Panel data analysis: dynamics
Day 4.

David Pupovac

What did we do up until now?

- **<u>Day 1. Non-spherical errors</u>** - we have emphasized that the most common problems are:
  - panel heteroskedasticity
  - contemporaneous correlation of the errors
  - serially correlation of errors

- Solutions: **FGLS, PCSE and clustered sandwich estimator**

- **<u>Day 2. Fixed effects</u>**

$$y_{i,t} = \boldsymbol{\alpha_i} + x_{i,t}\beta + \varepsilon_{i,t}$$

- **<u>Day 3. Random effects</u>**

$$y_{i,t} = \alpha + x_{i,t}\beta + \boldsymbol{\nu_i} + \varepsilon_{i,t}$$

  - If you need to apply RE to a small sample:
    Swamy-Arora Estimator
  - If you have to apply RE but there is correlation of x and ν_i:
    Hausman-Taylor Estimator
  - If there is heterogeneity in both intercepts and slopes:
    Random Coefficients Estimator

When do you want to use:

Fixed effects

- If the data exhaust the population
- If data is TSCS
- If you are predominantly interested in variation over time
- If there is little variation between units
- If you believe there are omitted variables you need to control for
- If there is correlation of intercepts with explanatory variables
- If your variables are time variant
- If you are not concerned with variance of the estimates

Random effects

- If your data is a sampled
- If data CSTS
- If you are also interested in variation over cross-sections
- If there is little variation within units
- If your model is properly specified with all relevant variables included
- If there is no correlation of intercepts (now in error term) with explanatory variables
- If your variables are time invariant
- If you are not concerned with bias of the estimates

BEWARE: this is not a definitive guide!
Use tests, compare estimators, think…

What will we do today?

- We will discuss the modeling of dynamics in TSCS/CSTS setting

- We will discuss extensions from continuous variables to other types of dependent variables (hopefully…)

Building blocks of stochastic models

- Now we are going to address some of the most important concepts of **<u>time series</u>**, in particular:
  - White noise
  - Stationarity
  - Random walk
  - Random walk with drift
  - Deterministic trend

White noise

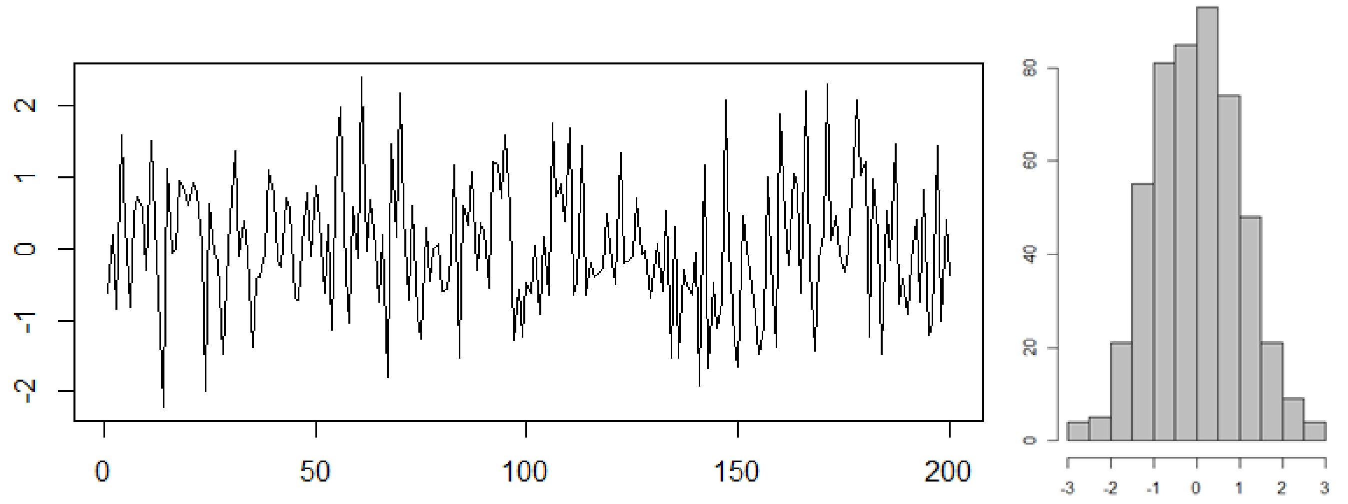
- If our model encapsulates most of the deterministic features of the time series, our residual error series should appear to be a realization of independent random variables from some probability distribution. We build models up from a model of independent random variation which is called **discrete white noise**.

