# Applying Tidy Finance with Python to Vietnam

Mike

2026-02-01

# Table of contents

**Preface** **3**
Motivation . . . 3
Why Emerging Markets Require Different Empirical Infrastructure . . . 4
Reproducibility as a Research Design Principle . . . 4
Vietnam as a Case, Not an Exception . . . 5
Data Access . . . 5
Contribution and Audience . . . 6
Structure of the Book . . . 6

**1 Institutional Background and Market Structure of Vietnam's Equity Market** **7**
1.1 Evolution of Vietnam's Equity Market . . . 7
1.2 Exchange Structure and Trading Mechanisms . . . 8
1.3 Listing Requirements and Firm Characteristics . . . 8
1.4 Investor Composition and Trading Behavior . . . 8
1.5 Regulatory Environment and Market Frictions . . . 9
1.6 Implications for Empirical Design . . . 9
1.7 Summary . . . 10

**2 Constructing and Analyzing Equity Return Series** **11**
2.1 Data Access and Preparation . . . 11
2.1.1 Prerequisites for API Access . . . 11
2.1.2 Checking Your IP Address . . . 13
2.1.3 Fetching the Dataset . . . 14
2.2 Examining a Single Equity . . . 15
2.3 From Prices to Returns . . . 17
2.4 Limiting the Influence of Extreme Returns . . . 18
2.5 Distributional Features of Returns . . . 18
2.6 Expanding to a Market Cross-Section . . . 20
2.7 Aggregating Returns Across Time . . . 23
2.8 Aggregation Across Firms: Trading Activity . . . 24
2.9 Summary . . . 26

**3 Modern Portfolio Theory** **27**
3.0.1 The Core Insight: Diversification as a Free Lunch . . . 27
3.0.2 The Mean-Variance Framework . . . 28

3.1 The Asset Universe: Setting Up the Problem . . . 28
3.1.1 The Two Stages of Portfolio Selection . . . 29
3.1.2 Loading and Preparing the Data . . . 29
3.1.3 Computing Expected Returns . . . 31
3.1.4 Computing Volatilities . . . 32
3.1.5 Visualizing the Risk-Return Trade-off . . . 32
3.2 The Variance-Covariance Matrix: Capturing Asset Interactions . . . 34
3.2.1 Why Correlations Matter . . . 34
3.2.2 Computing the Variance-Covariance Matrix . . . 34
3.2.3 Interpreting the Variance-Covariance Matrix . . . 35
3.3 The Minimum-Variance Portfolio . . . 36
3.3.1 Motivation: Risk Minimization as a Benchmark . . . 36
3.3.2 The Optimization Problem . . . 37
3.3.3 The Analytical Solution . . . 37
3.3.4 Implementation . . . 38
3.3.5 Visualizing the Minimum-Variance Weights . . . 38
3.3.6 Portfolio Performance . . . 39
3.4 Efficient Portfolios: Balancing Risk and Return . . . 40
3.4.1 The Investor's Trade-off . . . 40
3.4.2 Setting the Target Return . . . 41
3.4.3 The Analytical Solution . . . 41
3.4.4 Implementation . . . 42
3.4.5 Comparing the Portfolios . . . 42
3.4.6 The Role of Risk Aversion . . . 43
3.5 The Efficient Frontier: The Menu of Optimal Portfolios . . . 44
3.5.1 The Mutual Fund Separation Theorem . . . 44
3.5.2 Proof of the Separation Theorem . . . 44
3.5.3 Computing the Efficient Frontier . . . 45
3.5.4 Visualizing the Efficient Frontier . . . 45
3.6 Key Takeaways . . . 47

**4 The Capital Asset Pricing Model 48**
4.1 From Efficient Portfolios to Equilibrium Prices . . . 48
4.2 Systematic versus Idiosyncratic Risk . . . 49
4.2.1 Idiosyncratic Risk: Diversifiable and Unrewarded . . . 49
4.2.2 Systematic Risk: Undiversifiable and Priced . . . 49
4.2.3 A Simple Illustration . . . 49
4.3 Data Preparation . . . 50
4.3.1 Computing Monthly Returns . . . 52
4.4 The Risk-Free Asset and the Investment Opportunity Set . . . 52
4.4.1 Adding a Risk-Free Asset . . . 52
4.4.2 Portfolio Return with a Risk-Free Asset . . . 53
4.4.3 Portfolio Variance . . . 53

4.4.4 Setting Up the Risk-Free Rate . . . 53
4.5 The Tangency Portfolio: Where Everyone Invests . . . 55
4.5.1 Deriving the Optimal Risky Portfolio . . . 55
4.5.2 The Tangency Portfolio . . . 56
4.5.3 The Sharpe Ratio and the Capital Market Line . . . 56
4.5.4 Computing the Tangency Portfolio . . . 57
4.5.5 Visualizing the Efficient Frontier with a Risk-Free Asset . . . 58
4.6 The CAPM Equation: Risk and Expected Return . . . 60
4.6.1 From Individual Optimization to Market Equilibrium . . . 60
4.6.2 The Market Portfolio . . . 60
4.6.3 Deriving the CAPM Equation . . . 60
4.6.4 Interpreting Beta . . . 61
4.7 The Security Market Line . . . 62
4.8 Empirical Estimation of CAPM Parameters . . . 63
4.8.1 The Regression Framework . . . 63
4.8.2 Alpha: Risk-Adjusted Performance . . . 64
4.8.3 Loading Factor Data . . . 64
4.8.4 Running the Regressions . . . 65
4.8.5 Visualizing Alpha Estimates . . . 65
4.9 Limitations and Extensions . . . 67
4.9.1 The Market Portfolio Problem . . . 67
4.9.2 Time-Varying Betas . . . 68
4.9.3 Empirical Anomalies . . . 68
4.9.4 Multifactor Extensions . . . 68
4.10 Key Takeaways . . . 69

**5 Financial Statement Analysis 71**
5.1 From Market Prices to Fundamental Value . . . 71
5.2 The Three Financial Statements . . . 72
5.2.1 The Balance Sheet: A Snapshot of Financial Position . . . 72
5.2.2 The Income Statement: Performance Over Time . . . 73
5.2.3 The Cash Flow Statement: Following the Money . . . 74
5.2.4 Illustrating with FPT's Financial Statements . . . 74
5.3 Loading Financial Statement Data . . . 75
5.4 Liquidity Ratios: Can the Company Pay Its Bills? . . . 76
5.4.1 The Current Ratio . . . 76
5.4.2 The Quick Ratio . . . 77
5.4.3 The Cash Ratio . . . 77
5.4.4 Calculating Liquidity Ratios . . . 77
5.4.5 Cross-Sectional Comparison of Liquidity . . . 78
5.5 Leverage Ratios: How Is the Company Financed? . . . 79
5.5.1 Why Capital Structure Matters . . . 79
5.5.2 Debt-to-Equity Ratio . . . 80

5.5.3 Debt-to-Asset Ratio . . . 80
5.5.4 Interest Coverage Ratio . . . 80
5.5.5 Calculating Leverage Ratios . . . 81
5.5.6 Leverage Trends Over Time . . . 81
5.5.7 Cross-Sectional Leverage Comparison . . . 82
5.5.8 The Leverage-Coverage Trade-off . . . 83
5.6 Efficiency Ratios: How Well Are Assets Managed? . . . 85
5.6.1 Asset Turnover . . . 85
5.6.2 Inventory Turnover . . . 86
5.6.3 Receivables Turnover . . . 86
5.6.4 Calculating Efficiency Ratios . . . 86
5.7 Profitability Ratios: Is the Company Making Money? . . . 87
5.7.1 Gross Margin . . . 87
5.7.2 Profit Margin . . . 88
5.7.3 Return on Equity (ROE) . . . 88
5.7.4 The DuPont Decomposition . . . 88
5.7.5 Calculating Profitability Ratios . . . 89
5.7.6 Gross Margin Trends . . . 89
5.7.7 From Gross to Net: Where Do Profits Go? . . . 90
5.8 Combining Financial Ratios: A Holistic View . . . 91
5.8.1 Ranking Companies Across Categories . . . 92
5.9 Financial Ratios in Asset Pricing . . . 94
5.9.1 The Fama-French Factors . . . 94
5.9.2 Calculating Fama-French Variables . . . 95
5.9.3 Fama-French Factor Rankings . . . 96
5.10 Limitations and Practical Considerations . . . 97
5.10.1 Accounting Discretion . . . 97
5.10.2 Industry Comparability . . . 98
5.10.3 Point-in-Time Limitations . . . 98
5.10.4 Backward-Looking Nature . . . 98
5.10.5 Quality of Earnings . . . 98
5.11 Key Takeaways . . . 98

**6 Discounted Cash Flow Analysis 100**
6.1 What Is a Company Worth? . . . 100
6.1.1 Valuation Methods Overview . . . 100
6.1.2 The Three Pillars of DCF . . . 101
6.2 Understanding Free Cash Flow . . . 101
6.2.1 Why Free Cash Flow, Not Net Income? . . . 101
6.2.2 The Free Cash Flow Formula . . . 102
6.3 Loading Historical Financial Data . . . 102
6.3.1 Computing Historical Free Cash Flow . . . 103
6.3.2 Understanding the Historical Pattern . . . 105

6.4 Visualizing Historical Ratios . . . 106
6.5 Forecasting Free Cash Flow . . . 108
6.5.1 The Ratio-Based Forecasting Approach . . . 108
6.5.2 Setting Forecast Assumptions . . . 109
6.5.3 Forecasting Revenue Growth . . . 110
6.5.4 Building the Forecast . . . 111
6.6 Visualizing the Forecast . . . 112
6.7 Terminal Value: Capturing Long-Term Value . . . 117
6.7.1 The Perpetuity Growth Model . . . 117
6.7.2 Choosing the Perpetual Growth Rate . . . 118
6.7.3 Alternative: Exit Multiple Approach . . . 119
6.8 The Discount Rate: Weighted Average Cost of Capital . . . 119
6.8.1 Estimating WACC Components . . . 120
6.8.2 Using Industry WACC Data . . . 120
6.9 Computing Enterprise Value . . . 123
6.10 Sensitivity Analysis . . . 125
6.11 From Enterprise Value to Equity Value . . . 127
6.11.1 Implied Share Price . . . 128
6.12 Limitations and Practical Considerations . . . 128
6.12.1 Sensitivity to Assumptions . . . 129
6.12.2 Terminal Value Dominance . . . 129
6.12.3 Garbage In, Garbage Out . . . 129
6.12.4 Not Suitable for All Companies . . . 129
6.12.5 Complement with Other Methods . . . 129
6.13 Key Takeaways . . . 130

**7 Accessing and Managing VN Financial Data 131**
7.1 Overview of Vietnamese Financial Data Sources . . . 132
7.2 Stock Market Data . . . 134
7.2.1 Historical Price Data . . . 134
7.2.2 Fundamental Data and Financial Statements . . . 134
7.2.3 Corporate Actions and Events . . . 135
7.3 Market Indices and Benchmarks . . . 135
7.3.1 Index Constituent Data . . . 135
7.4 Macroeconomic Data from Vietnamese Sources . . . 136
7.4.1 Key Macroeconomic Indicators . . . 136
7.4.2 Risk-Free Rate Approximation . . . 137
7.5 Setting Up a Database for Vietnamese Financial Data . . . 138
7.5.1 Database Schema Design . . . 138
7.5.2 Storing Data . . . 138
7.6 Querying and Updating the Database . . . 139
7.6.1 Database Maintenance . . . 140

7.7 Alternative Data Sources for Vietnamese Markets . . . 141
7.7.1 Foreign Investor Flow Data . . . 141
7.7.2 News and Sentiment Data . . . 141
7.8 Key Takeaways . . . 141

**8 Datacore Data** **143**
8.1 Data Access Options . . . 143
8.2 Chapter Overview . . . 143
8.3 Setting Up the Environment . . . 144
8.4 Connecting to Datacore . . . 144
8.5 Company Fundamentals Data . . . 146
8.5.1 Understanding Vietnamese Financial Statements . . . 147
8.5.2 Downloading Fundamentals Data . . . 147
8.5.3 Cleaning and Standardizing Fundamentals . . . 148
8.5.4 Creating Standardized Variables . . . 149
8.5.5 Computing Book Equity and Profitability . . . 151
8.5.6 Computing Investment . . . 154
8.5.7 Computing Total Debt . . . 155
8.5.8 Applying Filters and Final Preparation . . . 156
8.5.9 Storing Fundamentals Data . . . 157
8.6 Stock Price Data . . . 157
8.6.1 Downloading Price Data . . . 157
8.6.2 Processing Price Data . . . 158
8.6.3 Computing Shares Outstanding and Market Capitalization . . . 159
8.6.4 Computing Returns and Excess Returns . . . 159
8.6.5 Storing Price Data . . . 168
8.7 Descriptive Statistics . . . 169
8.7.1 Market Evolution Over Time . . . 169
8.7.2 Market Capitalization Evolution . . . 170
8.7.3 Return Distribution . . . 171
8.7.4 Coverage of Book Equity . . . 172
8.8 Merging Stock and Fundamental Data . . . 174
8.9 Key Takeaways . . . 178

**9 Beta Estimation** **180**
9.1 Theoretical Foundation . . . 180
9.1.1 The Capital Asset Pricing Model . . . 180
9.1.2 Empirical Implementation . . . 181
9.2 Setting Up the Environment . . . 182
9.3 Loading and Preparing Data . . . 182
9.3.1 Stock Returns Data . . . 182
9.3.2 Company Information . . . 183
9.3.3 Market Excess Returns . . . 184

9.3.4 Merging Datasets . . . 184
9.3.5 Handling Outliers . . . 185
9.4 Estimating Beta for Individual Stocks . . . 186
9.4.1 Single Stock Example . . . 186
9.4.2 CAPM Estimation Function . . . 188
9.5 Rolling-Window Estimation . . . 190
9.5.1 Motivation for Rolling Windows . . . 190
9.5.2 Rolling Window Implementation . . . 190
9.5.3 Example: Rolling Betas for Selected Stocks . . . 192
9.5.4 Visualizing Rolling Betas . . . 193
9.6 Parallelized Estimation for the Full Market . . . 195
9.6.1 The Computational Challenge . . . 195
9.6.2 Setting Up Parallel Processing . . . 195
9.6.3 Parallel Beta Estimation . . . 195
9.6.4 Storing Results . . . 197
9.7 Beta Estimation Using Daily Returns . . . 197
9.7.1 Batch Processing for Daily Data . . . 198
9.8 Analyzing Beta Estimates . . . 202
9.8.1 Extracting Beta Estimates . . . 202
9.8.2 Summary Statistics . . . 202
9.8.3 Beta Distribution Across Industries . . . 203
9.8.4 Time Variation in Cross-Sectional Beta Distribution . . . 205
9.8.5 Coverage Analysis . . . 207
9.9 Comparing Monthly and Daily Beta Estimates . . . 208
9.10 Key Takeaways . . . 211

**10 Univariate Portfolio Sorts** **212**
10.1 Data Preparation . . . 212
10.2 Sorting by Market Beta . . . 213
10.3 Performance Evaluation . . . 214
10.4 Functional Programming for Portfolio Sorts . . . 216
10.5 More Performance Evaluation . . . 217
10.6 Security Market Line and Beta Portfolios . . . 219
10.7 Key Takeaways . . . 223

**11 Fama-French Factors** **225**
11.1 Theoretical Background . . . 225
11.1.1 The Evolution from CAPM to Multi-Factor Models . . . 225
11.1.2 The Five-Factor Extension . . . 226
11.1.3 Factor Construction Methodology . . . 226
11.2 Setting Up the Environment . . . 227
11.3 Data Preparation . . . 227
11.3.1 Loading Stock Returns . . . 227

11.3.2 Loading Company Fundamentals . . . 228
11.3.3 Constructing Sorting Variables . . . 228
11.3.4 Validating Sorting Variables . . . 231
11.3.5 Handling Outliers . . . 231
11.4 Portfolio Assignment Functions . . . 233
11.4.1 The Portfolio Assignment Function . . . 233
11.4.2 Assigning Portfolios for Three-Factor Model . . . 235
11.4.3 Validating Portfolio Assignments . . . 236
11.5 Fama-French Three-Factor Model (Monthly) . . . 238
11.5.1 Merging Portfolios with Returns . . . 238
11.5.2 Computing Value-Weighted Portfolio Returns . . . 239
11.5.3 Constructing SMB and HML Factors . . . 241
11.5.4 Computing the Market Factor . . . 242
11.5.5 Combining Three Factors . . . 243
11.5.6 Saving Three-Factor Data . . . 244
11.6 Fama-French Five-Factor Model (Monthly) . . . 245
11.6.1 Portfolio Assignments with Dependent Sorts . . . 245
11.6.2 Validating Five-Factor Portfolios . . . 246
11.6.3 Merging and Computing Portfolio Returns . . . 247
11.6.4 Constructing All Five Factors . . . 247
11.6.5 Factor Correlations . . . 250
11.6.6 Saving Five-Factor Data . . . 251
11.7 Daily Fama-French Factors . . . 251
11.7.1 Motivation for Daily Factors . . . 251
11.7.2 Loading Daily Returns . . . 252
11.7.3 Adding Market Cap for Daily Weighting . . . 252
11.7.4 Merging Daily Returns with Portfolios . . . 253
11.7.5 Computing Daily Three Factors . . . 255
11.7.6 Computing Daily Market Factor . . . 256
11.7.7 Combining Daily Three Factors . . . 257
11.7.8 Computing Daily Five Factors . . . 258
11.7.9 Saving Daily Factors . . . 261
11.8 Factor Validation and Diagnostics . . . 262
11.8.1 Cumulative Factor Returns . . . 263
11.8.2 Average Factor Premiums . . . 264
11.8.3 Comparing Monthly and Daily Factors . . . 265
11.9 Key Takeaways . . . 266

**12 Momentum Strategies** **268**
12.1 Theoretical Background . . . 268
12.1.1 The Momentum Effect . . . 268
12.1.2 Overlapping Portfolios and Calendar-Time Returns . . . 269
12.1.3 Theoretical Explanations . . . 269

12.1.4 Momentum in Emerging Markets . . . 270
12.2 Setting Up the Environment . . . 270
12.3 Data Preparation . . . 271
12.3.1 Loading Monthly Stock Returns . . . 271
12.3.2 Inspecting the Data . . . 272
12.3.3 Loading Factor Data . . . 273
12.3.4 Data Quality Filters . . . 274
12.4 Momentum Portfolio Construction . . . 274
12.4.1 Computing Formation Period Returns . . . 274
12.4.2 Assigning Momentum Decile Portfolios . . . 276
12.4.3 Defining Holding Period Dates . . . 279
12.4.4 Computing Portfolio Holding Period Returns . . . 279
12.4.5 Computing Equally Weighted Portfolio Returns . . . 280
12.5 Baseline Results: J=6, K=6 Strategy . . . 282
12.5.1 Summary Statistics by Momentum Decile . . . 282
12.5.2 The Long-Short Momentum Portfolio . . . 283
12.5.3 Cumulative Returns . . . 284
12.5.4 Monthly Return Distribution . . . 287
12.6 Extending to Multiple Formation and Holding Periods . . . 288
12.6.1 The J $\times$ K Grid . . . 288
12.6.2 Running the Full Grid . . . 291
12.6.3 Winners Portfolio Returns . . . 294
12.6.4 Losers Portfolio Returns . . . 295
12.6.5 Long-Short Momentum Returns . . . 296
12.6.6 Visualizing the Momentum Premium Across Specifications . . . 297
12.7 Risk-Adjusted Performance . . . 298
12.7.1 CAPM Alpha . . . 298
12.7.2 Fama-French Three-Factor Alpha . . . 299
12.7.3 Interpretation of Risk Exposures . . . 301
12.8 Momentum and Market States . . . 301
12.8.1 Conditional Performance . . . 301
12.9 Momentum Crashes . . . 304
12.9.1 Understanding Momentum Drawdowns . . . 304
12.9.2 Maximum Drawdown Analysis . . . 306
12.10 Value-Weighted Momentum Portfolios . . . 307
12.11 Daily Momentum Analysis . . . 313
12.11.1 Loading Daily Data . . . 313
12.11.2 Daily Returns of Monthly Momentum Portfolios . . . 314
12.11.3 Daily Cumulative Returns . . . 315
12.11.4 Annualized Risk Metrics from Daily Data . . . 316
12.11.5 Realized Volatility of Momentum Returns . . . 318
12.12 Saving Results to the Database . . . 319

12.13 Practical Considerations . . . 320
12.13.1 Transaction Costs . . . 320
12.13.2 Implementation Lag . . . 321
12.13.3 Survivorship Bias . . . 321
12.13.4 Small Sample Considerations . . . 321
12.14 Key Takeaways . . . 321

**13 Fama-MacBeth Regressions** **323**
13.1 The Econometric Framework . . . 323
13.1.1 Intuition: Why not Panel OLS? . . . 323
13.1.2 Mathematical Derivation . . . 324
13.2 Data Preparation . . . 325
13.3 Step 1: Cross-Sectional Regressions with WLS . . . 327
13.4 Step 2: Time-Series Aggregation & Hypothesis Testing . . . 328
13.4.1 Visualizing the Time-Varying Risk Premium . . . 330
13.5 Sanity Checks . . . 332
13.5.1 Time-Series Volatility Check . . . 332
13.5.2 Correlation of Characteristics (Multicollinearity) . . . 333

**14 Event Studies in Finance** **334**
14.0.1 Why Event Studies Matter . . . 334
14.1 Literature Review and Methodological Evolution . . . 335
14.1.1 The Classical Framework (1969-1985) . . . 335
14.1.2 Risk Model Refinements (1992-2015) . . . 335
14.1.3 Testing for Abnormal Returns (1976-2010) . . . 336
14.1.4 CARs versus BHARs . . . 337
14.1.5 Emerging Market Considerations . . . 337
14.2 Mathematical Framework . . . 337
14.2.1 Timeline and Windows . . . 337
14.2.2 Normal Return Models . . . 338
14.2.3 Aggregation: CARs and BHARs . . . 339
14.2.4 Standardized Returns . . . 339
14.2.5 Test Statistics . . . 339
14.3 Python Implementation . . . 341
14.3.1 Design Philosophy . . . 341
14.3.2 Setup and Imports . . . 341
14.3.3 Configuration . . . 343
14.3.4 Step 1: Trading Calendar Construction . . . 345
14.3.5 Step 2: Event Date Alignment . . . 346
14.3.6 Step 3: Data Extraction and Factor Merging . . . 347
14.3.7 Step 4: Risk Model Estimation . . . 349
14.3.8 Step 5: Abnormal Return Computation . . . 352
14.3.9 Step 6: Comprehensive Test Statistics . . . 354

14.3.10 Step 7: Publication-Ready Visualization . . . 357
14.3.11 The Master Pipeline . . . 359
14.4 Demonstration with Simulated Data . . . 361
14.4.1 Running the Full Pipeline . . . 363
14.4.2 Visualizing Results . . . 364
14.4.3 Complete Test Statistics . . . 366
14.4.4 Running Multiple Models for Robustness . . . 366
14.5 How to Use This Framework with Your Data . . . 368
14.5.1 Required Data Format . . . 368
14.5.2 Minimal Usage Example . . . 369
14.6 Demonstration with Vietnamese Market Data . . . 370
14.6.1 Loading the Data . . . 370
14.6.2 Creating Sample Events . . . 372
14.6.3 Daily Event Study: Fama-French 3-Factor Model . . . 373
14.6.4 Visualizing Daily Results . . . 374
14.6.5 Complete Test Statistics (Daily) . . . 376
14.6.6 Robustness: Multiple Risk Models (Daily) . . . 376
14.6.7 Robustness: Multiple Event Windows . . . 376
14.6.8 Monthly Event Study: Fama-French 3-Factor Model . . . 376
14.6.9 Daily Event Study: Fama-French 5-Factor Model . . . 382
14.6.10 Comparing FF3 vs FF5 Estimation Quality . . . 383
14.6.11 Event-Level Detail . . . 383
14.6.12 Daily Abnormal Return Dynamics . . . 383
14.6.13 Summary of Key Findings . . . 387
14.7 Practical Recommendations . . . 388

**15 Standardized Earnings Surprises (SUE) 390**
15.1 Methodology . . . 390
15.1.1 Method 1: Seasonal Random Walk . . . 390
15.1.2 Method 2: Exclusion of Special Items . . . 391
15.1.3 Method 3: Analyst Consensus . . . 391
15.2 Data Description . . . 391
15.2.1 Visualizing the Core Data . . . 392
15.3 Implementation . . . 392
15.3.1 Python Setup and Data Loading . . . 392
15.3.2 Calculation Logic . . . 394
15.4 Results and Analysis . . . 395
15.4.1 Tabular Results (FY 2024) . . . 395
15.4.2 Visualization . . . 396
15.5 Conclusion . . . 397

**16 Measuring Divergence of Investor Opinion 398**

**17 Theoretical Framework** **401**
17.1 The Miller (1977) Overpricing Hypothesis . . . 401
17.2 Alternative Theoretical Perspectives . . . 402
17.3 Relevance to the Vietnamese Market . . . 403

**18 Data Sources and Sample Construction** **405**
18.1 Data Sources . . . 405
18.2 Sample Construction . . . 405
18.3 Corporate Action Adjustments . . . 409
18.4 Trading Calendar Construction . . . 410

**19 Volume-Based DIVOP Proxies** **412**
19.1 Theoretical Motivation . . . 412
19.2 Unexplained Volume (DTO) . . . 412
19.2.1 Construction Methodology . . . 412
19.2.2 Vietnam-Specific Considerations for DTO . . . 414
19.3 Standardized Unexplained Volume (SUV) . . . 415
19.3.1 Construction Methodology . . . 415
19.3.2 Interpreting the SUV Regression Coefficients . . . 417

**20 Volatility-Based DIVOP Proxies** **419**
20.1 Total Return Volatility . . . 419
20.1.1 Theoretical Motivation . . . 419
20.1.2 Construction . . . 419
20.2 Idiosyncratic Volatility (IVOL) . . . 419
20.2.1 Vietnam-Specific Considerations for Volatility . . . 421

**21 Spread-Based and Liquidity DIVOP Proxies** **423**
21.1 Bid-Ask Spread (BASPREAD) . . . 423
21.1.1 Theoretical Motivation . . . 423
21.1.2 Construction . . . 423
21.2 Amihud Illiquidity (ILLIQ) . . . 423
21.2.1 Vietnam-Specific Considerations for Spread and Liquidity . . . 425

**22 Analyst Forecast Dispersion** **426**
22.1 Theoretical Motivation . . . 426
22.2 Data Challenges in Vietnam . . . 426
22.3 Construction Methodology . . . 426
22.4 Scaling Considerations . . . 429

**23 Cross-Sectional Correlations Among DIVOP Proxies** **430**
23.0.1 Expected Correlation Patterns . . . 431

**24 Descriptive Statistics and Cross-Sectional Properties** **432**
24.1 Summary Statistics . . . 432
24.2 DIVOP by Firm Characteristics . . . 433
24.3 Time-Series Evolution . . . 433

**25 Putting It All Together** **435**

**26 Empirical Applications** **437**
26.1 Application 1: DIVOP and the Cross-Section of Returns . . . 437
26.2 Application 2: DIVOP and Earnings Announcements . . . 438
26.3 Application 3: Composite DIVOP Index via PCA . . . 439

**27 Conclusion and Practical Recommendations** **441**

**28 P/E Ratio** **443**
28.0.1 Technical Schema and Variable Engineering . . . 443
28.1 Firm-Specific P/E Valuation Metrics . . . 443
28.1.1 Trailing P/E and the TTM Methodology . . . 443
28.1.2 Forward P/E and Analyst Forecast Reliability . . . 444
28.1.3 Cyclically Adjusted Price-Earnings (CAPE) Ratio . . . 444
28.1.4 Unlevered P/E and the Leibowitz Framework . . . 445
28.2 Market-Level Aggregation Techniques . . . 446
28.2.1 The VN-Index: Capitalization-Weighted Aggregation . . . 446
28.2.2 The VN30 and Free-Float Adjustments . . . 446
28.2.3 Median vs. Mean Aggregation . . . 447
28.3 Implementation . . . 447
28.3.1 Data Ingestion and Processing . . . 447
28.3.2 Shiller CAPE . . . 447
28.4 Macroeconomic and Regulatory Factors in Valuation . . . 448
28.4.1 The Role of Corporate Income Tax (CIT) . . . 448
28.4.2 Macroeconomic Determinants: Inflation and Interest Rates . . . 448
28.5 Synthesis of Valuation Dynamics . . . 449
28.6 Conclusions . . . 449

**29 Institutional Ownership Analytics in Vietnam** **450**
29.1 Institutional Ownership in Vietnam: A Distinct Landscape . . . 450
29.2 Data Infrastructure: DataCore.vn . . . 451
29.3 Data Pipeline . . . 454
29.3.1 Stock Price Data and Corporate Action Adjustments . . . 454
29.3.2 Ownership Structure Data . . . 461
29.4 Vietnam's Ownership Taxonomy . . . 464
29.4.1 The Five Ownership Categories . . . 464

29.5 Institutional Ownership Measures . . . 468
29.5.1 Ownership Ratio . . . 468
29.5.2 Concentration: Herfindahl-Hirschman Index . . . 469
29.5.3 Breadth of Ownership . . . 469
29.5.4 Time Series Visualization . . . 473
29.6 Foreign Ownership Dynamics . . . 473
29.6.1 Foreign Ownership Limits and the FOL Premium . . . 473
29.7 Institutional Trades . . . 481
29.7.1 Trade Inference in Vietnam . . . 481
29.8 Fund-Level Flows and Turnover . . . 487
29.8.1 Portfolio Assets and Returns from Fund Holdings . . . 487
29.8.2 Turnover Measures . . . 487
29.9 State Ownership Analysis . . . 490
29.9.1 Equitization and the Decline of State Ownership . . . 490
29.10 Modern Extensions . . . 492
29.10.1 Network Analysis of Co-Ownership . . . 492
29.10.2 ML-Enhanced Investor Classification . . . 495
29.10.3 Event Study: Ownership Disclosure Shocks . . . 498
29.11 Empirical Applications . . . 502
29.11.1 Application 1: Foreign Ownership and Stock Returns in Vietnam . . . 502
29.11.2 Application 2: State Divestiture and Value Creation . . . 503
29.11.3 Application 3: Institutional Herding in Vietnam . . . 505
29.12 Conclusion and Practical Recommendations . . . 507
29.12.1 Summary of Measures . . . 507
29.12.2 Data Quality Checklist for Vietnam . . . 508
29.12.3 Comparison with US Framework . . . 509

**30 Institutional Trades, Flows, and Turnover Ratios 510**
30.1 Measuring Institutional Ownership and Trading . . . 511
30.2 Trade Classification . . . 511
30.3 Turnover Measures . . . 512
30.4 Institutional Ownership in Emerging Markets . . . 512
30.5 Net Flows and Performance Attribution . . . 513

**31 Data Infrastructure 514**
31.1 Data Reader Class . . . 515

**32 Stock Price and Return Processing 518**
32.1 Price Data Extraction and Adjustment . . . 518
32.2 Monthly and Quarterly Price Processing . . . 521

**33 Ownership Data Processing 524**
33.1 Ownership Taxonomy . . . 524

33.2 Building the Holdings Panel . . . 525

**34 Institutional Ownership Metrics** **529**
34.1 Institutional Ownership Ratio . . . 529
34.2 Ownership Concentration: Herfindahl-Hirschman Index . . . 530
34.3 Ownership Breadth . . . 531
34.4 Implementation . . . 533
34.4.1 Trade Visualization . . . 535

**35 Portfolio Assets, Flows, and Returns** **538**
35.1 Total Assets and Portfolio Returns . . . 538
35.2 Aggregate Buys and Sales . . . 539

**36 Net Flows and Turnover Ratios** **541**
36.1 Net Flows . . . 541
36.2 Three Turnover Measures . . . 541
36.2.1 Turnover Summary Statistics . . . 543

**37 Foreign Ownership Analytics** **546**
37.1 FOL Utilization . . . 546
37.2 Room Premium Regression . . . 547

**38 Complete Pipeline** **549**

**39 Advanced Extensions** **552**
39.1 Herding Measures . . . 552
39.2 Demand Persistence . . . 553
39.3 Information Content of Trades . . . 554

**40 Empirical Applications** **556**
40.1 Application 1: Institutional Ownership Changes and Future Returns . . . 556
40.2 Application 2: Turnover and Performance . . . 557
40.3 Application 3: Foreign vs. Domestic Trading . . . 558

**41 Data Quality and Robustness** **561**
41.1 Common Pitfalls . . . 561
41.1.1 Corporate Action Misadjustment . . . 561
41.1.2 Disclosure Timing Mismatches . . . 561
41.1.3 Name Changes and Entity Mergers . . . 561
41.2 Validation Checks . . . 561

**42 Summary** **564**

**43 Conclusion** **565**
43.1 What you should take away . . . . . 565
43.1.1 Reproducibility is an identification strategy . . . . . 565
43.1.2 Vietnam rewards “microstructure humility” . . . . . 565
43.2 A reproducibility checklist you can actually use . . . . . 566

**44 Closing perspective** **567**

**References** **568**

# Preface

**Attribution**

This book is an independent derivative work inspired by reproducible research principles developed in **Tidy Finance**. It is not affiliated with, or officially provided by the creators of the original Tidy Finance books. All content, code, and empirical applications are original and tailored to the Vietnamese market.
This work builds directly on the methodological foundation established in:

- Scheuch, C., Voigt, S., & Weiss, P. (2023). *Tidy Finance with R*. Chapman and Hall/CRC. https://www.tidy-finance.org/r/ (Scheuch, Voigt, and Weiss 2023)
- Scheuch, C., Voigt, S., Weiss, P., & Frey, C. (2024). *Tidy Finance with Python*. Chapman and Hall/CRC. https://www.tidy-finance.org/python/ (Scheuch et al. 2024)

We gratefully acknowledge the Tidy Finance authors for developing an open, reproducible approach to empirical finance that made this market-specific adaptation possible.

## Motivation

Empirical finance has undergone a fundamental transformation over the past two decades. Advances in computational capacity, open-source statistical software, and data availability have reshaped how financial research is conducted, evaluated, and disseminated. Increasingly, credible empirical work is expected to be transparent, replicable, and extensible, with results generated through scripted workflows rather than manual intervention. Reproducibility, defined as the ability for independent researchers to regenerate empirical results using the same data and methods, has thus become a core norm in modern financial economics.

Despite this progress, the adoption of reproducible research practices has been uneven across markets. In developed financial systems, particularly those with long-established databases and standardized reporting regimes, reproducible empirical workflows are now commonplace. In contrast, research on emerging and frontier markets frequently relies on fragmented datasets, undocumented data cleaning procedures, and implicit institutional assumptions that are difficult to verify or extend. As a result, empirical findings in these markets are often fragile, non-comparable across studies, and costly to update as new data become available.

This book addresses that gap.

It develops a reproducible empirical finance framework designed explicitly for emerging and frontier markets, using Vietnam as a primary empirical case. Rather than adapting developed-market research pipelines post hoc, the book begins from the institutional and data realities of a fast-growing, retail-dominated, regulation-intensive market and builds methodological solutions accordingly. The objective is not merely to analyze Vietnam's financial markets, but to demonstrate how reproducible finance principles, as developed in the Tidy Finance framework, can be extended, stress-tested, and refined in environments characterized by data scarcity, institutional heterogeneity, and rapid structural change.

## Why Emerging Markets Require Different Empirical Infrastructure

Much of modern empirical finance implicitly assumes the existence of stable, high-frequency, institutionally harmonized datasets. These assumptions are rarely stated, yet they are deeply embedded in standard research designs: survivorship-free security histories, consistent accounting standards, unrestricted trading mechanisms, and deep institutional liquidity.

Emerging and frontier markets challenge each of these assumptions.

In Vietnam, as in many comparable economies, equity markets exhibit binding daily price limits, episodic trading halts, concentrated state ownership, and a predominance of retail investors. Financial disclosures reflect local accounting standards and evolving regulatory frameworks. Corporate actions are frequent, inconsistently documented, and occasionally revised ex post.

These characteristics are not inconveniences to be eliminated through aggressive data cleaning. They shape return dynamics, risk premia, factor construction, and statistical inference itself. An empirical framework that ignores these institutional features risks producing results that are internally inconsistent or externally misleading. A reproducible approach for emerging markets must therefore encode institutional context directly into data schemas, transformation logic, and modeling choices.

## Reproducibility as a Research Design Principle

In this book, reproducibility extends beyond the narrow notion of code availability. It is treated as an organizing principle governing the entire empirical research lifecycle.

First, all datasets are constructed from raw inputs through documented, deterministic transformations, ensuring clear data provenance. Second, empirical methods are implemented in a manner that makes modeling assumptions explicit and modifiable. Third, results are generated through scripted pipelines rather than interactive analysis, guaranteeing that updates to data or parameters propagate consistently throughout the analysis. Finally, empirical designs are

modular, allowing researchers to substitute markets, sample periods, or variable definitions without rewriting entire workflows.

This approach draws methodological inspiration from the broader reproducible research movement in economics and finance (e.g., Gentzkow and Shapiro 2014; Vilhuber 2020), while deliberately extending it beyond its original institutional and data environment. The goal is not to reproduce existing studies, but to enable new ones—particularly those that would otherwise be impractical due to fragmented data and institutional complexity.

## Vietnam as a Case, Not an Exception

Vietnam serves as the central empirical case throughout the book, but it is not treated as an idiosyncratic exception. Instead, it is presented as a representative example of a class of markets that occupy an intermediate position between frontier and emerging status: large enough to sustain active equity trading, yet still evolving in terms of regulation, disclosure quality, and investor composition.

By grounding methodological development in Vietnam's market structure, the book aims to produce insights that generalize to other contexts, including Southeast Asia, South Asia, Sub-Saharan Africa, and parts of Latin America. Each empirical chapter emphasizes which components are market-specific and which are portable, encouraging readers to adapt the framework rather than adopt it wholesale.

## Data Access

The empirical analyses in this book rely on Vietnamese equity market data provided by Datacore. To ensure reproducibility while respecting data licensing constraints, we provide the following resources:

- **Sample datasets**: A subset of anonymized data is available in DataCore's Sample Dataset for readers to run example code.
- **Data construction scripts**: All scripts used to clean and transform raw data are fully documented and available in the repository.
- **Replication guidance**: Readers with access to Vietnamese market data from commercial providers can use our scripts to construct equivalent datasets.

For questions about data access or replication, please contact the author.

## Contribution and Audience

This book makes three primary contributions.

First, it proposes a reproducible empirical finance framework explicitly designed for emerging and frontier markets, integrating institutional detail into data construction and model design. Second, it provides original empirical evidence on asset pricing, liquidity, and market microstructure in Vietnam using consistently constructed datasets. Third, it provides publication-ready, end-to-end research workflows suitable for academic research, policy analysis, and applied finance.

The intended audience includes graduate students in finance and economics, academic researchers studying non-developed markets, and practitioners interested in the systematic analysis of emerging-market equities. Familiarity with basic asset pricing theory and statistical programming is assumed, but no prior experience with Vietnam or similar markets is required.

## Structure of the Book

The chapters that follow progress from data infrastructure to empirical application. The book begins with an introduction to the data sources and infrastructure used throughout, followed by chapters on institutional context, data construction, and reproducible workflow design. Subsequent chapters develop asset pricing tests, liquidity measures, and market microstructure analyses tailored to Vietnam's equity market. Each chapter is designed to be self-contained, yet all are linked through a common data and code architecture to ensure internal consistency.

The book concludes by reflecting on the broader implications of reproducible empirical finance for emerging markets research and by outlining directions for future methodological and empirical work.

# 1 Institutional Background and Market Structure of Vietnam's Equity Market

Empirical analysis of financial markets is inseparable from institutional context. Market design, regulatory constraints, ownership structure, and investor composition shape observed prices, volumes, and returns. In developed markets, many of these features are sufficiently stable and standardized that they fade into the background of empirical research. In emerging markets, by contrast, institutional features are often first-order determinants of empirical outcomes.

This chapter provides the institutional foundation for the empirical analyses developed later in the book. It describes the structure of Vietnam's equity market, the regulatory environment governing trading and disclosure, and the characteristics of listed firms and investors. Rather than offering a purely descriptive account, the discussion emphasizes how institutional features map directly into data construction choices, modeling assumptions, and interpretation of empirical results.

## 1.1 Evolution of Vietnam's Equity Market

Vietnam's modern equity market is relatively young. Formal stock exchanges were established only in the early 2000s, as part of broader economic reforms aimed at transitioning from a centrally planned system toward a market-oriented economy. Since then, market capitalization, trading volume, and the number of listed firms have grown rapidly, albeit unevenly across sectors and time.

The pace of market development has been shaped by a combination of gradual privatization of state-owned enterprises, episodic regulatory reform, and sustained participation by retail investors. Unlike markets that evolved alongside large institutional investor bases, Vietnam's equity market matured in an environment where individual investors dominate trading activity and informational asymmetries remain substantial.

These features have important empirical implications. Return dynamics may reflect behavioral trading patterns, liquidity shocks can be amplified by coordinated retail activity, and firm-level information is incorporated into prices at varying speeds. A reproducible empirical framework must therefore be capable of capturing these dynamics without imposing assumptions derived from institutionally different markets.

## 1.2 Exchange Structure and Trading Mechanisms

Vietnam operates multiple equity exchanges, each with distinct listing requirements and trading rules. Trading is conducted through a centralized limit order book, with price-time priority determining execution. Importantly, daily price limits constrain the maximum allowable price movement for individual securities. These limits vary by exchange and security type and are binding during periods of heightened volatility.

Price limits introduce mechanical truncation in observed returns, clustering at upper and lower bounds, and persistence in price movements across days. From an empirical perspective, this challenges standard assumptions about continuous price adjustment and complicates volatility estimation, momentum measurement, and event-study design.

In this book, price limits are treated as structural features rather than anomalies. Data pipelines explicitly preserve limit-hit indicators, and empirical models are adapted to account for constrained price dynamics. This design choice reflects a broader principle: reproducibility in emerging markets requires preserving institutional signals rather than smoothing them away.

## 1.3 Listing Requirements and Firm Characteristics

Listed firms in Vietnam exhibit substantial heterogeneity in size, ownership structure, and disclosure quality. A defining characteristic of the market is the prevalence of firms with significant state ownership, either directly or through affiliated entities. State ownership affects governance, dividend policy, risk-taking behavior, and responsiveness to market signals.

Accounting disclosures follow Vietnamese Accounting Standards, which differ in important respects from international standards. While convergence efforts are ongoing, historical financial statements often reflect transitional rules, incomplete adoption of fair value accounting, and limited segment reporting. These features complicate cross-firm comparability and longitudinal analysis.

From a reproducible research standpoint, accounting variables cannot be treated as uniform primitives. Variable definitions, reporting lags, and restatement practices must be explicitly documented and encoded into data construction logic. Later chapters demonstrate how accounting data are harmonized in a transparent, version-controlled manner without obscuring underlying institutional differences.

## 1.4 Investor Composition and Trading Behavior

Retail investors dominate trading volume in Vietnam's equity market. Institutional investors, including domestic funds and foreign participants, play a growing but still secondary role.

This investor composition has implications for liquidity provision, price discovery, and market stability.

Retail-dominated markets tend to exhibit higher turnover, episodic herding behavior, and sensitivity to non-fundamental information. These patterns affect the interpretation of empirical results, particularly in studies of short-term return predictability, volume-return relations, and volatility clustering.

Rather than assuming institutional trading as the default, this book explicitly models liquidity and trading activity in a retail-centric environment. Measures of liquidity, for example, are chosen and constructed to remain meaningful in the presence of small trade sizes, intermittent trading, and order imbalances driven by individual investors.

## 1.5 Regulatory Environment and Market Frictions

Regulatory oversight of Vietnam's equity market has evolved alongside market development. Trading rules, disclosure requirements, and foreign ownership limits have been periodically revised, sometimes with limited backward compatibility. Regulatory changes can induce structural breaks in data that are not immediately apparent in raw time series.

Short-selling constraints, limited securities lending, and restrictions on derivative usage further distinguish Vietnam's market from developed counterparts. These frictions affect arbitrage activity and the feasibility of certain trading strategies, influencing observed return patterns and factor realizations.

A key principle of the empirical framework developed in this book is regulatory awareness. Data pipelines incorporate regulatory timelines, and empirical tests are designed to be robust to rule changes. This ensures that results are interpretable within the institutional regime in which they arise.

## 1.6 Implications for Empirical Design

The institutional features described in this chapter motivate several design choices that recur throughout the book:

1. **Data preservation over simplification**: Institutional constraints such as price limits and trading halts are retained and explicitly modeled.
2. **Modular variable construction**: Accounting and market variables are constructed through transparent functions that can be adjusted as standards evolve.
3. **Regime sensitivity**: Empirical analyses are structured to detect and accommodate regulatory and structural breaks.

4. **Context-aware interpretation**: Results are interpreted in light of market structure rather than benchmarked mechanically against developed-market findings.

## 1.7 Summary

Vietnam's equity market combines rapid growth with distinctive institutional features that challenge conventional empirical finance methods. Price limits, retail investor dominance, state ownership, and evolving regulation shape market outcomes in ways that cannot be ignored or abstracted away. For researchers working in such environments, reproducibility requires more than clean code and documented data—it requires embedding institutional context directly into empirical design.

# 2 Constructing and Analyzing Equity Return Series

This chapter develops a practical framework for transforming raw equity price records into return series suitable for empirical financial analysis. The focus is on methodological clarity and reproducibility, with particular attention to data issues that are prevalent in emerging equity markets such as Vietnam.

The discussion proceeds from individual stocks to a broad market cross-section, using constituents of the VN30 index as the primary empirical setting.

## 2.1 Data Access and Preparation

We begin by loading the core numerical and data manipulation libraries. These provide all functionality required for return construction without relying on specialized financial wrappers.

```
import pandas as pd
import numpy as np
```

For this project, we retrieve our historical price data using the DataCore API. If you wish to replicate this analysis or use the dataset for your own work, you will need to access the data through their platform.

### 2.1.1 Prerequisites for API Access

To run the code below, you need to configure a few things first:

1. **Obtain an API Key:** You must subscribe to the relevant dataset on the DataCore platform to receive a unique API key.
2. **Whitelist Your IP Address:** The API requires your IP address to be whitelisted for security.
    - **Local Machine:** If you are running this code on your personal computer, you generally need to whitelist your public IP address.

- **Cloud or Remote Sessions (e.g., HPC Open OnDemand):** If you are using a remote server such as those provided by DataCore, the server's IP address will change with each new session. You must retrieve the server's private/public IP for that specific session and whitelist it in your DataCore account settings before running the script.

3. **Set Environment Variables:** To keep your credentials secure, do not hardcode your API key into your scripts. Instead, save it as an environment variable named `datacore_api` on your machine.

*Note: If you only want to test the code performance, DataCore provides a preview endpoint that does not require an API key, though the data returned is limited.*

```
import requests
import pandas as pd

url = "https://gateway.datacore.vn/data/ds/preview"
params = {
    "dataSetCode": "fundamental_annual",
    "pageSize": 10000
}
headers = {
    "Accept": "application/json",
    "Origin": "https://datacore.vn",
    "Referer": "https://datacore.vn/"
}

response = requests.get(url, params=params, headers=headers)
data = response.json()

columns = data['data']['fields']
rows = data['data']['dataDetail']

df = pd.DataFrame(rows, columns=columns)
print(df.head())
print("Total rows:", len(df))
```

```
  symbol  year  total_current_asset ca_fin        ca_cce       ca_cash  \
0    CLL  2011         1.078338e+11   None  8.313178e+10  4.131776e+09
1    CLL  2012         2.360575e+10   None  8.003560e+09  4.003560e+09
2    CLL  2013         5.764370e+10   None  3.496426e+10  9.964256e+09
3    CLL  2014         4.973590e+10   None  1.718744e+10  1.718744e+10
4    CLL  2015         2.389115e+11   None  1.790364e+11  2.403638e+10
```

```
  ca_cash_inbank ca_cash_attransit  ca_cash_equivalent  ca_fin_invest  ...  \
0           None              None        7.900000e+10   0.000000e+00  ...
1           None              None        4.000000e+09   0.000000e+00  ...
2           None              None        2.500000e+10   0.000000e+00  ...
3           None              None        0.000000e+00   1.000000e+09  ...
4           None              None        1.550000e+11   1.000000e+09  ...

   operating_margin      roe      roa sector_pe sector_pb  sector_ps  \
0           0.35473  0.19560  0.10649   3.00213   0.26108    0.22170
1           0.45369  0.20337  0.13018   2.40335   0.26211    0.21745
2           0.45997  0.23459  0.16452   3.11089   0.41013    0.32436
3           0.40554  0.19984  0.14747   3.25886   0.46823    0.42146
4           0.33774  0.16525  0.12633   6.77337   0.81110    0.68401

   sector_eps  sector_ros  sector_roe  sector_roa
0  3341.54903     0.07385     0.08753     0.05039
1  4296.47005     0.09048     0.11392     0.06346
2  4024.74648     0.10427     0.13645     0.07415
3  5100.81019     0.12933     0.14956     0.08087
4  5216.38499     0.10098     0.11815     0.06234

[5 rows x 308 columns]
Total rows: 10
```

### 2.1.2 Checking Your IP Address

If you need to verify the IP address of the machine running your code (to whitelist it), you can use the following Python snippets.

**To find your Public IP:**

```python
import requests

try:
    public_ip = requests.get("https://api.ipify.org").text
    print(f"Public IP: {public_ip}")
except requests.exceptions.RequestException as e:
    print(f"Could not retrieve Public IP: {e}")
```

**To find your Private IP (useful for specific remote server setups):**

```python
import socket

def get_private_ip():
    try:
        s = socket.socket(socket.AF_INET, socket.SOCK_DGRAM)
        s.connect(("8.8.8.8", 80))
        private_ip = s.getsockname()[0]
        s.close()
        return private_ip
    except Exception as e:
        return f"Error: {e}"

print(f"Private IP: {get_private_ip()}")
```

### 2.1.3 Fetching the Dataset

The following script demonstrates how to securely authenticate and paginate through the DataCore API to retrieve the full `dataset_historical_price` dataset.

```python
# Convert the date column to proper datetime objects
prices["date"] = pd.to_datetime(prices["date"])

# Ensure price and ratio columns are numeric before calculation
prices["close_price"] = pd.to_numeric(prices["close_price"])
prices["adj_ratio"] = pd.to_numeric(prices["adj_ratio"])

# Calculate the adjusted close price
prices["adjusted_close"] = prices["close_price"] * prices["adj_ratio"]

# Rename columns to match standard conventions
prices = prices.rename(
    columns={
        "vol_total": "volume",
        "open_price": "open",
        "low_price": "low",
        "high_price": "high",
        "close_price": "close",
    }
)

# Sort the dataset logically by symbol and date
```

```
prices = prices.sort_values(["symbol", "date"])

print("Data manipulation complete. The dataset is ready for analysis.")
```

```
Data manipulation complete. The dataset is ready for analysis.
```

```
prices["date"] = pd.to_datetime(prices["date"])


prices["adjusted_close"] = prices["close_price"] * prices["adj_ratio"]


prices = prices.rename(
    columns={
        "vol_total": "volume",
        "open_price": "open",
        "low_price": "low",
        "high_price": "high",
        "close_price": "close",
    }
)

prices = prices.sort_values(["symbol", "date"])
```

Adjusted closing prices incorporate mechanical changes due to corporate actions such as cash dividends and stock splits. Using adjusted prices ensures that subsequent return calculations reflect investor-relevant performance rather than accounting artifacts.

## 2.2 Examining a Single Equity

To ground the discussion, we isolate the trading history of a single large-cap stock, FPT, over a long sample period.

```
import datetime as dt

start = pd.Timestamp("2000-01-01")
end = pd.Timestamp(dt.date.today().year - 1, 12, 31)
```

```
fpt = prices.loc[
    (prices["symbol"] == "FPT")
    & (prices["date"] >= start)
    & (prices["date"] <= end),
    ["date", "symbol", "volume", "open", "low", "high", "close", "adjusted_close"],
].copy()
```

This subset contains the standard daily market variables required for most empirical studies. Before computing returns, it is good practice to visually inspect the price series.

```
from plotnine import ggplot, aes, geom_line, labs
```

```
(
    ggplot(fpt, aes(x="date", y="adjusted_close"))
    + geom_line()
    + labs(title="Adjusted price path of FPT", x="", y="")
)
```

Figure 2.1: Prices are in VND, adjusted for dividend payments and stock splits.

## 2.3 From Prices to Returns

Most empirical asset pricing models are formulated in terms of returns rather than price levels. The simple daily return is defined as

$$r_t = \frac{p_t}{p_t - 1} - 1,$$

where $p_t$ denotes the adjusted closing price at the end of trading day $t$.

Before computing returns, we must address invalid price observations. In Vietnamese equity data, adjusted prices occasionally take the value zero. These entries typically arise from IPO placeholders, trading suspensions, or historical backfilling conventions and cannot be used to compute meaningful returns.

```
prices.loc[prices["adjusted_close"] <= 0, ["symbol", "date", "adjusted_close"]].head()
```

| | symbol | date | adjusted_close |
|---|---|---|---|
| 33886 | ADP | 2010-02-09 | 0.0 |
| 33887 | ADP | 2010-02-24 | 0.0 |
| 33888 | ADP | 2010-03-01 | 0.0 |
| 33889 | ADP | 2010-03-03 | 0.0 |
| 33890 | ADP | 2010-03-12 | 0.0 |

We therefore exclude non-positive adjusted prices and compute returns stock by stock. Correct chronological ordering is essential.

```
returns = (
    prices
    .loc[prices["adjusted_close"] > 0]
    .sort_values(["symbol", "date"])
    .assign(ret=lambda x: x.groupby("symbol")["adjusted_close"].pct_change())
    [["symbol", "date", "ret"]]
)
returns = returns.dropna(subset=["ret"])
```

The initial return for each stock is missing by construction, since no lagged price is available. These observations are mechanical and can safely be removed in most applications.

## 2.4 Limiting the Influence of Extreme Returns

Daily return series often contain extreme observations driven by data errors, thin trading, or abrupt price adjustments. A common approach is to winsorize returns using cross-sectional percentile cutoffs.

```
def winsorize_cs(df, column="ret", lower_q=0.01, upper_q=0.99):
    lo = df[column].quantile(lower_q)
    hi = df[column].quantile(upper_q)
    out = df.copy()
    out[column] = out[column].clip(lo, hi)
    return out

returns = winsorize_cs(returns)
```

Applying winsorization across the full cross-section limits the impact of extreme market-wide observations while preserving relative differences between firms. Winsorizing within each stock is rarely appropriate in panel settings and can severely distort illiquid securities.

## 2.5 Distributional Features of Returns

We next examine the empirical distribution of daily returns for FPT. The figure below also marks the historical 5 percent quantile, which provides a simple, non-parametric measure of downside risk.

```
from mizani.formatters import percent_format
from plotnine import geom_histogram, geom_vline, scale_x_continuous


fpt_ret = returns.loc[returns["symbol"] == "FPT"].copy()
q05 = fpt_ret["ret"].quantile(0.05)
```

```
(
    ggplot(fpt_ret, aes(x="ret"))
    + geom_histogram(bins=100)
    + geom_vline(xintercept=q05, linetype="dashed")
    + scale_x_continuous(labels=percent_format())
    + labs(title="Distribution of daily FPT returns", x="", y="")
)
```

Figure 2.2: The dotted vertical line indicates the historical five percent quantile.

Summary statistics offer a compact description of return behavior and should always be inspected before formal modeling.

```
returns["ret"].describe().round(3)
```

```
count    4305063.000
mean           0.000
std            0.035
min           -0.125
25%           -0.004
50%            0.000
75%            0.003
max            0.130
Name: ret, dtype: float64
```

Computing these statistics by calendar year can reveal periods of elevated volatility or structural change.

```
(
    returns
    .assign(year=lambda x: x["date"].dt.year)
    .groupby("year")["ret"]
    .describe()
    .round(3)
)
```

| year | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 2010 | 131548.0 | -0.001 | 0.036 | -0.125 | -0.021 | 0.0 | 0.018 | 0.13 |
| 2011 | 166826.0 | -0.003 | 0.033 | -0.125 | -0.020 | 0.0 | 0.011 | 0.13 |
| 2012 | 177938.0 | 0.000 | 0.033 | -0.125 | -0.012 | 0.0 | 0.015 | 0.13 |
| 2013 | 180417.0 | 0.001 | 0.033 | -0.125 | -0.004 | 0.0 | 0.008 | 0.13 |
| 2014 | 181907.0 | 0.001 | 0.034 | -0.125 | -0.008 | 0.0 | 0.011 | 0.13 |
| 2015 | 197881.0 | 0.000 | 0.033 | -0.125 | -0.006 | 0.0 | 0.005 | 0.13 |
| 2016 | 227896.0 | 0.000 | 0.035 | -0.125 | -0.005 | 0.0 | 0.003 | 0.13 |
| 2017 | 283642.0 | 0.001 | 0.034 | -0.125 | -0.002 | 0.0 | 0.001 | 0.13 |
| 2018 | 329887.0 | 0.000 | 0.035 | -0.125 | 0.000 | 0.0 | 0.000 | 0.13 |
| 2019 | 352754.0 | 0.000 | 0.033 | -0.125 | 0.000 | 0.0 | 0.000 | 0.13 |
| 2020 | 369367.0 | 0.001 | 0.035 | -0.125 | 0.000 | 0.0 | 0.000 | 0.13 |
| 2021 | 379415.0 | 0.002 | 0.038 | -0.125 | -0.005 | 0.0 | 0.007 | 0.13 |
| 2022 | 387050.0 | -0.001 | 0.038 | -0.125 | -0.008 | 0.0 | 0.004 | 0.13 |
| 2023 | 391605.0 | 0.001 | 0.034 | -0.125 | -0.002 | 0.0 | 0.002 | 0.13 |
| 2024 | 400379.0 | 0.000 | 0.031 | -0.125 | -0.002 | 0.0 | 0.000 | 0.13 |
| 2025 | 146551.0 | 0.000 | 0.037 | -0.125 | -0.004 | 0.0 | 0.002 | 0.13 |

## 2.6 Expanding to a Market Cross-Section

The same procedures apply naturally to a larger universe of stocks. We now restrict attention to the constituents of the VN30 index.

```
vn30 = [
    "ACB","BCM","BID","BVH","CTG","FPT","GAS","GVR","HDB","HPG",
    "MBB","MSN","MWG","PLX","POW","SAB","SHB","SSB","STB","TCB",
    "TPB","VCB","VHM","VIB","VIC","VJC","VNM","VPB","VRE","EIB",
]
```

```
prices_vn30 = prices.loc[prices["symbol"].isin(vn30)]
from plotnine import theme
```

```
(
    ggplot(prices_vn30, aes(x="date", y="adjusted_close", color="symbol"))
    + geom_line()
    + labs(title="Adjusted prices of VN30 constituents", x="", y="")
    + theme(legend_position="none")
)
```

Figure 2.3: Prices in VND, adjusted for dividend payments and stock splits.

