Week 1: Foundations of Financial Data Science

Data Science as Statistical Science

Part I: Data Science Is Statistical Science

What Is Data Science, Really?

Data science = the disciplined study of **variation** and **uncertainty** in data.

- **Variation**: what we model (differences across units, over time, between groups)
- **Uncertainty**: what we quantify (limits of what we can know from finite, noisy data)
- **Not** just coding, algorithms, or "AI" : those are tools in service of this study
- Finance amplifies both: noisy signals, non-stationary processes, strategic behaviour

Three Challenges of Statistical Inference

Gelman, Hill, and Vehtari (2020) frame all statistical work around three challenges:

1. **Sample → Population**: Can 100 stocks tell us about the market?
2. **Treatment → Control**: Did the strategy *cause* the improvement?
3. **Measurement → Construct**: Does volatility capture actual risk?

All three are prediction problems under uncertainty.

The Model Choice Dilemma

- Model A: five curated signals, 85% accuracy on training data
- Model B: fifty engineered signals, 92% on same data
- **Question**: Which model wins out-of-sample?
- Without disciplined validation, the higher score might just be noise

Learning Objectives

- Decide when complex models earn their variance by validating and regularising rigorously
- Use the bias-variance perspective to diagnose model behaviour for returns and risk
- Quantify and communicate uncertainty with the right statistical tools
- Contrast frequentist and Bayesian reasoning; pick the frame that answers the question
- Follow reproducible research practices: transparent code, documented assumptions, evidence-based claims

Part II: Variation : The Heart of Statistical Science

Financial Returns 101

- Simple return: $r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$; log return: $g_t = \ln P_t - \ln P_{t-1}$
- Excess return: asset return minus a benchmark (risk-free or market), e.g., $r_t^e = r_t - r_{f,t}$
- Why returns (not prices)? Stationarity and comparability across assets
- Our targets: market (MKT) and factor returns (e.g., MOM)

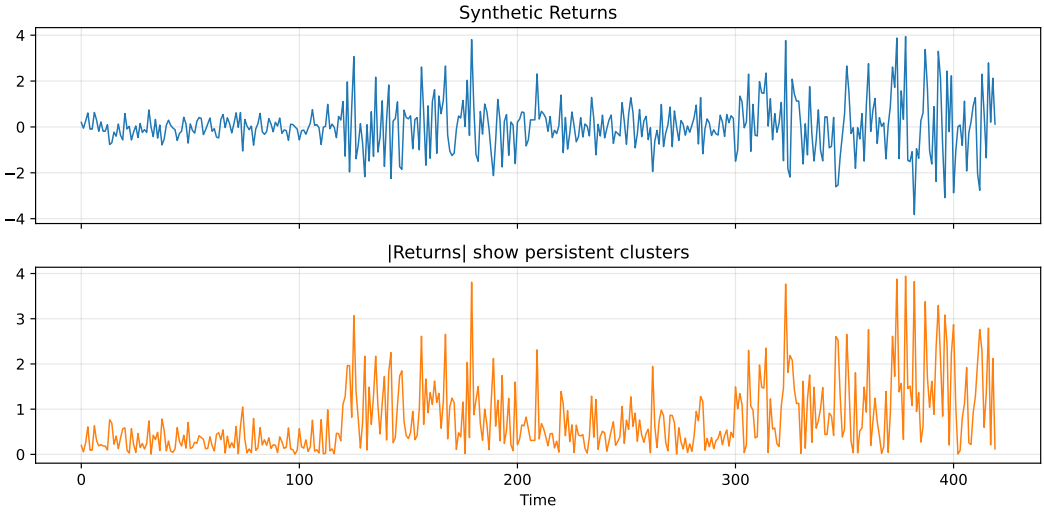
Stylised Facts: What Markets Give Us

- **Heavy tails** and mild skewness : extreme events happen more than Normal predicts
- **Weak linear autocorrelation** in raw returns : hard to forecast
- **Volatility clustering** and leverage effects : calm follows calm, storms follow storms
- **Time-varying factors**, non-stationarity, and microstructure noise

