Econometrics

Practical Session 24

Endterm 2024/25 – Walkthrough

Ricardo Gouveia-Mendes
rgouveiamendes@ucp.pt

Spring 2025-26

Católica-Lisbon School of Business and Economics

Multiple Choice

Q1 | Internal validity

A statistical analysis is **internally valid** if:

A. the statistical inferences about causal effects are valid for the population being studied

B. the hypothesised parameter value is inside the confidence interval

C. its inferences and conclusions can be generalised from the population and setting studied to other populations and settings

D. statistical inference is conducted inside the sample period

Q1 | Internal validity

A statistical analysis is **internally valid** if:

A. the statistical inferences about causal effects are valid for the population being studied

B. the hypothesised parameter value is inside the confidence interval

C. its inferences and conclusions can be generalised from the population and setting studied to other populations and settings

D. statistical inference is conducted inside the sample period

SOLUTION

Answer: A.

- **Internal validity:** inferences about causal effects are valid for the population **being studied** (no OVB, no simultaneity, correct SEs, ...)
- C – **External validity**: generalisability to other populations / settings
- B – confuses validity with a hypothesis-test outcome
- D – nonsensical

Q2 | Adding a regressor

By including another variable in the regression, you will:

A. decrease the regression R^2 if that variable is important

B. look at the t-statistic and include the variable only if it is significant at 1%

C. eliminate the possibility of omitted-variable bias from excluding that variable

D. decrease the variance of the estimator of the coefficients of interest

Q2 | Adding a regressor

By including another variable in the regression, you will:

A. decrease the regression R^2 if that variable is important

B. look at the t-statistic and include the variable only if it is significant at 1%

C. eliminate the possibility of omitted-variable bias from excluding that variable

D. decrease the variance of the estimator of the coefficients of interest

SOLUTION

Answer: D.

- A – R^2 **never decreases** when a regressor is added (it weakly increases)
- B – significance is not the criterion for inclusion; theory + OVB risk are
- C – only OVB from **that** variable is removed; other omitted variables remain
- D – a relevant control reduces residual variance → smaller SE on coefficients of interest

Q3 | Departures from stationarity

Departures from stationarity:

A. cannot be fixed

B. occur often in cross-sectional data

C. can be made to have less severe consequences by using log–log specifications

D. jeopardise forecasts and inference based on time-series regression

Q3 | Departures from stationarity

Departures from stationarity:

A. cannot be fixed

B. occur often in cross-sectional data

C. can be made to have less severe consequences by using log–log specifications

D. jeopardise forecasts and inference based on time-series regression

SOLUTION

Answer: D.

- Non-stationarity invalidates standard t/F inference and the forecast itself
- A is false – differencing, detrending, or breaking the sample often fixes it
- B is false – stationarity is a **time-series** concept
- C is false – logs help with variance, not with unit roots or breaks

Q4 | Endogenous instruments

If the instruments are **not** exogenous:

A. then, in order to conduct proper inference, it is essential that you use heteroskedasticity-robust standard errors

B. you cannot perform the first stage of TSLS

C. your model becomes over-identified

D. then TSLS is inconsistent

Q4 | Endogenous instruments

If the instruments are **not** exogenous:

A. then, in order to conduct proper inference, it is essential that you use heteroskedasticity-robust standard errors

B. you cannot perform the first stage of TSLS

C. your model becomes over-identified

D. then TSLS is inconsistent

SOLUTION

Answer: D.

- Consistency of TSLS requires $\text{cov}(z, u) = 0$ (the **exogeneity** / exclusion restriction)
- Failing this contaminates the IV estimator, no matter how large n gets
- A – Robust SEs fix nothing here
- B – the first stage is mechanical and still runs
- C – identification is about counts of instruments, not exogeneity

Q5 | Negative autocorrelation

Negative autocorrelation in the change of a variable implies that:

A. the variable contains only negative values

B. an increase in the variable in one period is, on average, associated with a **decrease** in the next

C. the series is not stable

D. the data are negatively trended

Q5 | Negative autocorrelation

Negative autocorrelation in the change of a variable implies that:

A. the variable contains only negative values

B. an increase in the variable in one period is, on average, associated with a **decrease** in the next

C. the series is not stable

D. the data are negatively trended

SOLUTION

Answer: B.

- Negative $\text{corr}(\Delta y_t, \Delta y_{t-1}) < 0$ → today's change reverses tomorrow on average
- A – the sign of the autocorrelation in the difference of a variable says nothing about the **sign** of the levels
- D – it also says nothing about the **trend**
- C – it is perfectly compatible with stability

Q6 | The AR(p) model

The AR(p) model:

A. can be represented as follows: $y_t = \beta_0 + \beta_1 x_t + \beta_p y_{t-p} + u_t$

B. is defined as $y_t = \beta_0 + \beta_p y_{t-p} + u_t$

C. can be written as $y_t = \beta_0 + \beta_1 y_{t-1} + u_{t-p}$

D. represents y_t as a linear function of p of its lagged values

Q6 | The AR(p) model

The AR(p) model:

A. can be represented as follows: $y_t = \beta_0 + \beta_1 x_t + \beta_p y_{t-p} + u_t$

B. is defined as $y_t = \beta_0 + \beta_p y_{t-p} + u_t$

C. can be written as $y_t = \beta_0 + \beta_1 y_{t-1} + u_{t-p}$

D. represents y_t as a linear function of p of its lagged values

SOLUTION

Answer: D.

