psi}'V(\boldsymbol{u})\boldsymbol{\Psi} = \boldsymbol{\Psi}'\boldsymbol{\Omega}\boldsymbol{\Psi} = \boldsymbol{\Psi}'(\boldsymbol{\Psi}')^{-1}\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi} = \boldsymbol{I}$$

Thus we have error terms fulfilling the classical assumptions. For this model we know due to Gauss-Markov theorem the most efficient estimator

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{\Psi}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{\Psi}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}$$

The covariance matrix of this estimator follows immediately

$$V(\hat{\boldsymbol{\beta}}_{GLS}) = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}$$

This estimator can also be obtained if we minimize the criterion function

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

which is the sum of squared residuals from the transformed model. The first order conditions for this optimization problem are

$$\boldsymbol{X}'\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}) = \boldsymbol{0}$$

If we compare it to the instrumental variables estimator (or more general to GMM) we see that setting $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}\boldsymbol{X}$ the moment conditions

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

follow. From that the method of moments yields the estimator

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} = (\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{W}'\boldsymbol{y} = \boldsymbol{\beta} + (\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{W}'\boldsymbol{u}$$

Therefore, the covariance matrix of $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ is

$$V(\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}) = (\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{W}'\boldsymbol{\Omega}\boldsymbol{W}(\boldsymbol{X}'\boldsymbol{W})^{-1}$$

Again, the sandwich form indicates that it is not efficient. Note, that setting $\boldsymbol{W} = \boldsymbol{X}$ we are back to the above analyzed OLS estimator.

To prove efficiency of $\hat{\boldsymbol{\beta}}_{GLS}$ we first show that $\boldsymbol{A} - \boldsymbol{B}$ is positive definite if and only if $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}$ is positive definite. To see that we start with a symmetric positive definite matrix $\boldsymbol{A}$ and a full column rank matrix $\boldsymbol{C}$. It follows that $\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C}$ is positive definite:

$$\boldsymbol{x}'\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C}\boldsymbol{x} = (\boldsymbol{x}'\boldsymbol{C}')\boldsymbol{A}(\boldsymbol{C}\boldsymbol{x}) = (\boldsymbol{C}\boldsymbol{x})'\boldsymbol{A}(\boldsymbol{C}\boldsymbol{x}) = \boldsymbol{w}'\boldsymbol{A}\boldsymbol{w} > 0 \quad \forall \boldsymbol{x} \neq \boldsymbol{0}$$

Note that full column rank implies $\boldsymbol{0} = \boldsymbol{w} = \boldsymbol{C}\boldsymbol{x} \Rightarrow \boldsymbol{x} = \boldsymbol{0}$. Additionally, we need the fact that if $\boldsymbol{A}$ is positive definite then $\boldsymbol{A}^{-1}$ is positive definite as well. As before,

we write $\boldsymbol{A}^{-1}$ as $\boldsymbol{A}^{-1} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$ implying $\boldsymbol{\Gamma}'\boldsymbol{A}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}'(\boldsymbol{\Gamma}')^{-1}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma} = \boldsymbol{I}$. Now we consider $(\boldsymbol{I} - \boldsymbol{A})$ which is positive definite if and only if $(\boldsymbol{A}^{-1} - \boldsymbol{I})$ is positive definite. To see that we pre- and postmultiply $(\boldsymbol{I} - \boldsymbol{A})$ with $\boldsymbol{\Gamma}'$ and $\boldsymbol{\Gamma}$, respectively:

$$\boldsymbol{\Gamma}'(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{\Gamma} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma} - \boldsymbol{\Gamma}'\boldsymbol{A}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma} - \boldsymbol{I}$$

This matrix has the same eigenvalues as $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' - \boldsymbol{I} = \boldsymbol{A}^{-1} - \boldsymbol{I}$:

$$(\boldsymbol{A}^{-1} - \boldsymbol{I})\boldsymbol{x} = \lambda\boldsymbol{x} \quad |\text{premult.}\boldsymbol{\Gamma}' \qquad \Leftrightarrow \qquad (\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\boldsymbol{\Gamma}' - \boldsymbol{\Gamma}')\boldsymbol{x} = \lambda(\boldsymbol{\Gamma}'\boldsymbol{x})$$

$$\Leftrightarrow \quad (\boldsymbol{\Gamma}'\boldsymbol{\Gamma} - \boldsymbol{I})(\boldsymbol{\Gamma}'\boldsymbol{x}) = \lambda(\boldsymbol{\Gamma}'\boldsymbol{x})$$

Now we turn to $\boldsymbol{A}$ and $\boldsymbol{B}$ both being symmetric and positive definite. If $\boldsymbol{A} - \boldsymbol{B}$ is positive definite then

$$\boldsymbol{\Gamma}'(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{\Gamma} = \boldsymbol{\Gamma}'\boldsymbol{A}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}'\boldsymbol{B}\boldsymbol{\Gamma} = \boldsymbol{I} - \boldsymbol{\Gamma}'\boldsymbol{B}\boldsymbol{\Gamma}$$

is positive definite and also $\boldsymbol{\Gamma}^{-1}\boldsymbol{B}^{-1}(\boldsymbol{\Gamma}')^{-1} - \boldsymbol{I}$ which implies positiv definiteness of

$$\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^{-1}\boldsymbol{B}^{-1}(\boldsymbol{\Gamma}')^{-1} - \boldsymbol{I})\boldsymbol{\Gamma}' = \boldsymbol{B}^{-1} - \boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}$$

We now consider the difference between the covariance matrices of $\hat{\boldsymbol{\beta}}_{GLS}$ and $\hat{\boldsymbol{\beta}}_{W}$:

$$V(\hat{\boldsymbol{\beta}}_{W}) - V(\hat{\boldsymbol{\beta}}_{GLS}) = (\boldsymbol{W}'\boldsymbol{X})^{-1}\boldsymbol{W}'\boldsymbol{\Omega}\boldsymbol{W}(\boldsymbol{X}'\boldsymbol{W})^{-1} - (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}$$

for which we want to show positive semi-definiteness. Thus we consider

$$\begin{aligned}
&\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{\Omega}\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{X} = \boldsymbol{X}'(\boldsymbol{\Psi}\boldsymbol{\Psi}' - \boldsymbol{W}(\boldsymbol{W}'(\boldsymbol{\Psi}')^{-1}\boldsymbol{\Psi}^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}')\boldsymbol{X} \\
&= \boldsymbol{X}'\boldsymbol{\Psi}(\boldsymbol{I} - \boldsymbol{\Psi}^{-1}\boldsymbol{W}(\boldsymbol{W}'(\boldsymbol{\Psi}')^{-1}\boldsymbol{\Psi}^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}'(\boldsymbol{\Psi}')^{-1})\boldsymbol{\Psi}'\boldsymbol{X} \\
&= \boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{M}_{\boldsymbol{\Psi}^{-1}\boldsymbol{W}}\boldsymbol{\Psi}'\boldsymbol{X}
\end{aligned}$$

Since $\boldsymbol{M}$ is idempotent the positive semi-definiteness follows immediately showing the efficiency of $\hat{\boldsymbol{\beta}}_{GLS}$.

6.2 Feasible GLS

The consistency and normality of $\hat{\boldsymbol{\beta}}_{GLS}$ follows immediately from the interpretation as an instrumental variables estimator (MM or GMM).

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}) \stackrel{a}{=} (\text{plim } \frac{1}{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}(\text{plim } n^{-1/2}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{u})$$

If we assume that $\boldsymbol{\Omega}$ depends on an $l-$dimensional vector $\boldsymbol{\gamma}$ such that

$$E(u_t^2) = \exp(\boldsymbol{Z}_t\boldsymbol{\gamma})$$

we obtain $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ which might be estimated by inserting an estimate of $\boldsymbol{\gamma}$: $\hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}})$.

To obtain a consistent estimate of $\boldsymbol{\gamma}$ we first need a consistent estimate of u_t which is obviously possible if we initially use OLS estimate $\hat{\boldsymbol{\beta}}_{OLS}$ and calculate the OLS residuals and then estimate

$$\log(u_t^2) = \boldsymbol{Z}_t\boldsymbol{\gamma} + \nu_t$$

A typical element of $\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}})$ then is

$$\hat{\omega}_t = (\exp(\boldsymbol{Z}_t\hat{\boldsymbol{\gamma}}))^{1/2}$$

Thus for the feasible GLS estimator $\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{y}$ we see that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{FGLS} - \boldsymbol{\beta}) \stackrel{a}{=} (\text{plim } \frac{1}{n}\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X})^{-1}(\text{plim } n^{-1/2}\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{u})$$

If we have

$$\text{plim } \frac{1}{n}\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X} = \text{plim } \frac{1}{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}$$

and

$$\text{plim } n^{-1/2}\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{u} = \text{plim } n^{-1/2}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{u}$$

then $\hat{\boldsymbol{\beta}}_{FGLS}$ and $\hat{\boldsymbol{\beta}}_{GLS}$ share the same asymptotic properties.

The consistency of $\hat{\boldsymbol{\beta}}_{FGLS}$ and $\hat{\boldsymbol{\beta}}_{GLS}$ hinges on consistent residual estimates which in turn require consistent OLS estimates. If $\boldsymbol{\Omega}$ is not diagonal the consistency of $\hat{\boldsymbol{\beta}}_{GLS}$ is lost if any element of $\boldsymbol{X}_t$ is a lagged dependent variable. The lagged dependent variables are still predetermined to current u_t but due to serial correlation of the errors they are not uncorrelated to u_t.

6.3 Heteroscedasticity

As seen above, if we assume for the main diagonal of $\boldsymbol{\Omega}$

$$\omega_t^2 = \exp(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma})$$

or more general

$$\omega_t^2 = h(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma})$$

and off-main diagonal elements equal to zero we obtain as a typical element of $\boldsymbol{\Omega}^{-1}$

$$\omega_t^{-2} = \frac{1}{h(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma})}$$

and therefore a typical element of $\boldsymbol{\Psi}$ as

$$\omega_t^{-1} = \left(\frac{1}{h(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma})}\right)^{1/2}$$

The transformed model then looks like

$$\omega_t^{-1} y_t = \omega_t^{-1}\boldsymbol{X}_t\boldsymbol{\beta} + \omega_t^{-1}u_t$$

This is the reason why it is sometimes called *weighted least squares* (WLS). Note that if $\boldsymbol{X}_t$ contained a constant, after transformation it becomes to $1/\omega_t$.

6.3.1 Testing for Heteroscedasticity

Assuming $E(u_t^2|\boldsymbol{Z}_t) = h(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma})$ we can write this as

$$u_t^2 = h(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma}) + \text{residual}$$

This is a non-linear regression for which we get

$$\begin{aligned} h'(\hat{\delta} + \boldsymbol{Z}_t\hat{\boldsymbol{\gamma}})\left(u_t^2 - h(\hat{\delta} + \boldsymbol{Z}_t\hat{\boldsymbol{\gamma}})\right) &= 0 \\ h'(\hat{\delta} + \boldsymbol{Z}_t\hat{\boldsymbol{\gamma}})\boldsymbol{Z}_t'\left(u_t^2 - h(\hat{\delta} + \boldsymbol{Z}_t\hat{\boldsymbol{\gamma}})\right) &= \mathbf{0} \end{aligned}$$

Thus an artificial regression of $u_t^2 - h(\hat{\delta} + \boldsymbol{Z}_t\hat{\boldsymbol{\gamma}})$ on $h'(\hat{\delta} + \boldsymbol{Z}_t\hat{\boldsymbol{\gamma}})$ and $h'(\hat{\delta} + \boldsymbol{Z}_t\hat{\boldsymbol{\gamma}})\boldsymbol{Z}_t$ should result in coefficients equal to zero:

$$u_t^2 - h(\hat{\delta} + \boldsymbol{Z}_t\hat{\boldsymbol{\gamma}}) = h'(\hat{\delta} + \boldsymbol{Z}_t\hat{\boldsymbol{\gamma}})b_\delta + h'(\hat{\delta} + \boldsymbol{Z}_t\hat{\boldsymbol{\gamma}})\boldsymbol{Z}_t\boldsymbol{b}_{\boldsymbol{\gamma}} + \text{residual}$$

Under the null hypothesis of no heteroscedasticity, i.e. $H_0: \boldsymbol{\gamma} = \mathbf{0}$, this regression reduces to

$$u_t^2 - h(\delta) = h'(\delta)b_\delta + h'(\delta)\boldsymbol{Z}_t\boldsymbol{b}_{\boldsymbol{\gamma}} + \text{residual}$$

The constant factor $h'(\delta)$ changes the coefficients, they do not change the explanatory power (also SSR) of the model. Thus a test for heteroscedasticity then simply is running the regression

$$\hat{u}_t^2 = b_\delta + \boldsymbol{Z}_t \boldsymbol{b}_{\boldsymbol{\gamma}} + \text{residual}$$

and performing an F-test for $\boldsymbol{b}_{\boldsymbol{\gamma}} = \boldsymbol{0}$. Note that it does not depend on the form of $h(\cdot)$.