Why These Facts Arise

- Information arrival is uneven; volatility is a mixture of states
- Traders update at different speeds, creating persistence in $|r|$
- Microstructure frictions distort high-frequency data
- Structural shifts (policy, technology) move the goalposts; expect regime change

Volatility Clustering Snapshot



Part III: Regression : Coefficients as Comparisons

The Classical Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

- OLS: minimises sum of squared residuals
- Under CLRM assumptions: OLS is **BLUE** (Best Linear Unbiased Estimator)

Property	Meaning
Best	Minimum variance among linear unbiased estimators
Linear	$\widehat{\beta}$ is linear function of Y
Unbiased	$\mathbb{E}[\widehat{\beta}] = \beta$ on average

Assumptions Ranked by Importance

Gelman, Hill, and Vehtari (2020) argue we focus on the wrong assumptions:

Rank	Assumption	Why It Matters
1	**Validity**	Does your model address your research question?
2	**Representativeness**	Is your sample representative of target population?
3	**Additivity & Linearity**	Most important *mathematical* assumption
4	**Independence of errors**	Violated in time series, spatial, multilevel data
5	**Equal variance**	Heteroscedasticity rarely changes conclusions
6	**Normality of errors**	"Barely important at all" for estimation

Coefficients as Comparisons, Not Effects

Example: ESG and Stock Returns

```
annual_return = 8.2 + 0.15 × ESG_score + 0.02 × market_cap + error
```

Comparison ( )	Effect ( )
"Firms with higher ESG scores have 15bp higher returns, on average"	"Improving ESG causes returns to increase by 15bp"

Why the Distinction Matters

High-ESG companies may differ systematically:

- Better management overall
- Stronger governance structures
- More patient investors
- Greater financial resources

The observed return difference could reflect these omitted factors, not ESG itself.

More Finance Examples

Research Finding	Comparison ( )	Effect ( )
$= 0.3$ on analyst coverage	More coverage $\rightarrow$ higher returns observed	Adding analysts *causes* higher returns
$=$ -0.05 on leverage	Leveraged firms $\rightarrow$ lower returns	Reducing leverage *increases* returns
$= 0.02$ on insider ownership	Higher ownership $\rightarrow$ better performance	Giving managers shares *improves* performance

Causal claims require different evidence: RCTs, IV, natural experiments.

The Regression Fallacy

> ⚠ Warning
>
> **Regression to the mean** is often mistaken for a causal effect.
> A company performs poorly last year, improves this year. Did the new CEO cause the improvement? Or would regression to the mean have produced similar results anyway?

Extreme observations contain signal + luck. On repetition, luck averages out.

When Assumptions Fail

Violation	OLS Unbiased?	Standard Errors Valid?	Remedy
Heteroscedasticity			White (HC) SEs
Autocorrelation			Newey-West (HAC)
Multicollinearity		(but imprecise)	Regularisation
Endogeneity			IV methods

Connection to Machine Learning

Tip

ML methods extend classical econometrics:

- **Regularisation** (ridge, lasso) $\rightarrow$ addresses multicollinearity
- **Tree-based methods** $\rightarrow$ handle nonlinearity and interactions automatically
- **Sequence learning** $\rightarrow$ explicitly models temporal dependencies

These are not departures from statistics : they are extensions.

Part IV: Uncertainty : What R^2 Can and Cannot Tell Us

The Limits of R^2

$$R^2 = 1 - \frac{\text{Residual variance}}{\text{Total variance}}$$

What R^2 tells us:

- How much variance our predictors explain
- Whether adding variables improves explanatory power

What R^2 does NOT tell us:

- Whether the model is *correct* (wrong model can have high R^2)
- Whether the model is *useful* for prediction (low R^2 can still be actionable)
- Whether relationships are *causal*

Low R^2 Can Still Be Informative

Gelman, Hill, and Vehtari (2020) note: predicting earnings from height yields R^2 0.10.

This means 90% of variance has nothing to do with height. **Yet the regression is still informative** : it reveals a genuine association.

In finance, R^2 values of 0.01-0.05 are common when predicting **returns**. This reflects fundamental difficulty, not model failure.