Returns for the VN30 universe are computed analogously.

```
returns_vn30 = (
    prices_vn30
    .sort_values(["symbol", "date"])
    .assign(ret=lambda x: x.groupby("symbol")["adjusted_close"].pct_change())
    [["symbol", "date", "ret"]]
    .dropna()
)
```

```
returns_vn30.groupby("symbol")["ret"].describe().round(3)
```

| symbol | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| ACB | 3822.0 | -0.000 | 0.023 | -0.407 | -0.006 | 0.0 | 0.007 | 0.097 |
| BCM | 1795.0 | 0.001 | 0.027 | -0.136 | -0.010 | 0.0 | 0.010 | 0.159 |
| BID | 2811.0 | 0.000 | 0.024 | -0.369 | -0.010 | 0.0 | 0.011 | 0.070 |
| BVH | 3825.0 | 0.000 | 0.024 | -0.097 | -0.012 | 0.0 | 0.012 | 0.070 |
| CTG | 3825.0 | 0.000 | 0.024 | -0.376 | -0.010 | 0.0 | 0.010 | 0.070 |
| EIB | 3825.0 | -0.000 | 0.022 | -0.302 | -0.008 | 0.0 | 0.008 | 0.070 |
| FPT | 3825.0 | -0.000 | 0.024 | -0.439 | -0.008 | 0.0 | 0.009 | 0.070 |
| GAS | 3236.0 | 0.000 | 0.022 | -0.289 | -0.009 | 0.0 | 0.010 | 0.070 |
| GVR | 1775.0 | 0.001 | 0.030 | -0.137 | -0.014 | 0.0 | 0.016 | 0.169 |
| HDB | 1828.0 | -0.001 | 0.028 | -0.391 | -0.009 | 0.0 | 0.010 | 0.070 |
| HPG | 3825.0 | -0.001 | 0.032 | -0.581 | -0.010 | 0.0 | 0.011 | 0.070 |
| MBB | 3371.0 | -0.000 | 0.023 | -0.473 | -0.008 | 0.0 | 0.008 | 0.069 |
| MSN | 3825.0 | 0.000 | 0.024 | -0.553 | -0.010 | 0.0 | 0.010 | 0.070 |
| MWG | 2701.0 | -0.000 | 0.035 | -0.751 | -0.009 | 0.0 | 0.011 | 0.070 |
| PLX | 2009.0 | -0.000 | 0.021 | -0.140 | -0.010 | 0.0 | 0.010 | 0.070 |
| POW | 1784.0 | 0.000 | 0.023 | -0.071 | -0.012 | 0.0 | 0.011 | 0.102 |
| SAB | 2100.0 | -0.000 | 0.024 | -0.745 | -0.008 | 0.0 | 0.007 | 0.070 |
| SHB | 3824.0 | -0.000 | 0.028 | -0.338 | -0.013 | 0.0 | 0.013 | 0.100 |
| SSB | 1029.0 | -0.000 | 0.023 | -0.292 | -0.005 | 0.0 | 0.004 | 0.070 |
| STB | 3825.0 | 0.000 | 0.024 | -0.321 | -0.010 | 0.0 | 0.010 | 0.070 |
| TCB | 1732.0 | -0.000 | 0.035 | -0.884 | -0.009 | 0.0 | 0.010 | 0.070 |
| TPB | 1761.0 | -0.001 | 0.029 | -0.477 | -0.009 | 0.0 | 0.009 | 0.070 |
| VCB | 3825.0 | -0.000 | 0.024 | -0.539 | -0.009 | 0.0 | 0.009 | 0.070 |
| VHM | 1744.0 | -0.000 | 0.024 | -0.419 | -0.009 | 0.0 | 0.008 | 0.070 |
| VIB | 2072.0 | -0.000 | 0.031 | -0.489 | -0.009 | 0.0 | 0.010 | 0.109 |
| VIC | 3825.0 | -0.000 | 0.027 | -0.673 | -0.008 | 0.0 | 0.008 | 0.070 |
| VJC | 2046.0 | -0.000 | 0.020 | -0.455 | -0.007 | 0.0 | 0.006 | 0.070 |
| VNM | 3825.0 | -0.000 | 0.023 | -0.547 | -0.007 | 0.0 | 0.007 | 0.070 |
| VPB | 1927.0 | -0.000 | 0.033 | -0.678 | -0.010 | 0.0 | 0.010 | 0.070 |
| VRE | 1871.0 | -0.000 | 0.024 | -0.295 | -0.012 | 0.0 | 0.011 | 0.070 |

## 2.7 Aggregating Returns Across Time

Financial variables are observed at different frequencies. While equity prices are recorded daily, many empirical questions require monthly or annual returns. Lower-frequency returns are constructed by compounding higher-frequency observations.

```
returns_monthly = (
    returns_vn30
    .assign(month=lambda x: x["date"].dt.to_period("M").dt.to_timestamp())
    .groupby(["symbol", "month"], as_index=False)
    .agg(ret=("ret", lambda x: np.prod(1 + x) - 1))
)
```

Comparing daily and monthly return distributions illustrates how aggregation dampens volatility and alters tail behavior.

```
from plotnine import facet_wrap

fpt_d = returns_vn30.loc[returns_vn30["symbol"] == "FPT"].assign(freq="Daily")
fpt_m = returns_monthly.loc[returns_monthly["symbol"] == "FPT"].assign(freq="Monthly")


fpt_both = pd.concat([
    fpt_d[["ret", "freq"]],
    fpt_m[["ret", "freq"]],
])
```

```
(
    ggplot(fpt_both, aes(x="ret"))
    + geom_histogram(bins=50)
    + scale_x_continuous(labels=percent_format())
    + labs(title="FPT returns at different frequencies", x="", y="")
    + facet_wrap("freq", scales="free")
)
```

Figure 2.4: Returns are based on prices adjusted for dividend payments and stock splits.

## 2.8 Aggregation Across Firms: Trading Activity

Aggregation is not limited to time. In some settings, it is informative to aggregate variables across firms. As an illustration, we compute total daily trading value for VN30 stocks by multiplying share volume by adjusted prices and summing across firms.

```
trading_value = (
    prices_vn30
    .assign(value=lambda x: x["volume"] * x["adjusted_close"] / 1e9)
    .groupby("date")["value"]
    .sum()
    .reset_index()
    .assign(value_lag=lambda x: x["value"].shift(1))
)
(
    ggplot(trading_value, aes(x="date", y="value"))
    + geom_line()
    + labs(title="Aggregate VN30 trading value (billion VND)", x="", y="")
)
```

Finally, we assess persistence in trading activity by comparing trading value on consecutive days.

```
from plotnine import geom_point, geom_abline
```

```
(
    ggplot(trading_value, aes(x="value_lag", y="value"))
    + geom_point()
    + geom_abline(intercept=0, slope=1, linetype="dashed")
    + labs(
        title="Persistence in VN30 trading value",
        x="Previous day",
        y="Current day",
    )
)
```

Figure 2.5: Total daily trading volume.

A strong alignment along the 45-degree line indicates that high-activity trading days tend to be followed by similarly active days, a common empirical regularity in equity markets.

## 2.9 Summary

This chapter established a reproducible workflow for transforming raw price data into return series, diagnosing common data issues, and aggregating information across time and firms. These steps provide the empirical foundation for subsequent analyses of risk, return predictability, and market dynamics in Vietnam's equity market.

# 3 Modern Portfolio Theory

In the previous chapter, we showed how to download and analyze stock market data with figures and summary statistics. Now, we turn to one of the most fundamental questions in finance: How should an investor allocate their wealth across assets that differ in expected returns, variance, and correlations to optimize their portfolio's performance?

This question might seem straightforward at first glance. Why not simply invest everything in the asset with the highest expected return? The answer lies in a profound insight that transformed financial economics: **risk matters, and it can be managed through diversification**.

Modern Portfolio Theory (MPT), introduced by Markowitz (1952), revolutionized investment decision-making by formalizing the trade-off between risk and expected return. Before Markowitz, investors largely thought about risk on a security-by-security basis. Markowitz's genius was recognizing that what matters is not the risk of individual securities in isolation, but how they contribute to the risk of the *entire portfolio*. This insight was so influential that it earned him the Sveriges Riksbank Prize in Economic Sciences in 1990 and laid the foundation for much of modern finance.

### 3.0.1 The Core Insight: Diversification as a Free Lunch

MPT relies on a crucial mathematical fact: portfolio risk depends not only on individual asset volatilities but also on the *correlations* between asset returns. This insight reveals the power of diversification—combining assets whose returns don't move in perfect lockstep can reduce overall portfolio risk without necessarily sacrificing expected return.

Consider a simple analogy: Imagine you run a business selling both sunscreen and umbrellas. On sunny days, sunscreen sales boom but umbrella sales suffer; on rainy days, the reverse happens. By selling both products, your total revenue becomes more stable than if you sold only one. The "correlation" between sunscreen and umbrella sales is negative, and combining them reduces the variance of your overall income. This is precisely the logic behind portfolio diversification.

The fruit basket analogy offers another perspective: If all you have are apples and they spoil, you lose everything. With a variety of fruits, some may spoil, but others will stay fresh. Diversification provides insurance against the idiosyncratic risks of individual assets.

### 3.0.2 The Mean-Variance Framework

At the heart of MPT is **mean-variance analysis**, which evaluates portfolios based on two dimensions:

1. **Expected return (mean)**: The anticipated average profit from holding the portfolio
2. **Risk (variance)**: The dispersion of possible returns around the expected value

The key assumption is that investors care only about these two moments of the return distribution. This assumption is exactly correct if returns are normally distributed, or if investors have quadratic utility functions. Even when these conditions don't hold precisely, mean-variance analysis often provides a good approximation to optimal portfolio choice.

By balancing expected return and risk, investors can construct portfolios that either maximize expected return for a given level of risk, or minimize risk for a desired level of expected return. In this chapter, we derive these optimal portfolio decisions analytically and implement the mean-variance approach computationally.

```
import pandas as pd
import numpy as np
import tidyfinance as tf

from plotnine import *
from mizani.formatters import percent_format
from adjustText import adjust_text
```

## 3.1 The Asset Universe: Setting Up the Problem

Suppose $N$ different risky assets are available to the investor. Each asset $i$ is characterized by:

- **Expected return** $\mu_i$: The anticipated profit from holding the asset for one period
- **Variance** $\sigma_i^2$: The dispersion of returns around the mean
- **Covariances** $\sigma_{ij}$: The degree to which asset $i$'s returns move together with asset $j$'s returns

The investor chooses **portfolio weights** $\omega_i$ for each asset $i$. These weights represent the fraction of total wealth invested in each asset. We impose the constraint that weights sum to one:

$$\sum_{i=1}^{N} \omega_i = 1$$

This "budget constraint" ensures that the investor is fully invested—there is no outside option such as keeping money under a mattress. Note that we allow weights to be negative (short selling) or greater than one (leverage), though in practice these positions may face constraints.

### 3.1.1 The Two Stages of Portfolio Selection

According to Markowitz (1952), portfolio selection involves two distinct stages:

1. **Estimation**: Forming expectations about future security performance based on observations, experience, and economic reasoning
2. **Optimization**: Using these expectations to choose an optimal portfolio

In practice, these stages cannot be fully separated. The estimation stage determines the inputs ($\mu$, $\Sigma$) that feed into the optimization stage. Poor estimation leads to poor portfolio choices, regardless of how sophisticated the optimization procedure.

To keep things conceptually clear, we focus primarily on the optimization stage in this chapter. We treat the expected returns and variance-covariance matrix as known, using historical data to compute reasonable proxies. In later chapters, we address the substantial challenges that arise from estimation uncertainty.

### 3.1.2 Loading and Preparing the Data

We work with the VN30 index constituents—the 30 largest and most liquid stocks on Vietnam's Ho Chi Minh Stock Exchange. This provides a realistic asset universe for a domestic Vietnamese investor.

```
vn30_symbols = [
    "ACB","BCM","BID","BVH","CTG","FPT","GAS","GVR","HDB","HPG",
    "MBB","MSN","MWG","PLX","POW","SAB","SHB","SSB","STB","TCB",
    "TPB","VCB","VHM","VIB","VIC","VJC","VNM","VPB","VRE","EIB"
]
```

We load the historical price data:

```
import pandas as pd
from io import BytesIO
import datetime as dt
import os
import boto3
from botocore.client import Config
```

```python
class ConnectMinio:
    def __init__(self):
        self.MINIO_ENDPOINT = os.environ["MINIO_ENDPOINT"]
        self.MINIO_ACCESS_KEY = os.environ["MINIO_ACCESS_KEY"]
        self.MINIO_SECRET_KEY = os.environ["MINIO_SECRET_KEY"]
        self.REGION = os.getenv("MINIO_REGION", "us-east-1")

        self.s3 = boto3.client(
            "s3",
            endpoint_url=self.MINIO_ENDPOINT,
            aws_access_key_id=self.MINIO_ACCESS_KEY,
            aws_secret_access_key=self.MINIO_SECRET_KEY,
            region_name=self.REGION,
            config=Config(signature_version="s3v4"),
        )

    def test_connection(self):
        resp = self.s3.list_buckets()
        print("Connected. Buckets:")
        for b in resp.get("Buckets", []):
            print(" -", b["Name"])

conn = ConnectMinio()
s3 = conn.s3
conn.test_connection()

bucket_name = os.environ["MINIO_BUCKET"]

prices = pd.read_csv(
    BytesIO(
        s3.get_object(
            Bucket=bucket_name,
            Key="historycal_price/dataset_historical_price.csv"
        )["Body"].read()
    ),
    low_memory=False
)

prices["date"] = pd.to_datetime(prices["date"])
prices["adjusted_close"] = prices["close_price"] * prices["adj_ratio"]
prices = prices.rename(columns={
    "vol_total": "volume",
```

```
    "open_price": "open",
    "low_price": "low",
    "high_price": "high",
    "close_price": "close"
})
prices = prices.sort_values(["symbol", "date"])
```

```
Connected. Buckets:
 - dsteam-data
 - rawbctc
```

We filter to keep only the VN30 constituents:

```
prices_daily = prices[prices["symbol"].isin(vn30_symbols)]
prices_daily[["date", "symbol", "adjusted_close"]].head(3)
```

| | date | symbol | adjusted_close |
|---|---|---|---|
| 18176 | 2010-01-04 | ACB | 329.408244 |
| 18177 | 2010-01-05 | ACB | 329.408244 |
| 18178 | 2010-01-06 | ACB | 320.258015 |

### 3.1.3 Computing Expected Returns

The sample mean return serves as our proxy for expected returns. For each asset $i$, we compute:

$$\hat{\mu}_i = \frac{1}{T}\sum_{t=1}^{T} r_{i,t}$$

where $r_{i,t}$ is the return of asset $i$ in period $t$, and $T$ is the total number of periods.

**Why monthly returns?** While daily data provides more observations, monthly returns offer several advantages for portfolio optimization. First, monthly returns are less noisy and exhibit weaker serial correlation. Second, monthly rebalancing is more realistic for most investors, avoiding excessive transaction costs. Third, the estimation error in mean returns is already substantial—using daily data doesn't materially improve the precision of mean estimates because the mean return scales with the horizon while estimation error scales with the square root of observations.

```
returns_monthly = (prices_daily
  .assign(
    date=prices_daily["date"].dt.to_period("M").dt.to_timestamp()
  )
  .groupby(["symbol", "date"], as_index=False)
  .agg(adjusted_close=("adjusted_close", "last"))
  .assign(
    ret=lambda x: x.groupby("symbol")["adjusted_close"].pct_change()
  )
)
```

### 3.1.4 Computing Volatilities

Individual asset risk in MPT is quantified using variance ($\sigma_i^2$) or its square root, the standard deviation or volatility ($\sigma_i$). We use the sample standard deviation as our proxy:

$$\hat{\sigma}_i = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(r_{i,t} - \hat{\mu}_i)^2}$$

Alternative risk measures exist, including Value-at-Risk, Expected Shortfall, and higher-order moments such as skewness and kurtosis. However, variance remains the workhorse measure in portfolio theory because of its mathematical tractability and the central role of the normal distribution in finance.

```
assets = (returns_monthly
  .groupby("symbol", as_index=False)
  .agg(
    mu=("ret", "mean"),
    sigma=("ret", "std")
  )
)
```

### 3.1.5 Visualizing the Risk-Return Trade-off

Figure 3.1 displays each asset's expected return (vertical axis) against its volatility (horizontal axis). This "mean-standard deviation" space is fundamental to portfolio theory.

```
assets_figure = (
  ggplot(
    assets,
    aes(x="sigma", y="mu", label="symbol")
  )
  + geom_point()
  + geom_text(adjust_text={"arrowprops": {"arrowstyle": "-"}})
  + scale_x_continuous(labels=percent_format())
  + scale_y_continuous(labels=percent_format())
  + labs(
      x="Volatility (Standard Deviation)",
      y="Expected Return",
      title="Expected returns and volatilities of VN30 index constituents"
  )
)
assets_figure.show()
```

Figure 3.1: Expected returns and volatilities based on monthly returns adjusted for dividend payments and stock splits.

Several observations emerge from this figure. First, there is substantial heterogeneity in both expected returns and volatilities across stocks. Second, the relationship between risk and return

is far from linear. Some high-volatility stocks have low or even negative expected returns. Third, most individual stocks appear to offer poor risk-return trade-offs. As we will see, portfolios can substantially improve upon these individual positions.

## 3.2 The Variance-Covariance Matrix: Capturing Asset Interactions

### 3.2.1 Why Correlations Matter

A key innovation of MPT is recognizing that portfolio risk depends critically on how assets move together. The **variance-covariance matrix** $\Sigma$ captures all pairwise interactions between asset returns.

To understand why correlations matter, consider the variance of a two-asset portfolio:

$$\sigma_p^2 = \omega_1^2\sigma_1^2 + \omega_2^2\sigma_2^2 + 2\omega_1\omega_2\sigma_{12}$$

The third term involves the covariance $\sigma_{12} = \rho_{12}\sigma_1\sigma_2$, where $\rho_{12}$ is the correlation coefficient. When $\rho_{12} < 1$, the portfolio variance is *less* than the weighted average of individual variances. When $\rho_{12} < 0$, the diversification benefit is even more pronounced.

This mathematical fact has profound implications: **You can reduce risk without reducing expected return** by combining assets that don't move perfectly together. This is sometimes called the "only free lunch in finance."

### 3.2.2 Computing the Variance-Covariance Matrix

We compute the sample covariance matrix as:

$$\hat{\sigma}_{ij} = \frac{1}{T-1}\sum_{t=1}^{T}(r_{i,t} - \hat{\mu}_i)(r_{j,t} - \hat{\mu}_j)$$

First, we reshape the returns data into a wide format with assets as columns:

```
returns_wide = (returns_monthly
  .pivot(index="date", columns="symbol", values="ret")
  .reset_index()
)

sigma = (returns_wide
  .drop(columns=["date"])
  .cov()
)
```

### 3.2.3 Interpreting the Variance-Covariance Matrix

The diagonal elements of $\Sigma$ are the variances of individual assets. The off-diagonal elements are covariances, which can be positive (assets tend to move together), negative (assets tend to move in opposite directions), or zero (no linear relationship).

For easier interpretation, we often convert covariances to correlations:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

Correlations are bounded between -1 and +1, making them easier to compare across asset pairs.

Figure 3.2 visualizes the variance-covariance matrix as a heatmap.

```
sigma_long = (sigma
  .reset_index()
  .melt(id_vars="symbol", var_name="symbol_b", value_name="value")
)

sigma_long["symbol_b"] = pd.Categorical(
  sigma_long["symbol_b"],
  categories=sigma_long["symbol_b"].unique()[::-1],
  ordered=True
)

sigma_figure = (
  ggplot(
    sigma_long,
    aes(x="symbol", y="symbol_b", fill="value")
  )
  + geom_tile()
  + labs(
      x="", y="", fill="(Co-)Variance",
      title="Sample variance-covariance matrix of VN30 index constituents"
    )
  + scale_fill_continuous(labels=percent_format())
  + theme(axis_text_x=element_text(angle=45, hjust=1))
)
sigma_figure.show()
```

Figure 3.2: Variances and covariances based on monthly returns adjusted for dividend payments and stock splits.

The heatmap reveals important patterns. The diagonal (variances) shows which stocks are most volatile. The off-diagonal patterns show which pairs of stocks tend to move together. In general, stocks within the same sector tend to have higher correlations with each other than with stocks from different sectors.

## 3.3 The Minimum-Variance Portfolio

### 3.3.1 Motivation: Risk Minimization as a Benchmark

Before considering expected returns, let's find the portfolio that minimizes risk entirely. This **minimum-variance portfolio (MVP)** serves as an important benchmark and reference point. It represents what an extremely risk-averse investor—one who cares only about minimizing volatility—would choose.

### 3.3.2 The Optimization Problem

The minimum-variance investor solves:

$$\min_{\omega} \omega' \Sigma \omega$$

subject to the constraint that weights sum to one:

$$\omega' \iota = 1$$

where $\iota$ is an $N \times 1$ vector of ones.

In words: minimize portfolio variance, subject to being fully invested.

### 3.3.3 The Analytical Solution

This is a classic constrained optimization problem that can be solved using Lagrange multipliers. The Lagrangian is:

$$\mathcal{L} = \omega' \Sigma \omega - \lambda(\omega' \iota - 1)$$

Taking the first-order condition with respect to $\omega$:

$$\frac{\partial \mathcal{L}}{\partial \omega} = 2\Sigma\omega - \lambda\iota = 0$$

Solving for $\omega$:

$$\omega = \frac{\lambda}{2}\Sigma^{-1}\iota$$

Using the constraint $\omega' \iota = 1$ to solve for $\lambda$:

$$\frac{\lambda}{2}\iota' \Sigma^{-1} \iota = 1 \implies \frac{\lambda}{2} = \frac{1}{\iota' \Sigma^{-1} \iota}$$

Substituting back:

$$\omega_{\text{mvp}} = \frac{\Sigma^{-1}\iota}{\iota' \Sigma^{-1} \iota}$$

This elegant formula shows that the minimum-variance weights depend only on the covariance matrix—expected returns play no role. The inverse covariance matrix $\Sigma^{-1}$ determines how much to invest in each asset based on its variance and its covariances with all other assets.

### 3.3.4 Implementation

```
iota = np.ones(sigma.shape[0])
sigma_inv = np.linalg.inv(sigma.values)
omega_mvp = (sigma_inv @ iota) / (iota @ sigma_inv @ iota)
```

### 3.3.5 Visualizing the Minimum-Variance Weights

Figure 3.3 displays the portfolio weights of the minimum-variance portfolio.

```
assets = assets.assign(omega_mvp=omega_mvp)

assets["symbol"] = pd.Categorical(
  assets["symbol"],
  categories=assets.sort_values("omega_mvp")["symbol"],
  ordered=True
)

omega_figure = (
  ggplot(
    assets,
    aes(y="omega_mvp", x="symbol", fill="omega_mvp>0")
  )
  + geom_col()
  + coord_flip()
  + scale_y_continuous(labels=percent_format())
  + labs(
      x="",
      y="Portfolio Weight",
      title="Minimum-variance portfolio weights"
  )
  + theme(legend_position="none")
)
omega_figure.show()
```

Figure 3.3: Weights are based on historical moments of monthly returns adjusted for dividend payments and stock splits.

Several features of the minimum-variance portfolio are noteworthy. First, many stocks receive zero or near-zero weights. Second, some stocks receive negative weights (short positions). These short positions are not a computational artifact, they reflect the optimizer's attempt to exploit correlations for risk reduction. Third, the weights are quite extreme (both large positive and large negative), which often indicates estimation error amplification, which is a topic we address in later chapters.

### 3.3.6 Portfolio Performance

Let's compute the expected return and volatility of the minimum-variance portfolio:

```
mu = assets["mu"].values
mu_mvp = omega_mvp @ mu
sigma_mvp = np.sqrt(omega_mvp @ sigma.values @ omega_mvp)

summary_mvp = pd.DataFrame({
  "mu": [mu_mvp],
  "sigma": [sigma_mvp],
  "type": ["Minimum-Variance Portfolio"]
```

```
})
summary_mvp
```

| | mu | sigma | type |
|---|---|---|---|
| 0 | -0.011424 | 0.043512 | Minimum-Variance Portfolio |

```
mu_mvp_fmt = f"{mu_mvp:.4f}"
sigma_mvp_fmt = f"{sigma_mvp:.4f}"
print(f"The MVP return is {mu_mvp_fmt} and volatility is {sigma_mvp_fmt}.")
```

```
The MVP return is -0.0114 and volatility is 0.0435.
```

If the expected return is negative, this is not a computational error. The minimum-variance portfolio minimizes risk without regard to expected returns. Because some assets in the sample have negative average returns, the risk-minimizing combination may inherit a negative expected return. This highlights a fundamental limitation of using historical sample means as estimates of expected returns: they are extremely noisy, and can lead to economically unintuitive results even when the optimization mathematics are working correctly.

## 3.4 Efficient Portfolios: Balancing Risk and Return

### 3.4.1 The Investor's Trade-off

In most cases, minimizing variance is not the investor's sole objective. A more realistic formulation allows the investor to trade off risk against expected return. The investor might be willing to accept higher portfolio variance in exchange for higher expected returns.

An **efficient portfolio** minimizes variance subject to earning at least some target expected return $\bar{\mu}$. Formally:

$$\min_{\omega} \omega' \Sigma \omega$$

subject to:

$$\omega' \iota = 1 \quad \text{(fully invested)}$$

$$\omega' \mu \geq \bar{\mu} \quad \text{(minimum return)}$$

When $\bar{\mu}$ exceeds the expected return of the minimum-variance portfolio, the investor accepts more risk to earn more return.

### 3.4.2 Setting the Target Return

For illustration, suppose the investor wants to earn at least the historical average return of the best-performing stock:

```
mu_bar = assets["mu"].max()
print(f"Target expected return: {mu_bar:.5f}")
```

```
Target expected return: 0.01886
```

This is an ambitious target—it means matching the return of the single highest-returning stock while benefiting from diversification to reduce risk.

### 3.4.3 The Analytical Solution

The constrained optimization problem with an inequality constraint on expected returns can be solved using the Karush-Kuhn-Tucker (KKT) conditions. At the optimum (assuming the return constraint binds), the solution is:

$$\omega_{\text{efp}} = \frac{\lambda^*}{2}\left(\Sigma^{-1}\mu - \frac{D}{C}\Sigma^{-1}\iota\right)$$

where:

- $C = \iota'\Sigma^{-1}\iota$ (a scalar measuring the "size" of the inverse covariance matrix)
- $D = \iota'\Sigma^{-1}\mu$ (capturing the interaction between expected returns and the inverse covariance matrix)
- $E = \mu'\Sigma^{-1}\mu$ (measuring the "signal" in expected returns weighted by inverse covariances)
- $\lambda^* = 2\frac{\bar{\mu}-D/C}{E-D^2/C}$ (the shadow price of the return constraint)

Alternatively, we can express the efficient portfolio as a linear combination of the minimum-variance portfolio and an "excess return" portfolio:

$$\omega_{\text{efp}} = \omega_{\text{mvp}} + \frac{\lambda^*}{2}\left(\Sigma^{-1}\mu - D \cdot \omega_{\text{mvp}}\right)$$

This representation reveals important intuition: the efficient portfolio starts from the minimum-variance portfolio and tilts toward higher-expected-return assets, with the tilt magnitude determined by $\lambda^*$.

### 3.4.4 Implementation

```
C = iota @ sigma_inv @ iota
D = iota @ sigma_inv @ mu
E = mu @ sigma_inv @ mu
lambda_tilde = 2 * (mu_bar - D / C) / (E - (D ** 2) / C)
omega_efp = omega_mvp + (lambda_tilde / 2) * (sigma_inv @ mu - D * omega_mvp)

mu_efp = omega_efp @ mu
sigma_efp = np.sqrt(omega_efp @ sigma.values @ omega_efp)

summary_efp = pd.DataFrame({
  "mu": [mu_efp],
  "sigma": [sigma_efp],
  "type": ["Efficient Portfolio"]
})
```

### 3.4.5 Comparing the Portfolios

Figure 3.4 plots both portfolios alongside the individual assets.

```
summaries = pd.concat(
  [assets, summary_mvp, summary_efp], ignore_index=True
)

summaries_figure = (
  ggplot(
    summaries,
    aes(x="sigma", y="mu")
  )
  + geom_point(data=summaries.query("type.isna()"))
  + geom_point(data=summaries.query("type.notna()"), color="#F21A00", size=3)
  + geom_label(aes(label="type"), adjust_text={"arrowprops": {"arrowstyle": "-"}})
  + scale_x_continuous(labels=percent_format())
  + scale_y_continuous(labels=percent_format())
  + labs(
      x="Volatility (Standard Deviation)",
      y="Expected Return",
      title="Efficient & minimum-variance portfolios"
    )
```

```
)
summaries_figure.show()
```

Figure 3.4: The big dots indicate the location of the minimum-variance and the efficient portfolio that delivers the expected return of the stock with the highest return, respectively. The small dots indicate the location of the individual constituents.

The figure demonstrates the substantial diversification benefits of portfolio optimization. The efficient portfolio achieves the same expected return as the highest-returning individual stock but with substantially lower volatility. This "free lunch" from diversification is the central insight of Modern Portfolio Theory.

### 3.4.6 The Role of Risk Aversion

The target return $\bar{\mu}$ implicitly reflects the investor's risk aversion. Less risk-averse investors choose higher $\bar{\mu}$, accepting more variance to earn more expected return. More risk-averse investors choose $\bar{\mu}$ closer to the minimum-variance portfolio's expected return.

Equivalently, the mean-variance framework can be derived from the optimal decisions of an investor with a mean-variance utility function:

$$U(\omega) = \omega'\mu - \frac{\gamma}{2}\omega'\Sigma\omega$$

where $\gamma$ is the coefficient of relative risk aversion. The Appendix shows there is a one-to-one mapping between $\gamma$ and $\bar{\mu}$, so both formulations yield identical efficient portfolios.

## 3.5 The Efficient Frontier: The Menu of Optimal Portfolios

The **efficient frontier** is the set of all portfolios for which no other portfolio offers higher expected return at the same or lower variance. Geometrically, it traces the upper boundary of achievable (volatility, expected return) combinations.

Every rational mean-variance investor should hold a portfolio on the efficient frontier. Portfolios below the frontier are "dominated," there exists another portfolio with either higher return for the same risk, or lower risk for the same return.

### 3.5.1 The Mutual Fund Separation Theorem

A remarkable result simplifies the construction of the efficient frontier. The **mutual fund separation theorem** (sometimes called the two-fund theorem) states that any efficient portfolio can be expressed as a linear combination of any two distinct efficient portfolios.

Formally, if $\omega_{\mu_1}$ and $\omega_{\mu_2}$ are efficient portfolios earning expected returns $\mu_1$ and $\mu_2$ respectively, then the portfolio:

$$\omega_{a\mu_1+(1-a)\mu_2} = a \cdot \omega_{\mu_1} + (1-a) \cdot \omega_{\mu_2}$$

is also efficient and earns expected return $a\mu_1 + (1-a)\mu_2$.

This result has profound practical implications: an investor needs access to only two efficient "mutual funds" to construct any portfolio on the efficient frontier. The specific funds don't matter—any two distinct efficient portfolios span the entire frontier.

### 3.5.2 Proof of the Separation Theorem

The proof follows directly from the analytical solution for efficient portfolios. Consider:

$$a \cdot \omega_{\mu_1} + (1-a) \cdot \omega_{\mu_2} = \left( \frac{a\mu_1 + (1-a)\mu_2 - D/C}{E - D^2/C} \right) \left( \Sigma^{-1}\mu - \frac{D}{C}\Sigma^{-1}\iota \right)$$

This expression has exactly the form of the efficient portfolio earning expected return $a\mu_1 + (1-a)\mu_2$, proving the theorem.

### 3.5.3 Computing the Efficient Frontier

Using the minimum-variance portfolio and our efficient portfolio as the two "funds," we can trace out the entire efficient frontier:

```
efficient_frontier = (
  pd.DataFrame({
    "a": np.arange(-1, 2.01, 0.01)
  })
  .assign(
    omega=lambda x: x["a"].map(lambda a: a * omega_efp + (1 - a) * omega_mvp)
  )
  .assign(
    mu=lambda x: x["omega"].map(lambda w: w @ mu),
    sigma=lambda x: x["omega"].map(lambda w: np.sqrt(w @ sigma @ w))
  )
)
```

Note that we allow $a$ to range from -1 to 2, which means some portfolios involve shorting one of the two basis funds and leveraging into the other. This traces out both the upper and lower portions of the frontier hyperbola.

### 3.5.4 Visualizing the Efficient Frontier

Figure 3.5 displays the efficient frontier alongside individual assets and the benchmark portfolios.

```
summaries = pd.concat(
  [summaries, efficient_frontier], ignore_index=True
)

summaries_figure = (
  ggplot(
    summaries,
    aes(x="sigma", y="mu")
  )
  + geom_point(data=summaries.query("type.isna()"))
  + geom_line(data=efficient_frontier, color="blue", alpha=0.7)
  + geom_point(data=summaries.query("type.notna()"), color="#F21A00", size=3)
  + geom_label(aes(label="type"), adjust_text={"arrowprops": {"arrowstyle": "-"}})
  + scale_x_continuous(labels=percent_format())
```

```
  + scale_y_continuous(labels=percent_format())
  + labs(
      x="Volatility (Standard Deviation)",
      y="Expected Return",
      title="The Efficient Frontier and VN30 Constituents"
    )
)
summaries_figure.show()
```

Figure 3.5: The big dots indicate the location of the minimum-variance and the efficient portfolio. The small dots indicate the location of the individual constituents.

The efficient frontier has a characteristic hyperbolic shape. The leftmost point is the minimum-variance portfolio. Moving up and to the right along the frontier, expected return increases but so does volatility. The upper portion of the hyperbola (above the minimum-variance portfolio) is the "efficient" part—these portfolios offer the highest return for each level of risk. The lower portion is "inefficient"—these portfolios are dominated by their mirror images on the upper portion.

The figure also reveals how dramatically diversification improves upon individual stock positions. Nearly all individual stocks lie well inside the efficient frontier, meaning investors can achieve the same expected return with much less risk, or much higher expected return with the same risk, simply by diversifying.

## 3.6 Key Takeaways

This chapter introduced the concepts of Modern Portfolio Theory. The main insights are:

1. **Portfolio risk depends on correlations**: The variance of a portfolio is not simply the weighted average of individual variances. Covariances between assets play a crucial role, creating opportunities for diversification.
2. **Diversification is the "only free lunch" in finance**: By combining assets that don't move perfectly together, investors can reduce risk without sacrificing expected return. This insight is the cornerstone of modern investment practice.
3. **The minimum-variance portfolio minimizes risk**: This portfolio depends only on the covariance matrix and serves as an important benchmark. It represents the least risky way to be fully invested in risky assets.
4. **Efficient portfolios balance risk and return**: By accepting more variance, investors can earn higher expected returns. Efficient portfolios are those that offer the best possible trade-off.
5. **The efficient frontier characterizes optimal portfolios**: This boundary in mean-standard deviation space represents the menu of optimal choices available to mean-variance investors.
6. **Two-fund separation simplifies implementation**: Any efficient portfolio can be constructed from any two distinct efficient portfolios, reducing the computational burden of portfolio optimization.

Looking ahead, several important complications arise in practice. First, the inputs to portfolio optimization (expected returns and covariances) must be estimated from data, and estimation error can dramatically affect portfolio performance. Second, real-world constraints such as transaction costs, short-sale restrictions, and position limits modify the optimization problem. Third, the assumption that investors care only about mean and variance may be too restrictive when returns are non-normal or when investors have more complex preferences. We address these extensions in subsequent chapters.

# 4 The Capital Asset Pricing Model

## 4.1 From Efficient Portfolios to Equilibrium Prices

The previous chapter on Modern Portfolio Theory (MPT) showed how an investor can construct portfolios that optimally trade off risk and expected return. But MPT leaves a crucial question unanswered: What determines the *expected returns* themselves? Why do some assets command higher risk premiums than others?

The Capital Asset Pricing Model (CAPM) answers this question. Developed simultaneously by Sharpe (1964), Lintner (1965), and Mossin (1966), the CAPM extends MPT to explain how assets should be priced in equilibrium when *all* investors follow mean-variance optimization principles. The CAPM's central insight is both elegant and counterintuitive: **not all risk is rewarded**. Only the component of risk that cannot be diversified away (i.e., systematic risk) commands a risk premium in equilibrium.

The CAPM remains the cornerstone of asset pricing theory, not because it perfectly describes reality, but because it provides the simplest coherent framework for understanding the relationship between risk and expected return. Every extension and alternative model in asset pricing (e.g., from the Fama-French factors to consumption-based pricing) builds upon or reacts against the CAPM's foundational logic.

In this chapter, we derive the CAPM from first principles, illustrate its theoretical underpinnings, and show how to estimate its parameters empirically. We download stock market data, estimate betas using regression analysis, and evaluate asset performance relative to model predictions.

```
import pandas as pd
import numpy as np
import tidyfinance as tf

from plotnine import *
from mizani.formatters import percent_format
from adjustText import adjust_text
```

## 4.2 Systematic versus Idiosyncratic Risk

Before diving into the mathematics, we need to understand the fundamental distinction that makes the CAPM work: the difference between *systematic* and *idiosyncratic* risk.

### 4.2.1 Idiosyncratic Risk: Diversifiable and Unrewarded

Consider events that affect individual companies but not the broader market: a CEO resigns unexpectedly, a product launch fails, earnings disappoint analysts, or a factory experiences a fire. These company-specific events can dramatically affect individual stock prices, but they tend to average out across a diversified portfolio. When one company has bad news, another often has good news; the shocks are largely uncorrelated.

This idiosyncratic (or firm-specific) risk can be eliminated through diversification. By holding a portfolio of many stocks, an investor can reduce idiosyncratic risk to nearly zero. Since this risk can be avoided at no cost, investors should not expect compensation for bearing it. In equilibrium, idiosyncratic risk earns no premium.

### 4.2.2 Systematic Risk: Undiversifiable and Priced

Systematic risk, by contrast, affects all assets simultaneously. Recessions, interest rate changes, geopolitical crises, and pandemics impact virtually every company to some degree. No amount of diversification can eliminate exposure to these economy-wide shocks, they are inherent to participating in the market.

Since systematic risk cannot be diversified away, investors genuinely dislike it. They must be compensated for bearing it. The CAPM formalizes this intuition: expected returns should depend only on systematic risk, not total risk. Two assets with identical total volatility can have very different expected returns if their systematic risk exposures differ.

### 4.2.3 A Simple Illustration

Imagine two stocks with identical 30% annual volatility. Stock A is a gold mining company whose returns move opposite to the overall market: it does well when the economy struggles and poorly when it booms. Stock B is a luxury retailer that amplifies market movements: soaring in good times and crashing in bad times.

Which stock should offer higher expected returns? Intuition might suggest they should be equal since both have the same volatility. But the CAPM says Stock B should offer substantially higher returns. Why? Because Stock B performs poorly precisely when investors' overall wealth is already down (during market crashes), making its returns particularly painful. Stock A, by contrast, provides insurance. Its strong performance during market downturns partially offsets

losses elsewhere in the portfolio. Investors value this insurance property and are willing to accept lower expected returns in exchange.

This is the CAPM's core insight: expected returns compensate investors for systematic risk exposure, measured by how an asset's returns co-move with the market portfolio.

## 4.3 Data Preparation

Building on our analysis from the previous chapter, we examine the VN30 constituents as our asset universe. We download and prepare monthly return data:

```
vn30_symbols = [
    "ACB", "BCM", "BID", "BVH", "CTG", "FPT", "GAS", "GVR", "HDB", "HPG",
    "MBB", "MSN", "MWG", "PLX", "POW", "SAB", "SHB", "SSB", "STB", "TCB",
    "TPB", "VCB", "VHM", "VIB", "VIC", "VJC", "VNM", "VPB", "VRE", "EIB"
]
```

```
import os
import boto3
from botocore.client import Config
from io import BytesIO

class ConnectMinio:
    def __init__(self):
        self.MINIO_ENDPOINT = os.environ["MINIO_ENDPOINT"]
        self.MINIO_ACCESS_KEY = os.environ["MINIO_ACCESS_KEY"]
        self.MINIO_SECRET_KEY = os.environ["MINIO_SECRET_KEY"]
        self.REGION = os.getenv("MINIO_REGION", "us-east-1")

        self.s3 = boto3.client(
            "s3",
            endpoint_url=self.MINIO_ENDPOINT,
            aws_access_key_id=self.MINIO_ACCESS_KEY,
            aws_secret_access_key=self.MINIO_SECRET_KEY,
            region_name=self.REGION,
            config=Config(signature_version="s3v4"),
        )

    def test_connection(self):
        resp = self.s3.list_buckets()
        print("Connected. Buckets:")
```

```
        for b in resp.get("Buckets", []):
            print(" -", b["Name"])


conn = ConnectMinio()
s3 = conn.s3
conn.test_connection()

bucket_name = os.environ["MINIO_BUCKET"]

prices = pd.read_csv(
    BytesIO(
        s3.get_object(
            Bucket=bucket_name,
            Key="historycal_price/dataset_historical_price.csv"
        )["Body"].read()
    ),
    low_memory=False
)
```

```
Connected. Buckets:
 - dsteam-data
 - rawbctc
```

We process the raw price data to compute adjusted closing prices and standardize column names:

```
prices["date"] = pd.to_datetime(prices["date"])
prices["adjusted_close"] = prices["close_price"] * prices["adj_ratio"]
prices = prices.rename(columns={
    "vol_total": "volume",
    "open_price": "open",
    "low_price": "low",
    "high_price": "high",
    "close_price": "close"
})
prices = prices.sort_values(["symbol", "date"])
```

```
prices_daily = prices[prices["symbol"].isin(vn30_symbols)]
prices_daily[["date", "symbol", "adjusted_close"]].head(3)
```

| | date | symbol | adjusted_close |
|---|---|---|---|
| 18176 | 2010-01-04 | ACB | 329.408244 |
| 18177 | 2010-01-05 | ACB | 329.408244 |
| 18178 | 2010-01-06 | ACB | 320.258015 |

### 4.3.1 Computing Monthly Returns

We aggregate daily prices to monthly frequency. Using monthly returns rather than daily returns offers several advantages for portfolio analysis: monthly returns exhibit less noise, better approximate normality, and reduce the impact of microstructure effects like bid-ask bounce.

```
returns_monthly = (prices_daily
    .assign(
        date=prices_daily["date"].dt.to_period("M").dt.to_timestamp("M")
    )
    .groupby(["symbol", "date"], as_index=False)
    .agg(adjusted_close=("adjusted_close", "last"))
    .sort_values(["symbol", "date"])
    .assign(
        ret=lambda x: x.groupby("symbol")["adjusted_close"].pct_change()
    )
)

returns_monthly.head(3)
```

| | symbol | date | adjusted_close | ret |
|---|---|---|---|---|
| 0 | ACB | 2010-01-31 | 291.975489 | NaN |
| 1 | ACB | 2010-02-28 | 303.621235 | 0.039886 |
| 2 | ACB | 2010-03-31 | 273.784658 | -0.098269 |

## 4.4 The Risk-Free Asset and the Investment Opportunity Set

### 4.4.1 Adding a Risk-Free Asset

The previous chapter on MPT considered portfolios composed entirely of risky assets, requiring that portfolio weights sum to one. The CAPM introduces a crucial new element: a **risk-free asset** that pays a constant interest rate $r_f$ with zero volatility.

This seemingly simple addition fundamentally transforms the investment opportunity set. With a risk-free asset available, investors can choose to park some wealth in the safe asset and invest the remainder in risky assets. They can also borrow at the risk-free rate to leverage their risky positions.

Let $\omega \in \mathbb{R}^N$ denote the portfolio weights in the $N$ risky assets. Unlike before, these weights need not sum to one. The remainder, $1 - \iota'\omega$ (where $\iota$ is a vector of ones), is invested in the risk-free asset.

### 4.4.2 Portfolio Return with a Risk-Free Asset

The expected return on this combined portfolio is:

$$\mu_\omega = \omega'\mu + (1 - \iota'\omega)r_f = r_f + \omega'(\mu - r_f) = r_f + \omega'\tilde{\mu}$$

where $\mu$ is the vector of expected returns on risky assets and $\tilde{\mu} = \mu - r_f$ denotes the vector of **excess returns** (returns above the risk-free rate).

This expression reveals an important decomposition: the portfolio's expected return equals the risk-free rate plus a risk premium determined by the exposure to risky assets.

### 4.4.3 Portfolio Variance

Since the risk-free asset has zero volatility and zero covariance with risky assets, only the risky portion contributes to portfolio variance:

$$\sigma_\omega^2 = \omega'\Sigma\omega$$

where $\Sigma$ is the variance-covariance matrix of risky asset returns. The portfolio's volatility (standard deviation) is:

$$\sigma_\omega = \sqrt{\omega'\Sigma\omega}$$

### 4.4.4 Setting Up the Risk-Free Rate

For a realistic proxy of the risk-free rate, we use the Vietnam government bond yield. Government bonds of stable economies are considered "risk-free" because the government can always print money to meet its obligations (though this may cause inflation).

```
all_dates = pd.date_range(
    start=returns_monthly["date"].min(),
    end=returns_monthly["date"].max(),
    freq="ME"
)

# Vietnam 10-Year Government Bond Yield (approximately 2.52% annualized)
rf_annual = 0.0252
rf_monthly_val = (1 + rf_annual)**(1/12) - 1

risk_free_monthly = pd.DataFrame({
    "date": all_dates,
    "risk_free": rf_monthly_val
})

risk_free_monthly["date"] = (
    pd.to_datetime(risk_free_monthly["date"])
    .dt.to_period("M")
    .dt.to_timestamp("M")
)

risk_free_monthly.head(3)
```

| | date | risk_free |
|---|---|---|
| 0 | 2010-01-31 | 0.002076 |
| 1 | 2010-02-28 | 0.002076 |
| 2 | 2010-03-31 | 0.002076 |

We merge the risk-free rate with our returns data and compute excess returns:

```
returns_monthly = returns_monthly.merge(
    risk_free_monthly[["date", "risk_free"]],
    on="date",
    how="left"
)

rf = risk_free_monthly["risk_free"].mean()

returns_monthly = (returns_monthly
    .assign(
```

```
        ret_excess=lambda x: x["ret"] - x["risk_free"]
    )
    .assign(
        ret_excess=lambda x: x["ret_excess"].clip(lower=-1)
    )
)

returns_monthly.head(3)
```

| | symbol | date | adjusted_close | ret | risk_free | ret_excess |
|---|---|---|---|---|---|---|
| 0 | ACB | 2010-01-31 | 291.975489 | NaN | 0.002076 | NaN |
| 1 | ACB | 2010-02-28 | 303.621235 | 0.039886 | 0.002076 | 0.037810 |
| 2 | ACB | 2010-03-31 | 273.784658 | -0.098269 | 0.002076 | -0.100345 |

## 4.5 The Tangency Portfolio: Where Everyone Invests

### 4.5.1 Deriving the Optimal Risky Portfolio

With a risk-free asset available, how should an investor allocate wealth across risky assets? Consider an investor who wants to achieve a target expected excess return $\bar{\mu}$ with minimum variance. The optimization problem becomes:

$$\min_{\omega} \omega' \Sigma \omega \quad \text{subject to} \quad \omega' \tilde{\mu} = \bar{\mu}$$

Using the Lagrangian method:

$$\mathcal{L}(\omega, \lambda) = \omega' \Sigma \omega - \lambda(\omega' \tilde{\mu} - \bar{\mu})$$

The first-order condition with respect to $\omega$ is:

$$\frac{\partial \mathcal{L}}{\partial \omega} = 2\Sigma\omega - \lambda\tilde{\mu} = 0$$

Solving for the optimal weights:

$$\omega^* = \frac{\lambda}{2}\Sigma^{-1}\tilde{\mu}$$

The constraint $\omega'\tilde{\mu} = \bar{\mu}$ determines $\lambda$:

$$\bar{\mu} = \tilde{\mu}'\omega^* = \frac{\lambda}{2}\tilde{\mu}'\Sigma^{-1}\tilde{\mu} \implies \lambda = \frac{2\bar{\mu}}{\tilde{\mu}'\Sigma^{-1}\tilde{\mu}}$$

Substituting back:

$$\omega^* = \frac{\bar{\mu}}{\tilde{\mu}'\Sigma^{-1}\tilde{\mu}}\Sigma^{-1}\tilde{\mu}$$

### 4.5.2 The Tangency Portfolio

Notice something remarkable: the *direction* of $\omega^*$ is always $\Sigma^{-1}\tilde{\mu}$, regardless of the target return $\bar{\mu}$. Only the *scale* changes. This means all investors, regardless of their risk preferences, hold the *same portfolio of risky assets*. They differ only in how much they allocate to this portfolio versus the risk-free asset.

To obtain the portfolio of risky assets that is fully invested (weights summing to one), we normalize:

$$\omega_{\tan} = \frac{\omega^*}{\iota'\omega^*} = \frac{\Sigma^{-1}(\mu - r_f)}{\iota'\Sigma^{-1}(\mu - r_f)}$$

This is called the **tangency portfolio** (or maximum Sharpe ratio portfolio) because it lies at the point where the efficient frontier is tangent to the capital market line.

### 4.5.3 The Sharpe Ratio and the Capital Market Line

The **Sharpe ratio** measures excess return per unit of volatility:

$$\text{Sharpe Ratio} = \frac{\mu_p - r_f}{\sigma_p}$$

The tangency portfolio maximizes the Sharpe ratio among all possible portfolios. Any combination of the risk-free asset and the tangency portfolio lies on a straight line in mean-standard deviation space, called the **Capital Market Line (CML)**:

$$\mu_p = r_f + \left(\frac{\mu_{\tan} - r_f}{\sigma_{\tan}}\right)\sigma_p$$

The slope of this line equals the Sharpe ratio of the tangency portfolio (i.e., the highest achievable Sharpe ratio).

### 4.5.4 Computing the Tangency Portfolio

Let's compute the tangency portfolio for our VN30 universe:

```
assets = (returns_monthly
    .groupby("symbol", as_index=False)
    .agg(
        mu=("ret", "mean"),
        sigma=("ret", "std")
    )
)

sigma = (returns_monthly
    .pivot(index="date", columns="symbol", values="ret")
    .cov()
)

mu = (returns_monthly
    .groupby("symbol")["ret"]
    .mean()
    .values
)
```

```
# Compute tangency portfolio weights
w_tan = np.linalg.solve(sigma, mu - rf)
w_tan = w_tan / np.sum(w_tan)

# Portfolio performance metrics
mu_w = w_tan.T @ mu
sigma_w = np.sqrt(w_tan.T @ sigma @ w_tan)

efficient_portfolios = pd.DataFrame([
    {"symbol": r"$\omega_{\mathrm{tan}}$", "mu": mu_w, "sigma": sigma_w},
    {"symbol": r"$r_f$", "mu": rf, "sigma": 0}
])

sharpe_ratio = (mu_w - rf) / sigma_w

print(f"Tangency Portfolio Sharpe Ratio: {sharpe_ratio:.4f}")
print(efficient_portfolios)
```

```
Tangency Portfolio Sharpe Ratio: -0.5552
```

```
                    symbol        mu     sigma
0  $\omega_{\mathrm{tan}}$ -0.041157  0.077866
1                    $r_f$  0.002076  0.000000
```

### 4.5.5 Visualizing the Efficient Frontier with a Risk-Free Asset

Figure 4.1 shows the efficient frontier when a risk-free asset is available. The frontier is now a straight line (the Capital Market Line) connecting the risk-free asset to the tangency portfolio and extending beyond.

```
efficient_portfolios_figure = (
    ggplot(efficient_portfolios, aes(x="sigma", y="mu"))
    + geom_point(data=assets)
    + geom_point(data=efficient_portfolios, color="blue", size=3)
    + geom_label(
        aes(label="symbol"),
        adjust_text={"arrowprops": {"arrowstyle": "-"}}
    )
    + scale_x_continuous(labels=percent_format())
    + scale_y_continuous(labels=percent_format())
    + labs(
        x="Volatility (Standard Deviation)",
        y="Expected Return",
        title="Efficient Frontier with Risk-Free Asset (VN30)"
    )
    + geom_abline(slope=sharpe_ratio, intercept=rf, linetype="dotted")
)

efficient_portfolios_figure.show()
```

Figure 4.1: The efficient frontier with a risk-free asset becomes a straight line (the Capital Market Line) connecting the risk-free rate to the tangency portfolio. Individual assets lie below this line, demonstrating the benefits of diversification.

You may notice that estimated expected returns appear quite low, some even negative. This is not a model failure but reflects the realities of estimation:

1. **Sample period matters**: If the estimation window includes market downturns (such as the 2022-2023 period), realized average returns can be near zero or negative. Mean-variance optimization takes sample means literally.
2. **Estimation noise in emerging markets**: With volatile emerging market data, sample means are dominated by noise. A few extremely bad months can push the average below the risk-free rate even if the long-run equity premium is positive.

This highlights a fundamental challenge in portfolio optimization: the inputs we observe (historical returns) are noisy estimates of the true parameters we need (expected future returns).

## 4.6 The CAPM Equation: Risk and Expected Return

### 4.6.1 From Individual Optimization to Market Equilibrium

So far, we've focused on one investor's optimization problem. The CAPM's power comes from considering what happens when *all* investors optimize simultaneously.

If all investors follow mean-variance optimization, they all hold some combination of the risk-free asset and the tangency portfolio. The only difference between investors is their risk tolerance. More risk-averse investors hold more of the risk-free asset, while risk-tolerant investors may even borrow at the risk-free rate to leverage their position in the tangency portfolio.

### 4.6.2 The Market Portfolio

In equilibrium, the total demand for each risky asset must equal its supply. Since all investors hold the same portfolio of risky assets (the tangency portfolio), the equilibrium portfolio weights must equal the *market capitalization weights*. The tangency portfolio *is* the market portfolio.

This insight has enormous practical implications: instead of estimating expected returns and covariances to compute the tangency portfolio, we can simply use the market portfolio (approximated by a broad market index) as a proxy.

### 4.6.3 Deriving the CAPM Equation

From the first-order conditions of the optimization problem, we derived that:

$$\tilde{\mu} = \frac{2}{\lambda}\Sigma\omega^*$$

Since $\omega^*$ is proportional to $\omega_{\tan}$, and in equilibrium $\omega_{\tan}$ equals the market portfolio $\omega_m$:

$$\tilde{\mu} = c \cdot \Sigma\omega_m$$

for some constant $c$. The $i$-th element of $\Sigma\omega_m$ is:

$$\sum_{j=1}^{N} \sigma_{ij}\omega_{m,j} = \mathrm{Cov}(r_i, r_m)$$

where $r_m = \sum_j \omega_{m,j} r_j$ is the return on the market portfolio.

For the market portfolio itself:

$$\tilde{\mu}_m = c \cdot \mathrm{Var}(r_m) = c \cdot \sigma_m^2$$

Therefore $c = \tilde{\mu}_m / \sigma_m^2$, and for any asset $i$:

$$\tilde{\mu}_i = \frac{\tilde{\mu}_m}{\sigma_m^2} \mathrm{Cov}(r_i, r_m) = \beta_i \tilde{\mu}_m$$

where:

$$\beta_i = \frac{\mathrm{Cov}(r_i, r_m)}{\mathrm{Var}(r_m)}$$

This is the famous **CAPM equation**:

$$E(r_i) - r_f = \beta_i [E(r_m) - r_f]$$

### 4.6.4 Interpreting Beta

Beta ($\beta_i$) measures an asset's **systematic risk** (i.e., its sensitivity to market movements). The interpretation is straightforward:

- $\beta = 1$: The asset moves one-for-one with the market (average systematic risk)
- $\beta > 1$: The asset amplifies market movements (aggressive, high systematic risk)
- $\beta < 1$: The asset dampens market movements (defensive, low systematic risk)
- $\beta < 0$: The asset moves opposite to the market (provides insurance)

The CAPM says that expected excess return is proportional to beta, not to total volatility. This explains why:

1. An asset with zero beta earns only the risk-free rate (i.e., its risk is entirely idiosyncratic).
2. An asset with beta of 1 earns the market risk premium
3. A negative-beta asset earns *less* than the risk-free rate (i.e., investors pay for its insurance properties).

## 4.7 The Security Market Line

The CAPM predicts a linear relationship between beta and expected return. This relationship is called the **Security Market Line (SML)**:

$$E(r_i) = r_f + \beta_i[E(r_m) - r_f]$$

Unlike the Capital Market Line (which plots expected return against total risk), the Security Market Line plots expected return against systematic risk (beta).

```
betas = (sigma @ w_tan) / (w_tan.T @ sigma @ w_tan)
assets["beta"] = betas.values

price_of_risk = float(w_tan.T @ mu - rf)

assets_figure = (
    ggplot(assets, aes(x="beta", y="mu"))
    + geom_point()
    + geom_abline(intercept=rf, slope=price_of_risk)
    + scale_y_continuous(labels=percent_format())
    + labs(
        x="Beta (Systematic Risk)",
        y="Expected Return",
        title="Security Market Line"
    )
    + annotate("text", x=0.05, y=rf + 0.001, label="Risk-free rate")
)

assets_figure.show()
```

You may observe that the estimated SML has a negative slope, which seems to contradict CAPM's prediction. This reflects a **negative estimated market risk premium** in our sample period (i.e., the market portfolio earned less than the risk-free rate).

When the market risk premium is negative, CAPM predicts that high-beta stocks should have *lower* expected returns than low-beta stocks. This is not a model failure, the model is behaving consistently. Rather, it reflects an unusual (but not impossible) sample period where risky assets underperformed safe assets.

This observation highlights an important distinction: CAPM describes *expected* returns in equilibrium, but *realized* returns over any particular period may differ substantially from expectations due to shocks and surprises.

## 4.8 Empirical Estimation of CAPM Parameters

### 4.8.1 The Regression Framework

In practice, we estimate CAPM parameters using time-series regression. The model implies:

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i(r_{m,t} - r_{f,t}) + \varepsilon_{i,t}$$

where:

- $r_{i,t}$: Return on asset $i$ at time $t$
- $r_{f,t}$: Risk-free rate at time $t$
- $r_{m,t}$: Market return at time $t$
- $\alpha_i$: Intercept (should be zero if CAPM holds)
- $\beta_i$: Systematic risk (slope coefficient)
- $\varepsilon_{i,t}$: Idiosyncratic shock (residual)

### 4.8.2 Alpha: Risk-Adjusted Performance

The intercept $\alpha_i$ measures **risk-adjusted performance**. If CAPM holds perfectly, alpha should be zero for all assets (i.e., any excess return is exactly compensated by systematic risk).

- $\alpha > 0$: The asset outperformed its CAPM-predicted return (positive abnormal return)
- $\alpha < 0$: The asset underperformed its CAPM-predicted return (negative abnormal return)

Positive alpha is the holy grail of active management: earning returns beyond what systematic risk exposure would justify.

### 4.8.3 Loading Factor Data

We use Fama-French market excess returns as our market portfolio proxy. These data provide a widely accepted benchmark that is already adjusted for the risk-free rate:

```
import sqlite3

tidy_finance = sqlite3.connect(database="data/tidy_finance_python.sqlite")

factors = pd.read_sql_query(
    sql="SELECT * FROM factors_ff5_monthly",
    con=tidy_finance,
    parse_dates={"date"}
)

factors.head(3)
```

| | date | smb | hml | rmw | cma | mkt_excess |
|---|---|---|---|---|---|---|
| 0 | 2011-07-31 | 0.029280 | -0.049915 | 0.013239 | -0.025244 | -0.067002 |
| 1 | 2011-08-31 | 0.022241 | -0.021097 | 0.001543 | -0.021929 | 0.049073 |
| 2 | 2011-09-30 | 0.026609 | 0.014323 | 0.025125 | 0.003444 | -0.017362 |

### 4.8.4 Running the Regressions

We estimate CAPM regressions for each stock in our universe:

```
import statsmodels.formula.api as smf

returns_excess_monthly = (returns_monthly
    .merge(factors, on="date", how="left")
    .assign(ret_excess=lambda x: x["ret"] - x["risk_free"])
)


def estimate_capm(data):
    model = smf.ols("ret_excess ~ mkt_excess", data=data).fit()
    result = pd.DataFrame({
        "coefficient": ["alpha", "beta"],
        "estimate": model.params.values,
        "t_statistic": model.tvalues.values
    })
    return result


capm_results = (returns_excess_monthly
    .groupby("symbol", group_keys=True)
    .apply(estimate_capm)
    .reset_index()
)

capm_results.head(4)
```

| | symbol | level_1 | coefficient | estimate | t_statistic |
|---|---|---|---|---|---|
| 0 | ACB | 0 | alpha | -0.000677 | -0.086189 |
| 1 | ACB | 1 | beta | 0.611184 | 4.518154 |
| 2 | BCM | 0 | alpha | 0.035585 | 2.410328 |
| 3 | BCM | 1 | beta | 1.062267 | 4.666808 |

### 4.8.5 Visualizing Alpha Estimates

Figure 4.2 shows the estimated alphas across our VN30 sample. Statistical significance (at the 95% level) is indicated by color.

```
alphas = (capm_results
    .query("coefficient == 'alpha'")
    .assign(is_significant=lambda x: np.abs(x["t_statistic"]) >= 1.96)
)

alphas["symbol"] = pd.Categorical(
    alphas["symbol"],
    categories=alphas.sort_values("estimate")["symbol"],
    ordered=True
)

alphas_figure = (
    ggplot(alphas, aes(y="estimate", x="symbol", fill="is_significant"))
    + geom_col()
    + scale_y_continuous(labels=percent_format())
    + coord_flip()
    + labs(
        x="",
        y="Estimated Alpha (Monthly)",
        fill="Significant at 95%?",
        title="Estimated CAPM Alphas for VN30 Index Constituents"
    )
)

alphas_figure.show()
```

Figure 4.2: Estimated CAPM alphas for VN30 index constituents. Color indicates statistical significance at the 95% confidence level. Most alphas are statistically indistinguishable from zero, consistent with CAPM predictions.

The distribution of alphas provides evidence on CAPM's empirical validity. If the model holds, we expect:

1. Most alphas close to zero
2. Few statistically significant alphas
3. Roughly equal numbers of positive and negative alphas

Systematic patterns in alphas, such as consistently positive alphas for certain types of stocks, would suggest the CAPM is incomplete and that additional risk factors may be needed.

## 4.9 Limitations and Extensions

### 4.9.1 The Market Portfolio Problem

A fundamental challenge in testing the CAPM is identifying the market portfolio. The theory requires a portfolio that includes *all* investable assets, not just stocks, but also bonds, real estate, private businesses, human capital, and even intangible assets. In practice, we use proxies like broad market indices (VNI, S&P 500), but these capture only publicly traded equities.

This limitation is profound. As Richard Roll famously argued, the CAPM is essentially untestable because the true market portfolio is unobservable. Any test of the CAPM is simultaneously a test of whether our proxy adequately represents the market.

### 4.9.2 Time-Varying Betas

The CAPM assumes that betas are constant over time, but this assumption rarely holds in practice. Companies undergo changes that affect their market sensitivity:

- **Capital structure changes**: Increasing leverage raises beta
- **Business model evolution**: Diversification into new industries can alter systematic risk
- **Market conditions**: Betas often increase during market stress

Conditional CAPM models (Jagannathan and Wang 1996) address this by allowing risk premiums and betas to vary with the business cycle.

### 4.9.3 Empirical Anomalies

Decades of empirical research have documented patterns in stock returns that CAPM cannot explain:

1. **Size effect**: Small-cap stocks tend to outperform large-cap stocks, even after adjusting for beta
2. **Value effect**: Stocks with high book-to-market ratios outperform growth stocks
3. **Momentum**: Stocks that performed well recently tend to continue performing well

These anomalies suggest that systematic risk has multiple dimensions beyond market exposure.

### 4.9.4 Multifactor Extensions

The limitations of CAPM have led to increasingly sophisticated asset pricing models. The **Fama-French three-factor model** (Eugene F. Fama and French 1992) adds two factors to capture size and value effects:

- **SMB (Small Minus Big)**: Returns on small stocks minus large stocks
- **HML (High Minus Low)**: Returns on value stocks minus growth stocks

The **Fama-French five-factor model** (Eugene F. Fama and French 2015) adds two more dimensions:

- **RMW (Robust Minus Weak)**: Returns on profitable firms minus unprofitable firms

- **CMA (Conservative Minus Aggressive)**: Returns on conservative investors minus aggressive investors

The **Carhart four-factor model** (Mark M. Carhart 1997b) adds momentum to the three-factor framework.

