Week 10: Factor Replication : Principles & Critical Analysis

Learning Objectives

- Explain factor replication as a research methodology for testing published findings
- Interpret HAC (Heteroskedasticity and Autocorrelation Consistent) standard errors and understand why time-series autocorrelation matters
- Evaluate factor performance using multiple metrics (Sharpe, alpha, robustness)
- Identify sources of selection bias and overfitting in factor research
- Apply critical thinking to assess whether documented factors are exploitable

Agenda

Part I : What is factor replication? Research methodology foundations
Part II : Statistical foundations: HAC errors, alpha tests, robustness
Part III : Selection bias and the replication crisis in finance
Part IV : Critical analysis: What makes interpretation rigorous?
Part V : Preparation for Coursework 2: Principles, not templates

Part I : Factor Replication as Research Methodology

What Are Factors?

Factors are the building blocks of modern quantitative investing. Rather than picking individual stocks, factor strategies systematically buy characteristics that historically generate excess returns.

Definition: Characteristics that explain cross-sectional variation in stock returns

Classic examples:

- **Value (HML)**: High Minus Low book-to-market : buy undervalued stocks, sell overvalued (1992)
- **Momentum (MOM)**: Buy past 6-12 month winners, sell losers (behavioural persistence) (1993)
- **Size (SMB)**: Small Minus Big : small-cap premium (though weakening post-publication) (1981; Fama and French 1992)
- **Quality (RMW)**: Robust Minus Weak profitability : sustainable competitive advantages (2013; Fama and French 2015)

Long-Short Construction: Zero-Investment Portfolios

Factors are constructed as **long-short portfolios**: simultaneously buying one group and selling another. This isolates factor exposure from market movements.

Mechanics:

- **Long leg**: Buy stocks with desired characteristic (e.g., high book-to-market = value)
- **Short leg**: Sell stocks with opposite characteristic (e.g., low book-to-market = growth)
- **Equal weights**: Long and short legs have equal dollar amounts
- **Net investment**: £0 (long purchases offset short sales)

Example: Value Factor (HML)

Component	Valuation	Action	Investment	Return
Long	Undervalued (high B/M)	Buy £100 value stocks	-£100	+£5 (5%)
Short	Overvalued (low	Sell £100	+£100	-£2 (-2%)

Factor Replication: What Does It Mean?

Replication is core scientific practice. In medicine, we demand multiple trials before approving drugs. In finance, if a factor works 1970-1990, we demand it works 1991-2020.

Replication = reproduce published findings using independent data or time periods

Why replicate?

- Test whether published results are real or data mining artifacts
- Assess out-of-sample performance (does it work on new data?)
- Evaluate economic significance (after costs, is there exploitable profit?)
- Understand robustness (does it work across markets, time periods, specifications?)

! Jensen, Kelly & Pedersen [-@jensen2024replication]: "Is There a Replication Crisis in Finance?"

- Tested 153 published factors using consistent methodology
- Many factors show 50% decline in out-of-sample performance
- Cross-region replication often fails (US factors don't work in Europe/Asia)

Factor Replication Workflow

This is a conceptual framework, not a mechanical recipe. The scaffold notebook implements these steps, but understanding **why** each matters separates a pass from a distinction.

Conceptual steps:

1. **Choose factor**: Select published factor with theoretical motivation
2. **Obtain data**: Download returns from JKP portal (https://jkpfactors.com)
3. **Descriptive analysis**: Mean, volatility, Sharpe ratio, cumulative returns
4. **Alpha test**: Regress factor on market using HAC standard errors
5. **Robustness checks**: Sample splits, subperiod analysis, cost adjustments
6. **Interpretation**: Is factor real? Exploitable after costs? What are limitations?

Each step requires judgment: what robustness checks matter depends on your specific factor

Part II : Statistical Foundations for Rigorous Replication

Signal and Noise in Financial Returns

Financial returns are inherently noisy. Even if a factor has true alpha, observed returns mix **signal** (predictable component) with **noise** (random variation). Standard errors help us distinguish signal from noise.

