

Regression with time-series data

Wooldridge ch. 10–12, with Swedish data

Magnus Lodefalk

Örebro University

Economics Research and Communication (NA310G)

13 October 2026



Who I am, and the thread through my courses

- Magnus Lodefalk: I co-lead the AI-Econ Lab at Örebro University, on how AI changes work and firms
- **One question:** does AI take over tasks, make workers more productive, or create new tasks?
- **One family of data:** firms' accounts, job ads asking for AI skills, and the AI exposure of occupations; your report uses them

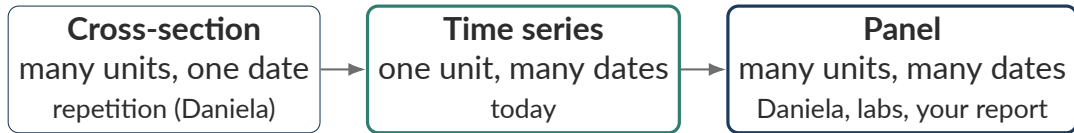
My part bridges to panel data, the core of the course, using the questions we study in the lab.

Today's goals

By the end of today you can:

1. State what OLS assumes with time series and why each assumption is stronger than in a cross-section; read static and distributed-lag models with trends (ch. 10)
2. Explain how large samples let us loosen those assumptions, and the price: stationary, weakly dependent data (ch. 11)
3. Recognise serial correlation and handle it, and know why your report needs `vce(cluster firmID)` (ch. 12)

Today: the stepping stone to panel data



- From the cross-section we keep OLS and its assumptions; time adds order, persistence and feedback
- A panel combines both (many firms, each a short time series): the focus of the course

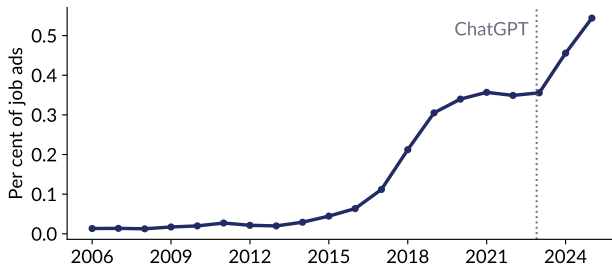
What you learn today about time is what makes panel data work in your report.

From last time

Answer on your own first, then compare with a neighbour

1. Which cross-section assumption gives you an unbiased slope?
2. Which assumption makes the observations independent of each other?

AI job ads rose 41-fold: is AI taking or creating tasks?



0.54 per cent of ads in 2025. One series, one strong trend: we return to what it can tell us at the end.

Source: AIEL Monitor (v1.6), Platsbanken ads via JobTech.

Key terms for chapter 10

- **Time series:** one unit observed at many dates, $y_1, y_2, \dots, y_T$, in order
- **Lag, lead:** x_{t-1} is the lag of x_t (last period); x_{t+1} is its lead (next period)
- **Trend:** a series that drifts in one direction over time
- **Serial correlation:** errors in different periods are correlated, for example this month's with last month's

From a cross-section to a time dimension

- **Cross-section:** many units at one date. Random sampling makes the observations independent, so their order carries no information
- **Time series:** one unit at many dates. There is no random sampling: we see one history, in order, and neighbouring periods resemble each other
- **Feedback:** today's y can change tomorrow's x , as when policy reacts to the economy

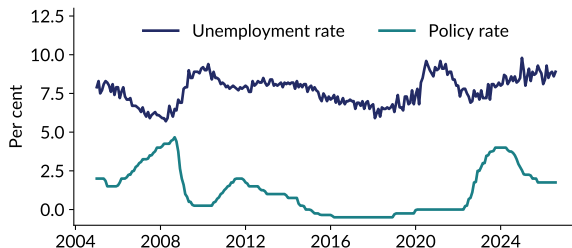
In symbols

Cross-section: $y_i = \beta_0 + \beta_1 x_i + u_i, i = 1, \dots, n$

Time series: $y_t = \beta_0 + \beta_1 x_t + u_t, t = 1, \dots, T$

Losing random sampling is the root of everything today: the assumptions must say directly what random sampling used to deliver for free.

