Econometrics

Practical Session 19

Nonstationarity: Structural Breaks

Ricardo Gouveia-Mendes
rgouveiamendes@ucp.pt

Spring 2025-26

Católica-Lisbon School of Business and Economics

Theoretical Wrap-up

A Second Type of Nonstationarity: Structural Breaks

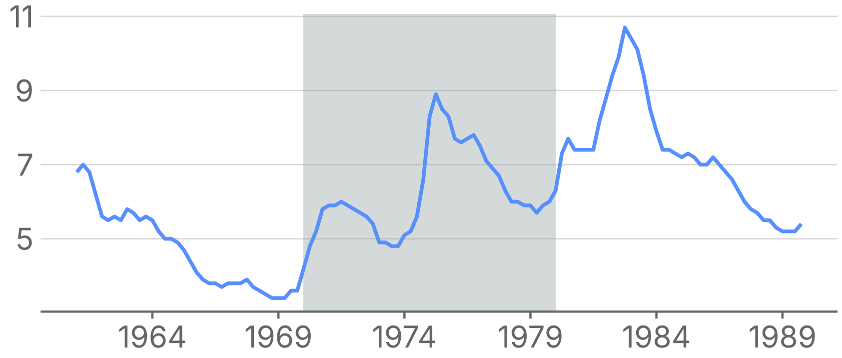
- So far, nonstationarity meant **unit roots**
- An equally important type: **the regression coefficients change over time**
- **Why does this matter for forecasting?**
  - Our models are trained on **historical data → estimates no longer apply**
  - This is an **external validity** problem: the model doesn't generalize beyond the training period

A Second Type of Nonstationarity: Structural Breaks

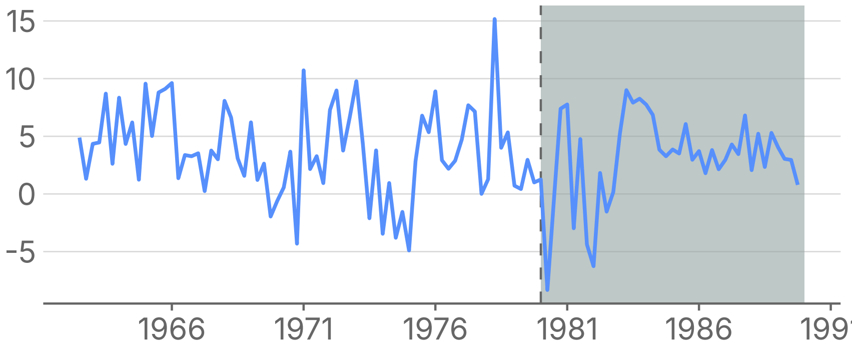
- **Examples:**
  - The Phillips curve (inflation-unemployment) appeared stable in the 1960s but broke down in the 1970s
  - The GDP-yield spread relationship was disrupted around 1980 (Volcker disinflation)
  - Financial models estimated before 2008 missed the crisis
- **Informal diagnostic:** do pseudo out-of-sample forecasts track actual values at the end of the sample?

A Second Type of Nonstationarity: Structural Breaks

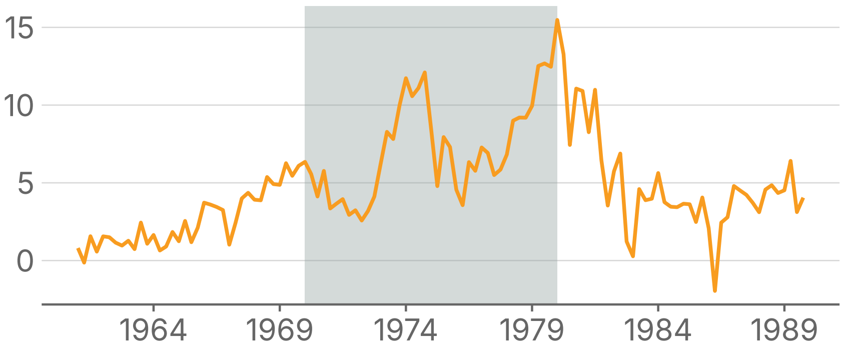
US Unemployment Rate
(%)