6.3.2 Heteroscedasticity-Consistent Covariance Matrix (HCCM)

Using $\hat{\boldsymbol{\beta}}_{OLS}$ we saw above that for $\boldsymbol{X}$ considered as fix and given $\boldsymbol{\Omega}$

$$V(\hat{\boldsymbol{\beta}}_{OLS}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Looking at $\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})$ we find that the asymptotic covariance matrix to be

$$\left(\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}\right) \left(\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{10}$$

In one of the probably most famous papers in econometrics White (1980) showed that the middle term can be consistently estimated by

$$\frac{1}{n}\sum_{t=1}^{n} \hat{u}_t^2 \boldsymbol{X}_t'\boldsymbol{X}_t$$

Thus an estimate for the asymptotic HCCM in (10) is

$$\widehat{V(\hat{\boldsymbol{\beta}}_{OLS})} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{t=1}^{n} \hat{u}_t^2 \boldsymbol{X}_t'\boldsymbol{X}_t\right) (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

6.4 Autoregressive Processes

In regression models for time series data it is an often encountered problem that due to various reasons error terms might be correlated.

6.4.1 AR(1) Process

The simplest autoregressive structure is the first order process written as

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim IID(0, \sigma^2) \qquad |\rho| < 1 \tag{11}$$

The requirement $|\rho| < 1$ is called **stationarity condition**. Weak stationarity of process u_t requires

- $E(u_t) = \mu$
- $\mathrm{Var}(u_t) = \sigma_u^2$
- $\mathrm{Cov}(u_t, u_{t-s}) = \gamma_s$

All these moments, of course they should exist, are independent of t. The covariance only depends on the number of lags considered but is also independent of t.

Substituting recursively $u_{t-1} = \rho u_{t-2} + \varepsilon_{t-1}$ in (14) again and again yields

$$u_t = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3} + \rho^4\varepsilon_{t-4} + \dots$$

Since the ε_i are all independent it is easy to compute the variance of error term u_t:

$$\sigma_u^2 \equiv \mathrm{Var}(u_t) = \sum_{i=0}^{\infty} \left(\rho^i\right)^2 \sigma_\varepsilon^2 = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} \left(\rho^2\right)^i = \frac{\sigma_\varepsilon^2}{1-\rho^2}$$

The requirement for the geometric series to converges is $|\rho^2| < 1$ which implies $|\rho| < 1$. The covariance of u_t and u_{t-1} is

$$\mathrm{Cov}(u_t, u_{t-1}) = E(u_t \cdot u_{t-1}) = E((\rho u_{t-1} + \varepsilon_t) \cdot u_{t-1}) = E(\rho u_{t-1}^2) = \rho\sigma_u^2$$

The covariance of u_t and u_{t-1} is

$$\mathrm{Cov}(u_t, u_{t-2}) = E(u_t \cdot u_{t-2}) = E((\rho^2 u_{t-2} + \varepsilon_t + \rho\varepsilon_{t-1}) \cdot u_{t-2}) = E(\rho^2 u_{t-2}^2) = \rho^2\sigma_u^2$$

Or more general: $\gamma_s = \mathrm{Cov}(u_t, t_{t-s}) = \rho^s\sigma_u^2$

As a result, the covariance matrix $\boldsymbol{\Omega}$ of $n-$vector $\boldsymbol{u}$ is

$$\boldsymbol{\Omega}(\rho) = \frac{\sigma_\varepsilon^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix} \tag{12}$$

6.4.2 Higher-Order Autoregressive Processes

We consider an AR(p) process

$$u_t = \alpha_1 u_{t-1} + \alpha_2 u_{t-2} + \ldots + \alpha_p u_{t-p} + \varepsilon_t \qquad \varepsilon_t \sim IID(0, \sigma_\varepsilon^2)$$

For these higher order processes it is useful to introduce the lag-operator L which has the properties $\mathrm{L}u_t = u_{t-1}$ and $\mathrm{LL}u_t = \mathrm{L}^2 u_t = u_{t-2}$. The AR(p) process can now be written as

$$(1 - \alpha_1\mathrm{L} - \alpha_2\mathrm{L}^2 - \ldots - \alpha_p\mathrm{L}^p)u_t = \varepsilon_t$$

Note, that the AR(1) process with $(1-\rho\mathrm{L})u_t = \varepsilon_t$ can be **inverted** to

$$u_t = (1-\rho\mathrm{L})^{-1}\varepsilon_t$$

Thus the term $(1-\rho\mathrm{L})^{-1}$ represents a geometric series of $(\rho\mathrm{L})$.

The notation using the lag-operator is helpful since it closely relates to the **characteristic equation**

$$1 - \alpha_1 z - \alpha_2 z^2 - \ldots - \alpha_p z^p = 0$$

If the roots of this characteristic equation, some of them might be complex, all lie outside the unit circle then the AR(p) process is stationary. A complex number $z = a + bi$ lies outside the unit circle if $a^2 + b^2 > 1$.

6.4.3 Testing for Serial Correlation

The most common test statistic to test for serial correlation is the Durbin-Watson DW-statistic:

$$d = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2} = \frac{\sum_{t=2}^{n}\hat{u}_t^2}{\sum_{t=1}^{n}\hat{u}_t^2} - 2 \cdot \underbrace{\frac{\sum \hat{u}_t\hat{u}_{t-1}}{\sum \hat{u}_t^2}}_{\hat{\rho}} + \frac{\sum_{t=2}^{n}\hat{u}_{t-1}^2}{\sum_{t=1}^{n}\hat{u}_t^2}$$

If we assume $\hat{u}_0 = 0$ the middle term is the sample correlation coefficient $\hat{\rho}$. Thus for $n \to \infty$ the first and the last term of the right hand side tend to 1 so that d is asymptotically equal to $2 - 2\hat{\rho}$.

As we have seen in section 2 the residual vector $\hat{\boldsymbol{u}}$ of a correctly specified model is equal to $\boldsymbol{M_X u}$. It implies that in finite samples even if the components in $\boldsymbol{u}$ are serially uncorrelated the residuals $\hat{\boldsymbol{u}}$ display a certain amount of serial correlation depending on matrix $\boldsymbol{X}$.

A critical value d for a given significance level thus depends on $\boldsymbol{X}$. But it has been shown that there are lower d_l and upper d_u bounds for it only depending on n and k. In practice, it is still quite common to reject $H_0 : \rho = 0$ if $d < d_l$, not reject if $d > d_u$, and be indecisive if $d_l < d < d_u$ although everyone could perform a Monte Carlo based test to overcome this problem.

6.4.4 Estimation

If we consider a model with AR(1) error term

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t \qquad u_t = \rho u_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim IID(0, \sigma^2) \qquad |\rho| < 1$$

we can replace $u_{t-1} = y_{t-1} - \boldsymbol{X}_{t-1}\boldsymbol{\beta}$ and obtain

$$y_t = \rho y_{t-1} + \boldsymbol{X}_t\boldsymbol{\beta} - \rho\boldsymbol{X}_{t-1}\boldsymbol{\beta} + \varepsilon_t$$

If we ignore the first observation this equation can be estimated using non-linear least squares.

An alternative approach is the feasible GLS estimator. We just have to find a decomposition of covariance matrix (12) fulfilling $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}'$ in a way that $\boldsymbol{\Psi}'\boldsymbol{u} = \boldsymbol{\varepsilon}$. For observations $t = 2, \ldots, n$ we know that

$$\varepsilon_t = u_t - \rho u_{t-1}$$

Thus the t^{th} row of $\boldsymbol{\Psi}'$ must have a 1 in the t^{th} position, $-\rho$ in the $(t-1)^{th}$ position, and 0 elsewhere. For the first row of $\boldsymbol{\Psi}'\boldsymbol{u}$ we use the unconditional variance of u_1 which from above is $\mathrm{Var}(u_1) = \sigma_\varepsilon^2/(1-\rho^2)$. Therefore, if we define

$$\varepsilon_1 = (1-\rho^2)^{1/2} u_1$$

it is obvious that ε_1 is independent of all ε_2, ε_3 etc. and additionally we obtain $\mathrm{Var}(\varepsilon_1) = (1-\rho^2)/(1-\rho^2) = 1$. As a result, we get

$$\boldsymbol{\Psi}' = \begin{pmatrix} (1-\rho^2)^{1/2} & 0 & \cdots & & & 0 \\ -\rho & 1 & 0 & \cdots & & 0 \\ 0 & -\rho & 1 & 0 & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & 0 & -\rho & 1 & 0 \\ 0 & \cdots & & 0 & -\rho & 1 \end{pmatrix}$$

The next step is calculating the transformed observations $\boldsymbol{\Psi}'\boldsymbol{y}$ and $\boldsymbol{\Psi}'\boldsymbol{X}$ and running an OLS regression for these transformed variables.

Whenever ρ is known this GLS estimator can be applied immediately. If it is unknown it can be estimated using $\hat{\boldsymbol{u}}$ from a preceding OLS estimation.

The FGLS residual $(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLS})$ can be used to re-estimate ρ and reiterate the GLS step. By succeedingly repeating these steps we come quite close to the non-linear least squares estimates.

A third alternative estimator is MLE if one is willing to assume $\varepsilon_t \sim IIN(0, \sigma_\varepsilon^2)$.

6.5 Panel Data

Panel data have two dimensions, a cross-section dimension and a time dimension, and can be seen as repeated cross-sections. Our model the becomes to

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + u_{it} \qquad i = 1, \ldots, m \quad t = 1, \ldots, T$$

where $\boldsymbol{X}_{it}$ is a $1 \times k$ vector of explanatory variables. There are now $n = mT$ observations.

- If $E(u_{it}|\boldsymbol{X}_{it}) = 0$ OLS yields unbiased but probably inefficient estimates if $\text{Cov}(u_{it}, u_{jt}) \neq 0$ or $\text{Cov}(u_{it}, u_{i,t+s}) \neq 0$.
- If $\boldsymbol{X}_{it}$ contains lagged dependent variables, OLS estimates are inconsistent if u_{it} are serially correlated.

Error Components

It is quite common to assume

$$u_{it} = e_t + \nu_i + \varepsilon_{it}$$

with e_t being independent across t, ν_i being independent across i, and ε_{it} being independent across i and t.

We will simplify by assuming e_t=0.

Fixed-Effects

If we assume ν_i to be non-stochastic or fixed but unknown to the researches they can be treated as parameters to be estimated. In such a case we can write the model as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

The covariance matrix for error $\boldsymbol{\varepsilon}$ then is $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma_\varepsilon^2 \boldsymbol{I}_n$. Vector $\boldsymbol{y}$ has y_{it} as typical element and $\boldsymbol{D}$ is an $n \times m$ matrix of dummy variables. A 1 in column j indicates

that this specific observation in that row belongs to cross-sectional unit j. There are T ones in each column of $\boldsymbol{D}$. The $m-$dimensional vector $\boldsymbol{\eta}$ contains the fixed effects ν_i.

The **fixed-effects** estimator of $\boldsymbol{\beta}$, sometimes called least squares dummy variables (LSDV) estimator, then is

$$\hat{\boldsymbol{\beta}}_{FE} = (\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{y}$$

$\boldsymbol{D}'\boldsymbol{y}$ yields an $m-$dimensional vector with typical element $\sum_t y_{it}$ and therefore

$$(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{y} = \begin{pmatrix} T^{-1}\sum_t y_{1t} \\ \vdots \\ T^{-1}\sum_t y_{mt} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_m \end{pmatrix}$$

Thus $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{y}$ has typical element $y_{it} - \bar{y}_i$. The $\hat{\boldsymbol{\beta}}_{FE}$ estimator makes only use of a variable's variation around the mean for each of the m groups and, therefore, it is called **within-group estimator**.

The estimator is unbiased since $\boldsymbol{X}$ and $\boldsymbol{D}$ are exogenous and also efficient. However, if some explanatory variables for an individual i do not vary over time then columns of $\boldsymbol{X}$ and $\boldsymbol{D}$ are collinear.

Random-Effects

If we treat ν_i to be stochastic but independent of ε_{it} we need to restrict that ν_i should be independent of $\boldsymbol{X}$. In that case, OLS estimates $\boldsymbol{\beta}$ are unbiased but inefficient. This is due to the fact that

$$\begin{aligned} \text{Var} &= \sigma_\nu^2 + \sigma_\varepsilon^2 \\ \text{Cov}(u_{it}, u_{is}) &= \sigma_\nu^2 \\ \text{Cov}(u_{it}, u_{js}) &= 0 \qquad \forall i \neq j \end{aligned}$$

Ordering the $n = mT$ observations by cross-sectional units, we can write the covariance Matrix $E(\boldsymbol{u}\boldsymbol{u}') = \boldsymbol{\Omega}$ as a block-diagonal matrix

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{pmatrix}$$

with $\boldsymbol{\Sigma} = \sigma_\varepsilon^2 \boldsymbol{I}_T + \sigma_\nu^2 \boldsymbol{\iota}\boldsymbol{\iota}'$.