Running example: does the rate move unemployment?



260 months, 2005 to 2026. Both persistent: high this month, high next month.

Source: SCB Labour Force Survey (seasonally adjusted, 15–74); Riksbank policy rate.

A static regression, run as in a cross-section

```
tsset t
regress unemployment policy_rate
```

unemployment	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
policy_rate	-.0655376	.0379629	-1.73	0.085	-.1402942	.009219
_cons	7.794248	.072552	107.43	0.000	7.651379	7.937118

- tsset tells Stata the rows are months in order
- A higher rate goes with **lower** unemployment, $t = -1.7$: the Riksbank raises the rate when the economy is strong

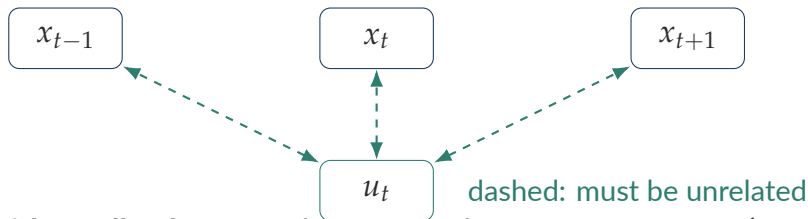
Can we trust this coefficient and this t? It depends on assumptions, next.

The assumptions, side by side

	Cross-section	Time series
Linear in parameters	MLR.1	TS.1
Random sampling	MLR.2	none: one ordered history
No perfect collinearity	MLR.3	TS.2
Zero conditional mean	MLR.4: $E(u_i \mathbf{x}_i) = 0$	TS.3: $E(u_t \mathbf{X}) = 0$
Constant variance	MLR.5	TS.4: $\text{Var}(u_t \mathbf{X}) = \sigma^2$
No serial correlation	(from random sampling)	TS.5: assumed outright
Normal errors	MLR.6	TS.6

TS.3 is **strict exogeneity** ($\mathbf{X}$: all regressors in all periods). TS.1–3: unbiased; TS.1–5: BLUE (best: the smallest variance among linear unbiased estimators); with TS.6, exact t and F tests.

Strict exogeneity in one picture



- **Fails with feedback:** unemployment today moves tomorrow's rate: u_t relates to x_{t+1}
- **Fails with a lagged y :** u_t is part of y_t , the regressor at $t + 1$
- **Related to a past x ?** Add x_{t-1} as a regressor

In symbols: key

Strict (TS.3): $E(u_t \mid x_1, \dots, x_T) = 0$

Contemporaneous (TS.3', ch. 11): $E(u_t \mid x_t) = 0$

Check your understanding

Answer on your own first, then compare with a neighbour

The Riksbank cuts the rate when unemployment rises. In our regression, which assumption fails, and what does that do to OLS?

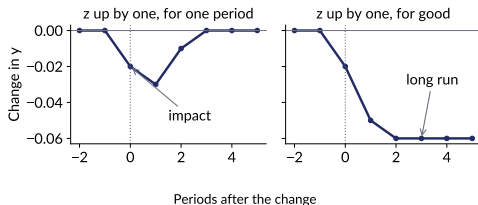
Two models: static and finite distributed lag

Static: $y_t = \beta_0 + \beta_1 z_t + u_t$

Distributed lag: $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t$

- **Impact propensity** δ_0 : the immediate change in y when z rises by one unit at t
- **Long-run propensity** $\delta_0 + \delta_1 + \delta_2$: how much higher y ends up, and stays, after a *permanent* one-unit increase in z

A temporary or a permanent change in z



In symbols: key

Temporary: y moves by $\delta_0, \delta_1, \delta_2$, then back

Permanent: $LRP = \delta_0 + \delta_1 + \delta_2$

A temporary change dies out; a permanent one builds up to the long-run propensity.

Source: Invented weights: $\delta_0 = -0.02, \delta_1 = -0.03, \delta_2 = -0.01$.