But where you look matters: Volatility (~25% R^2) and cross-sectional variation (~10% R^2) are more predictable. We'll see why in Week 3.

Interactions Require 4× Sample Size

A crucial but overlooked insight from Gelman, Hill, and Vehtari (2020):

> *Estimating interaction effects requires roughly* ***four times the sample size*** *of main effects at the same precision.*

Why? Standard error of interaction $2\times$ standard error of main effect.

Implications:

- Study powered for "size effect" is *underpowered* for "does size effect vary by industry?"
- Significant interactions in exploratory analysis are likely *overestimated*
- Studying "does momentum work differently in bull vs. bear markets?" requires substantially more data

Part V: Five Pitfalls of Statistical Significance

Pitfall 1: Significance $\neq$ Practical Importance

A result can be "statistically significant" yet trivially small.

Example: Strategy earns 0.001% excess return with SE 0.0003%

- t-statistic $\approx$ 3.3 (statistically significant!)
- Economically meaningless after transaction costs

Ask: "Is the effect large enough to matter?" not just "Is $p < 0.05$?"

Pitfall 2: Non-Significance $\neq$ Zero Effect

Failure to reject the null does not mean the effect is zero.

Example: Estimate of $5\% \pm 8\%$

- Not statistically significant (confidence interval includes zero)
- But consistent with *large positive* or *small negative* effect
- **Data are inconclusive**, not proof of no effect

Pitfall 3: Comparing Significant vs Non-Significant

This subtle error pervades finance research:

Scenario:

- Strategy A: significant alpha ($t = 2.1$)
- Strategy B: not significant ($t = 1.8$)

Wrong conclusion: A and B differ meaningfully

Why? To compare them, test the *difference* : SE is roughly $\sqrt{2}$ times larger.

The difference between "significant" and "not significant" is not itself significant.

Pitfall 4: P-Hacking and Forking Paths

With enough flexibility in:

- Data processing
- Variable selection
- Model specification
- Sample period choice

...researchers can achieve $p < 0.05$ from almost any dataset : even pure noise.

The problem is not always conscious “fishing” but the accumulation of small, defensible choices.

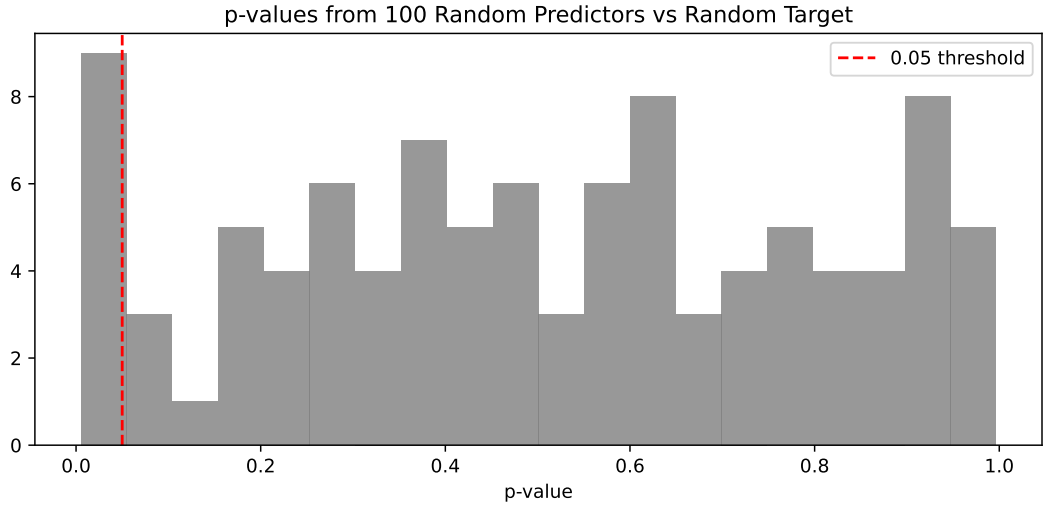
Pitfall 5: Publication Bias

When only "significant" results get published:

- Literature systematically overstates effect sizes
- A strategy that "works" in one study may be the lucky draw
- Meta-analyses inherit this bias

Remedy: Pre-registration, report all specifications tried, track trial counts.