Other theoretical developments include:

- **Consumption CAPM**: Links asset prices to macroeconomic consumption risk
- **Q-factor model** (Hou, Xue, and Zhang 2014): Derives factors from investment-based asset pricing theory
- **Arbitrage Pricing Theory**: Allows for multiple sources of systematic risk without specifying their identity

Despite its limitations, the CAPM remains valuable as a conceptual benchmark. Its core insight (i.e., only systematic, undiversifiable risk commands a premium) continues to inform how we think about risk and return.

## 4.10 Key Takeaways

This chapter introduced the Capital Asset Pricing Model and its implications for understanding the relationship between risk and expected return. The main insights are:

1. **Not all risk is rewarded**: The CAPM distinguishes between systematic risk (which cannot be diversified away and commands a premium) and idiosyncratic risk (which can be eliminated through diversification and earns no premium).

2. **The tangency portfolio is universal**: When a risk-free asset exists, all mean-variance investors hold the same portfolio of risky assets (i.e., the tangency or maximum Sharpe ratio portfolio). They differ only in how much they allocate to this portfolio versus the risk-free asset.

3. **In equilibrium, the tangency portfolio is the market portfolio**: Since all investors hold the same risky portfolio, and total demand must equal supply, the equilibrium portfolio weights are market capitalization weights.

4. **Expected returns depend on beta**: The CAPM equation states that expected excess return equals beta times the market risk premium. Beta measures covariance with the market portfolio, normalized by market variance.

5. **Alpha measures risk-adjusted performance**: Positive alpha indicates returns above what systematic risk would justify; negative alpha indicates underperformance.

6. **Empirical challenges exist**: Testing the CAPM requires identifying the market portfolio, which is unobservable in practice. Documented anomalies (size, value, momentum) suggest additional risk factors beyond market exposure.

7. **Extensions abound**: Multifactor models like Fama-French extend the CAPM framework by adding factors that capture dimensions of systematic risk the market factor misses.

The CAPM's elegance lies in its simplicity: a single factor (i.e., exposure to the market) should explain expected returns in equilibrium. While reality is more complex, this framework provides the foundation for all modern asset pricing theory.

# 5 Financial Statement Analysis

## 5.1 From Market Prices to Fundamental Value

The previous chapters focused on how financial markets price assets in equilibrium. The Capital Asset Pricing Model showed that expected returns depend on systematic risk exposure, while Modern Portfolio Theory demonstrated how to construct efficient portfolios. But these frameworks take expected returns and risk as given, they don't explain where these expectations come from.

Financial statement analysis addresses this gap. By examining a company's accounting records, investors can form independent assessments of firm value, identify mispriced securities, and understand the economic forces driving business performance. Financial statements provide the primary source of standardized information about a company's operations, financial position, and cash generation. Their legal requirements and standardized formats make them particularly valuable. Every publicly traded company must file them, creating a level playing field for analysis.

This chapter introduces the three primary financial statements: the balance sheet, income statement, and cash flow statement. We then demonstrate how to transform raw accounting data into meaningful financial ratios that facilitate comparison across companies and over time. These ratios serve multiple purposes: they enable investors to benchmark companies against peers, help creditors assess default risk, and provide inputs for asset pricing models like the Fama-French factors we will encounter in later chapters.

Our analysis combines theoretical frameworks with practical implementation using Vietnamese market data. By the end of this chapter, you will understand how to access financial statements, calculate key ratios across multiple categories, and interpret these metrics in context.

```
import pandas as pd
import numpy as np

from plotnine import *
from mizani.formatters import percent_format
from adjustText import adjust_text
```

## 5.2 The Three Financial Statements

Before diving into ratios and analysis, we need to understand the three interconnected statements that form the foundation of financial reporting. Each statement answers a different question about the company, and together they provide a comprehensive picture of financial health.

### 5.2.1 The Balance Sheet: A Snapshot of Financial Position

The balance sheet captures a company's financial position at a specific moment in time, think of it as a photograph rather than a movie. It lists everything the company owns (assets), everything it owes (liabilities), and the residual claim belonging to shareholders (equity). These three components are linked by the fundamental accounting equation:

$$\text{Assets} = \text{Liabilities} + \text{Equity}$$

This equation is not merely a definition, it reflects a core economic principle. A company's resources (assets) must be financed from somewhere: either borrowed from creditors (liabilities) or contributed by owners (equity). Every transaction affects both sides equally, maintaining the balance.

**Assets** represent resources the company controls that are expected to generate future economic benefits:

- **Current assets** can be converted to cash within one year: cash and equivalents, short-term investments, accounts receivable (money owed by customers), and inventory (raw materials, work-in-progress, and finished goods)
- **Non-current assets** support operations beyond one year: property, plant, and equipment (PP&E), long-term investments, and intangible assets like patents, trademarks, and goodwill

**Liabilities** encompass obligations to external parties:

- **Current liabilities** come due within one year: accounts payable (money owed to suppliers), short-term debt, accrued expenses, and the current portion of long-term debt
- **Non-current liabilities** extend beyond one year: long-term debt, bonds payable, pension obligations, and deferred tax liabilities

**Shareholders' equity** represents the owners' residual claim:

- **Common stock** and additional paid-in capital from share issuance
- **Retained earnings** (i.e., accumulated profits reinvested rather than distributed as dividends)
- **Treasury stock**: shares repurchased by the company

Understanding these categories is essential for ratio analysis. Current assets and liabilities determine short-term liquidity, while the mix of debt and equity reveals capital structure choices.

### 5.2.2 The Income Statement: Performance Over Time

While the balance sheet provides a snapshot, the income statement (also called the profit and loss statement, or P&L) measures financial performance over a period (e.g., a quarter or year). It follows a hierarchical structure that progressively captures different levels of profitability:

$$\text{Revenue} - \text{COGS} = \text{Gross Profit}$$

$$\text{Gross Profit} - \text{Operating Expenses} = \text{Operating Income (EBIT)}$$

$$\text{EBIT} - \text{Interest} - \text{Taxes} = \text{Net Income}$$

Each line reveals something different about the business:

- **Revenue (Sales)**: Total income from goods or services sold (i.e., the "top line")
- **Cost of Goods Sold (COGS)**: Direct costs of producing what was sold (materials, direct labor, manufacturing overhead)
- **Gross Profit**: Revenue minus COGS, measuring basic profitability from core operations
- **Operating Expenses**: Costs of running the business beyond production (selling, general & administrative expenses, research & development)
- **Operating Income (EBIT)**: Earnings Before Interest and Taxes, measuring profitability from operations before financing decisions and taxes
- **Interest Expense**: The cost of debt financing
- **Net Income**: The "bottom line" (i.e., total profit after all expenses)

The income statement's hierarchical structure allows analysts to identify where profitability problems originate. A company with strong gross margins but weak net income might have bloated overhead costs. One with weak gross margins faces fundamental pricing or production challenges.

### 5.2.3 The Cash Flow Statement: Following the Money

The cash flow statement bridges a critical gap: profitable companies can run out of cash, and unprofitable companies can generate positive cash flow. This happens because accrual accounting (used in the income statement) recognizes revenue when earned and expenses when incurred, not when cash changes hands.

The cash flow statement tracks actual cash movements, divided into three categories:

- **Operating activities**: Cash generated from core business operations. Starts with net income, then adjusts for non-cash items (depreciation, changes in working capital)
- **Investing activities**: Cash spent on or received from long-term investments (e.g., purchasing equipment, acquiring businesses, selling assets)
- **Financing activities**: Cash flows from capital structure decisions (e.g., issuing stock, borrowing, repaying debt, paying dividends, buying back shares)

A company can show strong net income while burning cash if it's building inventory, extending generous credit terms, or making large capital expenditures. Conversely, a company reporting losses might generate positive operating cash flow by collecting receivables faster than it pays suppliers.

### 5.2.4 Illustrating with FPT's Financial Statements

To see these concepts in practice, let's examine FPT Corporation's 2023 financial statements. FPT is one of Vietnam's largest technology companies, providing IT services, telecommunications, and education.

```
# Placeholder for FPT balance sheet visualization
# In practice, this would display the actual PDF or cleaned data
# from DataCore's acquisition pipeline

# Example structure of what the balance sheet data looks like:
# Assets: Current assets (cash, receivables, inventory) + Non-current assets (PP&E, intangibl
# Liabilities: Current liabilities (payables, short-term debt) + Non-current liabilities (lon
# Equity: Common stock + Retained earnings
```

The balance sheet demonstrates the fundamental accounting equation in action. FPT's assets (e.g., spanning cash, receivables, technology infrastructure, and intangible assets like software) exactly equal the sum of its liabilities and equity.

```
# Placeholder for FPT income statement visualization
# Shows the progression from revenue through various profit measures to net income
```

FPT's income statement reveals how the company transforms revenue into profit. The progression from gross profit through operating income to net income shows the impact of operating expenses, interest costs, and taxes.

```
# Placeholder for FPT cash flow statement visualization
# Reconciles net income with actual cash generation
```

The cash flow statement shows how FPT's reported profits translate into actual cash. Differences between net income and operating cash flow reveal the impact of working capital management and non-cash expenses.

## 5.3 Loading Financial Statement Data

We now turn to systematic analysis across multiple companies. We load financial statement data for the VN30 index constituents (i.e., the 30 largest and most liquid stocks on Vietnam's Ho Chi Minh Stock Exchange).

```
import sqlite3

tidy_finance = sqlite3.connect(database="data/tidy_finance_python.sqlite")

comp_vn = pd.read_sql_query(
    sql="SELECT * FROM comp_vn",
    con=tidy_finance,
    parse_dates={"datadate"}
)

comp_vn.head(3)
```

| | symbol | year | total_current_asset | ca_fin | ca_cce | ca_cash | ca_cash_inbank | ca_cash |
|---|---|---|---|---|---|---|---|---|
| 0 | AGF | 1998 | 8.845141e+10 | None | 5.469709e+09 | 0.000000e+00 | None | None |
| 1 | BBC | 1999 | 5.672574e+10 | None | 5.354939e+09 | 5.354939e+09 | None | None |
| 2 | AGF | 1999 | 9.558392e+10 | None | 2.609276e+09 | 0.000000e+00 | None | None |

```
vn30_symbols = [
    "ACB", "BCM", "BID", "BVH", "CTG", "FPT", "GAS", "GVR", "HDB", "HPG",
    "MBB", "MSN", "MWG", "PLX", "POW", "SAB", "SHB", "SSB", "STB", "TCB",
    "TPB", "VCB", "VHM", "VIB", "VIC", "VJC", "VNM", "VPB", "VRE", "EIB"
]
```

```
comp_vn30 = comp_vn[comp_vn["symbol"].isin(vn30_symbols)]
comp_vn30.head(3)
```

| | symbol | year | total_current_asset | ca_fin | ca_cce | ca_cash | ca_cash_inbank | ca_ca |
|---|---|---|---|---|---|---|---|---|
| 11 | FPT | 2002 | 5.098910e+11 | None | 1.027470e+11 | 0.000000e+00 | None | None |
| 17 | VNM | 2003 | 2.101406e+12 | None | 6.925924e+11 | 6.925924e+11 | None | None |
| 21 | FPT | 2003 | 9.171390e+11 | None | 7.995600e+10 | 0.000000e+00 | None | None |

This dataset provides the foundation for calculating financial ratios and conducting cross-sectional comparisons. Each row contains balance sheet, income statement, and cash flow items for a company-year observation.

## 5.4 Liquidity Ratios: Can the Company Pay Its Bills?

Liquidity ratios assess a company's ability to meet short-term obligations. These metrics matter most to creditors, suppliers, and employees who need assurance that the company can pay its bills. They're calculated using balance sheet items, comparing liquid assets against near-term liabilities.

### 5.4.1 The Current Ratio

The most basic liquidity measure compares all current assets to current liabilities:

$$\text{Current Ratio} = \frac{\text{Current Assets}}{\text{Current Liabilities}}$$

A ratio above one indicates the company has enough current assets to cover obligations due within one year. However, the interpretation depends heavily on the composition of current assets. A company with current assets tied up in slow-moving inventory is less liquid than one holding cash.

### 5.4.2 The Quick Ratio

The quick ratio (or "acid test") provides a more stringent measure by excluding inventory:

$$\text{Quick Ratio} = \frac{\text{Current Assets} - \text{Inventory}}{\text{Current Liabilities}}$$

Why exclude inventory? Inventory is typically the least liquid current asset. It must be sold (potentially at a discount) before generating cash. A company facing a liquidity crisis cannot easily convert raw materials or finished goods into immediate cash. The quick ratio answers: "Can we pay our bills without relying on inventory sales?"

### 5.4.3 The Cash Ratio

The most conservative liquidity measure focuses solely on the most liquid assets:

$$\text{Cash Ratio} = \frac{\text{Cash and Cash Equivalents}}{\text{Current Liabilities}}$$

While a ratio of one indicates robust liquidity, most companies maintain lower cash ratios to avoid holding excessive non-productive assets. Cash sitting in bank accounts could otherwise be invested in growth opportunities, returned to shareholders, or used to pay down costly debt.

### 5.4.4 Calculating Liquidity Ratios

Let's compute these ratios for our VN30 sample:

```
balance_sheet_statements = (comp_vn30
    .assign(
        fiscal_year=lambda x: x["year"].astype(int),

        # Current Ratio: Current Assets / Current Liabilities
        current_ratio=lambda x: x["act"] / x["lct"],

        # Quick Ratio: (Current Assets - Inventory) / Current Liabilities
        quick_ratio=lambda x: (x["act"] - x["inv"]) / x["lct"],

        # Cash Ratio: Cash and Equivalents / Current Liabilities
        cash_ratio=lambda x: x["ca_cce"] / x["lct"],

        label=lambda x: np.where(
```

```
            x["symbol"].isin(vn30_symbols), x["symbol"], np.nan
        )
    )
)

balance_sheet_statements.head(3)
```

| | symbol | year | total_current_asset | ca_fin | ca_cce | ca_cash | ca_cash_inbank | ca_ca |
|---|---|---|---|---|---|---|---|---|
| 11 | FPT | 2002 | 5.098910e+11 | None | 1.027470e+11 | 0.000000e+00 | None | None |
| 17 | VNM | 2003 | 2.101406e+12 | None | 6.925924e+11 | 6.925924e+11 | None | None |
| 21 | FPT | 2003 | 9.171390e+11 | None | 7.995600e+10 | 0.000000e+00 | None | None |

### 5.4.5 Cross-Sectional Comparison of Liquidity

Figure 5.1 compares liquidity ratios across companies for the most recent fiscal year. This cross-sectional view reveals how different business models and industries maintain different liquidity profiles.

```
liquidity_ratios = (balance_sheet_statements
    .query("year == 2023 & label.notna()")
    .get(["symbol", "current_ratio", "quick_ratio", "cash_ratio"])
    .melt(id_vars=["symbol"], var_name="name", value_name="value")
    .assign(
        name=lambda x: x["name"].str.replace("_", " ").str.title()
    )
)

liquidity_ratios_figure = (
    ggplot(liquidity_ratios, aes(y="value", x="name", fill="symbol"))
    + geom_col(position="dodge")
    + coord_flip()
    + labs(
        x="", y="Ratio Value", fill="",
        title="Liquidity Ratios for VN30 Stocks (2023)"
    )
)

liquidity_ratios_figure.show()
```

Figure 5.1: Liquidity ratios measure a company's ability to meet short-term obligations. Higher values indicate greater liquidity, though excessively high ratios may suggest inefficient use of assets.

Several patterns emerge from this comparison. Banks and financial institutions typically show different liquidity profiles than industrial companies due to their unique business models. Companies with high inventory (retailers, manufacturers) often show larger gaps between current and quick ratios.

## 5.5 Leverage Ratios: How Is the Company Financed?

Leverage ratios examine a company's capital structure (i.e., the mix of debt and equity financing). These metrics reveal financial risk and long-term solvency, helping investors understand how much of the company's operations are funded by borrowed money.

### 5.5.1 Why Capital Structure Matters

A company's financing choice involves fundamental trade-offs:

- **Debt** offers tax advantages (interest is deductible) and doesn't dilute ownership, but creates fixed obligations that must be met regardless of business performance

- **Equity** provides flexibility (no required payments) but dilutes existing shareholders and may be more expensive than debt

Companies with high leverage amplify both gains and losses. In good times, shareholders capture more upside because profits aren't shared with additional equity holders. In bad times, fixed interest payments can push the company toward distress. This is why beta (systematic risk) tends to increase with leverage.

### 5.5.2 Debt-to-Equity Ratio

This ratio indicates how much debt financing the company uses relative to shareholder investment:

$$\text{Debt-to-Equity} = \frac{\text{Total Debt}}{\text{Total Equity}}$$

A ratio of 1.0 means equal parts debt and equity financing. Higher ratios indicate more aggressive use of leverage, which can enhance returns in good times but increases bankruptcy risk.

### 5.5.3 Debt-to-Asset Ratio

This ratio shows what percentage of assets are financed through debt:

$$\text{Debt-to-Asset} = \frac{\text{Total Debt}}{\text{Total Assets}}$$

A ratio of 0.5 means half the company's assets are debt-financed. This metric is bounded between 0 and 1 (assuming positive equity), making it easier to compare across companies than the debt-to-equity ratio.

### 5.5.4 Interest Coverage Ratio

While the above ratios measure leverage levels, interest coverage assesses the ability to service that debt:

$$\text{Interest Coverage} = \frac{\text{EBIT}}{\text{Interest Expense}}$$

This ratio answers: "How many times over can current operating profits cover interest obligations?" A ratio below 1.0 means operating income doesn't cover interest payments, which is a dangerous position. Ratios above 3-5 generally indicate comfortable coverage.

### 5.5.5 Calculating Leverage Ratios

```
balance_sheet_statements = balance_sheet_statements.assign(
    debt_to_equity=lambda x: x["total_debt"] / x["total_equity"],
    debt_to_asset=lambda x: x["total_debt"] / x["at"]
)

income_statements = (comp_vn30
    .assign(
        year=lambda x: x["year"].astype(int),
        # Handle zero interest expense to avoid infinity
        interest_coverage=lambda x: np.where(
            x["cfo_interest_expense"] > 0,
            x["is_net_business_profit"] / x["cfo_interest_expense"],
            np.nan
        ),
        label=lambda x: np.where(
            x["symbol"].isin(vn30_symbols), x["symbol"], np.nan
        )
    )
)
```

### 5.5.6 Leverage Trends Over Time

Figure 5.2 tracks how debt-to-asset ratios have evolved over time. Time-series analysis reveals whether companies are becoming more or less leveraged.

```
debt_to_asset = balance_sheet_statements.query("symbol in @vn30_symbols")

debt_to_asset_figure = (
    ggplot(debt_to_asset, aes(x="year", y="debt_to_asset", color="symbol"))
    + geom_line(size=1)
    + scale_y_continuous(labels=percent_format())
    + labs(
        x="", y="Debt-to-Asset Ratio", color="",
        title="Debt-to-Asset Ratios of VN30 Stocks Over Time"
    )
)

debt_to_asset_figure.show()
```

Figure 5.2: Debt-to-asset ratios show the proportion of assets financed by debt. Changes over time reflect evolving capital structure strategies and market conditions.

### 5.5.7 Cross-Sectional Leverage Comparison

Figure 5.3 provides a snapshot of leverage across all VN30 constituents for the most recent year.

```
debt_to_asset_comparison = balance_sheet_statements.query("year == 2023")

debt_to_asset_comparison["symbol"] = pd.Categorical(
    debt_to_asset_comparison["symbol"],
    categories=debt_to_asset_comparison.sort_values("debt_to_asset")["symbol"],
    ordered=True
)

debt_to_asset_comparison_figure = (
    ggplot(
        debt_to_asset_comparison,
        aes(y="debt_to_asset", x="symbol", fill="label")
    )
    + geom_col()
```

```
    + coord_flip()
    + scale_y_continuous(labels=percent_format())
    + labs(
        x="", y="Debt-to-Asset Ratio", fill="",
        title="Debt-to-Asset Ratios of VN30 Stocks (2023)"
    )
    + theme(legend_position="none")
)

debt_to_asset_comparison_figure.show()
```

Figure 5.3: Cross-sectional comparison of debt-to-asset ratios reveals industry patterns and company-specific financing strategies.

### 5.5.8 The Leverage-Coverage Trade-off

Figure 5.4 examines the relationship between leverage levels and debt-servicing ability. Companies with higher debt loads should ideally have stronger interest coverage to maintain financial stability.

```
interest_coverage = (income_statements
    .query("year == 2023")
    .get(["symbol", "year", "interest_coverage"])
    .merge(balance_sheet_statements, on=["symbol", "year"], how="left")
)

interest_coverage_figure = (
    ggplot(
        interest_coverage,
        aes(x="debt_to_asset", y="interest_coverage", color="label")
    )
    + geom_point(size=2)
    + geom_label(
        aes(label="label"),
        adjust_text={"arrowprops": {"arrowstyle": "-"}}
    )
    + scale_x_continuous(labels=percent_format())
    + labs(
        x="Debt-to-Asset Ratio", y="Interest Coverage Ratio",
        title="Leverage versus Interest Coverage for VN30 Stocks (2023)"
    )
    + theme(legend_position="none")
)

interest_coverage_figure.show()
```

Figure 5.4: The relationship between leverage and interest coverage reveals whether companies can comfortably service their debt. High leverage with low coverage indicates elevated financial risk.

The scatter plot reveals important patterns. Companies in the upper-left quadrant (low leverage, high coverage) have conservative financing with ample debt capacity. Those in the lower-right (high leverage, low coverage) face elevated financial risk.

## 5.6 Efficiency Ratios: How Well Are Assets Managed?

Efficiency ratios measure how effectively a company utilizes its assets and manages operations. These metrics help identify whether management is extracting maximum value from the company's resource base.

### 5.6.1 Asset Turnover

This ratio measures how efficiently a company uses total assets to generate revenue:

$$\text{Asset Turnover} = \frac{\text{Revenue}}{\text{Total Assets}}$$

A higher ratio indicates more efficient asset utilization: the company generates more sales per dollar of assets. However, optimal levels vary dramatically across industries. Retailers with minimal fixed assets might achieve turnovers above 2.0, while capital-intensive manufacturers might operate below 0.5.

### 5.6.2 Inventory Turnover

For companies carrying inventory, this ratio reveals how quickly stock moves through the business:

$$\text{Inventory Turnover} = \frac{\text{Cost of Goods Sold}}{\text{Inventory}}$$

Higher turnover suggests efficient inventory management (i.e., goods don't sit on shelves collecting dust). However, extremely high turnover might indicate stockout risks, while very low turnover could signal obsolete inventory or overinvestment in working capital.

We use COGS rather than revenue in the numerator because inventory is recorded at cost, not selling price. Using revenue would overstate turnover for high-margin businesses.

### 5.6.3 Receivables Turnover

This ratio measures how effectively a company collects payments from customers:

$$\text{Receivables Turnover} = \frac{\text{Revenue}}{\text{Accounts Receivable}}$$

Higher turnover indicates faster collection (i.e., customers pay promptly). Converting this to "days sales outstanding" (365 / turnover) gives the average collection period in days. Companies must balance collection efficiency against the sales impact of restrictive credit policies.

### 5.6.4 Calculating Efficiency Ratios

```
combined_statements = (balance_sheet_statements
    .get([
        "symbol", "year", "label", "current_ratio", "quick_ratio",
        "cash_ratio", "debt_to_equity", "debt_to_asset", "total_asset",
        "total_equity"
    ])
    .merge(
```

```
        (income_statements
            .get([
                "symbol", "year", "interest_coverage", "is_revenue",
                "is_cogs", "selling_general_and_administrative_expenses",
                "is_interest_expense", "is_gross_profit", "is_eat"
            ])
        ),
        on=["symbol", "year"],
        how="left"
    )
    .merge(
        (comp_vn30
            .assign(year=lambda x: x["year"].astype(int))
            .get(["symbol", "year", "ca_total_inventory", "ca_acc_receiv"])
        ),
        on=["symbol", "year"],
        how="left"
    )
)

combined_statements = combined_statements.assign(
    asset_turnover=lambda x: x["is_revenue"] / x["total_asset"],
    inventory_turnover=lambda x: x["is_cogs"] / x["ca_total_inventory"],
    receivables_turnover=lambda x: x["is_revenue"] / x["ca_acc_receiv"]
)
```

Efficiency ratios vary dramatically across industries, making peer comparison essential. A grocery store and a shipbuilder will have fundamentally different asset and inventory dynamics.

## 5.7 Profitability Ratios: Is the Company Making Money?

Profitability ratios evaluate how effectively a company converts activity into earnings. These metrics directly measure financial success and are among the most closely watched indicators by investors.

### 5.7.1 Gross Margin

The gross margin reveals what percentage of revenue remains after direct production costs:

$$\text{Gross Margin} = \frac{\text{Gross Profit}}{\text{Revenue}} = \frac{\text{Revenue} - \text{COGS}}{\text{Revenue}}$$

Higher gross margins indicate stronger pricing power, more efficient production, or a favorable product mix. This metric is particularly useful for comparing companies within an industry, as it reveals relative efficiency in core operations before overhead costs.

### 5.7.2 Profit Margin

The profit margin shows what percentage of revenue ultimately becomes net income:

$$\text{Profit Margin} = \frac{\text{Net Income}}{\text{Revenue}}$$

This comprehensive measure accounts for all costs (e.g., production, operations, interest, and taxes). Higher profit margins suggest effective overall cost management. However, optimal margins vary by industry: software companies routinely achieve 20%+ margins, while grocery stores operate on razor-thin 2-3% margins.

### 5.7.3 Return on Equity (ROE)

ROE measures how efficiently a company uses shareholders' investment to generate profits:

$$\text{Return on Equity} = \frac{\text{Net Income}}{\text{Total Equity}}$$

This metric directly addresses what shareholders care about: returns on their invested capital. Higher ROE indicates more effective use of equity, though interpretation requires caution. High leverage can artificially inflate ROE by reducing the equity base (e.g., a company financed 90% by debt will show spectacular ROE on modest profits).

### 5.7.4 The DuPont Decomposition

The DuPont framework decomposes ROE into three components that reveal different aspects of performance:

$$\text{ROE} = \underbrace{\frac{\text{Net Income}}{\text{Revenue}}}_{\text{Profit Margin}} \times \underbrace{\frac{\text{Revenue}}{\text{Assets}}}_{\text{Asset Turnover}} \times \underbrace{\frac{\text{Assets}}{\text{Equity}}}_{\text{Leverage}}$$

This decomposition shows that high ROE can come from different sources: strong profit margins (pricing power, cost control), efficient asset use (high turnover), or aggressive leverage. Understanding which driver dominates helps assess sustainability. ROE driven by margins is generally more sustainable than ROE driven by leverage.

### 5.7.5 Calculating Profitability Ratios

```
combined_statements = combined_statements.assign(
    gross_margin=lambda x: x["is_gross_profit"] / x["is_revenue"],
    profit_margin=lambda x: x["is_eat"] / x["is_revenue"],
    after_tax_roe=lambda x: x["is_eat"] / x["total_equity"]
)
```

### 5.7.6 Gross Margin Trends

Figure 5.5 tracks gross margin evolution over time, revealing whether companies are maintaining pricing power and production efficiency.

```
gross_margins = combined_statements.query("symbol in @vn30_symbols")

gross_margins_figure = (
    ggplot(gross_margins, aes(x="year", y="gross_margin", color="symbol"))
    + geom_line()
    + scale_y_continuous(labels=percent_format())
    + labs(
        x="", y="Gross Margin", color="",
        title="Gross Margins for VN30 Stocks (2019-2023)"
    )
)

gross_margins_figure.show()
```

Figure 5.5: Gross margin trends reveal changes in pricing power and production efficiency. Declining margins may signal increased competition or rising input costs.

### 5.7.7 From Gross to Net: Where Do Profits Go?

Figure 5.6 examines the relationship between gross and profit margins. The gap between them reveals the impact of operating expenses, interest, and taxes.

```
profit_margins = combined_statements.query("year == 2023")

profit_margins_figure = (
    ggplot(
        profit_margins,
        aes(x="gross_margin", y="profit_margin", color="label")
    )
    + geom_point(size=2)
    + geom_label(
        aes(label="label"),
        adjust_text={"arrowprops": {"arrowstyle": "-"}}
    )
    + scale_x_continuous(labels=percent_format())
    + scale_y_continuous(labels=percent_format())
```

```
    + labs(
        x="Gross Margin", y="Profit Margin",
        title="Gross versus Profit Margins for VN30 Stocks (2023)"
    )
    + theme(legend_position="none")
)

profit_margins_figure.show()
```

Figure 5.6: Comparing gross and profit margins reveals how much of gross profit survives operating expenses, interest, and taxes. Companies far below the diagonal have high overhead relative to gross profit.

Companies along the diagonal convert gross profit to net income efficiently. Those well below the diagonal face high operating costs, interest burdens, or tax rates that erode profitability.

## 5.8 Combining Financial Ratios: A Holistic View

Individual ratios provide specific insights, but combining them offers a more complete picture. A company might excel in profitability while struggling with liquidity, or maintain conservative leverage while underperforming on efficiency.

### 5.8.1 Ranking Companies Across Categories

Figure 5.7 compares company rankings across four ratio categories. Rankings closer to 1 indicate better performance within each category, enabling quick identification of relative strengths and weaknesses.

```
financial_ratios = (combined_statements
    .query("year == 2023")
    .filter(
        items=["symbol"] + [
            col for col in combined_statements.columns
            if any(x in col for x in [
                "ratio", "margin", "roe", "_to_", "turnover", "interest_coverage"
            ])
        ]
    )
    .melt(id_vars=["symbol"], var_name="name", value_name="value")
    .assign(
        type=lambda x: np.select(
            [
                x["name"].isin(["current_ratio", "quick_ratio", "cash_ratio"]),
                x["name"].isin(["debt_to_equity", "debt_to_asset", "interest_coverage"]),
                x["name"].isin(["asset_turnover", "inventory_turnover", "receivables_turnover
                x["name"].isin(["gross_margin", "profit_margin", "after_tax_roe"]),
            ],
            [
                "Liquidity Ratios",
                "Leverage Ratios",
                "Efficiency Ratios",
                "Profitability Ratios"
            ],
            default="Other"
        )
    )
)

financial_ratios["rank"] = (financial_ratios
    .sort_values(["type", "name", "value"], ascending=[True, True, False])
    .groupby(["type", "name"])
    .cumcount() + 1
)

final_ranks = (financial_ratios
```

```
    .groupby(["symbol", "type"], as_index=False)
    .agg(rank=("rank", "mean"))
    .query("symbol in @vn30_symbols")
)

final_ranks_figure = (
    ggplot(final_ranks, aes(x="rank", y="type", color="symbol"))
    + geom_point(shape="^", size=4)
    + labs(
        x="Average Rank (Lower is Better)", y="", color="",
        title="Average Rank Across Financial Ratio Categories"
    )
    + coord_cartesian(xlim=[1, 30])
)

final_ranks_figure.show()
```

Figure 5.7: Ranking companies across multiple ratio categories reveals overall financial profiles. Companies with consistently low ranks across categories demonstrate broad-based financial strength.

The combined view reveals how different business strategies manifest in financial profiles. A

company might deliberately accept lower profitability rankings in exchange for stronger liquidity, or use aggressive leverage to boost returns at the cost of financial flexibility.

## 5.9 Financial Ratios in Asset Pricing

Beyond evaluating individual companies, financial ratios serve as crucial inputs for asset pricing models. The Fama-French five-factor model, which we explore in detail in Fama-French Factors, uses several accounting-based measures to explain cross-sectional variation in stock returns.

### 5.9.1 The Fama-French Factors

The model incorporates four company characteristics derived from financial statements:

**Size** is measured as the logarithm of market capitalization:

$$\text{Size} = \ln(\text{Market Cap})$$

This captures the empirical finding that smaller firms tend to outperform larger firms on a risk-adjusted basis (i.e., the "size premium").

**Book-to-Market** relates accounting value to market value:

$$\text{Book-to-Market} = \frac{\text{Book Equity}}{\text{Market Cap}}$$

High book-to-market stocks ("value" stocks) have historically outperformed low book-to-market stocks ("growth" stocks) (i.e., the "value premium").

**Operating Profitability** measures profit generation relative to equity:

$$\text{Profitability} = \frac{\text{Revenue} - \text{COGS} - \text{SG\&A} - \text{Interest}}{\text{Book Equity}}$$

More profitable firms tend to earn higher returns (i.e., the "profitability premium").

**Investment** captures asset growth:

$$\text{Investment} = \frac{\text{Total Assets}_t}{\text{Total Assets}_{t-1}} - 1$$

Firms investing aggressively tend to underperform conservative investors (i.e., the "investment premium").

### 5.9.2 Calculating Fama-French Variables

```
prices_monthly = pd.read_sql_query(
    sql="SELECT * FROM prices_monthly",
    con=tidy_finance,
    parse_dates={"datadate"}
)

# Use December prices for annual calculations
prices_december = (prices_monthly
    .assign(date=lambda x: pd.to_datetime(x["date"]))
    .query("date.dt.month == 12")
)
```

```
combined_statements_ff = (combined_statements
    .query("year == 2023")
    .merge(prices_december, on=["symbol", "year"], how="left")
    .merge(
        (balance_sheet_statements
            .query("year == 2022")
            .get(["symbol", "total_asset"])
            .rename(columns={"total_asset": "total_assets_lag"})
        ),
        on="symbol",
        how="left"
    )
    .assign(
        size=lambda x: np.log(x["mktcap"]),
        book_to_market=lambda x: x["total_equity"] / x["mktcap"],
        operating_profitability=lambda x: (
            (x["is_revenue"] - x["is_cogs"] -
             x["selling_general_and_administrative_expenses"] -
             x["is_interest_expense"]) / x["total_equity"]
        ),
        investment=lambda x: x["total_asset"] / x["total_assets_lag"] - 1
    )
)

combined_statements_ff.head(3)
```

| | symbol | year | label | current_ratio | quick_ratio | cash_ratio | debt_to_equity | debt_to_asset | tot |
|---|---|---|---|---|---|---|---|---|---|
| 0 | POW | 2023 | POW | 1.084255 | 1.084255 | 0.315089 | 0.0 | 0.0 | 7.0 |
| 1 | HPG | 2023 | HPG | 1.156655 | 1.156655 | 0.171324 | 0.0 | 0.0 | 1.8 |
| 2 | MWG | 2023 | MWG | 1.688604 | 1.688604 | 0.174408 | 0.0 | 0.0 | 6.0 |

### 5.9.3 Fama-French Factor Rankings

Figure 5.8 shows how VN30 companies rank on each Fama-French variable, connecting fundamental analysis to asset pricing.

```
factors_ranks = (combined_statements_ff
    .get(["symbol", "size", "book_to_market", "operating_profitability", "investment"])
    .rename(columns={
        "size": "Size",
        "book_to_market": "Book-to-Market",
        "operating_profitability": "Profitability",
        "investment": "Investment"
    })
    .melt(id_vars=["symbol"], var_name="name", value_name="value")
    .assign(
        rank=lambda x: (
            x.sort_values(["name", "value"], ascending=[True, False])
            .groupby("name")
            .cumcount() + 1
        )
    )
    .query("symbol in @vn30_symbols")
)

factors_ranks_figure = (
    ggplot(factors_ranks, aes(x="rank", y="name", color="symbol"))
    + geom_point(shape="^", size=4)
    + labs(
        x="Rank", y="", color="",
        title="Rank in Fama-French Variables for VN30 Stocks"
    )
    + coord_cartesian(xlim=[1, 30])
)

factors_ranks_figure.show()
```

Figure 5.8: Rankings on Fama-French variables connect financial statement analysis to asset pricing. According to factor models, smaller, higher book-to-market, more profitable, and lower-investment firms should earn higher expected returns.

These rankings have implications for expected returns according to factor models. A small, high book-to-market, highly profitable company with conservative investment should, in theory, earn higher risk-adjusted returns than its opposite.

## 5.10 Limitations and Practical Considerations

While financial ratios provide powerful analytical tools, several limitations deserve attention:

### 5.10.1 Accounting Discretion

Companies have significant discretion in how they apply accounting standards. Revenue recognition timing, depreciation methods, inventory valuation (FIFO vs. LIFO), and capitalization versus expensing decisions all affect reported numbers. Sophisticated analysis requires understanding these choices and their impact.

### 5.10.2 Industry Comparability

Ratios vary dramatically across industries. Comparing a bank's leverage to a retailer's is meaningless (e.g., banks naturally operate with much higher leverage due to their business model). Always benchmark against industry peers rather than absolute standards.

### 5.10.3 Point-in-Time Limitations

Balance sheet ratios capture a single moment, which may not represent typical conditions. Companies often "window dress" by temporarily improving metrics at reporting dates. Trend analysis and quarter-over-quarter comparisons can reveal such practices.

### 5.10.4 Backward-Looking Nature

Financial statements report historical results. Past profitability doesn't guarantee future performance, especially for companies in rapidly changing industries or facing disruption.

### 5.10.5 Quality of Earnings

Not all profits are created equal. Earnings driven by one-time gains, accounting adjustments, or aggressive revenue recognition may not recur. Cash flow analysis helps assess earnings quality. Profits that don't convert to cash warrant skepticism.

## 5.11 Key Takeaways

This chapter introduced financial statement analysis as a tool for understanding company fundamentals. The main insights are:

1. **Three statements, three perspectives**: The balance sheet shows financial position at a point in time, the income statement measures performance over a period, and the cash flow statement tracks actual cash movements. Together, they provide a complete picture of financial health.
2. **Liquidity ratios assess short-term survival**: Current, quick, and cash ratios measure the ability to meet near-term obligations. Higher ratios indicate greater liquidity but may suggest inefficient asset use.
3. **Leverage ratios reveal capital structure risk**: Debt-to-equity, debt-to-asset, and interest coverage ratios show how the company finances operations and whether it can service its debt. Higher leverage amplifies both returns and risk.

4. **Efficiency ratios measure management effectiveness**: Asset turnover, inventory turnover, and receivables turnover reveal how well the company converts resources into revenue. Industry context is essential for interpretation.
5. **Profitability ratios quantify financial success**: Gross margin, profit margin, and ROE measure the ability to generate earnings. The DuPont decomposition reveals whether ROE comes from margins, turnover, or leverage.
6. **Ratios connect to asset pricing**: Financial statement variables like book-to-market, profitability, and investment form the basis of factor models that explain cross-sectional return differences.
7. **Context matters for interpretation**: Ratios must be compared against industry peers, tracked over time, and considered alongside qualitative factors. No single ratio tells the complete story.

Looking ahead, subsequent chapters will explore how these fundamental variables interact with market prices in asset pricing models, and how to construct factor portfolios based on financial statement characteristics.

# 6 Discounted Cash Flow Analysis

## 6.1 What Is a Company Worth?

The previous chapters examined how markets price securities in equilibrium and how financial statements reveal company fundamentals. But these approaches leave a central question unanswered: What is the *intrinsic value* of a business, independent of its current market price?

Discounted Cash Flow (DCF) analysis answers this question by valuing a company based on its ability to generate cash for investors. The core insight is simple: a business is worth the present value of all future cash it will produce. This principle that value equals discounted future cash flows underlies virtually all of finance, from bond pricing to real estate valuation.

DCF analysis stands apart from other valuation approaches in three important ways. First, it explicitly accounts for the **time value of money** (i.e., the principle that a dollar today is worth more than a dollar tomorrow). By discounting future cash flows at an appropriate rate, we incorporate both time preferences and risk. Second, DCF is **forward-looking**, making it particularly suitable for companies where historical performance may not reflect future potential. Third, DCF is **flexible** enough to accommodate various business models and capital structures, making it applicable across industries and company sizes.

### 6.1.1 Valuation Methods Overview

Company valuation methods broadly fall into three categories:

- **Market-based approaches** compare companies using relative metrics like Price-to-Earnings or EV/EBITDA ratios. These are quick but assume comparable companies are fairly valued.
- **Asset-based methods** focus on the net value of tangible and intangible assets. These work well for liquidation scenarios but miss going-concern value.
- **Income-based techniques** value companies based on their ability to generate future cash flows. DCF is the most rigorous income-based method.

We focus on DCF because it forces analysts to make explicit assumptions about growth, profitability, and risk. These assumptions are often hidden in other methods. Even when DCF

isn't the final word on valuation, the discipline of building a DCF model deepens understanding of what drives value.

### 6.1.2 The Three Pillars of DCF

Every DCF analysis rests on three components:

1. **Free Cash Flow (FCF) forecasts**: The expected future cash available for distribution to investors after operating expenses, taxes, and investments
2. **Terminal value**: The company's value beyond the explicit forecast period, often representing a majority of total valuation
3. **Discount rate**: Typically the Weighted Average Cost of Capital (WACC), which adjusts future cash flows to present value by incorporating risk and capital structure

We make simplifying assumptions throughout this chapter. In particular, we assume firms conduct only operating activities (i.e., financial statements do not include non-operating items like excess cash or investment securities). Real-world valuations require valuing these separately. Entire textbooks are devoted to valuation nuances; our goal is to establish the conceptual framework and practical implementation.

```
import pandas as pd
import numpy as np
import statsmodels.formula.api as smf

from plotnine import *
from mizani.formatters import percent_format, comma_format
from itertools import product
```

## 6.2 Understanding Free Cash Flow

Before diving into calculations, we need to understand what Free Cash Flow represents and why it matters for valuation.

### 6.2.1 Why Free Cash Flow, Not Net Income?

Accountants report net income, but DCF uses free cash flow. Why the difference?

Net income includes non-cash items (like depreciation) and ignores cash needs (like capital expenditures and working capital investments). A company can report strong profits while burning cash, or generate substantial cash while reporting losses. Free cash flow captures what

actually matters for valuation: the cash available to distribute to all capital providers (both debt holders and equity holders) after funding operations and investments.

### 6.2.2 The Free Cash Flow Formula

We calculate FCF using the following formula:

$$\text{FCF} = \text{EBIT} \times (1 - \tau) + \text{D\&A} - \Delta\text{WC} - \text{CAPEX}$$

where:

- **EBIT** (Earnings Before Interest and Taxes): Core operating profit before financing costs and taxes
- $\tau$: Corporate tax rate applied to operating profits
- **D&A** (Depreciation & Amortization): Non-cash charges that reduce reported earnings but don't consume cash
- **ΔWC** (Change in Working Capital): Cash tied up in (or released from) operations (increases in receivables and inventory consume cash, while increases in payables provide cash)
- **CAPEX** (Capital Expenditures): Investments in long-term assets required to maintain and grow operations

An alternative formulation starts from EBIT directly:

$$\text{FCF} = \text{EBIT} + \text{D\&A} - \text{Taxes} - \Delta\text{WC} - \text{CAPEX}$$

Both formulations yield the same result when taxes are calculated consistently. The key insight is that FCF represents cash generated from operations after all reinvestment needs (i.e., cash that could theoretically be distributed to investors without impairing the business).

## 6.3 Loading Historical Financial Data

We use FPT Corporation, one of Vietnam's largest technology companies, as our case study. FPT provides IT services, telecommunications, and education. It's a diversified business with meaningful capital requirements and growth potential.

```python
import sqlite3

tidy_finance = sqlite3.connect(database="data/tidy_finance_python.sqlite")

comp_vn = pd.read_sql_query(
    sql="SELECT * FROM comp_vn",
    con=tidy_finance,
    parse_dates={"date"}
)

# Filter to FPT and examine the data structure
fpt_data = comp_vn[comp_vn["symbol"] == "FPT"].copy()
fpt_data["year"] = fpt_data["year"].astype(int)
fpt_data = fpt_data.sort_values("year").reset_index(drop=True)

print(f"Available years: {fpt_data['year'].min()} to {fpt_data['year'].max()}")
print(f"Number of observations: {len(fpt_data)}")
```

```
Available years: 2002 to 2023
Number of observations: 22
```

### 6.3.1 Computing Historical Free Cash Flow

Let's calculate the components needed for FCF from the financial statement data:

```python
# Extract and compute FCF components
historical_data = (fpt_data
    .assign(
        # Revenue for ratio calculations
        revenue=lambda x: x["is_net_revenue"],

        # EBIT = Earnings before interest and taxes
        # Approximate as EBT + Interest Expense
        ebit=lambda x: x["is_ebt"] + x["is_interest_expense"],

        # Tax payments (use actual tax expense)
        taxes=lambda x: x["is_cit_expense"],

        # Depreciation and amortization (non-cash add-back)
        depreciation=lambda x: x["cfo_depreciation"],
```

```
        # Change in working capital components
        # Positive delta_wc means cash is consumed (tied up in working capital)
        delta_working_capital=lambda x: (
            x["cfo_receive"] +      # Change in receivables
            x["cfo_inventory"] -    # Change in inventory
            x["cfo_payale"]         # Change in payables (negative = cash source)
        ),

        # Capital expenditures
        capex=lambda x: x["capex"]
    )
    .loc[:, [
        "year", "revenue", "ebit", "taxes", "depreciation",
        "delta_working_capital", "capex"
    ]]
)

# Calculate Free Cash Flow
historical_data["fcf"] = (
    historical_data["ebit"]
    - historical_data["taxes"]
    + historical_data["depreciation"]
    - historical_data["delta_working_capital"]
    - historical_data["capex"]
)

historical_data
```

| | year | revenue | ebit | taxes | depreciation | delta_working_capital | capex |
|---|---|---|---|---|---|---|---|
| 0 | 2002 | 1.514961e+12 | 2.698700e+10 | 0.000000e+00 | 1.261500e+10 | -2.561760e+11 | 2.202800 |
| 1 | 2003 | 4.148298e+12 | 5.676100e+10 | 0.000000e+00 | 1.837700e+10 | -5.078740e+11 | 3.753300 |
| 2 | 2004 | 8.734781e+12 | 2.145902e+11 | 1.795700e+10 | 2.947900e+10 | -4.280270e+11 | 5.252100 |
| 3 | 2005 | 1.410079e+13 | 3.753490e+11 | 4.251500e+10 | 5.381700e+10 | -4.471110e+11 | 1.428320 |
| 4 | 2006 | 2.139975e+13 | 6.672593e+11 | 7.368682e+10 | 1.068192e+11 | -1.173099e+12 | 2.459780 |
| 5 | 2007 | 1.349889e+13 | 1.071941e+12 | 1.487146e+11 | 1.709335e+11 | -1.873794e+12 | 4.802762 |
| 6 | 2008 | 1.638184e+13 | 1.320573e+12 | 1.890384e+11 | 2.395799e+11 | -1.419506e+11 | 6.690461 |
| 7 | 2009 | 1.840403e+13 | 1.807221e+12 | 2.916482e+11 | 3.041813e+11 | -8.065011e+11 | 7.632280 |
| 8 | 2010 | 2.001730e+13 | 2.261341e+12 | 3.314359e+11 | 3.294060e+11 | -2.360993e+12 | 8.672138 |
| 9 | 2011 | 2.537025e+13 | 2.751044e+12 | 4.223952e+11 | 3.759567e+11 | -2.099380e+12 | 4.524081 |
| 10 | 2012 | 2.459430e+13 | 2.635219e+12 | 4.210738e+11 | 3.995598e+11 | 8.043763e+11 | 7.083318 |
| 11 | 2013 | 2.702789e+13 | 2.690568e+12 | 4.503170e+11 | 4.429860e+11 | -1.947751e+12 | 9.110216 |

| | year | revenue | ebit | taxes | depreciation | delta_working_capital | capex |
|---|---|---|---|---|---|---|---|
| 12 | 2014 | 3.264466e+13 | 2.625389e+12 | 3.800994e+11 | 5.472736e+11 | -3.078130e+12 | 1.417399 |
| 13 | 2015 | 3.795970e+13 | 3.113651e+12 | 4.130641e+11 | 7.328801e+11 | -1.951778e+12 | 1.974295 |
| 14 | 2016 | 3.953147e+13 | 3.388085e+12 | 4.382078e+11 | 9.334397e+11 | -9.242713e+11 | 1.428472 |
| 15 | 2017 | 4.265861e+13 | 4.623663e+12 | 7.270039e+11 | 1.039417e+12 | -4.638788e+12 | 1.100498 |
| 16 | 2018 | 2.321354e+13 | 4.095947e+12 | 6.236054e+11 | 1.164692e+12 | -1.033438e+12 | 2.452902 |
| 17 | 2019 | 2.771696e+13 | 5.023518e+12 | 7.528183e+11 | 1.354613e+12 | -5.308818e+11 | 3.230818 |
| 18 | 2020 | 2.983040e+13 | 5.648794e+12 | 8.397114e+11 | 1.490607e+12 | -8.040730e+11 | 3.014322 |
| 19 | 2021 | 3.565726e+13 | 6.821202e+12 | 9.879053e+11 | 1.643916e+12 | -2.821825e+12 | 2.908134 |
| 20 | 2022 | 4.400953e+13 | 8.308009e+12 | 1.170940e+12 | 1.833064e+12 | -3.746661e+12 | 3.209581 |
| 21 | 2023 | 5.261790e+13 | 1.003565e+13 | 1.414956e+12 | 2.286514e+12 | -2.147304e+12 | 3.948982 |

### 6.3.2 Understanding the Historical Pattern

Before forecasting, we should understand the historical trends in FCF and its components:

```
# Calculate key ratios relative to revenue
historical_ratios = (historical_data
    .assign(
        # Revenue growth (year-over-year)
        revenue_growth=lambda x: x["revenue"].pct_change(),

        # Operating margin: EBIT as % of revenue
        operating_margin=lambda x: x["ebit"] / x["revenue"],

        # Depreciation as % of revenue
        depreciation_margin=lambda x: x["depreciation"] / x["revenue"],

        # Tax rate (taxes as % of revenue, for simplicity)
        tax_margin=lambda x: x["taxes"] / x["revenue"],

        # Working capital intensity
        working_capital_margin=lambda x: x["delta_working_capital"] / x["revenue"],

        # Capital intensity
        capex_margin=lambda x: x["capex"] / x["revenue"],

        # FCF margin
        fcf_margin=lambda x: x["fcf"] / x["revenue"]
    )
)
```

```
# Display key metrics
display_cols = [
    "year", "revenue_growth", "operating_margin", "depreciation_margin",
    "tax_margin", "working_capital_margin", "capex_margin", "fcf_margin"
]

historical_ratios[display_cols].round(3)
```

| | year | revenue_growth | operating_margin | depreciation_margin | tax_margin | working_capital_mar |
|---|---|---|---|---|---|---|
| 0 | 2002 | NaN | 0.018 | 0.008 | 0.000 | -0.169 |
| 1 | 2003 | 1.738 | 0.014 | 0.004 | 0.000 | -0.122 |
| 2 | 2004 | 1.106 | 0.025 | 0.003 | 0.002 | -0.049 |
| 3 | 2005 | 0.614 | 0.027 | 0.004 | 0.003 | -0.032 |
| 4 | 2006 | 0.518 | 0.031 | 0.005 | 0.003 | -0.055 |
| 5 | 2007 | -0.369 | 0.079 | 0.013 | 0.011 | -0.139 |
| 6 | 2008 | 0.214 | 0.081 | 0.015 | 0.012 | -0.009 |
| 7 | 2009 | 0.123 | 0.098 | 0.017 | 0.016 | -0.044 |
| 8 | 2010 | 0.088 | 0.113 | 0.016 | 0.017 | -0.118 |
| 9 | 2011 | 0.267 | 0.108 | 0.015 | 0.017 | -0.083 |
| 10 | 2012 | -0.031 | 0.107 | 0.016 | 0.017 | 0.033 |
| 11 | 2013 | 0.099 | 0.100 | 0.016 | 0.017 | -0.072 |
| 12 | 2014 | 0.208 | 0.080 | 0.017 | 0.012 | -0.094 |
| 13 | 2015 | 0.163 | 0.082 | 0.019 | 0.011 | -0.051 |
| 14 | 2016 | 0.041 | 0.086 | 0.024 | 0.011 | -0.023 |
| 15 | 2017 | 0.079 | 0.108 | 0.024 | 0.017 | -0.109 |
| 16 | 2018 | -0.456 | 0.176 | 0.050 | 0.027 | -0.045 |
| 17 | 2019 | 0.194 | 0.181 | 0.049 | 0.027 | -0.019 |
| 18 | 2020 | 0.076 | 0.189 | 0.050 | 0.028 | -0.027 |
| 19 | 2021 | 0.195 | 0.191 | 0.046 | 0.028 | -0.079 |
| 20 | 2022 | 0.234 | 0.189 | 0.042 | 0.027 | -0.085 |
| 21 | 2023 | 0.196 | 0.191 | 0.043 | 0.027 | -0.041 |

## 6.4 Visualizing Historical Ratios

Figure 6.1 shows the historical evolution of key financial ratios that drive FCF. Understanding these patterns helps inform our forecasts.

```
# Prepare data for plotting
ratio_columns = [
    "operating_margin", "depreciation_margin", "tax_margin",
    "working_capital_margin", "capex_margin"
]

ratios_long = (historical_ratios
    .melt(
        id_vars=["year"],
        value_vars=ratio_columns,
        var_name="ratio",
        value_name="value"
    )
    .assign(
        ratio=lambda x: x["ratio"].str.replace("_", " ").str.title()
    )
)

ratios_figure = (
    ggplot(ratios_long, aes(x="year", y="value", color="ratio"))
    + geom_line(size=1)
    + geom_point(size=2)
    + scale_y_continuous(labels=percent_format())
    + labs(
        x="", y="Ratio (% of Revenue)", color="",
        title="Key Financial Ratios of FPT Over Time"
    )
    + theme(legend_position="right")
)

ratios_figure.show()
```

Figure 6.1: Historical financial ratios reveal the operating characteristics of FPT. These patterns inform our forecast assumptions.

Several patterns emerge from the historical data. Operating margins show the profitability of core operations. Depreciation margins indicate asset intensity. CAPEX margins reveal investment requirements. Working capital margins can be volatile, reflecting changes in credit terms and inventory management.

## 6.5 Forecasting Free Cash Flow

With historical patterns established, we now project FCF into the future. This requires forecasting both revenue growth and the ratios that convert revenue into cash flow.

### 6.5.1 The Ratio-Based Forecasting Approach

We use a ratio-based approach that links all FCF components to revenue. This makes forecasting tractable: rather than projecting absolute dollar amounts for each component, we forecast (1) revenue growth and (2) how each component scales with revenue.

This approach embeds a key assumption: that the relationship between revenue and FCF components remains stable. In reality, operating leverage, investment needs, and working

capital requirements may change as companies mature. Sophisticated valuations model these dynamics explicitly.

### 6.5.2 Setting Forecast Assumptions

For our five-year forecast, we make the following assumptions about FPT's financial ratios. These should reflect industry analysis, company guidance, and competitive dynamics. Here we use estimates for illustration:

```
# Define the forecast horizon
last_historical_year = historical_data["year"].max()
forecast_years = list(range(last_historical_year + 1, last_historical_year + 6))
n_forecast_years = len(forecast_years)

print(f"Forecast period: {forecast_years[0]} to {forecast_years[-1]}")

# Define forecast ratios
# In practice, these would come from detailed analysis
forecast_assumptions = pd.DataFrame({
    "year": forecast_years,
    # Operating margin: slight improvement as scale increases
    "operating_margin": [0.12, 0.125, 0.13, 0.13, 0.135],
    # Depreciation: stable as % of revenue
    "depreciation_margin": [0.03, 0.03, 0.03, 0.028, 0.028],
    # Tax rate: stable
    "tax_margin": [0.02, 0.02, 0.02, 0.02, 0.02],
    # Working capital: modest cash consumption
    "working_capital_margin": [0.01, 0.01, 0.008, 0.008, 0.008],
    # CAPEX: declining as % of revenue as growth moderates
    "capex_margin": [0.05, 0.048, 0.045, 0.042, 0.04]
})

forecast_assumptions
```

```
Forecast period: 2024 to 2028
```

| | year | operating_margin | depreciation_margin | tax_margin | working_capital_margin | capex_margin |
|---|---|---|---|---|---|---|
| 0 | 2024 | 0.120 | 0.030 | 0.02 | 0.010 | 0.050 |
| 1 | 2025 | 0.125 | 0.030 | 0.02 | 0.010 | 0.048 |
| 2 | 2026 | 0.130 | 0.030 | 0.02 | 0.008 | 0.045 |

| | year | operating_margin | depreciation_margin | tax_margin | working_capital_margin | capex_margin |
|---|---|---|---|---|---|---|
| 3 | 2027 | 0.130 | 0.028 | 0.02 | 0.008 | 0.042 |
| 4 | 2028 | 0.135 | 0.028 | 0.02 | 0.008 | 0.040 |

### 6.5.3 Forecasting Revenue Growth

Revenue growth is often the most important and most uncertain assumption in DCF analysis. We demonstrate two approaches: using historical averages and linking growth to macroeconomic forecasts.

**Approach 1: Historical Average**

A simple approach uses the historical average growth rate:

```
historical_growth = historical_ratios["revenue_growth"].dropna()
avg_historical_growth = historical_growth.mean()

print(f"Average historical revenue growth: {avg_historical_growth:.1%}")
```

```
Average historical revenue growth: 25.2%
```

**Approach 2: GDP-Linked Growth**

A more sophisticated approach links company growth to GDP forecasts from institutions like the IMF. This captures the intuition that company revenues often move with broader economic activity.

```
# Vietnam GDP growth forecasts (illustrative, based on IMF WEO style projections)
# In practice, download from IMF WEO database
gdp_forecasts = pd.DataFrame({
    "year": forecast_years,
    "gdp_growth": [0.065, 0.063, 0.060, 0.058, 0.055]  # Gradually declining to long-term
})

# Assume FPT grows at a premium to GDP (tech sector outperformance)
# This premium should reflect company-specific factors
growth_premium = 0.05  # 5 percentage points above GDP

forecast_assumptions = forecast_assumptions.merge(gdp_forecasts, on="year")
forecast_assumptions["revenue_growth"] = (
    forecast_assumptions["gdp_growth"] + growth_premium
```

```
)

forecast_assumptions[["year", "gdp_growth", "revenue_growth"]]
```

| | year | gdp_growth | revenue_growth |
|---|---|---|---|
| 0 | 2024 | 0.065 | 0.115 |
| 1 | 2025 | 0.063 | 0.113 |
| 2 | 2026 | 0.060 | 0.110 |
| 3 | 2027 | 0.058 | 0.108 |
| 4 | 2028 | 0.055 | 0.105 |

### 6.5.4 Building the Forecast

Now we combine our assumptions to project revenue and FCF:

```
# Get the last historical revenue as our starting point
last_revenue = historical_data.loc[
    historical_data["year"] == last_historical_year, "revenue"
].values[0]

print(f"Last historical revenue ({last_historical_year}): {last_revenue/1e12:.2f} trillion VI

# Project revenue forward
forecast_data = forecast_assumptions.copy()
forecast_data["revenue"] = None

# Calculate revenue for each forecast year
for i, row in forecast_data.iterrows():
    if i == 0:
        # First forecast year: grow from last historical
        forecast_data.loc[i, "revenue"] = last_revenue * (1 + row["revenue_growth"])
    else:
        # Subsequent years: grow from previous forecast
        prev_revenue = forecast_data.loc[i-1, "revenue"]
        forecast_data.loc[i, "revenue"] = prev_revenue * (1 + row["revenue_growth"])

# Convert revenue to numeric
forecast_data["revenue"] = forecast_data["revenue"].astype(float)

# Calculate FCF components from ratios
```

```
forecast_data["ebit"] = forecast_data["operating_margin"] * forecast_data["revenue"]
forecast_data["depreciation"] = forecast_data["depreciation_margin"] * forecast_data["revenue
forecast_data["taxes"] = forecast_data["tax_margin"] * forecast_data["revenue"]
forecast_data["delta_working_capital"] = forecast_data["working_capital_margin"] * forecast_
forecast_data["capex"] = forecast_data["capex_margin"] * forecast_data["revenue"]

# Calculate FCF
forecast_data["fcf"] = (
    forecast_data["ebit"]
    - forecast_data["taxes"]
    + forecast_data["depreciation"]
    - forecast_data["delta_working_capital"]
    - forecast_data["capex"]
)

forecast_data[["year", "revenue", "ebit", "fcf"]].round(0)
```

```
Last historical revenue (2023): 52.62 trillion VND
```

| | year | revenue | ebit | fcf |
|---|---|---|---|---|
| 0 | 2024 | 5.866896e+13 | 7.040275e+12 | 4.106827e+12 |
| 1 | 2025 | 6.529855e+13 | 8.162319e+12 | 5.027988e+12 |
| 2 | 2026 | 7.248139e+13 | 9.422581e+12 | 6.305881e+12 |
| 3 | 2027 | 8.030938e+13 | 1.044022e+13 | 7.067226e+12 |
| 4 | 2028 | 8.874187e+13 | 1.198015e+13 | 8.430477e+12 |

## 6.6 Visualizing the Forecast

Figure 6.2 compares our forecast ratios with historical values, showing the transition from realized to projected performance.

```
# Prepare historical data for plotting
historical_plot = (historical_ratios
    .loc[:, ["year", "operating_margin", "depreciation_margin",
             "tax_margin", "working_capital_margin", "capex_margin"]]
    .assign(type="Historical")
)
```

```
# Prepare forecast data for plotting
forecast_plot = (forecast_data
    .loc[:, ["year", "operating_margin", "depreciation_margin",
             "tax_margin", "working_capital_margin", "capex_margin"]]
    .assign(type="Forecast")
)

# Combine
combined_ratios = pd.concat([historical_plot, forecast_plot], ignore_index=True)

# Reshape for plotting
combined_long = combined_ratios.melt(
    id_vars=["year", "type"],
    var_name="ratio",
    value_name="value"
)

combined_long["type"] = pd.Categorical(
    combined_long["type"],
    categories=["Historical", "Forecast"]
)

forecast_ratios_figure = (
    ggplot(combined_long, aes(x="year", y="value", color="ratio", linetype="type"))
    + geom_line(size=1)
    + geom_point(size=2)
    + scale_y_continuous(labels=percent_format())
    + labs(
        x="", y="Ratio (% of Revenue)", color="", linetype="",
        title="Historical and Forecast Financial Ratios for FPT"
    )
    + theme(legend_position="right")
)

forecast_ratios_figure.show()
```

Figure 6.2: Historical ratios (solid lines) and forecast assumptions (dashed lines) for key financial metrics. The forecast period begins after the last historical observation.

