

Econometrics

Practical Session 16

ADL Models and Forecast Intervals



Ricardo Gouveia-Mendes
rgouveiamendes@ucp.pt

Spring 2025-26

Católica-Lisbon School of Business and Economics

Theoretical Wrap-up

The ADL(p, q) Model

- **Autoregressive Distributed Lag** model adds q lags of a predictor x_t to an AR(p) model:

$$y_t = \underbrace{\beta_0 + \beta_1 y_{t-1} + \dots + \beta_p y_{t-p}}_{\text{autoregressive part}} + \underbrace{\delta_1 x_{t-1} + \dots + \delta_q x_{t-q}}_{\text{distributed lag part}} + \varepsilon_t$$

$$\hat{y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 y_T + \dots + \hat{\beta}_p y_{T-p+1} + \hat{\delta}_1 x_T + \dots + \hat{\delta}_q x_{T-q+1}$$

- Are past values of x **useful for predicting y above and beyond** past values of y ? **Does x Granger-causes y ?**
→ **Joint F -test:** $H_0 : \delta_1 = \delta_2 = \dots = \delta_q = 0$
- **Granger causality is about forecasting**, not economic causality

How Good a Forecast is?

- We know for sure **there will be a forecast error**:

$$\text{MSFE} = \mathbb{E}[(y_{T+h} - \hat{y}_{T+h})^2] \neq 0$$

- The problem is to **estimate it before the actual realization of y_{T+h}**
- Three approaches:

1. **Standard Error of Regression (SER):** (*h=1 only*)

$$\widehat{\text{MSFE}}_1 = \hat{\sigma}^2 = \frac{\text{SSR}}{T - p - 1}$$

In-sample estimate → **optimism bias** because OLS minimizes the in-sample residuals (systematically smaller than errors on new data)

How Good a Forecast is?

2. Final Prediction Error (FPE): ($h=1$ only)

$$\widehat{\text{MSFE}}_1 = \frac{T + p + 1}{T} \hat{\sigma}^2$$

In-sample estimate that corrects for the **optimism bias** of SER

3. Pseudo out-of-sample (POOS): split the sample into training ($t = 1, \dots, T_0$) and hold-out ($t = T_0 + 1, \dots, T$):

$$\widehat{\text{MSFE}}_h = \frac{1}{P_h} \sum_{t=T_0}^{T-h} (y_{t+h} - \hat{y}_{t+h|t})^2, \quad P_h = T - h - T_0$$

Lose h observations from the end of the hold-out

Forecast Intervals

- A **point forecast** without uncertainty quantification is **incomplete**
- Two **sources of uncertainty** in forecast errors:
 1. **Shock**: the next ε_{T+1} is unknown even with perfect coefficients
 2. **Estimation**: OLS estimates are efficient $\rightarrow \lim_{T \rightarrow +\infty} \hat{\beta} = \beta$
- **MSFE grows with h** :
 - At $h = 1$: ε_{T+1} is unknown
 - At $h = 2$: ε_{T+2} is unknown, and $\hat{y}_{T+1|T}$ is itself uncertain
 - **Uncertainty compounds** with each additional step
 $\rightarrow \widehat{MSFE}_1 \leq \widehat{MSFE}_2 \leq \dots$, converging to $\text{Var}(y)$ as $h \rightarrow +\infty$

- $(1 - \alpha)\%$ **forecast confidence interval** at horizon h :

$$\hat{y}_{T+h|T} \pm Z_{\alpha/2} \times \hat{\sigma}_{f(h)}, \quad \hat{\sigma}_{f(h)} = \sqrt{\widehat{\text{MSFE}}_h}$$

- Forecast intervals **widen as h grows**
- **In R:** `predict(..., interval = "prediction")` computes exact intervals:

$$\hat{y}_{T+1|T} \pm Z_{\alpha/2} \times \hat{\sigma} \sqrt{1 + \mathbf{x}_{\text{new}}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_{\text{new}}}$$