The Multiple Testing Alarm



Even pure noise yields “significant” hits when you try enough ideas.

Part VI: The Bias-Variance Tradeoff

When Complexity Wins

- High-dimensional signals can reduce bias enough to dominate variance when markets are noisy and persistent (Kelly, Malamud, and Zhou (2024))
- Richer feature spaces capture stylised facts (volatility clustering, asymmetry) that simple models miss
- In finance, complexity often rescues weak signals : but only with disciplined regularisation

Kelly-Malamud-Zhou (2024): Key Finding

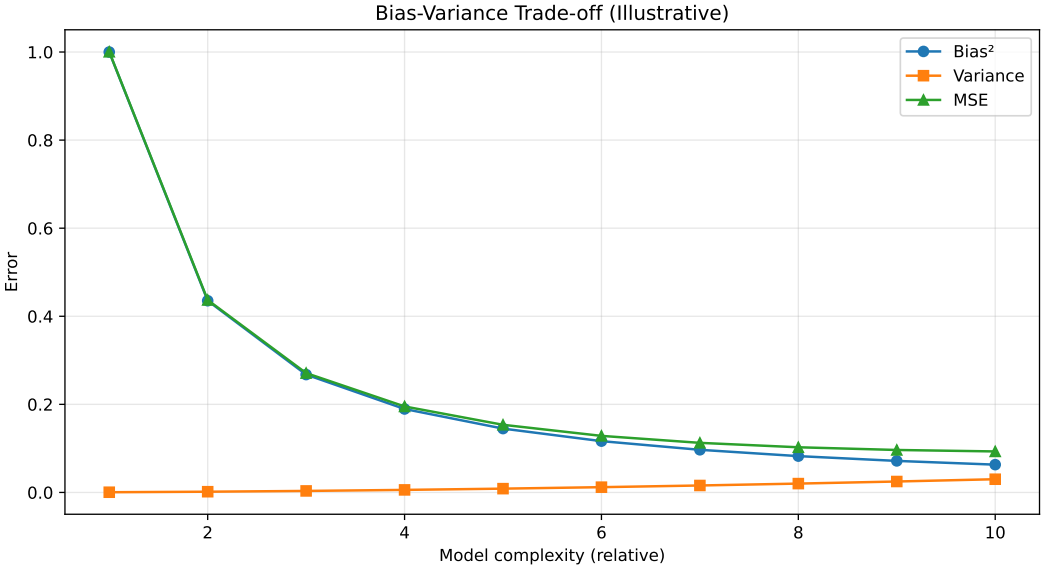
Setting: Return prediction with high-dimensional predictors

Core result: Under realistic financial DGPs, bias reduction from richer models can outweigh variance increases $\rightarrow$ better out-of-sample performance

Caveat: Complexity must be disciplined:

- Regularisation
- Time-aware validation
- Governance guardrails

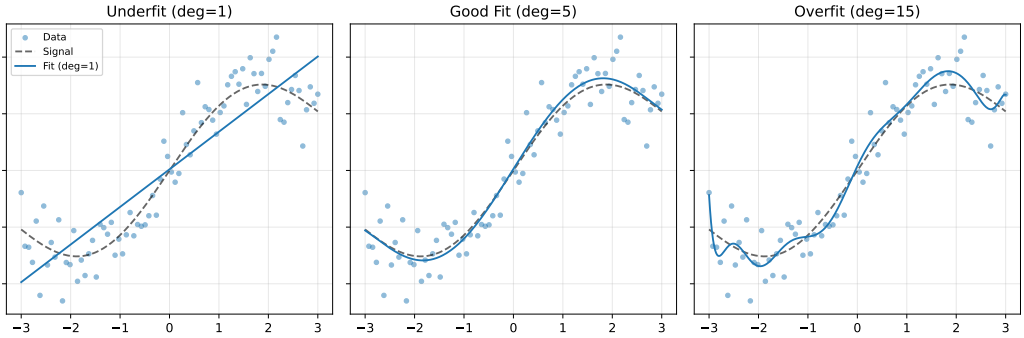
Bias-Variance Map (Illustrative)



Decision Flow: Should We Go Complex?