Figure 6.3 shows the revenue growth trajectory, comparing historical performance with our GDP-linked forecasts.

```
# Prepare growth data
historical_growth_df = (historical_ratios
    .loc[:, ["year", "revenue_growth"]]
    .dropna()
    .assign(type="Historical")
)

forecast_growth_df = (forecast_data
    .loc[:, ["year", "revenue_growth", "gdp_growth"]]
    .assign(type="Forecast")
)

# Combine for revenue growth
growth_combined = pd.concat([
    historical_growth_df,
    forecast_growth_df[["year", "revenue_growth", "type"]]
], ignore_index=True)
```

```
growth_combined["type"] = pd.Categorical(
    growth_combined["type"],
    categories=["Historical", "Forecast"]
)

growth_figure = (
    ggplot(growth_combined, aes(x="year", y="revenue_growth", linetype="type"))
    + geom_line(size=1, color="steelblue")
    + geom_point(size=2, color="steelblue")
    + scale_y_continuous(labels=percent_format())
    + labs(
        x="", y="Revenue Growth Rate", linetype="",
        title="Historical and Forecast Revenue Growth for FPT"
    )
)

growth_figure.show()
```

Figure 6.3: Revenue growth rates: historical (realized) and forecast (GDP-linked with company premium). The forecast assumes FPT grows at a premium to Vietnam's GDP growth.

Figure 6.4 presents the resulting FCF projections alongside historical values.

```
# Combine historical and forecast FCF
fcf_historical = (historical_data
    .loc[:, ["year", "fcf"]]
    .assign(type="Historical")
)

fcf_forecast = (forecast_data
    .loc[:, ["year", "fcf"]]
    .assign(type="Forecast")
)

fcf_combined = pd.concat([fcf_historical, fcf_forecast], ignore_index=True)
fcf_combined["type"] = pd.Categorical(
    fcf_combined["type"],
    categories=["Historical", "Forecast"]
)

fcf_figure = (
    ggplot(fcf_combined, aes(x="year", y="fcf/1e12", fill="type"))
    + geom_col()
    + labs(
        x="", y="Free Cash Flow (Trillion VND)", fill="",
        title="Historical and Forecast Free Cash Flow for FPT"
    )
)

fcf_figure.show()
```

Figure 6.4: Free Cash Flow: historical (realized) and forecast (projected). The forecast reflects our assumptions about revenue growth and operating ratios.

## 6.7 Terminal Value: Capturing Long-Term Value

A critical component of DCF analysis is the **terminal value** (or continuation value), which represents the company's value beyond the explicit forecast period. In most valuations, terminal value constitutes 50-80% of total enterprise value, making its estimation particularly important.

### 6.7.1 The Perpetuity Growth Model

The most common approach is the **Perpetuity Growth Model** (also called the Gordon Growth Model), which assumes FCF grows at a constant rate forever:

$$TV_T = \frac{FCF_{T+1}}{r-g} = \frac{FCF_T \times (1+g)}{r-g}$$

where:

- $TV_T$: Terminal value at the end of year $T$

- $FCF_T$: Free cash flow in the final forecast year
- $g$: Perpetual growth rate
- $r$: Discount rate (WACC)

### 6.7.2 Choosing the Perpetual Growth Rate

The perpetual growth rate $g$ should reflect long-term sustainable growth. Key considerations:

1. **No company can grow faster than the economy forever**. If it did, the company would eventually become larger than GDP, which is an impossibility. Long-term GDP growth (nominal, including inflation) provides an upper bound.
2. **Mature companies typically grow at or below GDP growth**. The perpetual growth rate should reflect the company in its "steady state," not its current high-growth phase.
3. **For Vietnam**, long-term nominal GDP growth might be 6-8% given current development stage, but this will moderate over time. A perpetual growth rate of 3-5% is often reasonable.

```python
def compute_terminal_value(last_fcf, growth_rate, discount_rate):
    """
    Compute terminal value using the perpetuity growth model.

    Parameters:
    -----------
    last_fcf : float
        Free cash flow in the final forecast year
    growth_rate : float
        Perpetual growth rate (g)
    discount_rate : float
        Discount rate / WACC (r)

    Returns:
    --------
    float : Terminal value
    """
    if discount_rate <= growth_rate:
        raise ValueError("Discount rate must exceed growth rate for finite terminal value")

    return last_fcf * (1 + growth_rate) / (discount_rate - growth_rate)
```

```
# Example calculation
last_fcf = forecast_data["fcf"].iloc[-1]
perpetual_growth = 0.04  # 4% perpetual growth
discount_rate = 0.10     # 10% WACC (placeholder)

terminal_value = compute_terminal_value(last_fcf, perpetual_growth, discount_rate)

print(f"Last forecast FCF: {last_fcf/1e12:.2f} trillion VND")
print(f"Terminal value (at {perpetual_growth:.0%} growth, {discount_rate:.0%} WACC): {termina
```

```
Last forecast FCF: 8.43 trillion VND
Terminal value (at 4% growth, 10% WACC): 146.1 trillion VND
```

### 6.7.3 Alternative: Exit Multiple Approach

Practitioners often cross-check terminal value using the **exit multiple approach**, which assumes the company is sold at the end of the forecast period at a multiple of EBITDA, EBIT, or revenue comparable to similar companies today.

For example, if comparable companies trade at 10x EBITDA, the terminal value would be:

$$TV_T = \text{EBITDA}_T \times \text{Exit Multiple}$$

This approach is simpler but embeds the assumption that current market multiples will persist (a strong assumption that may not hold).

## 6.8 The Discount Rate: Weighted Average Cost of Capital

The discount rate converts future cash flows to present value. For FCF (which goes to all capital providers), we use the **Weighted Average Cost of Capital (WACC)**:

$$WACC = \frac{E}{E+D} \times r_E + \frac{D}{E+D} \times r_D \times (1-\tau)$$

where:

- $E$: Market value of equity
- $D$: Market value of debt
- $r_E$: Cost of equity (typically estimated using CAPM)
- $r_D$: Cost of debt (pre-tax)

- $\tau$: Corporate tax rate

The $(1-\tau)$ term on debt reflects the tax shield. Interest payments are tax-deductible, reducing the effective cost of debt.

### 6.8.1 Estimating WACC Components

**Cost of Equity** is typically estimated using the Capital Asset Pricing Model. See the CAPM chapter:

$$r_E = r_f + \beta \times (r_m - r_f)$$

where $r_f$ is the risk-free rate, $\beta$ measures systematic risk, and $(r_m - r_f)$ is the market risk premium.

**Cost of Debt** can be estimated from:

- Interest expense divided by total debt (effective rate)
- Yields on the company's traded bonds
- Yields on bonds with similar credit ratings

**Capital Structure Weights** should use market values when available. For equity, market capitalization is straightforward. For debt, book value is often used when market values aren't observable.

### 6.8.2 Using Industry WACC Data

Professor Aswath Damodaran at NYU Stern maintains comprehensive industry WACC and country risk premium data. We use these datasets to estimate appropriate discount rates for Vietnamese companies.

#### 6.8.2.1 Downloading the Data

The following code downloads the required datasets. Run this manually when you need to update the data (typically annually), but it is not executed during book rendering to avoid dependency on external servers.

```python
import requests
from pathlib import Path

# Create data directory if needed
data_dir = Path("data")
data_dir.mkdir(exist_ok=True)

# Download WACC data
wacc_url = "https://pages.stern.nyu.edu/~adamodar/pc/datasets/wacc.xls"
response = requests.get(wacc_url, timeout=30)
response.raise_for_status()
(data_dir / "damodaran_wacc.xls").write_bytes(response.content)
print("Downloaded: damodaran_wacc.xls")

# Download country risk premium data
crp_url = "https://pages.stern.nyu.edu/~adamodar/pc/datasets/ctryprem.xlsx"
response = requests.get(crp_url, timeout=30)
response.raise_for_status()
(data_dir / "damodaran_crp.xlsx").write_bytes(response.content)
print("Downloaded: damodaran_crp.xlsx")
```

#### 6.8.2.2 Industry WACC

We extract the cost of capital for the Computer Services industry, which most closely matches FPT's business profile:

```python
import pandas as pd

# Read local WACC data
wacc_data = pd.read_excel(
    "data/damodaran_wacc.xls",
    sheet_name=1,
    skiprows=18
)

# Find WACC for Computer Services
industry_wacc = wacc_data.loc[
    wacc_data["Industry Name"] == "Computer Services",
    "Cost of Capital"
].values[0]
```

```
print(f"Industry WACC (Computer Services): {industry_wacc:.2%}")
```

```
Industry WACC (Computer Services): 7.83%
```

#### 6.8.2.3 Country Risk Premium

For Vietnamese companies, we must adjust for country-specific risk. Damodaran calculates country risk premiums based on sovereign credit ratings and relative equity market volatility:

```
# Read local country risk premium data
crp_data = pd.read_excel(
    "data/damodaran_crp.xlsx",
    sheet_name="ERPs by country",
    skiprows=7
)

# Find Vietnam's country risk premium
vietnam_row = crp_data[
    crp_data.iloc[:, 0].str.contains("Vietnam", case=False, na=False)
]

country_risk_premium = vietnam_row.iloc[0, 8]
print(f"Vietnam Country Risk Premium: {country_risk_premium:.2%}")
```

```
Vietnam Country Risk Premium: 2.09%
```

#### 6.8.2.4 Adjusted WACC for Vietnam

Combining the industry benchmark with the country risk premium gives us an appropriate discount rate:

```
wacc_vietnam = industry_wacc + country_risk_premium

print(f"Industry WACC (US):        {industry_wacc:.2%}")
print(f"Country Risk Premium:      {country_risk_premium:.2%}")
print(f"Adjusted WACC (Vietnam):   {wacc_vietnam:.2%}")

wacc = wacc_vietnam
```

```
Industry WACC (US):        7.83%
Country Risk Premium:      2.09%
Adjusted WACC (Vietnam):   9.92%
```

**Note**: This approach assumes Vietnamese companies in the same industry face similar operating risks as their US counterparts, with the country risk premium capturing additional macroeconomic and political risks. For company-specific analysis, further adjustments for leverage differences and firm-specific risk factors may be warranted.

## 6.9 Computing Enterprise Value

With all components in place, we can now compute enterprise value. The DCF formula is:

$$\text{Enterprise Value} = \sum_{t=1}^{T} \frac{FCF_t}{(1+WACC)^t} + \frac{TV_T}{(1+WACC)^T}$$

The first term is the present value of forecast-period cash flows; the second is the present value of terminal value.

```python
def compute_dcf_value(fcf_series, wacc, perpetual_growth):
    """
    Compute enterprise value using DCF analysis.

    Parameters:
    -----------
    fcf_series : array-like
        Free cash flows for forecast period
    wacc : float
        Weighted average cost of capital
    perpetual_growth : float
        Perpetual growth rate for terminal value

    Returns:
    --------
    dict : Components of DCF valuation
    """
    fcf = np.array(fcf_series)
    n_years = len(fcf)

    # Discount factors
```

```
    discount_factors = (1 + wacc) ** np.arange(1, n_years + 1)

    # Present value of forecast period cash flows
    pv_fcf = fcf / discount_factors
    pv_fcf_total = pv_fcf.sum()

    # Terminal value and its present value
    terminal_value = compute_terminal_value(fcf[-1], perpetual_growth, wacc)
    pv_terminal = terminal_value / discount_factors[-1]

    # Total enterprise value
    enterprise_value = pv_fcf_total + pv_terminal

    return {
        "pv_fcf": pv_fcf_total,
        "terminal_value": terminal_value,
        "pv_terminal": pv_terminal,
        "enterprise_value": enterprise_value,
        "terminal_pct": pv_terminal / enterprise_value
    }


# Compute DCF value
perpetual_growth = 0.04  # 4% perpetual growth

dcf_result = compute_dcf_value(
    fcf_series=forecast_data["fcf"].values,
    wacc=wacc,
    perpetual_growth=perpetual_growth
)

print("DCF Valuation Results")
print("=" * 50)
print(f"PV of Forecast Period FCF: {dcf_result['pv_fcf']/1e12:.1f} trillion VND")
print(f"Terminal Value: {dcf_result['terminal_value']/1e12:.1f} trillion VND")
print(f"PV of Terminal Value: {dcf_result['pv_terminal']/1e12:.1f} trillion VND")
print(f"Enterprise Value: {dcf_result['enterprise_value']/1e12:.1f} trillion VND")
print(f"Terminal Value as % of EV: {dcf_result['terminal_pct']:.1%}")
```

```
DCF Valuation Results
==================================================
PV of Forecast Period FCF: 22.7 trillion VND
```

```
Terminal Value: 148.1 trillion VND
PV of Terminal Value: 92.3 trillion VND
Enterprise Value: 115.1 trillion VND
Terminal Value as % of EV: 80.2%
```

Note that terminal value often represents 60-80% of enterprise value. This highlights the importance of terminal value assumptions and the inherent uncertainty in DCF analysis.

## 6.10 Sensitivity Analysis

Given the uncertainty in DCF inputs, sensitivity analysis is essential. We examine how enterprise value changes with different assumptions about WACC and perpetual growth.

```python
# Define ranges for sensitivity analysis
wacc_range = np.arange(0.08, 0.14, 0.01)  # 8% to 13%
growth_range = np.arange(0.02, 0.06, 0.01)  # 2% to 5%

# Create all combinations
sensitivity_results = []

for w in wacc_range:
    for g in growth_range:
        if w > g:  # Must have WACC > growth for valid terminal value
            result = compute_dcf_value(
                fcf_series=forecast_data["fcf"].values,
                wacc=w,
                perpetual_growth=g
            )
            sensitivity_results.append({
                "wacc": w,
                "growth_rate": g,
                "enterprise_value": result["enterprise_value"] / 1e12  # In trillions
            })

sensitivity_df = pd.DataFrame(sensitivity_results)

# Create heatmap
sensitivity_figure = (
    ggplot(sensitivity_df, aes(x="wacc", y="growth_rate", fill="enterprise_value"))
    + geom_tile()
    + geom_text(
```

```
        aes(label="enterprise_value"),
        format_string="{:.0f}",
        color="white",
        size=9
    )
    + scale_x_continuous(labels=percent_format())
    + scale_y_continuous(labels=percent_format())
    + scale_fill_gradient(low="darkblue", high="lightblue")
    + labs(
        x="WACC", y="Perpetual Growth Rate",
        fill="EV\n(Trillion VND)",
        title="DCF Sensitivity: Enterprise Value by WACC and Growth Rate"
    )
)

sensitivity_figure.show()
```

Figure 6.5: Sensitivity of enterprise value to WACC and perpetual growth rate assumptions. Small changes in these inputs can substantially affect valuation.

The sensitivity analysis reveals several important insights:

1. **Valuation is highly sensitive to inputs**: Small changes in WACC or growth rate produce large changes in enterprise value. A 1 percentage point change in WACC can shift value by 20% or more.
2. **The relationship is non-linear**: The impact of growth rate changes is amplified at lower WACCs because the terminal value formula has $(r - g)$ in the denominator.
3. **Reasonable people can disagree**: Given input uncertainty, DCF should be thought of as producing a *range* of values, not a single precise number.

## 6.11 From Enterprise Value to Equity Value

Our DCF analysis yields **enterprise value** (i.e., the total value of the company's operations to all capital providers). To determine **equity value** (what shareholders own), we must adjust for the claims of debt holders and any non-operating assets:

$$\text{Equity Value} = \text{Enterprise Value} + \text{Non-Operating Assets} - \text{Debt}$$

**Non-Operating Assets** include:

- Excess cash beyond operating needs
- Marketable securities
- Non-core real estate or investments

**Debt** includes:

- Short-term debt
- Long-term debt
- Capital lease obligations
- Preferred stock (if treated as debt-like)

```
# Get most recent balance sheet data for FPT
latest_year = fpt_data["year"].max()
latest_data = fpt_data[fpt_data["year"] == latest_year].iloc[0]

# Extract debt and cash (column names may vary)
total_debt = latest_data.get("total_debt", 0)
cash = latest_data.get("ca_cce", 0)

# Compute equity value
enterprise_value = dcf_result["enterprise_value"]
equity_value = enterprise_value - total_debt + cash
```

```
print("From Enterprise Value to Equity Value")
print("=" * 50)
print(f"Enterprise Value: {enterprise_value/1e12:.1f} trillion VND")
print(f"Less: Total Debt: {total_debt/1e12:.1f} trillion VND")
print(f"Plus: Cash: {cash/1e12:.1f} trillion VND")
print(f"Equity Value: {equity_value/1e12:.1f} trillion VND")
```

```
From Enterprise Value to Equity Value
==================================================
Enterprise Value: 115.1 trillion VND
Less: Total Debt: 0.0 trillion VND
Plus: Cash: 8.3 trillion VND
Equity Value: 123.3 trillion VND
```

### 6.11.1 Implied Share Price

If we know the number of shares outstanding, we can compute an implied share price:

```
# Get shares outstanding (this would come from market data)
# Using placeholder - in practice, get from exchange data
shares_outstanding = latest_data.get("total_equity", equity_value) / 25000  # Rough estimate

implied_price = equity_value / shares_outstanding

print(f"\nImplied Share Price: {implied_price:,.0f} VND")
```

```
Implied Share Price: 103,008 VND
```

Comparing the implied price to the current market price tells us whether the stock appears under- or overvalued according to our DCF model.

## 6.12 Limitations and Practical Considerations

DCF analysis is powerful but has important limitations:

### 6.12.1 Sensitivity to Assumptions

As our sensitivity analysis showed, small changes in inputs produce large changes in value. This is particularly problematic because the most influential inputs (long-term growth, WACC) are the hardest to estimate accurately.

### 6.12.2 Terminal Value Dominance

Terminal value often represents 60-80% of total value, yet it's based on assumptions about the very distant future. This concentrates valuation risk in the most uncertain component.

### 6.12.3 Garbage In, Garbage Out

DCF is only as good as its inputs. Unrealistic growth assumptions, optimistic margins, or inappropriate discount rates produce meaningless valuations. The discipline of DCF lies in forcing analysts to justify their assumptions.

### 6.12.4 Not Suitable for All Companies

DCF works best for companies with:

- Positive and predictable cash flows
- Stable or predictably changing margins
- Reasonable visibility into future operations

It struggles with:

- Early-stage companies with no profits
- Highly cyclical businesses
- Companies undergoing major transitions
- Financial institutions (which require different approaches)

### 6.12.5 Complement with Other Methods

Wise practitioners use DCF alongside other valuation methods:

- **Comparable company analysis**: How do similar companies trade?
- **Precedent transactions**: What have acquirers paid for similar businesses?
- **Sum-of-the-parts**: Value divisions separately and add

When methods converge, confidence increases. When they diverge, it prompts investigation into why.

## 6.13 Key Takeaways

This chapter introduced Discounted Cash Flow analysis as a framework for intrinsic valuation. The main insights are:

1. **Free Cash Flow is the foundation**: FCF represents cash available to all investors after operating expenses, taxes, and investments. It differs from net income by excluding non-cash items and including investment needs.
2. **Ratio-based forecasting links components to revenue**: By expressing FCF components as percentages of revenue, we can systematically forecast cash flows based on revenue growth assumptions and operating ratio projections.
3. **Terminal value captures long-term value**: The perpetuity growth model assumes FCF grows at a constant rate forever. The perpetual growth rate should not exceed long-term economic growth.
4. **WACC is the appropriate discount rate**: The Weighted Average Cost of Capital reflects the blended cost of debt and equity financing, adjusted for the tax shield on interest.
5. **DCF produces enterprise value**: To derive equity value, subtract debt and add non-operating assets. Dividing by shares outstanding yields an implied share price.
6. **Sensitivity analysis is essential**: Given input uncertainty, presenting a range of values based on different assumptions is more honest than a single point estimate.
7. **DCF complements other methods**: No single valuation method is definitive. Cross-checking DCF with market multiples and transaction comparables provides a more complete picture.

The true value of DCF analysis lies not in producing a precise number but in forcing rigorous thinking about what drives company value. The process of building a DCF model (i.e., forecasting growth, projecting margins, estimating risk) develops deep understanding of the business being valued.

# 7 Accessing and Managing VN Financial Data

This chapter provides a guide to organizing, accessing, and managing financial data specifically tailored for the Vietnamese market. While global financial databases such as CRSP and Compustat serve as standard resources for developed markets, emerging markets like Vietnam require a different approach due to unique data sources, market structures, and regulatory environments. Understanding these nuances is essential for conducting rigorous empirical research on Vietnamese equities, bonds, and macroeconomic indicators.

Vietnam's financial market has experienced remarkable growth since the establishment of the Ho Chi Minh City Stock Exchange (HOSE) in 2000 and the Hanoi Stock Exchange (HNX) in 2005. Today, the market comprises over 1,600 listed companies across three trading venues: HOSE for large-cap stocks, HNX for mid-cap stocks, and UPCoM (Unlisted Public Company Market) for smaller companies transitioning to formal listing. This diversity creates both opportunities and challenges for financial researchers seeking comprehensive coverage of the Vietnamese equity universe.

The Vietnamese market presents several distinctive characteristics that researchers must account for. Foreign ownership limits (typically 49% for most sectors, with exceptions for banking and certain strategic industries), trading band restrictions (e.g., currently $\pm$ 7% for HOSE and $\pm$ 10% for HNX), and the T+2 settlement cycle all influence market microstructure and return dynamics. Additionally, the market operates in Vietnamese Dong (VND), requiring careful attention to currency effects when comparing results with international studies.

We begin by loading the essential Python packages that facilitate data acquisition and management throughout this chapter.

```
import pandas as pd
import numpy as np
import requests
from datetime import datetime, timedelta
import json
import sqlite3
```

We also define the date range for our data collection, which spans from the early days of the Vietnamese stock market to the present. This extended timeframe allows us to capture the market's evolution through various economic cycles, including the 2008 global financial crisis, the 2011-2012 domestic banking crisis, and the COVID-19 pandemic period.

```
start_date = "2000-07-28"  # HOSE establishment date
end_date = "2024-12-31"
```

---

## 7.1 Overview of Vietnamese Financial Data Sources

Before diving into the technical implementation, it is valuable to understand the landscape of financial data providers serving the Vietnamese market. Unlike developed markets where a few dominant providers (Bloomberg, Refinitiv, FactSet) offer comprehensive coverage, Vietnamese financial data has historically been fragmented across multiple sources, each with distinct strengths and limitations.

The primary sources of Vietnamese financial data include official exchange feeds from HOSE and HNX, which provide real-time and historical trading data. The State Securities Commission of Vietnam (SSC) publishes regulatory filings, corporate announcements, and market statistics. Commercial data vendors such as FiinGroup, StoxPlus (now part of FiinGroup), and VNDirect offer curated datasets with varying levels of coverage and data quality. Additionally, the State Bank of Vietnam (SBV) and the General Statistics Office (GSO) provide macroeconomic indicators essential for asset pricing research.

For academic researchers, this fragmentation traditionally involved difficult trade-offs between cost, coverage, data quality, and ease of access. Commercial providers like FiinGroup offer clean, standardized data but require subscription fees that may be prohibitive for individual researchers and smaller institutions. Open-source alternatives provide free access but often require substantial data cleaning and validation efforts. Manually collecting data from government websites is time-consuming and prone to inconsistencies.

Fortunately, this landscape has improved significantly with the emergence of **Datacore** as a unified data platform for Vietnamese financial markets. In our experience working with Vietnamese financial data across multiple research projects, Datacore has proven to be the most practical solution for academic research. The platform consolidates data from multiple sources, including stock prices, corporate fundamentals, market indices, macroeconomic indicators, and alternative data, into a single, accessible interface with a well-documented API.

What distinguishes Datacore from traditional commercial providers like FiinGroup extends beyond mere data aggregation. While FiinGroup has long been the institutional incumbent, several factors make Datacore particularly attractive for rigorous empirical research:

1. **API-First Architecture**: Datacore was built from the ground up for programmatic access, making it seamlessly integrable with Python, R, and other research workflows. FiinGroup's data access, by contrast, often requires manual downloads or cumbersome Excel-based interfaces that impede reproducibility.

2. **Cost Efficiency**: Academic researchers frequently operate under budget constraints. Datacore offers competitive pricing structures that make comprehensive market coverage accessible without the substantial subscription fees associated with legacy providers.
3. **Corporate Action Handling**: One persistent challenge with Vietnamese data is accurate adjustment for stock splits, bonus shares, and rights issues. Datacore implements transparent adjustment methodologies with clear documentation, whereas legacy providers often apply adjustments inconsistently or without adequate explanation.
4. **Update Frequency**: Datacore maintains near real-time data updates with clear timestamps, enabling event study research and timely portfolio rebalancing. Traditional providers often suffer from publication lags that can compromise research requiring current data.
5. **Coverage Breadth**: Beyond standard price and fundamental data, Datacore integrates alternative data, and macroeconomic indicators into a unified schema. This eliminates the need to merge datasets from multiple sources, which is a process that introduces potential errors and consumes valuable research time.

Throughout this chapter, we leverage Datacore as our primary data source. By centralizing our data acquisition through a single platform, we benefit from consistent data formats, reliable corporate action adjustments, and comprehensive market coverage spanning HOSE, HNX, and UPCoM. The code examples that follow demonstrate how straightforward Vietnamese financial research becomes when data access friction is minimized.

The following table summarizes the key data sources for Vietnamese financial research:

Table 7.1: Vietnamese Financial Data Sources

| Data Source | Coverage | Access Type | Key Strengths | Limitations |
|---|---|---|---|---|
| **Datacore** | Prices, fundamentals, indices, macro, derivatives | API | Unified platform, programmatic access, comprehensive coverage, transparent methodology | Newer platform |
| FiinGroup | Full market coverage | Commercial | Established reputation, institutional adoption | High cost, manual access, limited API |
| HOSE/HNX websites | Official exchange data | Free (manual) | Authoritative, real-time | No API, manual collection required |

| Data Source | Coverage | Access Type | Key Strengths | Limitations |
|---|---|---|---|---|
| GSO (gso.gov.vn) | Macroeconomic indicators | Free (manual) | Official government statistics | Infrequent updates, no API |
| SBV (sbv.gov.vn) | Monetary policy, rates | Free (manual) | Central bank data | Manual download only |
| CafeF/VnExpress | News, announcements | Free | Market sentiment, events | Unstructured, requires NLP processing |

---

## 7.2 Stock Market Data

The resulting DataFrame contains essential security identifiers including the ticker symbol, company name in both Vietnamese and English, exchange listing, industry classification according to the Vietnam Standard Industrial Classification (VSIC), and various flags indicating special status such as foreign ownership restrictions or trading suspensions.

### 7.2.1 Historical Price Data

### 7.2.2 Fundamental Data and Financial Statements

Beyond price data, fundamental analysis requires access to corporate financial statements including balance sheets, income statements, and cash flow statements. Vietnamese publicly listed companies are required to publish quarterly and annual financial reports according to Vietnamese Accounting Standards (VAS), which differ in certain respects from International Financial Reporting Standards (IFRS). Understanding these differences is important when comparing Vietnamese firms with international peers or applying models developed using US or European data.

Key differences between VAS and IFRS that affect financial analysis include:

1. **Revenue recognition**: VAS allows more flexibility in timing of revenue recognition compared to IFRS 15
2. **Financial instruments**: VAS has less comprehensive guidance on fair value measurement
3. **Lease accounting**: VAS does not require operating lease capitalization as under IFRS 16
4. **Goodwill**: VAS requires amortization while IFRS requires impairment testing only

### 7.2.3 Corporate Actions and Events

Accurate treatment of corporate actions is essential for computing correct returns and maintaining data integrity. Vietnamese companies frequently engage in corporate actions including cash dividends, stock dividends (bonus shares), rights issues, and stock splits.

## 7.3 Market Indices and Benchmarks

Constructing appropriate benchmarks is fundamental to performance evaluation and factor model estimation. The Vietnamese market features several indices that serve different purposes in financial research.

Table 7.2: Vietnamese Market Indices

| Index | Exchange | Description | Use Case |
|---|---|---|---|
| VN-Index | HOSE | All HOSE-listed stocks | Broad market benchmark |
| VN30-Index | HOSE | 30 largest, most liquid | Investable benchmark |
| HNX-Index | HNX | All HNX-listed stocks | Mid-cap benchmark |
| HNX30-Index | HNX | 30 largest HNX stocks | HNX large-cap |
| VNAllShare | Combined | HOSE + HNX | Total market |
| VN100 | Combined | Top 100 stocks | Large/mid-cap |

The VN-Index, which tracks all stocks listed on HOSE, is the most widely followed benchmark and serves as the primary gauge of overall market performance. The HNX-Index covers stocks on the Hanoi exchange, while the VN30-Index tracks the thirty largest and most liquid stocks on HOSE.

For asset pricing research, the VN30-Index is particularly valuable as it represents the investable universe for institutional investors and serves as the underlying for Vietnam's most liquid derivatives contracts. The constituent stocks are reviewed semi-annually based on market capitalization, liquidity, and free-float requirements.

```
# Retrieve VN-Index historical data
```

### 7.3.1 Index Constituent Data

For factor model construction and portfolio analysis, access to index constituent lists and their weights is essential. While official constituent data requires subscription to exchange data feeds, we can approximate index membership using market capitalization and liquidity filters.

## 7.4 Macroeconomic Data from Vietnamese Sources

Asset pricing models often incorporate macroeconomic variables as predictors of expected returns or as state variables in conditional models. For the Vietnamese market, relevant macroeconomic data comes primarily from two sources: the General Statistics Office (GSO) and the State Bank of Vietnam (SBV).

### 7.4.1 Key Macroeconomic Indicators

The following macroeconomic variables are particularly relevant for Vietnamese financial research:

1. **Consumer Price Index (CPI)**: Essential for computing real returns and inflation-adjusted valuations. Vietnam experienced periods of high inflation, particularly during 2008 and 2011 when annual CPI exceeded 20%.
2. **Industrial Production Index (IPI)**: Proxy for economic activity and business cycle conditions.
3. **Money Supply (M2)**: Indicator of monetary policy stance and liquidity conditions.
4. **Credit Growth**: Bank lending growth, a key driver of economic activity in Vietnam's bank-dominated financial system.
5. **USD/VND Exchange Rate**: Critical for international investors and companies with foreign currency exposure.
6. **Foreign Direct Investment (FDI)**: Indicator of international capital flows and economic confidence.
7. **Trade Balance**: Export and import dynamics affecting corporate earnings.

Unfortunately, unlike the US Federal Reserve's FRED database, Vietnamese macroeconomic data is not available through standardized APIs. Researchers must typically download data manually from GSO and SBV websites or use web scraping techniques.

```
# Structure for Vietnamese macroeconomic data
```

### 7.4.2 Risk-Free Rate Approximation

Determining an appropriate risk-free rate for Vietnam presents challenges not encountered in developed markets. Unlike the US Treasury market, Vietnam's government bond market is relatively illiquid with limited secondary trading. Several alternatives exist:

1. **SBV Refinancing Rate**: The policy rate set by the State Bank of Vietnam. Not directly investable but reflects monetary policy stance.
2. **Government Bond Yields**: One-year or longer-term government bond yields from auction results. More investable but less liquid than US Treasuries.
3. **Interbank Rates**: Overnight or term interbank lending rates. Reflect short-term funding costs but include credit risk.
4. **Adjusted US Rate**: US Treasury rate plus expected VND depreciation, following uncovered interest rate parity.

```python
def calculate_risk_free_rate(macro_data, method="refinancing"):
    """
    Calculate risk-free rate proxy for Vietnamese market.

    Parameters
    ----------
    macro_data : pd.DataFrame
        DataFrame with macroeconomic data
    method : str
        Method for risk-free rate: 'refinancing', 'bond', or 'adjusted_us'

    Returns
    -------
    pd.DataFrame
        DataFrame with date and monthly risk-free rate
    """
    if method == "refinancing":
        # Use SBV refinancing rate, convert annual to monthly
        rf = macro_data[["date", "refinancing_rate"]].copy()
        rf["rf_monthly"] = rf["refinancing_rate"] / 12 / 100

    elif method == "adjusted_us":
        # US rate + expected VND depreciation
        # Requires additional data on US rates and exchange rate expectations
        pass
```

```
    return rf[["date", "rf_monthly"]]
```

## 7.5 Setting Up a Database for Vietnamese Financial Data

Managing financial data across multiple sources and formats requires a systematic approach to data storage. We recommend using SQLite as the primary database engine for several reasons: it requires no server setup, stores the entire database in a single portable file, supports standard SQL queries, and integrates seamlessly with Python through the built-in sqlite3 module.

### 7.5.1 Database Schema Design

Our database schema is designed to support efficient queries for common research tasks while maintaining data integrity. We create separate tables for different data types with appropriate relationships.

```
import os
import sqlite3

# Create data directory if it doesn't exist
if not os.path.exists("data"):
    os.makedirs("data")

# Initialize SQLite database connection
tidy_finance_python = sqlite3.connect(
    "data/tidy_finance_python.sqlite"
)
```

### 7.5.2 Storing Data

With the database schema established, we can store our collected data using pandas' to_sql() method.

```
# Store stock listing data
common_stocks.to_sql(
    name="stock_master",
    con=tidy_finance_python,
    if_exists="replace",
    index=False
```

```
)

# Store stock price data
stock_prices.to_sql(
    name="stock_prices_daily",
    con=tidy_finance_python,
    if_exists="replace",
    index=False
)

# Store market indices
vn_index.to_sql(
    name="market_indices",
    con=tidy_finance_python,
    if_exists="replace",
    index=False
)

# Store factor returns
factors_vietnam.to_sql(
    name="factors_monthly",
    con=tidy_finance_python,
    if_exists="replace",
    index=False
)
```

## 7.6 Querying and Updating the Database

Once data is stored in the database, retrieval is straightforward using SQL queries. The pandas read_sql_query() function executes a SQL statement and returns the results as a DataFrame.

```
# Query stock prices for specific symbols and date range
query = """
SELECT date, symbol, close, volume
FROM stock_prices_daily
WHERE symbol IN ('VNM', 'VIC', 'FPT', 'VHM', 'VCB')
  AND date >= '2020-01-01'
ORDER BY symbol, date
"""

selected_stocks = pd.read_sql_query(
```

```
    sql=query,
    con=tidy_finance_python,
    parse_dates=["date"]
)

# Query factor data merged with market returns
query_factors = """
SELECT f.date, f.mkt_rf, f.smb, f.hml, f.rf,
       m.cpi_yoy, m.credit_growth
FROM factors_monthly f
LEFT JOIN macro_monthly m ON f.date = m.date
WHERE f.date >= '2015-01-01'
ORDER BY f.date
"""

factor_data = pd.read_sql_query(
    sql=query_factors,
    con=tidy_finance_python,
    parse_dates=["date"]
)
```

### 7.6.1 Database Maintenance

Regular database maintenance ensures optimal performance and data integrity.

```
# Optimize database
tidy_finance_python.execute("VACUUM")

# Check database integrity
integrity_check = pd.read_sql_query(
    "PRAGMA integrity_check",
    tidy_finance_python
)
print(f"Integrity check: {integrity_check.iloc[0, 0]}")

# Get database statistics
table_stats = pd.read_sql_query("""
    SELECT name,
           (SELECT COUNT(*) FROM stock_prices_daily) as price_rows,
           (SELECT COUNT(*) FROM stock_master) as stock_count,
           (SELECT COUNT(*) FROM factors_monthly) as factor_months
```

```
    FROM sqlite_master
    WHERE type='table' AND name='stock_master'
""", tidy_finance_python)

print(table_stats)

# Close connection when done
tidy_finance_python.close()
```

## 7.7 Alternative Data Sources for Vietnamese Markets

Beyond traditional price and fundamental data, researchers increasingly incorporate alternative data sources to gain unique insights into market dynamics.

### 7.7.1 Foreign Investor Flow Data

Foreign investor flow data is particularly valuable given the significant role of foreign capital in Vietnamese equity markets. The State Securities Commission publishes daily foreign ownership statistics by security.

### 7.7.2 News and Sentiment Data

Media sentiment from Vietnamese financial news sources offers another research avenue. Major outlets such as CafeF, VnExpress Finance, and Vietstock publish real-time news that can be analyzed for market sentiment.

## 7.8 Key Takeaways

1. **Market Structure Understanding**: The Vietnamese financial market operates across three exchanges (HOSE, HNX, UPCoM) with distinct characteristics including foreign ownership limits, trading band restrictions, and a T+2 settlement cycle. Researchers must account for these institutional features in empirical analysis.
2. **Macroeconomic Data Challenges**: Unlike developed markets with standardized APIs (e.g., FRED), Vietnamese macroeconomic data requires manual collection from government sources (GSO, SBV). Researchers should plan for this additional data gathering effort and implement systematic data management practices.

3. **Database-Centric Workflow**: SQLite provides an efficient and portable database solution for managing Vietnamese financial data across research projects. The structured database approach enables reproducible research workflows, efficient queries, and easy data sharing among collaborators.

4. **Data Quality Imperative**: Data quality validation is especially important for emerging market data. Implementing systematic checks for missing values, extreme returns, duplicate entries, and cross-source validation helps ensure research reliability and reproducibility.

5. **Alternative Data Opportunities**: Foreign investor flows, corporate announcements, and media sentiment provide unique research opportunities in the Vietnamese market that can complement traditional price and fundamental analysis. These data sources can reveal insights about market dynamics not captured in standard datasets.

6. **Continuous Maintenance**: Financial databases require ongoing maintenance including incremental updates, integrity checks, and optimization. Establishing systematic update procedures ensures data currency and database performance over time.

# 8 Datacore Data

This chapter introduces Datacore, Vietnam's data platform for academic, corporate, and government research. Datacore provides comprehensive financial and economic datasets, including historical trading data, company fundamentals, and macroeconomic indicators essential for reproducible finance research. We use Datacore as the primary data source throughout this book.

## 8.1 Data Access Options

Readers can access the data used in this book through several channels:

1. **Institutional subscription**: Many universities and research institutions subscribe to Datacore. Check with your library or research office for access credentials. If your institution does not yet have a subscription, consider requesting one through your library's acquisition process—Datacore offers institutional pricing for academic use.
2. **Demo datasets**: Datacore provides demo datasets that allow you to run the code examples in this book with sample data.

## 8.2 Chapter Overview

The chapter is organized as follows. We first establish the connection to Datacore's cloud storage infrastructure. Then, we download and prepare company fundamentals data, including balance sheet items, income statement variables, and derived metrics essential for asset pricing research. Next, we retrieve and process stock price data, computing returns, market capitalizations, and excess returns. We conclude by merging these datasets and providing descriptive statistics that characterize the Vietnamese equity market.

## 8.3 Setting Up the Environment

We begin by loading the Python packages used throughout this chapter. The core packages include `pandas` for data manipulation, `numpy` for numerical operations, and `sqlite3` for local database management. We also import visualization libraries for creating publication-quality figures.

```
import pandas as pd
import numpy as np
import sqlite3
from datetime import datetime
from io import BytesIO

from plotnine import *
from mizani.formatters import comma_format, percent_format
```

We establish a connection to our local SQLite database, which serves as the central repository for all processed data. This database was introduced in the previous chapter and will store the cleaned datasets for use in subsequent analyses.

```
tidy_finance = sqlite3.connect(database="data/tidy_finance_python.sqlite")
```

We define the date range for our data collection. The Vietnamese stock market began operations in July 2000 with the establishment of the Ho Chi Minh City Stock Exchange (HOSE), so our sample period starts from 2000 and extends through the end of 2024.

```
start_date = "2000-01-01"
end_date = "2024-12-31"
```

## 8.4 Connecting to Datacore

Datacore delivers data through a cloud-based object storage system built on MinIO, an S3-compatible storage infrastructure. This architecture enables efficient, programmatic access to large datasets without the limitations of traditional database connections. To access the data, you need credentials provided by Datacore upon subscription: an endpoint URL, access key, and secret key.

The following class establishes the connection to Datacore's storage system. The credentials are stored as environment variables for security, following best practices for credential management in research computing environments.

```python
import os
import boto3
from botocore.client import Config

class DatacoreConnection:
    """
    Connection handler for Datacore's MinIO-based storage system.

    This class manages authentication and provides methods for
    accessing financial datasets stored in Datacore's cloud infrastructure.

    Attributes
    ----------
    s3 : boto3.client
        S3-compatible client for interacting with Datacore storage
    """

    def __init__(self):
        """Initialize connection using environment variables."""
        self.MINIO_ENDPOINT = os.environ["MINIO_ENDPOINT"]
        self.MINIO_ACCESS_KEY = os.environ["MINIO_ACCESS_KEY"]
        self.MINIO_SECRET_KEY = os.environ["MINIO_SECRET_KEY"]
        self.REGION = os.getenv("MINIO_REGION", "us-east-1")

        self.s3 = boto3.client(
            "s3",
            endpoint_url=self.MINIO_ENDPOINT,
            aws_access_key_id=self.MINIO_ACCESS_KEY,
            aws_secret_access_key=self.MINIO_SECRET_KEY,
            region_name=self.REGION,
            config=Config(signature_version="s3v4"),
        )

    def test_connection(self):
        """Verify connection by listing available buckets."""
        response = self.s3.list_buckets()
        print("Connected successfully. Available buckets:")
        for bucket in response.get("Buckets", []):
            print(f"  - {bucket['Name']}")

    def list_objects(self, bucket_name, prefix=""):
        """List objects in a bucket with optional prefix filter."""
```

```
        response = self.s3.list_objects_v2(
            Bucket=bucket_name,
            Prefix=prefix
        )
        return [obj["Key"] for obj in response.get("Contents", [])]

    def read_excel(self, bucket_name, key):
        """Read an Excel file from Datacore storage."""
        obj = self.s3.get_object(Bucket=bucket_name, Key=key)
        return pd.read_excel(BytesIO(obj["Body"].read()))

    def read_csv(self, bucket_name, key, **kwargs):
        """Read a CSV file from Datacore storage."""
        obj = self.s3.get_object(Bucket=bucket_name, Key=key)
        return pd.read_csv(BytesIO(obj["Body"].read()), **kwargs)
```

With the connection class defined, we can establish a connection and verify access to Datacore's data repositories.

```
# Initialize connection
conn = DatacoreConnection()
conn.test_connection()

# Get bucket name from environment
bucket_name = os.environ["MINIO_BUCKET"]
```

```
Connected successfully. Available buckets:
  - dsteam-data
  - rawbctc
```

## 8.5 Company Fundamentals Data

Firm accounting data are essential for portfolio analyses, factor construction, and valuation studies. Datacore hosts comprehensive fundamentals data for Vietnamese listed companies, including annual and quarterly financial statements prepared according to Vietnamese Accounting Standards (VAS).

### 8.5.1 Understanding Vietnamese Financial Statements

Before processing the data, it is important to understand the structure of Vietnamese financial reports. Vietnamese companies follow VAS, which shares similarities with International Financial Reporting Standards (IFRS) but has notable differences:

1. **Fiscal Year**: Most Vietnamese companies use a calendar fiscal year ending December 31, though some companies (particularly in retail and agriculture) use different fiscal year-ends.
2. **Reporting Frequency**: Listed companies must publish quarterly financial statements within 20 days of quarter-end and annual audited statements within 90 days of fiscal year-end.
3. **Industry-Specific Formats**: Companies in banking, insurance, and securities sectors follow specialized reporting formats that differ from the standard industrial format.
4. **Currency**: All figures are reported in Vietnamese Dong (VND). Given the large nominal values (millions to trillions of VND), we often scale figures to millions or billions for readability.

### 8.5.2 Downloading Fundamentals Data

Datacore organizes fundamentals data in Excel files partitioned by time period for efficient access. We download and concatenate these files to create a comprehensive dataset spanning our sample period.

```
# Define paths to fundamentals data files
fundamentals_paths = [
    "fundamental_annual_1767674486317/fundamental_annual_1.xlsx",
    "fundamental_annual_1767674486317/fundamental_annual_2.xlsx",
    "fundamental_annual_1767674486317/fundamental_annual_3.xlsx",
]

# Download and combine all files
fundamentals_list = []
for path in fundamentals_paths:
    df_temp = conn.read_excel(bucket_name, path)
    fundamentals_list.append(df_temp)
    print(f"Downloaded: {path} ({len(df_temp):,} rows)")

df_fundamentals_raw = pd.concat(fundamentals_list, ignore_index=True)
print(f"\nTotal observations: {len(df_fundamentals_raw):,}")
```

```
Downloaded: fundamental_annual_1767674486317/fundamental_annual_1.xlsx (10,000 rows)

Downloaded: fundamental_annual_1767674486317/fundamental_annual_2.xlsx (10,000 rows)

Downloaded: fundamental_annual_1767674486317/fundamental_annual_3.xlsx (2,821 rows)

Total observations: 22,821
```

### 8.5.3 Cleaning and Standardizing Fundamentals

The raw fundamentals data requires several cleaning steps to ensure consistency and usability. We standardize variable names, handle missing values, and create derived variables commonly used in asset pricing research.

```python
def clean_fundamentals(df):
    """
    Clean and standardize company fundamentals data.

    Parameters
    ----------
    df : pd.DataFrame
        Raw fundamentals data from Datacore

    Returns
    -------
    pd.DataFrame
        Cleaned fundamentals with standardized column names
    """
    df = df.copy()

    # Standardize identifiers
    df["symbol"] = df["symbol"].astype(str).str.upper().str.strip()
    df["year"] = pd.to_numeric(df["year"], errors="coerce").astype("Int64")

    # Drop rows with missing identifiers
    df = df.dropna(subset=["symbol", "year"])

    # Define columns that should be numeric
    numeric_columns = [
        "total_asset", "total_equity", "total_liabilities",
        "total_current_asset", "total_current_liabilities",
```

```
        "is_net_revenue", "is_cogs", "is_manage_expense",
        "is_interest_expense", "is_eat", "is_net_business_profit",
        "na_tax_deferred", "nl_tax_deferred", "e_preferred_stock",
        "capex", "total_cfo", "ca_cce", "ca_total_inventory",
        "ca_acc_receiv", "cfo_interest_expense", "basic_eps",
        "is_shareholders_eat", "cl_loan", "cl_finlease",
        "cl_due_long_debt", "nl_loan", "nl_finlease",
        "is_cos_of_sales", "e_equity"
    ]

    for col in numeric_columns:
        if col in df.columns:
            df[col] = pd.to_numeric(df[col], errors="coerce")

    # Handle duplicates: keep row with most non-missing values
    df["_completeness"] = df.notna().sum(axis=1)
    df = (df
        .sort_values(["symbol", "year", "_completeness"])
        .drop_duplicates(subset=["symbol", "year"], keep="last")
        .drop(columns="_completeness")
        .reset_index(drop=True)
    )

    return df

df_fundamentals = clean_fundamentals(df_fundamentals_raw)
print(f"After cleaning: {len(df_fundamentals):,} firm-year observations")
print(f"Unique firms: {df_fundamentals['symbol'].nunique():,}")
```

```
After cleaning: 21,232 firm-year observations
Unique firms: 1,554
```

### 8.5.4 Creating Standardized Variables

To facilitate comparison with international studies and ensure compatibility with standard asset pricing methodologies, we create variables following conventions established in the academic literature. We map Vietnamese financial statement items to their Compustat equivalents where possible.

```python
def create_standard_variables(df):
    """
    Create standardized financial variables for asset pricing research.

    This function maps Vietnamese financial statement items to standard
    variable names used in the academic finance literature, following
    conventions from Fama and French (1992, 1993, 2015).

    Parameters
    ----------
    df : pd.DataFrame
        Cleaned fundamentals data

    Returns
    -------
    pd.DataFrame
        Fundamentals with standardized variables added
    """
    df = df.copy()

    # Fiscal date (assume December year-end)
    df["datadate"] = pd.to_datetime(df["year"].astype(str) + "-12-31")

    # === Balance Sheet Items ===
    df["at"] = df["total_asset"]                    # Total assets
    df["lt"] = df["total_liabilities"]              # Total liabilities
    df["seq"] = df["total_equity"]                  # Stockholders' equity
    df["act"] = df["total_current_asset"]           # Current assets
    df["lct"] = df["total_current_liabilities"]     # Current liabilities

    # Common equity (fallback to total equity if not available)
    df["ceq"] = df.get("e_equity", df["seq"])

    # === Deferred Taxes ===
    df["txditc"] = df.get("na_tax_deferred", 0).fillna(0)  # Deferred tax assets
    df["txdb"] = df.get("nl_tax_deferred", 0).fillna(0)    # Deferred tax liab.
    df["itcb"] = 0  # Investment tax credit (rare in Vietnam)

    # === Preferred Stock ===
    pref = df.get("e_preferred_stock", 0)
    if isinstance(pref, pd.Series):
        pref = pref.fillna(0)
```

```
    df["pstk"] = pref
    df["pstkrv"] = pref  # Redemption value
    df["pstkl"] = pref   # Liquidating value

    # === Income Statement Items ===
    df["sale"] = df["is_net_revenue"]                        # Net sales/revenue
    df["cogs"] = df.get("is_cogs", 0).fillna(0)              # Cost of goods sold
    df["xsga"] = df.get("is_manage_expense", 0).fillna(0)    # SG&A expenses
    df["xint"] = df.get("is_interest_expense", 0).fillna(0)  # Interest expense
    df["ni"] = df.get("is_eat", np.nan)                      # Net income
    df["oibdp"] = df.get("is_net_business_profit", np.nan)   # Operating income

    # === Cash Flow Items ===
    df["oancf"] = df.get("total_cfo", np.nan)  # Operating cash flow
    df["capx"] = df.get("capex", np.nan)       # Capital expenditures

    return df

df_fundamentals = create_standard_variables(df_fundamentals)
```

### 8.5.5 Computing Book Equity and Profitability

Book equity is a crucial variable for value investing strategies and the construction of HML (High Minus Low) factor portfolios. We follow the definition from Kenneth French's data library, which accounts for deferred taxes and preferred stock.

```
def compute_book_equity(df):
    """
    Compute book equity following Fama-French conventions.

    Book equity = Stockholders' equity
                  + Deferred taxes and investment tax credit
                  - Preferred stock

    Negative or zero book equity is set to missing, as book-to-market
    ratios are undefined for such firms.

    Parameters
    ----------
    df : pd.DataFrame
        Fundamentals with standardized variables
```

```
    Returns
    -------
    pd.DataFrame
        Fundamentals with book equity (be) added
    """
    df = df.copy()

    # Primary measure: stockholders' equity
    # Fallback 1: common equity + preferred stock
    # Fallback 2: total assets - total liabilities
    seq_measure = (df["seq"]
        .combine_first(df["ceq"] + df["pstk"])
        .combine_first(df["at"] - df["lt"])
    )

    # Add deferred taxes
    deferred_taxes = (df["txditc"]
        .combine_first(df["txdb"] + df["itcb"])
        .fillna(0)
    )

    # Subtract preferred stock (use redemption value as primary)
    preferred = (df["pstkrv"]
        .combine_first(df["pstkl"])
        .combine_first(df["pstk"])
        .fillna(0)
    )

    # Book equity calculation
    df["be"] = seq_measure + deferred_taxes - preferred

    # Set non-positive book equity to missing
    df["be"] = df["be"].where(df["be"] > 0, np.nan)

    return df

df_fundamentals = compute_book_equity(df_fundamentals)

# Summary statistics for book equity
print("Book Equity Summary Statistics (in million VND):")
print(df_fundamentals["be"].describe().round(2))
```

```
Book Equity Summary Statistics (in million VND):
count    2.023500e+04
mean     1.031884e+12
std      4.705269e+12
min      4.404402e+07
25%      7.267610e+10
50%      1.803885e+11
75%      5.304653e+11
max      1.836314e+14
Name: be, dtype: float64
```

Operating profitability, introduced by Eugene F. Fama and French (2015), measures a firm's profits relative to its book equity. Firms with higher operating profitability tend to have higher expected returns.

```python
def compute_profitability(df):
    """
    Compute operating profitability following Fama-French (2015).

    Operating profitability = (Revenue - COGS - SG&A - Interest) / Book Equity

    Parameters
    ----------
    df : pd.DataFrame
        Fundamentals with book equity computed

    Returns
    -------
    pd.DataFrame
        Fundamentals with operating profitability (op) added
    """
    df = df.copy()

    # Operating profit before taxes
    operating_profit = (
        df["sale"]
        - df["cogs"].fillna(0)
        - df["xsga"].fillna(0)
        - df["xint"].fillna(0)
    )

    # Scale by book equity
```

```
    df["op"] = operating_profit / df["be"]

    # Winsorize extreme values (outside 1st and 99th percentiles)
    lower = df["op"].quantile(0.01)
    upper = df["op"].quantile(0.99)
    df["op"] = df["op"].clip(lower=lower, upper=upper)

    return df

df_fundamentals = compute_profitability(df_fundamentals)
```

### 8.5.6 Computing Investment

Investment, measured as asset growth, captures firms' investment behavior. Eugene F. Fama and French (2015) document that firms with high asset growth (aggressive investment) tend to have lower future returns.

```
def compute_investment(df):
    """
    Compute investment (asset growth) following Fama-French (2015).

    Investment = (Total Assets_t / Total Assets_{t-1}) - 1

    Parameters
    ----------
    df : pd.DataFrame
        Fundamentals data

    Returns
    -------
    pd.DataFrame
        Fundamentals with investment (inv) added
    """
    df = df.copy()

    # Create lagged assets
    df_lag = (df[["symbol", "year", "at"]]
        .assign(year=lambda x: x["year"] + 1)
        .rename(columns={"at": "at_lag"})
    )
```

```
    # Merge lagged values
    df = df.merge(df_lag, on=["symbol", "year"], how="left")

    # Compute investment (asset growth)
    df["inv"] = df["at"] / df["at_lag"] - 1

    # Set to missing if lagged assets non-positive
    df["inv"] = df["inv"].where(df["at_lag"] > 0, np.nan)

    return df

df_fundamentals = compute_investment(df_fundamentals)
```

### 8.5.7 Computing Total Debt

In Vietnamese financial statements, total liabilities include non-interest-bearing items such as accounts payable and tax payables. For leverage analysis, we compute total interest-bearing debt by aggregating loan and lease obligations.

```
def compute_total_debt(df):
    """
    Compute total interest-bearing debt.

    Total Debt = Short-term loans + Finance leases (current)
                 + Current portion of long-term debt
                 + Long-term loans + Finance leases (non-current)

    Parameters
    ----------
    df : pd.DataFrame
        Fundamentals data

    Returns
    -------
    pd.DataFrame
        Fundamentals with total_debt added
    """
    df = df.copy()

    df["total_debt"] = (
        df.get("cl_loan", 0).fillna(0) +           # Short-term bank loans
```

```
        df.get("cl_finlease", 0).fillna(0) +       # Current finance leases
        df.get("cl_due_long_debt", 0).fillna(0) +  # Current portion LT debt
        df.get("nl_loan", 0).fillna(0) +           # Long-term bank loans
        df.get("nl_finlease", 0).fillna(0)         # Non-current finance leases
    )

    return df

df_fundamentals = compute_total_debt(df_fundamentals)
```

### 8.5.8 Applying Filters and Final Preparation

We apply standard filters to ensure data quality: requiring positive assets, non-negative sales, and presence of core variables needed for portfolio construction.

```
# Keep only observations with required variables
required_vars = ["at", "lt", "seq", "sale"]
comp_vn = df_fundamentals.dropna(subset=required_vars)

# Apply quality filters
comp_vn = comp_vn.query("at > 0")      # Positive assets
comp_vn = comp_vn.query("sale >= 0")   # Non-negative sales

# Keep last observation per firm-year (in case of restatements)
comp_vn = (comp_vn
    .sort_values("datadate")
    .groupby(["symbol", "year"])
    .tail(1)
    .reset_index(drop=True)
)

# Diagnostic summary
print(f"Final sample: {len(comp_vn):,} firm-year observations")
print(f"Unique firms: {comp_vn['symbol'].nunique():,}")
print(f"Sample period: {comp_vn['year'].min()} - {comp_vn['year'].max()}")
```

```
Final sample: 20,091 firm-year observations
Unique firms: 1,502
Sample period: 1998 - 2023
```

### 8.5.9 Storing Fundamentals Data

We store the prepared fundamentals data in our local SQLite database for use in subsequent chapters.

```
comp_vn.to_sql(
    name="comp_vn",
    con=tidy_finance,
    if_exists="replace",
    index=False
)

print("Company fundamentals saved to database.")
```

```
Company fundamentals saved to database.
```

## 8.6 Stock Price Data

Stock price data forms the foundation of return-based analyses in empirical finance. Datacore provides comprehensive historical price data for all securities traded on HOSE, HNX, and UPCoM, including adjusted prices that account for corporate actions.

### 8.6.1 Downloading Price Data

We download the historical price data from Datacore's storage system. The data includes daily observations with open, high, low, close prices, trading volume, and adjustment factors.

```
# Download historical price data
prices_raw = conn.read_csv(
    bucket_name,
    "historycal_price/dataset_historical_price.csv",
    low_memory=False
)

print(f"Downloaded {len(prices_raw):,} daily price observations")
print(f"Date range: {prices_raw['date'].min()} to {prices_raw['date'].max()}")
```

```
Downloaded 4,307,791 daily price observations
Date range: 2010-01-04 to 2025-05-12
```

### 8.6.2 Processing Price Data

We clean the price data and compute adjusted prices that account for stock splits, stock dividends, and other corporate actions.

```
def process_price_data(df):
    """
    Process raw price data from Datacore.
    """
    df = df.copy()

    # Parse dates
    df["date"] = pd.to_datetime(df["date"])

    # Standardize column names
    df = df.rename(columns={
        "open_price": "open",
        "high_price": "high",
        "low_price": "low",
        "close_price": "close",
        "vol_total": "volume"
    })

    # Compute adjusted close price
    df["adjusted_close"] = df["close"] * df["adj_ratio"]

    # Standardize symbol
    df["symbol"] = df["symbol"].astype(str).str.upper().str.strip()

    # Sort for return calculation
    df = df.sort_values(["symbol", "date"])

    # Add year and month
    df["year"] = df["date"].dt.year
    df["month"] = df["date"].dt.month

    return df

prices = process_price_data(prices_raw)
```

### 8.6.3 Computing Shares Outstanding and Market Capitalization

Market capitalization is computed as the product of price and shares outstanding. Since Datacore provides earnings per share and net income, we can infer shares outstanding from these variables.

```
def compute_shares_outstanding(fundamentals_df):
    """
    Compute shares outstanding from fundamentals.
    """
    shares = fundamentals_df.copy()
    shares["shrout"] = shares["is_shareholders_eat"] / shares["basic_eps"]
    shares = shares[["symbol", "year", "shrout"]].dropna()

    return shares

shares_outstanding = compute_shares_outstanding(df_fundamentals)
```

```
def add_market_cap(df, shares_df):
    """
    Add market capitalization to price data.
    """
    df = df.merge(shares_df, on=["symbol", "year"], how="left")

    # Compute market cap (in million VND)
    df["mktcap"] = (df["close"] * df["shrout"]) / 1_000_000

    # Set zero or negative market cap to missing
    df["mktcap"] = df["mktcap"].where(df["mktcap"] > 0, np.nan)

    return df

prices = add_market_cap(prices, shares_outstanding)
```

### 8.6.4 Computing Returns and Excess Returns

We compute returns using adjusted closing prices to ensure returns correctly reflect total shareholder returns including dividends and corporate actions.