US Real GDP Growth
(%, annualized)

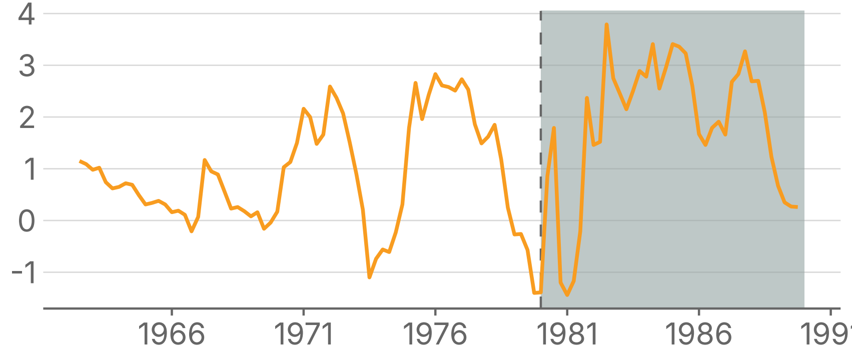


US Inflation Rate
(%, annualized)



Source: FRED, series UNRATE, CPIAUCSL.

US Yield Spread: 10Y – 3M
(pp)



Source: FRED. Dashed line: 1980Q1 (Volcker).

Case I: Known Break Date — The Chow Test

- Suppose a break at date τ is **known in advance** (e.g., a policy change, a law, a major event)
- **Strategy:** estimate a **fully interacted regression** → allow all coefficients to differ before and after τ:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \gamma_0 D_{t(\tau)} + \gamma_1 [D_{t(\tau)} \times Y_{t-1}] + \gamma_2 [D_{t(\tau)} \times X_{t-1}] + u_t$$

  where $D_{t(\tau)} = 1$ if $t \geq \tau$, else 0
- **Chow test:** F-test of $H_0 : \gamma_0 = \gamma_1 = \gamma_2 = 0$ → if we reject: the regression function changed at τ → break confirmed
- **Problem:** in practice we often do **not** know τ in advance

Case II: Unknown Break Date — The QLR Test

- **Data snooping:** plot the data, notice a kink in 1980, and run a Chow test
  - A Chow test at a **pre-specified τ** has a 5% false-rejection rate under the null → $F(\tau)$ follows a standard F distribution
  - But you **implicitly picked the one where F is highest** → you are **testing the maximum**, not a single statistic
  - The maximum of many **F-statistics is larger than any single one**, even with **no break** → so the **standard critical values rejects far more** than 5% of the time
- **Solution:** run a Chow $F(\tau)$ at **every candidate date** in the trimmed range, then **report the maximum** and use **higher critical values**

Case II: Unknown Break Date — The QLR Test

- The **Quandt Likelihood Ratio (QLR)** (*a.k.a.* sup-Wald) statistic:

$$\text{QLR} = \max_{\tau \in [\tau_0, \tau_1]} F(\tau)$$

  where $F(\tau)$ is the Chow F-statistic for a break at date τ

- **The trimmings** guarantee a minimum **usable sub-samples before or after the candidate date** if that date is close to beginning or the end of the sample

QLR Critical Values

- The QLR has a non-standard distribution: critical values depend on the **number of restrictions q and the trimming fraction**
- Critical values from S&W Table 15.5 (15% trimming):

Restrictions q	10%	5%	1%
1	7.12	8.68	12.16
2	5.00	5.86	7.78
3	4.09	4.71	6.02
4	3.59	4.09	5.12
5	3.26	3.66	4.53

QLR Critical Values

Restrictions q	10%	5%	1%
6	3.02	3.37	4.12
7	2.84	3.15	3.82
8	2.69	2.98	3.57
9	2.58	2.84	3.38
10	2.48	2.71	3.23

Informal Diagnostic: POOS Forecasting

- The **QLR test has low power** toward the **end of the sample**, because of trimming
- But end-of-sample breaks are often the most practically important!
- **Pseudo out-of-sample (POOS) forecasting** as a complement:
  1. Estimate the model on a growing training window (recursive estimation)
  2. At each t, forecast $t + 1$ using only data up to t
  3. Plot actual vs. forecast: does the model track recently?