The challenge:

- **Signal**: True factor alpha (e.g., value stocks genuinely outperform)
- **Noise**: Random variation (luck, market shocks, measurement error)
- **Observed return** = Signal + Noise

Why this matters:

- With 20 years of monthly data (240 observations), noise can create spurious patterns
- A factor might appear significant just by chance (noise masquerading as signal)
- Standard errors quantify how much noise contaminates our signal estimate

Example: True alpha = 0% (no factor), but observed alpha = 0.5% monthly
→ Is this signal (real factor) or noise (lucky sample)?

Measuring Signal-to-Noise: Methodology

To quantify signal vs noise, we decompose return variance into **predictable** (signal) and **unpredictable** (noise) components using conditional expectations.

Econometrically Rigorous Construction:

For returns r_t, we define signal as the **predictable component** conditional on available information:

1. **Conditional Expectation Model**: $E[r_t|\mathcal{J}_{t-1}] = \alpha + \beta \cdot \text{market}_t$
   - Predictable component based on market exposure (CAPM)
   - Captures time-varying expected returns, not just constant mean
2. **Variance Decomposition**:
   - **Total Variance**: $\text{Var}(r_t) = \sigma^2$
   - **Signal Variance**: $\text{Var}(E[r_t|\mathcal{J}_{t-1}])$ (variance of conditional expectation)
   - **Noise Variance**: $\text{Var}(r_t - E[r_t|\mathcal{J}_{t-1}])$ (variance of residuals)
   - **Signal Fraction**: R^2 from prediction model (proportion of variance explained)
   - **Noise Fraction**: $1 - R^2$ (proportion unexplained)

Why This Is More Rigorous:

Real-World Signal-to-Noise: Bloomberg Data

Using real financial data (2018-2025) from Bloomberg Terminal, we apply the **econometrically rigorous** approach: signal = predictable component from CAPM regression.

Method: Regress each asset on market (SPY) to extract conditional expectation $E[r_t|\text{market}_t]$

Signal Fraction = R^2 from CAPM (variance explained by market exposure)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

_csv_path = data_root / "bloomberg_database" / "signal_noise_metrics.csv"
if not _csv_path.exists():
    _csv_path = Path("data/bloomberg_database/signal_noise_metrics.csv")
metrics = pd.read_csv(_csv_path)

# Select example assets
```

Why Standard Errors Matter

Standard errors quantify uncertainty in estimates. In factor replication, we're testing whether observed alpha is **signal** (true factor premium) or **noise** (random variation).

Connection to signal-to-noise analysis:

- **Individual stocks** (AAPL): R^2 0.55 $\rightarrow$ 45% noise $\rightarrow$ larger standard errors
- **Factors** (long-short portfolios): R^2 0.1-0.2 $\rightarrow$ 80-90% noise $\rightarrow$ **much larger** standard errors
- **Implication**: Factor alpha estimates are less precise than stock alpha estimates

Statistical significance = "Is observed alpha signal or noise?"

t-statistic = Alpha / Standard Error

- $|t| > 1.96 \rightarrow$ statistically significant at 5% level (conventional threshold)
- $|t| < 1.96 \rightarrow$ cannot reject null hypothesis (could be random chance)
- Harvey (2017) recommends $t > 3$ for finance (multiple testing correction)

Why factors need higher t-statistics: With 80-90% noise fraction, standard errors are large. Need $t > 3$ to confidently distinguish signal from noise.

How Standard Errors Are Constructed

Standard errors measure both **estimation precision** and **sampling variability**. They are calculated from the variance-covariance matrix (error variance ÷ sample size), but their frequentist interpretation is as the standard deviation of the sampling distribution under hypothetical repeated sampling. High noise variance → large standard errors → imprecise estimates.

Basic OLS standard error formula:

For regression coefficient $\hat{\beta}$:

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}}$$

where $\hat{\sigma}^2$ is estimated error variance.

Key components:

- **Error variance ($\hat{\sigma}^2$)**: How much returns deviate from predicted values

Time-Series Data: Autocorrelation Problem

Financial time-series violate a key OLS assumption: **independence of observations**. If monthly returns are correlated, you don't have 120 independent observations over 10 years: you have fewer "effective" observations.

Financial returns exhibit serial correlation:

- Momentum: Positive returns predict future positive returns (6-12 months)
- Volatility clustering: High volatility today predicts high volatility tomorrow
- Market regimes: Bull and bear markets persist over time

Problem for inference:

- Standard OLS assumes observations are independent (and uncorrelated)
- Autocorrelation breaks this assumption
- Result: OLS **understates standard errors** $\rightarrow$ inflates t-statistics $\rightarrow$ **false positives**

⚠ Impact: HAC (Newey-West) standard errors typically 1.5-2$\times$ larger than OLS for monthly factors

Detecting Autocorrelation: Bloomberg Data Evidence

- Using real financial data, we can **measure** autocorrelation and test whether it's statistically significant. This demonstrates why HAC corrections are essential.

> **Autocorrelation Function (ACF)**: Correlation between r_t and r_{t-k} for lags $k = 1, 2, ...$

Ljung-Box Test: Tests null hypothesis of no autocorrelation up to lag k

- H : No autocorrelation (= = ... = = 0)
- If p-value < 0.05 → reject H → autocorrelation present → OLS SEs are wrong

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.stats.diagnostic import acorr_ljungbox
from statsmodels.graphics.tsaplots import plot_acf

df = load_bloomberg()
```

