Stochastic processes, stationarity and autocorrelation

Individual stochastic processes

- A stochastic process is an ordered sequence of random variables: $\ldots y_0, y_1, \ldots, y_t, y_{t+1}, \ldots$ or $\{y_t\}$ for short.
- Each panel member is characterised by a different stochastic process, so add subscript *i*: $\{y_{it}\}$
- Assume two things:
  - (A1) individuals *i* and *j* initially sampled at random at (say) time $t = 0$ $\Rightarrow$ joint distribution of y_{i0} and y_{j0} is $f_0(y_{i0}) \times f_0(y_{j0})$
  - (A2) after selection, there is no cross-individual influence $\Rightarrow$ joint distribution of $(y_{i1}, y_{j1} \mid y_{i0}, y_{j0})$ is $f_1(y_{i1} \mid y_{i0}) \times f_1(y_{j1} \mid y_{j0})$
- Implies that the whole sequence $\{y_{it}\}$ for individual *i* is statistically independent of the sequence $\{y_{jt}\}$ for any other individual *j*

Stationarity

- A stochastic process $\{y_t\}$ is stationary (specifically 2nd-order or covariance-stationary) if:
  - the mean, $E(y_t)$, is the same for all periods *t*
  - the variance, $Var(y_t)$, is the same for all periods *t*
  - each covariance, $Cov(y_t, y_s)$, only depends on the time separation $|t\text{-}s|$, not on the particular times *t* and *s* at which *y* is observed
- Stationarity $\Rightarrow$ in any sequence of consecutive periods we happen to observe, the process $\{y_t\}$ will tend to look much the same – at least in terms of its first and second moments
- Most panel data don't follow stationary processes – e.g. earnings, BMI, happiness, etc. tend to follow clear trends in level and volatility through the lifecourse
- But, stationary processes are useful as building blocks to construct realistic models

Five important stochastic processes

- *White noise* $\{\varepsilon_{it}\}$ is a sequence of independent random variables, each with the same mean 0 and variance σ_ε^2
- An *Equi-correlated* or *exchangeable* process can be expressed as the sum of a time-invariant random variable u and white noise ε_t

  $y_{it} = u_i + \varepsilon_{it}$
- A kth-order *autoregressive process*, or AR(k), involves feedback from k past values and a white noise innovation:

  $y_{it} = \rho_1 y_{it-1} + \ldots + \rho_k y_{it-k} + \varepsilon_{it}$
- A kth-order *moving average process*, or MA(k), is an average of k consecutive values from a white noise process:

  $y_{it} = \varepsilon_{it} + \theta_1 \varepsilon_{it-1} + \ldots + \theta_k \varepsilon_{it-k} + \varepsilon_{it}$
- A *random walk* is a ***non-stationary*** process that cumulates white noise innovations:

  without drift: $y_{it} = y_{it-1} + \varepsilon_{it}$

  with drift: $y_{it} = \alpha + y_{it-1} + \varepsilon_{it}$

Autocovariances and autocorrelations

For a stationary stochastic process $\{y_{it}\}$, define the jth-order autocovariance, $\gamma(j)$, and autocorrelation, $\rho(j)$ as:

$c(j) = \text{cov}(y_{it}, y_{it-j})$

$r(j) = \text{corr}(y_{it}, y_{it-j})$

For non-stationary processes, $c(j)$ and $r(j)$ may depend on t.

Non-stationary example: random walk

$y_{it} = \varepsilon_{it} + \varepsilon_{it-1} + \ldots + \varepsilon_{i1}$ (initial condition: $y_{i0} = 0$)

Variance: $\sigma_y^2 = t\sigma_\varepsilon^2 \to \infty$ as t increases

Autocovariance: $c(j) = (t-j)\,\sigma_\varepsilon^2$

Autocorrelation: $r(j) = (t-j) / [t\,(t-j)]^{1/2} \to 1$ as t increases, for all j

Autocovariances and autocorrelations

An equicorrelated process, $y_{it} = u_i + \varepsilon_{it}$, has variance and serial correlation:

$\sigma_y^2 = \sigma_u^2 + \sigma_\varepsilon^2$

$c(j) = \sigma_u^2$

$r(j) = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2)$ $\Rightarrow$ ***doesn't decay with increasing*** j

An AR(1) process, $y_{it} = \rho\, y_{it-1} + \varepsilon_{it}$, has variance and serial correlation:

$\sigma_y^2 = \sigma_\varepsilon^2/(1-\rho^2)$

$c(j) = \sigma_y^2 \rho^j$

$r(j) = \rho^j$ $\Rightarrow$ ***geometric decay***

An MA(1) process, $y_t = \varepsilon_t + \theta\,\varepsilon_{t-1}$, has variance and serial correlation:

$\sigma_y^2 = \sigma_\varepsilon^2(1+\theta^2)$

$c(1) = \sigma_\varepsilon^2\theta;\ c(j) = 0$ for all $j > 1$

$r(1) = \theta / (1+\theta^2)\,;\ r(j) = 0$ for all $j > 1$ $\Rightarrow$ ***abrupt decay***