Exercises

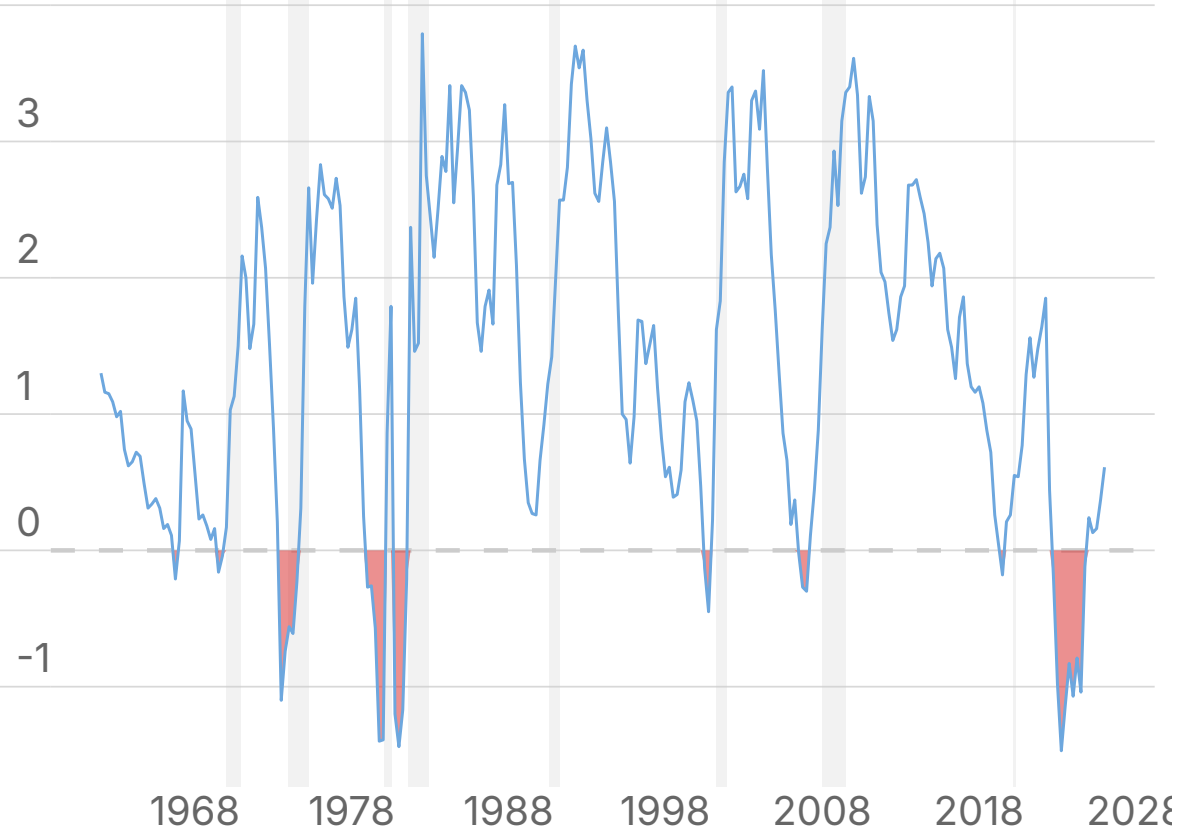
Exercise 1 | GDP Growth and the Term Spread

(S&W Ch. 15, with updated data from FRED, 1962Q1–2025Q4)

- Term spread: $\text{Spread}_t = r_t^{10Y} - r_t^{3M}$
 - An inverted yield curve historically precedes recessions
 - We want to test whether the spread improves forecasts of annualized GDP growth rate GR_t
- a)** The term spread (10Y – 3M Treasury yield) is plotted below, with shaded regions marking yield curve inversions. When do inversions cluster? What pattern connects them to economic activity?

Exercise 1 | GDP Growth and the Term Spread

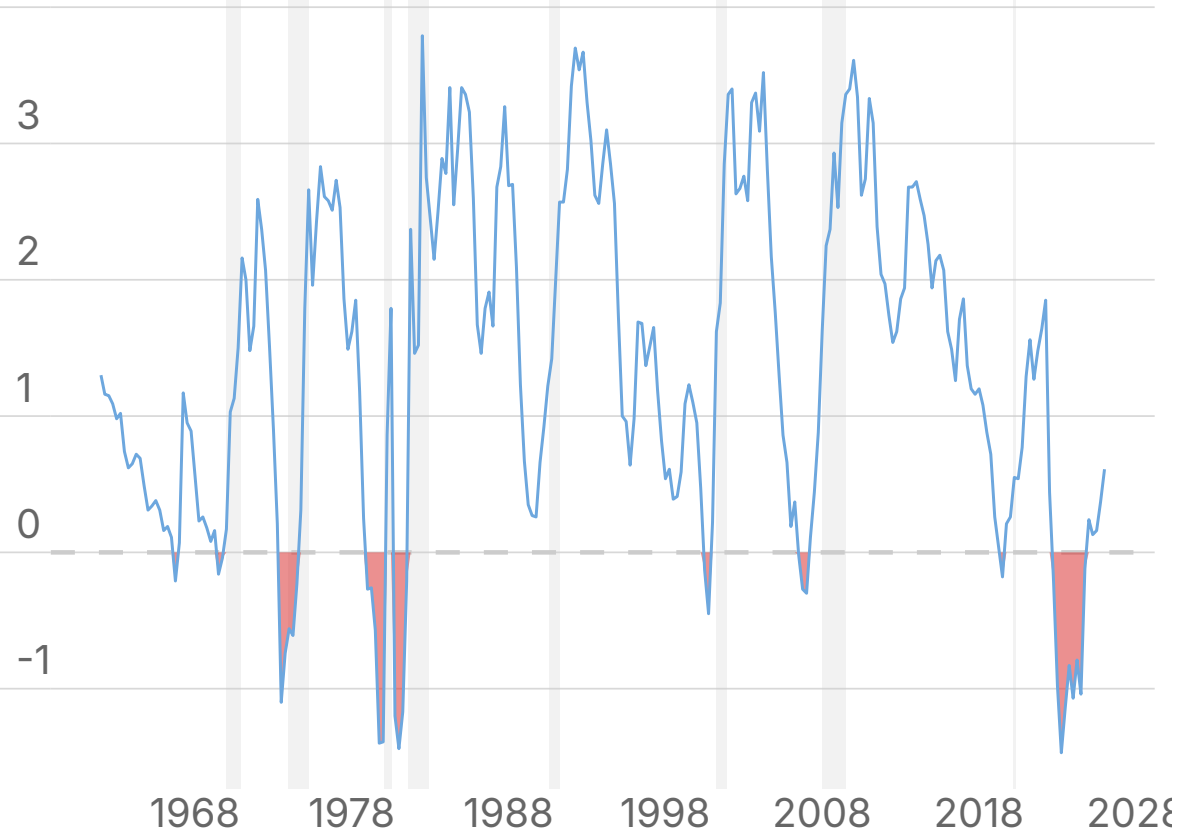
US Term Spread: 10Y – 3M (pp)