HAC vs OLS Standard Errors: Practical Impact

Let's compare OLS and HAC standard errors using a CAPM regression on real Bloomberg data. The difference shows why HAC is essential.

Methodology: Regress asset returns on market (SPY) using both OLS and HAC standard errors

```
import pandas as pd
import numpy as np
import statsmodels.api as sm

df = load_bloomberg()

# Get market (SPY) and asset (AAPL) returns
spy_data = df[df['ticker'] == 'SPY'].sort_values('date')
aapl_data = df[df['ticker'] == 'AAPL'].sort_values('date')

# Merge to align dates
merged = pd.merge(
```

Alpha Tests: CAPM Regression

The CAPM alpha test decomposes factor returns into two components: market exposure ( ) and excess return beyond market ( ). **Only alpha matters**: beta just tells you market risk.

Capital Asset Pricing Model regression:

$$R_{factor,t} = \alpha + \beta \cdot R_{market,t} + \varepsilon_t$$

Interpretation:

- **Alpha ( )**: Excess return not explained by market exposure ("skill" component)
- **Beta ( )**: Factor's sensitivity to market movements
- **R^2**: Fraction of factor variance explained by market
- **Null hypothesis**: $= 0$ (no excess return beyond market)

Example: Momentum earns 1.2% monthly, beta = 0.2, market earns 0.8% monthly

$\rightarrow$ CAPM predicts momentum return = $+ 0.2 \times 0.8\% = + 0.16\%$
$\rightarrow$ Observed return 1.2%, so **$= 1.04\%$ monthly** (if HAC $t > 1.96$, it's significant)

Robustness: Why One Test Isn't Enough

A single significant result is weak evidence. Researchers have many degrees of freedom: what Gelman and Loken (2014) call the "garden of forking paths": if you try enough specifications, one will appear significant by chance. Robustness guards against false discoveries.

Robustness checks test if results hold under alternative specifications:

- **Sample split**: Does factor work in first half AND second half? (minimum check)
- **Subperiod analysis**: Does alpha remain positive in each decade?
- **Alternative construction**: Tertiles vs. quintiles, value-weighted vs. equal-weighted?
- **Cross-region**: US factors often don't replicate in Europe/Asia (Jensen, Kelly, and Pedersen 2024)
- **Transaction costs**: Is net alpha positive after 0.2-0.5% monthly costs?

! Ethical Econometrics

Don't cherry-pick checks that passed. If factor works 2000-2010 but not 2010-2020, **report both**. Selective reporting is a breach of research ethics: transparent,

Part III : Selection Bias and the Replication Crisis

The Multiple Testing Problem

This is the core problem creating the replication crisis. With 5% significance threshold, testing 100 hypotheses generates ~5 false positives even if all nulls are true.

Academic research process (the problem):

1. Researcher tests 50 potential factors
2. 45 don't work (0, not significant)
3. 5 appear significant (> 0, t > 2) **by chance** (5% false positive rate)
4. Researcher publishes the 5 "successful" factors
5. Failed tests go in file drawer (never published)
6. Journals prefer positive results; null results don't advance careers

Result: Published literature massively overrepresents spurious findings

Harvey (2017) estimates over 300 equity factors published, but only ~10-15 are genuinely robust (95% are questionable)

Simulation: The Multiple Testing Problem in Action

Setup: 1,000 researchers each test 10 factors. **All factors are pure noise** (true $= 0$). At 5% significance level, how many "discoveries" emerge?