GLS yields an efficient estimator if $\boldsymbol{\Sigma}$ is known. Problems with the consistency of FGLS arise since $n \to \infty$ can be achieved by fixed m and increasing T or the reverse.

The counterpart to the within-group estimator is the **between-group estimator** stemming from the fact that $\boldsymbol{y} = (\boldsymbol{M_D} + \boldsymbol{P_D})\boldsymbol{y} = \boldsymbol{M_D}\boldsymbol{y} + \boldsymbol{P_D}\boldsymbol{y}$:

$$\hat{\boldsymbol{\beta}}_{BG} = (\boldsymbol{X}'\boldsymbol{P_D}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P_D}\boldsymbol{y}$$

As we have seen before, vector $\boldsymbol{P_D}\boldsymbol{y}$ contains the m group sample means, however, replicated T times. The regression

$$\boldsymbol{P_D}\boldsymbol{y} = \boldsymbol{P_D}\boldsymbol{X}\boldsymbol{\beta} + \text{residual}$$

is a basis for estimating σ_ν^2 whereas the fixed effects model is appropriate to estimate σ_ε^2.

It is insightful to note that the OLS estimator mentioned at the beginning of this section is a weighted average of $\hat{\boldsymbol{\beta}}_{BG}$ and $\hat{\boldsymbol{\beta}}_{FE}$:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{M_D}\boldsymbol{y} + \boldsymbol{P_D}\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M_D}\boldsymbol{X}\hat{\boldsymbol{\beta}}_{FE} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P_D}\boldsymbol{X}\hat{\boldsymbol{\beta}}_{BG} \end{aligned}$$

6.5.1 Instrumental Variables Estimator

Writing the fixed-effects estimator alternatively as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{FE} &= \left(\sum_{i=1}^{m}\sum_{t=1}^{T}(\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_i)'(\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_i)\right)^{-1}\sum_{i=1}^{m}\sum_{t=1}^{T}(\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_i)'(y_{it}-\bar{y}_i)\\ &= \left(\sum_{i=1}^{m}\sum_{t=1}^{T}(\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_i)'\boldsymbol{X}_{it}\right)^{-1}\sum_{i=1}^{m}\sum_{t=1}^{T}(\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_i)'y_{it}\end{aligned}$$

the estimator can be interpreted as an instrumental variables estimator for the model

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + (\nu_i + \varepsilon_{it})$$

where $\boldsymbol{X}_{it}$ is instrumented by $\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_i$. Since $E((\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_i)'\nu_i) = \mathbf{0}$ by construction, consistency of this estimator is given provided $E((\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_i)'\varepsilon_{it}) = \mathbf{0}$. If we have components in $\boldsymbol{X}_{it}$ that are obviously uncorrelated with ν_i there is no need for instrumenting those. This allows incorporation of certain time-invariant variables.

An extension to that is a generalized version by noting that not only $\boldsymbol{X}_{it}-\bar{\boldsymbol{X}}_i$ is a valid instrument for $\boldsymbol{X}_{it}$ but also $\boldsymbol{X}_{i1}-\bar{\boldsymbol{X}}_i$ up to $\boldsymbol{X}_{i,t-1}-\bar{\boldsymbol{X}}_i$ resulting in more instruments than unknown parameters.

6.5.2 Dynamic Linear Models

Extending our linear panel model by a lagged dependent variable leads to dynamic model

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + \gamma y_{i,t-1} + (\nu_i + \varepsilon_{it})$$

First we study a simpler version of that autoregressive panel model that has no exogenous variables

$$y_{it} = \gamma y_{i,t-1} + (\nu_i + \varepsilon_{it})$$

with $|\gamma| < 1$. The fixed effects estimator for γ is

$$\hat{\gamma}_{FE} = \frac{\sum_{i=1}^{m}\sum_{t=2}^{T}(y_{it}-\bar{y}_i)(y_{i,t-1}-\bar{y}_{i,-1})}{\sum_{i=1}^{m}\sum_{t=2}^{T}(y_{i,t-1}-\bar{y}_{i,-1})^2}$$

where $\bar{y}_i = (T-1)^{-1}\sum_{t=2}^{T} y_{it}$ and $\bar{y}_{i,-1} = (T-1)^{-1}\sum_{t=2}^{T} y_{i,t-1}$. For **fixed** T and increasing m it is an inconsistent estimate since

$$\varepsilon_{it} = y_{it} - \gamma y_{i,t-1} - \nu_i \quad \Rightarrow \quad \sum_{t=2}^{T}\varepsilon_{it} = \sum_{t=2}^{T} y_{it} - \gamma\sum_{t=2}^{T} y_{i,t-1} - \sum_{t=2}^{T}\nu_i$$

$$\Rightarrow \quad \bar{\varepsilon}_{it} = \bar{y}_{it} - \gamma\bar{y}_{i,-1} - \nu_i \quad \Rightarrow \quad \varepsilon_{it} - \bar{\varepsilon}_i = (y_{it} - \bar{y}_i) - \gamma(y_{i,t-1} - \bar{y}_{i,-1})$$

$$\Rightarrow \quad \operatorname*{plim}_{m\to\infty}(\varepsilon_{it} - \bar{\varepsilon}_i)(y_{i,t-1} - \bar{y}_{i,-1}) \neq 0$$

One possible solution is a different data transformation to cancel out ν_i:

$$y_{it} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + (\varepsilon_{it} - \varepsilon_{i,t-1})$$

OLS estimation of this transformation is obviously inconsistent since $y_{i,t-1}$ is correlated with $\varepsilon_{i,t-1}$. However, e.g. $y_{i,t-2}$ is uncorrelated to both error terms and is therefore a valid instrument if ε_{it} is serially uncorrelated. This suggests an IV estimator

$$\gamma_{IV} = \frac{\sum_{i=1}^{m}\sum_{t=3}^{T} y_{i,t-2}(y_{it} - y_{i,t-1})}{\sum_{i=1}^{m}\sum_{t=3}^{T} y_{i,t-2}(y_{i,t-1} - y_{i,t-2})}$$

Of course, $(y_{i,t-2} - y_{i,t-3})$ could be used as an instrument as well but suffers from losing one sample period. Again, if ε_{it} is serially uncorrelated this estimator is also consistent since

$$E((y_{i,t-2} - y_{i,t-3})(\varepsilon_{it} - \varepsilon_{i,t-1})) = 0$$

This is a moment condition just like $E((y_{i,t-2})(\varepsilon_{it} - \varepsilon_{i,t-1})) = 0$ so that a GMM approach combining both should be more efficient. Considering a specific t we have $t-2$ moment conditions which can be exploited.

6.6 Systems of Regression Equations

In this section we focus on **multivariate models** in which more than one regression equation are considered. There are multiple endogenous variables to be explained simultaneously.

6.6.1 Seemingly Unrelated Regressions

We consider g endogenous variables, indexed by i, each of them having a linear regression function:

$$\underset{(n\times k_i)}{\boldsymbol{y}_i} = \boldsymbol{X}_i\,\boldsymbol{\beta}_i + \boldsymbol{u}_i \qquad E(\boldsymbol{u}_i\boldsymbol{u}_i') = \sigma_{ii}\boldsymbol{I}$$

In this respect the g regressions are unrelated leading to the common name *seemingly unrelated regressions* (SUR). However, contemporaneous covariances across these equations are allowed:

$$E(u_{ti}u_{tj}) = \sigma_{ij} \qquad \text{for all } t\,, \qquad E(u_{ti}u_{sj}) = 0 \qquad \text{for all } t \neq s$$

Defining an $1 \times g$ vector $\boldsymbol{U}_t$ with typical element u_{ti} and with this an $n \times g$ matrix of error terms $\boldsymbol{U}$ (note that $\boldsymbol{U} = [\boldsymbol{u}_1 \vdots \boldsymbol{u}_2 \vdots \ \ldots\ \vdots \boldsymbol{u}_g]$) we can denote

$$E(\boldsymbol{U}_t'\boldsymbol{U}_t) = \frac{1}{n}E(\boldsymbol{U}'\boldsymbol{U}) = \boldsymbol{\Sigma}$$

In this system we possibly have $\sum_{i=1}^{g} k_i = k$ regressors which are not all distinct (e.g. ι is most likely included in all $\boldsymbol{X}_i$). We combine all non-redundant regressors into (full column rank) matrix $\boldsymbol{X}$ having $l < k$ columns. We then can either assume strict exogeneity $E(\boldsymbol{U} \mid \boldsymbol{X}) = \boldsymbol{O}$ or the weaker condition $E(\boldsymbol{U}_t \mid \boldsymbol{X}_t) = \boldsymbol{0}$. Note that $\boldsymbol{X}_t$ now denotes the t^{th} row of $\boldsymbol{X}$ whereas before $\boldsymbol{X}_i$ denotes an $n \times k_i$ matrix.

The entire SUR system can be written compactly as

$$\boldsymbol{y}_\bullet = \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet + \boldsymbol{u}_\bullet$$

where $\boldsymbol{y}_\bullet$ is a gn vector resulting from stacking $\boldsymbol{y}_1$ through $\boldsymbol{y}_g$. Matrix $\boldsymbol{X}_\bullet$ is block-diagonal of dimension $gn \times k$:

$$\boldsymbol{X}_\bullet = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{X}_2 & \cdots & \boldsymbol{O} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{X}_g \end{pmatrix}$$

The covariance matrix of $\boldsymbol{u}_\bullet$, in which the columns of $\boldsymbol{U}$ are vertically stacked, is

$$E(\boldsymbol{u}_\bullet \boldsymbol{u}_\bullet') = \begin{pmatrix} E(\boldsymbol{u}_1\boldsymbol{u}_1') & \cdots & E(\boldsymbol{u}_1\boldsymbol{u}_g') \\ \vdots & \vdots & \vdots \\ E(\boldsymbol{u}_g\boldsymbol{u}_1') & \cdots & E(\boldsymbol{u}_g\boldsymbol{u}_g') \end{pmatrix} = \begin{pmatrix} \sigma_{11}\boldsymbol{I} & \cdots & \sigma_{1g}\boldsymbol{I} \\ \vdots & \vdots & \vdots \\ \sigma_{g1}\boldsymbol{I} & \vdots & \sigma_{gg}\boldsymbol{I} \end{pmatrix} = \boldsymbol{\Sigma}_\bullet$$

This matrix can be written as a *Kronecker product* $\boldsymbol{\Sigma}_\bullet = \boldsymbol{\Sigma} \otimes \boldsymbol{I}$. For $p \times q$ matrix $\boldsymbol{A}$ and an $r \times s$ matrix $\boldsymbol{B}$ the resulting Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$ is of dimension $pr \times qs$ and has as typical elements $r \times s$ blocks $a_{ij}\boldsymbol{B}$. The general properties are (conformable matrices required):

$$\begin{aligned} (\boldsymbol{A} \otimes \boldsymbol{B})' &= \boldsymbol{A}' \otimes \boldsymbol{B}' \\ (\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) &= (\boldsymbol{AC}) \otimes (\boldsymbol{BD}) \\ (\boldsymbol{A} \otimes \boldsymbol{B})^{-1} &= \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1} \end{aligned}$$

GLS Estimation

Using this notation the GLS estimator for a given matrix $\boldsymbol{\Sigma}$ then is:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\bullet^{GLS} &= (\boldsymbol{X}_\bullet'\boldsymbol{\Sigma}_\bullet^{-1}\boldsymbol{X}_\bullet)^{-1}\boldsymbol{X}_\bullet'\boldsymbol{\Sigma}_\bullet^{-1}\boldsymbol{y}_\bullet \\ &= (\boldsymbol{X}_\bullet'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I})\boldsymbol{X}_\bullet)^{-1}\boldsymbol{X}_\bullet'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I})\boldsymbol{y}_\bullet \end{aligned}$$

The covariance matrix of that estimator thus is

$$V(\hat{\boldsymbol{\beta}}_\bullet^{GLS}) = (\boldsymbol{X}_\bullet'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I})\boldsymbol{X}_\bullet)^{-1}$$

Feasible GLS first uses the inefficient but at least consistent OLS estimator $\hat{\boldsymbol{\beta}}_\bullet^{OLS} = (\boldsymbol{X}_\bullet'\boldsymbol{X}_\bullet)^{-1}\boldsymbol{X}_\bullet'\boldsymbol{y}_\bullet$ to compute the residuals $\hat{\boldsymbol{U}}$ and in turn estimating $\boldsymbol{\Sigma}$ by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\hat{\boldsymbol{U}}'\hat{\boldsymbol{U}}$$

This estimator replaces $\boldsymbol{\Sigma}$ in the above formulas resulting in $\hat{\boldsymbol{\beta}}_{\bullet}^{FGLS}$

GMM Estimation

Considering a single equation isolated we can interpret $E(\boldsymbol{X}_i\boldsymbol{u}_i) = \mathbf{0}$ as exactly identifying moment conditions resulting in the OLS estimator for that isolated equation. For all the other regressors in the other equations not included in $\boldsymbol{X}_i$ the moment condition with $\boldsymbol{u}_i$ should hold as well. Therefore, we can state gl theoretical moment conditions

$$E(\boldsymbol{X}'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i)) = \mathbf{0} \qquad\qquad i = 1, \ldots, g$$

What we can learn from this is that if all the g equations use exactly the same regressors $\boldsymbol{X} = \boldsymbol{X}_i$ the empirical counterpart of these moment conditions is

$$\boldsymbol{X}'(\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}_i) = \mathbf{0} \qquad\qquad i = 1, \ldots, g$$

and thus OLS estimation for each equation separately is efficient.