- AR(p) : $y_t = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + u_t$
- A – adds an exogenous x_t → that is an ADL, not an AR
- B – drops the intermediate lags
- C – lags the error, not the regressor

Q7 | First difference of log

The first difference of log y_t equals:

A. the difference between the lead and the lag of y

B. approximately the growth rate of y when the growth rate is small

C. the growth rate of y exactly

D. the first difference of y

Q7 | First difference of log

The first difference of log y_t equals:

A. the difference between the lead and the lag of y

B. approximately the growth rate of y when the growth rate is small

C. the growth rate of y exactly

D. the first difference of y

SOLUTION

Answer: B.

- $\Delta \log y_t = \log y_t - \log y_{t-1} = \log(y_t/y_{t-1}) \approx (y_t - y_{t-1})/y_{t-1}$
- The approximation is good for **small** growth rates (say, < 10%); it deteriorates as the growth rate grows

Q8 | What IV fixes

The IV estimator can be used to potentially eliminate bias resulting from:

A. heteroskedasticity

B. serial correlation

C. multicollinearity

D. errors in variables

Q8 | What IV fixes

The IV estimator can be used to potentially eliminate bias resulting from:

A. heteroskedasticity

B. serial correlation

C. multicollinearity

D. errors in variables

SOLUTION

Answer: D.

- IV fixes problems that make $\text{cov}(x, u) \neq 0$: **omitted variables**, **simultaneous causality**, and **measurement error** (errors in variables)
- A and B are SE problems, not bias problems (fix with robust / HAC SEs)
- C is a precision problem, not a bias problem

Q9 | The random walk

The random walk model is an example of a:

A. stochastic trend model

B. deterministic trend model

C. stationary model

D. binomial model

Q9 | The random walk

The random walk model is an example of a:

A. stochastic trend model

B. deterministic trend model

C. stationary model

D. binomial model

SOLUTION

Answer: A.

- $y_t = y_{t-1} + u_t$ → the level depends on the **cumulated history of shocks**: $y_t = y_0 + \sum_{s=1}^{t} u_s$
- The trend is random (stochastic), not a deterministic function of t
- It has a unit root → non-stationary

Q10 | IBM volume and price

You explain the number of IBM shares traded per day in 2005 using the closing price of the share. This is an example of:

A. invalid inference due to a small sample size

B. simultaneous causality

C. sample selection bias since you should analyse more than one stock

D. a situation where homoskedasticity-only standard errors should be used since you only analyse one company

Q10 | IBM volume and price

You explain the number of IBM shares traded per day in 2005 using the closing price of the share. This is an example of:

A. invalid inference due to a small sample size

B. simultaneous causality

C. sample selection bias since you should analyse more than one stock

D. a situation where homoskedasticity-only standard errors should be used since you only analyse one company

SOLUTION

Answer: B.

- Volume and price are **jointly determined by supply and demand** on the same day
- A – 250 trading days is not a small sample
- C – Restricting to one stock is a **scope** decision, not selection on the dependent variable
- D – Sample size has nothing to do with the choice between robust and homoskedastic SEs

Essay Questions

Q11 | Simultaneity & IV

Consider the population regression model relating y_i and x_i:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$
$$x_i = \alpha_0 + \alpha_1 y_i + v_i$$

where z_i is a valid instrument for x_i.

a) Show OLS is inconsistent

Q11 | Simultaneity & IV

Consider the population regression model relating y_i and x_i:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$
$$x_i = \alpha_0 + \alpha_1 y_i + v_i$$

where z_i is a valid instrument for x_i.

a) Show OLS is inconsistent

SOLUTION

- OLS gives $\hat{\beta}_1^{OLS} = \text{cov}(x, y)/\text{var}(x)$, with probability limit

$$\lim \hat{\beta}_1^{OLS} \xrightarrow{p} \beta_1 + \frac{\text{cov}(x, u)}{\text{var}(x)}.$$

- Compute $\text{cov}(x, u)$ using the second equation:

$$\begin{aligned}\text{cov}(x, u) &= \text{cov}(\alpha_0 + \alpha_1 y + v, u) \\ &= \alpha_1 \text{ cov}(y, u) + \text{cov}(v, u).\end{aligned}$$

- And:

$$\begin{aligned}\text{cov}(y, u) &= \text{cov}(\beta_0 + \beta_1 x + u, u) \\ &= \beta_1 \text{ cov}(x, u) + \text{var}(u).\end{aligned}$$

Q11 | Simultaneity & IV

Q11 | Simultaneity & IV

- Substituting (and assuming $\text{cov}(v, u) = 0$):

$$\text{cov}(x, u) = \alpha_1[\beta_1 \text{ cov}(x, u) + \text{var}(u)]$$

$$\text{cov}(x, u) = \frac{\alpha_1 \text{ var}(u)}{1 - \alpha_1\beta_1} \neq 0,$$

  as long as $\alpha_1 \neq 0$ and $\alpha_1\beta_1 \neq 1$.