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

np.random.seed(42)

# Simulation parameters
n_researchers = 1000
factors_per_researcher = 10
n_months = 240  # 20 years of monthly data
significance_level = 0.05
true_alpha = 0.0  # ALL factors are noise (null is true)

# Simulate: each factor is pure noise
```

Guarding Against Selection Bias

Selection bias is hard to eliminate but can be mitigated with rigorous practices. These separate good research from bad.

Best practices in research:

- **Pre-registration**: Specify hypothesis before seeing data (medical trials standard)
- **Out-of-sample testing**: Test on data not available when factor was published
- **Cross-region replication**: Factors should work globally if they're real
- **Multiple testing corrections**: Use $t > 3$ threshold (Harvey 2017) instead of $t > 2$
- **Economic theory**: Value/momentum have theoretical foundations; "vowel tickers outperform" doesn't

i For Coursework 2: Intellectual Honesty Earns Marks

- If you tried 3 factors, disclose that (don't pretend you tested only 1)
- Report robustness failures, not just successes
- If alpha is $t = 2.1$, acknowledge it's marginal, not "strong evidence"
- **The 35% Critical Analysis component explicitly rewards honest**

Part IV : Critical Analysis: What Makes Interpretation Rigorous?

Beyond Reporting Numbers: Ask Questions

Weak analysis (reporting):

"Value factor earns 0.5% monthly alpha with $t = 2.3$ (significant). Sharpe ratio is 0.4. Results are robust to sample split."

Strong analysis (interpretation):

"Value earns 0.5% monthly alpha (6% annualised). This is economically meaningful but modest. Statistical significance ($t = 2.3$) suggests it's not pure luck, but close to threshold. Sample split shows alpha is stable (0.6% first half, 0.4% second half), increasing confidence. However, transaction costs (~0.2% monthly for value rebalancing) would reduce net alpha to 0.3% (3.6% annualised). Is 3.6% net alpha sufficient to compensate for tracking error and implementation frictions? Original paper reported 8% annualised: our replication shows 50% lower alpha, consistent with post-publication decline documented by Jensen et al. (2024)."

Interpreting Your Factor Results

When you analyse your factor's performance, interrogate your conclusions:

- **Statistical vs economic significance**: A t-stat of 2.1 clears the 1.96 threshold: but how confident would you be investing real money on that evidence?
- **Scale matters**: What does 0.1% monthly alpha actually mean for an investor over a year? Is that worth pursuing?
- **Benchmarking performance**: If the market delivers a Sharpe ratio around 0.4, what should you conclude about a factor with Sharpe of 0.3?
- **Robustness integrity**: If some of your robustness tests pass and others fail, what story does that tell about your factor?
- **From paper to portfolio**: What happens between calculating returns on a spreadsheet and actually implementing a trading strategy?

i The Implementation Gap

Academic factor returns assume frictionless trading. Real portfolios face transaction costs, market impact, and timing constraints. How might these affect your conclusions?

Part V : Preparation for Coursework 2: Principles, Not Templates

What the Scaffold Provides vs. What You Must Provide

The scaffold notebook is deliberately comprehensive: we want you to focus on **understanding and interpretation**, not debugging code. The 35% Critical Analysis component is where marks are won or lost.

Scaffold provides (execution):

- Working code for data loading, alpha regression, robustness checks
- All necessary functions pre-written (HAC standard errors, sample splits)
- Publication-quality tables and figures ready for your report

You must provide (interpretation grounded in YOUR results):

- **Numerical engagement**: "My alpha is X bp/month (t = Y). The original paper found Z bp. This N% difference likely reflects..."
- **Specific robustness narrative**: Which tests passed? Which failed? What does *that specific pattern* tell you?
- **Your judgment, defended**: Would you invest £10,000 of your own money in this factor? Why or why not, given YOUR numbers?
- **Process reflection**: What did you expect to find? What surprised you? What

Questions to Ask About YOUR Results

When you have your output, interrogate it. These questions connect today's principles to YOUR specific analysis.

About your methodology:

- Did your robustness tests pass or fail? What does *that specific pattern* suggest?
- How does your sample period compare to the original paper's? Does that explain any differences?

About your statistics:

- Is your t-stat comfortably above 2, or hovering near the threshold? What's the practical difference?
- Your alpha is X bp monthly. What does that mean for a £10m portfolio over a year?

About your judgment:

- Given YOUR numbers, would you recommend this factor to a pension fund? Why or why not?
- If your results are weaker than the original paper, is that replication failure: or

Using AI Tools Appropriately

AI tools like ChatGPT and Copilot are permitted: but how you use them determines whether they help or hurt your work.

AI can help you:

- Understand concepts: "Explain HAC standard errors in simple terms"
- Debug code: "Why is this pandas merge failing?"
- Learn techniques: "Show me how to calculate Newey-West standard errors"

AI cannot help you:

- Interpret YOUR specific results: It doesn't know your alpha is 0.28% with t = 1.9
- Explain YOUR robustness pattern: It can't see that your early-sample passed but late-sample failed
- Defend YOUR recommendation: Generic "factors can be useful" isn't a position

⚠ The Specificity Test

If your critical analysis section could have been written without ever looking at your

AI Use: What Helps vs. What Hurts

Appropriate use (helps your learning):

Task	Example prompt	Why it's fine
Concept clarification	"Explain why autocorrelation inflates t-statistics"	Builds understanding
Code debugging	"Why does my HAC calculation give NaN?"	Technical problem-solving
Writing feedback	"Is this paragraph clear?"	Improves communication

Problematic use (hurts your marks):

Task	Example prompt	Why it fails
Generic interpretation	"Write a limitations section for a factor study"	Not specific to YOUR results
Boilerplate	"Explain what alpha significance	Could apply to ANY study

Academic Integrity: Detection and Verification

To maintain fairness for all students, I have developed a **multi-model GenAI detection architecture** that analyses submission patterns across multiple dimensions.

What this means:

- All coursework submissions are processed through this system
- The system flags submissions with characteristics suggesting over-reliance on AI-generated content
- Flags are reviewed by me personally: the system assists, it doesn't decide

! Oral Examination Rights

I reserve the right to **orally examine** any student whose submission is flagged by this system. You may be asked to explain your analysis, walk through your reasoning, and demonstrate understanding of your own work.

This is not about catching you out: it's about ensuring your degree means something. Students who genuinely engage with their analysis have nothing to worry about.

Demonstration: HAC Standard Errors in Practice

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.regression.linear_model import OLS
from statsmodels.stats.sandwich_covariance import cov_hac

# Simulate factor returns with autocorrelation (illustrative)
np.random.seed(42)
n = 240  # 20 years monthly
market = np.random.normal(0.008, 0.04, n)
# Factor with positive alpha and autocorrelation
factor = 0.005 + 0.3 * market + np.random.normal(0, 0.03, n)
for i in range(1, n):
    factor[i] += 0.3 * factor[i-1]  # Autocorrelation

# Regression
X = sm.add_constant(market)
```