Gray: NBER recessions. Red: inverted yield curve. Source: FRED

Exercise 1 | GDP Growth and the Term Spread

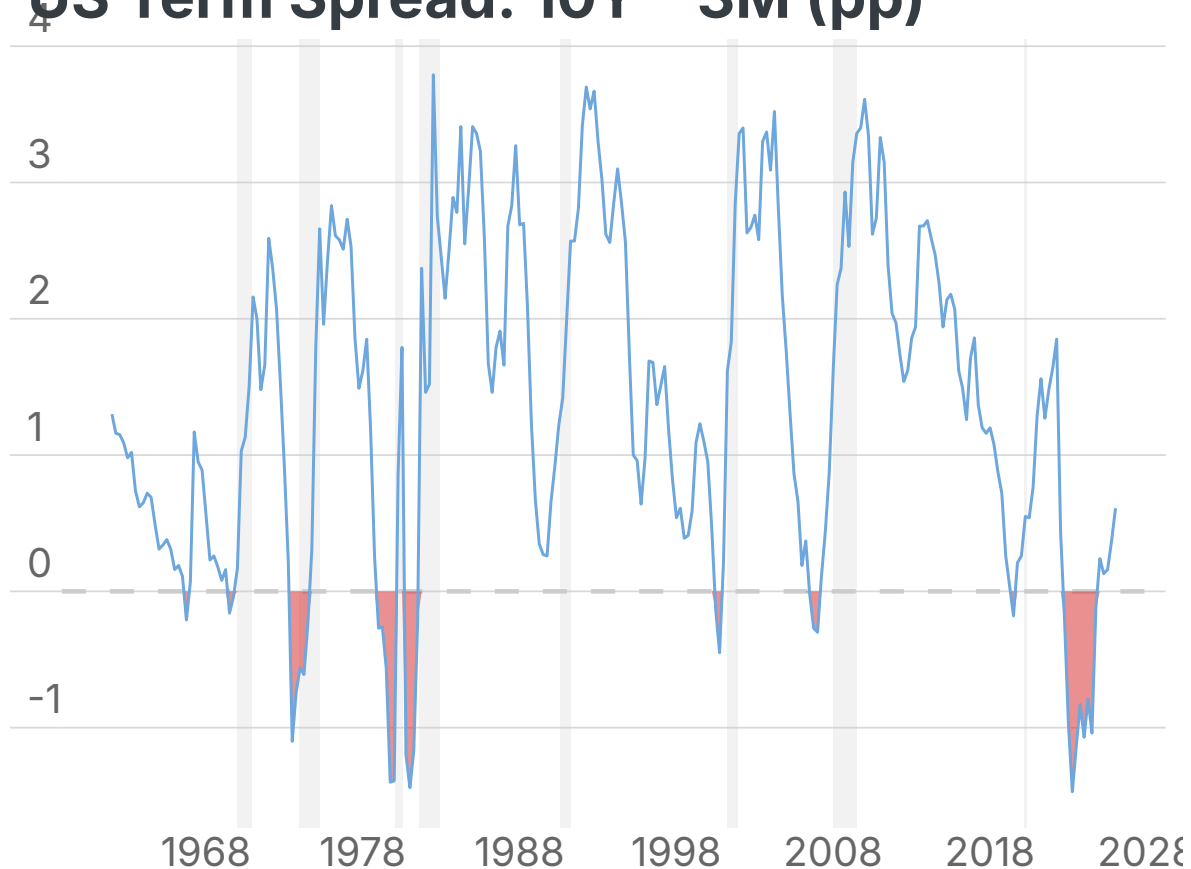
US Term Spread: 10Y – 3M (pp)



Gray: NBER recessions. Red: inverted yield curve. Source: FRED

Exercise 1 | GDP Growth and the Term Spread

US Term Spread: 10Y – 3M (pp)