- Therefore:

$$\lim \hat{\beta}_1^{\text{OLS}} \xrightarrow{p} \beta_1 + \delta \neq \beta_1,$$

$$\delta = \frac{\alpha_1 \text{ var}(u)}{(1 - \alpha_1\beta_1) \text{ var}(x)}$$

- The OLS estimator is **biased and inconsistent** whenever the reverse-causality channel is active ($\alpha_1 \neq 0$).

Q11 | Simultaneity & IV

- Substituting (and assuming $\text{cov}(v, u) = 0$):

$$\text{cov}(x, u) = \alpha_1[\beta_1 \text{ cov}(x, u) + \text{var}(u)]$$

$$\text{cov}(x, u) = \frac{\alpha_1 \text{ var}(u)}{1 - \alpha_1\beta_1} \neq 0,$$

  as long as $\alpha_1 \neq 0$ and $\alpha_1\beta_1 \neq 1$.

- Therefore:

$$\lim \hat{\beta}_1^{\text{OLS}} \xrightarrow{p} \beta_1 + \delta \neq \beta_1,$$

$$\delta = \frac{\alpha_1 \text{ var}(u)}{(1 - \alpha_1\beta_1) \text{ var}(x)}$$

- The OLS estimator is **biased and inconsistent** whenever the reverse-causality channel is active ($\alpha_1 \neq 0$).

b) How is this type of bias called?

A. sample selection bias

B. simultaneous causality bias

C. omitted variable bias

D. measurement error bias

Q11 | Simultaneity & IV

- Substituting (and assuming $\text{cov}(v, u) = 0$):

$$\text{cov}(x, u) = \alpha_1[\beta_1 \text{ cov}(x, u) + \text{var}(u)]$$

$$\text{cov}(x, u) = \frac{\alpha_1 \text{ var}(u)}{1 - \alpha_1\beta_1} \neq 0,$$

  as long as $\alpha_1 \neq 0$ and $\alpha_1\beta_1 \neq 1$.

- Therefore:

$$\text{lim } \hat{\beta}_1^{OLS} \xrightarrow{p} \beta_1 + \delta \neq \beta_1,$$

$$\delta = \frac{\alpha_1 \text{ var}(u)}{(1 - \alpha_1\beta_1) \text{ var}(x)}$$

- The OLS estimator is **biased and inconsistent** whenever the reverse-causality channel is active ($\alpha_1 \neq 0$).

b) How is this type of bias called?

A. sample selection bias

B. simultaneous causality bias

C. omitted variable bias

D. measurement error bias

Answer: B.

x causes *y* **and** *y* causes *x* → each one is correlated with the other equation's error.

Q11 | Simultaneity & IV

c) How would you construct a consistent estimator for β_1?

Q11 | Simultaneity & IV

c) How would you construct a consistent estimator for β_1?

SOLUTION

Use **Two-Stage Least Squares** with z_i as instrument:

1. **First stage:** regress x_i on z_i (and any exogenous controls):

$$x_i = \pi_0 + \pi_1 z_i + e_i, \quad \rightarrow \quad \hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$$

2. **Second stage:** regress y_i on $\hat{x}_i$:

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \tilde{u}_i$$

Q11 | Simultaneity & IV

c) How would you construct a consistent estimator for β_1?

SOLUTION

Use **Two-Stage Least Squares** with z_i as instrument:

1. **First stage:** regress x_i on z_i (and any exogenous controls):

$$x_i = \pi_0 + \pi_1 z_i + e_i, \quad \rightarrow \quad \hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$$

2. **Second stage:** regress y_i on $\hat{x}_i$:

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \tilde{u}_i$$

- Equivalently, in closed form:

$$\hat{\beta}_1^{\text{TSLS}} = \frac{\text{cov}(z, y)}{\text{cov}(z, x)}.$$

- z replaces the **endogenous** variation in x with the **exogenous** variation predicted by z
- Consistent provided z is **relevant** ($\text{cov}(z, x) \neq 0$) and **exogenous** ($\text{cov}(z, u) = 0$)

Q11 | Simultaneity & IV

d) You estimate the model:

$$x_i = \gamma_0 + \gamma_1 z_i + e_i$$

and find $\text{cov}(z_i, e_i) \neq 0$. Is it a problem? If yes, how do you solve it? If no, why not?

Q11 | Simultaneity & IV

d) You estimate the model:

$$x_i = \gamma_0 + \gamma_1 z_i + e_i$$

and find $\text{cov}(z_i, e_i) \neq 0$. Is it a problem? If yes, how do you solve it? If no, why not?

SOLUTION

- **No – and in fact this cannot happen.**
- OLS residuals are **orthogonal to every regressor** in the sample by construction:

$$\sum_{i=1}^{n} z_i \hat{e}_i = 0 \quad \rightarrow \quad \text{cov}(z_i, \hat{e}_i) = 0.$$

Q11 | Simultaneity & IV

d) You estimate the model:

$$x_i = \gamma_0 + \gamma_1 z_i + e_i$$

and find $\text{cov}(z_i, e_i) \neq 0$. Is it a problem? If yes, how do you solve it? If no, why not?