Trends: control for them

```
gen trend = _n  
regress unemployment policy_rate trend
```

policy_rate	-.0406122	.0367605	-1.10	0.270
trend	.0034995	.0007201	4.86	0.000

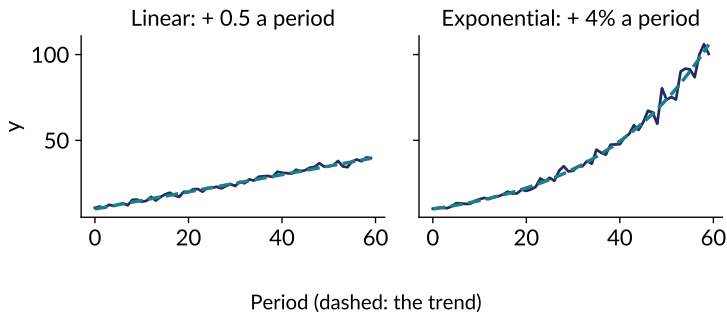
- With a trend the coefficient goes from -0.07 to -0.04 (the same as detrending both series)
- With a trending y , R^2 is inflated; Wooldridge computes it on the detrended y

In symbols

Linear: $y_t = \alpha_0 + \alpha_1 t + e_t$

Exponential: $\ln y_t = \beta_0 + \beta_1 t + e_t$, $\beta_1 \approx$ growth per period

Linear and exponential trends



A linear trend adds the same amount each period; an exponential trend adds the same per cent. Use logs for the second.

Source: Simulated series.

Seasons and events

- **Seasonality:** monthly series rise and fall with the season; add one dummy per month less one (`i.month`), or use a seasonally adjusted series, as SCB's here
- **Event dummies:** 1 during an event and 0 otherwise, for example the pandemic months
- **Logs** work as in a cross-section: $\ln y$ on $\ln z$ gives an elasticity

In symbols

$$y_t = \beta_0 + \delta_1 feb_t + \cdots + \delta_{11} dec_t + \beta_1 x_t + u_t \quad (\text{January is the base month})$$

Same logic as a trend: control for what moves both series, so it is not read as an effect.

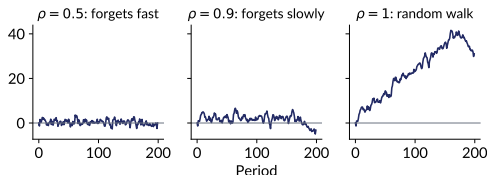
Key terms for chapter 11

- **Stationary** (covariance stationary): stable over time: same mean and variance; $\text{Corr}(y_t, y_{t+h})$ depends only on h
- **Weakly dependent**: that correlation dies out as h grows: the series forgets its past
- **Trend-stationary**: stationary and weakly dependent around a trend; meets TS.1' with the trend included
- **Random walk, I(1)**: today = yesterday + a new shock; shocks never fade
- **Consistent**: close to the truth for large T (unbiased: right on average)

In symbols

Stationary: $E(x_t) = \mu$, $\text{Var}(x_t) = \sigma^2$, $\text{Cov}(x_t, x_{t+h})$ depends on h only

Three series, the same shocks, three memories



In symbols: key

AR(1): $y_t = \rho y_{t-1} + e_t$, $\text{Corr}(y_t, y_{t+h}) = \rho^h \rightarrow 0$ if $|\rho| < 1$

Random walk ($\rho = 1$): $\text{Var}(y_t) = \sigma_e^2 t$, growing with time

Weakly dependent: shocks and policy changes fade. Highly persistent: they last.

Source: Simulated: 200 periods, the same random shocks in all three panels.

The trade: weaker assumptions, larger samples

	Finite sample (ch. 10)	Large sample (ch. 11)
Data	any series	stationary, weakly dependent
Exogeneity	strict, all periods	contemporaneous, same period
Lagged y as regressor	violates TS.3	allowed
Normal errors	needed for exact tests	not needed
OLS is	unbiased	consistent, may be biased

Contemporaneous exogeneity keeps only the arrow at t in the timeline picture. The price: data that are stable and forget the past.

