Test Power

Lecture 4

Piotr Dybka

Applied econometrics

What is stationarity? – a reminder

- In the strict sense, stationarity means that for the stochastic process y_t the joint probability distribution of $\{y_t, y_{t+1}, \ldots, y_T\}$ is the same as that of any shifted set of observations $\{y_{t+m}, y_{t+m+1}, \ldots, y_{T+m}\}$.
- The problem is that stationarity defined this way (*strict* stationarity) cannot be verified from observed data. This is why, empirically, we test for *weak* stationarity, which holds when the expected value, variance, and covariance of the process are constant over time:

$$\mu_t = \mu$$

$$\sigma_t^2 = \sigma^2$$

$$cov(y_t, y_{t+s})_t = cov(y_t, y_{t+s}) \quad \text{for } s = 1, 2, \ldots$$

The order of integration of a time series

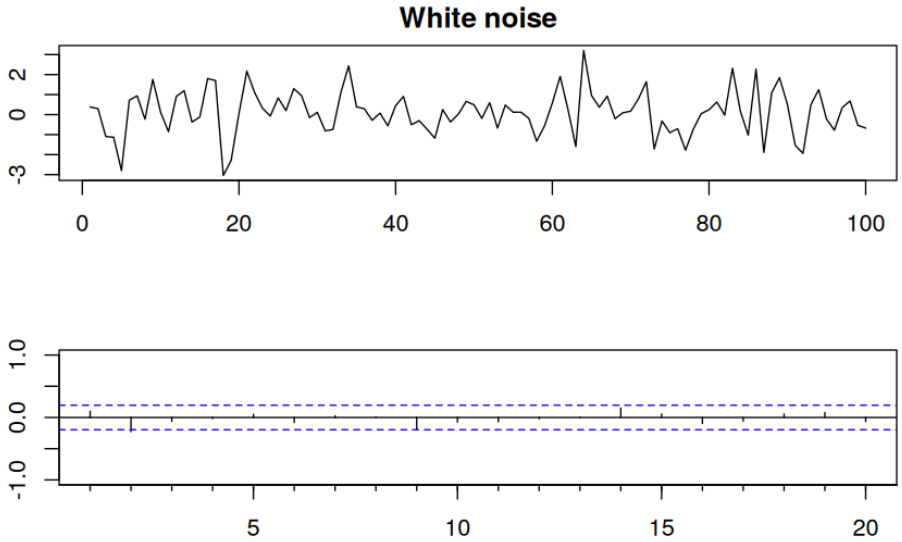
The order of integration is denoted $I(K)$. To determine the order of integration of a series:

1. We test y_t for stationarity. If the series is stationary, we have $y_t \sim I(0)$. If it is non-stationary, we analyse the stationarity of Δy_t, and so on.
2. If for $k = 1, 2, \ldots, K-1$ the variable $\Delta^k y_t$ is non-stationary, while $\Delta^K y_t$ is stationary, then the series y_t is non-stationary, integrated of order K, which we write as $y_t \sim I(K)$.

White noise

White noise (WN):

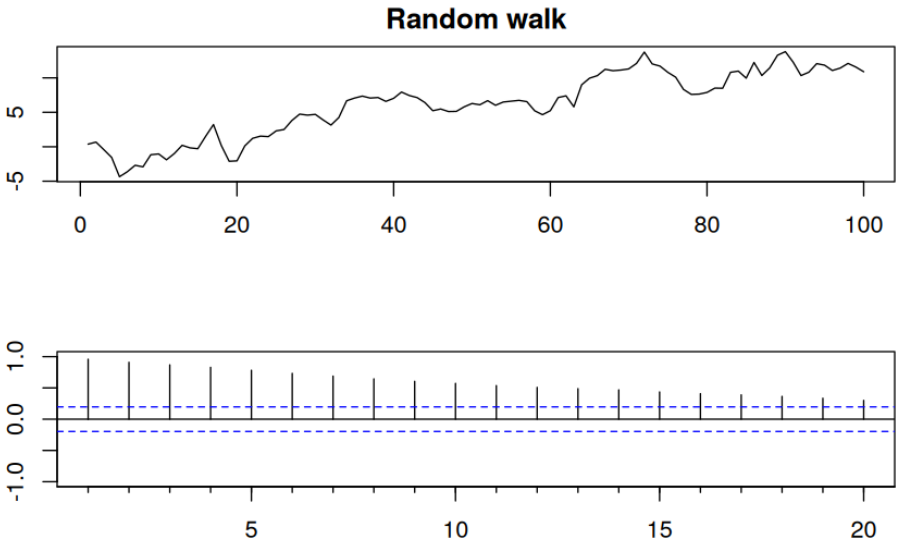
- For any $s \neq t$, the observations y_t and y_s are independent of one another, have the same probability distribution, and their expected values equal 0.
- Example: repeated draws from a normal distribution.
- White noise is stationary.