SOLUTION

- **No – and in fact this cannot happen.**
- OLS residuals are **orthogonal to every regressor** in the sample by construction:

$$\sum_{i=1}^{n} z_i \hat{e}_i = 0 \quad \rightarrow \quad \text{cov}(z_i, \hat{e}_i) = 0.$$

- If you observe a non-zero sample covariance, you have a computational error
- At the **population** level, the **linear projection** of x on z is the best linear predictor — the line $\gamma_0 + \gamma_1 z$ that minimises $\mathbb{E}\left[(x - \gamma_0 - \gamma_1 z)^2\right]$, with coefficients

$$\gamma_1 = \text{cov}(z, x)/\text{var}(z), \quad \gamma_0 = \mathbb{E}[x] - \gamma_1 \mathbb{E}[z]$$

- The projection error $e = x - \gamma_0 - \gamma_1 z$ is uncorrelated with z **by construction** (it's the FOC of the minimisation)
- No causal or structural claim is made → no exogeneity assumption needed. The IV exogeneity assumption applies to the **structural** error u, not to e

Q11 | Simultaneity & IV

WHY EXOGENEITY IS NOT NEEDED IN THE 1ST STAGE TSLS?

- Imagine x has another determinant w:

$$x = \gamma_0 + \gamma_1 z + \gamma_2 w + v,$$
$$\text{cov}(z, v) = \text{cov}(w, v) = 0$$

- Omit w and project x on z:

$$x = \gamma_0^{\text{proj}} + \gamma_1^{\text{proj}} z + e$$

$$\gamma_1^{\text{proj}} = \frac{\text{cov}(z, x)}{\text{var}(z)} = \gamma_1 + \gamma_2 \cdot \frac{\text{cov}(z, w)}{\text{var}(z)}$$

- The projection error:

$$e = x - \gamma_0^{\text{proj}} - \gamma_1^{\text{proj}} z =$$
$$= \left(\gamma_0 - \gamma_0^{\text{proj}}\right) + \left(\gamma_1 - \gamma_1^{\text{proj}}\right) z + \gamma_2 w + v$$

contains w. Yet:

$$\text{cov}(z, e) = -\gamma_2 \cdot \text{cov}(z, w) +$$
$$+\gamma_2 \cdot \text{cov}(z, w) = 0$$

by construction.

Q11 | Simultaneity & IV

e) Which assumption(s) can you test from the output shown in Table 1? What is your conclusion?

Dep. var.	*x* (1)	*y* (2)	*y* (3)
Constant	0.00*** (0.00)	−0.22*** (0.013)	−0.22*** (0.015)
z	0.04 (0.023)		
$\hat{x}$		1.38*** (0.002)	
x			1.38*** (0.003)
N	1000	1000	1000

Q11 | Simultaneity & IV

Q11 | Simultaneity & IV

SOLUTION

- Column (1) is the first stage of TSLS:

$$x = \pi_0 + \pi_1 z + e$$

- We can test **instrument relevance**
- Coefficient on z: $\hat{\pi}_1 = 0.04$, SE = 0.023
- $t = 0.04/0.023 \approx 1.74$ → $F = t^2 \approx 3.0$
- **Rule of thumb:** $F > 10$ is the threshold for a strong instrument
- $F \approx 3 \ll 10$ → **the instrument is weak** → the TSLS estimate in column (2) is unreliable, and its inference is invalid

Q11 | Simultaneity & IV

SOLUTION

- Column (1) is the first stage of TSLS:

$$x = \pi_0 + \pi_1 z + e$$

- We can test **instrument relevance**
- Coefficient on z: $\hat{\pi}_1 = 0.04$, SE = 0.023
- $t = 0.04/0.023 \approx 1.74$ → $F = t^2 \approx 3.0$
- **Rule of thumb:** $F > 10$ is the threshold for a strong instrument
- $F \approx 3 \ll 10$ → **the instrument is weak** → the TSLS estimate in column (2) is unreliable, and its inference is invalid

f) Why are the standard errors from model (2) and (3) different? Give your best guess.

Q11 | Simultaneity & IV

SOLUTION

- Column (1) is the first stage of TSLS:

$$x = \pi_0 + \pi_1 z + e$$

- We can test **instrument relevance**
- Coefficient on z: $\hat{\pi}_1 = 0.04$, SE = 0.023
- $t = 0.04/0.023 \approx 1.74$ → $F = t^2 \approx 3.0$
- **Rule of thumb:** $F > 10$ is the threshold for a strong instrument
- $F \approx 3 \ll 10$ → **the instrument is weak** → the TSLS estimate in column (2) is unreliable, and its inference is invalid

f) Why are the standard errors from model (2) and (3) different? Give your best guess.