Gray: NBER recessions. Red: inverted yield curve. Source: FRED

KEY TAKEAWAYS

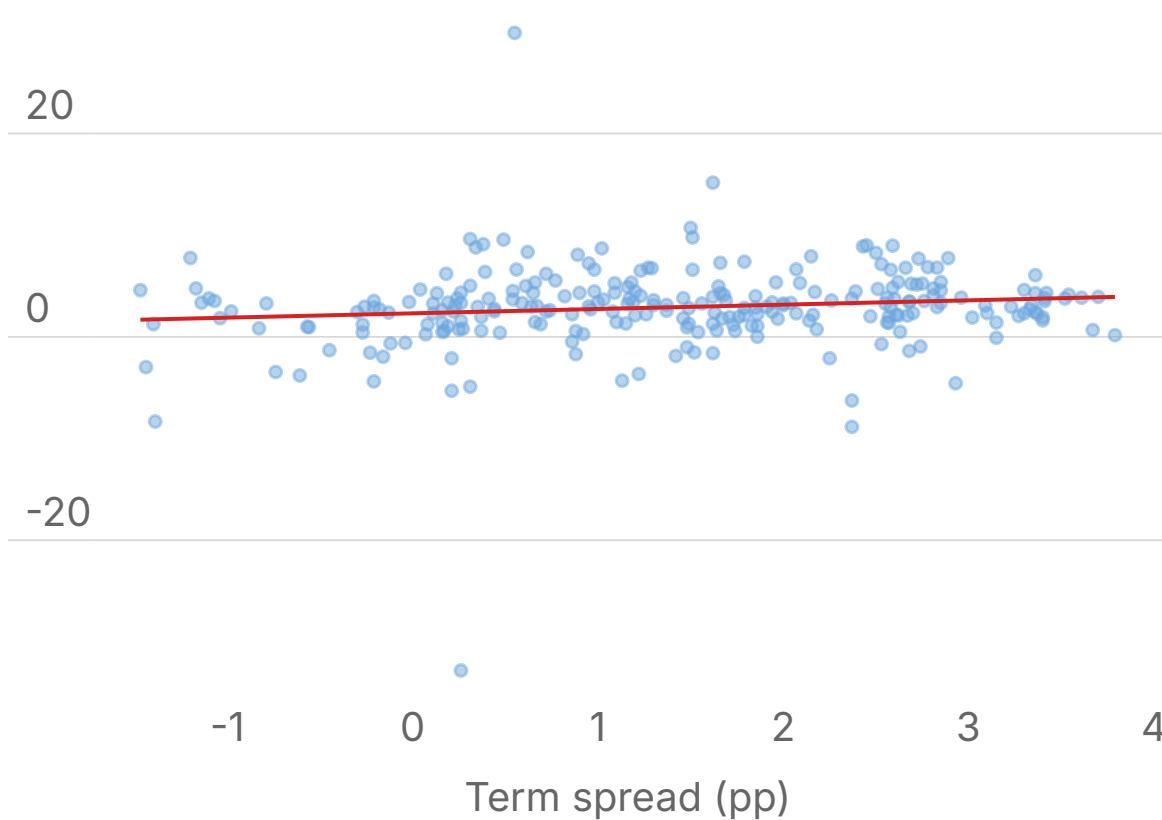
- Inversions before recessions: **1979–80, 2000, 2006–07, 2019, 2022–24**
- Each inversion was followed by slower growth or a downturn
- The timing lag is variable → **not** a precise timing signal

Exercise 1 | GDP Growth and the Term Spread

b) The scatter plot below shows the term spread against next-quarter GDP growth. The slope is positive but barely significant ($p = 0.045$, $R^2 = 0.016$). What does the low R^2 tell you? Does a low R^2 mean the spread is useless for forecasting?

Exercise 1 | GDP Growth and the Term Spread

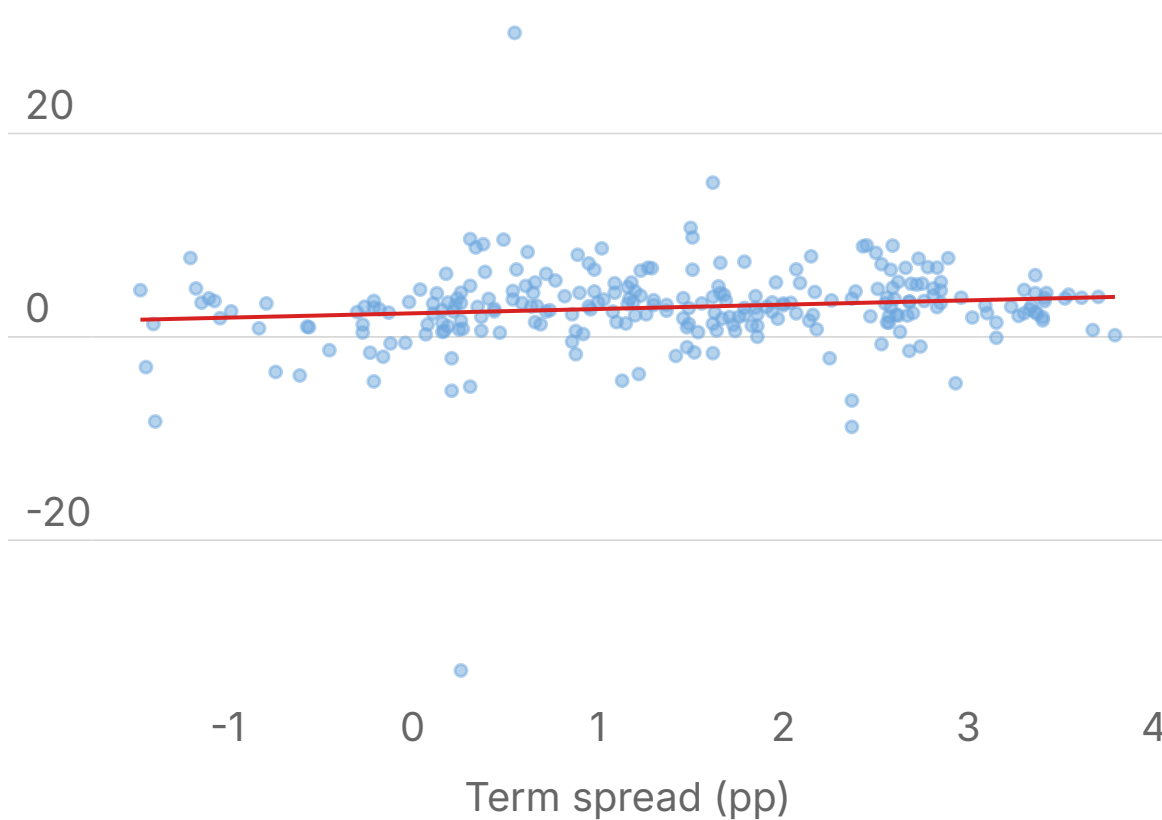
Term Spread and Next-Quarter GDP Gro



OLS: slope = 0.43 (se = 0.21)

