POLS 6481. Research Design and Quantitative Methods II
Lecture 13. Time Series Data I: Stationarity, Autoregression, and Unit Roots
Wooldridge, *Introductory Econometrics*, 1.3 + 10.1—10.4 + 11.1 + 11.3 + 18.2

Outline

1. What is Time Series Data
2. Static Time Series Regression Models
3. Example of a Stationary Series: *Wine Prices*
4. Example of a Non-Stationary Series: *Manatee Deaths*
5. Defining Stationary and Non-Stationary Series
6. The Run-Test
7. Auto-regression and Unit Roots
8. The Dickey-Fuller Test for a Unit Root

POLS 6481. Research Design and Quantitative Methods II
Lecture 13. Time Series Data I: Stationarity, Autoregression, and Unit Roots
Wooldridge, *Introductory Econometrics*, 1.3 + 10.1 + 10.2a

1. What is Time Series Data

"observations on a variable or several variables over time" (p. 8)

Assumption TS.1: The stochastic process $\{(x_{t1},..., x_{tk}, y_t), t = 1,..., T\}$ obeys a linear relationship

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t$$

where $\{u_t: t = 1, ..., T\}$ is the sequence of disturbances.

"common frequencies are daily, weekly, monthly, quarterly, and annually… Many weekly, monthly, and quarterly economic time series display a strong **seasonal pattern**, which can be an important factor in a time series analysis" (p. 8)

Unlike cross-sectional data, which are organized arbitrarily, time series data have a **temporal ordering**; you cannot simply rearrange observations within a dataset.

"Because past events can influence future events … time is an important dimension in time series data. … A key feature of time series data that makes them more difficult to analyze is that economic **observations can rarely be assumed to be independent across time**…

For example, knowing something about the GDP from last quarter tells us quite a bit about the likely range of the GDP during this quarter, because GDP tends to remain fairly stable from one quarter to the next." (p. 8)

POLS 6481. Research Design and Quantitative Methods II
Lecture 13. Time Series Data I: Stationarity, Autoregression, and Unit Roots
Wooldridge, *Introductory Econometrics*, 10.2a + 10.3

2. The Static Time Series Regression Model

* Emphasis on modeling the dependency of a variable on current values of other variables

Examples:

The Static Phillips Curve (current inflation depends on current unemployment)

$$inf_t = \beta_0 + \beta_1 unem_t + u_t$$

The Static Economic Model of Crime (also depends on current unemployment, etc.)

$$mrdrte_t = \beta_0 + \beta_1 convrte_t + \beta_2 unem_t + \beta_3 yngmle_t + u_t$$

A static model of manatee deaths due to boating accidents (depends on number of boats):

$$kills_t = \beta_0 + \beta_1 boats_t + \ldots$$

A static model of wine vintage prices (with a trend due to age, depends on warmth):

$$price_t = \beta_0 + \beta_1 age_t + \beta_2 temperature_t \ldots$$

POLS 6481. Research Design and Quantitative Methods II
Lecture 13. Time Series Data I: Stationarity, Autoregression, and Unit Roots
Wooldridge, *Introductory Econometrics*, 10.2a + 10.3

2. The Static Time Series Regression Model

* Emphasis on modeling the dependency of a variable on current values of other variables

Assumptions:

Assumption TS.1: Linearity (see prior slide)

Assumption TS.2: No perfect collinearity

Assumption TS.3: Exogeneity (at same period) $E(u_t|\mathbf{x}_t) = 0$

or Strict Exogeneity (at all periods) $E(u_t|\mathbf{X}) = 0$

Assumption TS.4: Homoskedasticity $Var(u_t|\mathbf{X}) = Var(u_t) = \sigma^2$

Assumption TS.5: No serial correlation $Corr(u_t, u_s|\mathbf{X}) = 0, \ t \neq s$

Under these assumptions, estimates of the coefficients, sampling variances, and the error variance are unbiased (Theorems 10.1, 10.2, and 10.3) and best (Theorem 10.4)

However, these assumptions (primarily TS.3 and TS.5) are frequently violated. Strict exogeneity rules out feedback from values of *y* to future values of *x*'s

POLS 6481. Research Design and Quantitative Methods II
Lecture 13. Time Series Data I: Stationarity, Autoregression, and Unit Roots
Fair, *Predicting Presidential Elections and Other Things*, ch. 6

3. Example of a Stationary Series: *Wine Prices*

Data: wine.dta
Script: Lecture 13 wine.R

N = 27 years from 1952 to 1980; missing 1954 and 1956 ("by 1990, the 1954 and 1956 vintages were no longer traded much; they were not very good.")