Informal Diagnostic: POOS Forecasting

- **What to look for:**
  - If forecasts systematically under- or over-shoot in recent observations → likely end-of-sample break
  - If errors in recent periods are much larger than earlier → model may have changed
- This is an **informal diagnostic**, not a formal test → complement it with QLR

Exercises

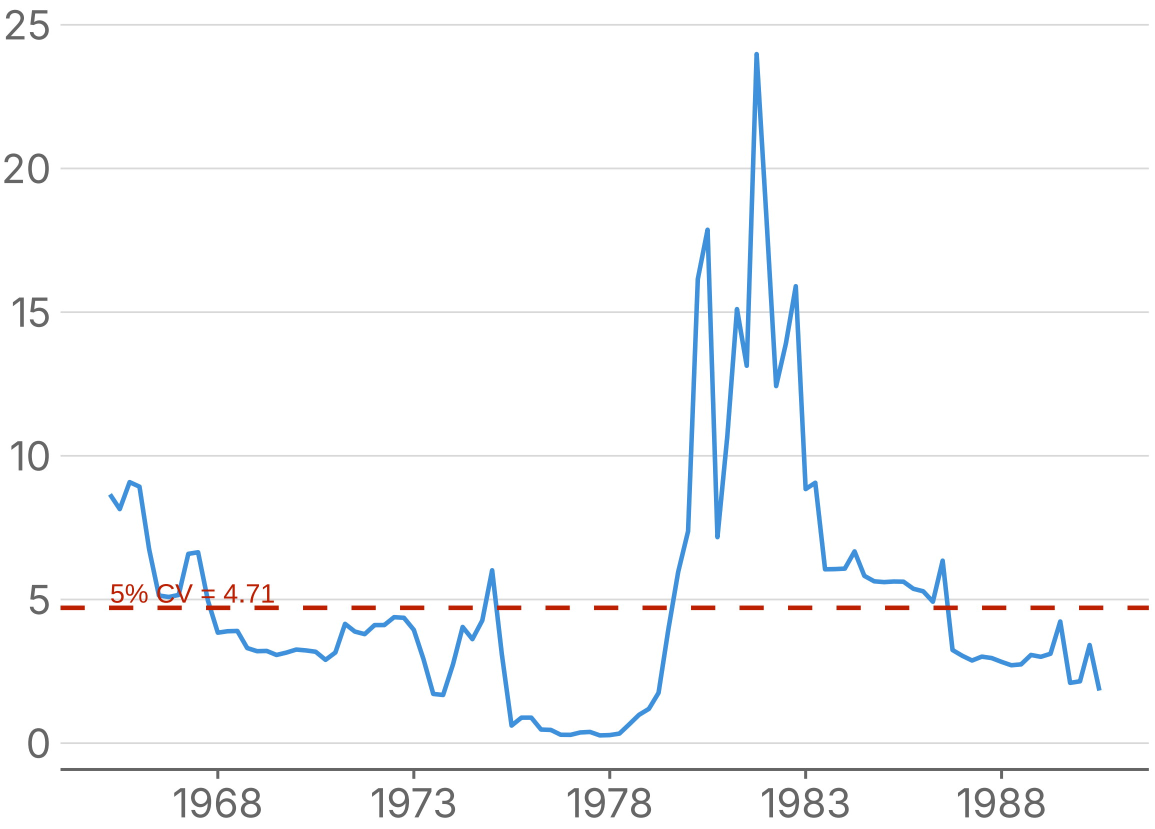
Exercise 1 | A Clean Structural Break: US Inflation (1960–1995)

Using quarterly CPI data (FRED: CPIAUCSL), we fit an AR(2) to annualised US inflation over 1960Q1–1995Q4 and apply the QLR test.

The Chow $F(\tau)$ statistics over the trimmed sample are plotted below. The dashed line marks the 5% critical value for $q = 3$ (intercept + 2 lags).