Exercise 1 | GDP Growth and the Term Spread

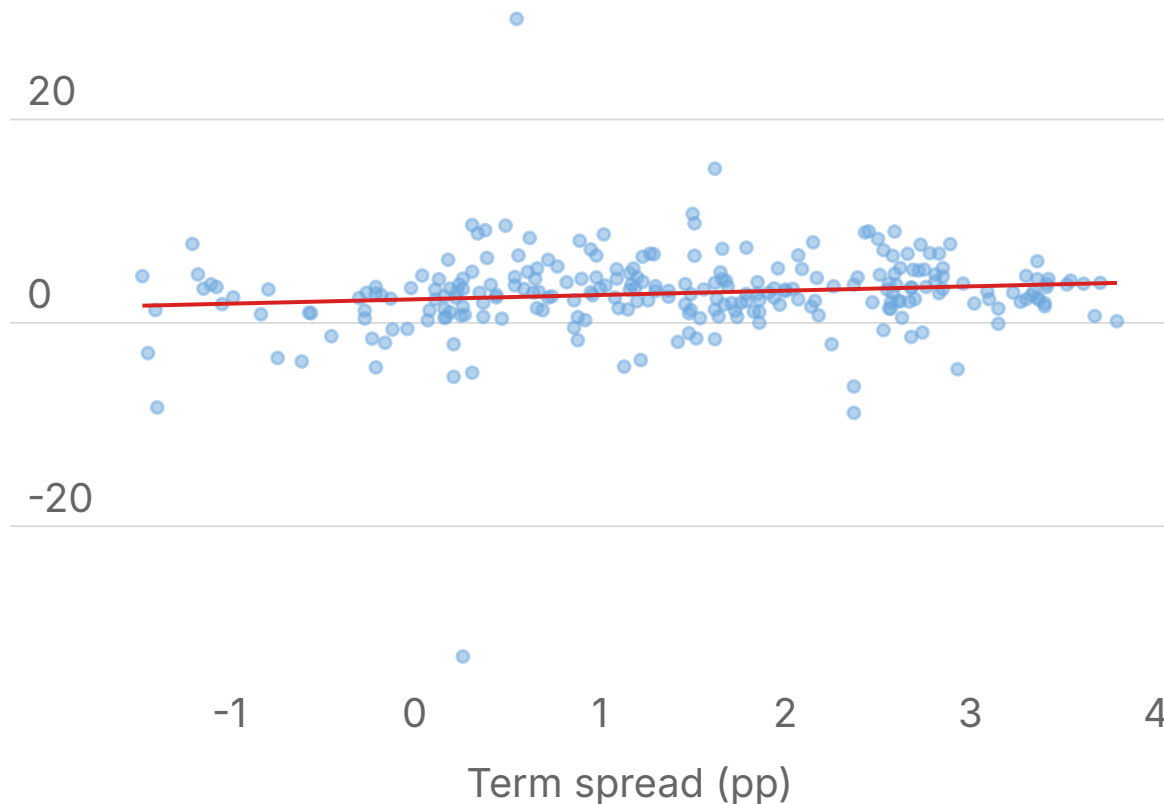
Term Spread and Next-Quarter GDP Gro



OLS: slope = 0.43 (se = 0.21)

Exercise 1 | GDP Growth and the Term Spread

Term Spread and Next-Quarter GDP Gro



OLS: slope = 0.43 (se = 0.21)

KEY TAKEAWAYS

- Positive slope, barely significant ($p = 0.045$, $R^2 = 0.016$)
- On its own the spread explains very little
- May still help **in combination** with AR lags

Exercise 1 | GDP Growth and the Term Spread

c) The ADL(2,2) is estimated for GDP growth on the term spread. Are the spread lags individually significant? Are they jointly significant? What does the joint test tell us about Granger causality?

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.957	0.465	4.21	<0.001	***
GR_l1	-0.010	0.063	-0.16	0.874	
GR_l2	0.079	0.063	1.27	0.206	
Spread_l1	-0.502	0.498	-1.01	0.315	
Spread_l2	1.034	0.503	2.06	0.041	*

RSE = 4.16 $R^2 = 0.042$

Exercise 1 | GDP Growth and the Term Spread

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
251	4444					
249	4299	2	144.5	4.19	0.016	*

Exercise 1 | GDP Growth and the Term Spread

c) The ADL(2,2) is estimated for GDP growth on the term spread. Are the spread lags individually significant? Are they jointly significant? What does the joint test tell us about Granger causality?

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.957	0.465	4.21	<0.001	***
GR_l1	-0.010	0.063	-0.16	0.874	
GR_l2	0.079	0.063	1.27	0.206	
Spread_l1	-0.502	0.498	-1.01	0.315	
Spread_l2	1.034	0.503	2.06	0.041	*

RSE = 4.16 $R^2 = 0.042$

Exercise 1 | GDP Growth and the Term Spread

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
251	4444					
249	4299	2	144.5	4.19	0.016	*

Exercise 1 | GDP Growth and the Term Spread

c) The ADL(2,2) is estimated for GDP growth on the term spread. Are the spread lags individually significant? Are they jointly significant? What does the joint test tell us about Granger causality?