The definition above is *strong* (i.i.d.) white noise. In practice, a *weak* definition is often enough: $E(y_t) = 0$, $Var(y_t) = \sigma^2$, $cov(y_t, y_s) = 0$ for $s \neq t$ – without requiring full independence or identical distributions.

Random walk

$$y_t = y_{t-1} + \epsilon_t, \qquad \epsilon_t \sim WN$$



- The random walk process is non-stationary:

$$y_t = \sum_{s=1}^{t} \epsilon_s + y_0$$

- Hence

$$E(y_t) = y_0$$

$$Var(y_t) = t\sigma^2$$

$$Cov(y_t, y_{t+s}) = t\sigma^2$$

A highly persistent series

- Suppose we have the series

$$y_t = \rho y_{t-1} + \epsilon_t, \qquad \epsilon_t \sim WN,\ \rho \in (-1, 1)$$

- This series is also stationary – it is characterised by reversion to the mean (the expected value).
- However, for large values $0.9 < |\rho| < 1$ it is characterised by **high persistence** (slow reversion to the mean).
- Such a series is stationary, yet formal statistical tests often struggle to correctly identify stationarity – we will see this in concrete numbers later in the lecture.

Half-life

- For the stationary process $(y_t - \mu) = \rho(y_{t-1} - \mu) + \epsilon_t$, i.e. for $\rho \in (-1, 1)$, y_t reverts to its expected value μ.
- The speed of reversion to the mean is described by the parameter $\delta = \rho - 1$.
- For example, for $\rho = 0.9$ (i.e. $\delta = -0.1$) the deviation from the mean decays at a rate of 10% per period.
- The half-life (hl) tells us after how much time half of a variable's deviation from its equilibrium level has died out:

$$hl = \frac{\ln(0.5)}{\ln(\rho)}$$

- For $\rho = 0.9$ the value of hl is about 6.5 periods.

Type I and Type II errors

A reminder from statistics:

- A Type I error occurs when we reject H_0 even though it is true.
- A Type II error occurs when we do not reject H_0 even though it is false.

		True state	
		H_0 true	H_0 false
Decision from sample	Fail to reject H_0	correct decision	Type II error
	Reject H_0	Type I error	correct decision

Note: we denote the probability of a Type II error by β. **Test power is not the same thing as a Type II error** – power is $1 - \beta$, i.e. the probability of correctly rejecting a false H_0. The higher the power, the less often we make a Type II error.

Test power – a formal definition

- **Significance level** $\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$ – we fix this in advance (usually 0.05 or 0.01); it is the probability of a Type I error.
- **Test power** $1 - \beta = P(\text{reject } H_0 \mid H_0 \text{ false})$ – this is *not* something we fix in advance: it depends on the specific parameter value within H_1, on the sample size N, and on the specification of the test itself.
- In the context of stationarity tests: the power of the ADF test is the probability of correctly rejecting H_0 (non-stationarity) when the series is *genuinely* stationary, for a **specific** value of $\rho < 1$.
- Power is **not a single number** for a given test – it is a *function*: it depends on how far H_1 (here: ρ) departs from H_0 (here: $\rho = 1$), on N, and on the specification of the test regression.
- In this lecture we will check the power of the ADF and KPSS tests **empirically**, using Monte Carlo simulation – we will generate many artificial series with known ρ and check how often the test reaches the correct decision.