SOLUTION

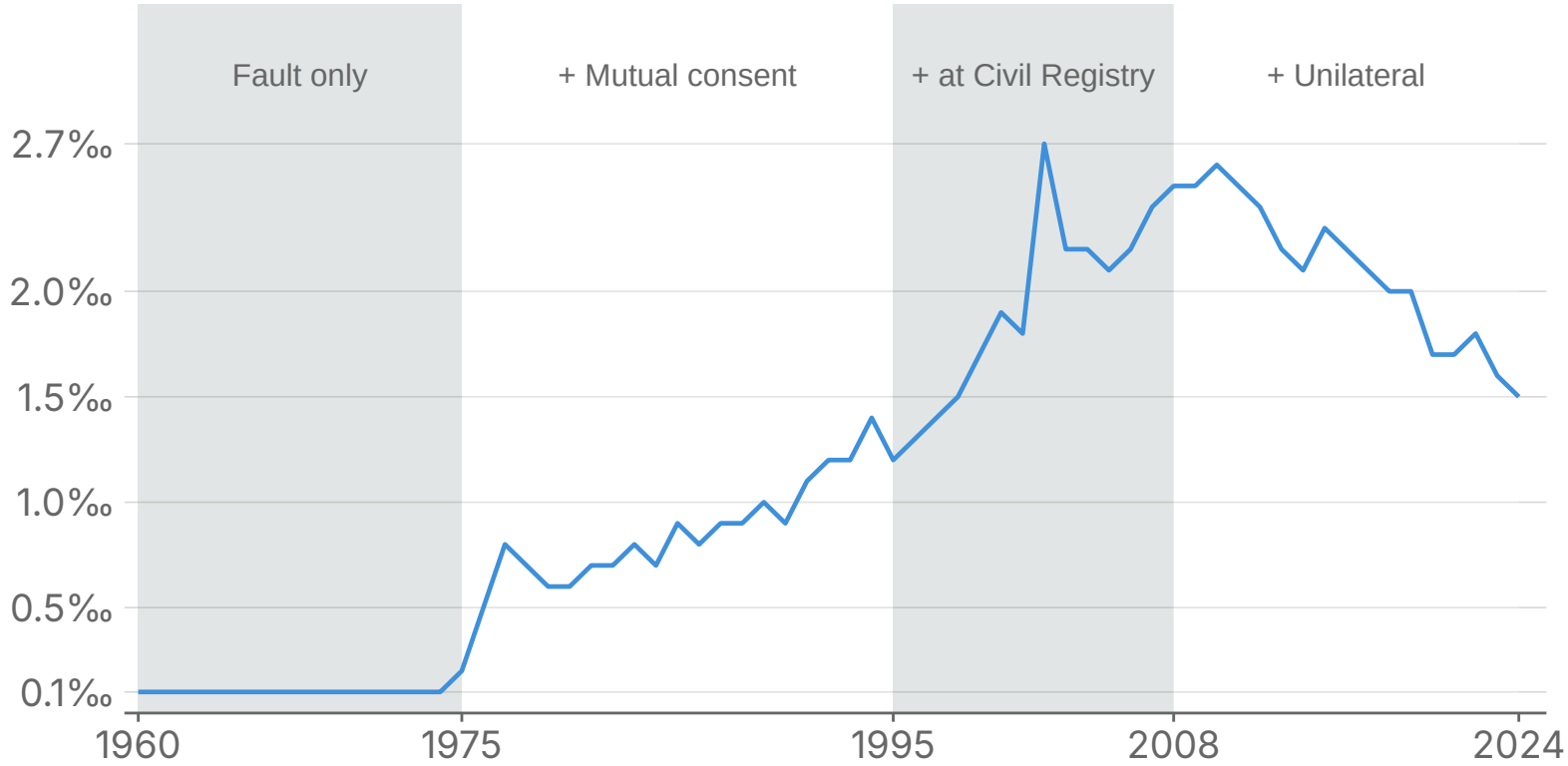
- The second stage OLS ignores first-stage estimation uncertainty
- Proper TSLS software inflates the SE to account for the fact that $\hat{x}$ is estimated, not observed
- With a **weak** instrument (part e), the correct TSLS SE should be **larger** than the OLS SE – not smaller

Q12 | Divorce rate & stationarity

Figure 1 shows the divorce rate in Portugal, per thousand people. The figure also shows 4 different periods:

- **Period 1 (1961–1974):** divorce available only for civil marriages on fault grounds; Catholic marriages were indissoluble under the 1940 Concordat.
- **Period 2 (1975–1994):** Law 4/76 introduced mutual-consent divorce after one year of *de-facto* separation and broadened fault grounds for contested cases.
- **Period 3 (1995–2007):** Law 84/95 allowed mutual-consent divorces to be processed cheaply at the civil registry and cut the required separation period; property-division rules were clarified, sharply reducing transaction costs.
- **Period 4 (2008–present):** Law 61/2008 created true unilateral no-fault divorce based on irretrievable breakdown; one spouse can file after one year of separation, and the other cannot block it.

Q12 | Divorce rate & stationarity



Source: PORDATA — Taxa bruta de divorcialidade (INE / DGPJ-MJ).

Note: This plot shows the number of divorces per thousand people, by year.

Q12 | Divorce rate & stationarity

Using this data, we ran three different models, shown in Table 2, where y_t is the rate of divorce in year t:

Dep. var.	y_t (1)	Δy_t (2)	$\Delta y_t - \Delta y_{t-1}$ (3)
Constant	0.0640** (0.0243)	0.0280 (0.0188)	0.0280 (0.0188)
y_{t-1}	0.9671*** (0.0255)		
Δy_{t-1}		0.3682 (0.2432)	−0.6318** (0.2432)
N	62	61	61
R^2	0.959	0.134	0.313

Note: model (3) is just model (2) with Δy_{t-1} subtracted from both sides – it is the **Dickey–Fuller** form for testing a unit root in Δy_t.

Q12 | Divorce rate & stationarity

a) Using Table 2, what is the first-order autocorrelation of Δy_t?

Q12 | Divorce rate & stationarity

a) Using Table 2, what is the first-order autocorrelation of Δy_t?

b) Is y_t in model (1) stationary? Why or why not?