- White noise variable w_t has to satisfy :

- $E[w_t] = 0$ mean of zero
- $Var[w_t] = \sigma^2$ constant, finite variance
- $Cov[w_t, w_{t+k}] = 0 \ \forall k \neq 0$ no correlation with past of future values

- Gaussian white noise has additional attribute of being normally distributed

White noise



- $y_t = \mu + w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} \ldots$ general linear process
- General linear model converts white noise to autocorrelated series where dynamic structure of series is completely determined by ψ.
- μ is constant, but we are typically not concerned with this parameter, as it does not affects dynamics (time-dependent aspects of the model)
- White noise series is **stationary**.

Stationarity

- Stationarity means that the statistical properties of the process do not change over time.

- Two main concepts are:

➢ **Strictly stationary** – if the joint distribution of random variables of the stochastic process is independent of time. As probability distribution of time series does not depend on t, mean, variance and all higher moments are independent of t.

➢ **Weakly stationary (covariance stationary)** – if mean is constant, variance is finite and the relation between y_t and, say, y_{t-j} depends on distance between them in time, but not on their absolute location on the time line.

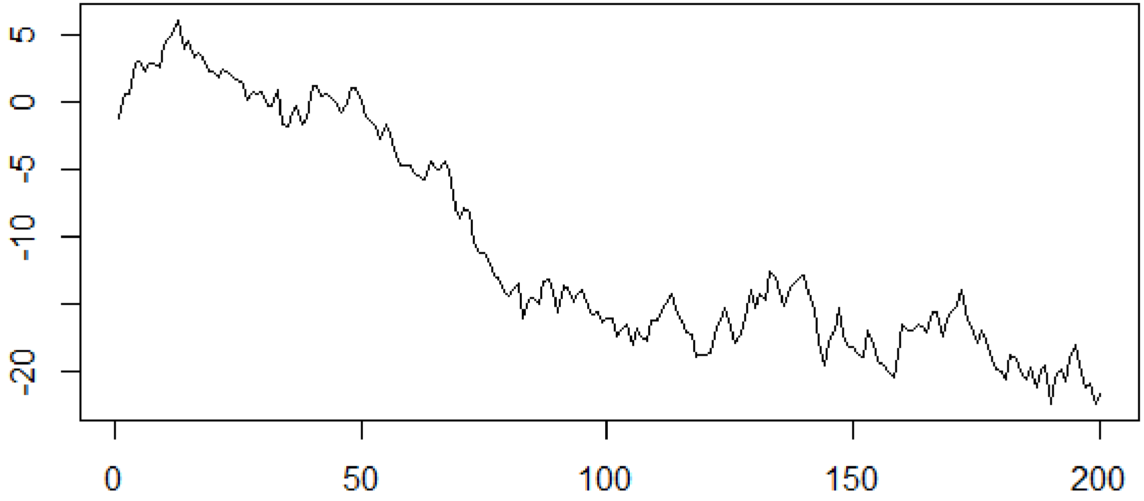
  – we typically require weakly stationary variables

Random walk (i.e. stochastic trend)

- Fundamental non-stationary model based on discrete white noise is called the **random walk**
- The series y_t is a random walk if:

$$\boldsymbol{y_t = y_{t-1} + w_t}$$

- where $\boldsymbol{w_t}$ is a white noise series. Thus, effect of $\boldsymbol{y_{t-1}}$ persists in $\boldsymbol{y_t}$



Random walk

$$y_t = y_{t-1} + w_t$$

$$y_t = y_{t-2} + w_{t-1} + w_t$$
$$y_t = y_{t-3} + w_{t-2} + w_{t-1} + w_t$$
$$y_t = y_0 + \sum_{i}^{t-1} w_i$$

- The random walk is constant in mean. The variance increases without limit as t increases. The covariance is a function of time. So, the process is non-stationary