The Dickey-Fuller test

The most popular stationarity test, with the hypotheses:

$$H_0: \text{ the series is non-stationary} \qquad H_1: \text{ the series is stationary}$$

- In the DF test we assume that y_t is generated by an AR(1) process:

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

- However, estimating the parameter ρ directly from this equation can lead to incorrect conclusions because of spurious regression, so instead the following equation is estimated:

$$\Delta y_t = \delta y_{t-1} + \epsilon_t$$

- Where $\delta = (\rho - 1)$.

$$H_0: \ \delta = 0 \iff \rho = 1 \qquad H_1: \ \delta < 0 \iff \rho < 1$$

The Dickey-Fuller test

- To avoid the risk of autocorrelation in the error term of the test regression (which would bias the estimator), the DF test can be extended to the augmented ADF test:

$$\Delta y_t = \delta y_{t-1} + \sum_{i=1}^{P} \beta_i \, \Delta y_{t-i} + \epsilon_t$$

- The DF (ADF) test statistic is given by

$$DF = \frac{\hat{\delta}}{S_\delta}$$

- This statistic does not follow a standard t-distribution – under H_0 (when $\rho = 1$) its distribution is a functional of a Wiener process and has no simple analytical form. Critical values are obtained by numerical (simulation) methods and tabulated (Dickey, Fuller 1979; MacKinnon 1996).

Long-run variance

For y_t a stationary process, for the statistic $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$ the expected value is $E(\bar{y}) = \mu$, and we can compute the variance $\hat{\gamma}_0$ as:

$$Var(\bar{y}) = \frac{1}{T}\left(\gamma_0 + 2\sum_{s=1}^{T}\left(1 - \frac{s}{T}\right)\gamma_s\right)$$

Where $\gamma_s = cov(y_t, y_{t-s})$. Under zero autocorrelation the formula simplifies to $Var(\bar{y}) = \frac{\gamma_0}{T}$. The long-run variance is the limit of the expression above as $T \to \infty$.

For a finite sample it can be estimated using the Newey-West method:

$$\hat{\gamma}_\infty = \hat{\gamma}_0 + 2\sum_{s=1}^{L}\kappa_{s,L}\hat{\gamma}_s$$

$$\kappa_{s,L} = 1 - \frac{s}{L+1}$$

L is the **maximum number of lags included** (the so-called truncation parameter, or *bandwidth*) – it should not be confused with the lag operator from ARMA notation.

The KPSS test

The hypotheses (note – reversed compared to the previous tests!):

$$H_0: \text{ the series is stationary} \qquad H_1: \text{ the series is non-stationary}$$

- The hypotheses are reversed relative to DF/ADF.
- It uses information about the long-run variance of the error term.

1. We estimate a regression of y_t on a constant (possibly a trend) and save the residuals e_t.
2. We compute the partial sums $S_t = \sum_{i=1}^{t} e_i$ for each period ($t = 1, 2, \ldots, T$).
3. Using the Newey-West method we estimate the long-run variance of the residuals $\hat{\gamma}_\infty$.
4. We compute the test statistic:

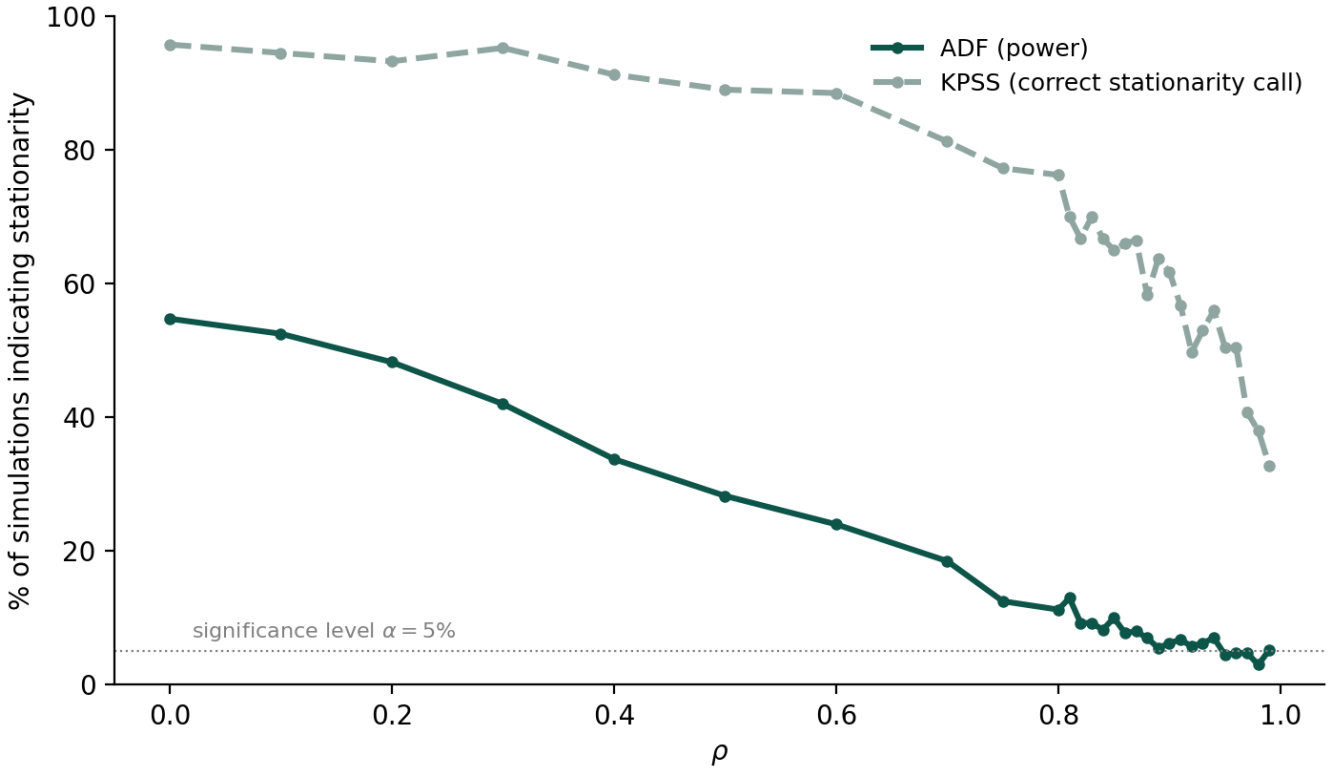
$$KPSS = \frac{1}{T^2} \frac{\sum_{i=1}^{T} S_i^2}{\hat{\gamma}_\infty}$$