#### 8.6.4.1 Creating Daily Dataset

1. Sequential version

```
def create_daily_dataset(df, annual_rf=0.04):
    """
    Create daily price dataset with returns and excess returns.
    """
    df = df.copy()

    # Sort by symbol and date (critical for correct return calculation)
    df = df.sort_values(["symbol", "date"]).reset_index(drop=True)

    # Remove duplicate dates within each symbol (keep last observation)
    df = df.drop_duplicates(subset=["symbol", "date"], keep="last")

    # Compute daily returns
    df["ret"] = df.groupby("symbol")["adjusted_close"].pct_change()

    # Cap extreme negative returns
    df["ret"] = df["ret"].clip(lower=-0.99)

    # Daily risk-free rate (assuming 252 trading days)
    df["risk_free"] = annual_rf / 252

    # Excess returns
    df["ret_excess"] = df["ret"] - df["risk_free"]
    df["ret_excess"] = df["ret_excess"].clip(lower=-1.0)

    # Lagged market cap
    df["mktcap_lag"] = df.groupby("symbol")["mktcap"].shift(1)

    return df

prices_daily = create_daily_dataset(prices)
```

2. Parallel version

```
from joblib import Parallel, delayed
import os

def process_daily_symbol(symbol_df, annual_rf=0.04):
```

```python
    """
    Process a single symbol's daily data.
    """
    df = symbol_df.copy()

    # Sort by date (critical for correct return calculation)
    df = df.sort_values("date").reset_index(drop=True)

    # Remove duplicate dates (keep last observation if duplicates exist)
    df = df.drop_duplicates(subset=["date"], keep="last")

    # Compute daily returns
    df["ret"] = df["adjusted_close"].pct_change()

    # Replace infinite values with NaN
    df["ret"] = df["ret"].replace([np.inf, -np.inf], np.nan)

    # Cap extreme negative returns
    df["ret"] = df["ret"].clip(lower=-0.99)

    # Daily risk-free rate
    df["risk_free"] = annual_rf / 252

    # Excess returns
    df["ret_excess"] = df["ret"] - df["risk_free"]
    df["ret_excess"] = df["ret_excess"].clip(lower=-1.0)

    # Lagged market cap
    df["mktcap_lag"] = df["mktcap"].shift(1)

    return df

def create_daily_dataset_parallel(df, annual_rf=0.04):
    """
    Create daily price dataset using parallel processing.
    """
    # Ensure data is sorted before splitting
    df = df.sort_values(["symbol", "date"])

    # Split by symbol
    symbol_groups = [group for _, group in df.groupby("symbol")]
```

```python
    n_jobs = max(1, os.cpu_count() - 1)
    print(f"Processing {len(symbol_groups):,} symbols using {n_jobs} cores...")

    results = Parallel(n_jobs=n_jobs, verbose=1)(
        delayed(process_daily_symbol)(group, annual_rf)
        for group in symbol_groups
    )

    return pd.concat(results, ignore_index=True)

prices_daily = create_daily_dataset_parallel(prices)

# Quick validation
print("\nValidation checks:")
print(f"Any duplicate (symbol, date): {prices_daily.duplicated(subset=['symbol', 'date']).sum
print(f"Sample of non-zero returns:")
print(prices_daily[prices_daily["ret"] != 0][["symbol", "date", "adjusted_close", "ret"]].hea


prices_daily.query("symbol == 'FPT'")[["symbol", "date", "adjusted_close", "ret"]].head(3)
```

```
Processing 1,837 symbols using 23 cores...


Validation checks:
Any duplicate (symbol, date): 0
Sample of non-zero returns:
   symbol       date  adjusted_close       ret
0     A32 2018-10-23       44.574418       NaN
27    A32 2018-11-29       55.072640  0.235521
30    A32 2018-12-04       48.188560 -0.125000
43    A32 2018-12-21       51.974804  0.078571
49    A32 2019-01-02       55.072640  0.059603
53    A32 2019-01-08       50.030370 -0.091557
74    A32 2019-02-13       44.289180 -0.114754
75    A32 2019-02-14       41.008500 -0.074074
78    A32 2019-02-19       36.087480 -0.120000
91    A32 2019-03-08       41.336568  0.145455
```

| | symbol | date | adjusted_close | ret |
|---|---|---|---|---|
| 1146076 | FPT | 2010-01-04 | 1170.9885 | NaN |
| 1146077 | FPT | 2010-01-05 | 1170.9885 | 0.000000 |
| 1146078 | FPT | 2010-01-06 | 1149.6978 | -0.018182 |

```
# Select columns
daily_columns = [
    "symbol", "date", "year", "month",
    "open", "high", "low", "close", "volume",
    "adjusted_close", "shrout", "mktcap", "mktcap_lag",
    "ret", "risk_free", "ret_excess"
]
prices_daily = prices_daily[daily_columns]

# Remove observations with missing essential variables
prices_daily = prices_daily.dropna(subset=["ret_excess", "mktcap", "mktcap_lag"])

print("Daily Return Summary Statistics:")
print(prices_daily["ret"].describe().round(4))
print(f"\nFinal daily sample: {len(prices_daily):,} observations")
```

```
Daily Return Summary Statistics:
count    3.462157e+06
mean     3.000000e-04
std      4.480000e-02
min     -9.900000e-01
25%     -4.900000e-03
50%      0.000000e+00
75%      4.000000e-03
max      3.250000e+01
Name: ret, dtype: float64

Final daily sample: 3,462,157 observations
```

#### 8.6.4.2 Creating Monthly Dataset

For monthly returns, we compute returns directly from month-end adjusted prices rather than compounding daily returns. This avoids compounding errors from missing days and is the standard approach in empirical finance.

1. Sequential version

```python
def create_monthly_dataset(df, annual_rf=0.04):
    """
    Create monthly price dataset with returns computed from
    month-end to month-end adjusted prices.
    """
    df = df.copy()

    # Sort by symbol and date (critical for correct return calculation)
    df = df.sort_values(["symbol", "date"]).reset_index(drop=True)

    # Remove duplicate dates within each symbol (keep last observation)
    df = df.drop_duplicates(subset=["symbol", "date"], keep="last")

    # Get month-end observations
    monthly = (df
        .groupby("symbol")
        .resample("ME", on="date")
        .agg({
            "open": "first",           # First day open
            "high": "max",             # Monthly high
            "low": "min",              # Monthly low
            "close": "last",           # Last day close
            "volume": "sum",           # Total monthly volume
            "adjusted_close": "last",  # Month-end adjusted price
            "shrout": "last",          # Month-end shares outstanding
            "mktcap": "last",          # Month-end market cap
            "year": "last",
            "month": "last"
        })
        .reset_index()
    )

    # Remove duplicate (symbol, date) after resampling (safety check)
    monthly = monthly.drop_duplicates(subset=["symbol", "date"], keep="last")

    # Sort again after resampling
    monthly = monthly.sort_values(["symbol", "date"]).reset_index(drop=True)

    # Compute monthly returns from month-end to month-end adjusted prices
    monthly["ret"] = monthly.groupby("symbol")["adjusted_close"].pct_change()

    # Cap extreme returns
```

```
    monthly["ret"] = monthly["ret"].clip(lower=-0.99)

    # Monthly risk-free rate
    monthly["risk_free"] = annual_rf / 12

    # Excess returns
    monthly["ret_excess"] = monthly["ret"] - monthly["risk_free"]
    monthly["ret_excess"] = monthly["ret_excess"].clip(lower=-1.0)

    # Lagged market cap for portfolio weighting
    monthly["mktcap_lag"] = monthly.groupby("symbol")["mktcap"].shift(1)

    return monthly

prices_monthly = create_monthly_dataset(prices)
```

2. Parallel version

```
from joblib import Parallel, delayed
import os

def process_monthly_symbol(symbol_df, annual_rf=0.04):
    """
    Process a single symbol's data to monthly frequency.
    """
    df = symbol_df.copy()

    # Sort by date (critical for correct return calculation)
    df = df.sort_values("date").reset_index(drop=True)

    # Remove duplicate dates (keep last observation if duplicates exist)
    df = df.drop_duplicates(subset=["date"], keep="last")

    # Set date as index for resampling
    df = df.set_index("date")

    # Resample to monthly
    monthly = df.resample("ME").agg({
        "symbol": "last",
        "open": "first",
        "high": "max",
        "low": "min",
```

```
        "close": "last",
        "volume": "sum",
        "adjusted_close": "last",
        "shrout": "last",
        "mktcap": "last",
        "year": "last",
        "month": "last"
    }).reset_index()

    # Remove rows where symbol is NaN (months with no trading)
    monthly = monthly.dropna(subset=["symbol"])

    # Sort by date
    monthly = monthly.sort_values("date").reset_index(drop=True)

    # Compute monthly returns
    monthly["ret"] = monthly["adjusted_close"].pct_change()

    # Replace infinite values with NaN
    monthly["ret"] = monthly["ret"].replace([np.inf, -np.inf], np.nan)

    # Cap extreme returns
    monthly["ret"] = monthly["ret"].clip(lower=-0.99)

    # Monthly risk-free rate
    monthly["risk_free"] = annual_rf / 12

    # Excess returns
    monthly["ret_excess"] = monthly["ret"] - monthly["risk_free"]
    monthly["ret_excess"] = monthly["ret_excess"].clip(lower=-1.0)

    # Lagged market cap
    monthly["mktcap_lag"] = monthly["mktcap"].shift(1)

    return monthly

def create_monthly_dataset_parallel(df, annual_rf=0.04):
    """
    Create monthly price dataset using parallel processing.
    """
    # Ensure data is sorted before splitting
    df = df.sort_values(["symbol", "date"])
```

```
    # Split by symbol
    symbol_groups = [group for _, group in df.groupby("symbol")]

    n_jobs = max(1, os.cpu_count() - 1)
    print(f"Processing {len(symbol_groups):,} symbols using {n_jobs} cores...")

    results = Parallel(n_jobs=n_jobs, verbose=1)(
        delayed(process_monthly_symbol)(group, annual_rf)
        for group in symbol_groups
    )

    return pd.concat(results, ignore_index=True)

prices_monthly = create_monthly_dataset_parallel(prices)

# Validation checks
print("\nValidation checks:")
print(f"Any duplicate (symbol, date): {prices_monthly.duplicated(subset=['symbol', 'date']).s
print(f"\nSample of non-zero returns:")
print(prices_monthly[prices_monthly["ret"] != 0][["symbol", "date", "adjusted_close", "ret"]]

prices_monthly.query("symbol == 'FPT'")[["symbol", "date", "adjusted_close", "ret"]].head(3)
```

```
Processing 1,837 symbols using 23 cores...


Validation checks:
Any duplicate (symbol, date): 0

Sample of non-zero returns:
   symbol       date  adjusted_close       ret
0     A32 2018-10-31       44.574418       NaN
1     A32 2018-11-30       55.072640  0.235521
2     A32 2018-12-31       51.974804 -0.056250
3     A32 2019-01-31       50.030370 -0.037411
4     A32 2019-02-28       36.087480 -0.278689
5     A32 2019-03-31       41.828670  0.159091
7     A32 2019-05-31       43.304976  0.035294
8     A32 2019-06-30       35.929125 -0.170323
9     A32 2019-07-31       37.525975  0.044444
10    A32 2019-08-31       38.324400  0.021277
```

| | symbol | date | adjusted_close | ret |
|---|---|---|---|---|
| 55963 | FPT | 2010-01-31 | 1092.9226 | NaN |
| 55964 | FPT | 2010-02-28 | 1107.1164 | 0.012987 |
| 55965 | FPT | 2010-03-31 | 1185.1823 | 0.070513 |

```
# Select columns (same structure as daily)
monthly_columns = [
    "symbol", "date", "year", "month",
    "open", "high", "low", "close", "volume",
    "adjusted_close", "shrout", "mktcap", "mktcap_lag",
    "ret", "risk_free", "ret_excess"
]
prices_monthly = prices_monthly[monthly_columns]

# Remove observations with missing essential variables
prices_monthly = prices_monthly.dropna(subset=["ret_excess", "mktcap", "mktcap_lag"])

print("Monthly Return Summary Statistics:")
print(prices_monthly["ret"].describe().round(4))
print(f"\nFinal monthly sample: {len(prices_monthly):,} observations")
```

```
Monthly Return Summary Statistics:
count    165499.0000
mean          0.0042
std           0.1862
min          -0.9900
25%          -0.0703
50%           0.0000
75%           0.0553
max          12.7500
Name: ret, dtype: float64

Final monthly sample: 165,499 observations
```

### 8.6.5 Storing Price Data

```
prices_daily.to_sql(
    name="prices_daily",
    con=tidy_finance,
```

```
    if_exists="replace",
    index=False
)
print("Daily price data saved to database.")

prices_monthly.to_sql(
    name="prices_monthly",
    con=tidy_finance,
    if_exists="replace",
    index=False
)
print("Monthly price data saved to database.")
```

## 8.7 Descriptive Statistics

Before proceeding to asset pricing analyses, we examine the characteristics of our sample to understand the Vietnamese equity market's evolution and composition.

### 8.7.1 Market Evolution Over Time

We first examine how the number of listed securities has grown over time.

```
securities_over_time = (prices_monthly
    .groupby("date")
    .agg(
        n_securities=("symbol", "nunique"),
        total_mktcap=("mktcap", "sum")
    )
    .reset_index()
)
```

```
securities_figure = (
    ggplot(securities_over_time, aes(x="date", y="n_securities"))
    + geom_line(color="steelblue", size=1)
    + labs(
        x="",
        y="Number of Securities",
        title="Growth of Vietnamese Stock Market"
    )
```

```
    + scale_x_datetime(date_breaks="2 years", date_labels="%Y")
    + scale_y_continuous(labels=comma_format())
    + theme_minimal()
)
securities_figure.show()
```

Figure 8.1: The figure shows the monthly number of securities in the Vietnamese stock market sample.

### 8.7.2 Market Capitalization Evolution

The aggregate market capitalization reflects the overall size and development of the Vietnamese equity market.

```
mktcap_figure = (
    ggplot(securities_over_time, aes(x="date", y="total_mktcap / 1000"))
    + geom_line(color="darkgreen", size=1)
    + labs(
        x="",
        y="Market Cap (Trillion VND)",
        title="Total Market Capitalization of Vietnamese Equities"
```

```
    )
    + scale_x_datetime(date_breaks="2 years", date_labels="%Y")
    + scale_y_continuous(labels=comma_format())
    + theme_minimal()
)
mktcap_figure.show()
```

Figure 8.2: The figure shows the total market capitalization of Vietnamese listed companies over time.

### 8.7.3 Return Distribution

Understanding the distribution of monthly returns helps identify potential data quality issues and characterize market risk.

```
return_distribution = (
    ggplot(prices_monthly, aes(x="ret_excess"))
    + geom_histogram(
        binwidth=0.02,
        fill="steelblue",
        color="white",
```

```
        alpha=0.7
    )
    + labs(
        x="Monthly Excess Return",
        y="Frequency",
        title="Distribution of Monthly Excess Returns"
    )
    + scale_x_continuous(limits=(-0.5, 0.5))
    + theme_minimal()
)
return_distribution.show()
```

Figure 8.3: Distribution of monthly excess returns for Vietnamese stocks.

### 8.7.4 Coverage of Book Equity

Book equity is essential for constructing value portfolios. We examine what fraction of our sample has book equity data available over time.

```
# Merge prices with fundamentals
coverage_data = (prices_monthly
```

```
    .assign(year=lambda x: x["date"].dt.year)
    .groupby(["symbol", "year"])
    .tail(1)
    .merge(comp_vn[["symbol", "year", "be"]],
           on=["symbol", "year"],
           how="left")
)

# Compute coverage by year
be_coverage = (coverage_data
    .groupby("year")
    .apply(lambda x: pd.Series({
        "share_with_be": x["be"].notna().mean()
    }))
    .reset_index()
)

coverage_figure = (
    ggplot(be_coverage, aes(x="year", y="share_with_be"))
    + geom_line(color="darkorange", size=1)
    + geom_point(color="darkorange", size=2)
    + labs(
        x="Year",
        y="Share with Book Equity",
        title="Coverage of Book Equity Data"
    )
    + scale_y_continuous(labels=percent_format(), limits=(0, 1))
    + theme_minimal()
)
coverage_figure.show()
```

Figure 8.4: Share of securities with available book equity data by year.

## 8.8 Merging Stock and Fundamental Data

The final step links price data with fundamental data using the stock symbol as the common identifier. This merged dataset forms the basis for constructing portfolios sorted on firm characteristics.

```
# Example: Create merged dataset for end-of-June each year
merged_data = (prices_monthly
    .query("month == 6")
    .merge(
        comp_vn[["symbol", "year", "be", "op", "inv", "at"]],
        on=["symbol", "year"],
        how="left",
        suffixes=("", "_fundamental")
    )
)

# Convert BE from VND to BILLION VND
merged_data["be"] = merged_data["be"] / 1e9
```

```python
# Compute book-to-market ratio
merged_data["bm"] = merged_data["be"] / merged_data["mktcap"]

merged_data.loc[
    (merged_data["bm"] <= 0) |
    (merged_data["bm"] > 20),
    "bm"
] = pd.NA


merged_data["bm"].describe(percentiles=[.01, .1, .5, .9, .99])

print(f"Merged observations: {len(merged_data):,}")
print(f"With book-to-market: {merged_data['bm'].notna().sum():,}")
merged_data.head(3)
merged_data.describe()
merged_data
```

```
Merged observations: 13,756
With book-to-market: 12,859
```

| | symbol | date | year | month | open | high | low | close | volume | adjusted_close | ... | m |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | A32 | 2019-06-30 | 2019.0 | 6.0 | 26.4 | 26.4 | 21.0 | 22.5 | 3700 | 35.929125 | ... | 15 |
| 1 | A32 | 2020-06-30 | 2020.0 | 6.0 | 25.0 | 26.3 | 24.5 | 26.3 | 7500 | 38.811173 | ... | 17 |
| 2 | A32 | 2021-06-30 | 2021.0 | 6.0 | 30.2 | 37.0 | 29.5 | 32.0 | 78400 | 45.363520 | ... | 21 |
| 3 | A32 | 2022-06-30 | 2022.0 | 6.0 | 30.9 | 35.5 | 25.0 | 35.3 | 15200 | 47.503210 | ... | 24 |
| 4 | A32 | 2023-06-30 | 2023.0 | 6.0 | 30.1 | 33.5 | 29.2 | 29.4 | 2400 | 35.064204 | ... | 19 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 13751 | YTC | 2019-06-30 | 2019.0 | 6.0 | 70.0 | 79.9 | 70.0 | 79.9 | 38900 | 171.451817 | ... | 24 |
| 13752 | YTC | 2020-06-30 | 2020.0 | 6.0 | 88.5 | 88.5 | 77.0 | 87.0 | 150640 | 180.966960 | ... | 26 |
| 13753 | YTC | 2021-06-30 | 2021.0 | 6.0 | 76.0 | 115.5 | 61.0 | 61.0 | 34100 | 126.884880 | ... | 18 |
| 13754 | YTC | 2022-06-30 | 2022.0 | 6.0 | 68.0 | 68.0 | 65.0 | 65.5 | 200 | 136.245240 | ... | 20 |
| 13755 | YTC | 2023-06-30 | 2023.0 | 6.0 | 59.0 | 59.0 | 59.0 | 59.0 | 49545 | 122.724720 | ... | 18 |

```python
from plotnine import *
import numpy as np

bm_plot_data = (
    merged_data[["bm"]]
      .dropna()
```

```
      .assign(bm_plot=lambda x: x["bm"].clip(upper=10))
)

(
    ggplot(bm_plot_data, aes(x="bm_plot")) +
    geom_histogram(bins=80) +
    labs(
        title="Distribution of Book to Market Ratios",
        x="Book to Market (capped at 10 for plotting)",
        y="Number of firms"
    ) +
    theme_minimal()
)
```

```
size_plot_data = (
    merged_data[["mktcap_lag"]]
      .dropna()
      .assign(log_size=lambda x: np.log(x["mktcap_lag"]))
)

(
    ggplot(size_plot_data, aes(x="log_size")) +
```

```
    geom_histogram(bins=80) +
    labs(
        title="Distribution of Log Market Capitalization",
        x="Log Market Cap",
        y="Number of firms"
    ) +
    theme_minimal()
)
```

```
scatter_data = (
    merged_data[["be", "mktcap_lag"]]
      .dropna()
      .assign(
          log_be=lambda x: np.log(x["be"]),
          log_me=lambda x: np.log(x["mktcap_lag"])
      )
)

(
    ggplot(scatter_data, aes(x="log_me", y="log_be")) +
    geom_point(alpha=0.2) +
    labs(
```

```
        title="Log Book Equity vs Log Market Equity",
        x="Log Market Cap",
        y="Log Book Equity"
    ) +
    theme_minimal()
)
```

## 8.9 Key Takeaways

1. **Datacore provides unified access** to Vietnamese financial data through a modern cloud-based infrastructure, eliminating the need to aggregate data from multiple fragmented sources.
2. **Company fundamentals** from Datacore include comprehensive balance sheet, income statement, and cash flow data prepared according to Vietnamese Accounting Standards, which we map to standard variables used in international research.
3. **Book equity computation** follows the Fama-French methodology, accounting for deferred taxes and preferred stock to ensure comparability with US-based studies.
4. **Stock price data** includes adjustment factors for corporate actions, enabling accurate return calculations over long horizons.

5. **Monthly frequency** is standard for asset pricing research, reducing noise while maintaining sufficient observations for statistical inference.
6. **Risk-free rate approximation** uses Vietnamese government bond yields as a proxy, given the absence of a standardized short-term rate series comparable to US Treasury bills.
7. **Data quality validation** through descriptive statistics and visualization helps identify potential issues before conducting formal analyses.
8. **Batch processing** enables efficient handling of large daily datasets that would otherwise exceed memory constraints.

# 9 Beta Estimation

This chapter introduces one of the most fundamental concepts in financial economics: the exposure of an individual stock to systematic market risk. According to the Capital Asset Pricing Model (CAPM) developed by Sharpe (1964), Lintner (1965), and Mossin (1966), cross-sectional variation in expected asset returns should be determined by the covariance between an asset's excess return and the excess return on the market portfolio. The regression coefficient that captures this relationship (commonly known as market beta) serves as the cornerstone of modern portfolio theory and remains widely used in practice for cost of capital estimation, performance attribution, and risk management.

In this chapter, we develop a complete framework for estimating market betas for Vietnamese stocks. We begin with a conceptual overview of the CAPM and its empirical implementation. We then demonstrate beta estimation using ordinary least squares regression, first for individual stocks and then scaled to the entire market using rolling-window estimation. To handle the computational demands of estimating betas for hundreds of stocks across many time periods, we introduce parallelization techniques that dramatically reduce processing time. Finally, we compare beta estimates derived from monthly versus daily returns and examine how betas vary across industries and over time in the Vietnamese market.

The chapter leverages several important computational concepts that extend beyond beta estimation itself. Rolling-window estimation is a technique applicable to any time-varying parameter, while parallelization provides a general solution for computationally intensive tasks that can be divided into independent subtasks.

## 9.1 Theoretical Foundation

### 9.1.1 The Capital Asset Pricing Model

The CAPM provides a theoretical framework linking expected returns to systematic risk. Under the model's assumptions—including mean-variance optimizing investors, homogeneous expectations, and frictionless markets—the expected excess return on any asset $i$ is proportional to its covariance with the market portfolio:

$$E[r_i - r_f] = \beta_i \cdot E[r_m - r_f]$$

where $r_i$ is the return on asset $i$, $r_f$ is the risk-free rate, $r_m$ is the return on the market portfolio, and $\beta_i$ is defined as:

$$\beta_i = \frac{\text{Cov}(r_i, r_m)}{\text{Var}(r_m)}$$

The market beta $\beta_i$ measures the sensitivity of asset $i$'s returns to market movements. A beta greater than one indicates the asset amplifies market movements, while a beta less than one indicates dampened sensitivity. A beta of zero would imply no systematic risk exposure, leaving only idiosyncratic risk that can be diversified away.

### 9.1.2 Empirical Implementation

In practice, we estimate beta by regressing excess stock returns on excess market returns:

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i(r_{m,t} - r_{f,t}) + \varepsilon_{i,t} \tag{9.1}$$

where $\alpha_i$ represents abnormal return (Jensen's alpha), $\beta_i$ is the market beta we seek to estimate, and $\varepsilon_{i,t}$ is the idiosyncratic error term. Under the CAPM, $\alpha_i$ should equal zero for all assets—any non-zero alpha represents a deviation from the model's predictions.

Several practical considerations affect beta estimation:

1. **Estimation Window**: Longer windows provide more observations and thus more precise estimates, but may include outdated information if betas change over time. Common choices range from 36 to 60 months for monthly data.
2. **Return Frequency**: Monthly returns reduce noise but provide fewer observations. Daily returns offer more data points but may introduce microstructure effects and non-synchronous trading biases.
3. **Market Proxy**: The theoretical market portfolio includes all assets, but in practice we use a broad equity index. For Vietnam, we use the value-weighted market return constructed from our stock universe.
4. **Minimum Observations**: Requiring a minimum number of observations (e.g., 48 out of 60 months) helps avoid unreliable estimates from sparse data.

## 9.2 Setting Up the Environment

We begin by loading the necessary Python packages. The core packages handle data manipulation, statistical modeling, and database operations. We also import parallelization tools that will be essential when scaling our estimation to the full market.

```
import pandas as pd
import numpy as np
import sqlite3
import statsmodels.formula.api as smf
from scipy.stats.mstats import winsorize

from plotnine import *
from mizani.formatters import percent_format, comma_format
from joblib import Parallel, delayed, cpu_count
from dateutil.relativedelta import relativedelta
```

We connect to our SQLite database containing the processed Vietnamese financial data from previous chapters.

```
tidy_finance = sqlite3.connect(database="data/tidy_finance_python.sqlite")
```

## 9.3 Loading and Preparing Data

### 9.3.1 Stock Returns Data

We load the monthly stock returns data prepared in the Datacore chapter. The data includes excess returns (returns minus the risk-free rate) for all Vietnamese listed stocks.

```
prices_monthly = pd.read_sql_query(
    sql="""
        SELECT symbol, date, ret_excess
        FROM prices_monthly
    """,
    con=tidy_finance,
    parse_dates={"date"}
)

# Add year for merging with fundamentals
prices_monthly["year"] = prices_monthly["date"].dt.year
```

```
print(f"Loaded {len(prices_monthly):,} monthly observations")
print(f"Covering {prices_monthly['symbol'].nunique():,} unique stocks")
print(f"Date range: {prices_monthly['date'].min():%Y-%m} to {prices_monthly['date'].max():%Y-
```

```
Loaded 209,495 monthly observations
Covering 1,837 unique stocks
Date range: 2010-01 to 2025-05
```

```
prices_daily = pd.read_sql_query(
    sql="""
        SELECT symbol, date, ret_excess
        FROM prices_daily
    """,
    con=tidy_finance,
    parse_dates={"date"}
)
```

### 9.3.2 Company Information

We load company information to enable industry-level analysis of beta estimates.

```
comp_vn = pd.read_sql_query(
    sql="""
        SELECT symbol, datadate, icb_name_vi
        FROM comp_vn
    """,
    con=tidy_finance,
    parse_dates={"datadate"}
)

# Extract year for merging
comp_vn["year"] = comp_vn["datadate"].dt.year

print(f"Company data: {comp_vn['symbol'].nunique():,} firms")
```

```
Company data: 1,502 firms
```

### 9.3.3 Market Excess Returns

For the market portfolio proxy, we use the value-weighted market excess return. If you have constructed Fama-French factors in a previous chapter, load them here. Otherwise, we can construct a simple market return from our stock data.

```
# Option 1: Load pre-computed market factor
factors_ff3_monthly = pd.read_sql_query(
    sql="SELECT date, mkt_excess FROM factors_ff3_monthly",
    con=tidy_finance,
    parse_dates={"date"}
)

# Option 2: Construct market return from stock data (if factors not available)
# This computes the value-weighted average return across all stocks
def compute_market_return(prices_df):
    """
    Compute value-weighted market return from individual stock returns.

    Parameters
    ----------
    prices_df : pd.DataFrame
        Stock returns with mktcap_lag for weighting

    Returns
    -------
    pd.DataFrame
        Monthly market excess returns
    """
    market_return = (prices_df
        .groupby("date")
        .apply(lambda x: np.average(x["ret_excess"], weights=x["mktcap_lag"]))
        .reset_index(name="mkt_excess")
    )
    return market_return
```

### 9.3.4 Merging Datasets

We combine the stock returns with market returns and company information to create our estimation dataset.

```python
# Merge stock returns with market returns
prices_monthly = prices_monthly.merge(
    factors_ff3_monthly,
    on="date",
    how="left"
)

# Merge with company information for industry classification
prices_monthly = prices_monthly.merge(
    comp_vn[["symbol", "year", "icb_name_vi"]],
    on=["symbol", "year"],
    how="left"
)

# Remove observations with missing data
prices_monthly = prices_monthly.dropna(subset=["ret_excess", "mkt_excess"])

print(f"Final estimation sample: {len(prices_monthly):,} observations")
```

```
Final estimation sample: 169,983 observations
```

### 9.3.5 Handling Outliers

Extreme returns can unduly influence regression estimates. We apply winsorization to limit the impact of outliers while preserving the general distribution of returns. Winsorization at the 1% level replaces values below the 1st percentile with the 1st percentile value, and values above the 99th percentile with the 99th percentile value.

```python
def winsorize_returns(df, columns, limits=(0.01, 0.01)):
    """
    Apply winsorization to return columns to limit outlier influence.

    Parameters
    ----------
    df : pd.DataFrame
        DataFrame containing return columns
    columns : list
        Column names to winsorize
    limits : tuple
        Lower and upper percentile limits for winsorization
```

```
    Returns
    -------
    pd.DataFrame
        DataFrame with winsorized columns
    """
    df = df.copy()
    for col in columns:
        df[col] = winsorize(df[col], limits=limits)
    return df

prices_monthly = winsorize_returns(
    prices_monthly,
    columns=["ret_excess", "mkt_excess"],
    limits=(0.01, 0.01)
)

print("Return distributions after winsorization:")
print(prices_monthly[["ret_excess", "mkt_excess"]].describe().round(4))
```

```
Return distributions after winsorization:
        ret_excess   mkt_excess
count  169983.0000  169983.0000
mean        0.0011      -0.0102
std         0.1548       0.0579
min        -0.4078      -0.1794
25%        -0.0700      -0.0384
50%        -0.0033      -0.0084
75%         0.0531       0.0219
max         0.6117       0.1221
```

## 9.4 Estimating Beta for Individual Stocks

### 9.4.1 Single Stock Example

Before scaling to the full market, we demonstrate beta estimation for a single well-known Vietnamese stock. We use Vingroup (VIC), one of the largest conglomerates in Vietnam with significant exposure to real estate, retail, and automotive sectors.

```python
# Filter data for Vingroup
vic_data = prices_monthly.query("symbol == 'VIC'").copy()

print(f"VIC observations: {len(vic_data)}")
print(f"Date range: {vic_data['date'].min():%Y-%m} to {vic_data['date'].max():%Y-%m}")
```

```
VIC observations: 150
Date range: 2011-07 to 2023-12
```

We estimate the CAPM regression using ordinary least squares via the `statsmodels` package. The formula interface provides a convenient way to specify regression models.

```python
# Estimate CAPM for Vingroup
model_vic = smf.ols(
    formula="ret_excess ~ mkt_excess",
    data=vic_data
).fit()

# Display regression results
print(model_vic.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             ret_excess   R-squared:                       0.153
Model:                            OLS   Adj. R-squared:                  0.147
Method:                 Least Squares   F-statistic:                     26.67
Date:                Wed, 04 Feb 2026   Prob (F-statistic):           7.66e-
07
Time:                        10:20:06   Log-Likelihood:                 131.96
No. Observations:                 150   AIC:                            -
259.9
Df Residuals:                     148   BIC:                            -
253.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.0075      0.008     -0.895      0.372      -0.024       0.009
mkt_excess     0.7503      0.145      5.164      0.000       0.463       1.037
==============================================================================
```

```
Omnibus:                       39.111   Durbin-Watson:                   2.039
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              107.620
Skew:                          -1.015   Prob(JB):                     4.27e-
24
Kurtosis:                       6.619   Cond. No.                         17.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The regression output provides several important pieces of information:

- **Beta (mkt_excess coefficient)**: The estimated market sensitivity. A beta above 1 indicates VIC amplifies market movements.
- **Alpha (Intercept)**: The abnormal return not explained by market exposure. Under CAPM, this should be zero.
- **R-squared**: The proportion of return variation explained by market movements.
- **t-statistics**: Test whether coefficients differ significantly from zero.

```python
# Extract key estimates
coefficients = model_vic.summary2().tables[1]

print("\nKey estimates for Vingroup (VIC):")
print(f"  Beta:  {coefficients.loc['mkt_excess', 'Coef.']:.3f}")
print(f"  Alpha: {coefficients.loc['Intercept', 'Coef.']:.4f}")
print(f"  R²:    {model_vic.rsquared:.3f}")
```

```
Key estimates for Vingroup (VIC):
  Beta:  0.750
  Alpha: -0.0075
  R²:    0.153
```

### 9.4.2 CAPM Estimation Function

We create a reusable function that estimates the CAPM and returns results in a standardized format. The function includes a minimum observations requirement to avoid unreliable estimates from sparse data.

```python
def estimate_capm(data, min_obs=48):
    """
    Estimate CAPM regression and return coefficients.

    This function regresses excess stock returns on excess market returns
    and extracts the coefficient estimates along with t-statistics.

    Parameters
    ----------
    data : pd.DataFrame
        DataFrame with 'ret_excess' and 'mkt_excess' columns
    min_obs : int
        Minimum number of observations required for estimation

    Returns
    -------
    pd.DataFrame
        DataFrame with coefficient estimates and t-statistics,
        or empty DataFrame if insufficient observations
    """
    if len(data) < min_obs:
        return pd.DataFrame()

    try:
        # Estimate OLS regression
        model = smf.ols(
            formula="ret_excess ~ mkt_excess",
            data=data
        ).fit()

        # Extract coefficient table
        coef_table = model.summary2().tables[1]

        # Format results
        results = pd.DataFrame({
            "coefficient": ["alpha", "beta"],
            "estimate": [
                coef_table.loc["Intercept", "Coef."],
                coef_table.loc["mkt_excess", "Coef."]
            ],
            "t_statistic": [
                coef_table.loc["Intercept", "t"],
```

```
                coef_table.loc["mkt_excess", "t"]
            ],
            "r_squared": model.rsquared
        })

        return results

    except Exception as e:
        # Return empty DataFrame if estimation fails
        return pd.DataFrame()
```

## 9.5 Rolling-Window Estimation

### 9.5.1 Motivation for Rolling Windows

Stock betas are not constant over time. A company's business mix, leverage, and operating environment evolve, causing its systematic risk exposure to change. To capture this time variation, we use rolling-window estimation: at each point in time, we estimate beta using only data from a fixed lookback period (e.g., the past 60 months).

Rolling-window estimation involves a trade-off:

- **Longer windows** provide more observations and thus more precise estimates, but may include stale information.
- **Shorter windows** are more responsive to changes but produce noisier estimates.

A common choice in academic research is 60 months (5 years) of monthly data, requiring at least 48 valid observations for estimation.

### 9.5.2 Rolling Window Implementation

The following function implements rolling-window CAPM estimation. For each month in the sample, it looks back over the specified window and estimates beta using all available data within that window.

```
def roll_capm_estimation(data, look_back=60, min_obs=48):
    """
    Perform rolling-window CAPM estimation.

    This function slides a window across time, estimating the CAPM
    regression at each point using the most recent 'look_back' months
```

```
    of data.

    Parameters
    ----------
    data : pd.DataFrame
        DataFrame with 'date', 'ret_excess', and 'mkt_excess' columns
    look_back : int
        Number of months in the estimation window
    min_obs : int
        Minimum observations required within each window

    Returns
    -------
    pd.DataFrame
        Time series of coefficient estimates with dates
    """
    # Ensure data is sorted by date
    data = data.sort_values("date").copy()

    # Get unique dates
    dates = data["date"].drop_duplicates().sort_values()

    # Container for results
    results = []

    # Slide window across dates
    for i in range(look_back - 1, len(dates)):
        # Define window boundaries
        end_date = dates.iloc[i]
        start_date = end_date - relativedelta(months=look_back - 1)

        # Extract data within window
        window_data = data.query("date >= @start_date and date <= @end_date")

        # Estimate CAPM for this window
        window_results = estimate_capm(window_data, min_obs=min_obs)

        if not window_results.empty:
            window_results["date"] = end_date
            results.append(window_results)

    # Combine all results
```

```
    if results:
        return pd.concat(results, ignore_index=True)
    else:
        return pd.DataFrame()
```

### 9.5.3 Example: Rolling Betas for Selected Stocks

We demonstrate rolling-window estimation for several well-known Vietnamese stocks spanning different industries.

```
# Define example stocks
examples = pd.DataFrame({
    "symbol": ["FPT", "VNM", "VIC", "HPG", "VCB"],
    "company": [
        "FPT Corporation",      # Technology
        "Vinamilk",             # Consumer goods
        "Vingroup",             # Real estate/conglomerate
        "Hoa Phat Group",       # Steel/materials
        "Vietcombank"           # Banking
    ]
})

# Check data availability for each example
data_availability = (prices_monthly
    .query("symbol in @examples['symbol']")
    .groupby("symbol")
    .agg(
        n_obs=("date", "count"),
        first_date=("date", "min"),
        last_date=("date", "max")
    )
    .reset_index()
)

print("Data availability for example stocks:")
print(data_availability)
```

```
Data availability for example stocks:
  symbol  n_obs first_date  last_date
0    FPT    150 2011-07-31 2023-12-31
1    HPG    150 2011-07-31 2023-12-31
```

```
2    VCB    150 2011-07-31 2023-12-31
3    VIC    150 2011-07-31 2023-12-31
4    VNM    150 2011-07-31 2023-12-31
```

```
# Estimate rolling betas for example stocks
example_data = prices_monthly.query("symbol in @examples['symbol']")

capm_examples = (example_data
    .groupby("symbol", group_keys=True)
    .apply(lambda x: roll_capm_estimation(x), include_groups=False)
    .reset_index()
    .drop(columns="level_1", errors="ignore")
)

# Filter to beta estimates only
beta_examples = (capm_examples
    .query("coefficient == 'beta'")
    .merge(examples, on="symbol")
)

print(f"Rolling beta estimates: {len(beta_examples):,} observations")
```

```
Rolling beta estimates: 455 observations
```

### 9.5.4 Visualizing Rolling Betas

Figure 9.1 displays the time series of beta estimates for our example stocks. The figure reveals how systematic risk exposure evolves differently across industries.

```
rolling_beta_figure = (
    ggplot(
        beta_examples,
        aes(x="date", y="estimate", color="company")
    )
    + geom_line(size=0.8)
    + geom_hline(yintercept=1, linetype="dashed", color="gray", alpha=0.7)
    + labs(
        x="",
        y="Beta",
        color="",
        title="Rolling Beta Estimates (60-Month Window)"
```

```
    )
    + scale_x_datetime(date_breaks="2 years", date_labels="%Y")
    + theme_minimal()
    + theme(legend_position="bottom")
)
rolling_beta_figure.show()
```

Figure 9.1: Monthly rolling beta estimates for selected Vietnamese stocks using a 60-month estimation window. Different industries exhibit distinct patterns of market sensitivity over time.

Several patterns emerge from the figure:

1. **Industry differences**: Technology and banking stocks may exhibit different beta patterns than real estate or consumer goods companies.
2. **Time variation**: Betas are not constant. They respond to changes in business conditions, leverage, and market regimes.
3. **Crisis periods**: Market stress periods (e.g., 2008 financial crisis, 2020 COVID-19) often see beta estimates change as correlations across stocks increase.

## 9.6 Parallelized Estimation for the Full Market

### 9.6.1 The Computational Challenge

Estimating rolling betas for all stocks in our database is computationally intensive. With hundreds of stocks, each requiring rolling estimation across many time periods, sequential processing would take considerable time. Fortunately, beta estimation for different stocks is independent (i.e., the estimate for stock A does not depend on the estimate for stock B). This independence makes the problem ideal for parallelization.

### 9.6.2 Setting Up Parallel Processing

We use the `joblib` library to distribute computation across multiple CPU cores. The `Parallel` class manages worker processes, while `delayed` wraps function calls for deferred execution.

```
# Determine available cores (reserve one for system operations)
n_cores = max(1, cpu_count() - 1)
print(f"Available cores for parallel processing: {n_cores}")
```

```
Available cores for parallel processing: 3
```

### 9.6.3 Parallel Beta Estimation

The following code estimates rolling betas for all stocks in parallel. Each stock is processed independently by a separate worker.

```
def estimate_all_betas_parallel(data, n_cores, look_back=60, min_obs=48):
    """
    Estimate rolling betas for all stocks using parallel processing.

    Parameters
    ----------
    data : pd.DataFrame
        Full dataset with all stocks
    n_cores : int
        Number of CPU cores to use
    look_back : int
        Months in estimation window
    min_obs : int
        Minimum observations required
```

```
    Returns
    -------
    pd.DataFrame
        Beta estimates for all stocks and dates
    """
    # Group data by stock
    grouped = data.groupby("symbol", group_keys=False)

    # Define worker function
    def process_stock(name, group):
        result = roll_capm_estimation(group, look_back=look_back, min_obs=min_obs)
        if not result.empty:
            result["symbol"] = name
        return result

    # Execute in parallel
    results = Parallel(n_jobs=n_cores, verbose=1)(
        delayed(process_stock)(name, group)
        for name, group in grouped
    )

    # Combine results
    results = [r for r in results if not r.empty]
    if results:
        return pd.concat(results, ignore_index=True)
    else:
        return pd.DataFrame()
```

```
# Estimate betas for all stocks
print("Estimating rolling betas for all stocks...")
capm_monthly = estimate_all_betas_parallel(
    prices_monthly,
    n_cores=n_cores,
    look_back=60,
    min_obs=48
)

print(f"\nCompleted: {len(capm_monthly):,} coefficient estimates")
print(f"Unique stocks: {capm_monthly['symbol'].nunique():,}")
```

### 9.6.4 Storing Results

We save the CAPM estimates to our database for use in subsequent chapters.

```
capm_monthly.to_sql(
    name="capm_monthly",
    con=tidy_finance,
    if_exists="replace",
    index=False
)

print("CAPM estimates saved to database.")
```

For subsequent analysis, we load the pre-computed estimates:

```
capm_monthly = pd.read_sql_query(
    sql="SELECT * FROM capm_monthly",
    con=tidy_finance,
    parse_dates={"date"}
)

print(f"Loaded {len(capm_monthly):,} CAPM estimates")
```

```
Loaded 161,580 CAPM estimates
```

## 9.7 Beta Estimation Using Daily Returns

While monthly returns are standard in academic research, some applications benefit from higher-frequency data:

- **Shorter estimation windows**: Daily data allows meaningful estimation over shorter periods (e.g., 3 months rather than 5 years).
- **More responsive estimates**: Daily betas capture changes more quickly.
- **Event studies**: High-frequency betas are useful for analyzing market reactions to specific events.

However, daily data introduces additional challenges:

- **Microstructure noise**: Bid-ask bounce and other trading frictions add noise to returns.
- **Non-synchronous trading**: Less liquid stocks may not trade every day, biasing beta estimates downward.
- **Computational burden**: Daily data is roughly 21 times larger than monthly data.

### 9.7.1 Batch Processing for Daily Data

Given the size of daily data, we process stocks in batches to manage memory constraints. This approach loads and processes a subset of stocks, saves results, and proceeds to the next batch.

```
def compute_market_return_daily(tidy_finance):
    """
    Compute daily value-weighted market excess return from stock data.
    """
    # Load daily prices with market cap for weighting
    prices_daily_full = pd.read_sql_query(
        sql="""
            SELECT p.symbol, p.date, p.ret_excess, m.mktcap_lag
            FROM prices_daily p
            LEFT JOIN prices_monthly m ON p.symbol = m.symbol
                AND strftime('%Y-%m', p.date) = strftime('%Y-%m', m.date)
        """,
        con=tidy_finance,
        parse_dates={"date"}
    )

    # Compute value-weighted market return each day
    mkt_daily = (prices_daily_full
        .dropna(subset=["ret_excess", "mktcap_lag"])
        .groupby("date")
        .apply(lambda x: np.average(x["ret_excess"], weights=x["mktcap_lag"]))
        .reset_index(name="mkt_excess")
    )

    return mkt_daily


def roll_capm_estimation_daily(data, look_back_days=1260, min_obs=1000):
    """
    Perform rolling-window CAPM estimation using daily data.

    Parameters
    ----------
    data : pd.DataFrame
        DataFrame with 'date', 'ret_excess', and 'mkt_excess' columns
    look_back_days : int
        Number of trading days in the estimation window
    min_obs : int
```

```
        Minimum daily observations required within each window

    Returns
    -------
    pd.DataFrame
        Time series of coefficient estimates with dates
    """
    data = data.sort_values("date").copy()
    dates = data["date"].drop_duplicates().sort_values().reset_index(drop=True)

    results = []

    for i in range(look_back_days - 1, len(dates)):
        end_date = dates.iloc[i]
        start_idx = max(0, i - look_back_days + 1)
        start_date = dates.iloc[start_idx]

        window_data = data.query("date >= @start_date and date <= @end_date")
        window_results = estimate_capm(window_data, min_obs=min_obs)

        if not window_results.empty:
            window_results["date"] = end_date
            results.append(window_results)

    if results:
        return pd.concat(results, ignore_index=True)
    else:
        return pd.DataFrame()


def estimate_daily_betas_batch(symbols, tidy_finance, n_cores, batch_size=500,
                               look_back_days=1260, min_obs=1000):
    """
    Estimate rolling betas from daily data using batch processing.
    """
    # First, compute or load market return
    print("Computing daily market excess returns...")
    mkt_daily = compute_market_return_daily(tidy_finance)
    print(f"Market returns: {len(mkt_daily)} days")

    n_batches = int(np.ceil(len(symbols) / batch_size))
    all_results = []
```

```
for j in range(n_batches):
    batch_start = j * batch_size
    batch_end = min((j + 1) * batch_size, len(symbols))
    batch_symbols = symbols[batch_start:batch_end]

    symbol_list = ", ".join(f"'{s}'" for s in batch_symbols)

    query = f"""
        SELECT symbol, date, ret_excess
        FROM prices_daily
        WHERE symbol IN ({symbol_list})
    """

    prices_daily_batch = pd.read_sql_query(
        sql=query,
        con=tidy_finance,
        parse_dates={"date"}
    )

    # Merge with market excess return
    prices_daily_batch = prices_daily_batch.merge(
        mkt_daily,
        on="date",
        how="inner"
    )

    # Group by symbol and estimate betas
    grouped = prices_daily_batch.groupby("symbol", group_keys=False)

    # Parallel estimation
    batch_results = Parallel(n_jobs=n_cores)(
        delayed(lambda name, group:
            roll_capm_estimation_daily(group, look_back_days=look_back_days, min_obs=min_
            .assign(symbol=name)
        )(name, group)
        for name, group in grouped
    )

    batch_results = [r for r in batch_results if r is not None and not r.empty]

    if batch_results:
        all_results.append(pd.concat(batch_results, ignore_index=True))
```

```
        print(f"Batch {j+1}/{n_batches} complete")

    if all_results:
        return pd.concat(all_results, ignore_index=True)
    else:
        return pd.DataFrame()
```

```
symbols = prices_monthly["symbol"].unique().tolist()

capm_daily = estimate_daily_betas_batch(
    symbols=symbols,
    tidy_finance=tidy_finance,
    n_cores=n_cores,
    batch_size=500,
    look_back_days=1260,  # ~5 years of trading days
    min_obs=1000
)

print(f"Daily beta estimates: {len(capm_daily):,}")
```

```
capm_daily.to_sql(
    name="capm_daily",
    con=tidy_finance,
    if_exists="replace",
    index=False
)

print("CAPM estimates saved to database.")
```

For subsequent analysis, we load the pre-computed estimates:

```
capm_daily = pd.read_sql_query(
    sql="SELECT * FROM capm_daily",
    con=tidy_finance,
    parse_dates={"date"}
)

print(f"Loaded {len(capm_daily):,} CAPM estimates")
```

```
Loaded 3,394,490 CAPM estimates
```

## 9.8 Analyzing Beta Estimates

### 9.8.1 Extracting Beta Estimates

We extract the beta coefficient estimates from our CAPM results for analysis.

```
# Extract monthly betas
beta_monthly = (capm_monthly
    .query("coefficient == 'beta'")
    .rename(columns={"estimate": "beta"})
    [["symbol", "date", "beta"]]
    .assign(frequency="monthly")
)

# Save to database
beta_monthly.to_sql(
    name="beta_monthly",
    con=tidy_finance,
    if_exists="replace",
    index=False
)

print(f"Monthly betas: {len(beta_monthly):,} observations")
print(f"Unique stocks: {beta_monthly['symbol'].nunique():,}")
```

```
Monthly betas: 80,790 observations
Unique stocks: 1,383
```

```
# Load pre-computed betas
beta_monthly = pd.read_sql_query(
    sql="SELECT * FROM beta_monthly",
    con=tidy_finance,
    parse_dates={"date"}
)
```

### 9.8.2 Summary Statistics

We examine the distribution of beta estimates to verify their reasonableness.

```
print("Beta Summary Statistics:")
print(beta_monthly["beta"].describe().round(3))

# Additional diagnostics
print(f"\nStocks with negative average beta: {(beta_monthly.groupby('symbol')['beta'].mean()
print(f"Stocks with beta > 2: {(beta_monthly.groupby('symbol')['beta'].mean() > 2).sum()}")
```

```
Beta Summary Statistics:
count    80790.000
mean         0.501
std          0.539
min         -1.345
25%          0.130
50%          0.447
75%          0.832
max          2.678
Name: beta, dtype: float64

Stocks with negative average beta: 177
Stocks with beta > 2: 5
```

### 9.8.3 Beta Distribution Across Industries

Different industries have different exposures to systematic market risk based on their business models, operating leverage, and financial leverage. Figure 9.2 shows the distribution of firm-level average betas across Vietnamese industries.

```
# Merge betas with industry information
beta_with_industry = (beta_monthly
    .merge(
        prices_monthly[["symbol", "date", "icb_name_vi"]].drop_duplicates(),
        on=["symbol", "date"],
        how="left"
    )
    .dropna(subset=["icb_name_vi"])
)

# Compute firm-level average beta by industry
beta_by_industry = (beta_with_industry
    .groupby(["icb_name_vi", "symbol"])["beta"]
    .mean()
```

```
    .reset_index()
)

# Order industries by median beta
industry_order = (beta_by_industry
    .groupby("icb_name_vi")["beta"]
    .median()
    .sort_values()
    .index.tolist()
)

# Select top 10 industries by number of firms for clearer visualization
top_industries = (beta_by_industry
    .groupby("icb_name_vi")
    .size()
    .nlargest(10)
    .index.tolist()
)

beta_by_industry_filtered = beta_by_industry.query("icb_name_vi in @top_industries")
```

```
beta_industry_figure = (
    ggplot(
        beta_by_industry_filtered,
        aes(x="icb_name_vi", y="beta")
    )
    + geom_boxplot(fill="steelblue", alpha=0.7)
    + geom_hline(yintercept=1, linetype="dashed", color="red", alpha=0.7)
    + coord_flip()
    + labs(
        x="",
        y="Beta",
        title="Beta Distribution by Industry"
    )
    + theme_minimal()
)
beta_industry_figure.show()
```

Figure 9.2: Distribution of firm-level average betas across Vietnamese industries. Box plots show the median, interquartile range, and outliers for each industry.

### 9.8.4 Time Variation in Cross-Sectional Beta Distribution

Betas vary not only across stocks but also over time. Figure 9.3 shows how the cross-sectional distribution of betas has evolved in the Vietnamese market.

```
# Compute monthly quantiles
beta_quantiles = (beta_monthly
    .groupby("date")["beta"]
    .quantile(q=np.arange(0.1, 1.0, 0.1))
    .reset_index()
    .rename(columns={"level_1": "quantile"})
    .assign(quantile=lambda x: (x["quantile"] * 100).astype(int).astype(str) + "%")
)

beta_quantiles_figure = (
    ggplot(
        beta_quantiles,
        aes(x="date", y="beta", color="quantile")
    )
```

```
    + geom_line(alpha=0.8)
    + geom_hline(yintercept=1, linetype="dashed", color="gray")
    + labs(
        x="",
        y="Beta",
        color="Quantile",
        title="Cross-Sectional Distribution of Betas Over Time"
    )
    + scale_x_datetime(date_breaks="2 years", date_labels="%Y")
    + theme_minimal()
)
beta_quantiles_figure.show()
```

Figure 9.3: Monthly quantiles of beta estimates over time. Each line represents a decile of the cross-sectional beta distribution.

The figure reveals several interesting patterns:

1. **Level shifts**: The entire distribution of betas can shift over time, reflecting changes in market-wide correlation.
2. **Dispersion changes**: During market stress, the spread between high and low beta stocks may change as correlations move.

3. **Trends**: Some periods show trending behavior in betas, possibly reflecting structural changes in the economy.

### 9.8.5 Coverage Analysis

We verify that our estimation procedure produces reasonable coverage across the sample. Figure 9.4 shows the fraction of stocks with available beta estimates over time.

```
# Count stocks with and without betas
coverage = (prices_monthly
    .groupby("date")["symbol"]
    .nunique()
    .reset_index(name="total_stocks")
    .merge(
        beta_monthly.groupby("date")["symbol"].nunique().reset_index(name="with_beta"),
        on="date",
        how="left"
    )
    .fillna(0)
    .assign(coverage=lambda x: x["with_beta"] / x["total_stocks"])
)

coverage_figure = (
    ggplot(coverage, aes(x="date", y="coverage"))
    + geom_line(color="steelblue", size=1)
    + labs(
        x="",
        y="Share with Beta Estimate",
        title="Beta Estimation Coverage Over Time"
    )
    + scale_y_continuous(labels=percent_format(), limits=(0, 1))
    + scale_x_datetime(date_breaks="2 years", date_labels="%Y")
    + theme_minimal()
)
coverage_figure.show()
```

Figure 9.4: Share of stocks with available beta estimates over time. Coverage increases as more stocks accumulate sufficient return history.

Coverage is lower in early years because stocks need sufficient return history (at least 48 months) before their betas can be estimated. As the market matures and stocks accumulate longer histories, coverage approaches 100%.

## 9.9 Comparing Monthly and Daily Beta Estimates

When both monthly and daily beta estimates are available, we can compare them to understand how estimation frequency affects results.

```
# Combine monthly and daily estimates
beta_daily = (capm_daily
    .query("coefficient == 'beta'")
    .rename(columns={"estimate": "beta"})
    [["symbol", "date", "beta"]]
    .assign(frequency="daily")
)

beta_combined = pd.concat([beta_monthly, beta_daily], ignore_index=True)
```

```
# Filter to example stocks
beta_comparison = (beta_combined
    .merge(examples, on="symbol")
    .query("symbol in ['VIC', 'FPT']")  # Select two for clarity
)

comparison_figure = (
    ggplot(
        beta_comparison,
        aes(x="date", y="beta", color="frequency", linetype="frequency")
    )
    + geom_line(size=0.8)
    + facet_wrap("~company", ncol=1)
    + labs(
        x="",
        y="Beta",
        color="Data Frequency",
        linetype="Data Frequency",
        title="Monthly vs Daily Beta Estimates"
    )
    + scale_x_datetime(date_breaks="2 years", date_labels="%Y")
    + theme_minimal()
    + theme(legend_position="bottom")
)
comparison_figure.show()
```

Figure 9.5: Comparison of beta estimates using monthly versus daily returns for selected stocks. Daily estimates are smoother due to more observations per estimation window.

The comparison reveals that daily-based estimates are generally smoother due to the larger number of observations in each window. However, the level and trend of estimates are similar across frequencies, providing validation that both approaches capture the same underlying systematic risk exposure.

```
# Correlation between monthly and daily estimates
correlation_data = (beta_combined
    .pivot_table(index=["symbol", "date"], columns="frequency", values="beta")
    .dropna()
)

print(f"Correlation between monthly and daily betas: {correlation_data.corr().iloc[0,1]:.3f}"
```

```
Correlation between monthly and daily betas: 0.745
```

Table 9.1: Theoretical Reasons for Imperfect Correlation

| Factor | Effect |
|---|---|
| Non-synchronous trading | Daily betas can be biased downward for illiquid stocks |
| Microstructure noise | Bid-ask bounce adds noise to daily estimates |
| Different effective windows | Same calendar period but ~20x more observations for daily |
| Mean reversion speed | Daily captures faster-moving risk dynamics |

Table 9.1 shows several reasons why we might observe imperfect correlation.

## 9.10 Key Takeaways

1. **CAPM beta** measures a stock's sensitivity to systematic market risk and is fundamental to modern portfolio theory, cost of capital estimation, and risk management.
2. **Rolling-window estimation** captures time variation in betas, which reflects changes in companies' business models, leverage, and market conditions.
3. **Parallelization** dramatically reduces computation time for large-scale estimation tasks by distributing work across multiple CPU cores.
4. **Estimation choices matter**: Window length, return frequency, and minimum observation requirements all affect beta estimates. Researchers should choose parameters appropriate for their specific application.
5. **Industry patterns**: Vietnamese stocks show systematic differences in market sensitivity across industries, with cyclical sectors exhibiting higher betas than defensive sectors.
6. **Time variation**: The cross-sectional distribution of betas in Vietnam has evolved over time, with notable shifts during market stress periods.
7. **Frequency comparison**: Monthly and daily beta estimates are positively correlated but not identical. Daily estimates are smoother while monthly estimates may better capture lower-frequency variation.
8. **Data quality checks**: Coverage analysis and summary statistics help identify potential issues in estimation procedures before using results in downstream analyses.

# 10 Univariate Portfolio Sorts

In this chapter, we dive into portfolio sorts, one of the most widely used statistical methodologies in empirical asset pricing (e.g., Bali, Engle, and Murray 2016). The key application of portfolio sorts is to examine whether one or more variables can predict future excess returns. In general, the idea is to sort individual stocks into portfolios, where the stocks within each portfolio are similar with respect to a sorting variable, such as firm size. The different portfolios then represent well-diversified investments that differ in the level of the sorting variable. You can then attribute the differences in the return distribution to the impact of the sorting variable. We start by introducing univariate portfolio sorts (which sort based on only one characteristic) and tackle bivariate sorting in Value and Bivariate Sorts.

A univariate portfolio sort considers only one sorting variable $x_{i,t-1}$. Here, $i$ denotes the stock and $t-1$ indicates that the characteristic is observable by investors at time $t$. The objective is to assess the cross-sectional relation between $x_{i,t-1}$ and, typically, stock excess returns $r_{i,t}$ at time $t$ as the outcome variable. To illustrate how portfolio sorts work, we use estimates for market betas from the previous chapter as our sorting variable.

```
import pandas as pd
import numpy as np
import sqlite3
import statsmodels.api as sm

from plotnine import *
from mizani.formatters import percent_format
from regtabletotext import prettify_result
```

## 10.1 Data Preparation

We start with loading the required data from our SQLite database introduced in Accessing and Managing Financial Data and DataCore Data. In particular, we use the monthly stock price data as our asset universe. Once we form our portfolios, we use the market factor returns to compute the risk-adjusted performance (i.e., alpha). `beta` is the dataframe with market betas computed in the previous chapter.

```
tidy_finance = sqlite3.connect(database="data/tidy_finance_python.sqlite")

prices_monthly = (pd.read_sql_query(
    sql="SELECT symbol, date, ret_excess, mktcap_lag FROM prices_monthly",
    con=tidy_finance,
    parse_dates={"date"})
)

factors_ff3_monthly = pd.read_sql_query(
  sql="SELECT date, mkt_excess FROM factors_ff3_monthly",
  con=tidy_finance,
  parse_dates={"date"}
)

beta = pd.read_sql_query(
  sql="SELECT symbol, date, beta FROM beta_monthly",
  con=tidy_finance,
  parse_dates={"date"}
)
```

## 10.2 Sorting by Market Beta

Next, we merge our sorting variable with the return data. We use the one-month *lagged* betas as a sorting variable to ensure that the sorts rely only on information available when we create the portfolios. To lag stock beta by one month, we add one month to the current date and join the resulting information with our return data. This procedure ensures that month $t$ information is available in month $t + 1$. You may be tempted to simply use a call such as `prices_monthly['beta_lag'] = prices_monthly.groupby('symbol')['beta'].shift(1)` instead. This procedure, however, does not work correctly if there are implicit missing values in the time series.

```
beta_lag = (beta
  .assign(date=lambda x: x["date"]+pd.DateOffset(months=1))
  .get(["symbol", "date", "beta"])
  .rename(columns={"beta": "beta_lag"})
  .dropna()
)

data_for_sorts = (prices_monthly
  .merge(beta_lag, how="inner", on=["symbol", "date"])
```

```
  .dropna()
)
```

The first step to conduct portfolio sorts is to calculate periodic breakpoints that you can use to group the stocks into portfolios. For simplicity, we start with the median lagged market beta as the single breakpoint. We then compute the value-weighted returns for each of the two resulting portfolios, which means that the lagged market capitalization determines the weight in `np.average()`.

```
beta_portfolios = (
    data_for_sorts
    .groupby("date")
    .apply(lambda x: (
        x.assign(
            portfolio=pd.qcut(x["beta_lag"], q=[0, 0.5, 1], labels=["low", "high"]),
            date=x.name
        )
    ))
    .reset_index(drop=True)
    .groupby(["portfolio", "date"])
    .apply(lambda x: np.average(x["ret_excess"], weights=x["mktcap_lag"]))
    .reset_index(name="ret")
    .dropna(subset=['ret'])
)
beta_portfolios.head()
```

| | portfolio | date | ret |
|---|---|---|---|
| 0 | low | 2016-08-31 | -0.016507 |
| 1 | low | 2016-09-30 | -0.089155 |
| 2 | low | 2016-11-30 | -0.015080 |
| 3 | low | 2017-01-31 | 0.004113 |
| 4 | low | 2017-02-28 | 0.035101 |

## 10.3 Performance Evaluation

We can construct a long-short strategy based on the two portfolios: buy the high-beta portfolio and, at the same time, short the low-beta portfolio. Thereby, the overall position in the market is net-zero.

```
beta_longshort = (beta_portfolios
  .pivot_table(index="date", columns="portfolio", values="ret")
  .reset_index()
  .assign(long_short=lambda x: x["high"]-x["low"])
)
```

We compute the average return and the corresponding standard error to test whether the long-short portfolio yields on average positive or negative excess returns. In the asset pricing literature, one typically adjusts for autocorrelation by using Newey and West (1987) $t$-statistics to test the null hypothesis that average portfolio excess returns are equal to zero. One necessary input for Newey-West standard errors is a chosen bandwidth based on the number of lags employed for the estimation. Researchers often default to choosing a pre-specified lag length of six months (which is not a data-driven approach). We do so in the `fit()` function by indicating the `cov_type` as `HAC` and providing the maximum lag length through an additional keywords dictionary.