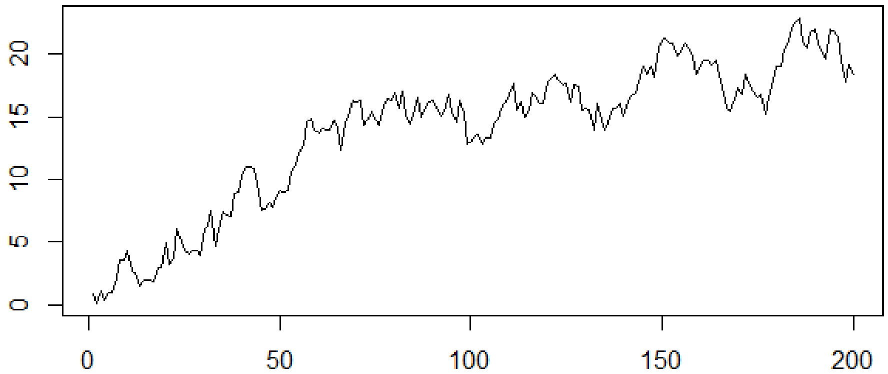
$$E[y_t] = 0$$
$$Cov[y_t, y_{t+k}] = \gamma_k(t)$$
$$Var[y_t] = t\sigma^2$$

- Random walk model is prototype of the **unit root process**.

Random walk with drift

- Sometimes the value of series is expected to increase despite volatility. The random walk model can be adapted to allow for this by including a drift parameter δ. Random walk with drift is defined as *:*

$$y_t = \delta + y_{t-1} + w_t$$
$$y_t = \delta + \delta + y_{t-2} + w_{t-1} + w_t$$
$$y_t = \delta + \delta + \delta + y_{t-3} + w_{t-2} + w_{t-1} + w_t$$
$$y_t = \delta t + y_0 + \sum_{i}^{t-1} w_i$$

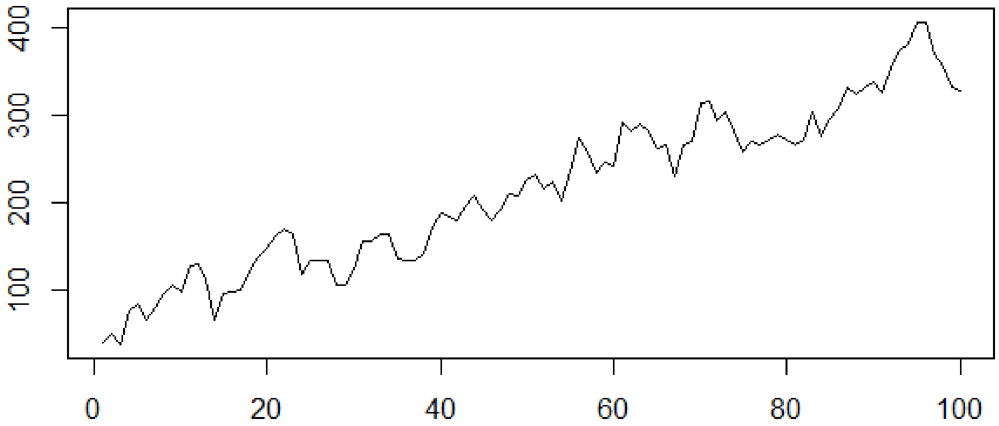


- With drift not only that the series is not stationary in the variance but it is not stationary in the mean.

$$E[y_t] = \delta t$$
$$Var[y_t] = t\sigma^2$$

Deterministic trend

$$y_t = \beta t + w_t$$



- Series is stationary in variance however the series is not stationary in mean. So the series is not statationary

$$E[y_t] = \beta t$$
$$Var[y_t] = \sigma^2$$

- The difference between deterministic trend and random walk with drift is that later is not stationary in variance. Even if they have same error term, deterministic trend will have less variation around the trend line in comparison to random walk with drift.

Stationary vs. non-stationary data

- Stationarity is an assumption underlying many statistical procedures used in time series analysis. Non-stationary series can be transformed to stationary series in two ways:

- **Difference stationary:** The mean trend is stochastic. Differencing adjacent terms of a series can transform a non-stationary series to a stationary series.
  - Differencing is defined as:

$$\Delta y_t = y_t - y_{t-1}$$

  - As the differenced series needs to be aggregated (or 'integrated') to recover the original series, the underlying stochastic process is called autoregressive integrated moving average - **ARIMA**

- **Trend stationary:** The mean trend is deterministic. Once the trend is estimated and removed from the data, the residual series is a stationary stochastic process.
  - Deterministic trend is most easily subtracted from a series by estimating

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

  - You can also address this non-parametrically using filters (e.g. moving average)

ARMA/ARIMA

- A time series can be explained either by its history or by contemporaneous and past shocks. This is the goal of ARIMA - autoregressive integrated moving average model

- This method models both **stationary (AR, MA, ARMA**) and **non-stationary (ARIMA)** univariate time series

- ARIMA modeling is about modeling the correlation patterns observed in the data.