How persistent are our two series?

```
regress unemployment L.unemployment  
regress policy_rate L.policy_rate
```

```
unemployment  L1. |    .9029951    .0272727  
policy_rate    L1. |    .9938652    .0067321
```

- The policy rate is close to a random walk
- Unemployment is persistent but weakly dependent: 0.90 per month is about 0.29 per year, and a shock halves in about 6.8 months

In symbols

Half-life: $\rho^h = 0.5 \Rightarrow h = \ln 0.5 / \ln \rho$

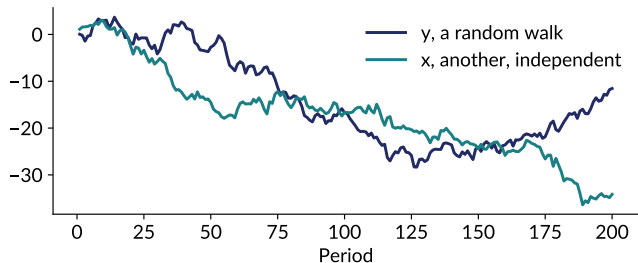
How close to 1 is "too close" depends on the frequency of the data.

Two unrelated random walks: how often is the slope significant?

Your bet first: hands up. The answer is on the next slide.

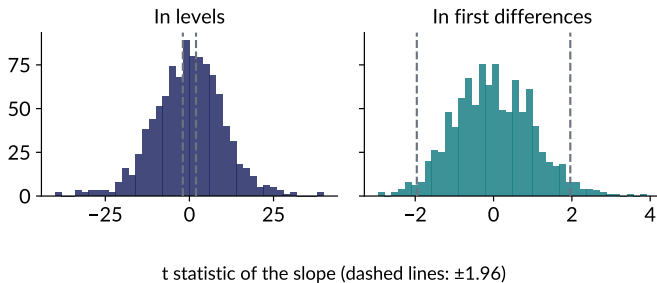
- A. In about 5% of 1,000 regressions (200 periods each), as it should be
- B. In about 25%
- C. In most of them

Two independent random walks



They wander, and over a stretch they seem to move together. Regressing y on x gives $t = 13.6$.

Spurious regression: 1,000 times



In levels 84 per cent of the t statistics fall outside ± 1.96 ; in first differences 5 per cent, as they should.

Why we expected 5 per cent, and why it failed

- **The instinct:** with no true relation, a 5% test rejects in 5% of samples. That is what the significance level means
- **But only if** the t statistic follows the t distribution, which needs stationary, weakly dependent data
- A random walk never forgets: its shocks pile up, and so does the error in $y_t = \beta_0 + \beta_1 x_t + u_t$. The standard error is far too small, and t is far too large
- In first differences each series is just its new shocks: weakly dependent again, and 5% is back

The 5% rule is a promise of the assumptions. With random walks in levels, the promise is void.

Highly persistent series: difference them

- Differencing removes both: a random walk becomes just the new shock, $I(0)$, and a linear trend becomes the constant
- **Rule of thumb:** compute the first-order autocorrelation (of the detrended series if it trends); in annual data, if it is above about 0.9, use first differences
- A random walk need not trend, and adding a trend does not fix it: difference

In symbols: key

Random walk $y_t = y_{t-1} + e_t \Rightarrow \Delta y_t = y_t - y_{t-1} = e_t$

Look at the series; if highly persistent, difference it; then OLS with robust standard errors.

In differences, with a year of lags: imprecise

```
newey D.unemployment ///  
      L(0/12).D.policy_rate, lag(12)  
lincom D.policy_rate + LD.policy_rate ///  
      + ... + L12D.policy_rate
```

(1)	-.0659871	.1882445	-0.35	0.726
-----	-----------	----------	-------	-------

- D. means the change since last month; L(0/12) . lags 0 to 12
- Long-run propensity -0.07, 95% confidence interval -0.44 to 0.30

Too imprecise to say. And if the Riksbank reacts within the month, even contemporaneous exogeneity fails: that needs a research design, not more lags.