```
model_fit = (sm.OLS.from_formula(
    formula="long_short ~ 1",
    data=beta_longshort
  )
  .fit(cov_type="HAC", cov_kwds={"maxlags": 6})
)
prettify_result(model_fit)
```

```
OLS Model:
long_short ~ 1

Coefficients:
           Estimate  Std. Error  t-Statistic  p-Value
Intercept     0.006       0.006        1.018    0.309

Summary statistics:
- Number of observations: 53
- R-squared: 0.000, Adjusted R-squared: 0.000
- F-statistic not available
```

The results indicate that we cannot reject the null hypothesis of average returns being equal to zero. Our portfolio strategy using the median as a breakpoint does not yield any abnormal returns. Is this finding surprising if you reconsider the CAPM? It certainly is. The CAPM predicts that high-beta stocks should yield higher expected returns. Our portfolio sort implicitly mimics an investment strategy that finances high-beta stocks by shorting low-beta stocks.

Therefore, one should expect that the average excess returns yield a return that is above the risk-free rate.

## 10.4 Functional Programming for Portfolio Sorts

Now, we take portfolio sorts to the next level. We want to be able to sort stocks into an arbitrary number of portfolios. For this case, functional programming is very handy: we define a function that gives us flexibility concerning which variable to use for the sorting, denoted by `sorting_variable`. We use `np.quantile()` to compute breakpoints for `n_portfolios`. Then, we assign portfolios to stocks using the `pd.cut()` function. The output of the following function is a new column that contains the number of the portfolio to which a stock belongs.

In some applications, the variable used for the sorting might be clustered (e.g., at a lower bound of 0). Then, multiple breakpoints may be identical, leading to empty portfolios. Similarly, some portfolios might have a very small number of stocks at the beginning of the sample. Cases where the number of portfolio constituents differs substantially due to the distribution of the characteristics require careful consideration and, depending on the application, might require customized sorting approaches.

```
def assign_portfolio(data, sorting_variable, n_portfolios):
    """Assign portfolios to a bin between breakpoints."""

    breakpoints = np.quantile(
      data[sorting_variable].dropna(),
      np.linspace(0, 1, n_portfolios + 1),
      method="linear"
    )

    assigned_portfolios = pd.cut(
      data[sorting_variable],
      bins=breakpoints,
      labels=range(1, breakpoints.size),
      include_lowest=True,
      right=False
    )

    return assigned_portfolios
```

We can use the above function to sort stocks into ten portfolios each month using lagged betas and compute value-weighted returns for each portfolio. Note that we transform the portfolio column to a factor variable because it provides more convenience for the figure construction below.

```
beta_portfolios = (data_for_sorts
  .groupby("date")
  .apply(lambda x: x.assign(
      portfolio=assign_portfolio(x, "beta_lag", 10)
    ), include_groups=False
  )
  .reset_index(level="date")
  .groupby(["portfolio", "date"])
  .apply(lambda x: pd.Series({
      "ret": np.average(x["ret_excess"], weights=x["mktcap_lag"])
    }), include_groups=False
  )
  .reset_index()
  .merge(factors_ff3_monthly, how="left", on="date")
)
```

## 10.5 More Performance Evaluation

In the next step, we compute summary statistics for each beta portfolio. Namely, we compute CAPM-adjusted alphas, the beta of each beta portfolio, and average returns.

```
beta_portfolios_summary = (beta_portfolios
  .groupby("portfolio")
  .apply(lambda x: pd.Series({
      "alpha": sm.OLS.from_formula(
          formula="ret ~ 1 + mkt_excess",
          data=x
        ).fit().params["Intercept"],
      "beta": sm.OLS.from_formula(
          formula="ret ~ 1 + mkt_excess",
          data=x
        ).fit().params["mkt_excess"],
      "ret": x["ret"].mean()
    }), include_groups=False
  )
  .reset_index()
)
beta_portfolios_summary
```

| | portfolio | alpha | beta | ret |
|---|---|---|---|---|
| 0 | 1 | 0.007031 | 0.356225 | 0.003892 |
| 1 | 2 | -0.004215 | 0.227882 | -0.005509 |
| 2 | 3 | -0.010169 | 0.390216 | -0.011241 |
| 3 | 4 | -0.014205 | 0.388342 | -0.015732 |
| 4 | 5 | -0.007220 | 0.616833 | -0.009723 |
| 5 | 6 | 0.010648 | 0.976502 | 0.006577 |
| 6 | 7 | -0.010739 | 0.679145 | -0.012645 |
| 7 | 8 | -0.002757 | 0.991774 | -0.006963 |
| 8 | 9 | -0.003448 | 0.904886 | -0.007174 |
| 9 | 10 | 0.010675 | 1.376356 | 0.004944 |

Figure 10.1 illustrates the CAPM alphas of beta-sorted portfolios.

```
beta_portfolios_figure = (
  ggplot(
    beta_portfolios_summary,
    aes(x="portfolio", y="alpha", fill="portfolio")
  )
  + geom_bar(stat="identity")
  + labs(
      x="Portfolio", y="CAPM alpha", fill="Portfolio",
      title="CAPM alphas of beta-sorted portfolios"
    )
  + scale_y_continuous(labels=percent_format())
  + theme(legend_position="none")
)
beta_portfolios_figure.show()
```

Unlike the well-documented "betting against beta" anomaly in US markets, where low-beta portfolios exhibit positive alphas and high-beta portfolios exhibit negative alphas in a monotonic pattern, the Vietnamese market shows no clear relationship between beta and risk-adjusted returns. The alphas fluctuate without a discernible trend across deciles. This lack of pattern likely reflects the limited sample period rather than a definitive conclusion about beta pricing in Vietnam. With such a short time series, the portfolio-level CAPM regressions contain substantial estimation noise, making it difficult to detect subtle anomalies. Longer sample periods would be needed to draw reliable conclusions about whether the low-beta anomaly exists in the Vietnamese equity market.

Figure 10.1: The figure shows CAPM alphas of beta-sorted portfolios. Portfolios are sorted into deciles each month based on their estimated CAPM beta. The bar charts indicate the CAPM alpha of the resulting portfolio returns during the sample period.

## 10.6 Security Market Line and Beta Portfolios

The CAPM predicts that our portfolios should lie on the security market line (SML). The slope of the SML is equal to the market risk premium and reflects the risk-return trade-off at any given time. Figure 10.2 illustrates the security market line: We see that (not surprisingly) the high-beta portfolio returns have a high correlation with the market returns. However, it seems like the average excess returns for high-beta stocks are lower than what the security market line implies would be an "appropriate" compensation for the high market risk.

```
sml_capm = (sm.OLS.from_formula(
    formula="ret ~ 1 + beta",
    data=beta_portfolios_summary
  )
  .fit()
  .params
)

sml_figure = (
```

```
  ggplot(
    beta_portfolios_summary,
    aes(x="beta", y="ret", color="factor(portfolio)")
  )
  + geom_point()
  + geom_abline(
      intercept=0, slope=factors_ff3_monthly["mkt_excess"].mean(), linetype="solid"
    )
  + geom_abline(
      intercept=sml_capm["Intercept"], slope=sml_capm["beta"], linetype="dashed"
    )
  + labs(
      x="Beta", y="Excess return", color="Portfolio",
      title="Average portfolio excess returns and beta estimates"
    )
  + scale_x_continuous(limits=(0, 2))
  + scale_y_continuous(labels=percent_format())
)
sml_figure.show()
```

To provide more evidence against the CAPM predictions, we again form a long-short strategy that buys the high-beta portfolio and shorts the low-beta portfolio.

```
beta_longshort = (beta_portfolios
  .assign(
    portfolio=lambda x: (
      x["portfolio"].apply(
        lambda y: "high" if y == x["portfolio"].max()
        else ("low" if y == x["portfolio"].min()
        else y)
      )
    )
  )
  .query("portfolio in ['low', 'high']")
  .pivot_table(index="date", columns="portfolio", values="ret")
  .assign(long_short=lambda x: x["high"]-x["low"])
  .merge(factors_ff3_monthly, how="left", on="date")
)
```

Again, the resulting long-short strategy does not exhibit statistically significant returns.

Figure 10.2: The figure shows average portfolio excess returns and beta estimates. Excess returns are computed as CAPM alphas of the beta-sorted portfolios. The horizontal axis indicates the CAPM beta of the resulting beta-sorted portfolio return time series. The dashed line indicates the slope coefficient of a linear regression of excess returns on portfolio betas.

```
model_fit = (sm.OLS.from_formula(
    formula="long_short ~ 1",
    data=beta_longshort
  )
  .fit(cov_type="HAC", cov_kwds={"maxlags": 1})
)
prettify_result(model_fit)
```

```
OLS Model:
long_short ~ 1

Coefficients:
           Estimate  Std. Error  t-Statistic  p-Value
Intercept     0.001       0.011        0.095    0.924

Summary statistics:
```

```
- Number of observations: 53
- R-squared: 0.000, Adjusted R-squared: 0.000
- F-statistic not available
```

However, controlling for the effect of beta, the long-short portfolio yields a CAPM-adjusted alpha. The results can provide evidence regarding the validity of the CAPM in the Vietnamese market. The betting-against-beta factor has been documented extensively in developed markets (Frazzini and Pedersen 2014). Betting-against-beta corresponds to a strategy that shorts high-beta stocks and takes a (levered) long position in low-beta stocks. If borrowing constraints prevent investors from taking positions on the security market line, they are instead incentivized to buy high-beta stocks, which leads to a relatively higher price (and therefore lower expected returns than implied by the CAPM) for such high-beta stocks. As a result, the betting-against-beta strategy earns from providing liquidity to capital-constrained investors with lower risk aversion.

```python
model_fit = (sm.OLS.from_formula(
    formula="long_short ~ 1 + mkt_excess",
    data=beta_longshort
  )
  .fit(cov_type="HAC", cov_kwds={"maxlags": 1})
)
prettify_result(model_fit)
```

```
OLS Model:
long_short ~ 1 + mkt_excess

Coefficients:
            Estimate  Std. Error  t-Statistic  p-Value
Intercept      0.004       0.009        0.394    0.694
mkt_excess     1.020       0.116        8.784    0.000

Summary statistics:
- Number of observations: 52
- R-squared: 0.527, Adjusted R-squared: 0.518
- F-statistic: 77.156 on 1 and 50 DF, p-value: 0.000
```

Figure 10.3 shows the annual returns of the extreme beta portfolios we are mainly interested in. The figure illustrates the patterns over the sample period; each portfolio exhibits periods with positive and negative annual returns.

```
beta_longshort_year = (beta_longshort
  .assign(year=lambda x: x["date"].dt.year)
  .groupby("year")
  .aggregate(
    low=("low", lambda x: (1+x).prod()-1),
    high=("high", lambda x: (1+x).prod()-1),
    long_short=("long_short", lambda x: (1+x).prod()-1)
  )
  .reset_index()
  .melt(id_vars="year", var_name="name", value_name="value")
)

beta_longshort_figure = (
  ggplot(
    beta_longshort_year,
    aes(x="year", y="value", fill="name")
  )
  + geom_col(position="dodge")
  + facet_wrap("~name", ncol=1)
  + labs(x="", y="", title="Annual returns of beta portfolios")
  + scale_y_continuous(labels=percent_format())
  + theme(legend_position="none")
)
beta_longshort_figure.show()
```

The high-beta portfolio and low-beta portfolio both exhibit substantial year-to-year variation. The long-short portfolio, which goes long high-beta stocks and short low-beta stocks, shows no consistent pattern of positive returns. This erratic performance reinforces our earlier finding that the beta-return relationship in the Vietnamese market does not conform to theoretical CAPM predictions during our sample period. The high volatility of annual long-short returns highlights the substantial risk inherent in such a strategy, particularly in an emerging market context with a limited sample period.

## 10.7 Key Takeaways

1. Univariate portfolio sorts assess whether a single firm characteristic, like lagged market beta, can predict future excess returns.
2. Portfolios are formed each month using quantile breakpoints, with returns computed using value-weighted averages to reflect realistic investment strategies.

Figure 10.3: The figure shows annual returns of beta portfolios. We construct portfolios by sorting stocks into high and low based on their estimated CAPM beta. Long short indicates a strategy that goes long into high beta stocks and short low beta stocks.

3. A long-short strategy based on beta-sorted portfolios fails to generate significant positive excess returns in the Vietnamese market, contradicting CAPM predictions that higher beta should yield higher returns.

4. The analysis provides a framework for examining the "betting against beta" anomaly, where low-beta portfolios may deliver higher alphas than high-beta portfolios, offering evidence regarding the validity of the CAPM.

5. The functional programming capabilities of Python enable scalable and flexible portfolio sorting, making it easy to analyze multiple characteristics and portfolio configurations.

6. Emerging markets like Vietnam may exhibit different beta-return relationships compared to developed markets, highlighting the importance of conducting market-specific empirical analysis rather than assuming universal applicability of asset pricing anomalies.

# 11 Fama-French Factors

This chapter provides a replication of the Fama-French factor portfolios for the Vietnamese stock market. The Fama-French factor models represent a cornerstone of empirical asset pricing, originating from the seminal work of Eugene F. Fama and French (1992) and later extended in Eugene F. Fama and French (2015). These models have transformed how academics and practitioners understand the cross-section of expected stock returns, moving beyond the single-factor Capital Asset Pricing Model to incorporate multiple sources of systematic risk.

We construct both the three-factor and five-factor models at monthly and daily frequencies. The monthly factors serve as the foundation for most asset pricing tests and portfolio analyses, while the daily factors enable higher-frequency applications including short-horizon event studies, market microstructure research, and daily beta estimation. By constructing factors at both frequencies, we create a complete toolkit for empirical finance research in the Vietnamese market.

The chapter proceeds as follows. We first discuss the theoretical motivation for each factor and the economic intuition behind the Fama-French methodology. We then prepare the necessary data, merging stock returns with accounting characteristics. Next, we implement the portfolio sorting procedures that form the basis of factor construction, carefully following the original Fama-French protocols while adapting them for Vietnamese market characteristics. We construct the three-factor model (market, size, and value) before extending to the five-factor model (adding profitability and investment). Finally, we construct daily factors and validate our replicated factors through various diagnostic checks.

## 11.1 Theoretical Background

### 11.1.1 The Evolution from CAPM to Multi-Factor Models

The Capital Asset Pricing Model of Sharpe (1964) posits that a single factor—the market portfolio—should explain all cross-sectional variation in expected returns. However, decades of empirical research have documented persistent patterns that CAPM cannot explain. Eugene F. Fama and French (1992) demonstrated that two firm characteristics—size and book-to-market ratio—capture substantial variation in average returns that the market beta leaves unexplained.

Small firms tend to earn higher returns than large firms, a pattern known as the size effect. Similarly, firms with high book-to-market ratios (value stocks) tend to outperform firms with low book-to-market ratios (growth stocks), known as the value premium. The three-factor model formalizes these observations by constructing tradeable factor portfolios:

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i^{MKT}(r_{m,t} - r_{f,t}) + \beta_i^{SMB} \cdot SMB_t + \beta_i^{HML} \cdot HML_t + \varepsilon_{i,t} \tag{11.1}$$

where:

- $r_{i,t} - r_{f,t}$ is the excess return on asset $i$
- $r_{m,t} - r_{f,t}$ is the market excess return
- $SMB_t$ (Small Minus Big) is the size factor
- $HML_t$ (High Minus Low) is the value factor

### 11.1.2 The Five-Factor Extension

Eugene F. Fama and French (2015) extended the model to include two additional factors motivated by the dividend discount model. Firms with higher profitability should have higher expected returns (all else equal), and firms with aggressive investment policies should have lower expected returns:

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i^{MKT} MKT_t + \beta_i^{SMB} SMB_t + \beta_i^{HML} HML_t + \beta_i^{RMW} RMW_t + \beta_i^{CMA} CMA_t + \varepsilon_{i,t} \tag{11.2}$$

where:

- $RMW_t$ (Robust Minus Weak) is the profitability factor
- $CMA_t$ (Conservative Minus Aggressive) is the investment factor

### 11.1.3 Factor Construction Methodology

The Fama-French methodology constructs factors through double-sorted portfolios:

1. **Size sorts**: Stocks are divided into Small and Big groups based on median market capitalization.
2. **Characteristic sorts**: Within each size group, stocks are sorted into terciles based on book-to-market (for HML), operating profitability (for RMW), or investment (for CMA).
3. **Factor returns**: Factors are computed as the difference between average returns of portfolios with high versus low characteristic values, averaging across size groups to neutralize size effects.

4. **Timing**: Portfolios are formed at the end of June each year using accounting data from the prior fiscal year, ensuring all information was publicly available at formation.

## 11.2 Setting Up the Environment

We load the required Python packages for data manipulation, statistical analysis, and visualization.

```
import pandas as pd
import numpy as np
import sqlite3
import statsmodels.formula.api as smf
from scipy.stats.mstats import winsorize
import matplotlib.pyplot as plt

from plotnine import *
from mizani.formatters import percent_format, comma_format
```

We connect to our SQLite database containing the processed Vietnamese financial data.

```
tidy_finance = sqlite3.connect(database="data/tidy_finance_python.sqlite")
```

## 11.3 Data Preparation

### 11.3.1 Loading Stock Returns

We load the monthly stock returns data, which includes excess returns, market capitalization, and the risk-free rate. These variables are essential for computing value-weighted portfolio returns and factor premiums.

```
prices_monthly = pd.read_sql_query(
    sql="""
        SELECT symbol, date, ret_excess, mktcap, mktcap_lag, risk_free
        FROM prices_monthly
    """,
    con=tidy_finance,
    parse_dates={"date"}
).dropna()
```

```
print(f"Monthly returns: {len(prices_monthly):,} observations")
print(f"Unique stocks: {prices_monthly['symbol'].nunique():,}")
print(f"Date range: {prices_monthly['date'].min():%Y-%m} to {prices_monthly['date'].max():%Y-
```

```
Monthly returns: 165,499 observations
Unique stocks: 1,457
Date range: 2010-02 to 2023-12
```

### 11.3.2 Loading Company Fundamentals

We load the company fundamentals data containing book equity, operating profitability, and investment—the characteristics needed for constructing the Fama-French factors.

```
comp_vn = pd.read_sql_query(
    sql="""
        SELECT symbol, datadate, be, op, inv
        FROM comp_vn
    """,
    con=tidy_finance,
    parse_dates={"datadate"}
).dropna()

print(f"Fundamentals: {len(comp_vn):,} firm-year observations")
print(f"Unique firms: {comp_vn['symbol'].nunique():,}")
```

```
Fundamentals: 18,108 firm-year observations
Unique firms: 1,496
```

### 11.3.3 Constructing Sorting Variables

Following Fama-French conventions, we construct the sorting variables with careful attention to timing. The key principles are:

1. **Size (June Market Cap)**: We use market capitalization at the end of June of year $t$ to sort stocks into size groups. This ensures we capture the firm's size at the moment of portfolio formation.
2. **Book-to-Market Ratio**: We use book equity from fiscal year $t-1$ divided by market equity at the end of December $t-1$. This creates a six-month gap between the accounting data and portfolio formation, ensuring the information was publicly available.

3. **Portfolio Formation Date**: Portfolios are formed on July 1st and held for twelve months until the following June.

```python
def construct_sorting_variables(prices_monthly, comp_vn):
    """
    Construct sorting variables following Fama-French methodology.

    Parameters
    ----------
    prices_monthly : pd.DataFrame
        Monthly stock returns with market cap
    comp_vn : pd.DataFrame
        Company fundamentals with book equity, profitability, investment

    Returns
    -------
    pd.DataFrame
        Sorting variables aligned with July 1st formation dates
    """

    # 1. Size: June market capitalization
    # Portfolio formation is July 1st, so we use June market cap
    size = (prices_monthly
        .query("date.dt.month == 6")
        .assign(
            sorting_date=lambda x: x["date"] + pd.offsets.MonthBegin(1)
        )
        [["symbol", "sorting_date", "mktcap"]]
        .rename(columns={"mktcap": "size"})
    )

    print(f"Size observations: {len(size):,}")

    # 2. Market Equity: December market cap for B/M calculation
    # December t-1 market cap is used with fiscal year t-1 book equity
    # This is then used for July t portfolio formation
    market_equity = (prices_monthly
        .query("date.dt.month == 12")
        .assign(
            # December year t-1 maps to July year t formation
            sorting_date=lambda x: x["date"] + pd.offsets.MonthBegin(7)
        )
        [["symbol", "sorting_date", "mktcap"]]
```

```
        .rename(columns={"mktcap": "me"})
    )

    print(f"Market equity observations: {len(market_equity):,}")

    # 3. Book-to-Market and other characteristics
    # Fiscal year t-1 data is used for July t portfolio formation
    book_to_market = (comp_vn
        .assign(
            # Fiscal year-end + 6 months = July formation
            sorting_date=lambda x: pd.to_datetime(
                (x["datadate"].dt.year + 1).astype(str) + "-07-01"
            )
        )
        .merge(market_equity, on=["symbol", "sorting_date"], how="inner")
        .assign(
            # Scale book equity to match market equity units
            # BE is in VND, ME is in millions VND
            bm=lambda x: x["be"] / (x["me"] * 1e9)
        )
        [["symbol", "sorting_date", "me", "bm", "op", "inv"]]
    )

    print(f"Book-to-market observations: {len(book_to_market):,}")

    # 4. Merge size with characteristics
    sorting_variables = (size
        .merge(book_to_market, on=["symbol", "sorting_date"], how="inner")
        .dropna()
        .drop_duplicates(subset=["symbol", "sorting_date"])
    )

    return sorting_variables

sorting_variables = construct_sorting_variables(prices_monthly, comp_vn)

print(f"\nFinal sorting variables: {len(sorting_variables):,} stock-years")
print(f"Sorting date range: {sorting_variables['sorting_date'].min():%Y-%m} to {sorting_varia
```

```
Size observations: 13,756
Market equity observations: 14,286
Book-to-market observations: 13,389
```

```
Final sorting variables: 12,046 stock-years
Sorting date range: 2011-07 to 2023-07
```

### 11.3.4 Validating Sorting Variables

Before proceeding, we validate that our sorting variables have reasonable distributions. The book-to-market ratio should center around 1.0 for a typical market, though emerging markets may differ.

```python
print("Sorting Variable Summary Statistics:")
print(sorting_variables[["size", "bm", "op", "inv"]].describe().round(4))

# Check for extreme values that might indicate data issues
print(f"\nB/M Median: {sorting_variables['bm'].median():.4f}")
print(f"B/M 1st percentile: {sorting_variables['bm'].quantile(0.01):.4f}")
print(f"B/M 99th percentile: {sorting_variables['bm'].quantile(0.99):.4f}")
```

```
Sorting Variable Summary Statistics:
               size          bm          op         inv
count   12046.0000  12046.0000  12046.0000  12046.0000
mean     2225.5648      1.7033      0.1852      0.1322
std     14680.4225      3.8683      0.2782      2.5410
min         0.4864      0.0014     -0.7529     -0.9569
25%        62.7556      0.7595      0.0309     -0.0495
50%       182.6410      1.1849      0.1367      0.0342
75%       641.8896      1.8853      0.2952      0.1582
max    426020.9817    272.1893      1.4256    261.3355

B/M Median: 1.1849
B/M 1st percentile: 0.1710
B/M 99th percentile: 8.0262
```

### 11.3.5 Handling Outliers

Extreme values in sorting characteristics can distort portfolio assignments and factor returns. We apply winsorization to limit the influence of outliers while preserving the general ranking of stocks.

```python
# Check BEFORE winsorization
print("BEFORE Winsorization:")
print(sorting_variables[["size", "bm", "op", "inv"]].describe().round(4))

# Apply winsorization
def winsorize_characteristics(df, columns, limits=(0.01, 0.99)):
    """
    Apply winsorization using pandas clip.
    """
    df = df.copy()
    for col in columns:
        if col in df.columns:
            lower = df[col].quantile(limits[0])
            upper = df[col].quantile(limits[1])
            df[col] = df[col].clip(lower=lower, upper=upper)
            print(f"  {col}: clipped to [{lower:.4f}, {upper:.4f}]")
    return df

sorting_variables = winsorize_characteristics(
    sorting_variables,
    columns=["bm", "op", "inv"],  # Don't winsorize size
    limits=(0.01, 0.99)
)

# Check AFTER winsorization
print("\nAFTER Winsorization:")
print(sorting_variables[["size", "bm", "op", "inv"]].describe().round(4))
```

```
BEFORE Winsorization:
              size          bm          op         inv
count   12046.0000  12046.0000  12046.0000  12046.0000
mean     2225.5648      1.7033      0.1852      0.1322
std     14680.4225      3.8683      0.2782      2.5410
min         0.4864      0.0014     -0.7529     -0.9569
25%        62.7556      0.7595      0.0309     -0.0495
50%       182.6410      1.1849      0.1367      0.0342
75%       641.8896      1.8853      0.2952      0.1582
max    426020.9817    272.1893      1.4256    261.3355
  bm: clipped to [0.1710, 8.0262]
  op: clipped to [-0.7319, 1.2192]
  inv: clipped to [-0.3990, 1.5195]
```

```
AFTER Winsorization:
              size          bm          op         inv
count   12046.0000  12046.0000  12046.0000  12046.0000
mean     2225.5648      1.5544      0.1837      0.0894
std     14680.4225      1.2843      0.2705      0.2721
min         0.4864      0.1710     -0.7319     -0.3990
25%        62.7556      0.7595      0.0309     -0.0495
50%       182.6410      1.1849      0.1367      0.0342
75%       641.8896      1.8853      0.2952      0.1582
max    426020.9817      8.0262      1.2192      1.5195
```

## 11.4 Portfolio Assignment Functions

### 11.4.1 The Portfolio Assignment Function

We create a flexible function for assigning stocks to portfolios based on quantile breakpoints. This function handles both independent sorts (where breakpoints are computed across all stocks) and dependent sorts (where breakpoints are computed within subgroups).

```python
def assign_portfolio(data, sorting_variable, percentiles):
    """Assign portfolios to a bin according to a sorting variable."""

    # Get the values
    values = data[sorting_variable].dropna()

    if len(values) == 0:
        return pd.Series([np.nan] * len(data), index=data.index)

    # Calculate breakpoints
    breakpoints = values.quantile(percentiles, interpolation="linear")

    # Handle duplicate breakpoints by using unique values
    unique_breakpoints = np.unique(breakpoints)

    # If all values are the same, assign all to portfolio 1
    if len(unique_breakpoints) <= 1:
        return pd.Series([1] * len(data), index=data.index)

    # Set boundaries to -inf and +inf
    unique_breakpoints.iloc[0] = -np.inf
    unique_breakpoints.iloc[unique_breakpoints.size-1] = np.inf
```

```
    # Assign to bins
    assigned = pd.cut(
        data[sorting_variable],
        bins=unique_breakpoints,
        labels=pd.Series(range(1, breakpoints.size)),
        include_lowest=True,
        right=False
    )

    return assigned
```

```
# Check the distribution of characteristics BEFORE portfolio assignment
print("Operating Profitability Distribution:")
print(sorting_variables["op"].describe())
print(f"\nUnique OP values: {sorting_variables['op'].nunique()}")

print("\nInvestment Distribution:")
print(sorting_variables["inv"].describe())
print(f"\nUnique INV values: {sorting_variables['inv'].nunique()}")

# Check breakpoints for a specific date
test_date = sorting_variables["sorting_date"].iloc[0]
test_data = sorting_variables.query("sorting_date == @test_date")

print(f"\nBreakpoints for {test_date}:")
print(f"OP 30th percentile: {test_data['op'].quantile(0.3):.4f}")
print(f"OP 70th percentile: {test_data['op'].quantile(0.7):.4f}")
print(f"INV 30th percentile: {test_data['inv'].quantile(0.3):.4f}")
print(f"INV 70th percentile: {test_data['inv'].quantile(0.7):.4f}")
```

```
Operating Profitability Distribution:
count    12046.000000
mean         0.183738
std          0.270509
min         -0.731888
25%          0.030913
50%          0.136675
75%          0.295185
max          1.219223
Name: op, dtype: float64
```

```
Unique OP values: 11804

Investment Distribution:
count    12046.000000
mean         0.089388
std          0.272147
min         -0.399042
25%         -0.049497
50%          0.034157
75%          0.158155
max          1.519474
Name: inv, dtype: float64

Unique INV values: 11805

Breakpoints for 2019-07-01 00:00:00:
OP 30th percentile: 0.0541
OP 70th percentile: 0.2566
INV 30th percentile: -0.0343
INV 70th percentile: 0.1116
```

### 11.4.2 Assigning Portfolios for Three-Factor Model

For the three-factor model, we perform independent double sorts on size and book-to-market. Size is split at the median (2 groups), and book-to-market is split at the 30th and 70th percentiles (3 groups), creating 6 portfolios.

```python
def assign_ff3_portfolios(sorting_variables):
    """
    Assign portfolios for Fama-French three-factor model.
    Independent 2x3 sort on size and book-to-market.
    """
    df = sorting_variables.copy()

    # Independent size sort (median split)
    df["portfolio_size"] = df.groupby("sorting_date")["size"].transform(
        lambda x: pd.qcut(x, q=[0, 0.5, 1], labels=[1, 2], duplicates='drop')
    )

    # Independent B/M sort (30/70 split)
    df["portfolio_bm"] = df.groupby("sorting_date")["bm"].transform(
        lambda x: pd.qcut(x, q=[0, 0.3, 0.7, 1], labels=[1, 2, 3], duplicates='drop')
```

```
    )

    return df

# Assign portfolios
portfolios_ff3 = assign_ff3_portfolios(sorting_variables)

# Validate
print("FF3 Book-to-Market by Portfolio (should be INCREASING):")
print(portfolios_ff3.groupby("portfolio_bm", observed=True)["bm"].median().round(4))


print("Three-Factor Portfolio Assignments:")
print(portfolios_ff3[["symbol", "sorting_date", "portfolio_size", "portfolio_bm"]].head(10))
```

```
FF3 Book-to-Market by Portfolio (should be INCREASING):
portfolio_bm
1    0.5836
2    1.1891
3    2.4552
Name: bm, dtype: float64
Three-Factor Portfolio Assignments:
  symbol sorting_date portfolio_size portfolio_bm
0    A32   2019-07-01              1            2
1    A32   2020-07-01              1            2
2    A32   2021-07-01              1            2
3    A32   2022-07-01              1            3
4    A32   2023-07-01              1            2
5    AAA   2011-07-01              2            2
6    AAA   2012-07-01              2            3
7    AAA   2013-07-01              2            3
8    AAA   2014-07-01              2            2
9    AAA   2015-07-01              2            3
```

### 11.4.3 Validating Portfolio Assignments

We verify that the portfolio assignments create the expected 2×3 grid with reasonable stock counts in each cell.

```python
# Check portfolio distribution for most recent year
latest_date = portfolios_ff3["sorting_date"].max()

portfolio_counts = (portfolios_ff3
    .query("sorting_date == @latest_date")
    .groupby(["portfolio_size", "portfolio_bm"], observed=True)
    .size()
    .unstack(fill_value=0)
)

print(f"Portfolio Counts for {latest_date:%Y-%m}:")
print(portfolio_counts)

# Verify characteristic monotonicity
print("\nBook-to-Market by Portfolio (should be increasing):")
print(portfolios_ff3.groupby("portfolio_bm", observed=True)["bm"].median().round(4))
```

```
Portfolio Counts for 2023-07:
portfolio_bm      1    2    3
portfolio_size
1               113  271  263
2               275  246  125

Book-to-Market by Portfolio (should be increasing):
portfolio_bm
1    0.5836
2    1.1891
3    2.4552
Name: bm, dtype: float64
```

We verify that for a single stock, the portfolio assignment remains constant between July of one year and June of the next.

```python
# Trace a single symbol (e.g., 'A32') across a formation window
persistence_check = (portfolios_ff3
    .query("symbol == 'A32' & sorting_date >= '2022-01-01' & sorting_date <= '2023-12-31'")
    .sort_values("sorting_date")
    [['symbol', 'sorting_date', 'portfolio_size', 'portfolio_bm']]
)
print("\nTemporal Persistence Check (Symbol A32):")
print(persistence_check.head(15))
```

```
Temporal Persistence Check (Symbol A32):
  symbol sorting_date portfolio_size portfolio_bm
3    A32   2022-07-01              1            3
4    A32   2023-07-01              1            2
```

## 11.5 Fama-French Three-Factor Model (Monthly)

### 11.5.1 Merging Portfolios with Returns

We merge the portfolio assignments with monthly returns. The key insight is that portfolios formed in July of year $t$ are held through June of year $t+1$. We implement this by computing a `sorting_date` for each monthly return observation.

```python
def merge_portfolios_with_returns(prices_monthly, portfolio_assignments):
    """
    Merge portfolio assignments with monthly returns.

    Portfolios formed in July t are held through June t+1.

    Parameters
    ----------
    prices_monthly : pd.DataFrame
        Monthly stock returns
    portfolio_assignments : pd.DataFrame
        Portfolio assignments with sorting_date

    Returns
    -------
    pd.DataFrame
        Returns merged with portfolio assignments
    """
    portfolios = (prices_monthly
        .assign(
            # Map each return month to its portfolio formation date
            sorting_date=lambda x: pd.to_datetime(
                np.where(
                    x["date"].dt.month <= 6,
                    (x["date"].dt.year - 1).astype(str) + "-07-01",
                    x["date"].dt.year.astype(str) + "-07-01"
                )
```

```
            )
        )
        .merge(
            portfolio_assignments,
            on=["symbol", "sorting_date"],
            how="inner"
        )
    )

    return portfolios

portfolios_monthly_ff3 = merge_portfolios_with_returns(
    prices_monthly,
    portfolios_ff3[["symbol", "sorting_date", "portfolio_size", "portfolio_bm"]]
)


print(f"Merged observations: {len(portfolios_monthly_ff3):,}")
```

```
Merged observations: 136,444
```

### 11.5.2 Computing Value-Weighted Portfolio Returns

We compute value-weighted returns for each of the six portfolios. Value-weighting uses lagged market capitalization to avoid look-ahead bias.

```
def compute_portfolio_returns(data, grouping_vars):
    """
    Compute value-weighted portfolio returns.

    Parameters
    ----------
    data : pd.DataFrame
        Returns data with portfolio assignments and mktcap_lag
    grouping_vars : list
        Variables defining portfolio groups

    Returns
    -------
    pd.DataFrame
        Value-weighted returns for each portfolio-date
```

```python
    """
    portfolio_returns = (data
        .groupby(grouping_vars + ["date"], observed=True)
        .apply(lambda x: pd.Series({
            "ret": np.average(x["ret_excess"], weights=x["mktcap_lag"]),
            "n_stocks": len(x)
        }))
        .reset_index()
    )

    return portfolio_returns


# Compute portfolio returns
portfolio_returns_ff3 = compute_portfolio_returns(
    portfolios_monthly_ff3,
    ["portfolio_size", "portfolio_bm"]
)

print("Portfolio Returns Summary:")
print(portfolio_returns_ff3.groupby(["portfolio_size", "portfolio_bm"], observed=True)["ret"]
```

```
Portfolio Returns Summary:
                              count    mean     std     min     25%     50%  \
portfolio_size portfolio_bm
1              1              150.0 -0.0052  0.0379 -0.1268 -0.0262 -0.0058
               2              150.0 -0.0025  0.0414 -0.1080 -0.0236 -0.0053
               3              150.0  0.0039  0.0601 -0.1612 -0.0269 -0.0004
2              1              150.0 -0.0124  0.0594 -0.2222 -0.0449 -0.0107
               2              150.0 -0.0021  0.0671 -0.1701 -0.0403 -0.0046
               3              150.0  0.0024  0.0879 -0.2359 -0.0527 -0.0012

                                 75%     max
portfolio_size portfolio_bm
1              1              0.0161  0.0849
               2              0.0180  0.1195
               3              0.0314  0.2015
2              1              0.0210  0.1741
               2              0.0317  0.1770
               3              0.0437  0.2124
```

### 11.5.3 Constructing SMB and HML Factors

We now construct the SMB and HML factors from the portfolio returns.

**SMB (Small Minus Big)**: Average return of three small portfolios minus average return of three big portfolios.

**HML (High Minus Low)**: Average return of two high B/M portfolios minus average return of two low B/M portfolios.

```
def construct_ff3_factors(portfolio_returns):
    """
    Construct Fama-French three factors from portfolio returns.

    Parameters
    ----------
    portfolio_returns : pd.DataFrame
        Value-weighted returns for 2x3 portfolios

    Returns
    -------
    pd.DataFrame
        Monthly SMB and HML factors
    """
    factors = (portfolio_returns
        .groupby("date")
        .apply(lambda x: pd.Series({
            # SMB: Small minus Big (average across B/M groups)
            "smb": (
                x.loc[x["portfolio_size"] == 1, "ret"].mean() -
                x.loc[x["portfolio_size"] == 2, "ret"].mean()
            ),
            # HML: High minus Low B/M (average across size groups)
            "hml": (
                x.loc[x["portfolio_bm"] == 3, "ret"].mean() -
                x.loc[x["portfolio_bm"] == 1, "ret"].mean()
            )
        }))
        .reset_index()
    )

    return factors
```

```
factors_smb_hml = construct_ff3_factors(portfolio_returns_ff3)

print("SMB and HML Factors:")
print(factors_smb_hml.head(10))
```

```
SMB and HML Factors:
        date       smb       hml
0 2011-07-31 -0.007768  0.002754
1 2011-08-31 -0.067309  0.011474
2 2011-09-30  0.014884  0.022854
3 2011-10-31 -0.003743  0.001631
4 2011-11-30  0.063234  0.009103
5 2011-12-31  0.014571  0.015280
6 2012-01-31 -0.026080  0.009672
7 2012-02-29 -0.035721  0.005474
8 2012-03-31 -0.002344  0.032477
9 2012-04-30 -0.033391  0.074191
```

### 11.5.4 Computing the Market Factor

The market factor is the value-weighted return of all stocks minus the risk-free rate. We compute this independently from the sorted portfolios.

```
def compute_market_factor(prices_monthly):
    """
    Compute value-weighted market excess return.

    Parameters
    ----------
    prices_monthly : pd.DataFrame
        Monthly stock returns with mktcap_lag

    Returns
    -------
    pd.DataFrame
        Monthly market excess return
    """
    market_factor = (prices_monthly
        .groupby("date")
        .apply(lambda x: pd.Series({
            "mkt_excess": np.average(x["ret_excess"], weights=x["mktcap_lag"]),
```

```
            "n_stocks": len(x)
        }), include_groups=False)
        .reset_index()
    )

    return market_factor

market_factor = compute_market_factor(prices_monthly)

print("Market Factor Summary:")
print(market_factor["mkt_excess"].describe().round(4))
```

```
Market Factor Summary:
count    167.0000
mean      -0.0123
std        0.0595
min       -0.2149
25%       -0.0394
50%       -0.0106
75%        0.0200
max        0.1677
Name: mkt_excess, dtype: float64
```

### 11.5.5 Combining Three Factors

We combine SMB, HML, and the market factor into the complete three-factor dataset.

```
factors_ff3_monthly = (factors_smb_hml
    .merge(market_factor[["date", "mkt_excess"]], on="date", how="inner")
)

# Add risk-free rate for completeness
rf_monthly = (prices_monthly
    .groupby("date")["risk_free"]
    .first()
    .reset_index()
)

factors_ff3_monthly = factors_ff3_monthly.merge(rf_monthly, on="date", how="left")

print("Fama-French Three Factors (Monthly):")
```

```
print(factors_ff3_monthly.head(10))

print("\nFactor Summary Statistics:")
print(factors_ff3_monthly[["mkt_excess", "smb", "hml"]].describe().round(4))
```

```
Fama-French Three Factors (Monthly):
        date       smb       hml  mkt_excess  risk_free
0 2011-07-31 -0.007768  0.002754   -0.078748   0.003333
1 2011-08-31 -0.067309  0.011474    0.029906   0.003333
2 2011-09-30  0.014884  0.022854   -0.002173   0.003333
3 2011-10-31 -0.003743  0.001631   -0.014005   0.003333
4 2011-11-30  0.063234  0.009103   -0.179410   0.003333
5 2011-12-31  0.014571  0.015280   -0.094802   0.003333
6 2012-01-31 -0.026080  0.009672    0.081273   0.003333
7 2012-02-29 -0.035721  0.005474    0.069655   0.003333
8 2012-03-31 -0.002344  0.032477    0.029005   0.003333
9 2012-04-30 -0.033391  0.074191    0.048791   0.003333

Factor Summary Statistics:
       mkt_excess       smb       hml
count    150.0000  150.0000  150.0000
mean      -0.0101    0.0027    0.0120
std        0.0586    0.0420    0.0535
min       -0.2149   -0.1599   -0.1284
25%       -0.0380   -0.0175   -0.0160
50%       -0.0095    0.0070    0.0043
75%        0.0214    0.0261    0.0340
max        0.1677    0.1175    0.1618
```

### 11.5.6 Saving Three-Factor Data

```
factors_ff3_monthly.to_sql(
    name="factors_ff3_monthly",
    con=tidy_finance,
    if_exists="replace",
    index=False
)

print("Three-factor monthly data saved to database.")
```

```
Three-factor monthly data saved to database.
```

## 11.6 Fama-French Five-Factor Model (Monthly)

### 11.6.1 Portfolio Assignments with Dependent Sorts

For the five-factor model, we use dependent sorts: size is sorted independently, but profitability and investment are sorted within size groups. This controls for the correlation between size and these characteristics.

```python
def assign_ff5_portfolios(sorting_variables):
    """
    Assign portfolios for Fama-French five-factor model.
    """
    df = sorting_variables.copy()

    # Independent size sort
    df["portfolio_size"] = df.groupby("sorting_date")["size"].transform(
        lambda x: pd.qcut(x, q=[0, 0.5, 1], labels=[1, 2], duplicates='drop')
    )

    # Dependent sorts within size groups
    df["portfolio_bm"] = df.groupby(["sorting_date", "portfolio_size"])["bm"].transform(
        lambda x: pd.qcut(x, q=[0, 0.3, 0.7, 1], labels=[1, 2, 3], duplicates='drop')
    )

    df["portfolio_op"] = df.groupby(["sorting_date", "portfolio_size"])["op"].transform(
        lambda x: pd.qcut(x, q=[0, 0.3, 0.7, 1], labels=[1, 2, 3], duplicates='drop')
    )

    df["portfolio_inv"] = df.groupby(["sorting_date", "portfolio_size"])["inv"].transform(
        lambda x: pd.qcut(x, q=[0, 0.3, 0.7, 1], labels=[1, 2, 3], duplicates='drop')
    )

    return df

# Run
portfolios_ff5 = assign_ff5_portfolios(sorting_variables)
```

### 11.6.2 Validating Five-Factor Portfolios

```
# Check characteristic monotonicity for each dimension
print("Profitability by Portfolio (should be increasing):")
print(portfolios_ff5.groupby("portfolio_op", observed=True)["op"].median().round(4))

print("\nInvestment by Portfolio (should be increasing):")
print(portfolios_ff5.groupby("portfolio_inv", observed=True)["inv"].median().round(4))

# Check portfolio counts
print("\nStocks per Size/Profitability Bin:")
print(portfolios_ff5.groupby(["portfolio_size", "portfolio_op"], observed=True).size().unsta

# Check number of unique firms per year
(portfolios_ff5
 .groupby("sorting_date")["symbol"]
 .nunique())
```

```
Profitability by Portfolio (should be increasing):
portfolio_op
1   -0.0053
2    0.1366
3    0.4098
Name: op, dtype: float64

Investment by Portfolio (should be increasing):
portfolio_inv
1   -0.1012
2    0.0329
3    0.2568
Name: inv, dtype: float64

Stocks per Size/Profitability Bin:
portfolio_op       1     2     3
portfolio_size
1               1812  2403  1811
2               1811  2401  1808


sorting_date
2011-07-01     556
2012-07-01     632
```

```
2013-07-01     643
2014-07-01     650
2015-07-01     669
2016-07-01     737
2017-07-01     842
2018-07-01    1073
2019-07-01    1183
2020-07-01    1224
2021-07-01    1258
2022-07-01    1286
2023-07-01    1293
Name: symbol, dtype: int64
```

### 11.6.3 Merging and Computing Portfolio Returns

```python
# Merge with returns
portfolios_monthly_ff5 = merge_portfolios_with_returns(
    prices_monthly,
    portfolios_ff5[["symbol", "sorting_date", "portfolio_size",
                    "portfolio_bm", "portfolio_op", "portfolio_inv"]]
)

print(f"Five-factor merged observations: {len(portfolios_monthly_ff5):,}")
```

```
Five-factor merged observations: 136,444
```

### 11.6.4 Constructing All Five Factors

We construct each factor from the appropriate portfolio sorts.

```python
def construct_ff5_factors(portfolios_monthly):
    """
    Construct Fama-French five factors from portfolio data.

    Parameters
    ----------
    portfolios_monthly : pd.DataFrame
        Monthly returns with all portfolio assignments
```

```
    Returns
    -------
    pd.DataFrame
        Monthly five-factor returns
    """

    # HML: Value factor from B/M sorts
    portfolios_bm = (portfolios_monthly
        .groupby(["portfolio_size", "portfolio_bm", "date"], observed=True)
        .apply(lambda x: pd.Series({
            "ret": np.average(x["ret_excess"], weights=x["mktcap_lag"])
        }))
        .reset_index()
    )

    factors_hml = (portfolios_bm
        .groupby("date")
        .apply(lambda x: pd.Series({
            "hml": (x.loc[x["portfolio_bm"] == 3, "ret"].mean() -
                    x.loc[x["portfolio_bm"] == 1, "ret"].mean())
        }))
        .reset_index()
    )

    # RMW: Profitability factor from OP sorts
    portfolios_op = (portfolios_monthly
        .groupby(["portfolio_size", "portfolio_op", "date"], observed=True)
        .apply(lambda x: pd.Series({
            "ret": np.average(x["ret_excess"], weights=x["mktcap_lag"])
        }))
        .reset_index()
    )

    factors_rmw = (portfolios_op
        .groupby("date")
        .apply(lambda x: pd.Series({
            "rmw": (x.loc[x["portfolio_op"] == 3, "ret"].mean() -
                    x.loc[x["portfolio_op"] == 1, "ret"].mean())
        }))
        .reset_index()
    )
```

```
    # CMA: Investment factor from INV sorts
    # Note: CMA is Conservative minus Aggressive (low inv - high inv)
    portfolios_inv = (portfolios_monthly
        .groupby(["portfolio_size", "portfolio_inv", "date"], observed=True)
        .apply(lambda x: pd.Series({
            "ret": np.average(x["ret_excess"], weights=x["mktcap_lag"])
        }))
        .reset_index()
    )

    factors_cma = (portfolios_inv
        .groupby("date")
        .apply(lambda x: pd.Series({
            "cma": (x.loc[x["portfolio_inv"] == 1, "ret"].mean() -
                   x.loc[x["portfolio_inv"] == 3, "ret"].mean())
        }))
        .reset_index()
    )

    # SMB: Size factor (average across all characteristic portfolios)
    all_portfolios = pd.concat([portfolios_bm, portfolios_op, portfolios_inv])

    factors_smb = (all_portfolios
        .groupby("date")
        .apply(lambda x: pd.Series({
            "smb": (x.loc[x["portfolio_size"] == 1, "ret"].mean() -
                   x.loc[x["portfolio_size"] == 2, "ret"].mean())
        }))
        .reset_index()
    )

    # Combine all factors
    factors = (factors_smb
        .merge(factors_hml, on="date", how="outer")
        .merge(factors_rmw, on="date", how="outer")
        .merge(factors_cma, on="date", how="outer")
    )

    return factors

factors_ff5 = construct_ff5_factors(portfolios_monthly_ff5)
```

```python
# Add market factor
factors_ff5_monthly = (factors_ff5
    .merge(market_factor[["date", "mkt_excess"]], on="date", how="inner")
    .merge(rf_monthly, on="date", how="left")
)

print("Fama-French Five Factors (Monthly):")
print(factors_ff5_monthly.head(10))

print("\nFactor Summary Statistics:")
print(factors_ff5_monthly[["mkt_excess", "smb", "hml", "rmw", "cma"]].describe().round(4))
```

```
Fama-French Five Factors (Monthly):
        date       smb       hml       rmw       cma  mkt_excess  risk_free
0 2011-07-31 -0.015907 -0.002812  0.060525  0.045291   -0.078748   0.003333
1 2011-08-31 -0.061842  0.006189 -0.022700 -0.023177    0.029906   0.003333
2 2011-09-30  0.014387  0.024301 -0.006005  0.003588   -0.002173   0.003333
3 2011-10-31 -0.006958 -0.006940  0.026694  0.003649   -0.014005   0.003333
4 2011-11-30  0.074369  0.015617 -0.058766  0.044214   -0.179410   0.003333
5 2011-12-31  0.006687  0.022494  0.062655  0.052444   -0.094802   0.003333
6 2012-01-31 -0.016254  0.010513 -0.042191 -0.067170    0.081273   0.003333
7 2012-02-29 -0.026606  0.024465 -0.030849 -0.036383    0.069655   0.003333
8 2012-03-31  0.005096  0.050930 -0.018441  0.043488    0.029005   0.003333
9 2012-04-30  0.000712  0.058214 -0.061434  0.009233    0.048791   0.003333

Factor Summary Statistics:
       mkt_excess       smb       hml       rmw       cma
count    150.0000  150.0000  150.0000  150.0000  150.0000
mean      -0.0101    0.0077    0.0115   -0.0047    0.0083
std        0.0586    0.0419    0.0518    0.0477    0.0335
min       -0.2149   -0.1522   -0.1283   -0.2126   -0.0814
25%       -0.0380   -0.0137   -0.0126   -0.0308   -0.0131
50%       -0.0095    0.0104    0.0046    0.0010    0.0067
75%        0.0214    0.0316    0.0323    0.0178    0.0289
max        0.1677    0.1284    0.1510    0.1297    0.1331
```

### 11.6.5 Factor Correlations

We examine correlations between factors, which should generally be low for the factors to capture distinct sources of risk.

```
print("Factor Correlation Matrix:")
correlation_matrix = factors_ff5_monthly[["mkt_excess", "smb", "hml", "rmw", "cma"]].corr()
print(correlation_matrix.round(3))
```

```
Factor Correlation Matrix:
            mkt_excess    smb    hml    rmw    cma
mkt_excess       1.000 -0.712  0.230 -0.006 -0.104
smb             -0.712  1.000  0.256 -0.373  0.246
hml              0.230  0.256  1.000 -0.694  0.479
rmw             -0.006 -0.373 -0.694  1.000 -0.352
cma             -0.104  0.246  0.479 -0.352  1.000
```

### 11.6.6 Saving Five-Factor Data

```
factors_ff5_monthly.to_sql(
    name="factors_ff5_monthly",
    con=tidy_finance,
    if_exists="replace",
    index=False
)

print("Five-factor monthly data saved to database.")
```

```
Five-factor monthly data saved to database.
```

## 11.7 Daily Fama-French Factors

### 11.7.1 Motivation for Daily Factors

Daily factors are essential for several applications:

1. **Daily beta estimation**: CAPM regressions using daily data require daily market excess returns.
2. **Event studies**: Measuring abnormal returns around corporate events requires daily factor adjustments.
3. **High-frequency research**: Market microstructure studies need daily or intraday factor data.

The construction methodology mirrors the monthly approach, but we compute portfolio returns at daily frequency while maintaining the same annual portfolio formation dates.

### 11.7.2 Loading Daily Returns

```
# Load daily price data
prices_daily = pd.read_sql_query(
    sql="""
        SELECT symbol, date, ret_excess
        FROM prices_daily
    """,
    con=tidy_finance,
    parse_dates={"date"}
)

print(f"Daily returns: {len(prices_daily):,} observations")
print(f"Date range: {prices_daily['date'].min():%Y-%m-%d} to {prices_daily['date'].max():%Y-%
```

```
Daily returns: 3,462,157 observations
Date range: 2010-01-05 to 2023-12-29
```

### 11.7.3 Adding Market Cap for Daily Weighting

For value-weighted daily returns, we need market capitalization. We use the most recent monthly market cap as the weight for daily returns within that month.

```
# Get monthly market cap to use as weights for daily returns
mktcap_monthly = (prices_monthly
    [["symbol", "date", "mktcap_lag"]]
    .assign(year_month=lambda x: x["date"].dt.to_period("M"))
)

# Add year_month to daily data for merging
prices_daily = (prices_daily
    .assign(year_month=lambda x: x["date"].dt.to_period("M"))
    .merge(
        mktcap_monthly[["symbol", "year_month", "mktcap_lag"]],
        on=["symbol", "year_month"],
        how="left"
    )
    .dropna(subset=["ret_excess", "mktcap_lag"])
)

print(f"Daily returns with weights: {len(prices_daily):,} observations")
```

```
Daily returns with weights: 3,443,815 observations
```

### 11.7.4 Merging Daily Returns with Portfolios

We use the same portfolio assignments (formed annually in July) for daily returns.

```
# Step 1: Ensure portfolios_ff3 has correct format
portfolios_ff3_clean = portfolios_ff3[["symbol", "sorting_date", "portfolio_size", "portfolio
portfolios_ff3_clean["sorting_date"] = pd.to_datetime(portfolios_ff3_clean["sorting_date"])

print("Portfolio sorting dates:")
print(portfolios_ff3_clean["sorting_date"].unique()[:5])

# Step 2: Create sorting_date for daily data
prices_daily_with_sort = prices_daily.copy()
prices_daily_with_sort["sorting_date"] = prices_daily_with_sort["date"].apply(
    lambda x: pd.Timestamp(f"{x.year}-07-01") if x.month > 6 else pd.Timestamp(f"{x.year - 1
)

print("\nDaily sorting dates:")
print(prices_daily_with_sort["sorting_date"].unique()[:5])

# Step 3: Merge
portfolios_daily_ff3 = prices_daily_with_sort.merge(
    portfolios_ff3_clean,
    on=["symbol", "sorting_date"],
    how="inner"
)

print(f"\nMerged daily observations: {len(portfolios_daily_ff3):,}")
print(f"Unique dates: {portfolios_daily_ff3['date'].nunique():,}")

# Step 4: Verify portfolio distribution
print("\nPortfolio distribution in daily data:")
print(portfolios_daily_ff3.groupby(["portfolio_size", "portfolio_bm"], observed=True).size()
```

```
Portfolio sorting dates:
<DatetimeArray>
['2019-07-01 00:00:00', '2020-07-01 00:00:00', '2021-07-01 00:00:00',
 '2022-07-01 00:00:00', '2023-07-01 00:00:00']
Length: 5, dtype: datetime64[us]
```

```
Daily sorting dates:
<DatetimeArray>
['2018-07-01 00:00:00', '2019-07-01 00:00:00', '2020-07-01 00:00:00',
 '2021-07-01 00:00:00', '2022-07-01 00:00:00']
Length: 5, dtype: datetime64[us]

Merged daily observations: 2,843,570
Unique dates: 3,126

Portfolio distribution in daily data:
portfolio_bm         1       2       3
portfolio_size
1               218040  585561  617327
2               636152  552114  234376
```

```python
# Diagnostic: Check the daily portfolio merge
print("="*50)
print("DIAGNOSTIC: Daily Portfolio Merge")
print("="*50)

print(f"\nDaily prices rows: {len(prices_daily):,}")
print(f"Daily FF3 portfolios rows: {len(portfolios_daily_ff3):,}")
print(f"Match rate: {len(portfolios_daily_ff3)/len(prices_daily)*100:.1f}%")

# Check portfolio distribution in daily data
print("\nDaily portfolio distribution:")
print(portfolios_daily_ff3.groupby(["portfolio_size", "portfolio_bm"], observed=True).size()

# Check a specific date
test_date = portfolios_daily_ff3["date"].iloc[1000]
print(f"\nSample date: {test_date}")
print(portfolios_daily_ff3.query("date == @test_date").groupby(["portfolio_size", "portfolio_
```

```
==================================================
DIAGNOSTIC: Daily Portfolio Merge
==================================================

Daily prices rows: 3,443,815
Daily FF3 portfolios rows: 2,843,570
Match rate: 82.6%
```

```
Daily portfolio distribution:
portfolio_bm         1       2       3
portfolio_size
1               218040  585561  617327
2               636152  552114  234376

Sample date: 2023-06-28 00:00:00
portfolio_bm      1    2    3
portfolio_size
1                 93  232  312
2                291  280   67
```

### 11.7.5 Computing Daily Three Factors

```python
def compute_daily_ff3_factors(portfolios_daily):
    """
    Compute daily Fama-French three factors.

    Parameters
    ----------
    portfolios_daily : pd.DataFrame
        Daily returns with portfolio assignments

    Returns
    -------
    pd.DataFrame
        Daily SMB and HML factors
    """
    # Compute daily portfolio returns
    portfolio_returns = (portfolios_daily
        .groupby(["portfolio_size", "portfolio_bm", "date"], observed=True)
        .apply(lambda x: pd.Series({
            "ret": np.average(x["ret_excess"], weights=x["mktcap_lag"])
        }))
        .reset_index()
    )

    # Compute factors
    factors = (portfolio_returns
        .groupby("date")
        .apply(lambda x: pd.Series({
```

```
            "smb": (x.loc[x["portfolio_size"] == 1, "ret"].mean() -
                   x.loc[x["portfolio_size"] == 2, "ret"].mean()),
            "hml": (x.loc[x["portfolio_bm"] == 3, "ret"].mean() -
                   x.loc[x["portfolio_bm"] == 1, "ret"].mean())
        }))
        .reset_index()
    )

    return factors

factors_daily_smb_hml = compute_daily_ff3_factors(portfolios_daily_ff3)

print(f"Daily factor observations: {len(factors_daily_smb_hml):,}")
print(factors_daily_smb_hml.head(10))
```

```
Daily factor observations: 3,126
        date       smb       hml
0 2011-07-01  0.008587  0.000967
1 2011-07-04  0.005099 -0.001099
2 2011-07-05 -0.009088  0.010152
3 2011-07-06  0.004875 -0.003918
4 2011-07-07 -0.011239 -0.000584
5 2011-07-08  0.005636 -0.008003
6 2011-07-11  0.003940  0.006172
7 2011-07-12  0.003205  0.006543
8 2011-07-13 -0.000097 -0.001134
9 2011-07-14 -0.001248  0.001669
```

### 11.7.6 Computing Daily Market Factor

```
def compute_daily_market_factor(prices_daily):
    """
    Compute daily value-weighted market excess return.