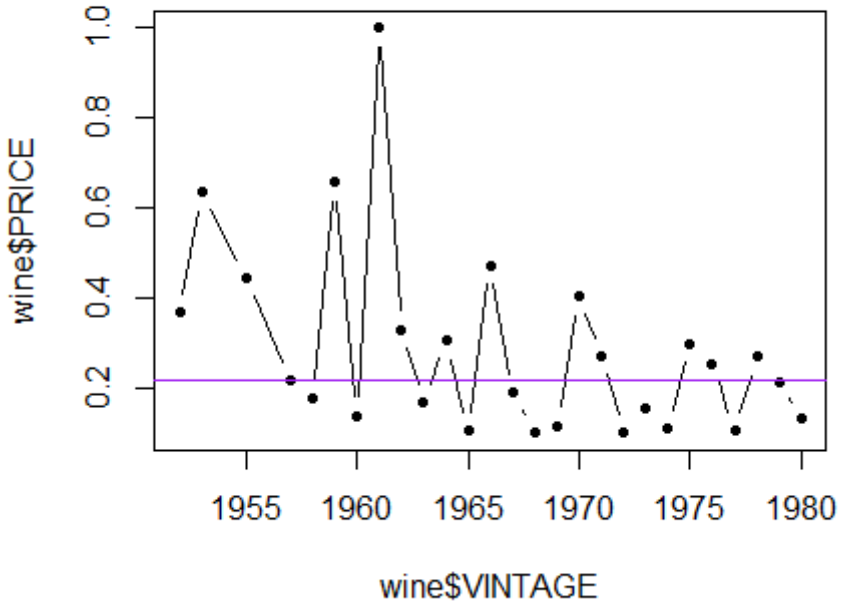
y is coded as the average price of six Bordeaux wines (six *premier cru* châteaux in Médoc), relative to the average price of 1961 vintage (1961 = 1)

Explanatory Variables:

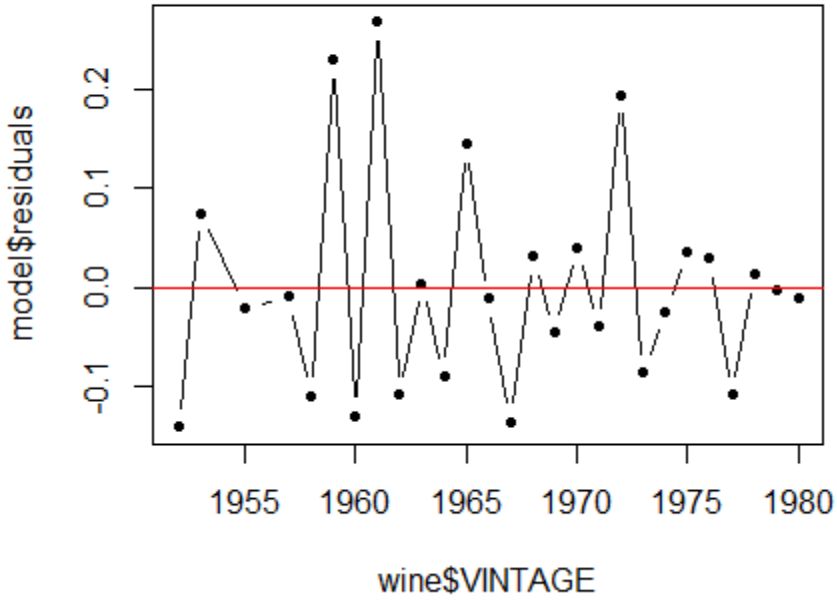
- the *age* of the vintage in years (cellaring is expensive, so older wines should cost more);
- the temperature of the region during the growing season, i.e., from April and September, in degrees Celsius (warmer years are supposed to yield better grapes);
- the average rainfall between August and September in mL (less rain during the peak growing season is desirable); and
- the average rainfall during the preceding October-March period in mL (more rain between growing seasons are desirable)