For the general case we can write $\boldsymbol{X}'(\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}_i) = \mathbf{0}$ for $i = 1, \ldots, g$ more compactly as

$$(\boldsymbol{I}_g \otimes \boldsymbol{X})'(\boldsymbol{y}_{\bullet} - \boldsymbol{X}_{\bullet}\boldsymbol{\beta}_{\bullet}) = \boldsymbol{O}$$

By the same line of argument as in section 6.1 the over-identifying moments in $\boldsymbol{X}$ need to be weighted to obtain an efficient GMM estimator. The optimally weighted instruments are thus again $(\boldsymbol{I}_g \otimes \boldsymbol{X})'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_n) = (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}')$ leading to optimally weighted moments

$$(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}')\boldsymbol{u}_{\bullet} = (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}')(\boldsymbol{y}_{\bullet} - \boldsymbol{X}_{\bullet}\boldsymbol{\beta}_{\bullet})$$

having covariance matrix

$$(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}')(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}) = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}'\boldsymbol{X}$$

implying the criterion function of the fully efficient GMM:

$$(\boldsymbol{y}_{\bullet} - \boldsymbol{X}_{\bullet}\boldsymbol{\beta}_{\bullet})'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X})(\boldsymbol{\Sigma} \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1})(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}')(\boldsymbol{y}_{\bullet} - \boldsymbol{X}_{\bullet}\boldsymbol{\beta}_{\bullet})$$

This reduces to $(\boldsymbol{y}_{\bullet} - \boldsymbol{X}_{\bullet}\boldsymbol{\beta}_{\bullet})'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{X}})(\boldsymbol{y}_{\bullet} - \boldsymbol{X}_{\bullet}\boldsymbol{\beta}_{\bullet})$ leading to

$$\hat{\boldsymbol{\beta}}_{\bullet}^{GMM} = \left(\boldsymbol{X}_{\bullet}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{X}_{\bullet}\right)^{-1}\boldsymbol{X}_{\bullet}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}_{\bullet}$$

Now we can see that the resulting estimator is equivalent to $\hat{\boldsymbol{\beta}}_{\bullet}^{GLS}$.

6.6.2 Linear Simultaneous Equation Models

The g equations are now extended by allowing some of the regressor variables to be endogenous variables from other equations:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{u}_i = \boldsymbol{Z}_i\boldsymbol{\beta}_{1i} + \boldsymbol{Y}_i\boldsymbol{\beta}_{2i} + \boldsymbol{u}_i$$

The $\boldsymbol{Z}_i$ are the exogenous or predetermined variables. All the distinct columns of these matrices $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_g$ are again summarized in the $n \times l$ matrix $\boldsymbol{Z}$. It must be possible to solve for the endogenous variables as functions of the predetermined variables $\boldsymbol{Z}$ which must be linear functions

$$\boldsymbol{y}_i = \boldsymbol{Z}\boldsymbol{\pi}_i + \text{residuals}$$

Thus $\boldsymbol{Z}$ serve as instruments for $\boldsymbol{y}_i$ by using $\boldsymbol{P}_{\boldsymbol{Z}}$ and of course for $\boldsymbol{Z}_i$. We can use

$$\hat{\boldsymbol{X}}_i = [\boldsymbol{Z}_i \vdots \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Y}_i] = \boldsymbol{P}_{\boldsymbol{Z}}[\boldsymbol{Z}_i \vdots \boldsymbol{Y}_i] = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_i$$

as conditional expectations of the regressors. By analogy to the previous subsection, the efficient GMM estimator is

$$\hat{\boldsymbol{\beta}}_{\bullet}^{GMM} = \left(\boldsymbol{X}_{\bullet}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{X}_{\bullet}\right)^{-1} \boldsymbol{X}_{\bullet}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{y}_{\bullet}$$

Feasible GMM or 3SLS

If the matrix $\boldsymbol{\Sigma}$ is unknown it can be replaced by a consistent estimate. Thus we first estimate $\boldsymbol{\beta}_{\bullet}$ by an instrumental variable approach using the moment conditions

$$\boldsymbol{X}_{\bullet}'(\boldsymbol{I}_g \otimes \boldsymbol{P}_{\boldsymbol{Z}})'(\boldsymbol{y}_{\bullet} - \boldsymbol{X}_{\bullet}\boldsymbol{\beta}_{\bullet}) = \boldsymbol{O}$$

leading to a consistent but inefficient estimator which is usually termed as 2SLS

$$\hat{\boldsymbol{\beta}}_{\bullet}^{2SLS} = \left(\boldsymbol{X}_{\bullet}'(\boldsymbol{I}_g \otimes \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{X}_{\bullet}\right)^{-1} \boldsymbol{X}_{\bullet}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{y}_{\bullet}$$

Using this estimator we can compute residuals and as above estimate the contemporaneous covariance matrix which we denote as $\hat{\boldsymbol{\Sigma}}_{2SLS}$. In the next stage we compute the above efficient GMM estimator which is usually called 3SLS:

$$\hat{\boldsymbol{\beta}}_{\bullet}^{3SLS} = \left(\boldsymbol{X}_{\bullet}'(\hat{\boldsymbol{\Sigma}}_{2SLS}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{X}_{\bullet}\right)^{-1} \boldsymbol{X}_{\bullet}'(\hat{\boldsymbol{\Sigma}}_{2SLS}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{y}_{\bullet}$$

7 Discrete and Limited Dependent Variables

7.1 Binary Response Models

Binary response models are characterized by the fact that the dependent value y_t can only take on two values usually coded 0 and 1. As usual in econometrics the conditional mean of y_t given $\boldsymbol{X}_t$ is the main focus.

A binomially distributed random variable $Z \sim B(1, p)$ has the moments

$$E(Z) = p \qquad \text{Var}(Z) = p(1-p)$$

Thus we consider $p_t \equiv P(y_t = 1|\boldsymbol{X}_t) = E(y_t|\boldsymbol{X}_t)$ which only can vary between 0 and 1 requiring a mapping of $\boldsymbol{X}_t$ and $\boldsymbol{\beta}$ on to the unit interval. Therefore, the usual specification is

$$p_t \equiv P(y_t = 1|\boldsymbol{X}_t) = E(y_t|\boldsymbol{X}_t) = F(\boldsymbol{X}_t\boldsymbol{\beta})$$

CDF's fulfill the requirement for such a function $F(\cdot)$. Common choices are the cdf of the normal distribution $F(\boldsymbol{X}_t\boldsymbol{\beta}) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta})$ (probit model) or the logistic distribution (logit model). Note, that since cdf's are non-linear the impact of a change in one of the regressors is not constant anymore

$$\frac{\partial p_t}{\partial x_{tj}} = \frac{\partial E(y_t|\boldsymbol{X}_t)}{\partial x_{tj}} = \frac{\partial F(\boldsymbol{X}_t\boldsymbol{\beta})}{\partial x_{tj}} = f(\boldsymbol{X}_t\boldsymbol{\beta}) \cdot \beta_j$$

Maximum Likelihood Estimation

Using the indicator y_t we can write the probability function compactly as

$$P(y_t) = p_t^{y_t} \cdot (1 - p_t)^{1-y_t}$$

Using the expression for p_t the likelihood function then is

$$f(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) = \prod_{t=1}^{n} F(\boldsymbol{X}_t\boldsymbol{\beta})^{y_t} \cdot (1 - F(\boldsymbol{X}_t\boldsymbol{\beta}))^{1-y_t}$$

from which we obtain the loglikelihood function

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) = \sum_{t=1}^{n} y_t \log F(\boldsymbol{X}_t\boldsymbol{\beta}) + (1 - y_t) \log(1 - F(\boldsymbol{X}_t\boldsymbol{\beta}))$$

Taking first derivatives

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \sum_{t=1}^{n} \frac{y_t}{F(\boldsymbol{X}_t\boldsymbol{\beta})} f(\boldsymbol{X}_t\boldsymbol{\beta})\boldsymbol{X}_t' + \frac{1-y_t}{1-F(\boldsymbol{X}_t\boldsymbol{\beta})}(-f(\boldsymbol{X}_t\boldsymbol{\beta})\boldsymbol{X}_t') \\ &= \sum_{t=1}^{n} \frac{(y_t - F(\boldsymbol{X}_t\boldsymbol{\beta}))f(\boldsymbol{X}_t\boldsymbol{\beta})}{F(\boldsymbol{X}_t\boldsymbol{\beta})(1-F(\boldsymbol{X}_t\boldsymbol{\beta}))}\boldsymbol{X}_t' \qquad (13)\end{aligned}$$

and second derivatives (using short-hand $F_t = F(\boldsymbol{X}_t\boldsymbol{\beta})$ and $f_t = f(\boldsymbol{X}_t\boldsymbol{\beta})$)

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = \sum_{t=1}^{n} \frac{(-f_t^2 + (y_t - F_t)\cdot df_t/dz)F_t(1-F_t) - (y_t - F_t)f_t(f_t - 2F_tf_t)}{(F_t(1-F_t))^2}\boldsymbol{X}_t'\boldsymbol{X}_t$$

Taking expectations of that Hessian and multiply it with -1 yields the information matrix

$$\boldsymbol{I}(\boldsymbol{\beta}) = \sum_{t=1}^{n} \frac{f_t^2}{F_t(1-F_t)}\boldsymbol{X}_t'\boldsymbol{X}_t$$

Probit Model

The density function of a standard normal distribution is

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

It is easy to see that the first derivative is $d\phi(z)/dz = \phi(z)\cdot(-z)$.

The conditional probability or conditional mean in the probit model $P(y_t = 1|\boldsymbol{X}_t) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta})$ can be derived alternatively assuming a latent variable y_t^* and its generating process

$$y_t^* = \boldsymbol{X}_t\boldsymbol{\beta} + u_t \qquad u_t \sim IIN(0,1)$$

We only observe whether the latent variable is larger than a given threshold (0 in this specific case):

$$y_t = \begin{cases} 1 & \text{if } y_t^* > 0 \\ 0 & \text{if } y_t^* \leq 0 \end{cases}$$

From this we derive

$$P(y_t = 1) = P(y_t^* > 0) = P(\boldsymbol{X}_t\boldsymbol{\beta} + u_t > 0) = 1 - P(u_t < -\boldsymbol{X}_t\boldsymbol{\beta}) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta})$$

Replacing $F(\boldsymbol{X}_t\boldsymbol{\beta})$ by $\Phi(\boldsymbol{X}_t\boldsymbol{\beta})$ and $f(\boldsymbol{X}_t\boldsymbol{\beta})$ by $\phi(\boldsymbol{X}_t\boldsymbol{\beta})$ in (13) and setting these equations equal to $\mathbf{0}$ defines the MLE.

Logit Model

The cdf for the logistic distribution is given by

$$\Lambda(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$$

We see that this distribution is symmetric around 0 since $\Lambda(-z) = 1 - \Lambda(z)$. The pdf has the form

$$\lambda(z) = \frac{e^z}{(1+e^z)^2} = \Lambda(z)\Lambda(-z) = \Lambda(z)(1-\Lambda(z))$$

Using these expressions replacing F and f,respectively, in (13) we obtain

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{t=1}^{n} (y_t - \Lambda(\boldsymbol{X}_t\boldsymbol{\beta}))\boldsymbol{X}_t'$$

Thus the Hessian simplifies to

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = \sum_{t=1}^{n} -\lambda(\boldsymbol{X}_t\boldsymbol{\beta})\boldsymbol{X}_t'\boldsymbol{X}_t = -\sum_{t=1}^{n} \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})(1-\Lambda(\boldsymbol{X}_t\boldsymbol{\beta}))\boldsymbol{X}_t'\boldsymbol{X}_t$$

In this case the Hessian does not depend on y_t so that the negative of it is already the information matrix. If we define a diagonal matrix $\boldsymbol{\Psi}$ with typical diagonal element $p_t(1-p_t) = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})(1-\Lambda(\boldsymbol{X}_t\boldsymbol{\beta}))$ the Hessian can be written as

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}$$

from which we can deduce negative definiteness if $\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$ shows variation across individuals. The loglikelihood function is therefore globally concave with a unique maximum.