    Parameters
    ----------
    prices_daily : pd.DataFrame
        Daily returns with mktcap_lag

    Returns
```

```
    -------
    pd.DataFrame
        Daily market excess return
    """
    market_daily = (prices_daily
        .groupby("date")
        .apply(lambda x: pd.Series({
            "mkt_excess": np.average(x["ret_excess"], weights=x["mktcap_lag"])
        }))
        .reset_index()
    )

    return market_daily

market_factor_daily = compute_daily_market_factor(prices_daily)

print(f"Daily market factor: {len(market_factor_daily):,} days")
```

```
Daily market factor: 3,474 days
```

### 11.7.7 Combining Daily Three Factors

```
factors_ff3_daily = (factors_daily_smb_hml
    .merge(market_factor_daily, on="date", how="inner")
)

# Add risk-free rate (use monthly rate / 21 as daily approximation, or load actual daily rate
factors_ff3_daily["risk_free"] = 0.04 / 252  # Approximate daily risk-free

print("Daily Fama-French Three Factors:")
print(factors_ff3_daily.head(10))

print("\nDaily Factor Summary Statistics:")
print(factors_ff3_daily[["mkt_excess", "smb", "hml"]].describe().round(6))
```

```
Daily Fama-French Three Factors:
        date       smb       hml  mkt_excess  risk_free
0 2011-07-01  0.008587  0.000967   -0.019862   0.000159
1 2011-07-04  0.005099 -0.001099   -0.000633   0.000159
2 2011-07-05 -0.009088  0.010152    0.013314   0.000159
```

```
3 2011-07-06  0.004875 -0.003918   -0.008045   0.000159
4 2011-07-07 -0.011239 -0.000584    0.003391   0.000159
5 2011-07-08  0.005636 -0.008003    0.000218   0.000159
6 2011-07-11  0.003940  0.006172   -0.013393   0.000159
7 2011-07-12  0.003205  0.006543   -0.017505   0.000159
8 2011-07-13 -0.000097 -0.001134    0.000767   0.000159
9 2011-07-14 -0.001248  0.001669   -0.000695   0.000159

Daily Factor Summary Statistics:
        mkt_excess          smb          hml
count  3126.000000  3126.000000  3126.000000
mean     -0.000479     0.000236     0.000594
std       0.011269     0.008488     0.008585
min      -0.070268    -0.032671    -0.039418
25%      -0.005074    -0.004882    -0.003941
50%       0.000350    -0.000106     0.000522
75%       0.005531     0.004307     0.005233
max       0.043386     0.042686     0.083889
```

### 11.7.8 Computing Daily Five Factors

```python
# Step 1: Clean portfolios
portfolios_ff5_clean = portfolios_ff5[["symbol", "sorting_date", "portfolio_size",
                                        "portfolio_bm", "portfolio_op", "portfolio_inv"]].co
portfolios_ff5_clean["sorting_date"] = pd.to_datetime(portfolios_ff5_clean["sorting_date"])

# Step 2: Merge with daily prices
portfolios_daily_ff5 = prices_daily_with_sort.merge(
    portfolios_ff5_clean,
    on=["symbol", "sorting_date"],
    how="inner"
)

print(f"FF5 Daily merged observations: {len(portfolios_daily_ff5):,}")
```

```
FF5 Daily merged observations: 2,843,570
```

```python
def compute_daily_ff5_factors(portfolios_daily):
    """Compute daily Fama-French five factors."""
```

```python
# HML from B/M sorts
portfolios_bm = (portfolios_daily
    .groupby(["portfolio_size", "portfolio_bm", "date"], observed=True)
    .apply(lambda x: pd.Series({
        "ret": np.average(x["ret_excess"], weights=x["mktcap_lag"])
    }))
    .reset_index()
)

factors_hml = (portfolios_bm
    .groupby("date")
    .apply(lambda x: pd.Series({
        "hml": (x.loc[x["portfolio_bm"] == 3, "ret"].mean() -
                x.loc[x["portfolio_bm"] == 1, "ret"].mean())
    }))
    .reset_index()
)

# RMW from OP sorts
portfolios_op = (portfolios_daily
    .groupby(["portfolio_size", "portfolio_op", "date"], observed=True)
    .apply(lambda x: pd.Series({
        "ret": np.average(x["ret_excess"], weights=x["mktcap_lag"])
    }))
    .reset_index()
)

factors_rmw = (portfolios_op
    .groupby("date")
    .apply(lambda x: pd.Series({
        "rmw": (x.loc[x["portfolio_op"] == 3, "ret"].mean() -
                x.loc[x["portfolio_op"] == 1, "ret"].mean())
    }))
    .reset_index()
)

# CMA from INV sorts (note: low minus high)
portfolios_inv = (portfolios_daily
    .groupby(["portfolio_size", "portfolio_inv", "date"], observed=True)
    .apply(lambda x: pd.Series({
        "ret": np.average(x["ret_excess"], weights=x["mktcap_lag"])
    }))
```

```
        .reset_index()
    )

    factors_cma = (portfolios_inv
        .groupby("date")
        .apply(lambda x: pd.Series({
            "cma": (x.loc[x["portfolio_inv"] == 1, "ret"].mean() -
                   x.loc[x["portfolio_inv"] == 3, "ret"].mean())
        }))
        .reset_index()
    )

    # SMB from all sorts
    all_portfolios = pd.concat([portfolios_bm, portfolios_op, portfolios_inv])

    factors_smb = (all_portfolios
        .groupby("date")
        .apply(lambda x: pd.Series({
            "smb": (x.loc[x["portfolio_size"] == 1, "ret"].mean() -
                   x.loc[x["portfolio_size"] == 2, "ret"].mean())
        }))
        .reset_index()
    )

    # Combine
    factors = (factors_smb
        .merge(factors_hml, on="date", how="outer")
        .merge(factors_rmw, on="date", how="outer")
        .merge(factors_cma, on="date", how="outer")
    )

    return factors

# Compute daily FF5 factors
factors_daily_ff5 = compute_daily_ff5_factors(portfolios_daily_ff5)

# Add market factor
factors_ff5_daily = (factors_daily_ff5
    .merge(market_factor_daily, on="date", how="inner")
)
factors_ff5_daily["risk_free"] = 0.04 / 252
```

```
print("Daily Fama-French Five Factors:")
print(factors_ff5_daily.head(10))

print("\nDaily Five-Factor Summary Statistics:")
print(factors_ff5_daily[["mkt_excess", "smb", "hml", "rmw", "cma"]].describe().round(6))
```

```
Daily Fama-French Five Factors:
        date       smb       hml       rmw       cma  mkt_excess  risk_free
0 2011-07-01  0.006295  0.002515  0.013140  0.007680   -0.019862   0.000159
1 2011-07-04  0.002880 -0.002875  0.006560 -0.004886   -0.000633   0.000159
2 2011-07-05 -0.004260  0.009864 -0.012158 -0.004470    0.013314   0.000159
3 2011-07-06  0.001544 -0.009847  0.012977  0.006286   -0.008045   0.000159
4 2011-07-07 -0.009789 -0.003988 -0.000197 -0.006995    0.003391   0.000159
5 2011-07-08  0.001537 -0.006700  0.010841 -0.007661    0.000218   0.000159
6 2011-07-11  0.005396  0.004747  0.000655  0.013375   -0.013393   0.000159
7 2011-07-12  0.004759  0.007367  0.001989  0.014669   -0.017505   0.000159
8 2011-07-13 -0.000009  0.001110 -0.002052 -0.001633    0.000767   0.000159
9 2011-07-14 -0.001668  0.002916 -0.005427  0.005388   -0.000695   0.000159

Daily Five-Factor Summary Statistics:
        mkt_excess          smb          hml          rmw          cma
count  3126.000000  3126.000000  3126.000000  3126.000000  3126.000000
mean     -0.000479     0.000484     0.000549    -0.000136     0.000413
std       0.011269     0.008033     0.008312     0.008538     0.006756
min      -0.070268    -0.036283    -0.039155    -0.154013    -0.047698
25%      -0.005074    -0.004122    -0.003681    -0.004212    -0.003364
50%       0.000350     0.000105     0.000384     0.000030     0.000162
75%       0.005531     0.004358     0.004819     0.004036     0.003800
max       0.043386     0.060307     0.086269     0.102001     0.089907
```

### 11.7.9 Saving Daily Factors

```
factors_ff3_daily.to_sql(
    name="factors_ff3_daily",
    con=tidy_finance,
    if_exists="replace",
    index=False
)

factors_ff5_daily.to_sql(
```

```
    name="factors_ff5_daily",
    con=tidy_finance,
    if_exists="replace",
    index=False
)

print("Daily factor data saved to database.")
```

```
Daily factor data saved to database.
```

## 11.8 Factor Validation and Diagnostics

```
# Verify all tables are in database
print("\n" + "="*50)
print("DATABASE SUMMARY")
print("="*50)

tables = ["factors_ff3_monthly", "factors_ff5_monthly",
          "factors_ff3_daily", "factors_ff5_daily"]

for table in tables:
    df = pd.read_sql_query(f"SELECT COUNT(*) as n FROM {table}", con=tidy_finance)
    print(f"{table}: {df['n'].iloc[0]:,} observations")

# Correlation check: Monthly vs Daily (aggregated)
print("\n" + "="*50)
print("MONTHLY VS DAILY CONSISTENCY CHECK")
print("="*50)

factors_daily_agg = (factors_ff3_daily
    .assign(year_month=lambda x: x["date"].dt.to_period("M"))
    .groupby("year_month")[["mkt_excess", "smb", "hml"]]
    .sum()
    .reset_index()
)

factors_monthly_check = (factors_ff3_monthly
    .assign(year_month=lambda x: x["date"].dt.to_period("M"))
)
```

```
comparison = factors_monthly_check.merge(
    factors_daily_agg, on="year_month", suffixes=("_monthly", "_daily")
)

for factor in ["mkt_excess", "smb", "hml"]:
    corr = comparison[f"{factor}_monthly"].corr(comparison[f"{factor}_daily"])
    print(f"{factor}: Monthly-Daily correlation = {corr:.4f}")
```

```
==================================================
DATABASE SUMMARY
==================================================
factors_ff3_monthly: 150 observations
factors_ff5_monthly: 150 observations
factors_ff3_daily: 3,126 observations
factors_ff5_daily: 3,126 observations

==================================================
MONTHLY VS DAILY CONSISTENCY CHECK
==================================================
mkt_excess: Monthly-Daily correlation = 0.9980
smb: Monthly-Daily correlation = 0.9953
hml: Monthly-Daily correlation = 0.9936
```

### 11.8.1 Cumulative Factor Returns

We visualize the cumulative performance of each factor to assess whether the factors generate meaningful premiums over time.

```
# Compute cumulative returns
factors_cumulative = (factors_ff5_monthly
    .set_index("date")
    [["mkt_excess", "smb", "hml", "rmw", "cma"]]
    .add(1)
    .cumprod()
)

# Plot
fig, ax = plt.subplots(figsize=(12, 6))
factors_cumulative.plot(ax=ax)
```

```
ax.set_title("Cumulative Factor Returns (Vietnam)")
ax.set_xlabel("")
ax.set_ylabel("Growth of $1")
ax.legend(title="Factor")
ax.axhline(y=1, color='gray', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()
```

Figure 11.1: Cumulative returns of Fama-French factors for the Vietnamese market. The figure shows the growth of $1 invested in each factor portfolio.

### 11.8.2 Average Factor Premiums

We compute annualized average factor premiums and their statistical significance.

```
# Annualized average returns (monthly returns * 12)
factor_premiums = (factors_ff5_monthly
    [["mkt_excess", "smb", "hml", "rmw", "cma"]]
    .mean() * 12 * 100  # Annualized percentage
)

# Standard errors
factor_se = (factors_ff5_monthly
    [["mkt_excess", "smb", "hml", "rmw", "cma"]]
```

```
    .std() / np.sqrt(len(factors_ff5_monthly)) * np.sqrt(12) * 100
)

# T-statistics
factor_tstat = factor_premiums / factor_se

print("Annualized Factor Premiums (%):")
print(factor_premiums.round(2))

print("\nT-Statistics:")
print(factor_tstat.round(2))
```

```
Annualized Factor Premiums (%):
mkt_excess   -12.09
smb            9.19
hml           13.75
rmw           -5.69
cma            9.94
dtype: float64

T-Statistics:
mkt_excess    -7.30
smb            7.76
hml            9.38
rmw           -4.22
cma           10.49
dtype: float64
```

### 11.8.3 Comparing Monthly and Daily Factors

We verify consistency between monthly and daily factors by computing correlations.

```
# Aggregate daily factors to monthly for comparison
factors_daily_monthly = (factors_ff5_daily
    .assign(year_month=lambda x: x["date"].dt.to_period("M"))
    .groupby("year_month")
    [["mkt_excess", "smb", "hml", "rmw", "cma"]]
    .sum()  # Sum daily returns to get monthly
    .reset_index()
)
```

```python
# Merge with actual monthly factors
comparison = (factors_ff5_monthly
    .assign(year_month=lambda x: x["date"].dt.to_period("M"))
    .merge(
        factors_daily_monthly,
        on="year_month",
        suffixes=("_monthly", "_daily")
    )
)

# Correlations
for factor in ["mkt_excess", "smb", "hml", "rmw", "cma"]:
    corr = comparison[f"{factor}_monthly"].corr(comparison[f"{factor}_daily"])
    print(f"{factor}: Monthly-Daily correlation = {corr:.4f}")
```

```
mkt_excess: Monthly-Daily correlation = 0.9980
smb: Monthly-Daily correlation = 0.9950
hml: Monthly-Daily correlation = 0.9948
rmw: Monthly-Daily correlation = 0.9929
cma: Monthly-Daily correlation = 0.9884
```

## 11.9 Key Takeaways

1. **Factor Models Explained**: The Fama-French three-factor model adds size (SMB) and value (HML) factors to the CAPM, while the five-factor model further includes profitability (RMW) and investment (CMA) factors.
2. **Construction Methodology**: Factors are constructed through double-sorted portfolios with careful attention to timing. Portfolios are formed in July using accounting data from the prior fiscal year to ensure information was publicly available.
3. **Independent vs. Dependent Sorts**: The three-factor model uses independent sorts on size and book-to-market, creating a 2×3 grid. The five-factor model uses dependent sorts where characteristics are sorted within size groups.
4. **Value-Weighted Returns**: Portfolio returns are computed using value-weighting with lagged market capitalization to avoid look-ahead bias.
5. **Daily Factors**: Daily factors use the same annual portfolio assignments but compute returns at daily frequency, enabling higher-frequency applications like daily beta estimation.

6. **Market Factor**: The market factor is computed independently as the value-weighted return of all stocks minus the risk-free rate.
7. **Validation**: Factor quality can be assessed through characteristic monotonicity, portfolio diversification, cumulative returns, and consistency between daily and monthly frequencies.
8. **Vietnamese Market Adaptation**: While following the original Fama-French methodology, we adapt for Vietnamese market characteristics including VAS accounting standards, reporting timelines, and currency units.

# 12 Momentum Strategies

Momentum is one of the most robust and pervasive anomalies in financial economics. Stocks that have performed well over the past three to twelve months tend to continue performing well over the subsequent three to twelve months, and stocks that have performed poorly tend to continue underperforming. This pattern, first documented by Jegadeesh and Titman (1993), has been replicated across virtually every equity market, asset class, and time period examined, earning it a central place in the canon of empirical asset pricing.

In this chapter, we provide an implementation of momentum strategies following the methodology of Jegadeesh and Titman (1993). We construct momentum portfolios based on past cumulative returns, evaluate their performance across different formation and holding period combinations, and examine whether the momentum premium exists in the Vietnamese equity market..

## 12.1 Theoretical Background

### 12.1.1 The Momentum Effect

The momentum effect refers to the tendency of assets with high recent returns to continue generating high returns, and assets with low recent returns to continue generating low returns. More formally, if we define the cumulative return of stock $i$ over the past $J$ months as

$$R_{i,t-J:t-1} = \prod_{s=t-J}^{t-1} (1 + r_{i,s}) - 1 \tag{12.1}$$

then momentum predicts a positive cross-sectional relationship between $R_{i,t-J:t-1}$ and future returns $r_{i,t}$. That is, stocks in the top decile of past returns (winners) should outperform stocks in the bottom decile (losers) in subsequent months.

The Jegadeesh and Titman (1993) framework parameterizes momentum strategies by two key dimensions:

1. **Formation period** ($J$): The number of months over which past returns are computed to rank stocks. Typical values range from 3 to 12 months.

2. **Holding period** ($K$): The number of months for which the momentum portfolios are held after formation. Typical values also range from 3 to 12 months.

A strategy is therefore characterized by the pair $(J, K)$. For example, a $(6, 6)$ strategy ranks stocks based on their cumulative returns over the past 6 months and holds the resulting portfolios for 6 months.

### 12.1.2 Overlapping Portfolios and Calendar-Time Returns

A critical implementation detail in Jegadeesh and Titman (1993) is the use of overlapping portfolios. In each month $t$, a new set of portfolios is formed based on the most recent $J$-month returns. However, portfolios formed in months $t-1$, $t-2$, ..., $t-K+1$ are still within their holding periods and therefore remain active. The return of the momentum strategy in month $t$ is thus the equally weighted average across all $K$ active cohorts:

$$r_{p,t} = \frac{1}{K} \sum_{k=0}^{K-1} r_{p,t}^{(t-k)} \tag{12.2}$$

where $r_{p,t}^{(t-k)}$ denotes the return in month $t$ of the portfolio formed in month $t-k$. This overlapping portfolio approach serves two purposes. First, it reduces the impact of any single formation date on the strategy's performance. Second, it produces a monthly return series that can be analyzed using standard time-series methods, even when the holding period $K$ exceeds one month.

### 12.1.3 Theoretical Explanations

The academic literature offers two broad classes of explanations for the momentum effect.

1. **Behavioral explanations** attribute momentum to systematic cognitive biases among investors. Daniel, Hirshleifer, and Subrahmanyam (1998) propose a model based on investor overconfidence and biased self-attribution: investors overweight private signals and attribute confirming outcomes to their own skill, leading to initial underreaction followed by delayed overreaction. Hong and Stein (1999) develop a model in which information diffuses gradually across heterogeneous investors, generating underreaction to firm-specific news. Barberis, Shleifer, and Vishny (1998) formalize a model combining conservatism bias (slow updating of beliefs in response to new evidence) and representativeness heuristic (extrapolation of recent trends).
2. **Risk-based explanations** argue that momentum profits represent compensation for systematic risk that varies over time. Johnson (2002) show that momentum can arise in a rational framework if expected returns are stochastic and time-varying. Grundy and Martin (2001) document that momentum portfolios have substantial time-varying factor

exposures, suggesting that at least part of the momentum premium reflects dynamic risk. However, standard risk models such as the Fama-French three-factor model have generally struggled to explain momentum returns, which motivated the development of explicit momentum factors such as the Winners-Minus-Losers (WML) factor in Mark M. Carhart (1997a).

### 12.1.4 Momentum in Emerging Markets

The behavior of momentum in emerging markets differs substantially from developed markets, and understanding these differences is essential for interpreting our Vietnamese market results.

Rouwenhorst (1998) provides early evidence that momentum profits exist in European markets. Chan, Hameed, and Tong (2000) extend the analysis to international equity markets and find that momentum profits are present in most developed markets but are weaker or absent in several Asian markets. Chui, Titman, and Wei (2010) offer a cultural explanation, documenting that momentum profits are positively related to a country's degree of individualism as measured by Hofstede (2001). Countries with collectivist cultures, including many in East and Southeast Asia, tend to exhibit weaker momentum effects.

Several market microstructure features common in emerging markets may attenuate momentum:

- **Trading band limits** constrain daily price movements, potentially slowing the adjustment process that generates momentum. Vietnam's HOSE imposes a $\pm 7\%$ daily limit, while HNX allows $\pm 10\%$.
- **Lower liquidity and higher transaction costs** can erode momentum profits, as documented by Lesmond, Schill, and Zhou (2004).
- **Foreign ownership limits** segment the investor base, potentially altering the information diffusion dynamics that underlie momentum.
- **Shorter market history** limits the statistical power available to detect the effect.

These considerations motivate our careful empirical analysis of whether, and to what extent, momentum manifests in Vietnamese equities.

## 12.2 Setting Up the Environment

We begin by loading the necessary Python packages. The core packages include `pandas` for data manipulation, `numpy` for numerical operations, and `sqlite3` for database connectivity. We also import `plotnine` for creating publication-quality figures and `scipy` for statistical tests.

```
import pandas as pd
import numpy as np
import sqlite3
from datetime import datetime
from itertools import product

from plotnine import *
from mizani.formatters import comma_format, percent_format
from scipy import stats
```

We connect to our SQLite database, which stores the cleaned datasets prepared in Accessing and Managing Financial Data and DataCore Data.

```
tidy_finance = sqlite3.connect(
    database="data/tidy_finance_python.sqlite"
)
```

## 12.3 Data Preparation

### 12.3.1 Loading Monthly Stock Returns

We load the monthly stock price data from our database. The `prices_monthly` table contains adjusted returns, market capitalizations, and other variables for all stocks listed on HOSE and HNX.

```
prices_monthly = pd.read_sql_query(
    # can add "exchange" variable
    sql="""
        SELECT symbol, date, ret, ret_excess, mktcap, mktcap_lag, risk_free
        FROM prices_monthly
    """,
    con=tidy_finance,
    parse_dates={"date"}
).dropna()

print(f"Total monthly observations: {len(prices_monthly):,}")
print(f"Unique stocks: {prices_monthly['symbol'].nunique():,}")
print(f"Date range: {prices_monthly['date'].min().date()} "
      f"to {prices_monthly['date'].max().date()}")
```

```
Total monthly observations: 165,499
Unique stocks: 1,457
Date range: 2010-02-28 to 2023-12-31
```

### 12.3.2 Inspecting the Data

Before proceeding with portfolio construction, we examine the key variables in our dataset. Table 12.1 presents the summary statistics for the main variables used in momentum portfolio construction.

```
summary_stats = (prices_monthly
    [["ret", "ret_excess", "mktcap"]]
    .describe()
    .T
    .round(4)
)
summary_stats
```

Table 12.1: Summary statistics for monthly stock return data. The table reports the count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum for monthly returns (ret), excess returns (ret_excess), and market capitalization (mktcap, in billion VND).

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| ret | 165499.0 | 0.0042 | 0.1862 | -0.9900 | -0.0703 | 0.0000 | 0.0553 | 12.7500 |
| ret_excess | 165499.0 | 0.0008 | 0.1862 | -0.9933 | -0.0736 | -0.0033 | 0.0519 | 12.7467 |
| mktcap | 165499.0 | 2183.1646 | 13983.9977 | 0.3536 | 60.3728 | 180.6224 | 660.0000 | 463886.6454 |

We also examine the cross-sectional distribution of stocks over time, which is important for understanding whether we have sufficient breadth for decile portfolio construction.

```
stock_counts = (prices_monthly
    .groupby("date")["symbol"]
    .nunique()
    .reset_index()
    .rename(columns={"symbol": "n_stocks"})
)

plot_stock_counts = (
    ggplot(stock_counts, aes(x="date", y="n_stocks")) +
```

```
    geom_line(color="#1f77b4") +
    labs(
        x="", y="Number of stocks",
        title="Cross-Sectional Breadth Over Time"
    ) +
    theme_minimal() +
    theme(figure_size=(10, 5))
)
plot_stock_counts
```

Figure 12.1: Number of stocks with available return data over time. The figure shows the monthly cross-section of stocks available for momentum portfolio construction. A minimum number of stocks is needed to form meaningful decile portfolios.

### 12.3.3 Loading Factor Data

We also load the Fama-French factor returns for risk-adjusted performance evaluation. These factors were constructed in Fama-French Factors.

```
factors_ff3_monthly = pd.read_sql_query(
    sql="SELECT date, mkt_excess, smb, hml, risk_free FROM factors_ff3_monthly",
    con=tidy_finance,
    parse_dates={"date"}
```

```
)

print(f"Factor observations: {len(factors_ff3_monthly):,}")
print(f"Date range: {factors_ff3_monthly['date'].min().date()} "
      f"to {factors_ff3_monthly['date'].max().date()}")
```

```
Factor observations: 150
Date range: 2011-07-31 to 2023-12-31
```

### 12.3.4 Data Quality Filters

Momentum strategies require continuous return histories for the formation period. We apply several filters to ensure data quality:

1. **Minimum price filter**: We exclude penny stocks with prices below 1,000 VND, which are subject to extreme microstructure noise.
2. **Return availability**: We require non-missing returns for the entire formation period.
3. **Market capitalization**: We require positive lagged market capitalization for portfolio weighting.

```
# Filter for stocks with positive market cap and non-missing returns
prices_clean = (prices_monthly
    .dropna(subset=["ret", "mktcap"])
    .query("mktcap > 0")
    .sort_values(["symbol", "date"])
    .reset_index(drop=True)
)

print(f"Observations after filtering: {len(prices_clean):,}")
print(f"Stocks after filtering: {prices_clean['symbol'].nunique():,}")
```

```
Observations after filtering: 165,499
Stocks after filtering: 1,457
```

## 12.4 Momentum Portfolio Construction

### 12.4.1 Computing Formation Period Returns

The first step in constructing momentum portfolios is to compute cumulative returns over the formation period. For a formation period of $J$ months, we compute the cumulative return for

each stock $i$ at the end of month $t$ as specified in Equation 12.1.

We implement this using a rolling product of gross returns. A key detail is that we require non-missing returns for all $J$ months in the formation window. If any monthly return is missing, the cumulative return is set to missing, and the stock is excluded from portfolio formation in that month.

```
def compute_formation_returns(data, J):
    """
    Compute J-month cumulative returns for momentum portfolio formation.

    Parameters
    ----------
    data : pd.DataFrame
        Panel data with columns 'symbol', 'date', 'ret'.
    J : int
        Formation period length in months (typically 3-12).

    Returns
    -------
    pd.DataFrame
        Original data augmented with 'cum_return' column.
    """
    df = data.sort_values(["symbol", "date"]).copy()

    # Compute rolling J-month cumulative return
    # Using gross returns: (1+r1)*(1+r2)*...*(1+rJ) - 1
    df["gross_ret"] = 1 + df["ret"]

    df["cum_return"] = (
        df.groupby("symbol")["gross_ret"]
        .rolling(window=J, min_periods=J)
        .apply(np.prod, raw=True)
        .reset_index(level=0, drop=True)
        - 1
    )

    df = df.drop(columns=["gross_ret"])

    return df
```

Let us apply this function for our baseline case with $J = 6$ months:

```python
J = 6  # Formation period: 6 months

prices_with_cumret = compute_formation_returns(prices_clean, J)

# Check the result
print(f"Observations with valid {J}-month cumulative returns: "
      f"{prices_with_cumret['cum_return'].notna().sum():,}")
print(f"\nCumulative return distribution:")
print(prices_with_cumret["cum_return"].describe().round(4))
```

```
Observations with valid 6-month cumulative returns: 158,227

Cumulative return distribution:
count    158227.0000
mean          0.0171
std           0.5053
min          -0.9999
25%          -0.2196
50%          -0.0400
75%           0.1404
max          35.7136
Name: cum_return, dtype: float64
```

### 12.4.2 Assigning Momentum Decile Portfolios

Each month, we sort all stocks with valid cumulative returns into decile portfolios based on their past $J$-month performance. Portfolio 1 contains the bottom decile (losers) and portfolio 10 contains the top decile (winners).

```python
def assign_momentum_portfolios(data, n_portfolios=10):
    """
    Assign stocks to momentum portfolios based on formation-period returns.

    For each cross-section (month), stocks are sorted into
    n_portfolios quantile groups according to their cumulative return.

    Portfolio 1 = lowest past returns (Losers)
    Portfolio n_portfolios = highest past returns (Winners)

    This implementation uses `groupby().transform()` rather than
    `groupby().apply()` to preserve the original DataFrame structure
```

```
    and avoid index mutation issues.

    Parameters
    ----------
    data : pd.DataFrame
        Panel data containing at minimum:
        - 'symbol' : stock identifier
        - 'date' : formation month
        - 'cum_return' : cumulative return over formation window

    n_portfolios : int, optional
        Number of portfolios to form (default = 10 for deciles).

    Returns
    -------
    pd.DataFrame
        Original data augmented with:
        - 'momr' : integer momentum rank (1 to n_portfolios)
    """

    # Drop observations without formation-period returns
    # These cannot be ranked into portfolios
    df = data.dropna(subset=["cum_return"]).copy()

    def safe_qcut(x):
        """
        Assign quantile portfolio labels within a single cross-section.

        Uses pd.qcut to create approximately equal-sized portfolios.
        If too few unique return values exist (which can happen in
        small markets or illiquid samples), fall back to rank-based
        binning to ensure portfolios are still formed.
        """
        try:
            # Standard quantile sorting
            return pd.qcut(
                x,
                q=n_portfolios,
                labels=range(1, n_portfolios + 1),
                duplicates="drop"  # prevents crash if quantile edges duplicate
            )
        except ValueError:
```

```
            # Fallback: rank then evenly cut into bins
            return pd.cut(
                x.rank(method="first"),
                bins=n_portfolios,
                labels=range(1, n_portfolios + 1)
            )

    # Cross-sectional portfolio assignment:
    # For each month, compute portfolio ranks based on cum_return.
    # transform ensures:
    # - original row order preserved
    # - no index mutation
    # - no loss of 'date' column
    df["momr"] = (
        df.groupby("date")["cum_return"]
          .transform(safe_qcut)
          .astype(int)
    )

    return df
```

```
portfolios = assign_momentum_portfolios(prices_with_cumret)

print(f"Stocks assigned to portfolios: {len(portfolios):,}")
print(f"\nPortfolio distribution (should be approximately equal):")
print(portfolios.groupby("momr")["symbol"].count().to_frame("count"))
```

```
Stocks assigned to portfolios: 158,227

Portfolio distribution (should be approximately equal):
      count
momr
1     15894
2     15815
3     16021
4     15759
5     15738
6     16568
7     15288
8     15582
9     15697
10    15865
```

### 12.4.3 Defining Holding Period Dates

After forming portfolios at the end of month $t$, we hold them from the beginning of month $t+1$ through the end of month $t+K$. This one-month gap between the formation period and the start of the holding period is standard in the literature and avoids the well-documented short-term reversal effect at the one-month horizon (Jegadeesh 1990).

```
K = 6  # Holding period: 6 months

portfolios_with_dates = portfolios.copy()
portfolios_with_dates = portfolios_with_dates.rename(
    columns={"date": "form_date"}
)

# Define holding period start and end dates
portfolios_with_dates["hdate1"] = (
    portfolios_with_dates["form_date"] + pd.offsets.MonthBegin(1)
)
portfolios_with_dates["hdate2"] = (
    portfolios_with_dates["form_date"] + pd.offsets.MonthEnd(K)
)

print(portfolios_with_dates[
    ["symbol", "form_date", "cum_return", "momr", "hdate1", "hdate2"]
].head(10))
```

```
   symbol  form_date  cum_return  momr     hdate1     hdate2
5     A32 2019-04-30   -0.061599     4 2019-05-01 2019-10-31
6     A32 2019-05-31   -0.213675     2 2019-06-01 2019-11-30
7     A32 2019-06-30   -0.308720     2 2019-07-01 2019-12-31
8     A32 2019-07-31   -0.249936     2 2019-08-01 2020-01-31
9     A32 2019-08-31    0.061986     8 2019-09-01 2020-02-29
10    A32 2019-09-30    0.030751     8 2019-10-01 2020-03-31
11    A32 2019-10-31    0.030751     8 2019-11-01 2020-04-30
12    A32 2019-11-30    0.032486     8 2019-12-01 2020-05-31
13    A32 2019-12-31    0.180073     9 2020-01-01 2020-06-30
14    A32 2020-01-31    0.412322    10 2020-02-01 2020-07-31
```

### 12.4.4 Computing Portfolio Holding Period Returns

We now compute the returns of each portfolio during its holding period. For each stock-formation date combination, we merge in the actual returns realized during the holding window.

This produces a panel of stock returns indexed by formation date, holding date, and momentum portfolio rank.

```
# Merge: for each portfolio assignment, get all returns during holding period
portfolio_returns = portfolios_with_dates.merge(
    prices_clean[["symbol", "date", "ret"]].rename(
        columns={"ret": "hret", "date": "hdate"}
    ),
    on="symbol",
    how="inner"
)

# Keep only returns within the holding period
portfolio_returns = portfolio_returns.query(
    "hdate >= hdate1 and hdate <= hdate2"
)

print(f"Portfolio-holding period observations: {len(portfolio_returns):,}")
```

```
Portfolio-holding period observations: 918,072
```

### 12.4.5 Computing Equally Weighted Portfolio Returns

The Jegadeesh and Titman (1993) methodology uses equally weighted portfolios. For each holding month, each momentum decile has $K$ active cohorts (one formed in each of the past $K$ months). We first compute the equally weighted return within each cohort, then average across cohorts to get the final monthly portfolio return.

```
# Step 1: Average return within each cohort (form_date × momr × hdate)
cohort_returns = (portfolio_returns
    .groupby(["hdate", "momr", "form_date"])
    .agg(cohort_ret=("hret", "mean"))
    .reset_index()
)

# Step 2: Average across cohorts for each momentum portfolio × month
ewret = (cohort_returns
    .groupby(["hdate", "momr"])
    .agg(
        ewret=("cohort_ret", "mean"),
        ewret_std=("cohort_ret", "std"),
```

```
        n_cohorts=("cohort_ret", "count")
    )
    .reset_index()
    .rename(columns={"hdate": "date"})
)

print(f"Monthly portfolio return observations: {len(ewret):,}")
print(f"Date range: {ewret['date'].min().date()} to {ewret['date'].max().date()}")
```

```
Monthly portfolio return observations: 1,610
Date range: 2010-08-31 to 2023-12-31
```

We should verify that we have the expected number of active cohorts per month. Once the strategy has been running for at least $K$ months, each momentum decile should have exactly $K$ active cohorts.

```
cohort_check = (ewret
    .groupby("date")["n_cohorts"]
    .mean()
    .reset_index()
    .rename(columns={"n_cohorts": "avg_cohorts"})
)

plot_cohorts = (
    ggplot(cohort_check, aes(x="date", y="avg_cohorts")) +
    geom_line(color="#1f77b4") +
    geom_hline(yintercept=K, linetype="dashed", color="red") +
    labs(
        x="", y="Average number of cohorts",
        title=f"Active Cohorts per Portfolio (K={K})"
    ) +
    theme_minimal() +
    theme(figure_size=(10, 4))
)
plot_cohorts
```

Figure 12.2: Number of active cohorts per momentum portfolio over time. After an initial ramp-up period of K months, each portfolio should have exactly K active cohorts contributing to its monthly return.

## 12.5 Baseline Results: J=6, K=6 Strategy

### 12.5.1 Summary Statistics by Momentum Decile

We now examine the average monthly returns for each momentum decile portfolio. Table 12.2 presents the mean, $t$-statistic, and $p$-value for each of the ten portfolios.

```
def compute_portfolio_stats(group):
    """Compute mean, t-stat, and p-value for a return series."""
    n = len(group)
    mean_ret = group.mean()
    std_ret = group.std()
    t_stat = mean_ret / (std_ret / np.sqrt(n)) if std_ret > 0 else np.nan
    p_val = 2 * (1 - stats.t.cdf(abs(t_stat), df=n-1)) if not np.isnan(t_stat) else np.nan
    return pd.Series({
        "N": n,
        "Mean (%)": mean_ret * 100,
        "Std (%)": std_ret * 100,
        "t-stat": t_stat,
        "p-value": p_val
    })

momentum_stats = (ewret
```

```
    .groupby("momr")["ewret"]
    .apply(compute_portfolio_stats)
    .unstack()
    .round(4)
)

momentum_stats
```

Table 12.2: Average monthly returns of momentum decile portfolios for the J=6, K=6 strategy. Portfolio 1 contains past losers and portfolio 10 contains past winners. The table reports the number of months, mean monthly return, t-statistic, and p-value for each decile.

| momr | N | Mean (%) | Std (%) | t-stat | p-value |
|---|---|---|---|---|---|
| 1 | 161.0 | 1.4190 | 6.5319 | 2.7565 | 0.0065 |
| 2 | 161.0 | 0.6602 | 5.9895 | 1.3985 | 0.1639 |
| 3 | 161.0 | 0.3658 | 5.5080 | 0.8427 | 0.4007 |
| 4 | 161.0 | 0.1623 | 5.0085 | 0.4112 | 0.6815 |
| 5 | 161.0 | 0.0111 | 4.7561 | 0.0297 | 0.9763 |
| 6 | 161.0 | 0.0408 | 4.6864 | 0.1104 | 0.9122 |
| 7 | 161.0 | -0.0674 | 4.8881 | -0.1749 | 0.8614 |
| 8 | 161.0 | -0.2066 | 4.9208 | -0.5329 | 0.5949 |
| 9 | 161.0 | -0.2228 | 5.3013 | -0.5332 | 0.5946 |
| 10 | 161.0 | -0.6812 | 5.7131 | -1.5130 | 0.1323 |

### 12.5.2 The Long-Short Momentum Portfolio

The key test of the momentum effect is whether the spread between winners and losers—the long-short momentum portfolio—generates statistically significant positive returns. We construct this spread portfolio by going long the top decile (portfolio 10) and short the bottom decile (portfolio 1).

```
# Pivot to wide format
ewret_wide = ewret.pivot(
    index="date", columns="momr", values="ewret"
).reset_index()

ewret_wide.columns = ["date"] + [f"port{i}" for i in range(1, 11)]
```

```
# Compute long-short return
ewret_wide["winners"] = ewret_wide["port10"]
ewret_wide["losers"] = ewret_wide["port1"]
ewret_wide["long_short"] = ewret_wide["winners"] - ewret_wide["losers"]
```

```
ls_stats = pd.DataFrame()
for col in ["winners", "losers", "long_short"]:
    series = ewret_wide[col].dropna()
    n = len(series)
    mean_val = series.mean()
    std_val = series.std()
    t_val = mean_val / (std_val / np.sqrt(n))
    p_val = 2 * (1 - stats.t.cdf(abs(t_val), df=n-1))
    ls_stats[col] = [n, mean_val * 100, std_val * 100, t_val, p_val]

ls_stats.index = ["N", "Mean (%)", "Std (%)", "t-stat", "p-value"]
ls_stats.columns = ["Winners", "Losers", "Long-Short"]
ls_stats = ls_stats.round(4)
ls_stats
```

Table 12.3: Performance of the momentum long-short strategy (J=6, K=6). Winners is the top decile, Losers is the bottom decile, and Long-Short is Winners minus Losers. The table reports the number of months, mean monthly return, t-statistic, and p-value.

| | Winners | Losers | Long-Short |
|---|---|---|---|
| N | 161.0000 | 161.0000 | 161.0000 |
| Mean (%) | -0.6812 | 1.4190 | -2.1002 |
| Std (%) | 5.7131 | 6.5319 | 4.8969 |
| t-stat | -1.5130 | 2.7565 | -5.4420 |
| p-value | 0.1323 | 0.0065 | 0.0000 |

### 12.5.3 Cumulative Returns

The time-series evolution of cumulative returns provides visual evidence of the momentum strategy's performance. Figure 12.3 shows the cumulative returns of the winner and loser portfolios, while Figure 12.4 shows the cumulative return of the long-short momentum strategy.

```
# Compute cumulative returns
ewret_wide = ewret_wide.sort_values("date").reset_index(drop=True)
```

```
for col in ["winners", "losers", "long_short"]:
    ewret_wide[f"cumret_{col}"] = (1 + ewret_wide[col]).cumprod() - 1
```

```
cumret_plot_data = (ewret_wide
    [["date", "cumret_winners", "cumret_losers"]]
    .melt(id_vars="date", var_name="portfolio", value_name="cumret")
)
cumret_plot_data["portfolio"] = cumret_plot_data["portfolio"].map({
    "cumret_winners": "Winners (P10)",
    "cumret_losers": "Losers (P1)"
})

plot_cumret = (
    ggplot(cumret_plot_data,
           aes(x="date", y="cumret", color="portfolio")) +
    geom_line(size=1) +
    scale_y_continuous(labels=percent_format()) +
    scale_color_manual(values=["#d62728", "#1f77b4"]) +
    labs(
        x="", y="Cumulative return",
        color="Portfolio",
        title="Cumulative Returns: Winners vs. Losers"
    ) +
    theme_minimal() +
    theme(
        figure_size=(10, 6),
        legend_position="bottom"
    )
)
plot_cumret
```

Figure 12.3: Cumulative returns of momentum winner and loser portfolios. The winner portfolio (blue) contains stocks in the top decile of past 6-month returns, and the loser portfolio (red) contains stocks in the bottom decile. The spread between the two lines reflects the cumulative momentum premium.

```
plot_ls = (
    ggplot(ewret_wide, aes(x="date", y="cumret_long_short")) +
    geom_line(size=1, color="#2ca02c") +
    geom_hline(yintercept=0, linetype="dashed", color="gray") +
    scale_y_continuous(labels=percent_format()) +
    labs(
        x="", y="Cumulative return",
        title="Cumulative Return: Long-Short Momentum Strategy"
    ) +
    theme_minimal() +
    theme(figure_size=(10, 6))
)
plot_ls
```

Figure 12.4: Cumulative return of the momentum long-short strategy (Winners minus Losers). Upward-sloping periods indicate momentum profits; sharp drawdowns often coincide with market reversals or crisis episodes.

### 12.5.4 Monthly Return Distribution

Beyond the mean, the full distribution of monthly long-short returns provides insight into the risk profile of the momentum strategy. Figure 12.5 shows the histogram of monthly returns.

```
mean_ls = ewret_wide["long_short"].mean()

plot_dist = (
    ggplot(ewret_wide, aes(x="long_short")) +
    geom_histogram(bins=50, fill="#1f77b4", alpha=0.7) +
    geom_vline(xintercept=mean_ls, linetype="dashed", color="red") +
    scale_x_continuous(labels=percent_format()) +
    labs(
        x="Monthly long-short return",
        y="Frequency",
        title="Distribution of Monthly Momentum Returns"
    ) +
```

```
    theme_minimal() +
    theme(figure_size=(10, 5))
)
plot_dist
```

Figure 12.5: Distribution of monthly long-short momentum returns. The histogram shows the frequency distribution of monthly Winners-minus-Losers returns. The vertical dashed line indicates the mean monthly return.

## 12.6 Extending to Multiple Formation and Holding Periods

### 12.6.1 The J × K Grid

Jegadeesh and Titman (1993) evaluate momentum strategies across a comprehensive grid of formation periods $J \in \{3, 6, 9, 12\}$ and holding periods $K \in \{3, 6, 9, 12\}$. This produces 16 different strategy specifications, allowing us to assess the robustness of the momentum effect across different horizons.

We now implement a function that computes the full momentum strategy for any given $(J, K)$ pair, wrapping the steps developed above into a single reusable pipeline.

```python
from joblib import Parallel, delayed
import time

def momentum_strategy(data, J, K, n_portfolios=10):
    """
    Implement the full Jegadeesh-Titman momentum strategy.

    Parameters
    ----------
    data : pd.DataFrame
        Panel of stock returns with columns: symbol, date, ret.
    J : int
        Formation period in months.
    K : int
        Holding period in months.
    n_portfolios : int
        Number of portfolios (default: 10 for deciles).

    Returns
    -------
    pd.DataFrame
        Monthly returns for each momentum portfolio and the long-short spread.
    """
    # Step 1: Compute formation period cumulative returns
    df = data.sort_values(["symbol", "date"]).copy()
    df["gross_ret"] = 1 + df["ret"]
    df["cum_return"] = (
        df.groupby("symbol")["gross_ret"]
        .rolling(window=J, min_periods=J)
        .apply(np.prod, raw=True)
        .reset_index(level=0, drop=True)
        - 1
    )
    df = df.drop(columns=["gross_ret"]).dropna(subset=["cum_return"])

    # Step 2: Assign to momentum portfolios using transform
    df["momr"] = df.groupby("date", observed=True)["cum_return"].transform(
        lambda x: pd.qcut(
            x,
            q=n_portfolios,
            labels=range(1, n_portfolios + 1),
            duplicates="drop"
```

```
        )
    ).astype('Int64')

    # If any NaNs in momr, fill with rank-based assignment
    mask = df["momr"].isna()
    if mask.any():
        df.loc[mask, "momr"] = df.loc[mask].groupby("date")["cum_return"].transform(
            lambda x: pd.qcut(
                x.rank(method="first"),
                q=min(n_portfolios, len(x.unique())),
                labels=False,
                duplicates="drop"
            )
        ).astype('Int64') + 1

    # Step 3: Define holding period dates
    df = df.rename(columns={"date": "form_date"})
    df["hdate1"] = df["form_date"] + pd.offsets.MonthBegin(1)
    df["hdate2"] = df["form_date"] + pd.offsets.MonthEnd(K)

    # Step 4: Merge with holding period returns
    port_ret = df[["symbol", "form_date", "momr", "hdate1", "hdate2"]].merge(
        data[["symbol", "date", "ret"]].rename(
            columns={"ret": "hret", "date": "hdate"}
        ),
        on="symbol",
        how="inner"
    )

    # Use boolean indexing
    port_ret = port_ret[
        (port_ret["hdate"] >= port_ret["hdate1"]) &
        (port_ret["hdate"] <= port_ret["hdate2"])
    ]

    # Step 5: Compute equally weighted returns (two-stage averaging)
    cohort_ret = (port_ret
        .groupby(["hdate", "momr", "form_date"])
        .agg(cohort_ret=("hret", "mean"))
        .reset_index()
    )
```

```
    monthly_ret = (cohort_ret
        .groupby(["hdate", "momr"])
        .agg(ewret=("cohort_ret", "mean"))
        .reset_index()
        .rename(columns={"hdate": "date"})
    )

    # Step 6: Compute long-short spread
    wide = monthly_ret.pivot(
        index="date", columns="momr", values="ewret"
    ).reset_index()

    # Handle variable number of portfolios
    n_cols = len(wide.columns) - 1
    wide.columns = ["date"] + [f"port{i}" for i in range(1, n_cols + 1)]
    wide["winners"] = wide[f"port{n_cols}"]
    wide["losers"] = wide["port1"]
    wide["long_short"] = wide["winners"] - wide["losers"]

    return monthly_ret, wide
```

### 12.6.2 Running the Full Grid

We now run the momentum strategy for all 16 $(J, K)$ combinations. This is computationally intensive, so we store the key summary statistics for each specification.

#### 12.6.2.1 Sequential Calculation

```
J_values = [3, 6, 9, 12]
K_values = [3, 6, 9, 12]

results_grid = []

for J_val, K_val in product(J_values, K_values):
    print(f"Computing J={J_val}, K={K_val}...", end=" ")

    try:
        _, wide_result = momentum_strategy(prices_clean, J_val, K_val)
```

```
        for portfolio_name in ["winners", "losers", "long_short"]:
            series = wide_result[portfolio_name].dropna()
            n = len(series)
            mean_ret = series.mean()
            std_ret = series.std()
            t_stat = mean_ret / (std_ret / np.sqrt(n)) if std_ret > 0 else np.nan
            p_val = (2 * (1 - stats.t.cdf(abs(t_stat), df=n-1))
                     if not np.isnan(t_stat) else np.nan)

            results_grid.append({
                "J": J_val,
                "K": K_val,
                "portfolio": portfolio_name,
                "n_months": n,
                "mean_ret": mean_ret,
                "std_ret": std_ret,
                "t_stat": t_stat,
                "p_value": p_val
            })

        print("Done.")
    except Exception as e:
        print(f"Error: {e}")

results_df = pd.DataFrame(results_grid)
```

#### 12.6.2.2 Parallel Calculation

```
def compute_single_strategy(data, J, K):
    """
    Compute statistics for a single (J, K) strategy.
    Returns a list of result dicts, one per portfolio.
    """
    try:
        _, wide_result = momentum_strategy(data, J, K)

        results = []
        for portfolio_name in ["winners", "losers", "long_short"]:
            series = wide_result[portfolio_name].dropna()
            n = len(series)
```

```
            mean_ret = series.mean()
            std_ret = series.std()
            t_stat = mean_ret / (std_ret / np.sqrt(n)) if std_ret > 0 else np.nan
            p_val = (2 * (1 - stats.t.cdf(abs(t_stat), df=n-1))
                     if not np.isnan(t_stat) else np.nan)

            results.append({
                "J": J,
                "K": K,
                "portfolio": portfolio_name,
                "n_months": n,
                "mean_ret": mean_ret,
                "std_ret": std_ret,
                "t_stat": t_stat,
                "p_value": p_val
            })
        return results
    except Exception as e:
        print(f"Error in J={J}, K={K}: {e}")
        return []
```

```
J_values = [3, 6, 9, 12]
K_values = [3, 6, 9, 12]

# Create list of (J, K) pairs
params = list(product(J_values, K_values))

print(f"Running {len(params)} momentum strategies in parallel with 4 cores...")
start_time = time.time()

# Parallel execution with 4 workers
results_list = Parallel(n_jobs=4, verbose=10)(
    delayed(compute_single_strategy)(prices_clean, J, K)
    for J, K in params
)

# Flatten results
results_grid = [item for sublist in results_list for item in sublist]
results_df = pd.DataFrame(results_grid)

elapsed = time.time() - start_time
print(f"\nCompleted in {elapsed:.2f} seconds")
```

```
Running 16 momentum strategies in parallel with 4 cores...

[Parallel(n_jobs=4)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=4)]: Done   5 tasks      | elapsed:    8.8s
[Parallel(n_jobs=4)]: Done  11 out of  16 | elapsed:   15.5s remaining:    7.0s
[Parallel(n_jobs=4)]: Done  13 out of  16 | elapsed:   19.2s remaining:    4.4s


Completed in 21.04 seconds

[Parallel(n_jobs=4)]: Done  16 out of  16 | elapsed:   20.9s finished
```

```python
# Summary statistics
print("\nResults Summary by Formation Period:")
print(results_df.groupby("J").agg({
    "mean_ret": "mean",
    "std_ret": "mean",
    "t_stat": ["mean", lambda x: (x.abs() > 1.96).sum()],
    "n_months": "first"
}).round(6))
```

```
Results Summary by Formation Period:
    mean_ret   std_ret    t_stat            n_months
        mean      mean      mean <lambda_0>    first
J
6  -0.004661  0.055672 -1.517507          9      161
9  -0.003265  0.056520 -1.071904          9      158
12 -0.002759  0.058194 -0.931273          9      155
```

### 12.6.3 Winners Portfolio Returns

Table 12.4 presents the average monthly returns of the winner portfolio (portfolio 10) across all $(J, K)$ combinations.

```python
winners_grid = (results_df
    .query("portfolio == 'winners'")
    .assign(
        display=lambda x: (
            x["mean_ret"].apply(lambda v: f"{v*100:.2f}") +
```

```
            "\n(" + x["t_stat"].apply(lambda v: f"{v:.2f}") + ")"
        )
    )
    .pivot(index="J", columns="K", values="display")
)
winners_grid.columns = [f"K={k}" for k in winners_grid.columns]
winners_grid.index = [f"J={j}" for j in winners_grid.index]
winners_grid
```

Table 12.4: Average monthly returns (%) of the winner portfolio across formation periods (J) and holding periods (K). t-statistics are reported in parentheses.

| | K=3 | K=6 | K=9 | K=12 |
|---|---|---|---|---|
| J=6 | -1.18\n(-2.65) | -0.68\n(-1.51) | -0.55\n(-1.23) | -0.39\n(-0.87) |
| J=9 | -1.00\n(-2.36) | -0.51\n(-1.22) | -0.28\n(-0.68) | -0.16\n(-0.39) |
| J=12 | -0.88\n(-2.09) | -0.39\n(-0.91) | -0.24\n(-0.56) | -0.14\n(-0.34) |

### 12.6.4 Losers Portfolio Returns

Table 12.5 presents the corresponding results for the loser portfolio.

```
losers_grid = (results_df
    .query("portfolio == 'losers'")
    .assign(
        display=lambda x: (
            x["mean_ret"].apply(lambda v: f"{v*100:.2f}") +
            "\n(" + x["t_stat"].apply(lambda v: f"{v:.2f}") + ")"
        )
    )
    .pivot(index="J", columns="K", values="display")
)
losers_grid.columns = [f"K={k}" for k in losers_grid.columns]
losers_grid.index = [f"J={j}" for j in losers_grid.index]
losers_grid
```

Table 12.5: Average monthly returns (%) of the loser portfolio across formation periods (J) and holding periods (K). t-statistics are reported in parentheses.

| | K=3 | K=6 | K=9 | K=12 |
|---|---|---|---|---|
| J=6 | 1.87\n(3.49) | 1.42\n(2.76) | 1.10\n(2.19) | 0.99\n(1.99) |
| J=9 | 1.94\n(3.51) | 1.38\n(2.56) | 1.21\n(2.31) | 1.15\n(2.22) |
| J=12 | 1.57\n(2.71) | 1.29\n(2.28) | 1.19\n(2.13) | 1.15\n(2.09) |

### 12.6.5 Long-Short Momentum Returns

Table 12.6 presents the most important results: the average monthly returns of the long-short momentum portfolio across all specifications.

```
ls_grid = (results_df
    .query("portfolio == 'long_short'")
    .assign(
        display=lambda x: (
            x["mean_ret"].apply(lambda v: f"{v*100:.2f}") +
            "\n(" + x["t_stat"].apply(lambda v: f"{v:.2f}") + ")"
        )
    )
    .pivot(index="J", columns="K", values="display")
)
ls_grid.columns = [f"K={k}" for k in ls_grid.columns]
ls_grid.index = [f"J={j}" for j in ls_grid.index]
ls_grid
```

Table 12.6: Average monthly returns (%) of the momentum long-short portfolio (Winners minus Losers) across formation periods (J) and holding periods (K). t-statistics are reported in parentheses. This table replicates the format of Table 1 in Jegadeesh and Titman (1993).

| | K=3 | K=6 | K=9 | K=12 |
|---|---|---|---|---|
| J=6 | -3.04\n(-7.38) | -2.10\n(-5.44) | -1.65\n(-4.94) | -1.37\n(-4.61) |
| J=9 | -2.94\n(-6.41) | -1.89\n(-4.55) | -1.49\n(-3.98) | -1.31\n(-3.87) |
| J=12 | -2.45\n(-5.42) | -1.68\n(-3.92) | -1.43\n(-3.60) | -1.30\n(-3.55) |

### 12.6.6 Visualizing the Momentum Premium Across Specifications

Figure 12.6 provides a visual summary of the long-short momentum premium across all $(J, K)$ combinations.

```
ls_heatmap_data = (results_df
    .query("portfolio == 'long_short'")
    .assign(
        mean_pct=lambda x: x["mean_ret"] * 100,
        significant=lambda x: x["p_value"] < 0.05,
        label=lambda x: x.apply(
            lambda row: f"{row['mean_ret']*100:.2f}{'*' if row['p_value'] < 0.05 else ''}",
            axis=1
        )
    )
)

plot_heatmap = (
    ggplot(ls_heatmap_data,
           aes(x="K.astype(str)", y="J.astype(str)", fill="mean_pct")) +
    geom_tile(color="white", size=2) +
    geom_text(aes(label="label"), size=10, color="white") +
    scale_fill_gradient2(
        low="#d62728", mid="#f7f7f7", high="#1f77b4", midpoint=0,
        name="Monthly\nReturn (%)"
    ) +
    labs(
        x="Holding Period (K months)",
        y="Formation Period (J months)",
        title="Momentum Premium Across J×K Specifications"
    ) +
    theme_minimal() +
    theme(figure_size=(8, 6))
)
plot_heatmap
```

Figure 12.6: Heatmap of average monthly long-short momentum returns (%) across formation periods (J) and holding periods (K). Darker shades indicate higher momentum profits. Asterisks denote statistical significance at the 5% level.

## 12.7 Risk-Adjusted Performance

### 12.7.1 CAPM Alpha

A natural question is whether the momentum premium is explained by exposure to the market factor. We estimate the CAPM alpha of the long-short momentum portfolio by regressing its returns on the market excess return:

$$r_{\text{WML},t} = \alpha + \beta \cdot r_{\text{MKT},t} + \epsilon_t \tag{12.3}$$

```python
# Merge momentum returns with factor data (baseline J=6, K=6)
ewret_factors = (ewret_wide
    [["date", "winners", "losers", "long_short"]]
    .merge(factors_ff3_monthly, on="date", how="inner")
)

# CAPM regression for long-short portfolio
from statsmodels.formula.api import ols as ols_formula
import statsmodels.api as sm

X_capm = sm.add_constant(ewret_factors["mkt_excess"])
y_ls = ewret_factors["long_short"]

capm_model = sm.OLS(y_ls, X_capm).fit(cov_type="HAC", cov_kwds={"maxlags": 6})

print("CAPM Regression: Long-Short Momentum Returns")
print("=" * 60)
print(f"Alpha (monthly):  {capm_model.params['const']*100:.4f}% "
      f"(t={capm_model.tvalues['const']:.2f})")
print(f"Market Beta:      {capm_model.params['mkt_excess']:.4f} "
      f"(t={capm_model.tvalues['mkt_excess']:.2f})")
print(f"R-squared:        {capm_model.rsquared:.4f}")
print(f"N observations:   {capm_model.nobs:.0f}")
```

```
CAPM Regression: Long-Short Momentum Returns
============================================================
Alpha (monthly):  -2.2703% (t=-5.05)
Market Beta:      -0.1779 (t=-1.75)
R-squared:        0.0470
N observations:   150
```

### 12.7.2 Fama-French Three-Factor Alpha

We extend the risk adjustment to the Fama-French three-factor model, which includes the size (SMB) and value (HML) factors in addition to the market factor:

$$r_{\text{WML},t} = \alpha + \beta_1 \cdot r_{\text{MKT},t} + \beta_2 \cdot \text{SMB}_t + \beta_3 \cdot \text{HML}_t + \epsilon_t \tag{12.4}$$

```python
X_ff3 = sm.add_constant(
    ewret_factors[["mkt_excess", "smb", "hml"]]
)
y_ls = ewret_factors["long_short"]

ff3_model = sm.OLS(y_ls, X_ff3).fit(cov_type="HAC", cov_kwds={"maxlags": 6})

# Display results as a clean table
ff3_results = pd.DataFrame({
    "Coefficient": ff3_model.params,
    "Std Error": ff3_model.bse,
    "t-stat": ff3_model.tvalues,
    "p-value": ff3_model.pvalues
}).round(4)

ff3_results.index = ["Alpha", "MKT", "SMB", "HML"]
ff3_results
```

Table 12.7: Fama-French three-factor regression for the momentum long-short portfolio. The dependent variable is the monthly return of the Winners-minus-Losers portfolio. Alpha is the intercept, representing the risk-adjusted abnormal return. HAC standard errors with 6 lags are used.

| | Coefficient | Std Error | t-stat | p-value |
|---|---|---|---|---|
| Alpha | -0.0234 | 0.0041 | -5.7189 | 0.0000 |
| MKT | -0.1269 | 0.1558 | -0.8145 | 0.4154 |
| SMB | 0.1123 | 0.1882 | 0.5969 | 0.5506 |
| HML | 0.0716 | 0.1414 | 0.5065 | 0.6125 |

```python
print(f"\nR-squared: {ff3_model.rsquared:.4f}")
print(f"Adjusted R-squared: {ff3_model.rsquared_adj:.4f}")
print(f"Alpha (annualized): {ff3_model.params['const'] * 12 * 100:.2f}%")
```

```
R-squared: 0.0553
Adjusted R-squared: 0.0359
Alpha (annualized): -28.02%
```

### 12.7.3 Interpretation of Risk Exposures

The factor loadings from the three-factor regression reveal the risk characteristics of the momentum strategy in the Vietnamese market. Several patterns are commonly observed:

1. **Market beta**: Momentum portfolios typically have moderate market exposure. In the U.S., Grundy and Martin (2001) document that the market beta of the long-short portfolio is close to zero on average but highly time-varying, spiking during market reversals.
2. **Size exposure (SMB)**: Momentum strategies often load positively on the size factor, reflecting the tendency for smaller stocks to exhibit stronger momentum patterns.
3. **Value exposure (HML)**: The long-short momentum portfolio typically loads negatively on HML, indicating that winners tend to be growth stocks while losers tend to be value stocks. This creates a natural tension between momentum and value strategies.

## 12.8 Momentum and Market States

### 12.8.1 Conditional Performance

An important finding in the momentum literature is that momentum profits vary with market conditions. Cooper, Gutierrez Jr, and Hameed (2004) document that momentum strategies perform well following market gains (UP markets) but experience severe losses following market declines (DOWN markets). This asymmetry is particularly relevant for emerging markets, which experience more extreme market states.

We define market states based on the cumulative market return over the prior 12 months:

$$\text{Market State}_t = \begin{cases} \text{UP} & \text{if } \prod_{s=t-12}^{t-1}(1+r_{m,s}) - 1 > 0 \\ \text{DOWN} & \text{otherwise} \end{cases} \tag{12.5}$$

```
# Compute 12-month lagged market return
market_returns = factors_ff3_monthly[["date", "mkt_excess"]].copy()
market_returns = market_returns.sort_values("date")
market_returns["mkt_cum_12m"] = (
    (1 + market_returns["mkt_excess"])
    .rolling(window=12, min_periods=12)
    .apply(np.prod, raw=True)
    - 1
)
market_returns["market_state"] = np.where(
```

```
    market_returns["mkt_cum_12m"] > 0, "UP", "DOWN"
)

# Merge with momentum returns
ewret_states = ewret_wide[["date", "long_short"]].merge(
    market_returns[["date", "market_state"]], on="date", how="inner"
).dropna()
```

```
state_stats = []
for state in ["UP", "DOWN"]:
    subset = ewret_states.query(f"market_state == '{state}'")["long_short"]
    n = len(subset)
    mean_ret = subset.mean()
    std_ret = subset.std()
    t_stat = mean_ret / (std_ret / np.sqrt(n)) if std_ret > 0 else np.nan
    p_val = 2 * (1 - stats.t.cdf(abs(t_stat), df=n-1)) if not np.isnan(t_stat) else np.nan
    state_stats.append({
        "Market State": state,
        "N Months": n,
        "Mean (%)": mean_ret * 100,
        "Std (%)": std_ret * 100,
        "t-stat": t_stat,
        "p-value": p_val
    })

state_stats_df = pd.DataFrame(state_stats).round(4)
state_stats_df
```

Table 12.8: Momentum long-short returns conditional on market states. UP markets are defined as periods where the cumulative market return over the prior 12 months is positive; DOWN markets are periods with negative prior 12-month returns.

| | Market State | N Months | Mean (%) | Std (%) | t-stat | p-value |
|---|---|---|---|---|---|---|
| 0 | UP | 39 | -1.1972 | 4.5814 | -1.6320 | 0.1109 |
| 1 | DOWN | 111 | -2.4050 | 4.8669 | -5.2063 | 0.0000 |

```
plot_states = (
    ggplot(ewret_states, aes(x="market_state", y="long_short",
                             fill="market_state")) +
    geom_boxplot(alpha=0.7) +
```

```
    geom_hline(yintercept=0, linetype="dashed", color="gray") +
    scale_y_continuous(labels=percent_format()) +
    scale_fill_manual(values={"UP": "#1f77b4", "DOWN": "#d62728"}) +
    labs(
        x="Market State (Prior 12-Month Return)",
        y="Monthly Long-Short Return",
        title="Momentum Returns by Market State"
    ) +
    theme_minimal() +
    theme(
        figure_size=(8, 6),
        legend_position="none"
    )
)
plot_states
```

Figure 12.7: Distribution of monthly momentum returns by market state. The box plots show the distribution of long-short momentum returns separately for UP and DOWN market states, defined by the sign of the prior 12-month cumulative market return.

## 12.9 Momentum Crashes

### 12.9.1 Understanding Momentum Drawdowns

One of the most important risk characteristics of momentum strategies is their susceptibility to sudden, severe losses—known as momentum crashes. Daniel and Moskowitz (2016) document that momentum strategies experience infrequent but extreme losses, typically during market rebounds following bear markets. These crashes occur because the loser portfolio, which has been short, is heavily loaded with high-beta stocks that surge when markets reverse.

We identify the worst drawdowns of the momentum strategy and examine their market context.

```
worst_months = (ewret_wide
    [["date", "long_short", "winners", "losers"]]
    .merge(factors_ff3_monthly[["date", "mkt_excess"]], on="date", how="left")
    .sort_values("long_short")
    .head(10)
    .assign(
        long_short_pct=lambda x: (x["long_short"] * 100).round(2),
        winners_pct=lambda x: (x["winners"] * 100).round(2),
        losers_pct=lambda x: (x["losers"] * 100).round(2),
        mkt_pct=lambda x: (x["mkt_excess"] * 100).round(2)
    )
    [["date", "long_short_pct", "winners_pct", "losers_pct", "mkt_pct"]]
    .rename(columns={
        "date": "Date",
        "long_short_pct": "L/S (%)",
        "winners_pct": "Winners (%)",
        "losers_pct": "Losers (%)",
        "mkt_pct": "Market (%)"
    })
)
worst_months
```

Table 12.9: Ten worst months for the momentum long-short strategy. The table reports the date, long-short return, winner return, loser return, and concurrent market return for the ten months with the largest momentum losses.

| | Date | L/S (%) | Winners (%) | Losers (%) | Market (%) |
|---|---|---|---|---|---|
| 153 | 2023-05-31 | -16.39 | -2.69 | 13.70 | 2.78 |
| 39 | 2013-11-30 | -14.26 | 4.24 | 18.50 | 2.57 |
| 20 | 2012-04-30 | -13.68 | 1.51 | 15.19 | 4.88 |
| 78 | 2017-02-28 | -13.56 | 2.10 | 15.67 | 1.49 |
| 107 | 2019-07-31 | -13.15 | -2.46 | 10.69 | 1.61 |
| 140 | 2022-04-30 | -11.71 | -17.57 | -5.87 | -11.19 |
| 149 | 2023-01-31 | -11.53 | 1.37 | 12.90 | 6.96 |
| 28 | 2012-12-31 | -11.52 | 4.03 | 15.54 | -1.61 |
| 0 | 2010-08-31 | -11.31 | -25.03 | -13.72 | NaN |
| 10 | 2011-06-30 | -10.44 | -3.15 | 7.29 | NaN |

### 12.9.2 Maximum Drawdown Analysis

The maximum drawdown provides a measure of the worst peak-to-trough decline experienced by the strategy. This metric is particularly relevant for practitioners evaluating the risk of momentum strategies.