1. **Diagnose signal**: Do stylised facts or theory justify richer structure?
2. **Prototype** simple baseline; record validated error
3. **Escalate** complexity with shrinkage controls and time-aware validation
4. **Stress-test** governance: interpretability, robustness, operational load
5. **Document** evidence that complex model wins before promoting it

Underfit vs Good Fit vs Overfit



Part VII: Uncertainty Quantification

Frequentist vs Bayesian: A Pragmatic View

Frequentist	Bayesian
Parameters fixed, data repeatable	Parameters random, data observed
Control long-run error rates (tests, CIs)	Update beliefs via priors and likelihood
Great for regulatory benchmarks	Great when prior information and decisions matter
Confidence intervals have coverage guarantees	Credible intervals express posterior belief

Both use the same Kolmogorov axioms : choosing is about the question asked.

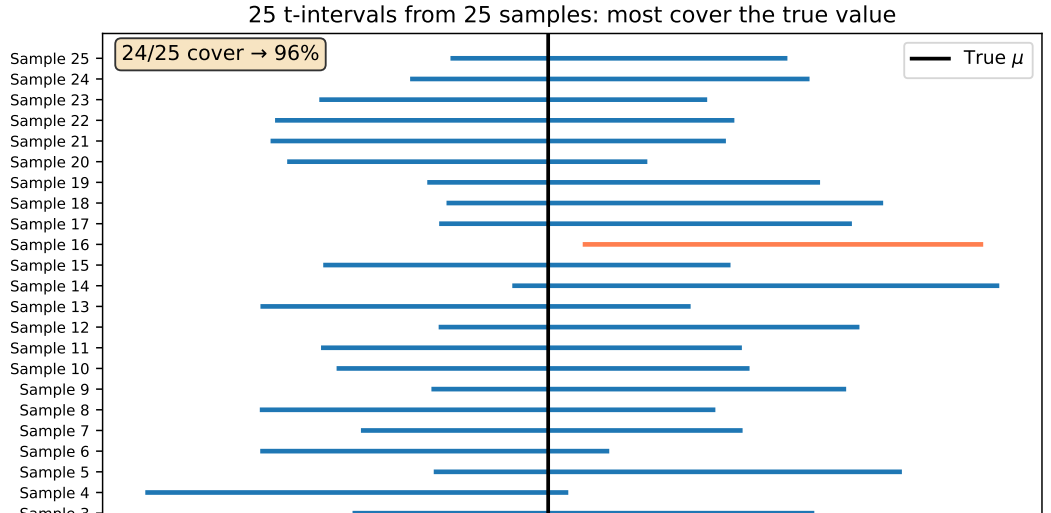
CI Belief

- Confidence interval: **procedure property**, not belief about the parameter
- 95% CI means: if we repeated sampling, 95% of intervals would contain the true value
- It does NOT mean: "I'm 95% confident the parameter is in this interval"

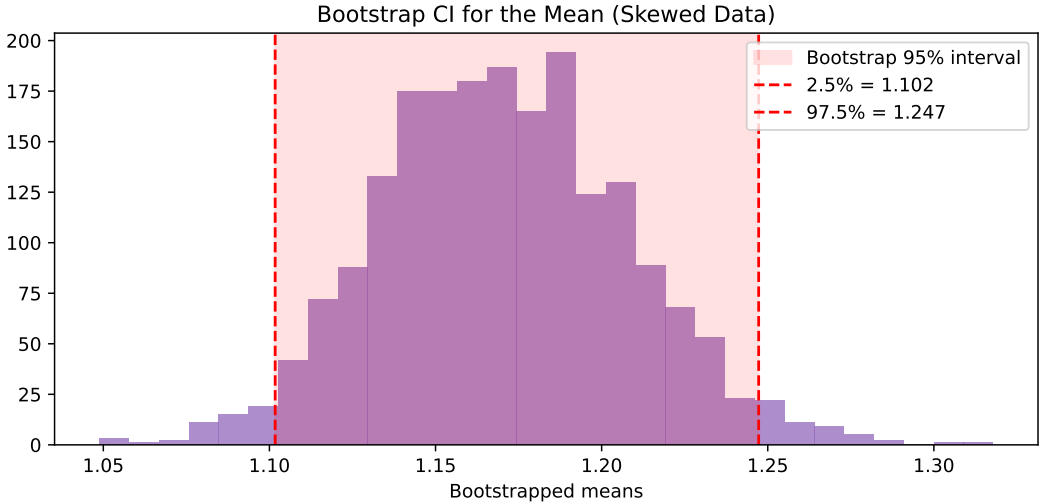
For belief statements, use Bayesian credible intervals.