NB $r(1)$ can never exceed 0.5 for an MA(1) process

Monte Carlo simulations

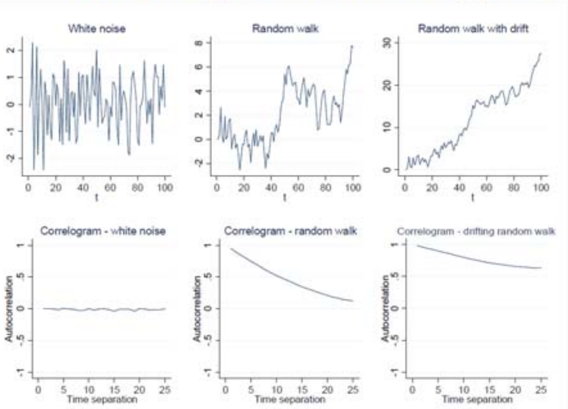
Six processes:

- White noise: $y_{it} = \varepsilon_{it}$
- Equi-correlated: $y_{it} = u_i + \varepsilon_{it}$
- AR(1): $y_{it} = 0.75\, y_{it\text{-}1} + \varepsilon_{it}$
- MA(1): $y_{it} = \varepsilon_{it} + 0.9\, \varepsilon_{it\text{-}1}$
- Random walk: $y_{it} = y_{it\text{-}1} + \varepsilon_{it}$
- Random walk with drift: $y_{it} = 0.2 + y_{it\text{-}1} + \varepsilon_{it}$

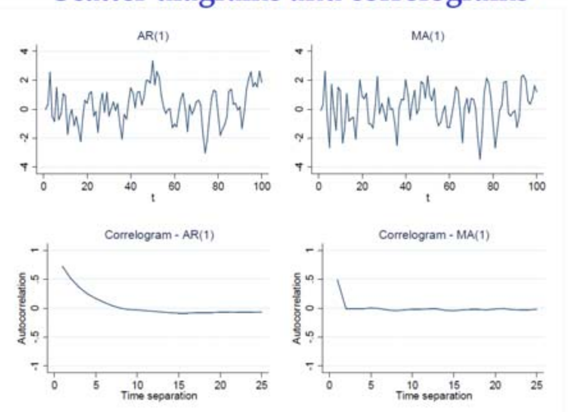
Method:

- generate pseudo-random N(0,1) values for u_i, ε_{i1} ... ε_{i100}
- from initial condition $y_{i0} = 0$, calculate sequence of y_{i1} ... y_{i100}
- scatter plot of $\{y_{it}\}$ for $i = 1$
- estimate and plot the first 25 elements of the correlogram

Scatter diagrams and correlograms

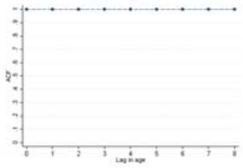


Scatter diagrams and correlograms

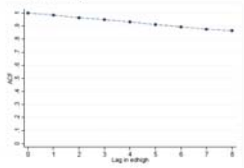


Some empirical correlograms (HILDA9)

(a) Age

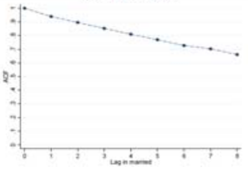


(b) Highest education level



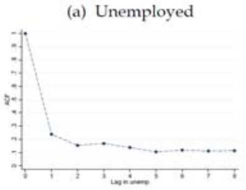
Method: correlate observed value at wave 9 with observed value at wave 9-*j*

(c) Married

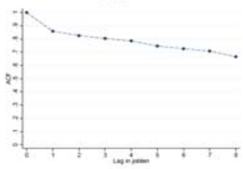


Some empirical correlograms (HILDA9)

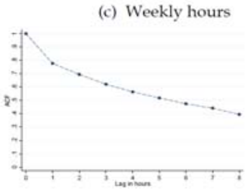
(a) Unemployed



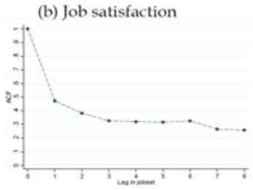
(b) Job tenure



(c) Weekly hours

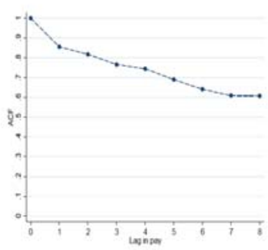


(b) Job satisfaction

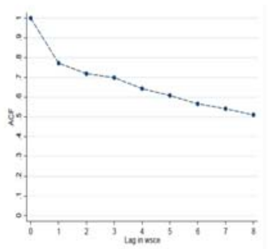


Some empirical correlograms (HILDA9)

Gross weekly earnings



(a) zeros excluded



(b) zeros included

Typical properties of real panel data

- Correlograms usually display slow decay, high persistence
- eg: autocorrelations at a separation of 8 years:
  - 1.0 (trivially) age, gender, etc.
  - 0.8-1.0 educational attainment
  - 0.6-0.8 marital status, job tenure
  - 0.4-0.6 weekly hours, earnings
  - 0.2-0.4 job satisfaction
  - < 0.2 unemployed
- Persistence ⇒ the total no. of observations, nT, greatly exaggerates the amount of information in a panel
- Should we assume that conditions external to the individual (i.e. the covariates) capture persistence adequately, or does behaviour display inherent persistence (e.g. habit, inertia)?