QLR: Chow F-statistic

AR(2) Inflation Model (1960–1995)



Is there evidence of a structural break? Does the plot show a **single clean break** or **pervasive instability**? At what approximate date does the maximum occur, and what economic event explains it?

Exercise 1 | A Clean Structural Break: US Inflation (1960–1995)

KEY TAKEAWAYS

- The F-statistic peaks sharply around **1979–1982** and stays **below the 5% critical value** (4.71 for $q = 3$) for most other dates → this is a **single, identifiable break**
- **Economic explanation:** the Volcker disinflation — Fed Chair Volcker raised the federal funds rate above 20% to crush inflation expectations; mean inflation fell from ~10% to ~3–4%
- **Shape:** one dominant spike with the F-statistic crossing the critical value only near the break → the QLR cleanly identifies one transition in inflation dynamics

Exercise 2 | Testing for a Structural Break in the ADL(2,2)

Using `us_quarterly.csv` and the ADL(2,2) model for GDP growth and the term spread (estimated in Session 16), we test whether the GDP–spread relationship has been stable over time.

a) The ADL(2,2) is estimated on the full sample (1962Q3–2026Q1). The QLR test with 15% trimming gives:

```
supF test

data:  GR ~ GR_l1 + GR_l2 + Spread_l1 + Spread_l2
sup.F = 33.2,   p-value < 0.001
```

The 5% critical value for $q = 5$ from S&W Table 15.5 is **3.66**. Is there evidence of a structural break? What does a very small p-value tell you here?

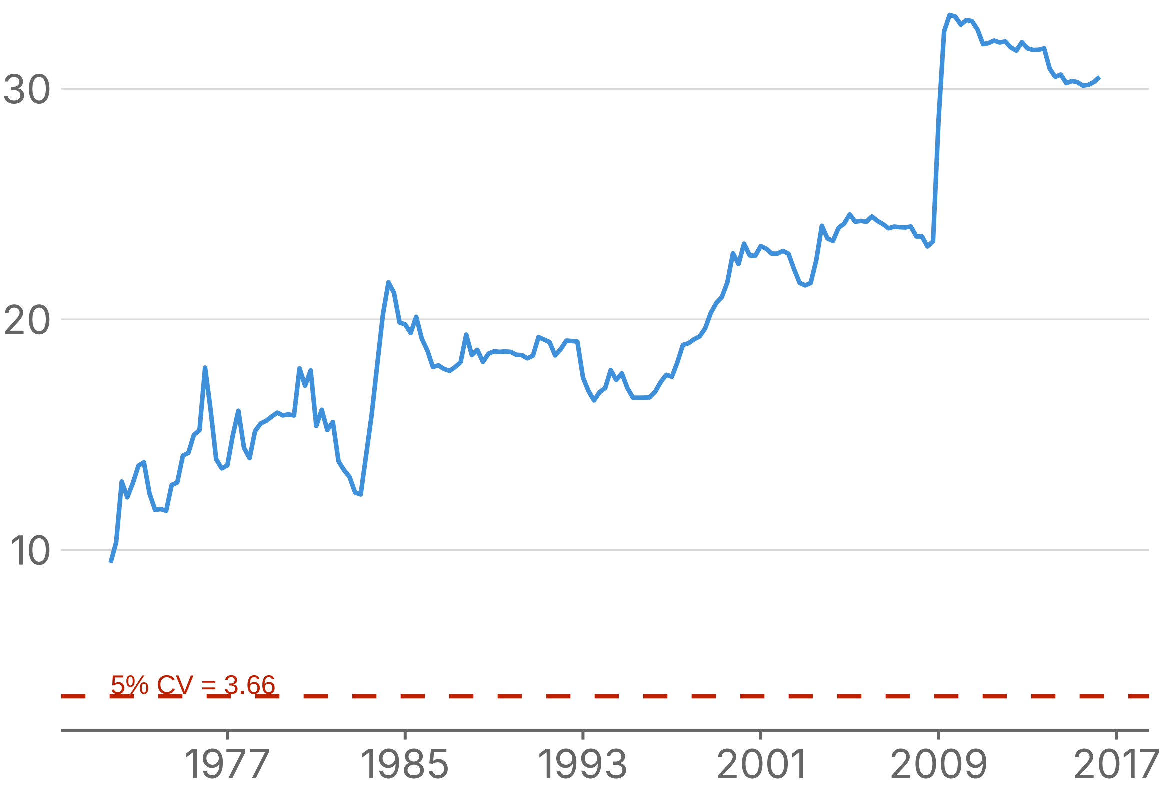
Exercise 2 | Testing for a Structural Break in the ADL(2,2)