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.957	0.465	4.21	<0.001	***
GR_l1	-0.010	0.063	-0.16	0.874	
GR_l2	0.079	0.063	1.27	0.206	
Spread_l1	-0.502	0.498	-1.01	0.315	
Spread_l2	1.034	0.503	2.06	0.041	*

RSE = 4.16 $R^2 = 0.042$

Exercise 1 | GDP Growth and the Term Spread

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
251	4444				
249	4299	2	144.5	4.19	0.016 *

KEY TAKEAWAYS

- AR lags remain insignificant individually
- Spread_l2 is significant at 5%: 1 pp wider spread two quarters ago → +1.03 pp faster growth

Exercise 1 | GDP Growth and the Term Spread

c) The ADL(2,2) is estimated for GDP growth on the term spread. Are the spread lags individually significant? Are they jointly significant? What does the joint test tell us about Granger causality?

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.957	0.465	4.21	<0.001	***
GR_l1	-0.010	0.063	-0.16	0.874	
GR_l2	0.079	0.063	1.27	0.206	
Spread_l1	-0.502	0.498	-1.01	0.315	
Spread_l2	1.034	0.503	2.06	0.041	*

RSE = 4.16 $R^2 = 0.042$

Exercise 1 | GDP Growth and the Term Spread

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
251	4444				
249	4299	2	144.5	4.19	0.016 *

KEY TAKEAWAYS

- AR lags remain insignificant individually
- Spread_l2 is significant at 5%: 1 pp wider spread two quarters ago → +1.03 pp faster growth

Jointly, spread lags are significant ($F = 4.19$, $p = 0.016$) → **the yield curve Granger-causes GDP growth**

Exercise 1 | GDP Growth and the Term Spread

d) The ADL forecast (2.27%) for 2026Q1 is lower than the AR(1) forecast (2.94%). What is driving the difference? What does this tell us about current yield curve conditions?

Exercise 1 | GDP Growth and the Term Spread

d) The ADL forecast (2.27%) for 2026Q1 is lower than the AR(1) forecast (2.94%). What is driving the difference? What does this tell us about current yield curve conditions?

Exercise 1 | GDP Growth and the Term Spread

d) The ADL forecast (2.27%) for 2026Q1 is lower than the AR(1) forecast (2.94%). What is driving the difference? What does this tell us about current yield curve conditions?

KEY TAKEAWAYS

- The ADL forecast is lower: 2.27 % vs. 2.94 %
- The recent spread inversion strikes predicted growth

Exercise 1 | GDP Growth and the Term Spread

e) The 95% prediction intervals for 2026Q1 are:

ADL(2,2):	fit = 2.27	lwr = -6.02	upr = 10.57
AR(1):	fit = 2.94	lwr = -5.36	upr = 11.25

What does this tell you about the practical value of a statistically significant predictor with $R^2 = 0.04$?

Exercise 1 | GDP Growth and the Term Spread

e) The 95% prediction intervals for 2026Q1 are:

ADL(2,2):	fit = 2.27	lwr = -6.02	upr = 10.57
AR(1):	fit = 2.94	lwr = -5.36	upr = 11.25

What does this tell you about the practical value of a statistically significant predictor with $R^2 = 0.04$?

Exercise 1 | GDP Growth and the Term Spread

e) The 95% prediction intervals for 2026Q1 are:

```
ADL(2,2):  fit = 2.27    lwr = -6.02    upr = 10.57
AR(1):     fit = 2.94    lwr = -5.36    upr = 11.25
```

What does this tell you about the practical value of a statistically significant predictor with $R^2 = 0.04$?

KEY TAKEAWAYS

- Both intervals are very wide → GDP growth is hard to forecast
- ADL interval is marginally narrower: the spread lags explain a small fraction of variance
- **Lesson:** even a statistically significant predictor may barely narrow the forecast interval if R^2 remains small

Exercise 1 | GDP Growth and the Term Spread

f) Does a significant joint F -test mean the yield curve **causes** recessions? How would you explain the predictive value of the spread to a non-technical audience?

Exercise 1 | GDP Growth and the Term Spread

f) Does a significant joint F -test mean the yield curve **causes** recessions? How would you explain the predictive value of the spread to a non-technical audience?

Exercise 1 | GDP Growth and the Term Spread

f) Does a significant joint F -test mean the yield curve **causes** recessions? How would you explain the predictive value of the spread to a non-technical audience?

KEY TAKEAWAYS

- A significant F -test shows **predictive content**, not causation
- **Correlation in time-series data is not causation**: a leading indicator may forecast without causing
- The spread forecasts growth because it **aggregates market expectations** about the future path of rates and economic activity
- To a non-technical audience: *“When short-term rates rise above long-term rates, investors expect the economy to slow — and historically, slower growth has tended to follow.”*

Exercise 2 | Out-of-Sample Forecast Evaluation

a) The AR(1), ADL(2,2), and random walk are trained on 1962Q1–2019Q4 and used to generate one-step-ahead forecasts over 2020Q1–2025Q4. Before seeing the RMSFE results: which model do you expect to perform best in the hold-out period (2020Q1–2025Q4)? What specific events make this period challenging for any model?