- *The goal:* ***to reduce a variable to "white noise"*** *- a stochastic series. ARMA/ARIMA is designed to identify and eliminate systematic sources of error*

(Main citation: George E.P. Box; Gwilym Jenkins (1970) Time Series Analysis: Forecasting and Control)

AR (autoregressive) process

- Linear combination of lagged observed variable is AR component – AR(p), ARMA(p,0) or ARIMA(p,0,0)

- A simple autoregressive model to capture a significant lag-1 autocorrelation is:

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

where is μ constant, w_t is withe noise error term, φ is autoregressive coefficient and $\varphi < 1$

- An AR process describes geometric (exponential) decay in the effects of shocks. As such some portion lasts forever. However, practically the impact can safely be treated as zero after few periods

- The series is an autoregressive process of order p, abbreviated to AR(p), if

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + w_t$$

MA (moving average) process

- Linear combination of unobservable white noise disturbances is MA process - MA(q), ARMA(0,q) or ARIMA(0,0,q)

- A simple moving average model MA(1) is defined as:

$$y_t = \mu + w_t + \vartheta w_{t-1}$$

where is μ constant, w_t is withe noise error term,ϑ is moving average coefficient

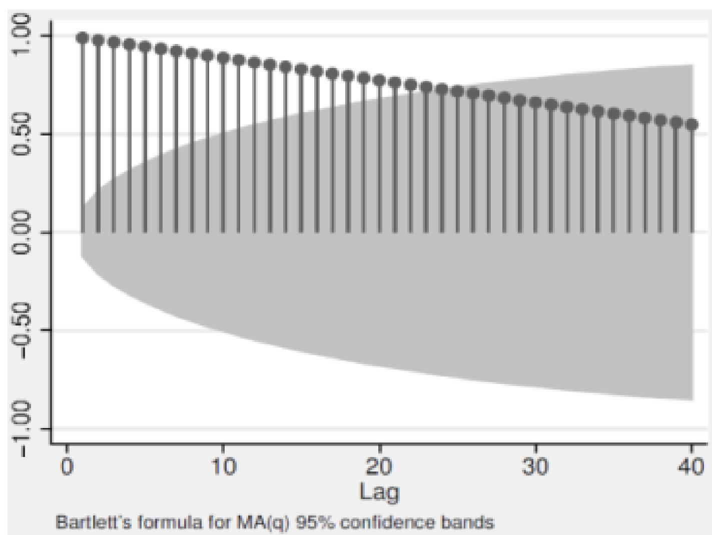
- Moving average of order q is defined as :

$$y_t = w_t + \ \vartheta_1 w_{t-1} + \vartheta_2 w_{t-2} + \cdots + \vartheta_q w_{t-q}$$

- Unlike AR process where autocorrelation function has a long tail, MA process rapidly decays once you pass the history.

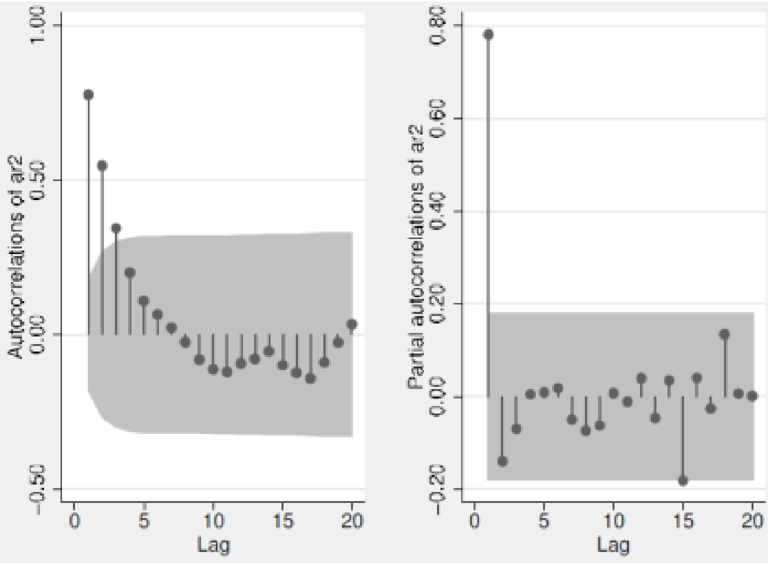
Identify order of differencing (integration)