What serial correlation does to OLS

- With strictly exogenous regressors, OLS is **still unbiased** (and consistent with weakly dependent data), and R^2 is still fine, but **no longer BLUE**
- The usual standard errors, t and F tests are **invalid**: with positive serial correlation, too small (correlated months carry less information)
- With a lagged y and AR(1) errors OLS is inconsistent: y_{t-1} contains u_{t-1} .
The cure is more lags: with enough lags, TS.3' holds and OLS is consistent

In symbols

With AR(1) errors $u_t = \rho u_{t-1} + e_t$ (and x demeaned):

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} + \frac{2\sigma^2}{SST_x^2} \sum_{t=1}^{T-1} \sum_{j=1}^{T-t} \rho^j x_t x_{t+j};$$
 the usual formula keeps only the first term

This month's residual on last month's: what slope?

Your bet first: hands up. The answer is on the next slide.

- A. About 0: errors are independent
- B. About 0.5
- C. Close to 1: errors are very persistent

The residuals have long runs above and below zero



```
regress u L.u policy_rate
```

In symbols

Test: $\hat{u}_t = \rho \hat{u}_{t-1} + \gamma x_t + e_t, H_0: \rho = 0$

Slope on the lagged residual 0.90: TS.5 fails badly.

Check your understanding

Answer on your own first, then compare with a neighbour

We now ask Stata for standard errors that allow for serial correlation. What will change: the coefficient, the standard error, or both?

The remedy: HAC (Newey–West) standard errors

```
newey unemployment policy_rate, lag(12)
```

OLS:	policy_rate		-.0655376	.0379629	-1.73
newey:	policy_rate		-.0655376	.1472251	-0.45

- Same coefficient; standard error 3.9 times larger (4 lags: only 2.5 times, too few here)
- **HAC** = robust to correlation over time. **FGLS** (prais) needs strict exogeneity: not here

In symbols

FGLS quasi-differences: $y_t - \rho y_{t-1} = \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + e_t$

Near-random walk: difference, then HAC. Consistent, valid tests in large samples; not unbiased, and no fix for feedback.

One idea, three settings: let the errors be correlated

Standard errors	Allow for	Stata
Robust (Daniela)	unequal error variance	<code>vce(robust)</code>
HAC (today)	+ correlation over time	<code>newey</code>
Clustered (report)	+ correlation within a firm	<code>vce(cluster firmID)</code>

Same coefficient every time; only the honesty of the standard error changes.

Your report: a panel is many short time series

- One firm observed 2018 to 2024 is a time series; the course data hold 1,534 of them
- A firm's error this year is related to its error last year: TS.5 fails within firms
- Firm fixed effects remove each firm's permanent level, not the persistence of its shocks: you still cluster
- Year effects ($i.year$) do for a panel what a trend does for one series

In symbols: key

$$y_{it} = a_i + \lambda_t + \beta x_{it} + u_{it}$$

$\text{Corr}(u_{it}, u_{is}) \neq 0$ within firm i : cluster by i

Bertrand, Duflo and Mullainathan (2004): ignoring this, up to 45% of invented state laws looked significant at the 5% level.

Clustering on the course data

```
xtset firmID year
xtreg ln_assets ai_firm i.year, fe
xtreg ln_assets ai_firm i.year, fe ///
    vce(cluster firmID)
```

default:	ai_firm		.1881241	.0277167	6.79
clustered:	ai_firm		.1881241	.0401257	4.69

- ai_firm = 1 if the firm posted an ad asking for AI skills that year
- Same coefficient; the clustered standard error is 1.4 times larger

In your report: cluster by firm.

Check your understanding

Answer on your own first, then compare with a neighbour

AI ads rose sharply after 2022, the same years as the interest-rate rise. Can the AI series alone tell us whether ChatGPT or the higher interest rate reduced hiring?

What a panel can do: our own study

- Swedish employers, monthly register data, 2023 to 2024-25
- Compare AI-exposed and less exposed employers, before and after ChatGPT

Young workers, 22-25 At AI-exposed employers the ratio of young to older workers fell 3.9 per cent more than at less exposed employers.

Variation across firms separates what one time series cannot. That is your report's design.