3. Example of a Stationary Series: *Wine Prices*

plot of dependent variable over time



plot of model's residuals over time



* Pay particular attention to the how often consecutive observations are above the average line or below the average line; count the number of "runs" (for part 6 – the "runs test"): after dropping years before 1956, there are 15 runs in left figure, 16 runs in right figure

4. Example of a Non-Stationary Series: *Manatee Deaths*

Data: manatees.csv
Script: Lecture 13 manatees.R

N = 33 years from 1977 to 2009

Kills is the number of manatees killed in a boating accident in a calendar year
Boats is the number of boats registered in the state of Florida (x 1,000)

Kills is autoregressive (examine correlation with $Kills_{t-1}$, correlogram, runs test, etc.)
Boats is autoregressive (examine correlation with $Boats_{t-1}$, correlogram, runs test, etc.)

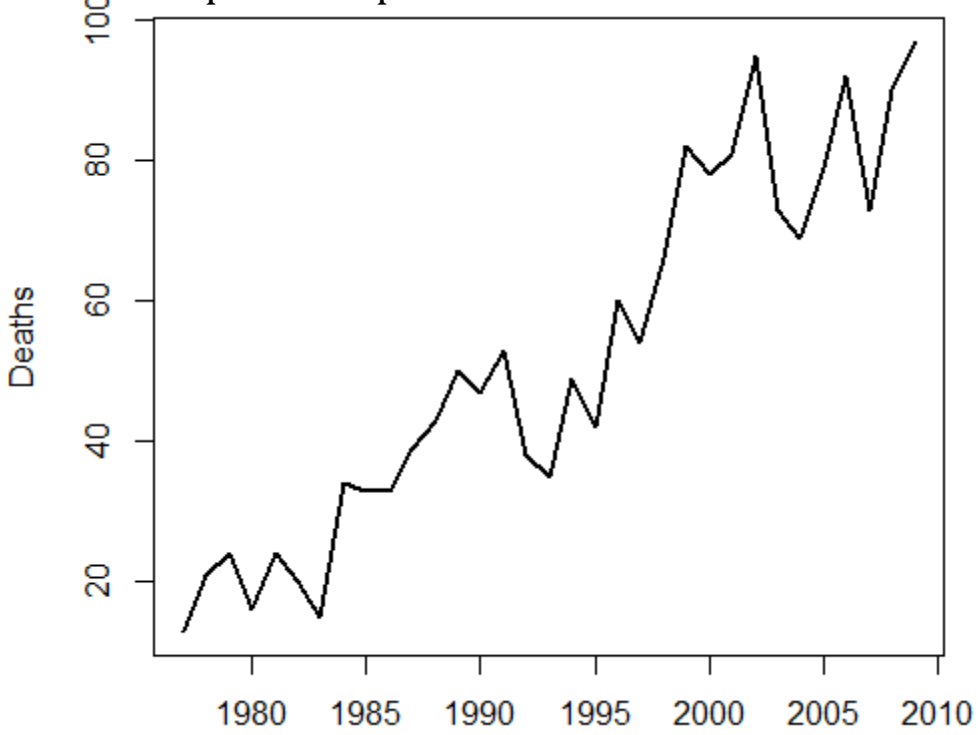
However, if you estimate a static model, the residual is not serially correlated!