KEY TAKEAWAYS

- QLR = 33.2 $\gg$ 3.66 (5% CV for $q = 5$) and $p < 0.001$: **reject stability overwhelmingly → the model is not stable** across the full sample
- **Conclusion:** there is strong evidence that the GDP–spread relationship changed at some point between 1962 and 2025

QLR: Chow F-statistic

ADL(2,2) GDP Growth Model



b) The sequence of Chow $F(\tau)$ statistics over the trimmed sample is plotted on the left. The dashed line marks the 5% critical value. At what approximate date does the maximum occur? What economic event might explain a break near that date?

Exercise 2 | Testing for a Structural Break in the ADL(2,2)

KEY TAKEAWAYS

- Every $F(\tau)$ in the trimmed range exceeds the 5% CV (3.66) → instability is **pervasive**
- Two dominant peaks stand out:
  - Around **1979–1981**: **Volcker disinflation** broke the normal spread–growth relationship
  - Around **2008–2009**: financial crisis and **zero-lower-bound policy** → the yield curve lost its predictive content when short rates were pinned near zero for years
- The maximum F falls near **2009Q3**, but since the **entire** sequence exceeds the critical value, no single split date resolves the instability → **multiple breaks**

Exercise 2 | Testing for a Structural Break in the ADL(2,2)

c) The ADL(2,2) is re-estimated on the pre-1980 (1962Q1–1979Q4) and post-1980 (1980Q1–2025Q4) sub-samples. The spread coefficients are:

Sub-sample	$\hat{\beta}_{\text{Spread},1}$	$\hat{\beta}_{\text{Spread},2}$
Pre-1980 ($n \approx 70$)	2.21 (1.07)	-0.94 (1.16)
Post-1980 ($n \approx 185$)	-0.96 (0.56)	1.45 (0.56)

How do the spread coefficients differ across sub-samples? Which sub-sample would you use for forecasting 2026Q1? What principle guides this choice?

Exercise 2 | Testing for a Structural Break in the ADL(2,2)

KEY TAKEAWAYS

- Pre-1980: the spread effect is large but very imprecisely estimated ($n \approx 70$, large SEs) → both coefficients are individually insignificant
- Post-1980: the spread coefficients have opposite signs ($\delta_1 = -0.96$, $\delta_2 = +1.45$) → a higher spread last quarter is associated with lower growth this quarter, but a higher spread two quarters ago is associated with higher growth; if the spread has been **persistently elevated** (both lags similar), the net effect is positive
- **For forecasting 2026Q1:** use the **post-1980** model → current interest rate regime
- **General principle:** when a structural break is confirmed, training on the wrong regime introduces systematic bias that outweighs the benefit of more observations

Exercise 2 | Testing for a Structural Break in the ADL(2,2)

d) A Chow test for a **known** break at the COVID recession (2020Q2), using a fully interacted regression, gives:

```
Linear hypothesis test

Hypothesis:
D = 0,  GR_l1_D = 0,  GR_l2_D = 0,  Spread_l1_D = 0,  Spread_l2_D = 0

F = 6.07,   Df = 5,   p-value < 0.001
```

Is there evidence that the ADL coefficients changed after COVID? What does this imply for the post-1980 model you chose in c)?