Summary

1. Ch. 10: without random sampling OLS needs strict exogeneity and no serial correlation, both strong; static and distributed-lag models; control for trends and seasons
2. Ch. 11: in large samples contemporaneous exogeneity is enough, at a price: stationary, weakly dependent data; difference highly persistent series
3. Ch. 12: serial correlation breaks the usual standard errors: HAC for one series, `vce(cluster firmID)` in your panel

Next: cross-section and time series meet in your panel.

Test yourself before next time

Without AI: this is the exam competence

1. Which cross-section assumption has no counterpart in time series, and what replaces it?
2. State strict exogeneity in words, and give two reasons it fails in time series
3. $y_t = \alpha_0 + 0.4z_t + 0.2z_{t-1} + u_t$: what are the impact and long-run propensities?
4. Why can two independent random walks give a significant regression slope, and what fixes it?
5. What does `newey` change compared with `regress`? Why cluster by firm in a panel?

Appendix

Reading guide: chapter 10

- **10-1 The nature of time-series data:** taught
- **10-2 Static and finite distributed-lag models:** taught; work the book's examples
- **10-3 Finite-sample properties (unbiasedness, Gauss–Markov, inference):** taught as results; read the theorems in the book
- **10-4 Functional form, dummy variables, index numbers:** touched; read on your own
- **10-5 Trends and seasonality:** taught; read detrending (10-5c) and R^2 with a trending y (10-5d)

Chapter 10 is the core of this lecture: know it well.

Reading guide: chapters 11 and 12

- **11-1 Stationary and weakly dependent series:** taught; MA(1) read on your own
- **11-2 Asymptotic properties of OLS:** taught as the "trade" table
- **11-3 Highly persistent series:** taught
- **11-4 Dynamically complete models; 11-5 homoskedasticity:** read on your own
- **12-1 OLS with serial correlation; 12-2 robust (HAC) inference:** taught
- **12-3 Testing; 12-4 FGLS; 12-5 differencing:** briefly
- **12-6 Heteroskedasticity in time series:** not covered

Wooldridge, *Introductory Econometrics*, 8th edition (2025).

Going further: a thesis in macroeconomics or finance

- **The book:** Wooldridge ch. 18: unit-root tests (18-2), cointegration and error-correction models (18-4), forecasting (18-5), and event studies with control groups (18-6)
- **In Stata:** `dfuller` (unit-root test), `arima`, `var`; all need `tsset`
- **Swedish data, free:** the Riksbank's interest and exchange rates, SCB's statistical database, and Serrano for firm accounts (free for Örebro University students)

Today is the introduction; chapter 18 is the next step if your thesis is about time.

The large-sample assumptions in full (ch. 11)

- **TS.1'**: linear, and the series are stationary and weakly dependent (trend-stationary series qualify)
- **TS.2'**: no perfect collinearity
- **TS.3'**, **contemporaneous exogeneity**: $E(u_t | \mathbf{x}_t) = 0$
- **TS.4'**: contemporaneous homoskedasticity; **TS.5'**: no serial correlation

Under TS.1'–TS.3' OLS is consistent; adding TS.4'–TS.5' the usual standard errors, t and F statistics are valid in large samples.

Dynamically complete models and the classic tests

- **Dynamically complete:** enough lags of y and z that no further lag helps predict y ; then the errors are not serially correlated
- **Strict stationarity:** the whole joint distribution is unchanged by a shift in time
- **Durbin-Watson** (`estat dwatson`) needs strict exogeneity and normal errors; **Breusch-Godfrey** (`estat bgodfrey`) does not

Distributed lags of a treatment are an event study

$$y_{it} = \alpha_i + \lambda_t + \sum_{k=-3}^3 \beta_k \mathbf{1}\{t - E_i = k\} + u_{it}$$

- E_i : the year firm i first posts an AI ad; k counts years from that event
- **Lags** ($k \geq 0$): how the outcome evolves after adoption
- **Leads** ($k < 0$): differences before adoption; if large, the comparison is suspect

The distributed-lag logic with firms as units: Wooldridge 18-6, and where this goes in your thesis.