- To determine order of differencing, ***d*** in ARIMA(p,d,q), look at autocorrelation function
- If autocorrelations collapse quickly toward zero, the time series is stationary
- If the magnitude of correlations declines approximately linearly at least one differencing is needed



Identifying order of AR(p)

AR(1)

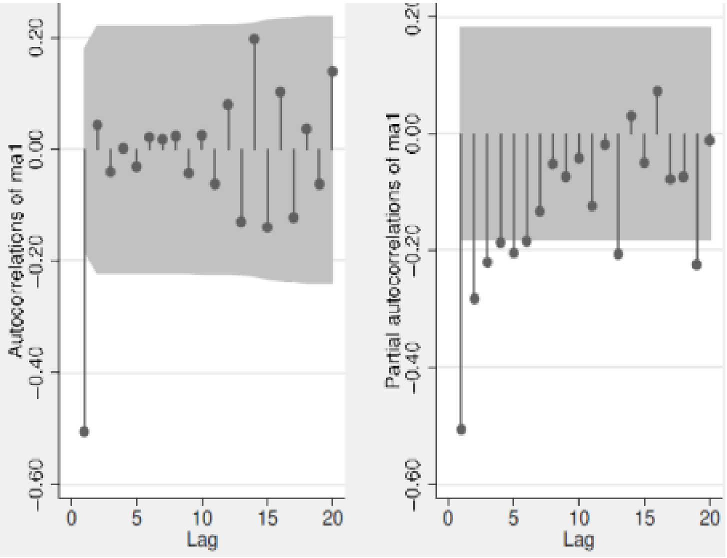


- In pure AR process **acf** dies out either exponentially (if roots are real), in damped oscillations (if roots are complex) or in combination of the two (if roots are both real and complex).

- In pure AR process **pacf** should lie outside confidence bands for first p lags of AR(p)

Identifying order of MA(q)

MA(1)



- In pure MA process **acf** should lie outside confidence bands for first q lags of MA(q).

- In pure MA process **pacf** dies out in some combination of exponentially decay and damped oscillations.

Order of AR(p) and MA(q) in ARIMA

- **The approach to identifying order of AR(p) and MA(q) can be both top-down and bottom-up.**
  - Top down would entail assessing reasonably large number of (p,q) parameters using **acf** and **pacf** graphs and successively pruning parameters on the basis of significance levels.
  - Bottom up approach would entail successively adding parameters until optimal ARIMA model is achieved.
- **The Box Jenkins emphasis on parsimony may cause tendency to fit models that are too simple**
- **Overfitting:**
  - The way to guard against too simple models it is overfitting - adding parameters to the model. AR and MA parameters should not be added simultaneously because of the risk of parameter redundancy. If the parameters are insignificant, more parsimonious specification is adequate.

Information criteria

- However, statistical significance may not be the best way to select the model.

- The modern approach to assessing the order of AR and MA orders in ARIMA is based on information criteria. Information criteria penalize models with too many parameters. Various criteria may be employed (and unfortunately they may produce different results). Some of criteria are defined bellow:

$$Akaike(AIC) = \ln(\hat{\sigma}^2) + \frac{2(p+q)}{t}$$

$$Bayesian\ (BIC) = \ln(\hat{\sigma}^2) + \frac{\ln(t) \times (p+q)}{t}$$

where $\hat{\sigma}^2$ signifies the estimated variance of an ARMA(p, q)-process

Unit root tests - Dickey-Fuller test

- The modern approach to identification of unit roots is based on testing

- Dickey and Fuller [1979] proposed the following test regression that is delineated from an assumed AR(1)-process of y_t:

$$y_t = \varphi y_{t-1} + w_t$$

- Logic: If we subtract y_{t-1} from both sides of the equation, we get:

$$y_t - y_{t-1} = \varphi y_{t-1} - y_{t-1} + w_t$$
$$\Delta y_{t-1} = (\varphi - 1) y_{t-1} + w_t$$
$$\Delta y_{t-1} = \pi y_{t-1} + w_t$$

under H_0 = process contains unit root $\pi = 0$ (that is, $\varphi = 1$)

- This test statistic does not have the familiar Student t distribution, so critical values have been calculated by simulation.