When y is I(1) and x is I(1) but $\hat{u}$ is I(0), we say that x and y are cointegrated.
($\hat{u}$ is the residual from a regression of y on x)

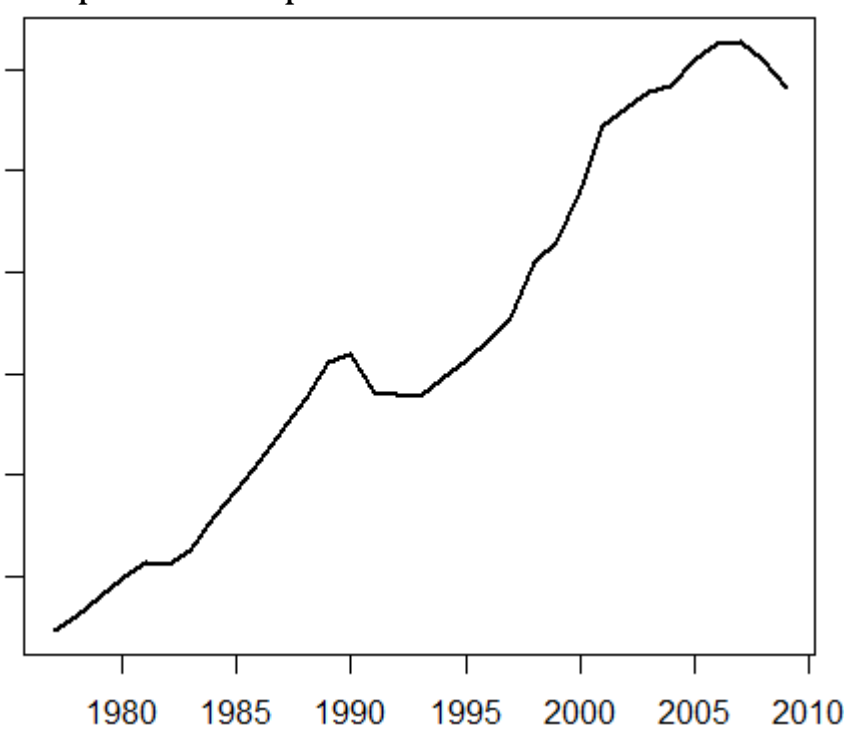
This might be due to spurious regression (see § 18.3) but it may not be (see § 18.4)

4. Example of a Non-Stationary Series: *Manatee Deaths*

plot of dependent variable over time



plot of independent variable over time



$Kills_t$ correlates with $Kills_{t-1}$ at $r = .91$

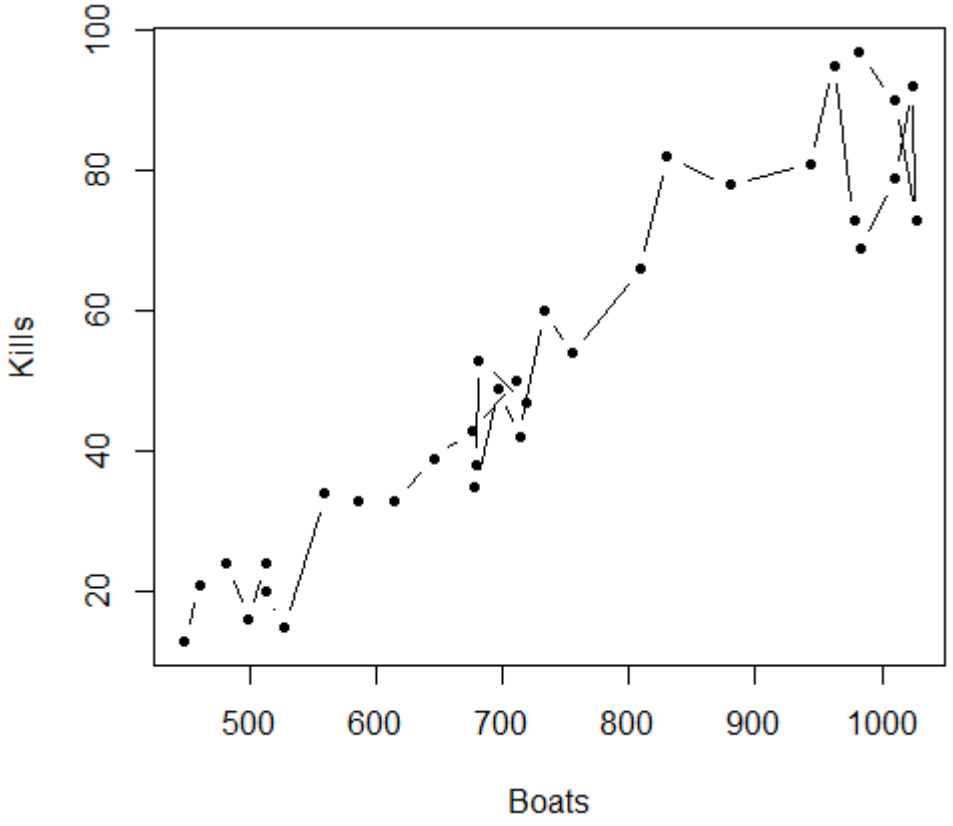
$Kills_t = 6.54 + 0.92\ Kills_{t-1}$ ($R^2 = .83$)