95% CI: Many Intervals, One True Value

95% = *in the long run*, 95% of such intervals contain the true parameter : a **procedure** property, not a probability about *this* interval.

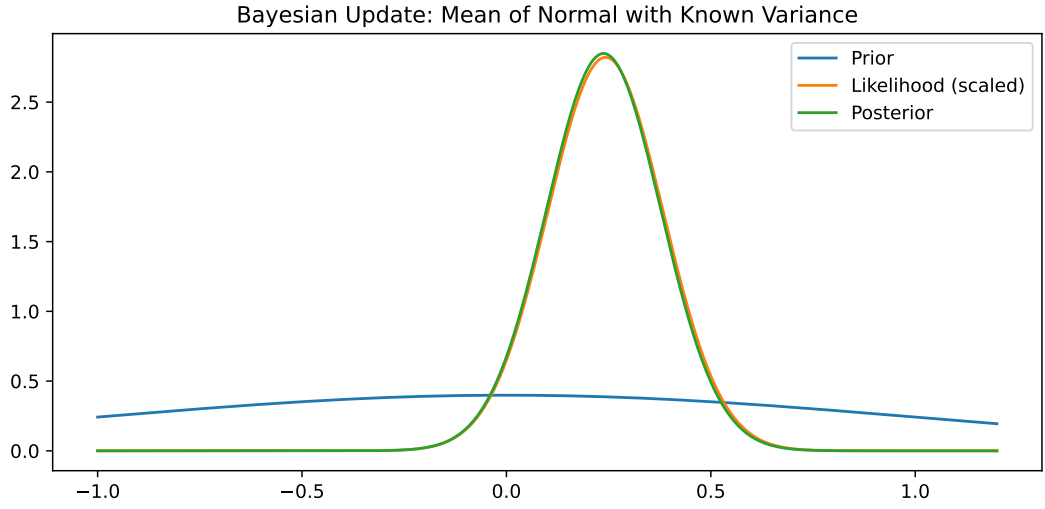


Bootstrap Confidence Interval



When textbook assumptions fail (e.g., skewed data), bootstrap provides evidence-based intervals.

Bayesian Update Example



Prior $\times$ Likelihood $\rightarrow$ Posterior: data updates our beliefs.

When Wide Credible Intervals Are Informative

Question: How much of return variance is predictable?

Fit Bayesian AR(1) to daily returns. Posterior for R^2 (squared autocorrelation):

Asset	Median R^2	95% Credible Interval
SPY	1.66%	[0.08%, 6.24%]
AAPL	0.73%	[0.02%, 2.94%]
BTCUSD	0.04%	[0.00%, 0.41%]

Wide intervals are the message, not a weakness:

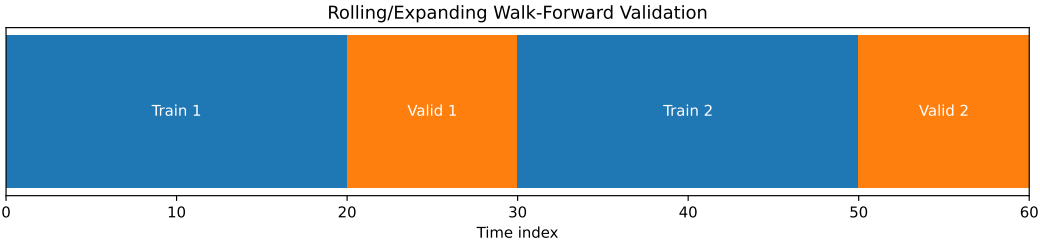
- Even at the **optimistic upper bound**, >93% of variance is noise
- We can't precisely measure how little predictability there is
- Point estimates hide this uncertainty; posteriors reveal it