Exercise 2 | Out-of-Sample Forecast Evaluation

a) The AR(1), ADL(2,2), and random walk are trained on 1962Q1–2019Q4 and used to generate one-step-ahead forecasts over 2020Q1–2025Q4. Before seeing the RMSFE results: which model do you expect to perform best in the hold-out period (2020Q1–2025Q4)? What specific events make this period challenging for any model?

Exercise 2 | Out-of-Sample Forecast Evaluation

a) The AR(1), ADL(2,2), and random walk are trained on 1962Q1–2019Q4 and used to generate one-step-ahead forecasts over 2020Q1–2025Q4. Before seeing the RMSFE results: which model do you expect to perform best in the hold-out period (2020Q1–2025Q4)? What specific events make this period challenging for any model?

KEY TAKEAWAYS

- The ADL(2,2) should outperform the AR(1) if the yield curve retains predictive power out-of-sample
- The COVID-19 pandemic (2020Q2–Q3) produced GDP swings far outside any historical range — no linear model trained on pre-2020 data could anticipate them
- The 2022 rate-hike cycle and subsequent spread inversion add further complexity

Exercise 2 | Out-of-Sample Forecast Evaluation

b) The RMSFE results are shown below. Does the ranking match your expectation from (a)? Why does excluding COVID quarters change all three RMSFEs dramatically? What does this tell us about model evaluation?

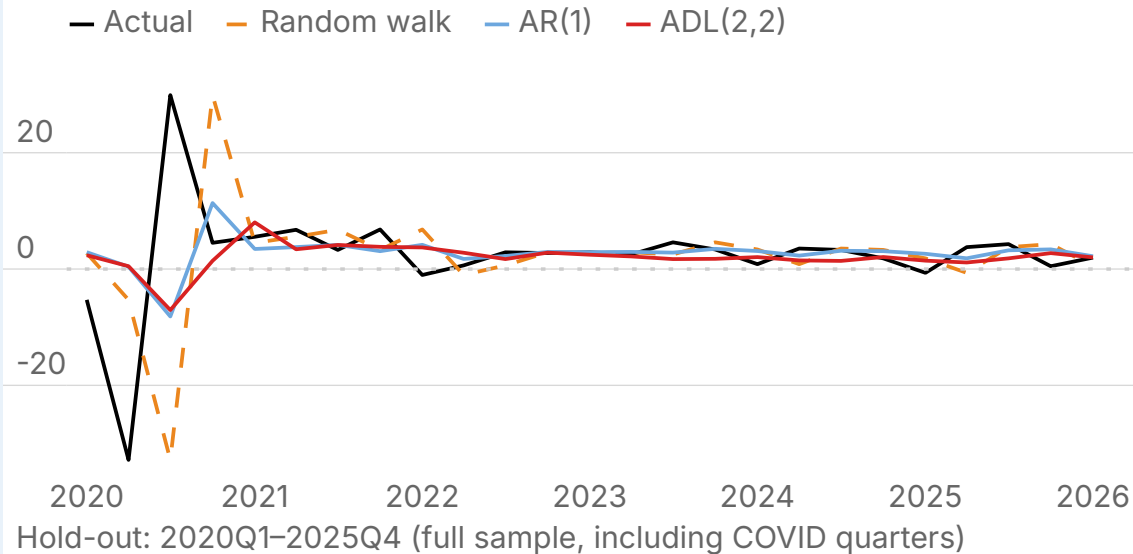
Model	Full hold-out	Excl. COVID
Random walk	15.2	6.28
AR(1)	10.7	3.06
ADL(2,2)	10.5	2.75

Exercise 2 | Out-of-Sample Forecast Evaluation

Exercise 2 | Out-of-Sample Forecast Evaluation

KEY TAKEAWAYS

POOS Forecasts vs Actual GDP Growth (% , annualized)



- ADL < AR(1) < Random walk in both cases → the ranking is robust
- COVID dominates all three errors → no linear AR model foresaw the pandemic
- **Lesson:** always report accuracy with and without extreme episodes; consistent ranking across both builds confidence

Exercise 2 | Out-of-Sample Forecast Evaluation

c) Of the 24 hold-out quarters, 21 (87.5%) fell within the AR(1)'s 95% prediction interval. Coverage is 87.5%, close to the nominal 95%. Which two quarters fell outside the band? What does this reveal about the limits of model-based forecast intervals?

Exercise 2 | Out-of-Sample Forecast Evaluation

c) Of the 24 hold-out quarters, 21 (87.5%) fell within the AR(1)'s 95% prediction interval. Coverage is 87.5%, close to the nominal 95%. Which two quarters fell outside the band? What does this reveal about the limits of model-based forecast intervals?

Exercise 2 | Out-of-Sample Forecast Evaluation

c) Of the 24 hold-out quarters, 21 (87.5%) fell within the AR(1)'s 95% prediction interval. Coverage is 87.5%, close to the nominal 95%. Which two quarters fell outside the band? What does this reveal about the limits of model-based forecast intervals?

KEY TAKEAWAYS

- ~88% of actual values fall within the 95% band: close to nominal coverage
- The two COVID quarters (2020Q2–Q3) fall **outside** the interval
- **Lesson:** forecast intervals are calibrated for “normal” volatility; tail events (pandemics, financial crises) are rarely captured by mechanical extrapolation

Exercise 2 | Out-of-Sample Forecast Evaluation

d) The 95% prediction interval for 2026Q1 spans about 16 percentage points. A policymaker asks: *"Is this model useful?"* How would you answer?