$Boats_t$ correlates with $Boats_{t-1}$ at $r = .99$

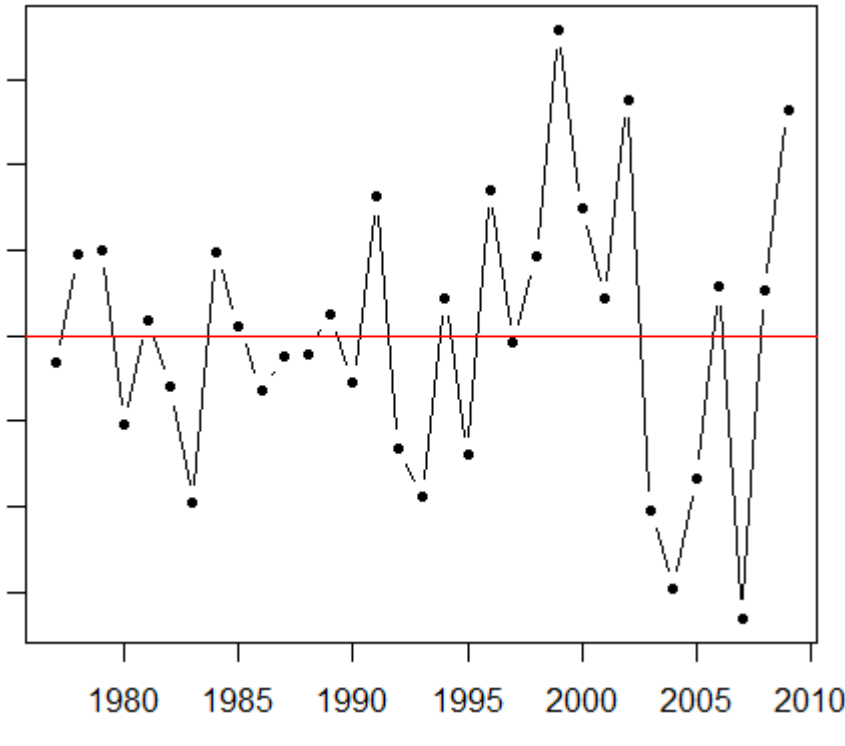
$Boats_t = 29.64 + 0.98\ Boats_{t-1}$ ($R^2 = .99$)

4. Example of a Non-Stationary Series: *Manatee Deaths*

scatterplot



plot of model's residuals over time



$Kills_t = -43.2 + 0.13\ Boats_t$ ($R^2 = .905$)

$\hat{u}_t = 0.088 + 0.092\ \hat{u}_{t-1}$ ($R^2 = .008$)

POLS 6481. Research Design and Quantitative Methods II
Lecture 13. Time Series Data I: Stationarity, Autoregression, and Unit Roots
Wooldridge, *Introductory Econometrics*, 11.1