Monte Carlo simulation of test power

How do we measure test power in practice, given that it depends on ρ, which is unknown in real data? **Answer**: we simulate data with a known ρ and check how often the test reaches the correct decision.

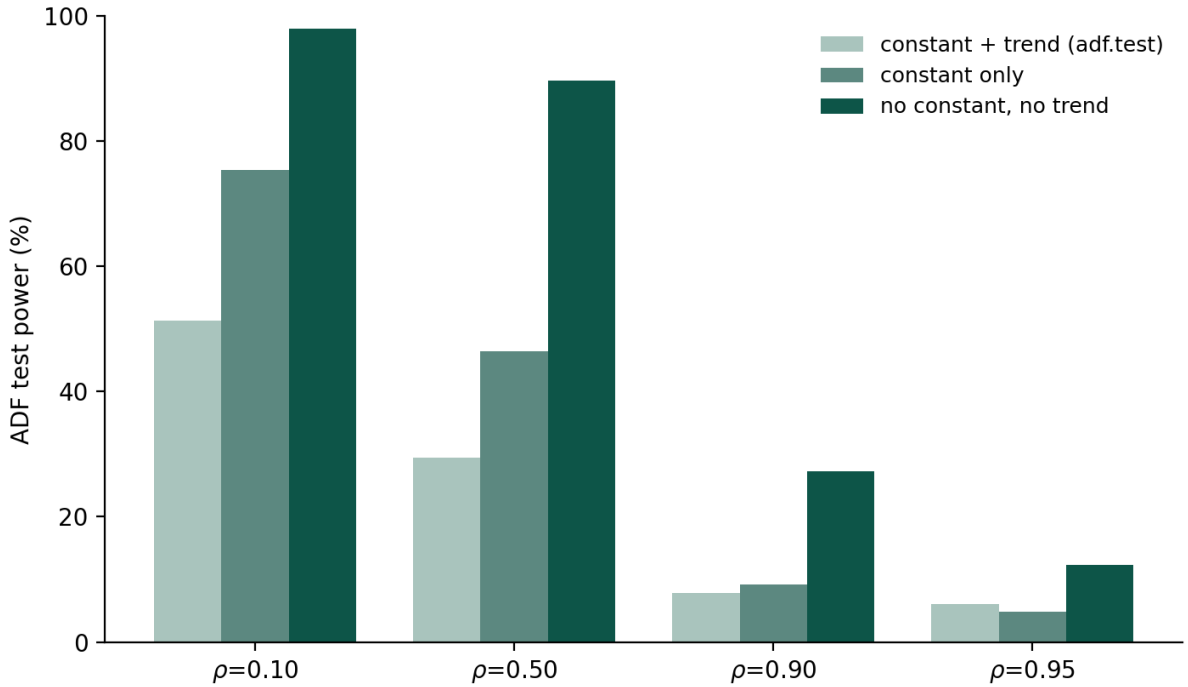
1. We fix ρ (e.g. $\rho = 0.95$) and the number of repetitions M (e.g. $M = 250$).
2. For each repetition $m = 1, \ldots, M$: we generate a series $y_t = \rho y_{t-1} + \epsilon_t$ ($n = 50$ observations) and compute the p-values of the ADF and KPSS tests for that series.
3. The share of repetitions in which ADF rejects H_0 (p-value < 0.05) is the **estimated power** of the ADF test for that ρ and that n.
4. Likewise for KPSS, except here H_0 is true (the process is stationary) – the share of repetitions in which KPSS *does not* reject H_0 tells us about the test's actual size, not its power in the strict sense (we could only speak of KPSS's power by simulating genuinely non-stationary data).
5. We repeat this for a grid of ρ values (e.g. $\rho = 0.90, 0.91, \ldots, 0.99$) to obtain a **power curve**.