SOLUTION

- In an AR(1), the slope coefficient **is** the first-order autocorrelation: $\text{corr}(\Delta y_t, \Delta y_{t-1}) = 0.3682$.
- In a stationary AR(1) $y_t = \beta_0 + \beta_1 y_{t-1} + u_t$ with $\text{cov}(u_t, y_{t-1}) = 0$:
  - $\text{var}(y_t) = \text{var}(y_{t-1}) \equiv \gamma_0$
  - $\text{cov}(y_t, y_{t-1}) = \beta_1 \cdot \text{var}(y_{t-1}) = \beta_1 \gamma_0$
  - $\rho_1 = \text{cov}(y_t, y_{t-1})/\gamma_0 = \beta_1$

Q12 | Divorce rate & stationarity

a) Using Table 2, what is the first-order autocorrelation of Δy_t?

SOLUTION

- In an AR(1), the slope coefficient **is** the first-order autocorrelation: $\text{corr}(\Delta y_t, \Delta y_{t-1}) = 0.3682$.
- In a stationary AR(1) $y_t = \beta_0 + \beta_1 y_{t-1} + u_t$ with $\text{cov}(u_t, y_{t-1}) = 0$:
  - $\text{var}(y_t) = \text{var}(y_{t-1}) \equiv \gamma_0$
  - $\text{cov}(y_t, y_{t-1}) = \beta_1 \cdot \text{var}(y_{t-1}) = \beta_1 \gamma_0$
  - $\rho_1 = \text{cov}(y_t, y_{t-1})/\gamma_0 = \beta_1$

b) Is y_t in model (1) stationary? Why or why not?

SOLUTION

$$y_t = 0.0640 + 0.9671\, y_{t-1} + e_t$$

Dickey–Fuller test:

$$\tau = \frac{0.9671 - 1}{0.0255} = -1.29.$$

- **Critical values** (S&W, with constant): −2.86 at 5%, −2.57 at 10%.
- $-1.29 > -2.86$ → we **cannot reject** the unit-root null → y_t **is non-stationary**.

Q12 | Divorce rate & stationarity

c) Is Δy_t in model (2) stationary? Why/ why not? Does it have a unit root?

Q12 | Divorce rate & stationarity

c) Is Δy_t in model (2) stationary? Why/ why not? Does it have a unit root?

SOLUTION

- Model (2) coefficient is 0.3682 – far below 1 – so heuristically Δy_t looks stationary
- **DF test:** rearrange model (2) by subtracting Δy_{t-1} from both sides. This is exactly **model (3)**:

$$\Delta y_t - \Delta y_{t-1} = 0.0280 - 0.6318 \Delta y_{t-1} + e_t.$$

$$\tau = -\frac{0.6318}{0.2432} = -2.60.$$

Q12 | Divorce rate & stationarity

c) Is Δy_t in model (2) stationary? Why/ why not? Does it have a unit root?

SOLUTION

- Model (2) coefficient is 0.3682 – far below 1 – so heuristically Δy_t looks stationary
- **DF test:** rearrange model (2) by subtracting Δy_{t-1} from both sides. This is exactly **model (3)**:

$$\Delta y_t - \Delta y_{t-1} = 0.0280 - 0.6318\Delta y_{t-1} + e_t.$$

$$\tau = -\frac{0.6318}{0.2432} = -2.60.$$

- Compare to DF critical values (with constant): –2.86 at 5%, –2.57 at 10%.
  - At 5%: –2.60 > –2.86 → **cannot reject** the unit-root null
  - At 10%: –2.60 < –2.57 → reject
- **Borderline.** The coefficient 0.3682 is well below 1, so practically Δy_t behaves like a stationary process, but the formal evidence against a unit root is weak → safer to work with $\Delta^2 y_t$

Q12 | Divorce rate & stationarity

d) If your goal is to predict y_t, should you delete the years before 1975?

Q12 | Divorce rate & stationarity

d) If your goal is to predict y_t, should you delete the years before 1975?

e) What is $\mathbb{E}[\Delta y_t]$?

SOLUTION

Yes, for two reasons:

1. Pre-1975 the divorce rate is near zero: Catholic marriages were indissoluble → **structural break**
2. For **forecasting** we want a model that reflects how the series evolves **today**

Q12 | Divorce rate & stationarity

d) If your goal is to predict y_t, should you delete the years before 1975?

SOLUTION

Yes, for two reasons:

1. Pre-1975 the divorce rate is near zero: Catholic marriages were indissoluble → **structural break**
2. For **forecasting** we want a model that reflects how the series evolves **today**

e) What is $\mathbb{E}[\Delta y_t]$?

SOLUTION

Assume Δy_t is stationary so that $\mathbb{E}[\Delta y_t] = \mathbb{E}[\Delta y_{t-1}] = \mu$. Taking expectations in model (2):

$$\mu = 0.0280 + 0.3682\mu \iff$$

$$\mu = \frac{0.0280}{1 - 0.3682} \approx 0.0443$$

Under the post-1975 regime, the divorce rate rises by **about 0.044 per thousand each year** on average.

f) Two consequences of non-stationarity (4 pts)

f) Provide two consequences of non-stationarity.

f) Two consequences of non-stationarity (4 pts)

f) Provide two consequences of non-stationarity.