It is interesting to note that $\Lambda(z)/(1-\Lambda(z)) = e^z$ and thus

$$\log\left(\frac{p_t}{1-p_t}\right) = \boldsymbol{X}_t\boldsymbol{\beta}$$

7.2 Ordered Discrete Response Models

We extend the latent variable model

$$y_t^* = \boldsymbol{X}_t\boldsymbol{\beta} + u_t \qquad\qquad u_t \sim IIN(0,\sigma^2)$$

by allowing more than two discrete responses, say r categories, so that the measurement equation becomes to

$$y_t = j \quad \Leftrightarrow \quad \gamma_{t,j-1} < y_t^* \leq \gamma_{t,j} \qquad \text{with} \quad \gamma_0 = -\infty \text{ and } \gamma_r = \infty$$

The probability for y_t being observed in category j:

$$\begin{aligned} P(y_t = j|\boldsymbol{X}_t) &= \int_{\gamma_{j-1}}^{\gamma_j} f(y_t^*)dy_t^* = \int_{-\infty}^{\gamma_j} f(y_t^*)dy_t^* - \int_{-\infty}^{\gamma_{j-1}} f(y_t^*)dy_t^* \\ &= P(y_t^* < \gamma_j) - P(y_t^* < \gamma_{j-1}) \\ &= F\left(\frac{\gamma_j - \boldsymbol{X}_t\boldsymbol{\beta}}{\sigma}\right) - F\left(\frac{\gamma_{j-1} - \boldsymbol{X}_t\boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

As implicitly done in the last section, some identifying restrictions are required: $\sigma = 1$ and in case $\boldsymbol{X}_t$ contains a constant e.g. $\gamma_1 = 0$.

If we define a set of dummy variables y_{ij} with

$$y_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{else} \end{cases}$$

we can write probabilities compactly as

$$P(y_i = k) = \prod_{j=1}^{r} P(y_i = j)^{y_{ij}}$$

resulting in a loglikelihood function

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{t=1}^{n} \sum_{j=1}^{r} y_{tj} \log P(y_t = j|\boldsymbol{X}_t)$$

7.3 Unordered Discrete Response Models

The discrete response categories in this class of models do not show any natural ordering scheme. These models are also referred to as *discrete choice* models.

Multinomial Logit Model

A natural extension for the multinomial case is by generalizing the logit case:

$$P(y_t = l) = \frac{\exp(\boldsymbol{W}_{tl}\boldsymbol{\beta}^l)}{\sum_{j=1}^{r} \exp(\boldsymbol{W}_{tj}\boldsymbol{\beta}^j)} \qquad \text{for } l = 1, \ldots, r$$

The (row) vector of regressors $\boldsymbol{W}_{tl}$ now contains individual specific as well as category specific variables. If there are only individual specific regressors, $\boldsymbol{W}_{tl}$ is replaced by $\boldsymbol{X}_t$. This shows that there must be category specific parameters $\boldsymbol{\beta}^l$. It also shows that not all parameter vectors are identified:

$$\begin{aligned} P(y_t = l) &= \frac{\exp(\boldsymbol{X}_t\boldsymbol{\beta}^l)}{\sum_{j=1}^r \exp(\boldsymbol{X}_t\boldsymbol{\beta}^j)} \cdot \frac{\exp(-\boldsymbol{X}_t\boldsymbol{\beta}^1)}{\exp(-\boldsymbol{X}_t\boldsymbol{\beta}^1)} = \frac{\exp(\boldsymbol{X}_t(\boldsymbol{\beta}^l - \boldsymbol{\beta}^1))}{\sum_{j=1}^r \exp(\boldsymbol{X}_t(\boldsymbol{\beta}^j - \boldsymbol{\beta}^1))} \\ &= \frac{\exp(\boldsymbol{X}_t(\boldsymbol{\beta}^l - \boldsymbol{\beta}^1))}{1 + \sum_{j=2}^r \exp(\boldsymbol{X}_t(\boldsymbol{\beta}^j - \boldsymbol{\beta}^1))} \end{aligned}$$

Thus we can restrict e.g. $\boldsymbol{\beta}^1 = \mathbf{0}$ which must be taken into account interpreting estimated parameters.

Another special case is the *conditional logit model* which uses individual specific choice characteristics $\boldsymbol{W}_{tl}$ and fixed $\boldsymbol{\beta}$

$$P(y_t = l) = \frac{\exp(\boldsymbol{W}_{tl}\boldsymbol{\beta})}{\sum_{j=1}^r \exp(\boldsymbol{W}_{tj}\boldsymbol{\beta})} \qquad \text{for } l = 1, \ldots, r$$

One important property of the multinomial logit model is the so-called *independence of irrelevant alternatives* (IIA) which can be seen from

$$\frac{P(y_t = l)}{P(y_t = j)} = \frac{\exp(\boldsymbol{W}_{tl}\boldsymbol{\beta}^l)}{\exp(\boldsymbol{W}_{tl}\boldsymbol{\beta}^j)}$$

The ratio of these probabilities does not depend on the regressors or parameters of other response categories. To avoid that sometimes unrealistic property this model is extended to the *nested logit model* where the r choices are partitioned into m disjoint subsets A_i so that

$$P(y_t = l | y_t \in A_i) = \frac{\exp(\boldsymbol{W}_{tl}\boldsymbol{\beta}^l/\theta_i)}{\sum_{j \in A_i} \exp(\boldsymbol{W}_{tj}\boldsymbol{\beta}^j/\theta_i)}$$

where θ_i is a scale parameter equal for all $l \in A_i$. These parameters detemines the probability of choosing an element of A_i.

Multinomial Pobit Model

A somewhat different foundation is by assuming a latent model

$$y_{tj}^* = \boldsymbol{W}_{tj}\boldsymbol{\beta}^j + u_{tj} \qquad \boldsymbol{u}_t \sim IIN(\mathbf{0}, \boldsymbol{\Omega})$$

where category j is observed and thus $y_{tj} = 1$ if $y^*_{tj} - y^*_{ti} \geq 0$ for all $i = 1, \ldots, r$. The latent model is often interpreted as describing the utility level y^*_{tj}. The choice with the highest utility is observed. Again, not all the parameter vectors are identified leading to the often used restriction $\boldsymbol{\beta}^1 = \mathbf{0}$.

The advantage of the multinomial probit model is the flexible covariance structure $\boldsymbol{\Omega}$ avoiding the IIA property if the categories are not modelled to be uncorrelated. However, the joint probabilities of an r dimensional normal distribution (actually only $r-1$ dimensions are required) are computationally burdensome since it involves approximation of $r-1$ fold integrals.

7.4 Count Data Models

For count data the most common distribution used is the Poisson distribution having the probability function

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!} \qquad y = 0, 1, 2, \ldots,$$

It follows that

$$E(Y) = \lambda \qquad \mathrm{Var}(Y) = \lambda$$

Modelling the mean of Y as a function of regressors requires its non-negativity. A common choice is

$$\lambda_t = \exp(\boldsymbol{X}_t\boldsymbol{\beta})$$

implying

$$P(Y_t = y) = \frac{\exp(-\exp(\boldsymbol{X}_t\boldsymbol{\beta}))\exp(\boldsymbol{X}_t\boldsymbol{\beta})^y}{y!} \qquad y = 0, 1, 2, \ldots,$$

The loglikelihood function is then obtained as

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) \sum_{t=1}^{n} (-\exp(\boldsymbol{X}_t\boldsymbol{\beta}) + y_t\boldsymbol{X}_t\boldsymbol{\beta} - \log y_t!)$$

resulting in the FOC

$$\frac{\partial \ell(\boldsymbol{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{t=1}^{n} (y_t - \exp(\boldsymbol{X}_t\boldsymbol{\beta}))\,\boldsymbol{X}'_t \;=\; \mathbf{0}$$

and the Hessian

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\sum_{t=1}^{n} \exp(\boldsymbol{X}_t\boldsymbol{\beta})\boldsymbol{X}_t'\boldsymbol{X}_t = -\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}$$

with diagonal matrix $\boldsymbol{\Psi}$ having typical diagonal element $\exp(\boldsymbol{X}_t\boldsymbol{\beta})$. The Hessian is negative definite if there is variation in the rows of $\boldsymbol{X}$ implying global concavity of $\ell(\boldsymbol{y}, \boldsymbol{\beta})$ and thus a unique maximum. Note the resemblance of FOC and Hessian to the binary logit model.

7.5 Truncated and Censored Models

Both situations can be modelled using the latent model

$$y_t^* = \boldsymbol{X}_t\boldsymbol{\beta} + u_t \qquad\qquad u_t \sim IIN(0, \sigma^2)$$

and the measurement equation

$$y_t = \begin{cases} y_t^* & \text{if } y_t^* > 0 \\ 0 & \text{else} \end{cases}$$

7.5.1 Truncated Regression Models

If only the observations $y_t^* > 0$ are observed the sample is truncated. Thus we get

$$\begin{aligned} P(y_t^* > 0) &= P(\boldsymbol{X}_t\boldsymbol{\beta} + u_t > 0) \\ &= 1 - P(u_t < -\boldsymbol{X}_t\boldsymbol{\beta}) = 1 - P(u_t/\sigma < -\boldsymbol{X}_t\boldsymbol{\beta}/\sigma) \\ &= 1 - \Phi(-\boldsymbol{X}_t\boldsymbol{\beta}/\sigma) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma) \end{aligned}$$

If all the latent values could be observed directly their density is equal to $\sigma^{-1}\phi((y_t^* - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma)$. Truncation requires that it has to be multiplied by a factor so that it integrates to one:

$$\frac{\sigma^{-1}\phi((y_t^* - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma)}{\Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma)}$$

Assume we have n_1 truncation points so that we have $n_2 = n - n_1$ observations. The loglikelihood function then looks like

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2) = -\frac{n_2}{2}\log(2\pi) - n_2 \log\sigma - \frac{1}{2\sigma^2}\sum_{t=1}^{n_2}(y_t^* - \boldsymbol{X}_t\boldsymbol{\beta})^2 - \sum_{t=1}^{n_2}\log\Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma)$$

Without the last term it is just the loglikelihood function of a classical normal regression. It shows that OLS estimation with the truncated sample yields inconsistent estimates since it neglects the last term in the loglikelihood function.

7.5.2 Censored Regression Models

If we observe all the y_t values including the n_1 zeros we call that the *Tobit model*. Note that if no observation is censored the loglikelihood function contribution of an observation

$$\ell_t(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2) = \sigma^{-1}\phi((y_t^* - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma)$$

would result. On the other hand, since

$$P(y_t = 0) = P(y_t^* \leq 0) = P(u_t \leq \boldsymbol{X}_t\boldsymbol{\beta}) = \Phi(-\boldsymbol{X}_t\boldsymbol{\beta}/\sigma)$$

we have loglikelihood contributions of the censored observations

$$\ell_t(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2) = \log\Phi((-\boldsymbol{X}_t\boldsymbol{\beta})/\sigma)$$

Combining both yields the loglikelihood function of the Tobit model

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2) = \sum_{y_t=0}\log\Phi((-\boldsymbol{X}_t\boldsymbol{\beta})/\sigma) + \sum_{y_t>0}\log\left(\sigma^{-1}\phi((y_t^* - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma)\right)$$

An OLS regression using the y_t information thus is inconsistent as well since it is equivalent to a loglikelihood function

$$\sum_{y_t=0}\log\left(\sigma^{-1}\phi((0 - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma)\right) + \sum_{y_t>0}\log\left(\sigma^{-1}\phi((y_t^* - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma)\right)$$

It is interesting to note that we can write the loglikelihood function as

$$\begin{aligned}\ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2) \;=\; & \sum_{y_t=0}\log\Phi((-\boldsymbol{X}_t\boldsymbol{\beta})/\sigma) + \sum_{y_t>0}\log\Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma) \\ & + \sum_{y_t>0}\log\left(\sigma^{-1}\phi((y_t^* - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma)\right) - \sum_{y_t>0}\log\Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma)\end{aligned}$$

The first two terms constitute the loglikelihood function of a binary probit just using the censored/not censored information while the last two terms are the loglikelihood function of the truncated sample

7.6 Sample Selectivity Models

Closely related to the tobit model is the sample selectivity model:

$$\begin{pmatrix} y_t^* \\ z_t^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_t\boldsymbol{\beta} \\ \boldsymbol{W}_t\boldsymbol{\gamma} \end{pmatrix} + \begin{pmatrix} u_t \\ v_t \end{pmatrix} \qquad \begin{pmatrix} u_t \\ v_t \end{pmatrix} \sim IIN\left(\mathbf{0}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right)$$