Part VIII: Validation and Governance

Time-Aware Validation Workflow

1. **Profile data**; lock in feature engineering rules
2. **Design** rolling or expanding splits; document dates
3. **Run** walk-forward validation; log metrics and diagnostics
4. **Review** stability, governance notes, decision memos

Walk-Forward Validation (Illustrative)



Common Pitfalls in Financial Modelling

- **Look-ahead/leakage**: future information in features or targets
- **Survivorship bias**: current constituents only; delistings omitted
- **Data snooping**: p-hacking; post-selection inference ignored
- **Unrealistic frictions**: zero costs; fills at non-tradable prices
- **Time alignment**: mismatched timestamps; timezone/holiday drift

Backtest Overfitting: CSCV and PBO

- **CSCV**: split time into contiguous folds; pick in-sample "winner," test OOS
- **PBO**: fraction of splits where the winner underperforms out-of-sample
- **High PBO** = fragile discovery; reconsider selection or improve validation

Model Validation: Governance Checklist

- Define target leakage tests before modelling
- Align validation window with business decision horizon
- Stress-test hyper-parameters and features for stability
- Keep benchmark models live (naive, linear) for accountability

Part IX: Practical Tools and Tips

Fake-Data Simulation

Before trusting real results, test your procedure on fake data:

1. Specify data-generating process with known parameters
2. Simulate data from this process
3. Apply your estimation procedure
4. Compare estimates to true values
5. Repeat many times to assess variability

If your procedure can't recover known effects from fake data, don't trust it with real data.

Ten Tips for Better Regression Modelling

1. Think about variation in the data
2. Forget statistical significance; focus on practical significance
3. Graph the data relevant to your analysis
4. Interpret coefficients as *comparisons*, not effects
5. Understand your methods using fake-data simulation

Ten Tips (continued)

6. Fit many models to understand sensitivity
7. Set up a computational workflow that supports iteration
8. Use transformations to improve linearity and reduce outlier influence
9. Do targeted causal inference only when supported by design
10. Learn methods through real worked examples

Part X: From Here to the Module

From Primer to Weeks 1-4

- **Week 1**: Apply complexity logic to cost persistence and regulatory data
- **Week 2**: Data quality, measurement, survivorship bias
- **Week 3**: Volatility modelling and time series foundations
- **Week 4**: Robo-advice models with reproducibility + governance checklists

Datasets We'll Use

- **Bloomberg database** (practice and labs)
- **JKP factors** (Coursework 2)
- **Simulated data** (for understanding methods)

Directed Learning & Reflection

- **Core**: Read Kelly, Malamud, and Zhou (2024) and Efron and Hastie (2016) overview; rerun one demo; summarise bias-variance insights
- **Optional extension**: Add Gelman, Hill, and Vehtari (2020) causal framing; write a short reflection on when complexity is virtuous

Recap: Data Science as Statistical Science

- Data science is **statistical science** : the disciplined study of variation and uncertainty
- Regression coefficients are **comparisons**, not effects
- Complexity can win, but only with **validation** and **governance**
- Communicate uncertainty honestly; avoid the five pitfalls of significance

Exit Ticket

- Write down one dataset where you will trial a complex model : and why
- Capture the validation design you will use to prove it beats the baseline
- Note a governance or ethical concern you will monitor as you scale

References

Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press. https://hastie.su.stanford.edu/CASI/.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge, UK: Cambridge University Press. https://avehtari.github.io/ROS-Examples/.

Kelly, Bryan T., Semyon Malamud, and Kangying Zhou. 2024. "The Virtue of Complexity in Return Prediction." *Journal of Finance* 79 (1): 459–503. https://doi.org/10.1111/jofi.13298.