Augmented Dickey-Fuller (ADF) test

- **Augmented Dickey-Fuller** includes lagged differences to soak up the serial correlation of order k . The extended test regression is also based on addition of β_1 (drift) and $\beta_2 t$ (deterministic trend) terms.

$$\Delta y_t = \beta_1 + \beta_2 t + \pi y_{t-1} + \sum_{j=1}^{k} \gamma_j \, \Delta y_{t-j} + \varepsilon_t$$

- **Phillips-Perron test** - allows for weak dependence and heterogeneity of the error process

- There other, more advanced tests such as:, Elliott-Rothenberg-Stock test, Schmidt-Phillips test, Kwiatkowski-Phillips-Schmidt-Shin test, etc. We will not discuss these.

Test of residuals

- Measure of well specified and accurately fitted time series model is evidence that the residual w_t is white noise.
  - Q test - portmanteau test for white noise
  - Bartlett’s test of nonrandom periodicity

Dynamic panel data models

- We will now consider a **dynamic panel data model**. A simple **static** model can be expressed as
$$y_{it} = a_i + \beta x_{it} + \varepsilon_{it}$$
- It implies that any changes are felt instantaneously with no delayed effect

- There are several ways to add dynamics to the static specification assuming that ***T* is large enough**

  – How large T should be? – there is no strict limit (maybe number of repeated observations should be $T \geq 10$); Beck and Katz use minimum of $T = 15$ in their simulations

- In principle, if T is small, we can't very effectively assess time-serial processes.

The dynamic panel bias (Hurwicz/Nickell bias)

- Additional reason for big T is dynamic panel bias or Nickell's bias (1981) (Hurwicz/Nickell bias) - Nickell's bias is the bias of the fixed effect estimator in a dynamic model. Consider a simple dynamic model:

$$y_{it} = a_i + \beta y_{it-1} + \varepsilon_{it}$$

- Fixed effect model is consistent for the static model whether the effects are fixed or random. However, fixed effect is inconsistent for a dynamic panel data model. More precisely:
  - If T is fixed, then a_i and β are ***inconsistent*** estimators no matter how large N is
  - However, if T tends to infinity, the fixed effect estimator of a_i and β are ***consistent***.

- **The bias of $\boldsymbol{\beta}$ is caused by having to eliminate the unknown unit effects $\boldsymbol{a_i}$ from each observation, which creates a correlation of order $(1/T)$ between the lagged dependent and other explanatory variables, and the residuals in the transformed model**

The dynamic panel bias

- Assuming a dynamic model:

$$y_{it} = a_i + \beta y_{it-1} + \varepsilon_{it}$$

- The transformed model would be:

$$y_{it} - \bar{y}_i = \beta\left(y_{it-1} - \bar{y}_{i,-1}\right) + (\varepsilon_{it} - \bar{\varepsilon_i})$$

- **For small $\boldsymbol{T}$, this bias is always negative if $\boldsymbol{\beta} > 0$ and the bias does not go to zero as $\boldsymbol{\beta}$ goes to zero.**
  - For instance, when $T = 2$, the asymptotic bias is equal to $-(1 + \beta)/2$, and when $T = 3$, it is equal to $-(2 + \beta)(1 + \beta)/2$. Even with $T = 10$ and $\beta = 0.5$, the asymptotic bias is -0.167.

- The dynamic fixed effects model remains biased with the introduction of exogenous variables if T is small, so if assume:

$$y_{it} = a_i + \beta_1 y_{it-1} + \beta_2 x_{it} + \varepsilon_{it}$$

in this case, both estimators $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are biased.

- **Because of this bias most panel data models do not include lagged dependent**. However, the bias disappears as T increases.