Measurement equations

$$y_t = \begin{cases} y_t^* & \text{if } z_t^* > 0 \\ \text{unobserved} & \text{otherwise} \end{cases}$$

$$z_t = \begin{cases} 1 & \text{if } z_t^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

A loglikelihood contribution can be written as

$$(1 - z_t)P(z_t = 0) + z_t P(z_t = 1) f(y_t^* | z_t = 1)$$

A joint density $f(y, z)$ can either be factorized as $f(y|z)f(z)$ or as $f(z|y)f(y)$ which is also true for our context. Thus we are concerned in finding an expression for $P(z_t = 1|y_t^*)$. In doing so we apply this factorization also to the joint distribution of u_t and v_t. The conditional distribution of a normal distribution again is normal. Therefore, we can write

$$v_t = \frac{\rho}{\sigma} u_t + \varepsilon_t \qquad \varepsilon_t \sim IIN(0, 1 - \rho^2)$$

To see that this is true we calculate the unconditional first two moments of v_t:

$$E(v_t) = \frac{\rho}{\sigma} E(u_t) + E(\varepsilon_t) = 0 \qquad \text{Var}(v_t) = \frac{\rho^2}{\sigma^2} \text{Var}(u_t) + (1 - \rho^2) = 1$$

Now we can calculate $P(z_t = 1|y_t^*)$:

$$P(z_t = 1|y_t^*) = \Phi\left(\frac{\boldsymbol{W}_t\boldsymbol{\gamma} + \rho(y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma}{(1 - \rho^2)^{1/2}}\right)$$

Combining everything yields the loglikelihhod function

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \sigma^2) = \sum_{z_t=0} \log \Phi(-\boldsymbol{W}_t\boldsymbol{\gamma}) \quad + \quad \sum_{z_t=1} \log \left(\frac{1}{\sigma} \phi((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma) \right) + \quad \sum_{z_t=1} \log \Phi \left(\frac{\boldsymbol{W}_t\boldsymbol{\gamma} + \rho(y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma}{(1-\rho^2)^{1/2}} \right)$$

In practice, the maximum likelihood estimation is often replaced by *Heckman's two-step estimator*.

7.7 Duration Models

Non-negative continuous variables are often modelled using gamma distribution, log-normal distribution, or Weibull distribution.

Gamma Distribution

The density function of a gamma distribution is

$$f(t) = \frac{\kappa^\theta}{\Gamma(\theta)} t^{\theta-1} \cdot \exp(-\kappa \cdot t) \qquad \kappa > 0,\ \theta > 0,\ t > 0$$

We obtain $\mathrm{E}(t) = \theta/\kappa$ and $\mathrm{Var}(t) = \theta/\kappa^2$. Note, if $\theta = 1$ we obtain the exponential distributions. The mean depends on two parameters so that it possible to model either of these as a function of explanatory variables. For instance, if we model $\kappa(\boldsymbol{X}_i) = \exp(\boldsymbol{X}_i\boldsymbol{\beta})$ we have

$$\mathrm{E}(t_i|\ \boldsymbol{X}_i) = \frac{\theta}{\exp(\boldsymbol{X}_i\boldsymbol{\beta})} = \theta \cdot \exp(-\boldsymbol{X}_i\boldsymbol{\beta})$$

$$\Rightarrow \frac{\partial \mathrm{E}(t_i|\ \boldsymbol{X}_i)}{\partial x_{ij}} = \underbrace{\theta \cdot \exp(-\boldsymbol{X}_i\boldsymbol{\beta})}_{>0} \cdot (-\beta_j)$$

The loglikelihood function then is

$$\ell(\boldsymbol{t}, \boldsymbol{\beta}, \theta) = \sum_{i=1}^{n} \theta \cdot \boldsymbol{X}_i\boldsymbol{\beta} - \log \Gamma(\theta) + (\theta - 1) \log t_i - t_i \cdot \exp(\boldsymbol{X}_i\boldsymbol{\beta})$$

Log-Normal Distribution

Density function:

$$f(t;\mu,\sigma^2) = \frac{1}{t\cdot\sigma\sqrt{2\pi}}\cdot\exp\left(-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right)$$

The mean value of the random variable t

$$\Rightarrow E(t) = \exp\left(\mu+\frac{\sigma^2}{2}\right)$$

again depends on two parameters allowing to specify either $\mu_i = \mu(\boldsymbol{X}_i)$ or $\sigma_i^2 = \sigma^2(\boldsymbol{X}_i)$. Usually the first alternative is chosen. With that the loglikelihood function looks like

$$\ell(\boldsymbol{t};\mu,\sigma^2) = \sum_{i=1}^{n} -\log(t_i\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(\log(t_i)-\boldsymbol{X}_i\boldsymbol{\beta})^2$$

Considering the F.O.C.

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{t};\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(\log(t_i)-\boldsymbol{X}_i\boldsymbol{\beta})\cdot(-\boldsymbol{X}_i)' \stackrel{!}{=} 0\\ \frac{\partial\ell(\boldsymbol{t};\boldsymbol{\beta},\sigma^2)}{\partial\sigma} &= \sum_{i=1}^{n} -\frac{1}{\sigma} - \frac{(-2)}{2\sigma^3}(\log(t_i)-\boldsymbol{X}_i\boldsymbol{\beta})^2 \stackrel{!}{=} 0\end{aligned}$$

we see from the first set of equations that this is simply an OLS regression of $\log(t_i)$ on $\boldsymbol{X}_i$ and thus $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{t}}$ with $\tilde{t_i} = \log(t_i)$. The second set yields

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(\ln(t_i)-\boldsymbol{X}_i\hat{\boldsymbol{\beta}})^2$$

Weibull Distribution

Density function:

$$f(t;\alpha,\theta) = \alpha\theta^\alpha t^{\alpha-1}\exp(-(\theta t)^\alpha) \qquad\qquad \alpha>0,\theta>0.$$

The mean value results

$$E(t) = \frac{1}{\theta}\cdot\Gamma\left(1-\frac{1}{\alpha}\right)$$

for which commonly $\theta_i = \theta(\boldsymbol{X}_i) = \exp(\boldsymbol{X}_i\boldsymbol{\beta})$ is specified.

Hazard Functions

As an alternative to modelling the mean value of a distribution of random variable T we can specify the hazard rate as a function of explanatory variables. For a distribution with density $f(t)$, cdf $F(t)$, and *survivor function* $S(t) = 1 - F(t)$ the hazard rate is defined

$$h(t) = \frac{f(t)}{S(t)}$$

For interpretation it is quite interesting to note that

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \cdot P(t \leq T \leq t + \Delta t | \, T \geq t) \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \cdot \frac{P(t \leq T \leq t + \Delta t \wedge T \geq t)}{P(T \geq t)} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \cdot \frac{F(t + \Delta t) - F(t)}{S(t)} = \frac{F'(t)}{S(t)} \end{aligned}$$

For the loglikelihood function we use the fact that $f(t) = h(t)S(t)$

$$\ell(\boldsymbol{t}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(t_i | \boldsymbol{X}_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log h(t_i | \boldsymbol{X}_i, \boldsymbol{\theta}) + \sum_{i=1}^{n} \log S(t_i | \boldsymbol{X}_i, \boldsymbol{\theta})$$

For instance, the Weibull distribution has cdf $F(t) = 1 - \exp(-(\theta t)^{\alpha})$ and therefore a hazard function $h(t) = \alpha \theta^{\alpha} t^{\alpha - 1}$. We see that if we model $\theta_i = \exp(\boldsymbol{X}_i \boldsymbol{\beta})$ it affects the mean value as well as the hazard function.

Right Censoring

At time t' not all the spells might have ended so that we have censored observation. In section 7.5.2 with left censoring (censoring of $y_t^* < 0$) we saw that the loglikelihood function has the form

$$\sum_{i \in C} \log F + \sum_{i \in U} \log f$$

Right censoring is therefore characterized by

$$\sum_{i \in U} \log f + \sum_{i \in C} \log(1 - F(t')) = \sum_{i \in U} \log f + \sum_{i \in C} \log S(t')$$

Thus the loglikelihood function for a sample of right censored observation then is

$$\begin{aligned}
\ell(\boldsymbol{t}, \boldsymbol{\theta}) &= \sum_{i \in U} \log f(t_i | \boldsymbol{X}_i, \boldsymbol{\theta}) + \sum_{i \in C} \log S(t') \\
&= \sum_{i \in U} \log h(t_i | \boldsymbol{X}_i, \boldsymbol{\theta}) + \sum_{i \in U} \log S(t_i | \boldsymbol{X}_i, \boldsymbol{\theta}) + \sum_{i \in C} \log S(t' | \boldsymbol{X}_i, \boldsymbol{\theta}) \\
&= \sum_{i \in U} \log h(t_i | \boldsymbol{X}_i, \boldsymbol{\theta}) + \sum_{i=1}^{n} \log S(t_i | \boldsymbol{X}_i, \boldsymbol{\theta})
\end{aligned}$$

8 Time Series Analysis

In section 6.4 we already introduced AR(p) processes to model time series behavior. We now use AR and Moving-Average (MA) processes to model a univariate time series of data $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)'$.

8.1 Stationary Univariate Processes

We already introduced the notion of weak stationarity and also the lag operator $Ly_t = y_{t-1}$. Besides AR(p) we also consider MA(q) which is given by

$$y_t = \varepsilon_t - \alpha_1\varepsilon_{t-1} - \alpha_2\varepsilon_{t-2} - \ldots - \alpha_q\varepsilon_{t-q} \qquad \varepsilon_t \sim IID(0, \sigma_\varepsilon^2)$$

Thus, the error term is assumed to be a stationary process called **White Noise**. It is easy to see that an MA(q) is always stationary since

$$\text{Var}(y_t) = (1 + \alpha_1^2 + \alpha_2^2 + \ldots + \alpha_q^2)\sigma_\varepsilon^2$$

We start with a AR(1) process

$$y_t = \theta y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim IID(0, \sigma_\varepsilon^2)$$

Repeated substitution yields:

$$\begin{aligned} y_t &= \theta y_{t-1} + \varepsilon_t \quad = \theta(\theta y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \theta^2 y_{t-2} + \varepsilon_t + \theta\varepsilon_{t-1} \\ &= \theta^2(\theta y_{t-3} + \varepsilon_{t-2}) + \varepsilon_t + \theta\varepsilon_{t-1} = \theta^3 y_{t-3} + \varepsilon_t + \theta\varepsilon_{t-1} + \theta^2\varepsilon_{t-2} \\ &\vdots \\ &= \theta^k y_{t-k} + \varepsilon_t + \theta\varepsilon_{t-1} + \theta^2\varepsilon_{t-2} + \ldots + \theta^{k-1}\varepsilon_{t-k+1} \end{aligned}$$

Provided that $|\theta| < 1$ the first term vanishes if $k \to \infty$ and we obtain an MA(∞) as a representation of the AR(1) process having

$$\text{Var}(y_t) = (1 + \alpha_1^2 + \alpha_2^2 + \ldots)\sigma_\varepsilon^2 = \sigma_\varepsilon^2 \sum_{j=0}^{\infty}(\alpha_j^j)^2 = \sigma_\varepsilon^2 \sum_{j=0}^{\infty}(\alpha_j^2)^j = \frac{\sigma_\varepsilon^2}{1 - \alpha_j^2}$$

This is called **inversion** of an AR(1) into an MA(∞) which can shortly be written by using the lag operator ($y_t = \theta L y_t + \varepsilon_t$)

$$(1 - \theta L)y_t = \varepsilon_t \quad \Leftrightarrow \quad y_t = (1 - \theta L)^{-1}\varepsilon_t = (1 + \theta L + \theta^2 L^2 + \theta^3 L^3 + \ldots)\varepsilon_t$$

Repeated substitution can also be done for an MA(1)

$$\begin{aligned}
\varepsilon_t &= y_t + \alpha\varepsilon_{t-1} \quad = y_t + \alpha(y_{t-1} + \alpha\varepsilon_{t-2}) = y_t + \alpha y_{t-1} + \alpha^2\varepsilon_{t-2} \\
&= y_t + \alpha y_{t-1} + \alpha^2(y_{t-2} + \alpha\varepsilon_{t-3}) = y_t + \alpha y_{t-1} + \alpha^2 y_{t-2} + \alpha^3\varepsilon_{t-3} \\
&\vdots \\
&= y_t + \alpha y_{t-1} + \alpha^2 y_{t-2} + \ldots + \alpha^k y_{t-k} + \alpha^{k+1}\varepsilon_{t-k-1}
\end{aligned}$$

Given that $|\alpha| < 1$ the MA(1) is represented by an AR(∞)

$$y_t = (1 - \alpha L)\varepsilon_t \quad \Leftrightarrow \quad \varepsilon_t = (1 - \alpha L)^{-1} y_t = (1 + \alpha L + \alpha^2 L^2 + \alpha^3 L^3 + \ldots)y_t$$

An AR(2) model

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t = \theta_1 L y_t + \theta_2 L^2 y_t + \varepsilon_t \quad \Leftrightarrow \quad (1 - \theta_1 L - \theta_2 L^2)y_t = \varepsilon_t$$

is stationary if the (possibly complex) roots of the characteristic equation

$$\theta(z) = 1 - \theta_1 z - \theta_2 z^2 = 0$$

all lie outside the unit circle. Thus the process $y_t = 1.4y_{t-1} - 0.9y_{t-2} + \varepsilon_t$ is stationary since the roots are 1.25 and 4 which is also true for $y_t = 1.05y_{t-1} - 0.2y_{t-2} + \varepsilon_t$ having complex roots outside the unit circle.