5. Defining Stationary and Non-Stationary Series

Stationary series are stable over time, for example:
"mean stationary" constant mean
"trend stationary" mean follows a time-trend; for wine data, we can use vintage
"covariance stationary" constant mean, constant variance, cov(x_t, x_{t+h}) depends only on h
"weak dependence" corr(x_t, x_{t+h}) goes to 0 as h goes to ∞

Non-stationary series are unstable over time, because errors accumulate; for example:
"random walk" $y_t = y_{t-1} + e_t$

$$\Rightarrow\ y_t = (y_{t-2}+e_{t-1})+e_t = \ldots = e_t+e_{t-1}+\ldots+e_1+y_0$$

"random walk with drift" $y_t = y_{t-1} + \alpha + e_t$

$$\Rightarrow\ y_t = \alpha_0 t + e_t + e_{t-1} + \ldots + e_1 + y_0$$

5. Defining Stationary and Non-Stationary Series

Non-stationary series

Four interesting facts about random walks …		… with drift:
Expected value of y at t:	$E(y_t) = E(y_0)$	$E(y_t) = E(y_0) + \alpha \cdot t$
Conditional expected value of y at t:	$E(y_{t+h} \mid y_t) = y_t$	$E(y_{t+h} \mid y_t) = y_t + \alpha \cdot h$
Correlations between values of y:	$Corr(y_t, y_{t+h}) = \sqrt{t/(t+h)}$	
Variance of y at t:	$Var(y_t) = \sigma_e^2 \cdot t$	
	$= Var(e_t) + Var(e_{t-1}) + \ldots + Var(e_1)$	

6. The Run-Test (nonparametric)

Generated from combinatorics; condition for using normal approximation (shown below) is number of 'failures' (n_f) and number of 'successes' (n_s) each must exceed 10

Mean number of runs: $$E(R) = \frac{2 \cdot n_f \cdot n_s}{n_f + n_s} + 1$$

Variance in number of runs: $$Var(R) = \frac{(2 \cdot n_f \cdot n_s) \cdot (2 \cdot n_f \cdot n_s - n_f - n_s)}{(n_f + n_s)^2 \cdot (n_f + n_s - 1)}$$

Equations reflect assumption of independence; actual number of runs observed in data = R

Test Statistic $$z = \frac{R - E(R)}{\sqrt{Var(R)}}$$

https://en.wikipedia.org/wiki/Wald–Wolfowitz_runs_test

Wald, A., and Wolfowitz, J. (1940) "On a Test Whether Two Samples Are From the Same Population," *Annals of Mathematical Statistics*, 11: 147–162

Geary, R.C. (1970) "Relative Efficiency of Count of Sign Changes for Assessing Residual Autoregression in Least Squares Regression," *Biometrika*, 57: 123–127

6. The Run-Test applied to the *Wine Prices* dependent variable (1957 to 1980 only)

Conditions: number of values below mean (n_f) = 15; number of values above mean (n_s) = 9

Mean number of runs: $E(R) = \frac{2 \cdot 15 \cdot 9}{15+9} + 1 = 12.25$

Variance in number of runs: $Var(R) = \frac{(2 \cdot 15 \cdot 9) \cdot (2 \cdot 15 \cdot 9 - 15 - 9)}{(15+9)^2 \cdot (15+9-1)} = \frac{270 \cdot 246}{24 \cdot 24 \cdot 23} = 5.0136$

Test statistic (R = 15) $z = \frac{15 - 12.25}{\sqrt{5.0136}} = 1.228$

Retention region for runs: $E(R) \pm 1.96 \cdot \sqrt{Var(R)}$
$= 12.25 \pm 1.96 \cdot 2.24$
$= 12.25 \pm 4.39$
$= [7.86 , 16.64]$

It may make sense to use a one-sided (left tail only) confidence interval here instead…;
if R is less than lower boundary of confidence interval, suspect serial auto-regression.
Use *runs*() program in *TSA* package in R; reports *p* value, E(R), n_f, n_s,
Use *runs.test*() program in *randtests* package in R; reports *z* statistic and *p* value.