Modeling dynamics

- Let's begin with stationary data (assuming sufficient T – i.e. TSCS data). The main issue is autocorrelated errors. We have already discussed some forms of dealing with autocorrelation.
  - **Panel corrected standard errors**
  - **(Feasible) generalized least squares**
  - **Clustered standard errors**

- In this setting we were dealing with dynamic aspects of data as nuisance

- However, we can explicitly model dynamics. Dynamic panel data models contain (at least) one lagged variable. We can distinguish between several basic models:

  - $y_{it} = \beta_1 x_{it} + \beta_2 x_{it-1} + \nu_{it}$ **finite distributed lag model**
  - $y_{it} = \beta x_{it} + \phi y_{it-1} + \nu_{it}$ **lagged dependent variable model**
  - $y_{it} = \beta x_{it} + \theta y_{it-1} + \gamma x_{it-1} + \nu_{it}$ **autoregressive distributed lag model**

Dynamics

- The most important issue is whether we think a change in some variable is felt only immediately or whether its impact is distributed over time. Each specification assumes specific dynamic structure:
  - The static specification assumes that all variables have an instantaneous and only an instantaneous impact.
  - Finite distributed lag model, assumes that the impact of x sets in over few periods but then dissipates completely, (e.g. if there is one lag, it takes two periods for the full effect of the change, but the effect dissipates after two periods)
  - The lagged dependent variable model assumes that x and the $error$ (equivalent to first-order moving average error process) have a declining geometric form
  - The autoregressive distributed lag, allows for maximum effect of x to happen with a delay as finite distributed lag, however lagged dependent variable assumes that both x and the $error$ have a declining geometric form

(The feasible generalized least squares models assumes x has only instantaneous effect, but the error decline in geometric form)

Differences between models

- As regards generalized least square and lagged dependent variable, the difference matters little when autocorrelation is small, as there is little difference between the immediate and long-run impact of X. As autocorrelation increases, the differences widen.

- Estimates using lagged dependent variable will pick up the effects of both cross-sectional and time-serial variation in Y, i.e., not just dynamics.

- Using lagged dependent variable one loses N cases, one precious observation for each (short) time series.

- Keep in mind that, models with lagged dependent variable are acceptable only if the resulting residuals are not serially correlated.

Discriminating Between Models

- We can use the fact that the autoregressive distributed lag model nests the lagged dependent variable and generalized least square models to test appropriate model. Distributed lag is specified as:

$$y_{it} = \beta x_{it} + \theta y_{it-1} + \gamma x_{it-1} + \nu_{it}$$

  - The lagged dependent variable model is specified as $y_{it} = \beta x_{it} + \theta y_{it-1} + \nu_{it}$ implying that $\gamma = 0$
  - On the other hand, the generalized least square model implies that $\gamma = -\theta\beta$

- Thus, we can estimate the full ADL model and test whether $\gamma = 0$ or $\gamma = -\theta\beta$. If both simplifications are rejected, we can retain the more complicated autoregressive distributed lag model.

- Even in the absence of a precise test, the autoregressive distributed lag model estimates will often indicate which simplification is not too costly to impose.

Modeling nonstationary data

- What can we do with nonstationary data?

- One approach is to induce stationarity by first (or higher order) differencing: $\Delta y_t = y_t - y_{t-1}$
  - This is the basis of ARIMA approach. Therefore, we would explain changes in y by changes in x, (which is likely to result in statistically insignificant results).
  - From more substantive point of view, objection may be that modeling in the first differences also throws out any long-run information about y and x

- We can also use error correction model:

$$\Delta y_{it} = \beta \Delta x_{it} - \lambda(y_{it-1} - kx_{it-1}) + \nu_{it}$$

Diagnostics

- Up to now we were modeling dynamics by including appropriate lagged values of the dependent variable and current and lagged values of the x's so that the resulting errors appear to be serially independent.

- Consequently, we must test if residuals are violating this assumption. Thus, we need to test for normality and for lack of autocorrelation.

- For this we will have to use methods and techniques discussed earlier.

The dynamic panel models

- What should we do in case of the panel (CSTS) data?

- We are interested in estimating the parameters of a model of the form:

$$y_{it} = \alpha_i + \gamma y_{it-1} + \beta x_{it} + \varepsilon_{it}$$

for $i = \{1, \ldots, N\}$ and $t = \{1, \ldots, T\}$ using datasets with large N and fixed T

- We have already discussed that in the estimating fixed effects model with lagged dependent variable will result in Nickell's bias. On the other hand, by construction, y_{it-1} is correlated with the unobserved individual-level effect α_i, so we can not use random effects.

- So we are left with first difference estimator:

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

- After differencing α_i are gone. However, the y_{it-1} in Δy_{it-1} is a function of the ε_{it-1} which is also in $\Delta\varepsilon_{it}$. So Δy_{it-1} is correlated with $\Delta\varepsilon_{it}$ by construction. So, we still have a problem of endogeneity.