The combination of AR(p) and MA(q) yields the ARMA(p,q) model

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \ldots + \theta_p y_{t-p} + \varepsilon_t - \alpha_1\varepsilon_{t-1} - \alpha_2\varepsilon_{t-2} - \ldots - \alpha_q\varepsilon_{t-q}$$

which can be be written as $\theta(L)y_t = \alpha(L)\varepsilon_t$ and thus (if invertability is given) $y_t = \theta(L)^{-1}\alpha(L)\varepsilon_t$ or $\alpha(L)^{-1}\theta(L)y_t = \varepsilon_t$.

An AR(p) process can be estimated using OLS regressing y_t on its lagged observations. As stated in chapter 2 the assumption of $E(\varepsilon_t | y_{t-1}, \ldots y_{t-p}) = 0$ is, however, not sufficient to ensure unbiasedness of the estimator but it ensures consistency (see section 3.2).

8.2 Stationary Multivariate Processes, VAR

Now we consider a vector of m time series $\boldsymbol{y}_t = (y_{1,t}, y_{2,t}, \ldots y_{m,t})'$ and model them simultaneously using a stationary autoregressive model

$$\boldsymbol{y}_t = \boldsymbol{\Theta}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t \qquad\qquad \boldsymbol{\varepsilon}_t \sim IID(\mathbf{0}, \boldsymbol{\Sigma})$$

As an example, a bivariate vector autoregressive model reads like

$$\begin{aligned} y_{1,t} &= \theta_{11} y_{1,t-1} + \theta_{12} y_{2,t-1} + \varepsilon_{1,t} \\ y_{2,t} &= \theta_{21} y_{1,t-1} + \theta_{22} y_{2,t-1} + \varepsilon_{2,t} \\ \Leftrightarrow \quad \begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} &= \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \end{aligned}$$

If e.g. $\theta_{22} = 0$ and we insert the second equation into the first, the system reduce to a univariate AR(2) process of $y_{1,t}$. The corresponding parameters are $\theta_1 = \theta_{11}$ and $\theta_2 = \theta_{12}\theta_{21}$, respectively.

To determine the joint stationarity of vector $\boldsymbol{y}_t$ we have to solve the characteristic equation obtained from $|\boldsymbol{I} - \boldsymbol{\Theta}_1 z| = 0$:

$$\begin{vmatrix} 1 - \theta_{11} z & -\theta_{12} z \\ -\theta_{21} z & 1 - \theta_{22} z \end{vmatrix} = (1 - \theta_{11} z)(1 - \theta_{22} z) - \theta_{12}\theta_{21} z^2 = 0$$

If all the roots of this equation lie outside the unit circle the multivariate process is jointly stationary. For example, for

$$\boldsymbol{\Theta}_1 = \begin{pmatrix} .8 & -.6 \\ .7 & .2 \end{pmatrix}$$

the complex roots of the characteristic function lie outside the unit circle so that the joint process is stationary. Another aspect in this context is **Granger causality**. In this simple bivariate context y_2 does not Granger cause y_1 if θ_{12} is equal to zero.

The generalization is the VAR(p) where vector $\boldsymbol{y}_t$ is explained by lagged vectors up to order p:

$$\begin{aligned} & \boldsymbol{y}_t = \boldsymbol{\Theta}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Theta}_2 \boldsymbol{y}_{t-2} + \ldots + \boldsymbol{\Theta}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t \\ \Leftrightarrow \quad & (\boldsymbol{I} - \boldsymbol{\Theta}_1 L - \boldsymbol{\Theta}_2 L^2 - \ldots - \boldsymbol{\Theta}_p L^p) \boldsymbol{y}_t = \boldsymbol{\varepsilon}_t \quad \Leftrightarrow \boldsymbol{\Theta}(L) \boldsymbol{y}_t = \boldsymbol{\varepsilon}_t \end{aligned}$$

The process is jointly stationary if the roots of the characteristic equation $|\boldsymbol{\Theta}(z)| = 0$ all lie outside the unit circle.

Estimation of a VAR(p) can be performed with a SUR. Since each equation depends on lags of all the m variables the regressor matrix is the same for each variable. Therefore, the VAR system can be estimated equation by equation using OLS to obtain asymptotically efficient estimators.

If the VAR(p) is invertible, i.e. $|\boldsymbol{\Theta}(1)| = |(\boldsymbol{I} - \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2 - \ldots - \boldsymbol{\Theta}_p)| \neq 0$, it can be written as a VMA(∞)

$$\boldsymbol{y}_t = \boldsymbol{\Theta}(L)^{-1}\boldsymbol{\varepsilon}_t = \boldsymbol{\varepsilon}_t + \boldsymbol{A}_1\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{A}_2\boldsymbol{\varepsilon}_{t-2} + \ldots$$

from which we derive

$$\frac{\partial \boldsymbol{y}_{t+s}}{\partial \boldsymbol{\varepsilon}_t'} = \boldsymbol{A}_s \qquad s > 0$$

Thus a change of the k-th component in $\boldsymbol{\varepsilon}_t$ by one unit we find the effects on $\boldsymbol{y}_{t+s}$ in the k-th column of $\boldsymbol{A}_s$. Therefore, we derive the dynamic effects of such a change on the j-th component of vector $\boldsymbol{y}$ by plotting the j,k-elements of matrices $\boldsymbol{I}, \boldsymbol{A}_1, \boldsymbol{A}_2, \ldots$ which is called the **impulse-response function**.

8.3 Nonstationary Time Series

8.3.1 Deterministic and Stochastic Trends

Economic time series like GDP, consumption, investments, etc. often show a positive trend. With a slightly modified AR(p)

$$y_t = \mu + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \ldots + \theta_p y_{t-p} + \varepsilon_t \tag{14}$$

we obtain for a stationary process an unconditional mean of

$$E(y_t) = \frac{\mu}{1 - \sum_{j=1}^{p} \theta_j}$$

which is obviously not a good description of the real data. The same is true for a modified MA(q) $y_t = \mu + \alpha(L)\varepsilon_t$ having an unconditional mean of μ. One way

to obtain a nonstationary series with an increasing unconditional mean is a **deterministic trend** model

$$y_t = \mu + \delta \cdot t + \alpha(L)\varepsilon_t$$

Due to the MA structure it is stationary around the trend. Another common model is the *random walk with drift*

$$y_t = \delta + y_{t-1} + \varepsilon_t \tag{15}$$

Note that $\Delta y_t = (1-L)y_t = y_t - y_{t-1} = \delta + \varepsilon_t$. From section 8.1 it is clear that the process with an assumed $y_0 = 0$ can be written as

$$y_t = \delta \cdot t + \sum_{i=1}^{t} \varepsilon_i$$

Comparing an AR(1) from (14) and (15) we find that the constant term has a different meaning once θ_1 becomes zero which also holds true for the variances. Due to that the process is said to have a **stochastic trend**. Processes like (15) are called *integrated* of order one, or $I(1)$, since the first differences are stationary, $I(0)$. It is interesting to consider the effects of differencing a stationary AR(1) process

$$y_t = \mu + \theta_1 y_{t-1} + \varepsilon_t \quad \Leftrightarrow \quad y_t - y_{t-1} = \mu + \theta_1 y_{t-1} - (\mu + \theta_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t$$

$$\Leftrightarrow \quad \Delta y_t = \theta_1 \Delta y_{t-1} - (1-L)\varepsilon_t$$

Thus the resulting ARMA(1,1) process contains a noninvertible MA(1) (unit root in the MA part) which is sometimes called an *over-differenced* process.

In the case of an integrated random walk with drift *and* trend

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

we get by setting $y_0 = 0$

$$\begin{aligned} y_1 &= \delta + \gamma \cdot 1 + \varepsilon_1 \\ y_2 &= 2\delta + \gamma \cdot (1+2) + \varepsilon_1 + \varepsilon_2 \\ &\vdots \\ y_t &= \delta \cdot t + \gamma \cdot \frac{t(t+1)}{2} + \sum_{i=1}^{t} \varepsilon_i \end{aligned}$$

The series shows a quadratic trend and an increasing variance whereas the differenced series is white noise plus a constant: $\Delta y_t = \gamma + \varepsilon_t$. Moreover, the differenced series of a process

$$y_t = \delta + \gamma t + \lambda t^2 + y_{t-1} + \varepsilon_t$$

has the form

$$\Delta y_t = (\gamma - \kappa) + (2\kappa) \cdot t + \varepsilon_t$$

8.3.2 Estimation of AR(1) Models

Estimation of a stationary AR(1) model which has no constant using OLS obtains

$$\hat{\theta} = \frac{\sum_{t=2}^{T} y_{t-1} y_t}{\sum_{t=2}^{T} y_{t-1}^2} = \frac{T^{-1} \sum_{t=2}^{T} y_{t-1}(\theta y_{t-1} + \varepsilon_t)}{T^{-1} \sum_{t=2}^{T} y_{t-1}^2} = \theta + \frac{T^{-1} \sum_{t=2}^{T} y_{t-1}\varepsilon_t}{T^{-1} \sum_{t=2}^{T} y_{t-1}^2}$$

Rearranging the first equation yields $\sum_{t=2}^{T} y_{t-1} y_t = \hat{\theta} \sum_{t=2}^{T} y_{t-1}^2$. The estimator's variance is

$$\mathrm{Var}(\hat{\theta}) = \frac{\sigma_\varepsilon^2}{\sum_{t=2}^{T} y_{t-1}^2}$$

so that via the estimate of σ_ε^2

$$\begin{aligned}
\hat{\sigma}_\varepsilon^2 &= T^{-1} \sum_{t=2}^{T} (y_t - \hat{\theta} y_{t-1})^2 = T^{-1} (\sum_{t=2}^{T} y_t^2 - 2\hat{\theta} \sum_{t=2}^{T} y_t y_{t-1} + \hat{\theta}^2 \sum_{t=2}^{T} y_{t-1}^2) \\
&= T^{-1} (\sum_{t=2}^{T} y_t^2 - \hat{\theta}^2 \sum_{t=2}^{T} y_{t-1}^2) = T^{-1} (\sum_{t=3}^{T+1} y_{t-1}^2 - \hat{\theta}^2 \sum_{t=2}^{T} y_{t-1}^2) \\
&\approx T^{-1} (1 - \hat{\theta}^2) \sum_{t=2}^{T} y_{t-1}^2
\end{aligned}$$

we get

$$\widehat{\mathrm{Var}(\hat{\theta})} \approx T^{-1}(1 - \hat{\theta}^2)$$

and thus using asymptotic reasoning

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{a} N(0, 1 - \theta^2)$$

Thus the t-statistic is asymptotically standard normally distributed. It can be shown that for the nonstationary AR(1) with $\theta = 1$ the convergence rate is T and not $\sqrt{T}$

so that under $H_0: \ \theta = 1$ it holds

$$T(\hat{\theta} - 1) \xrightarrow{a} N(0, V)$$

There is no closed analytical expression for variance V but it can be obtained using Monte Carlo methods. As a consequence the usual t-statistic

$$t = \frac{\sqrt{T}(\hat{\theta} - 1)}{\sqrt{1 - \hat{\theta}^2}} \tag{16}$$

is not asymptotically standard normal and needs to be simulated as well.

8.3.3 Unit Roots Tests

Several papers by Dickey and Fuller (1979) give critical values for the t-statistic in (16) of the one sided test of $H_0: \ \theta = 1$ for different T and different specifications. In order to obtain the test statistic immediately from a standard regression output the regression model is specified as

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

To allow for a trend stationary AR(1) as the alternative we could specify a constant and a time trend in the regression specification. Note that in the presence of a unit root it implies a deterministic linear and quadratic time trend as shown in section 8.3.1. To be precise, in such a situation the the null hypothesis should read like $H_0: \ \theta = 1$, *no constant, no trend* which could be tested jointly. Due to practical reasons, however, solely $\theta - 1 = 0$ is usually tested.