Next Steps: Week 11 Preview

Week 11 complements today by covering the prediction pathway (Coursework 2 Option B). Same principle-focused approach: understanding concepts, not copying templates.

Week 11 focus: Market prediction using factors

- Predict next-month market return using lagged factor data
- Compare OLS vs regularised models (ridge regression handles multicollinearity)
- Walk-forward validation for honest out-of-sample testing (prevents look-ahead bias)
- Evaluate predictive power: R^2 OOS, directional accuracy, economic value

Same pedagogical philosophy:

- Principles and understanding over step-by-step instructions
- Critical interpretation over mechanical execution
- Preparation for 35% Critical Analysis component

Connection: Replication (today) tests if factors exist. Prediction (next week) tests if factors forecast returns. Both require rigorous methodology and critical interpretation.

Summary: Week 10 Key Takeaways

Today provided foundational principles for factor replication. The core message: **understanding principles enables critical analysis**. Scaffold gives outputs; understanding gives interpretation.

Methodology:

- Factor replication tests if published findings are real (out-of-sample, robustness required)
- Jensen et al. (2024): Many factors show 50% decline in performance due to selection bias

Statistical foundations:

- HAC standard errors correct for autocorrelation (typically 1.5-2$\times$ larger than OLS)
- Alpha isolates excess returns beyond market exposure
- Robustness checks (sample split, subperiod, costs) separate signal from noise

Critical analysis earns marks:

- Interpret economically: 0.5% monthly = 6% annualised, but is it meaningful after