Exercise 2 | Testing for a Structural Break in the ADL(2,2)

CATÓLICA LISBON BUSINESS & ECONOMICS

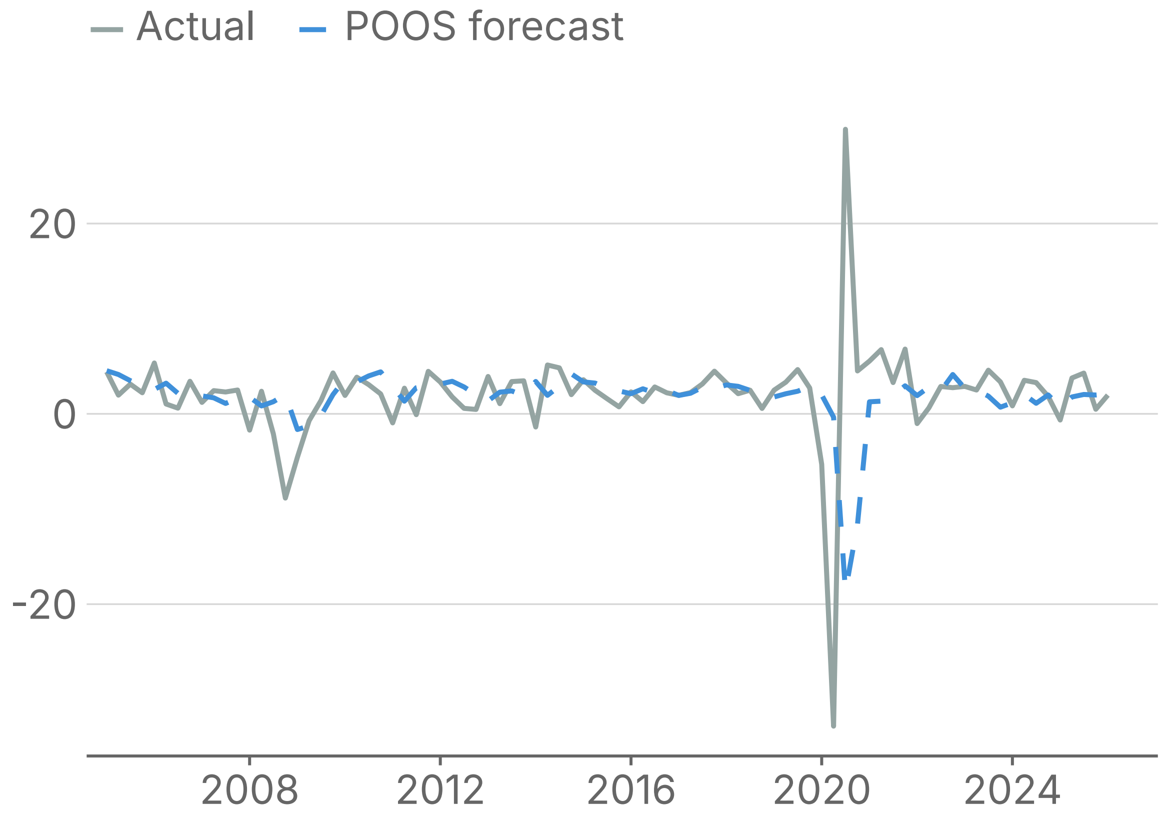
KEY TAKEAWAYS

- $F = 6.07 \gg 2.21$ (standard 5% critical value for $q = 5$): **strong evidence of a COVID break**
- The ADL coefficients shifted significantly after 2020Q2 → the model trained on 1980–2019 does not describe GDP dynamics post-COVID
- **Implication:** the post-1980 model in c) includes COVID-era observations that may have broken the relationship again
- **Practical response:** either (i) exclude 2020Q2–2020Q3 with dummy variables and re-estimate, or (ii) use only post-2020 data for forecasting → very few observations, parameter uncertainty becomes severe
- **There is no clean answer:** forecasting in the presence of a recent break requires judgment

Exercise 3 | POOS Analysis

POOS: ADL(2,2) Forecasts

US GDP Growth (%, post-1981)



a) Using a rolling training window from 1981Q1, pseudo out-of-sample one-step-ahead forecasts are computed for 2005Q1–2025Q4 and plotted on the left. Describe what you see. Does the model track actual GDP growth well across the full hold-out period?