Critical values for $\alpha = 5\%$ and different specifications are:

	No constant	Constant	Constant
T	No trend	No trend	Trend
25	-1.95	-3.00	-3.60
50	-1.95	-2.93	-3.50
500	-1.95	-2.87	-3.42
$\vdots$			
∞	-1.95	-2.86	-3.41

Higher Order AR Processes

An AR(2) model can be rewritten as

$$\begin{aligned} y_t &= \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t \\ \Leftrightarrow y_t - y_{t-1} &= \theta_1 y_{t-1} - y_{t-1} + \theta_2 y_{t-1} - \theta_2 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t \\ \Leftrightarrow \Delta y_t &= (\theta_1 + \theta_2 - 1) y_{t-1} - \theta_2 \Delta y_{t-1} + \varepsilon_t \end{aligned}$$

From the characteristic equation of this AR model $1 - \theta_1 z - \theta_2 z^2 = 0$ we see that if the roots are real values then the process is nonstationary if $z = 1$ and thus $1 - \theta_1 - \theta_2 = 0$. Fortunately, under the null $H_0 : \theta_1 + \theta_2 - 1 = 0$ the critical values of the t-statistic $(\hat{\theta}_1 + \hat{\theta}_2 - 1)/\sqrt{\text{Var}(\hat{\theta}_1 + \hat{\theta}_2)}$ are the same as for the Dickey-Fuller test given above. A generalization is called the **Augmented Dickey-Fuller test** (ADF) which has the form

$$\Delta y_t = \kappa y_{t-1} + \vartheta_1 \Delta y_{t-1} + \ldots + \vartheta_{p-1} \Delta y_{t-p+1} + \varepsilon_t$$

and we test $H_0 : \kappa = 0$ which is equivalent to $(\sum_{i=1}^{p} \theta_i \ - 1)$ against $H_1 \ : \ \kappa < 0$.

8.4 Cointegration

8.4.1 Error Correction model

A distributed lag model

$$Y_t = \mu + \sum_{i=0}^{m} \beta_i X_{t-i} + u_t$$

with β_i being proportional to probabilities of a geometric distribution with parameter λ

$$\beta_i = \beta \cdot P(t = i) = \beta \cdot \lambda^i \cdot (1 - \lambda)$$

after the Koyck-Transformation can be written as

$$Y_t = \mu(1 - \lambda) + \lambda Y_{t-1} + \beta(1 - \lambda) X_t + u_t - \lambda u_{t-1}$$

A more general form of these distributed lag models can be defined as

$$Y_t = \delta + \theta Y_{t-1} + \phi_0 X_t + \phi_1 X_{t-1} + \varepsilon_t \tag{17}$$

which usually called autoregressive distributed lag model ADL(1,1). Such a model with stationary Y_t and X_t shows an immediate effect of a change in X_t expressed by

$$\frac{\partial Y_t}{\partial X_t} = \phi_0$$

and an effect in next period of

$$\frac{\partial Y_{t+1}}{\partial X_t} = \theta \frac{\partial Y_t}{\partial X_t} + \phi_1 = \theta\phi_0 + \phi_1$$

and after two periods

$$\frac{\partial Y_{t+2}}{\partial X_t} = \theta \frac{\partial Y_{t+1}}{\partial X_t} = \theta(\theta\phi_0 + \phi_1)$$

and so on. If we impose $|\theta| < 1$ the long-run effect of a one unit change in X_t is

$$\begin{aligned} \phi_0 &+ (\theta\phi_0 + \phi_1) + \theta(\theta\phi_0 + \phi_1) + \theta^2(\theta\phi_0 + \phi_1) + \ldots \\ &= \phi_0 + (\theta\phi_0 + \phi_1)(1 + \theta + \theta^2 + \ldots) = \frac{\phi_0(1-\theta)}{1-\theta} + \frac{\theta\phi_0 + \phi_1}{1-\theta} \\ &= \frac{\phi_0 + \phi_1}{1-\theta} \end{aligned}$$

The long-run equilibrium of (17) can be written as

$$E(Y_t) = \frac{\delta}{1-\theta} + \frac{\phi_0 + \phi_1}{1-\theta} E(X_t) = \alpha + \beta E(X_t)$$

If we subtract Y_{t-1} on both sides of (17) and expand with $\phi_0 X_{t-1} - \phi_0 X_{t-1}$ we obtain

$$\Delta Y_t = \delta - (1-\theta) Y_{t-1} + \phi_0 \Delta X_t + (\phi_1 - \phi_0) X_{t-1} + \varepsilon_t$$

or

$$\Delta Y_t = \phi_0 \Delta X_t - (1-\theta)(Y_{t-1} - \alpha - \beta X_{t-1}) + \varepsilon_t$$

which is called the **error correction** formulation of the ADL(1,1). If Y_{t-1} is above its long-run equilibrium value $\alpha + \beta X_{t-1}$ a negative adjustment on Y_t is generated with $(1-\theta)$ being the adjustment speed.

8.4.2 Spurious Regression

Let us assume a deterministic trend in the two variables Y_t and X_t

$$\begin{aligned} Y_t &= a + b \cdot t + \varepsilon_{1,t} \qquad & \varepsilon_{1,t} \sim IID(0, \sigma^2_{\varepsilon_1}) \\ X_t &= c + d \cdot t + \varepsilon_{2,t} \qquad & \varepsilon_{2,t} \sim IID(0, \sigma^2_{\varepsilon_2}) \end{aligned}$$

with independent error terms. A regression of Y_t on X_t will result in usually highly significant regression coefficients and high R^2. In large samples the slope coefficient converges to b/d and the constant term to $a - c \cdot b/d$ due to the omitted variable t which is a **spurious regression**. However, regressing Y_t on X_t and t the coefficient is of X_t is mostly insignificant.

Considering two variables with stochastic trends like independent random walks

$$\begin{aligned} Y_t &= Y_{t-1} + \varepsilon_{1,t} \qquad & \varepsilon_{1,t} \sim IID(0, \sigma^2_{\varepsilon_1}) \\ X_t &= X_{t-1} + \varepsilon_{2,t} \qquad & \varepsilon_{2,t} \sim IID(0, \sigma^2_{\varepsilon_2}) \end{aligned}$$

we also obtain spurious results of high t- and F-statistics, a high R^2, and a Durbin-Watson statistic close to zero. The problem is that since Y_t and X_t are both $I(1)$, the error term ε_t is also $I(1)$ and therefore the t- and F-test statistics have no well-defined asymptotic distributions. However, a specification like (17) will obtain consistent estimates of $\theta = 1$ and $\phi_0 = \phi_1 = 0$ due to the residual being $I(0)$.

8.4.3 Cointegration of two Variables

If we have two random walk processes as described above where the two variables are each $I(1)$ and hence their respective first differences $I(0)$ but additionally there exists a value β such that $Y_t - \beta X_t$ is $I(0)$ then we describe those variables to be **cointegrated**. We call the vector $\boldsymbol{\beta} = (1; -\beta)'$ the cointegration vector for the variable vector $(Y_t; X_t)'$ since $\boldsymbol{\beta}'(Y_t; X_t)'$ is $I(0)$. The OLS estimates of the specification

$$Y_t = a + bX_t + e_t \tag{18}$$

are called *super consistent*. The intuition behind that is, that estimates $\hat{b} \neq \beta$ produces residuals that are not $I(0)$ and thus the sample variance of $\hat{e}_t$ is much larger than in the case of $\hat{b} = \beta$ in which the residuals are $I(0)$ and thus have smaller sample variance. Thus the estimates are forced to be closer to the true value than in a situation where the alternative is also $I(0)$. In fact, $\hat{b}$ is not $\sqrt{T}$ consistent for β but has a faster convergence rate of T.

Testing for cointegration requires in a first step separate unit root tests for Y_t and X_t, respectively. In a second step OLS estimates of (18) are computed and a Dickey-Fuller test for a unit root in the residuals is performed

$$\Delta\hat{e}_t = \varphi\hat{e}_t + \nu_t$$

If the null hypothesis $H_0 : \varphi = 0$ can be rejected the integrated variables Y_t and X_t are cointegrated. However, the Dickey-Fuller critical values need to be adjusted for the $\hat{e}_t$ are not a given time series but are estimated. Again simulation techniques can be used to determine those critical values for this case. For instance, for $T = 500$ the critical value for the usual test statistic of $\hat{\varphi}$ is -3.34 ($\alpha = 5\%$) instead of -2.87 for, say, Y_t.

Assuming a partial adjustment model for Y_t where the long-run target is $Y_t^* = \beta X_t$

$$\begin{aligned} \Delta Y_t &= \lambda_1 \Delta Y_t^* + \lambda_2(Y_{t-1} - Y_{t-1}^*) + \varepsilon_t \\ &= \lambda_1\beta\Delta X_t + \lambda_2(Y_{t-1} - \beta X_{t-1}) + \varepsilon_t \end{aligned}$$

where we get an error correction formulation where all the terms are $I(0)$. If we allow for a constant in the ECM it implies an additional deterministic trend in both variables.

8.4.4 Cointegration in a VAR

Extending the idea of cointegration to a k-vector of variables with more than two components each being $I(1)$ then there might be a $(k \times r)$ cointegration matrix $\boldsymbol{\beta}$ with rank $r \leq k-1$ such that $\boldsymbol{\beta}'\boldsymbol{y}_t$ describing various $I(0)$ equilibrium relations. We restrict our analysis to a VAR(3) process

$$\begin{aligned} & \boldsymbol{y}_t = \boldsymbol{\Theta}_1\boldsymbol{y}_{t-1} + \boldsymbol{\Theta}_2\boldsymbol{y}_{t-2} + \boldsymbol{\Theta}_3\boldsymbol{y}_{t-3} + \boldsymbol{\varepsilon}_t \\ \Leftrightarrow \quad & (\boldsymbol{I} - \boldsymbol{\Theta}_1 L - \boldsymbol{\Theta}_2 L^2 - \boldsymbol{\Theta}_3 L^3)\boldsymbol{y}_t = \boldsymbol{\varepsilon}_t \quad \Leftrightarrow \boldsymbol{\Theta}(L)\boldsymbol{y}_t = \boldsymbol{\varepsilon}_t \end{aligned}$$

We rewrite this in the manner of higher order univariate AR in 8.3.3.

$$\begin{aligned}\Delta \boldsymbol{y}_t &= (\boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2 - \boldsymbol{I})\boldsymbol{y}_{t-1} - \boldsymbol{\Theta}_2 \Delta \boldsymbol{y}_{t-1} + \boldsymbol{\Theta}_3 \boldsymbol{y}_{t-3} + \boldsymbol{\varepsilon}_t \\ &= (\boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2 + \boldsymbol{\Theta}_3 - \boldsymbol{I})\boldsymbol{y}_{t-1} - \boldsymbol{\Theta}_2 \Delta \boldsymbol{y}_{t-1} - \boldsymbol{\Theta}_3(\Delta \boldsymbol{y}_{t-1} + \Delta \boldsymbol{y}_{t-2}) + \boldsymbol{\varepsilon}_t \\ &= (\boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2 + \boldsymbol{\Theta}_3 - \boldsymbol{I})\boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_2 \Delta \boldsymbol{y}_{t-2} + \boldsymbol{\varepsilon}_t\end{aligned}$$

with $\boldsymbol{\Gamma}_1 = -\boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_3$ and $\boldsymbol{\Gamma}_2 = -\boldsymbol{\Theta}_3$. For a VAR(p) we obtain in the same fashion

$$\Delta \boldsymbol{y}_t = \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{\Pi} \boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t$$

with $\boldsymbol{\Pi} = -(\boldsymbol{I} - \boldsymbol{\Theta}_1 - \ldots - \boldsymbol{\Theta}_p) = -\boldsymbol{\Theta}(1)$. Now we can consider three cases:

- $\boldsymbol{\Pi} = \boldsymbol{0}$ which implies non-stationarity for the $\boldsymbol{y}_t$ components since $\boldsymbol{\Theta}(1) = \boldsymbol{0}$ and thus a stationary VAR(p-1) for $\Delta \boldsymbol{y}_t$.

- If all $\boldsymbol{y}_t$ components are stationary $\boldsymbol{\Pi} = -\boldsymbol{\Theta}(1)$ must have full rank and can be inverted. Thus a VMA representation $\boldsymbol{y}_t = \boldsymbol{\Theta}(L)^{-1}\boldsymbol{\varepsilon}_t$ is obtained

- $\boldsymbol{\Pi}$ is not zero and not invertible (having rank r) and can be written as $\boldsymbol{\Pi} = \boldsymbol{\gamma}\boldsymbol{\beta}'$ where the latter two matrices both have rank r. Thus we have r cointegrating relationships so that we obtain $\boldsymbol{z}_t = \boldsymbol{\beta}'\boldsymbol{y}_t$ which represent r stationary series. Inserting it in the last equation we can interpret it as a **vector error correction model** (VECM)

$$\Delta \boldsymbol{y}_t = \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{\gamma}\boldsymbol{\beta}' \boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t$$