SOLUTION

1. **Forecasts are unreliable.** The past distribution of y_t no longer describes the future, so out-of-sample predictions and their intervals are misleading
2. **Standard inference breaks down.** t and F statistics no longer follow their standard distributions → wrong p-values, wrong confidence intervals

f) Two consequences of non-stationarity (4 pts)

g) Assume the residuals are heteroskedasticity. If we estimated the models with homoskedastic only standard errors, we would have a problem of:

A. sample selection bias

B. consistency

C. external validity

D. internal validity

f) Two consequences of non-stationarity (4 pts)

g) Assume the residuals are heteroskedasticity. If we estimated the models with homoskedastic only standard errors, we would have a problem of:

A. sample selection bias

B. consistency

C. external validity

D. internal validity

SOLUTION

Answer: D – internal validity.

- Coefficients are still **consistent** (B is wrong)
- But the homoskedasticity-only SEs are **wrong** → *t*-stats, CIs, and *p*-values for **this** sample are invalid → internal validity fails
- Has nothing to do with sample selection (A) or generalisability to other populations (C)

Q13 | Forecasting US unemployment

You set out to forecast the unemployment rate in the United States (`UrateUS`), using quarterly data from 1960, first quarter, to 1999, fourth quarter.

a) The table below gives changes in the U.S. aggregate unemployment rate for the period 1999:I–2000:I together with levels of the current and lagged unemployment rates for 1999:I. Some numbers are omitted (–). What are the values x_1, x_2 and x_3? *Round to one decimal place.*

Q13 | Forecasting US unemployment

Quarter	Urate	Lag	ΔUrate
1999:I	4.3	4.4	–0.1
1999:II	–	4.3	0.0
1999:III	–	x_1	–0.1
1999:IV	–	x_2	–0.1
2000:I	–	x_3	–0.1

Q13 | Forecasting US unemployment

Quarter	Urate	Lag	ΔUrate
1999:I	4.3	4.4	–0.1
1999:II	–	4.3	0.0
1999:III	–	x_1	–0.1
1999:IV	–	x_2	–0.1
2000:I	–	x_3	–0.1

SOLUTION

Each lag is the previous quarter's **level**; each level is "previous lag + change".

- 1999:II: level = 4.3 + 0.0 = 4.3
- x_1 = lag for 1999:III = level in 1999:II = **4.3**. Then 1999:III level = 4.3 – 0.1 = 4.2
- x_2 = lag for 1999:IV = level in 1999:III = **4.2**. Then 1999:IV level = 4.2 – 0.1 = 4.1
- x_3 = lag for 2000:I = level in 1999:IV = **4.1**. Then 2000:I level = 4.1 – 0.1 = 4.0
- $x_1 = 4.3, \quad x_2 = 4.2, \quad x_3 = 4.1.$

Q13 | Forecasting US unemployment

You estimate an AR(4):

$$\Delta u_t = -0.005 + 0.663\Delta u_{t-1} - 0.082\Delta u_{t-2} + 0.106\Delta u_{t-3} - 0.176\Delta u_{t-4} + e_t.$$

b) Forecast for the unemployment level in 2000:III using the AR(4). *Round to two decimal places.*

Q13 | Forecasting US unemployment

You estimate an AR(4):

$$\Delta u_t = -0.005 + 0.663\Delta u_{t-1} - 0.082\Delta u_{t-2} + 0.106\Delta u_{t-3} - 0.176\Delta u_{t-4} + e_t.$$

b) Forecast for the unemployment level in 2000:III using the AR(4). *Round to two decimal places.*

SOLUTION

We need to iterate the AR(4) **twice**: first to forecast $\Delta\hat{u}_{2000:II}$, then $\Delta\hat{u}_{2000:III}$.

Step 1 – forecast $\Delta\hat{u}_{2000:II}$ using $(\Delta u_{2000:I}, \Delta u_{1999:IV}, \Delta u_{1999:III}, \Delta u_{1999:II}) = (-0.1, -0.1, -0.1, 0.0)$:

$$\Delta\hat{u}_{2000:II} = -0.005 + 0.663(-0.1) - 0.082(-0.1) + 0.106(-0.1) - 0.176(0.0)$$

$$= -0.005 - 0.0663 + 0.0082 - 0.0106 + 0 = -0.0737.$$

Q13 | Forecasting US unemployment

Step 2: forecast $\Delta\hat{u}_{2000:III}$:

Using $(\Delta\hat{u}_{2000:II}, \Delta u_{2000:I}, \Delta u_{1999:IV}, \Delta u_{1999:III}) = (-0.0737, -0.1, -0.1, -0.1)$:

$$\Delta\hat{u}_{2000:III} = -0.005 + 0.663(-0.0737) - 0.082(-0.1) + 0.106(-0.1) - 0.176(-0.1)$$

$$= -0.005 - 0.04886 + 0.0082 - 0.0106 + 0.0176 = -0.0387.$$

Step 3: convert back to levels:

$$\hat{u}_{2000:II} = u_{2000:I} + \Delta\hat{u}_{2000:II} = 4.0 - 0.0737 = 3.9263$$

$$\hat{u}_{2000:III} = \hat{u}_{2000:II} + \Delta\hat{u}_{2000:III} = 3.9263 - 0.0387 = 3.8876$$

Then: $\hat{u}_{2000:III} \approx 3.89\%$.