6. The Run-Test applied to the *Wine Prices* regression residuals (1957 to 1980 only)

Conditions: number of negative residuals (n_f) = 13; number of positive residuals (n_s) = 11

Mean number of runs: $E(R) = \frac{2 \cdot 13 \cdot 11}{13+11} + 1 = 12.917$

Variance in number of runs: $Var(R) = \frac{(2 \cdot 13 \cdot 11) \cdot (2 \cdot 13 \cdot 11 - 13 - 11)}{(13+11)^2 \cdot (13+11-1)} = \frac{286 \cdot 262}{24 \cdot 24 \cdot 23} = 5.656$

Test statistic ($R = 16$) $z = \frac{16 - 12.917}{\sqrt{5.656}} = 1.296$

Retention region for runs: $E(R) \pm 1.96 \cdot \sqrt{Var(R)}$
$= 12.92 \pm 1.96 \cdot 2.38$
$= 12.92 \pm 4.66$
$= [8.26 , 17.58]$

If R is less than lower boundary of confidence interval, suspect positive auto-correlation; if R exceeds upper boundary on confidence interval, suspect negative auto-correlation.

Use *runs.test*() program in *randtests* package in R; reports *z* statistic and *p* value.

POLS 6481. Research Design and Quantitative Methods II
Lecture 13. Time Series Data I: Stationarity, Autoregression, and Unit Roots
Wooldridge, *Introductory Econometrics*, 11.3b + 18.2

7. Auto-regression and Unit Roots

"Auto-regressive process of order one" AR(1) $y_t = \rho \cdot y_{t-1} + e_t$

Process begins at y_0 and evolves over time as $t = 1, 2, \ldots$

Process is stable if and only if $| \rho | < 1$ (in this case, $\text{corr}(y_t, y_{t+h}) = \rho^t$, which goes to 0…)

Process is unstable if $\rho = 1$, in which case consider using differences (Δy_t):
when y_t is integrated of order one [I(1)], Δy_t usually are I(0), or weakly dependent

Caveat: if y depends on any x that are I(1), then the model is *co-integrated*;
Wooldridge's 18.3–18.4 introduce the spurious regression problem and solutions

8. The Dickey-Fuller Test for a Unit Root

"Auto-regressive process of order one"… $y_t = \alpha + \rho \cdot y_{t-1} + e_t$
… has a "unit root" if $\rho = 1$

Tests for a unit root: regress y_t on y_{t-1} and test hypothesis $\rho = 1$
regress $\Delta y_t = (y_t - y_{t-1})$ on y_{t-1} and test hypothesis $\theta = 0$ (where $\theta = \rho - 1$)

Dickey-Fuller Test $$\Delta y_t = \alpha + \theta y_{t-1} + e_t$$

Augmented Dickey-Fuller test uses lagged values of Δy

$$\Delta y_t = \alpha + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + e_t$$

TABLE 18.2 Asymptotic Critical Values for Unit Root *t* Test: No Time Trend

Significance level	1%	2.5%	5%	10%
Critical value	−3.43	−3.12	−2.86	−2.57

(program *adf.test* in package *tseries*)

8. The Dickey-Fuller Test for a Unit Root

"Auto-regressive process of order one"… $y_t = \alpha + \rho \cdot y_{t-1} + e_t$
… has a "unit root" if $\rho = 1$

Tests for a unit root: regress y_t on y_{t-1} and test hypothesis $\rho = 1$
regress $\Delta y_t = (y_t - y_{t-1})$ on y_{t-1} and test hypothesis $\theta = 0$ (where $\theta = \rho - 1$)

Dickey-Fuller Test $$\Delta y_t = \alpha + \theta y_{t-1} + e_t$$

Augmented Dickey-Fuller test uses lagged values of Δy and can also include a time trend:

$$\Delta y_t = \alpha + \delta t + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_p \Delta y_{t-p} + e_t$$

TABLE 18.3 Asymptotic Critical Values for Unit Root *t* Test: Linear Time Trend

Significance level	1%	2.5%	5%	10%
Critical value	−3.96	−3.66	−3.41	−3.12

Under the alternative hypothesis of no unit root, the process is trend-stationary