Solutions

- What are the solutions?
- Consistent estimator of γ can be obtained by using:

  - IV approach (Anderson and Hsiao, 1982)
  - GMM approach (Arenallo and Bond, 1985)

- Let us start with Anderson's and Hsiao's suggestions:

  1. Anderson and Hsiao(1981) suggested a 2SLS estimator based on further lags of Δy_{it} as instruments for Δy_{it-1} **(for instance: $\boldsymbol{\Delta y_{it-2}}$)**
  2. Anderson and Hsiao(1981) also suggested a 2SLS estimator based on lagged <u>levels</u> of Δy_{it} as instruments for Δy_{it-1} **(for instance: $\boldsymbol{y_{it-2}}$)**

Instrumental variable approach

- The idea behind instrumental variables is to find a set of variables, instruments, that are both (1) correlated with the explanatory variables in the equation, and (2) uncorrelated with the disturbances. So, assume we have following expression:

$$y_{it} = \gamma y_{it-1} + \beta x_{it} + \rho z_i + \upsilon_i + \varepsilon_{it}$$

where x_{it} a vector of time-varying explanatory variables, z_i is a vector of time-invariant explanatory variables

  - **First step:** Take the first difference. The first difference of the model will swipe out the individual effects υ_i and the time invariant variables z_i:

$$\Delta y_{it} = \gamma \Delta y_{it-1} + \beta \Delta x_{it} + \Delta \varepsilon_{it}$$

  - **Second step:** Because y_{it-2} or $y_{it-2} - y_{it-3} = \Delta y_{it-2}$ are correlated with $y_{it-1} - y_{it-2} = \Delta y_{it-1}$ but are uncorrelated with $\varepsilon_{it} - \varepsilon_{it-1} = \Delta \varepsilon_{it}$ they can be used as an instrument for Δy_{it-1}. So, regress endogenous variable in the equation on all exogenous variables. Thus, we can estimate $\hat{\gamma}_{IV}$ and $\hat{\beta}_{IV}$
  - **Third step:** given the estimates $\hat{\gamma}_{IV}$ and $\hat{\beta}_{IV}$, we can deduce an estimate of ρ (the vector of parameters for the time invariant variables)

$$\bar{y}_i - \hat{\gamma}_{IV} \bar{y}_{i-1} - \hat{\beta}_{IV} \bar{x}_i = \rho z_i + \nu_i$$

  - **Step four:** given $\hat{\gamma}_{IV}$, $\hat{\beta}_{IV}$, and $\hat{\rho}$, we can estimate the variances σ_ε^2 and σ_υ^2

- The instrumental-variable estimators of γ, β , and σ_ε^2 are consistent when N or T or both tend to infinity. The estimators of ρ and σ_ε^2 are consistent only when N goes to infinity. They are inconsistent if N is fixed and T tends to infinity.

Generalized method of moments (GMM) Arellano-Bond estimation

- However, note that y_{it-2} or Δy_{it-2} are not the only instrument for Δy_{it-1}. In fact, Arellano and Bond (1991) noted that all y_{it-2-j}, j = 0, 1, .., satisfy the condition.

- Estimation is based on generalized method of moments (GMM). GMM is estimation method where the full shape of the distribution function of the data is not known, and therefore maximum likelihood estimation is not applicable The GMM estimation of dynamic panel model is also based on the first difference of the model (to swipe out the individual effects α_i and the time invariant variables z_i)

- Arellano and Bond(1991) showed how to construct estimators based on moment equations constructed from:
  1. Moment conditions using <u>lagged levels</u> of the dependent variable
  2. <u>First-differences</u> of strictly exogenous covariates

- As each exogenous variable contributes one instrument and the remaining instruments come from the $T - 2$ instruments available in periods $T = 3, 4, 5, 6, 7, 8, \ldots$ we have more instruments than parameters

GMM
Arellano-Bover/Blundell-Bond estimator

- Arellano and Bover (1995) and Blundell and Bond (1998) found that if the autoregressive process is too persistent, then the lagged-levels are weak instruments.

- Furthermore, if the ratio of the variance of the panel-level effect to the variance of unit error is too large, the Arellano and Bond estimator can perform poorly.

- Consequently, these authors proposed using additional moment conditions in which lagged differences of the dependent variable are orthogonal to levels of the disturbances. So, essentially we are extending the list of instruments.