Exercise 2 | Out-of-Sample Forecast Evaluation

d) The 95% prediction interval for 2026Q1 spans about 16 percentage points. A policymaker asks: *"Is this model useful?"* How would you answer?

Exercise 2 | Out-of-Sample Forecast Evaluation

d) The 95% prediction interval for 2026Q1 spans about 16 percentage points. A policymaker asks: *"Is this model useful?"* How would you answer?

KEY TAKEAWAYS

- **Useful for what?** The wide interval honestly represents genuine uncertainty
- The model is useful as a **baseline**: it tells us that absent other information, growth should be around 2.3%
- To narrow the interval, we need a model that explains more variation: higher-frequency data, better indicators
- **Key message**: report intervals, not just point forecasts; and set realistic expectations about what any model can do for a nearly-white-noise series

Exercise 3 | Putting it Together

a) The ADL(2,2) has $R^2 = 0.042$ in-sample but outperforms the AR(1) out-of-sample. How is this possible? What does it reveal about using R^2 for model selection?

Exercise 3 | Putting it Together

a) The ADL(2,2) has $R^2 = 0.042$ in-sample but outperforms the AR(1) out-of-sample. How is this possible? What does it reveal about using R^2 for model selection?

Exercise 3 | Putting it Together

a) The ADL(2,2) has $R^2 = 0.042$ in-sample but outperforms the AR(1) out-of-sample. How is this possible? What does it reveal about using R^2 for model selection?

KEY TAKEAWAYS

- R^2 measures in-sample fit; OLS minimizes training residuals, so adding any variable — even noise — raises R^2
- Out-of-sample RMSFE penalizes overfitting: spurious parameters that happen to fit the training data do not generalize
- The spread adds genuine predictive signal (small but real), which raises in-sample R^2 and also helps out-of-sample
- **Lesson:** R^2 is a poor criterion for model selection; prefer information criteria (AIC/BIC) or hold-out RMSFE

Exercise 3 | Putting it Together

b) A colleague says: *“The yield curve perfectly predicted every recession since 1970. Therefore we should use it to time economic policy.”* Identify two problems with this reasoning.

Exercise 3 | Putting it Together

b) A colleague says: *“The yield curve perfectly predicted every recession since 1970. Therefore we should use it to time economic policy.”* Identify two problems with this reasoning.

Exercise 3 | Putting it Together

b) A colleague says: *“The yield curve perfectly predicted every recession since 1970. Therefore we should use it to time economic policy.”* Identify two problems with this reasoning.

KEY TAKEAWAYS

- 1. In-sample overfitting / data mining:** the yield curve’s track record was identified by searching over many indicators; the few recessions in the sample make it easy to “predict” them after the fact
- 2. Confusion of statistical predictability with causality / policy relevance:** Granger causality means past spread values contain information about future growth, not that manipulating spreads controls growth; policy based on a leading indicator may alter the very correlation it relies on (Lucas critique)

Exercise 3 | Putting it Together

c) Your research project requires forecasting Saturday electricity consumption. You have a weekly $AR(7)$ and an ADL that adds yesterday's temperature as a predictor. Describe how you would choose between them using the framework from this session.

Exercise 3 | Putting it Together

Exercise 3 | Putting it Together

KEY TAKEAWAYS

1. **Split the data:** use all weeks up to a cut-off as training; reserve the last several Saturdays as the hold-out
2. **Estimate both models** on the training sample
3. **Generate one-step-ahead forecasts** for each hold-out Saturday and compute RMSFE for each model
4. **Check Granger causality:** run a joint F -test for the temperature lags in the ADL; if not significant and RMSFE gain is negligible, prefer the simpler AR(7)
5. **Report prediction intervals** from the chosen model; verify coverage against hold-out actuals
6. **Exclude anomalous weeks** (e.g., public holidays) and recheck — consistent ranking across both samples builds confidence

1. The ADL(2,2) has lower in-sample R^2 than an AR(4) with four lagged GDP growth terms, yet its out-of-sample RMSFE is lower. How is that possible?

1. The ADL(2,2) has lower in-sample R^2 than an AR(4) with four lagged GDP growth terms, yet its out-of-sample RMSFE is lower. How is that possible?

1. The ADL(2,2) has lower in-sample R^2 than an AR(4) with four lagged GDP growth terms, yet its out-of-sample RMSFE is lower. How is that possible?

KEY TAKEAWAYS

- R^2 measures in-sample fit
- Out-of-sample RMSFE penalises overfitting: noise parameters that fit the training data do not forecast well on new data
- **Lesson:** model selection by information criteria (BIC) or hold-out RMSFE is more reliable than comparing R^2 values

2. Think about the research project. How will you construct a forecast interval for Portuguese electricity consumption?

2. Think about the research project. How will you construct a forecast interval for Portuguese electricity consumption?

2. Think about the research project. How will you construct a forecast interval for Portuguese electricity consumption?

KEY TAKEAWAYS

- Use `predict(..., interval = "prediction")` from your best AR or ADL model
- **Report both the point forecast and the 95% interval**, though it is not strictly required by the project deliverables
- If your interval is implausibly narrow, check whether you evaluated it in-sample (overfitting) rather than on held-out data