Exercise 3 | POOS Analysis

KEY TAKEAWAYS

- The model tracks reasonably well in normal periods
- **Two episodes of catastrophic failure:**
  - 2008Q4–2009Q1: the financial crisis collapse is completely missed
  - 2020Q2: the COVID collapse of ≈ −31% is missed by an enormous margin
- **Lesson:** POOS performance is good evidence the model is stable in calm periods, but makes no promises in tail events

Exercise 3 | POOS Analysis

b) What do the POOS forecast errors at the 2008–09 financial crisis and the COVID recession tell us about model stability?

KEY TAKEAWAYS

- **2008–09 crisis:** the ADL model did not anticipate the sharp GDP collapse → POOS errors spike large. Financial crises are regime changes, not continuations of historical patterns; the model had no information to predict a new regime
- **COVID (2020Q2):** the –31% quarter is completely missed → no linear AR/ADL model trained on pre-COVID data has seen anything like it
- **Lesson:** poor POOS performance at specific episodes indicates the model failed to generalize → possibly a structural break, possibly a one-off outlier
- **Practical takeaway:** even without a formal QLR rejection, persistent large POOS errors are a warning sign that the model needs updating

Exercise 3 | POOS Analysis

c) The POOS RMSFE over the full hold-out period (2005Q1–2024Q4) is **6.98 pp**. When the crisis quarters (2008–09 and 2020) are excluded, it falls to **1.89 pp**. What does this tell you about where the forecast errors are concentrated? Does the model perform well in normal times?

KEY TAKEAWAYS

- In **normal times** the model performs well: **errors are small** (1.89 pp on average) and centered around zero, consistent with a stable post-1980 regime
- **Lesson:** a summary RMSFE statistic alone is misleading when a few extreme episodes drive the average; always inspect the plot alongside the number
- **Implication for 2026Q1:** if no crisis is anticipated, the post-1980 model's in-regime RMSFE (1.89 pp) is the relevant benchmark for forecast uncertainty

Discussion

1. The QLR test is designed to detect a single structural break. What would you do if you suspected **multiple** breaks over a long sample (e.g., 1960–2025)?

KEY TAKEAWAYS

- Multiple-break tests for *m* breaks simultaneously exist
- A **simpler practical approach**: look at the $F(\tau)$ plot → multiple peaks above the critical value point to multiple breaks
- **Rolling-window estimation** is a useful informal tool: estimate the model on a fixed-length rolling window and track whether coefficients drift over time
- For the **research project**: electricity consumption patterns likely shift with extreme events → check stability informally before finalizing your model

Discussion

2. What steps should you follow to implement a good time series modelling strategy?

KEY TAKEAWAYS

1. **Plot the series** → does it have a trend? Does variance grow? Is there seasonality?
2. **Test stationarity** → ADF test; transform if nonstationary (ΔY_t or $\Delta \ln Y_t$)
3. **Fit an AR(p)** → use ACF/PACF to guess p, confirm with BIC; this is your baseline
4. **Add predictors** → test whether external variables Granger-cause Y (joint F-test on lags of X)
5. **Evaluate forecasts out-of-sample** → compute RMSFE on a hold-out; compare to random walk and AR baseline
6. **Test stability** → QLR test + POOS plot; if breaks found, restrict to the relevant regime
7. **Report point forecast + 95% interval** → never report a point forecast alone

Discussion

3. The COVID quarter (2020Q2) showed GDP growth of approximately –31%. One student says: *"This is a structural break — we should split the sample at 2020Q2."* Another says: *"This is a one-off outlier — we should add a dummy variable for that quarter and keep the full sample."* Who is right?

Discussion

KEY TAKEAWAYS

- **Neither is simply right:** did the **coefficients** permanently changed (structural break) or it was just a single observation that was extreme (outlier)?
- **Empirical evidence leans toward outlier:** the economy bounced back rapidly (2020Q3 growth was +35%) → no permanent regime change
- But the **Chow test** in Exercise 2d rejects stability, so the picture is **genuinely ambiguous**
- **Practical recommendation:** add a 2020Q2–2020Q3 dummy pair, re-estimate, and check whether POOS performance improves. If the instability came from those two quarters, the dummy absorbs it without discarding useful data