Power curve: ADF and KPSS



Simulation results ($n = 50$, $M = 400$ for each ρ). ADF power declines systematically as ρ increases, and as $\rho \to 1$ it converges to the significance level itself ($\alpha = 5\%$) – the test practically **loses the ability to distinguish** a highly persistent stationary process from a non-stationary one. KPSS retains a higher share of correct calls for longer, but it too declines with ρ.

Deterministic specification and test power



`tseries::adf.test()` **always** fits a regression with a constant and a linear trend – this cannot be changed. Our simulated process has neither drift nor trend. Fitting an overly rich specification **sharply reduces power**: for $\rho = 0.1$, power falls from about 98% (no constant or trend) to about 51% (constant and trend) – on the exact same data!

Conclusion: the deterministic specification of a test should match the true nature of the data, not be chosen automatically.

Stationarity tests for panel data

For panel data, testing for stationarity is somewhat harder, because we have both a time and a cross-sectional dimension. In practice there are many different tests, with the hypotheses:

H_0 : the panel series are non-stationary $\qquad$ H_1 : the panel series are stationary

The idea behind these tests is to run separate tests (e.g. ADF) for each panel unit and average the result (e.g. Im, Pesaran, Shin (2003)), or to pool the data and compute the Levin, Lin, Chu (2002) test statistic:

1. We run the ADF test for each panel unit.
2. We save the residuals from the auxiliary ADF models.
3. We compute the ratio of long-run to short-run variance for each panel unit.
4. We compute the test statistic.

The Im, Pesaran, Shin (2003) test

The same hypotheses as for the Levin, Lin, Chu (2002) test:

H_0 : the panel series are non-stationary $\quad$ H_1 : the panel series are stationary

- The **key difference** from LLC is not just a matter of how results are averaged – it concerns the **form of the alternative hypothesis**:
  - **LLC**: assumes a **common** parameter ρ for all panel units. H_1: *all* countries have a stationary series (with the same ρ) – a restrictive assumption that is often unrealistic for heterogeneous panels.
  - **IPS**: allows for **different** ρ_i across units. H_1: *at least some* countries have a stationary series. The test statistic is a standardised average of the individual ADF statistics ($\bar{t}$), which still requires homogeneity of the alternative only in a limited sense.
- Consequence: when the panel is heterogeneous (some countries stationary, some not), LLC and IPS can give **conflicting conclusions** – we will see this in an example.

Panel tests – an example

A panel of 28 EU countries (including the United Kingdom, pre-Brexit), 1995–2017, six macroeconomic variables (`pmax=4`, `exo="intercept"`):

Variable	LLC (stat.)	LLC (p)	IPS (stat.)	IPS (p)
PAFP	−6.28	< 0.001	−3.59	< 0.001
SLE	−7.26	< 0.001	−1.65	0.049
X1	−2.86	0.002	0.63	0.734
FDIYS	−4.39	< 0.001	−1.73	0.042
GOVC1	−6.80	< 0.001	−3.77	< 0.001
POT	1.92	0.973	−0.51	0.305

- For most variables the two tests **agree**.
- For **X1** the tests give **conflicting conclusions**: LLC rejects H_0 (p=0.002, indicating stationarity), IPS does not reject (p=0.734, indicating non-stationarity). This is exactly the situation from the previous slide: if the panel is a mix of stationary and non-stationary countries, the assumption of a common ρ (LLC) can lead to rejecting H_0 even though some countries genuinely have a unit root.