Q13 | Forecasting US unemployment

c) What is the short-run effect (of one quarter) of a rise in the unemployment rate?

Q13 | Forecasting US unemployment

c) What is the short-run effect (of one quarter) of a rise in the unemployment rate?

SOLUTION

- The **one-quarter** impact of a shock: the coefficient on Δu_{t-1}, $\hat{\beta}_{\Delta u_{t-1}} = 0.663$
- A 1 pp increase in the **change** of unemployment in quarter $t - 1$ raises the **change** in t by 0.663 pp on average, holding the other lags fixed.

Q13 | Forecasting US unemployment

c) What is the short-run effect (of one quarter) of a rise in the unemployment rate?

SOLUTION

- The **one-quarter** impact of a shock: the coefficient on Δu_{t-1}, $\hat{\beta}_{\Delta u_{t-1}} = 0.663$
- A 1 pp increase in the **change** of unemployment in quarter $t - 1$ raises the **change** in t by 0.663 pp on average, holding the other lags fixed.

d) Information criteria and R^2 for different AR(p) are given:

p	BIC	AIC	$\hat{R}^2$
1	**0.158**	**0.1181**	0.393
2	0.185	0.125	0.397
3	0.217	0.138	0.400
4	0.218	0.1183	**0.416**

Q13 | Forecasting US unemployment

Which AR(p) model would you use for forecasting?

Q13 | Forecasting US unemployment

Which AR(p) model would you use for forecasting?

SOLUTION

- Both **BIC** and **AIC** are minimised at $p = 1$
- R^2 always (weakly) increases with more lags → never a model-selection criterion
- For forecasting, parsimony wins: **AR(1)**

Q14 | Funds rate & IV

You estimate the following quarterly regression for Portugal over the past 50 years:

$$\text{rate}_t = \gamma_0 + \gamma_1 \text{rate}_{t-1} + \beta_1 \text{inf}_{t-1} + \beta_2 \text{ur}_{t-1} + \beta_3 \text{pop}_{t-1} + e_t,$$

where rate_t is the government funds rate, inf_t inflation, ur_t unemployment, \$pop is the popularity of the government (approval rating −50%), and e_t is the error term.

Q14 | Funds rate & IV

a) If your goal is to measure the causal effect of pop_{t-1} on rate_t, do you have a problem? Explain.

Q14 | Funds rate & IV

a) If your goal is to measure the causal effect of pop_{t-1} on rate_t, do you have a problem? Explain.

SOLUTION

- **Yes – there is a simultaneity problem.**
- The government's funds rate **itself** moves the economy (output, unemployment, financial conditions) → it moves the government's **approval**
- So $\text{cov}(\text{pop}_{t-1}, e_t) \neq 0$
- OLS $\hat{\beta}_3$ is biased and inconsistent

b) Assume $\text{cov}(\text{pop}_{t-1}, e_t) \neq 0$. Suggest a valid instrument for pop_{t-1} and argue why it satisfies both relevance and exogeneity.

Q14 | Funds rate & IV

a) If your goal is to measure the causal effect of pop_{t-1} on $rate_t$, do you have a problem? Explain.

SOLUTION

- **Yes – there is a simultaneity problem.**
- The government's funds rate **itself** moves the economy (output, unemployment, financial conditions) → it moves the government's **approval**
- So $cov(pop_{t-1}, e_t) \neq 0$
- OLS $\hat{\beta}_3$ is biased and inconsistent

b) Assume $cov(pop_{t-1}, e_t) \neq 0$. Suggest a valid instrument for pop_{t-1} and argue why it satisfies both relevance and exogeneity.

SOLUTION

- The **resignation of a minister following a personal scandal**
- **Relevance:** scandals drop government's approval → first-stage F should be large
- **Exogeneity:** a personal scandal is plausibly orthogonal to the macro forces → it influences $rate_t$ only through approval

Q14 | Funds rate & IV

c) Assume z_i is a valid instrument. Write the first-stage equation.

Q14 | Funds rate & IV

c) Assume z_i is a valid instrument. Write the first-stage equation.

SOLUTION

$$\text{pop}_{t-1} = \pi_0 + \pi_1 z_{t-1} + \pi_2 \text{rate}_{t-1} + \pi_3 \text{inf}_{t-1} + \pi_4 \text{ur}_{t-1} + v_{t-1}.$$

- **All exogenous regressors from the structural equation must also be on the right** (otherwise the second stage is misspecified)
- Test $H_0 : \pi_1 = 0$ with an F-statistic → need $F > 10$ for a strong instrument

Q14 | Funds rate & IV

d) What is **not** problematic for forecasting?

A. the functional form of the regression being incorrect

B. a change in the variance of rate since the 90's

C. the regression having a low explanatory power

D. the presence of the endogenous variable pop

Q14 | Funds rate & IV

d) What is **not** problematic for forecasting?

A. the functional form of the regression being incorrect

B. a change in the variance of rate since the 90's

C. the regression having a low explanatory power

D. the presence of the endogenous variable pop

SOLUTION

Answer: D.

- **Forecasting** only requires that the regressors **predict** y – it does **not** require causal identification
- A (wrong functional form) → biased forecasts
- B (variance change) → structural break → external validity fails
- C (low R^2) → wide forecast intervals, low accuracy
- D matters only if we want to **interpret** β_3 as a causal effect