```
# Compute running maximum and drawdown
ewret_wide["cum_wealth"] = (1 + ewret_wide["long_short"]).cumprod()
ewret_wide["running_max"] = ewret_wide["cum_wealth"].cummax()
ewret_wide["drawdown"] = (
    ewret_wide["cum_wealth"] / ewret_wide["running_max"] - 1
)

max_dd = ewret_wide["drawdown"].min()
max_dd_date = ewret_wide.loc[ewret_wide["drawdown"].idxmin(), "date"]

print(f"Maximum drawdown: {max_dd*100:.2f}%")
print(f"Date of maximum drawdown: {max_dd_date.date()}")
```

```
Maximum drawdown: -97.08%
Date of maximum drawdown: 2023-10-31
```

```
plot_dd = (
    ggplot(ewret_wide, aes(x="date", y="drawdown")) +
    geom_area(fill="#d62728", alpha=0.5) +
    geom_line(color="#d62728", size=0.5) +
    scale_y_continuous(labels=percent_format()) +
    labs(
        x="", y="Drawdown",
        title="Momentum Strategy Drawdown"
    ) +
    theme_minimal() +
    theme(figure_size=(10, 5))
)
plot_dd
```

Figure 12.8: Drawdown of the momentum long-short strategy over time. The chart shows the percentage decline from the previous peak in cumulative wealth. Deeper drawdowns represent more severe momentum crashes.

## 12.10 Value-Weighted Momentum Portfolios

The baseline Jegadeesh and Titman (1993) implementation uses equally weighted portfolios. However, equally weighted returns can be dominated by small, illiquid stocks that may be difficult to trade in practice. Value-weighted portfolios, where each stock's contribution is proportional to its market capitalization, provide a more investable benchmark and are more representative of the returns that large investors could actually achieve.

```
def momentum_strategy_vw(data, J, K, n_portfolios=10):
    """
    Value-weighted momentum strategy implementation.

    Same as the equally-weighted version but uses lagged market
    capitalization as weights when computing portfolio returns.

    Parameters
    ----------
    data : pd.DataFrame
        Panel with columns: symbol, date, ret, mktcap_lag.
    J : int
```

```
        Formation period in months.
    K : int
        Holding period in months.
    n_portfolios : int
        Number of portfolios.

    Returns
    -------
    pd.DataFrame
        Monthly value-weighted portfolio returns.
    """
    # Step 1: Formation period returns
    df = data.sort_values(["symbol", "date"]).copy()
    df["gross_ret"] = 1 + df["ret"]
    df["cum_return"] = (
        df.groupby("symbol")["gross_ret"]
        .rolling(window=J, min_periods=J)
        .apply(np.prod, raw=True)
        .reset_index(level=0, drop=True)
        - 1
    )
    df = df.drop(columns=["gross_ret"]).dropna(subset=["cum_return"])

    # Step 2: Portfolio assignment using transform (fast)
    df["momr"] = df.groupby("date", observed=True)["cum_return"].transform(
        lambda x: pd.qcut(
            x,
            q=n_portfolios,
            labels=range(1, n_portfolios + 1),
            duplicates="drop"
        )
    ).astype('Int64')

    # Fill NaNs with rank-based assignment
    mask = df["momr"].isna()
    if mask.any():
        df.loc[mask, "momr"] = df.loc[mask].groupby("date")["cum_return"].transform(
            lambda x: pd.qcut(
                x.rank(method="first"),
                q=min(n_portfolios, len(x.unique())),
                labels=False,
                duplicates="drop"
```

```
        )
    ).astype('Int64') + 1

# Step 3: Holding period
df = df.rename(columns={"date": "form_date"})
df["hdate1"] = df["form_date"] + pd.offsets.MonthBegin(1)
df["hdate2"] = df["form_date"] + pd.offsets.MonthEnd(K)

# Step 4: Merge with holding period returns AND weights
port_ret = df[
    ["symbol", "form_date", "momr", "hdate1", "hdate2"]
].merge(
    data[["symbol", "date", "ret", "mktcap_lag"]].rename(
        columns={"ret": "hret", "date": "hdate"}
    ),
    on="symbol",
    how="inner"
)

# Use boolean indexing instead of query (faster)
port_ret = port_ret[
    (port_ret["hdate"] >= port_ret["hdate1"]) &
    (port_ret["hdate"] <= port_ret["hdate2"])
]
port_ret = port_ret.dropna(subset=["mktcap_lag"])
port_ret = port_ret[port_ret["mktcap_lag"] > 0]

# Step 5: Value-weighted returns within each cohort
def vw_mean(group):
    weights = group["mktcap_lag"]
    if weights.sum() == 0:
        return np.nan
    return np.average(group["hret"], weights=weights)

cohort_ret = (port_ret
    .groupby(["hdate", "momr", "form_date"])
    .apply(vw_mean, include_groups=False)
    .reset_index(name="cohort_ret")
)

monthly_ret = (cohort_ret
    .groupby(["hdate", "momr"])
```

```
        .agg(vwret=("cohort_ret", "mean"))
        .reset_index()
        .rename(columns={"hdate": "date"})
    )

    # Step 6: Long-short
    wide = monthly_ret.pivot(
        index="date", columns="momr", values="vwret"
    ).reset_index()

    # Handle variable number of portfolios
    n_cols = len(wide.columns) - 1
    wide.columns = ["date"] + [f"port{i}" for i in range(1, n_cols + 1)]
    wide["winners"] = wide[f"port{n_cols}"]
    wide["losers"] = wide["port1"]
    wide["long_short"] = wide["winners"] - wide["losers"]

    return monthly_ret, wide
```

```
# Run value-weighted J=6, K=6 strategy
_, vw_results = momentum_strategy_vw(prices_clean, J=6, K=6)

print("Value-Weighted Momentum Strategy (J=6, K=6)")
print("=" * 50)
print("\nPortfolio Statistics:")
print(vw_results[["winners", "losers", "long_short"]].describe().round(4))
```

```
Value-Weighted Momentum Strategy (J=6, K=6)
==================================================

Portfolio Statistics:
        winners    losers  long_short
count  161.0000  161.0000    161.0000
mean    -0.0129    0.0006     -0.0135
std      0.0763    0.0713      0.0653
min     -0.2519   -0.2178     -0.3813
25%     -0.0540   -0.0397     -0.0465
50%     -0.0067    0.0010     -0.0126
75%      0.0354    0.0443      0.0173
max      0.1765    0.2498      0.2498
```

```
comparison = []
for scheme, df in [("EW", ewret_wide), ("VW", vw_results)]:
    for col in ["winners", "losers", "long_short"]:
        series = df[col].dropna()
        n = len(series)
        mean_ret = series.mean()
        std_ret = series.std()
        t_stat = mean_ret / (std_ret / np.sqrt(n))
        p_val = 2 * (1 - stats.t.cdf(abs(t_stat), df=n-1))
        comparison.append({
            "Weighting": scheme,
            "Portfolio": col.replace("_", " ").title(),
            "Mean (%)": round(mean_ret * 100, 4),
            "Std (%)": round(std_ret * 100, 4),
            "t-stat": round(t_stat, 2),
            "p-value": round(p_val, 4)
        })

pd.DataFrame(comparison)
```

Table 12.10: Comparison of equally weighted (EW) and value-weighted (VW) momentum strategies for J=6, K=6. The table reports mean monthly returns, t-statistics, and p-values for winners, losers, and long-short portfolios under both weighting schemes.

| | Weighting | Portfolio | Mean (%) | Std (%) | t-stat | p-value |
|---|---|---|---|---|---|---|
| 0 | EW | Winners | -0.6812 | 5.7131 | -1.51 | 0.1323 |
| 1 | EW | Losers | 1.4190 | 6.5319 | 2.76 | 0.0065 |
| 2 | EW | Long Short | -2.1002 | 4.8969 | -5.44 | 0.0000 |
| 3 | VW | Winners | -1.2943 | 7.6274 | -2.15 | 0.0328 |
| 4 | VW | Losers | 0.0589 | 7.1275 | 0.10 | 0.9166 |
| 5 | VW | Long Short | -1.3533 | 6.5312 | -2.63 | 0.0094 |

```
vw_results = vw_results.sort_values("date")
vw_results["cumret_ls_vw"] = (1 + vw_results["long_short"]).cumprod() - 1

ew_data = (ewret_wide[["date", "long_short"]]
    .rename(columns={"long_short": "cumret"})
    .assign(scheme="Equally Weighted")
)
ew_data["cumret"] = (1 + ew_data["cumret"]).cumprod() - 1
```

```
vw_data = (vw_results[["date", "cumret_ls_vw"]]
    .rename(columns={"cumret_ls_vw": "cumret"})
    .assign(scheme="Value Weighted")
)

ew_vs_vw = pd.concat([ew_data, vw_data], ignore_index=True)

plot_ew_vw = (
    ggplot(ew_vs_vw, aes(x="date", y="cumret", color="scheme")) +
    geom_line(size=1) +
    scale_y_continuous(labels=percent_format()) +
    scale_color_manual(values=["#1f77b4", "#ff7f0e"]) +
    labs(
        x="", y="Cumulative return",
        color="Weighting Scheme",
        title="EW vs. VW Momentum Long-Short Strategy"
    ) +
    theme_minimal() +
    theme(
        figure_size=(10, 6),
        legend_position="bottom"
    )
)
plot_ew_vw
```

Figure 12.9: Cumulative returns of equally weighted (EW) versus value-weighted (VW) long-short momentum strategies. Differences between the two lines reflect the impact of firm size on momentum profitability.

## 12.11 Daily Momentum Analysis

While momentum strategies are typically evaluated at the monthly frequency following Jegadeesh and Titman (1993), analyzing daily return patterns provides additional insights into the dynamics of momentum profits. Daily data allows us to examine how momentum profits accrue within the holding period, measure intra-month volatility of the strategy, and compute more precise risk measures.

### 12.11.1 Loading Daily Data

```
prices_daily = pd.read_sql_query(
    sql=("SELECT symbol, date, ret, ret_excess, mktcap_lag "
         "FROM prices_daily"),
    con=tidy_finance,
    parse_dates={"date"}
```

```
)

print(f"Daily observations: {len(prices_daily):,}")
print(f"Unique stocks: {prices_daily['symbol'].nunique():,}")
print(f"Date range: {prices_daily['date'].min().date()} "
      f"to {prices_daily['date'].max().date()}")
```

```
Daily observations: 3,462,157
Unique stocks: 1,459
Date range: 2010-01-05 to 2023-12-29
```

### 12.11.2 Daily Returns of Monthly Momentum Portfolios

Rather than forming momentum portfolios at the daily frequency (which would require daily rebalancing and is impractical), we use the monthly portfolio assignments and track their daily returns. This gives us the daily return series of the monthly momentum strategy.

```
# # # Use the monthly portfolio assignments from the J=6 baseline
monthly_assignments = portfolios_with_dates[
    ["symbol", "form_date", "momr", "hdate1", "hdate2"]
].copy()

# Merge with daily returns
daily_mom_returns = monthly_assignments.merge(
    prices_daily[["symbol", "date", "ret"]].rename(
        columns={"ret": "dret", "date": "ddate"}
    ),
    on="symbol",
    how="inner"
)

# Use boolean indexing instead of query
daily_mom_returns = daily_mom_returns[
    (daily_mom_returns["ddate"] >= daily_mom_returns["hdate1"]) &
    (daily_mom_returns["ddate"] <= daily_mom_returns["hdate2"])
]

print(f"Daily portfolio return observations: {len(daily_mom_returns):,}")
```

```
Daily portfolio return observations: 19,112,536
```

```
# Compute daily equally weighted portfolio returns
# Stage 1: Average within each cohort
daily_cohort_ret = (daily_mom_returns
    .groupby(["ddate", "momr", "form_date"])
    .agg(cohort_ret=("dret", "mean"))
    .reset_index()
)

# Stage 2: Average across cohorts
daily_ewret = (daily_cohort_ret
    .groupby(["ddate", "momr"])
    .agg(ewret=("cohort_ret", "mean"))
    .reset_index()
    .rename(columns={"ddate": "date"})
)

# Compute daily long-short returns
daily_wide = daily_ewret.pivot(
    index="date", columns="momr", values="ewret"
).reset_index()
daily_wide.columns = ["date"] + [f"port{i}" for i in range(1, 11)]
daily_wide["winners"] = daily_wide["port10"]
daily_wide["losers"] = daily_wide["port1"]
daily_wide["long_short"] = daily_wide["winners"] - daily_wide["losers"]

print(f"Daily long-short return observations: {len(daily_wide):,}")
```

```
Daily long-short return observations: 3,352
```

### 12.11.3 Daily Cumulative Returns

```
daily_wide = daily_wide.sort_values("date")
daily_wide["cumret_ls"] = (1 + daily_wide["long_short"]).cumprod() - 1

plot_daily_cumret = (
    ggplot(daily_wide, aes(x="date", y="cumret_ls")) +
    geom_line(size=0.5, color="#2ca02c") +
    geom_hline(yintercept=0, linetype="dashed", color="gray") +
    scale_y_continuous(labels=percent_format()) +
    labs(
```

```
        x="", y="Cumulative return",
        title="Daily Cumulative Return: Momentum Long-Short Strategy"
    ) +
    theme_minimal() +
    theme(figure_size=(10, 6))
)
plot_daily_cumret
```

Figure 12.10: Cumulative daily returns of the momentum long-short strategy. This figure shows the same strategy as the monthly analysis but tracked at daily frequency, revealing intra-month dynamics and the precise timing of momentum gains and losses.

### 12.11.4 Annualized Risk Metrics from Daily Data

Daily data enables more precise estimation of risk metrics through higher-frequency sampling.

```
daily_ls = daily_wide["long_short"].dropna()

# Annualized metrics
```

```
ann_mean = daily_ls.mean() * 252
ann_vol = daily_ls.std() * np.sqrt(252)
sharpe = ann_mean / ann_vol if ann_vol > 0 else np.nan

# Drawdown
daily_wealth = (1 + daily_ls).cumprod()
daily_running_max = daily_wealth.cummax()
daily_dd = (daily_wealth / daily_running_max - 1).min()

# Higher moments
skew = daily_ls.skew()
kurt = daily_ls.kurtosis()

# VaR and CVaR
var_95 = daily_ls.quantile(0.05)
cvar_95 = daily_ls[daily_ls <= var_95].mean()

risk_metrics = pd.DataFrame({
    "Metric": [
        "Annualized Mean Return",
        "Annualized Volatility",
        "Sharpe Ratio",
        "Maximum Drawdown",
        "Skewness",
        "Excess Kurtosis",
        "Daily VaR (5%)",
        "Daily CVaR (5%)"
    ],
    "Value": [
        f"{ann_mean*100:.2f}%",
        f"{ann_vol*100:.2f}%",
        f"{sharpe:.2f}",
        f"{daily_dd*100:.2f}%",
        f"{skew:.2f}",
        f"{kurt:.2f}",
        f"{var_95*100:.2f}%",
        f"{cvar_95*100:.2f}%"
    ]
})
risk_metrics
```

Table 12.11: Annualized risk metrics for the momentum long-short strategy computed from daily returns. Volatility is annualized using the square root of 252 rule. The Sharpe ratio, maximum drawdown, skewness, and kurtosis provide a comprehensive risk profile.

| | Metric | Value |
|---|---|---|
| 0 | Annualized Mean Return | -26.90% |
| 1 | Annualized Volatility | 15.01% |
| 2 | Sharpe Ratio | -1.79 |
| 3 | Maximum Drawdown | -97.62% |
| 4 | Skewness | -11.23 |
| 5 | Excess Kurtosis | 378.11 |
| 6 | Daily VaR (5%) | -1.33% |
| 7 | Daily CVaR (5%) | -2.10% |

### 12.11.5 Realized Volatility of Momentum Returns

Using daily returns, we can compute the monthly realized volatility of the momentum strategy and examine how it varies over time.

```
daily_wide["year_month"] = daily_wide["date"].dt.to_period("M")

realized_vol = (daily_wide
    .groupby("year_month")["long_short"]
    .std()
    .reset_index()
    .rename(columns={"long_short": "realized_vol"})
)
realized_vol["date"] = realized_vol["year_month"].dt.to_timestamp()
realized_vol["realized_vol_ann"] = realized_vol["realized_vol"] * np.sqrt(252)

plot_rvol = (
    ggplot(realized_vol, aes(x="date", y="realized_vol_ann")) +
    geom_line(color="#d62728", size=0.8) +
    scale_y_continuous(labels=percent_format()) +
    labs(
        x="", y="Annualized Realized Volatility",
        title="Realized Volatility of Momentum Strategy"
    ) +
    theme_minimal() +
    theme(figure_size=(10, 5))
```

```
)
plot_rvol
```

Figure 12.11: Monthly realized volatility of the momentum long-short strategy, computed from daily returns within each month. Higher values indicate periods of greater uncertainty in momentum profits, often coinciding with market stress.

## 12.12 Saving Results to the Database

We save the momentum portfolio returns to our database for use in subsequent chapters, including factor model construction and portfolio optimization.

```
# Save monthly equally weighted momentum portfolio returns
ewret_to_save = ewret[["date", "momr", "ewret"]].copy()
ewret_to_save.to_sql(
    name="momentum_portfolios_monthly",
    con=tidy_finance,
    if_exists="replace",
    index=False
)
print(f"Saved {len(ewret_to_save):,} monthly momentum portfolio observations.")

# Save the long-short return series
```

```
momentum_factor = ewret_wide[["date", "long_short"]].dropna().copy()
momentum_factor = momentum_factor.rename(columns={"long_short": "wml"})
momentum_factor.to_sql(
    name="momentum_factor_monthly",
    con=tidy_finance,
    if_exists="replace",
    index=False
)
print(f"Saved {len(momentum_factor):,} monthly WML factor observations.")

# Save the daily long-short return series
daily_momentum_factor = daily_wide[["date", "long_short"]].dropna().copy()
daily_momentum_factor = daily_momentum_factor.rename(
    columns={"long_short": "wml"}
)
daily_momentum_factor.to_sql(
    name="momentum_factor_daily",
    con=tidy_finance,
    if_exists="replace",
    index=False
)
print(f"Saved {len(daily_momentum_factor):,} daily WML factor observations.")
```

```
Saved 1,610 monthly momentum portfolio observations.
Saved 161 monthly WML factor observations.
Saved 3,352 daily WML factor observations.
```

## 12.13 Practical Considerations

### 12.13.1 Transaction Costs

Momentum strategies involve substantial portfolio turnover, as stocks enter and exit the extreme decile portfolios each month. Korajczyk and Sadka (2004) examine whether momentum profits survive transaction costs and find that profitability declines significantly for large institutional investors, though smaller portfolios can still capture meaningful returns.

In the Vietnamese market, transaction costs include:

- **Brokerage commissions**: Typically 0.15%–0.25% of transaction value for institutional investors.
- **Exchange fees**: Approximately 0.03% per trade.

- **Market impact**: Particularly relevant for smaller, less liquid stocks that dominate the extreme momentum portfolios. Vietnam's lower liquidity compared to developed markets may amplify this cost.
- **Trading band limits**: The $\pm 7\%$ daily price limit on HOSE can prevent immediate execution of trades, introducing tracking error relative to the theoretical portfolio.

### 12.13.2 Implementation Lag

Our baseline implementation assumes that portfolios can be formed and rebalanced instantaneously at the end of each month. In practice, there is a lag between observing the formation period returns and executing the portfolio trades. The one-month gap between the formation period and the start of the holding period (Jegadeesh (1990)) partially addresses this concern, but practitioners should consider additional implementation delays.

### 12.13.3 Survivorship Bias

Our dataset from DataCore includes both active and delisted stocks, which mitigates survivorship bias. However, the treatment of delisted stocks can affect momentum results. Stocks that are delisted during the holding period may generate extreme returns (both positive for acquisitions and negative for failures). We retain delisted returns as reported in the database, which is consistent with the treatment in Jegadeesh and Titman (1993).

### 12.13.4 Small Sample Considerations

The Vietnamese stock market has a relatively short history compared to the U.S. market studied in Jegadeesh and Titman (1993). Our sample spans approximately two decades, compared to the nearly three decades in the original study. This shorter sample period implies wider confidence intervals and greater sensitivity to specific episodes (such as the 2007–2009 financial crisis, which had a severe impact on Vietnamese equities). Results should be interpreted with this caveat in mind.

## 12.14 Key Takeaways

This chapter has provided an implementation and analysis of momentum strategies in the Vietnamese equity market, following the methodology of Jegadeesh and Titman (1993). The main findings and methodological contributions are:

1. **Methodology**: We implemented the full Jegadeesh and Titman (1993) overlapping portfolio methodology, including the two-stage averaging procedure (within-cohort, then across-cohort) that handles the $K$ active portfolios in each month.

2. **Baseline results**: The $J = 6$, $K = 6$ strategy provides a natural benchmark. The spread between winner and loser portfolios reveals whether cross-sectional momentum exists in Vietnamese equities.

3. **Robustness across horizons**: By computing the full $J \times K$ grid with $J, K \in \{3, 6, 9, 12\}$, we assessed whether the momentum premium is robust to the choice of formation and holding periods, following the approach in Jegadeesh and Titman (1993) Table 1.

4. **Risk adjustment**: CAPM and Fama-French three-factor alphas measure whether the momentum premium is explained by standard risk factors, building on the analysis in Eugene F. Fama and French (1996) who show that their three-factor model fails to explain momentum.

5. **Market state dependence**: Following Cooper, Gutierrez Jr, and Hameed (2004), we examined whether momentum profits vary with market conditions, which is particularly relevant in emerging markets with pronounced boom-bust cycles.

6. **Value weighting**: The comparison of equally weighted and value-weighted strategies addresses practical implementability and isolates the role of firm size in driving momentum profits.

7. **Daily analysis**: By tracking momentum portfolios at the daily frequency, we computed precise risk metrics including realized volatility, maximum drawdown, VaR, and CVaR, providing a complete risk profile of the strategy.

8. **Emerging market context**: Throughout the analysis, we have highlighted features specific to the Vietnamese market—trading band limits, foreign ownership restrictions, shorter sample history, and higher transaction costs—that affect the interpretation and practical viability of momentum strategies.

# 13 Fama-MacBeth Regressions

In this chapter, we delve into the implementation of the Eugene F. Fama and MacBeth (1973) regression approach, a cornerstone of empirical asset pricing. While portfolio sorts provide a robust, non-parametric view of the relationship between characteristics and returns, they struggle when we need to control for multiple factors simultaneously. For instance, in the Vietnamese stock market (HOSE and HNX), small-cap stocks often exhibit high illiquidity. Does the "Size effect" exist because small stocks are risky, or simply because they are illiquid? Fama-MacBeth (FM) regressions allow us to disentangle these effects in a linear framework.

We will implement a version of the FM procedure, accounting for:

1. **Weighted Least Squares (WLS):** To prevent micro-cap stocks, which are prevalent and volatile in Vietnam, from dominating the estimates.
2. **Newey-West Adjustments:** To handle the serial correlation often observed in Vietnamese market risk premiums.

## 13.1 The Econometric Framework

The Fama-MacBeth procedure is essentially a two-step filter that separates the cross-sectional variation in returns from the time-series variation.

### 13.1.1 Intuition: Why not Panel OLS?

A naive approach would be to pool all data ($N$ stocks $\times$ $T$ months) and run a single Ordinary Least Squares (OLS) regression:

$$r_{i,t+1} = \alpha + \beta_{i,t}\lambda + \epsilon_{i,t+1}$$

However, this assumes that the error terms $\epsilon_{i,t+1}$ are independent across firms. In reality, stock returns are highly cross-sectionally correlated (if the VN-Index crashes, most stocks fall together). A pooled OLS would underestimate the standard errors, leading to "false positive" discoveries of risk factors. Fama-MacBeth solves this by running $T$ separate cross-sectional regressions, effectively treating each month as a single independent observation of the risk premium.

### 13.1.2 Mathematical Derivation

#### 13.1.2.1 Step 1: Cross-Sectional Regressions

For each month $t$, we estimate the premium $\lambda_{k,t}$ for $K$ factors. Let $r_{i,t+1}$ be the excess return of asset $i$ at time $t+1$. Let $\beta_{i,t}$ be a vector of $K$ characteristics (e.g., Market Beta, Book-to-Market, Size) known at time $t$.

The model for a specific month $t$ is:

$$\mathbf{r}_{t+1} = \mathbf{X}_t \lambda_{t+1} + \alpha_{t+1} + \epsilon_{t+1}$$

Where:

- $\mathbf{r}_{t+1}$ is an $N \times 1$ vector of returns.
- $\mathbf{X}_t$ is an $N \times (K+1)$ matrix of factor exposures (including a column of ones for the intercept).
- $\lambda_{t+1}$ is the vector of risk premiums realized in month $t+1$.

To use **Weighted Least Squares (WLS)**, We define a weighting matrix $\mathbf{W}_t$ (typically diagonal with market capitalizations). The estimator for month $t$ is:

$$\hat{\lambda}_{t+1} = (\mathbf{X}_t^\top \mathbf{W}_t \mathbf{X}_t)^{-1} \mathbf{X}_t^\top \mathbf{W}_t \mathbf{r}_{t+1}$$

#### 13.1.2.2 Step 2: Time-Series Aggregation

We now have a time-series of $T$ estimates: $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_T$. The final estimate of the risk premium is the time-series average:

$$\hat{\lambda}_k = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}_{k,t}$$

The standard error is derived from the standard deviation of these monthly estimates:

$$\sigma(\hat{\lambda}_k) = \sqrt{\frac{1}{T^2} \sum_{t=1}^{T} (\hat{\lambda}_{k,t} - \hat{\lambda}_k)^2}$$

## 13.2 Data Preparation

We utilize data from our local SQLite database. In Vietnam, the fiscal year typically ends in December, and audited reports are required by April. To ensure no look-ahead bias, we lag accounting data (Book Equity) to match returns starting in July (a 6-month conservative lag, similar to Fama-French, but adapted for Vietnamese reporting delays).

```
import pandas as pd
import numpy as np
import sqlite3
import statsmodels.formula.api as smf
import statsmodels.api as sm
from pandas.tseries.offsets import MonthEnd

# Connect to the Vietnamese data
tidy_finance = sqlite3.connect(database="data/tidy_finance_python.sqlite")

# Load Monthly Prices (HOSE & HNX)
prices_monthly = pd.read_sql_query(
  sql="SELECT symbol, date, ret_excess, mktcap, mktcap_lag FROM prices_monthly",
  con=tidy_finance,
  parse_dates={"date"}
)

# Load Book Equity (derived from Vietnamese Financial Statements)
comp_vn = pd.read_sql_query(
  sql="SELECT datadate, symbol, be FROM comp_vn",
  con=tidy_finance,
  parse_dates={"datadate"}
)

# Load Rolling Market Betas (Pre-calculated in Chapter 'Beta Estimation')
beta_monthly = pd.read_sql_query(
  sql="SELECT symbol, date, beta FROM beta_monthly",
  con=tidy_finance,
  parse_dates={"date"}
)
```

We construct our testing characteristics:

1. **(Market Beta):** The sensitivity to the VN-Index.
2. **Size (ln(ME)):** The natural log of market capitalization.

3. **Value (BM):** The ratio of Book Equity to Market Equity.

```
# Prepare Characteristics
characteristics = (
    comp_vn
    # Align reporting date to month end
    .assign(date=lambda x: pd.to_datetime(x["datadate"]) + MonthEnd(0))
    # Merge with price data to get Market Cap at fiscal year end
    .merge(prices_monthly, on=["symbol", "date"], how="left")
    .merge(beta_monthly, on=["symbol", "date"], how="left")
    .assign(
        # Compute Book-to-Market
        bm=lambda x: x["be"] / x["mktcap"],
        log_mktcap=lambda x: np.log(x["mktcap"]),
        # Create sorting date: Financials valid from July of year t+1
        sorting_date=lambda x: x["date"] + pd.DateOffset(months=6) + MonthEnd(0),
    )
    .get(["symbol", "bm", "beta", "sorting_date"])
    .dropna()
)

characteristics.head()
```

| | symbol | bm | beta | sorting_date |
|---|---|---|---|---|
| 8729 | VTV | 7.034945e+08 | 0.847809 | 2017-06-30 |
| 8732 | MTG | 2.670306e+09 | 1.140066 | 2017-06-30 |
| 8739 | MKV | 6.505031e+08 | -0.448319 | 2017-06-30 |
| 8740 | MIC | 1.243127e+09 | 0.772140 | 2017-06-30 |
| 8742 | MCP | 6.657350e+08 | 0.348139 | 2017-06-30 |

```
# Merge back to monthly return panel
data_fm = (prices_monthly
  .merge(characteristics,
         left_on=["symbol", "date"],
         right_on=["symbol", "sorting_date"],
         how="left")
#   .merge(beta_monthly, on=["symbol", "date"], how="left")
  .sort_values(["symbol", "date"])
)

# Forward fill characteristics for 12 months (valid until next report)
```

```python
data_fm[["bm"]] = data_fm.groupby("symbol")[["bm"]].ffill(limit=12)

# Log Market Cap is updated monthly
data_fm["log_mktcap"] = np.log(data_fm["mktcap"])

# Lead returns: We use characteristics at t to predict return at t+1
data_fm["ret_excess_lead"] = data_fm.groupby("symbol")["ret_excess"].shift(-1)

# Cleaning: Remove rows with missing future returns or characteristics
data_fm = data_fm.dropna(subset=["ret_excess_lead", "beta", "log_mktcap", "bm"])

print(data_fm.head())

print(f"Data ready: {len(data_fm):,} observations from {data_fm.date.min().date()} to {data_
```

```
    symbol       date  ret_excess       mktcap   mktcap_lag            bm  \
163    AAA 2017-06-30    0.129454  2078.455619  1834.816104  7.929854e+08
175    AAA 2018-06-30   -0.067690  2758.426126  2948.159140  8.161755e+08
187    AAA 2019-06-30    0.030469  3141.519560  3038.799575  1.389438e+09
199    AAA 2020-06-30   -0.035462  2311.250278  2387.972279  1.497272e+09
211    AAA 2021-06-30    0.275355  5423.280296  4241.283308  1.456989e+09

         beta sorting_date  log_mktcap  ret_excess_lead
163  1.479060   2017-06-30    7.639380        -0.051090
175  1.090411   2018-06-30    7.922416        -0.095926
187  1.099956   2019-06-30    8.052462        -0.027856
199  0.954144   2020-06-30    7.745544        -0.098769
211  1.245004   2021-06-30    8.598456        -0.175128
Data ready: 5,075 observations from 2017-06-30 to 2023-06-30
```

## 13.3 Step 1: Cross-Sectional Regressions with WLS

Hou, Xue, and Zhang (2020) argue that micro-cap stocks distorts inference because they have high transaction costs and idiosyncratic volatility. In Vietnam, this is exacerbated by "penny stock" speculation.

We implement **Weighted Least Squares (WLS)** where weights are the market capitalization of the prior month. This tests if the factors are priced in the *investable* universe, not just the equal-weighted average of tiny stocks.

```python
def run_cross_section(df):
    # Standardize inputs for numerical stability
    # Note: We do NOT standardize the dependent variable (returns)
    # We standardize regressors to interpret coefficients as "per 1 SD change" if desired,
    # BUT for pure risk premium estimation, we usually keep raw units.
    # Here we use raw units to interpret lambda as % return per unit of characteristic.

    # Define Weighted Least Squares
    model = smf.wls(
        formula="ret_excess_lead ~ beta + log_mktcap + bm",
        data=df,
        weights=df["mktcap_lag"] # Weight by size
    )
    results = model.fit()

    return results.params

# Apply to every month
risk_premiums = (data_fm
  .groupby("date")
  .apply(run_cross_section)
  .reset_index()
)

print(risk_premiums.head())
```

```
        date  Intercept      beta  log_mktcap            bm
0 2017-06-30  -0.089116 -0.063799    0.010284  2.897813e-11
1 2018-06-30  -0.023221 -0.008252    0.001890  1.377518e-11
2 2019-06-30  -0.079373  0.035622    0.006224 -8.139910e-12
3 2020-06-30  -0.031213 -0.114968    0.008999 -2.306768e-11
4 2021-06-30   0.081397 -0.011407   -0.007330 -5.211290e-11
```

## 13.4 Step 2: Time-Series Aggregation & Hypothesis Testing

We now possess the time-series of risk premiums. We calculate the arithmetic mean and the -statistics.

Crucially, we use **Newey-West (HAC)** standard errors. Risk premiums in Vietnam often exhibit autocorrelation (momentum in factor performance). A simple standard error formula would be invalid.

```python
def calculate_fama_macbeth_stats(df, lags=6):
    summary = []

    for col in ["Intercept", "beta", "log_mktcap", "bm"]:
        series = df[col]

        # 1. Point Estimate (Average Risk Premium)
        mean_premium = series.mean()

        # 2. Newey-West Standard Error
        # We regress the series on a constant (ones) to get the SE of the mean
        exog = sm.add_constant(np.ones(len(series)))
        nw_model = sm.OLS(series, exog).fit(
            cov_type='HAC', cov_kwds={'maxlags': lags}
        )

        se = nw_model.bse.iloc[0]
        t_stat = nw_model.tvalues.iloc[0]

        summary.append({
            "Factor": col,
            "Premium (%)": mean_premium * 100,
            "Std Error": se * 100,
            "t-statistic": t_stat,
            "Significance": "*" if abs(t_stat) > 1.96 else ""
        })

    return pd.DataFrame(summary)

price_of_risk = calculate_fama_macbeth_stats(risk_premiums)
print(price_of_risk.round(4))
```

```
       Factor  Premium (%)  Std Error  t-statistic Significance
0   Intercept      -1.8174     1.9117      -0.9507
1        beta      -1.7859     1.0407      -1.7161
2  log_mktcap       0.2347     0.2048       1.1457
3          bm      -0.0000     0.0000      -0.0928
```

### 13.4.1 Visualizing the Time-Varying Risk Premium

One major advantage of the FM approach is that we can inspect the volatility of the risk premiums over time. In Vietnam, we expect the "Size" premium to be highly volatile during periods of retail liquidity injection (e.g., 2020-2021).

```
import matplotlib.pyplot as plt
import matplotlib.ticker as mtick

# Calculate cumulative returns of the factors (as if they were tradable portfolios)
cumulative_premiums = (risk_premiums
    .set_index("date")
    .drop(columns=["Intercept"])
    .cumsum()
)

fig, ax = plt.subplots(figsize=(10, 6))
cumulative_premiums.plot(ax=ax, linewidth=2)
ax.set_title("Cumulative Risk Premiums in Vietnam (Fama-MacBeth)", fontsize=14)
ax.set_ylabel("Cumulative Coefficient Return")
ax.legend(title="Factor")
ax.grid(True, alpha=0.3)
plt.show()
```

Figure 13.1: Cumulative Risk Premiums in Vietnam.

1. **Market Beta:** In many empirical studies (including the US), the market beta premium is often insignificant or even negative (the "Betting Against Beta" anomaly). In Vietnam, if the -stat is , it implies the CAPM does not explain the cross-section of returns.
2. **Size (Log Mktcap):** A negative coefficient confirms the "Size Effect"—smaller firms have higher expected returns. However, using WLS often weakens this result compared to OLS, suggesting the size premium is concentrated in micro-caps.
3. **Value (BM):** A positive coefficient confirms the Value premium. In Vietnam, value stocks (high B/M) often outperform growth stocks, particularly in the manufacturing and banking sectors.

Figure 13.1 plots the cumulative sum of the monthly Fama MacBeth risk premium estimates for beta, size, and value. Because these lines cumulate estimated cross sectional prices of risk rather than actual portfolio returns, the figure should be interpreted as showing the time variation and persistence of estimated premia, not investable performance.

The beta premium displays a clear regime shift around 2020, with a sharp decline that only partially reverses afterward. This pattern suggests that the pricing of systematic risk in Vietnam is unstable over short samples and may be heavily influenced by episodic market conditions such as the post COVID retail trading boom. The size premium is comparatively smoother but small in magnitude, indicating only weak and time varying evidence that firm

size is priced in the cross section during this period. The value premium remains close to zero throughout, implying little consistent cross sectional reward to high book to market firms in this sample window.

Overall, the figure highlights that estimated risk premia in the Vietnamese market are highly time varying and sensitive to specific macro and market regimes, reinforcing the need for caution when drawing conclusions from short samples.

## 13.5 Sanity Checks

### 13.5.1 Time-Series Volatility Check

Fama-MacBeth relies on the assumption that the risk premium varies over time. If your `bm` premium is truly near zero every month, the method fails.

**Action:** Plot the time series of the estimated coefficients . You want to see "noise" around a mean. If you see a flat line or a single massive spike, your data is corrupted.

```
import matplotlib.pyplot as plt

# Plot the time series of the BM risk premium
fig, ax = plt.subplots(figsize=(10, 5))
risk_premiums["bm"].plot(ax=ax, title="Monthly Value Premium (BM) Coefficient")
ax.axhline(0, color="black", linestyle="--")
ax.set_ylabel("Slope Coefficient")
plt.show()
```

### 13.5.2 Correlation of Characteristics (Multicollinearity)

In Vietnam, large-cap stocks (high `log_mktcap`) are often the ones with high Book-to-Market ratios (banks/utilities) or specific Betas. If your factors are highly correlated, the Fama-MacBeth coefficients will be unstable and insignificant (low t-stats), even if the factors actually matter.

**Action:** Check the cross-sectional correlation.

```
# Check correlation of the characteristics
corr_matrix = data_fm[["beta", "log_mktcap", "bm"]].corr()
print(corr_matrix)
```

```
               beta  log_mktcap        bm
beta        1.000000    0.392776 -0.033748
log_mktcap  0.392776    1.000000 -0.203307
bm         -0.033748   -0.203307  1.000000
```

**Interpretation:**

- If correlation $> 0.7$ (absolute value), the regression struggles to distinguish between the two factors.
- For example, if `Size` and `Liquidity` are -0.8 correlated, the model cannot tell which one is driving the return, often resulting in both having insignificant t-stats.

# 14 Event Studies in Finance

Event studies constitute one of the most enduring and widely deployed empirical methodologies in financial economics. At their core, event studies measure the impact of a specific event on the value of a firm by examining **abnormal security returns** around the time the event occurs. The methodology rests on a simple premise: if capital markets are informationally efficient, the effect of an event will be reflected immediately in security prices, and any deviation from "normal" expected returns can be attributed to the event itself.

Since the pioneering work of Eugene F. Fama et al. (1969), who studied how stock prices adjust to new information around stock splits, event studies have become a cornerstone of empirical research across finance, accounting, economics, and law. Ball and Brown (2013) demonstrated that accounting earnings announcements convey information to the market, a finding that launched decades of research in accounting and disclosure. The methodology has since been refined through contributions by Brown and Warner (1980) and Brown and Warner (1985), who established the statistical properties of event study methods, and MacKinlay (1997) codified best practices that remain standard today.

The breadth of applications is remarkable. Event studies have been used to examine the wealth effects of mergers and acquisitions (Jensen and Ruback 1983; Andrade, Mitchell, and Stafford 2001), earnings announcements (Bernard and Thomas 1989), dividend changes (Aharony and Swary 1980), regulatory changes (Schwert 1981), executive turnover (Warner, Watts, and Wruck 1988), and macroeconomic announcements (Flannery and Protopapadakis 2002). In law and economics, event studies serve as the primary tool for measuring damages in securities fraud litigation (Mitchell and Netter 1993) and assessing the impact of regulatory interventions (Binder 1998). Kothari and Warner (2007) documented over 500 published event studies in the top five finance journals alone between 1974 and 2000.

### 14.0.1 Why Event Studies Matter

The enduring popularity of event studies stems from several compelling properties:

- **Direct measurement of economic significance.** Unlike regression-based approaches that estimate associations, event studies directly quantify the dollar impact of events on firm value. A cumulative abnormal return (CAR) of 3% for a firm with $10 billion market capitalization translates to $300 million in wealth creation, which is a tangible, economically meaningful magnitude.

- **Minimal maintained assumptions.** The methodology requires only semi-strong market efficiency (i.e., prices reflect publicly available information), a weaker assumption than many alternatives.
- **Statistical power.** Daily event studies have remarkable power to detect abnormal performance, even with modest sample sizes. Brown and Warner (1985) demonstrated that the market model detects abnormal returns of 1% or more with high reliability using samples as small as 20 securities.
- **Versatility.** The basic framework accommodates events that are firm-specific or market-wide, anticipated or surprising, and can be adapted to various asset classes and market structures.

---

## 14.1 Literature Review and Methodological Evolution

### 14.1.1 The Classical Framework (1969-1985)

The modern event study traces its origins to Eugene F. Fama et al. (1969), hereafter FFJR, who examined monthly stock returns around 940 stock splits between 1927 and 1959. Their key innovation was the use of the **market model** to decompose returns into expected (normal) and unexpected (abnormal) components.

Ball and Brown (2013) independently developed a similar approach to study earnings announcements, establishing the information content of accounting data. It was a finding with profound implications for both the efficient markets hypothesis and the relevance of financial reporting.

Brown and Warner (1980) provided the first systematic analysis of event study methodology using simulation. Their study of monthly data established several important results: (i) the simple market model performs at least as well as more complex models, (ii) value-weighted market indices can lead to misspecification when the sample is tilted toward smaller firms, and (iii) the standard cross-sectional test has well-specified size under the null hypothesis. Their follow-up study (Brown and Warner 1985) extended the analysis to daily data, documenting the importance of non-normality in daily returns and the increased power of daily versus monthly studies.

### 14.1.2 Risk Model Refinements (1992-2015)

The advent of the **Fama-French three-factor model** (Eugene F. Fama and French 1993a) represented a major advance in modeling expected returns. Adding size (SMB) and value (HML) factors to the market model improved the cross-sectional fit of expected returns considerably.

Mark M. Carhart (1997b) augmented this with a momentum factor (UMD), yielding the four-factor model that became standard in event studies through the 2000s. Eugene F. Fama and French (2015) subsequently introduced profitability (RMW) and investment (CMA) factors in their five-factor model.

The choice of risk model matters for event studies primarily in **long-horizon settings**. Kothari and Warner (2007) showed that for short-window studies (3-5 days), the market model and multi-factor models produce virtually identical results because the incremental factors explain very little daily return variation for individual firms. However, for event windows exceeding 20 trading days, model choice can materially affect inferences.

### 14.1.3 Testing for Abnormal Returns (1976-2010)

The statistical testing of abnormal returns has evolved considerably:

Table 14.1: Summary of major event study test statistics

| Test | Year | Key Property | Reference |
|---|---|---|---|
| **Patell Z** | 1976 | Standardizes by estimation-period $\sigma$; weights firms inversely by volatility | Patell (1976) |
| **Cross-Sectional** $t$ | 1980 | Allows event-induced variance change | Brown and Warner (1980) |
| **BMP** | 1991 | Robust to event-induced variance | Boehmer, Masumeci, and Poulsen (1991) |
| **Corrado Rank** | 1989 | Non-parametric; robust to non-normality | Corrado (1989) |
| **Generalized Sign** | 1992 | Non-parametric; uses estimation-window baseline | Cowan (1992) |
| **Kolari-Pynnönen** | 2010 | Accounts for cross-sectional dependence | Kolari and Pynnönen (2010) |
| **Skewness-Adjusted** | 1992 | Corrects for BHAR skewness | Hall (1992) |

### 14.1.4 CARs versus BHARs

Cumulative abnormal returns (CARs) sum daily abnormal returns, while buy-and-hold abnormal returns (BHARs) compound returns and subtract the compounded benchmark. Barber and Lyon (1997) demonstrated that BHARs better capture the actual investor experience, since investors earn compound, not cumulative, returns. However, Eugene F. Fama (1998) and Mitchell and Stafford (2000) showed that BHARs exhibit severe cross-sectional dependence and positive skewness. For **short event windows** (under 10 days), the difference between CARs and BHARs is negligible. For longer windows, both should be reported.

### 14.1.5 Emerging Market Considerations

Event studies in emerging markets face distinct challenges:

- **Thin trading.** Many emerging market securities trade infrequently, inducing bias in market model beta estimates. Scholes and Williams (1977) and Dimson (1979) proposed corrections using leading and lagging market returns.
- **Factor availability.** While Fama-French factors are readily available for developed markets, emerging market factors must often be constructed locally.
- **Market microstructure.** Price limits ($\pm$ 7% on HOSE, $\pm$ 10% on HNX, $\pm$ 15% on UPCOM in Vietnam), T+2 settlement, and the absence of short-selling affect the speed of price adjustment. Researchers should consider wider event windows to accommodate slower information incorporation (Bhattacharya et al. 2000; Griffin, Kelly, and Nardari 2010).

## 14.2 Mathematical Framework

This section presents the complete mathematical specification of the event study methodology. We follow the notation conventions of Campbell et al. (1998) and Kothari and Warner (2007).

### 14.2.1 Timeline and Windows

The event study timeline is defined relative to the event date, denoted $\tau = 0$. All dates are measured in **trading days**:

$$\underbrace{T_0 + 1, \dots, T_1}_{\text{Estimation Window (L days)}} \quad \underbrace{\quad}_{\text{Gap (G days)}} \quad \underbrace{\tau_1, \dots, 0, \dots, \tau_2}_{\text{Event Window (L days)}}$$

where:

- **Estimation window**: $L_1$ trading days over which the risk model parameters are estimated
- **Gap**: $G$ trading days separating estimation and event windows, preventing contamination by pre-event information leakage
- **Event window**: $L_2 = \tau_2 - \tau_1 + 1$ trading days centered around the event date

For example, with $L_1 = 150$, $G = 15$, $\tau_1 = -10$, $\tau_2 = +10$: the estimation window covers trading days $[-175, -25]$ relative to the event, and the event window covers $[-10, +10]$.

### 14.2.2 Normal Return Models

Let $R_{it}$ denote the return on security $i$ on trading day $t$, $R_{ft}$ the risk-free rate, and $R_{mt}$ the market return. We implement six models:

**Model 0: Market-Adjusted Returns.** Assumes $\beta_i = 1$ and $\alpha_i = 0$ for all firms:

$$AR_{it}^{MA} = R_{it} - R_{mt}$$

**Model 1: Market Model** (Sharpe 1964):

$$R_{it} = \alpha_i + \beta_i R_{mt} + \varepsilon_{it}, \quad E[\varepsilon_{it}] = 0, \quad \mathrm{Var}[\varepsilon_{it}] = \sigma_{\varepsilon_i}^2$$

$$AR_{it}^{MM} = R_{it} - \hat{\alpha}_i - \hat{\beta}_i R_{mt}$$

**Model 2: Fama-French Three-Factor** (Eugene F. Fama and French 1993a):

$$R_{it} - R_{ft} = \alpha_i + \beta_{i,1}(R_{mt} - R_{ft}) + \beta_{i,2} \cdot SMB_t + \beta_{i,3} \cdot HML_t + \varepsilon_{it}$$

**Model 3: Carhart Four-Factor** (Mark M. Carhart 1997b):

$$R_{it} - R_{ft} = \alpha_i + \beta_{i,1}(R_{mt} - R_{ft}) + \beta_{i,2} \cdot SMB_t + \beta_{i,3} \cdot HML_t + \beta_{i,4} \cdot UMD_t + \varepsilon_{it}$$

**Model 4: Fama-French Five-Factor** (Eugene F. Fama and French 2015):

$$R_{it} - R_{ft} = \alpha_i + \beta_{i,1}(R_{mt} - R_{ft}) + \beta_{i,2} \cdot SMB_t + \beta_{i,3} \cdot HML_t + \beta_{i,4} \cdot RMW_t + \beta_{i,5} \cdot CMA_t + \varepsilon_{it}$$

**Model 5: User-Specified Factor Model:**

$$R_{it} - R_{ft} = \alpha_i + \sum_{k=1}^{K} \beta_{i,k} F_{k,t} + \varepsilon_{it}$$

### 14.2.3 Aggregation: CARs and BHARs

**Cumulative Abnormal Returns** sum daily abnormal returns:

$$CAR_i(\tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau_2} AR_{it}, \qquad \overline{CAR}(\tau_1, \tau_2) = \frac{1}{N} \sum_{i=1}^{N} CAR_i(\tau_1, \tau_2)$$

**Buy-and-Hold Abnormal Returns** compound returns:

$$BHAR_i(\tau_1, \tau_2) = \prod_{t=\tau_1}^{\tau_2} (1 + R_{it}) - \prod_{t=\tau_1}^{\tau_2} (1 + \hat{E}[R_{it}])$$

### 14.2.4 Standardized Returns

The **standardized abnormal return** for firm $i$ on day $t$ is:

$$SAR_{it} = \frac{AR_{it}}{\hat{\sigma}_{\varepsilon_i}}$$

The **standardized cumulative abnormal return** is:

$$SCAR_i(\tau_1, \tau_2) = \frac{CAR_i(\tau_1, \tau_2)}{\hat{\sigma}_{\varepsilon_i} \sqrt{L_2}}$$

### 14.2.5 Test Statistics

Let $N$ denote the number of firm-event observations.

**Test 1: Cross-Sectional** $t$-Test. Allows event-induced variance; assumes cross-sectional independence:

$$t_{CS} = \frac{\overline{CAR}}{s_{CAR}/\sqrt{N}}, \quad s_{CAR} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (CAR_i - \overline{CAR})^2}$$

**Test 2: Patell Z-Test** (Patell 1976). Weights firms inversely by volatility:

$$Z_{Patell} = \frac{\sum_{i=1}^{N} SCAR_i}{\sqrt{\sum_{i=1}^{N} \frac{K_i - 2}{K_i - 4}}}$$

**Test 3: BMP Test** (Boehmer, Masumeci, and Poulsen 1991). Robust to event-induced variance:

$$t_{BMP} = \frac{\overline{SCAR}}{s_{SCAR}/\sqrt{N}}$$

**Test 4: Kolari-Pynnönen Adjusted BMP** (Kolari and Pynnönen 2010). Accounts for cross-sectional dependence:

$$t_{KP} = t_{BMP} \times \sqrt{\frac{1}{1 + (N-1)\bar{r}}}$$

where $\bar{r}$ is the mean pairwise cross-correlation of estimation-period residuals.

**Test 5: Generalized Sign Test** (Cowan 1992):

$$Z_{GSign} = \frac{\hat{p} - \hat{p}_0}{\sqrt{\hat{p}_0(1-\hat{p}_0)/N}}$$

**Test 6: Sign Test:**

$$Z_{Sign} = \frac{N^+ - 0.5N}{\sqrt{0.25N}}$$

**Test 7: Skewness-Adjusted** $t$-Test (Hall 1992):

$$t_{SA} = \sqrt{N}\left(\bar{z} + \frac{1}{3}\hat{\gamma}\bar{z}^2 + \frac{1}{27}\hat{\gamma}^2\bar{z}^3 + \frac{1}{6N}\hat{\gamma}\right)$$

**Test 8: Wilcoxon Signed-Rank Test:** A non-parametric test of whether the median CAR differs from zero.

The table below summarizes the assumptions of each test:

Table 14.2: Assumption requirements for event study test statistics

| Test | Event-Induced Variance | Cross-Sectional Independence | Normality |
|---|---|---|---|
| Cross-Sectional $t$ | Robust | Assumes | Assumes |
| Patell Z | Assumes no change | Assumes | Assumes |
| BMP | Robust | Assumes | Assumes |
| Kolari-Pynnönen | Robust | Robust | Assumes |

| Test | Event-Induced Variance | Cross-Sectional Independence | Normality |
|---|---|---|---|
| Generalized Sign | Robust | Assumes | Robust |
| Corrado Rank | Robust | Assumes | Robust |
| Skewness-Adjusted | Robust | Assumes | Partially |
| Wilcoxon | Robust | Assumes | Robust |

## 14.3 Python Implementation

### 14.3.1 Design Philosophy

Our implementation follows these principles:

1. **Modularity**: Each component (calendar, estimation, AR computation, testing) is a separate function.
2. **Vectorization**: All operations use pandas/numpy for performance on large datasets.
3. **Configurability**: All parameters are user-configurable via a dataclass.
4. **Transparency**: Intermediate outputs are preserved for inspection.
5. **Production-ready**: Comprehensive input validation, missing data handling, and edge cases.

### 14.3.2 Setup and Imports

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
from scipy import stats
from dataclasses import dataclass, field
from typing import Optional, List, Tuple
from enum import Enum
import warnings
import matplotlib.pyplot as plt
import matplotlib.ticker as mticker

warnings.filterwarnings('ignore')
pd.set_option('display.float_format', '{:.6f}'.format)
print("All libraries loaded.")
```

```
All libraries loaded.
```

```python
import pandas as pd
import sqlite3

tidy_finance = sqlite3.connect(database="data/tidy_finance_python.sqlite")

factors_ff3_daily = pd.read_sql_query(
    sql="SELECT * FROM factors_ff3_daily",
    con=tidy_finance,
    parse_dates=["date"]
)

factors_ff5_daily = pd.read_sql_query(
    sql="SELECT * FROM factors_ff5_daily",
    con=tidy_finance,
    parse_dates=["date"]
)

factors_ff3_monthly = pd.read_sql_query(
    sql="SELECT * FROM factors_ff3_monthly",
    con=tidy_finance,
    parse_dates=["date"]
)

factors_ff5_monthly = pd.read_sql_query(
    sql="SELECT * FROM factors_ff5_monthly",
    con=tidy_finance,
    parse_dates=["date"]
)
```

```python
prices_monthly = pd.read_sql_query(
    sql="""
        SELECT symbol, date, ret_excess, mktcap, mktcap_lag, risk_free
        FROM prices_monthly
    """,
    con=tidy_finance,
    parse_dates={"date"}
).dropna()

prices_daily = pd.read_sql_query(
    sql="""
        SELECT symbol, date, ret_excess, mktcap, mktcap_lag, risk_free
        FROM prices_daily
```

```
    """,
    con=tidy_finance,
    parse_dates={"date"}
).dropna()
```

### 14.3.3 Configuration

```
class RiskModel(Enum):
    """Supported risk models for expected return computation."""
    MARKET_ADJ = "market_adjusted"
    MARKET_MODEL = "market_model"
    FF3 = "ff3"
    CARHART = "carhart"
    FF5 = "ff5"
    CUSTOM = "custom"

@dataclass
class EventStudyConfig:
    """Complete configuration for an event study.

    Attributes
    ----------
    estimation_window : int
        Length of estimation period in trading days. Brown and Warner (1985)
        suggest  100 days; MacKinlay (1997) recommends 120 as standard.
    event_window_start : int
        Start of event window relative to event date (e.g., -10).
    event_window_end : int
        End of event window relative to event date (e.g., +10).
    gap : int
        Trading days between estimation and event windows. Prevents
        contamination from pre-event information leakage.
    min_estimation_obs : int
        Minimum non-missing returns required in estimation period.
    risk_model : RiskModel
        Risk model for computing expected returns.
    custom_factors : list
        Column names for user-specified factors (CUSTOM model only).
    thin_trading_adj : str or None
        None, 'scholes_williams', or 'dimson'.
```

```python
    dimson_lags : int
        Number of leads/lags for Dimson (1979) correction.
    """
    estimation_window: int = 150
    event_window_start: int = -10
    event_window_end: int = 10
    gap: int = 15
    min_estimation_obs: int = 120
    risk_model: RiskModel = RiskModel.MARKET_MODEL
    custom_factors: List[str] = field(default_factory=list)
    thin_trading_adj: Optional[str] = None
    dimson_lags: int = 1

    @property
    def event_window_length(self) -> int:
        return self.event_window_end - self.event_window_start + 1

    def validate(self):
        assert self.estimation_window > 0
        assert self.event_window_start <= self.event_window_end
        assert self.gap >= 0
        assert self.min_estimation_obs <= self.estimation_window
        if self.risk_model == RiskModel.CUSTOM:
            assert len(self.custom_factors) > 0
        return True

# Demonstrate
config_demo = EventStudyConfig(
    estimation_window=150, event_window_start=-10, event_window_end=10,
    gap=15, min_estimation_obs=120, risk_model=RiskModel.FF3
)
config_demo.validate()
print(f"Event window length: {config_demo.event_window_length} days")
print(f"Model: {config_demo.risk_model.value}")
```

```
Event window length: 21 days
Model: ff3
```

### 14.3.4 Step 1: Trading Calendar Construction

A correct trading calendar is fundamental. It maps any event date to the exact calendar dates for the start/end of estimation and event windows, accounting for weekends, holidays, and non-trading days.

```python
def build_trading_calendar(trading_dates, config):
    """Build a trading calendar mapping event dates to window boundaries.

    For each potential event date, identifies the calendar dates for the
    start/end of the estimation period and event window using only actual
    trading days.

    Parameters
    ----------
    trading_dates : array-like
        Sorted unique trading dates in the market.
    config : EventStudyConfig

    Returns
    -------
    pd.DataFrame with columns: estper_beg, estper_end, evtwin_beg,
        evtdate, evtwin_end, cal_index
    """
    dates = pd.Series(sorted(pd.to_datetime(trading_dates).unique()))
    n = len(dates)

    L1 = config.estimation_window
    G = config.gap
    s = config.event_window_start
    L2 = config.event_window_length

    # Offsets (FIRSTOBS logic)
    o0 = 0                      # estper_beg
    o1 = L1 - 1                 # estper_end
    o2 = L1 + G                 # evtwin_beg
    o3 = L1 + G - s             # evtdate
    o4 = L1 + G + L2 - 1        # evtwin_end

    max_offset = o4
    valid = n - max_offset
    if valid <= 0:
        raise ValueError(f"Need {max_offset+1} trading dates, have {n}")
```

```
    cal = pd.DataFrame({
        'estper_beg': dates.iloc[o0:o0+valid].values,
        'estper_end': dates.iloc[o1:o1+valid].values,
        'evtwin_beg': dates.iloc[o2:o2+valid].values,
        'evtdate':    dates.iloc[o3:o3+valid].values,
        'evtwin_end': dates.iloc[o4:o4+valid].values,
    })
    cal['cal_index'] = range(1, len(cal)+1)

    # Validate window lengths using a sample row
    idx = min(10, len(cal)-1)
    row = cal.iloc[idx]
    est_n = dates[(dates >= row['estper_beg']) & (dates <= row['estper_end'])].shape[0]
    evt_n = dates[(dates >= row['evtwin_beg']) & (dates <= row['evtwin_end'])].shape[0]
    assert est_n == L1, f"Estimation window: {est_n}  {L1}"
    assert evt_n == L2, f"Event window: {evt_n}  {L2}"

    return cal

# Demo
demo_dates = pd.bdate_range('2018-01-01', '2023-12-31', freq='B')
demo_cal = build_trading_calendar(demo_dates, config_demo)
print(f"Calendar: {len(demo_cal)} potential event dates")
print(demo_cal.head(3).to_string(index=False))
```

```
Calendar: 1380 potential event dates
estper_beg estper_end evtwin_beg    evtdate evtwin_end  cal_index
2018-01-01 2018-07-27 2018-08-20 2018-09-03 2018-09-17          1
2018-01-02 2018-07-30 2018-08-21 2018-09-04 2018-09-18          2
2018-01-03 2018-07-31 2018-08-22 2018-09-05 2018-09-19          3
```

### 14.3.5 Step 2: Event Date Alignment

When an event occurs on a non-trading day, align to the **next** available trading day.

```
def align_events(events, calendar, id_col='symbol', date_col='event_date'):
    """Align event dates to trading calendar.

    Non-trading-day events are shifted forward to the next trading day.
```

```
    Parameters
    ----------
    events : pd.DataFrame with [id_col, date_col] and optional 'group'
    calendar : pd.DataFrame from build_trading_calendar()

    Returns
    -------
    pd.DataFrame with window boundaries for each firm-event
    """
    events = events.copy()
    events[date_col] = pd.to_datetime(events[date_col])

    cal_dates = calendar[['evtdate']].drop_duplicates().sort_values('evtdate')

    merged = pd.merge_asof(
        events.sort_values(date_col),
        cal_dates.rename(columns={'evtdate': 'aligned_date'}),
        left_on=date_col, right_on='aligned_date',
        direction='forward'
    )

    result = merged.merge(calendar, left_on='aligned_date', right_on='evtdate', how='inner')

    shifted = (result[date_col] != result['evtdate']).sum()
    if shifted > 0:
        print(f"  {shifted} event(s) shifted to next trading day")

    result = result.rename(columns={date_col: 'original_date'})
    result = result.drop_duplicates(subset=[id_col, 'evtdate'])

    return result
```

### 14.3.6 Step 3: Data Extraction and Factor Merging

Extract returns for each security-event across the full estimation + event window and merge risk factors.

```
def extract_returns(aligned_events, prices, factors, config,
                    id_col='symbol', date_col='date', ret_col='ret',
                    mkt_col='mkt_excess', rf_col='risk_free'):
    """Extract stock returns and merge risk factors for each event.
```

```
    For each security-event, retrieves daily returns from estper_beg
    through evtwin_end and merges appropriate risk factors.
    """
    prices = prices.copy()
    factors = factors.copy()
    prices[date_col] = pd.to_datetime(prices[date_col])
    factors[date_col] = pd.to_datetime(factors[date_col])

    # Recover raw return from excess return if needed
    if ret_col not in prices.columns and 'ret_excess' in prices.columns:
        if rf_col in factors.columns:
            prices = prices.merge(factors[[date_col, rf_col]].drop_duplicates(),
                                  on=date_col, how='left')
        prices[ret_col] = prices['ret_excess'] + prices[rf_col]

    # Factor columns based on model
    model = config.risk_model
    fac_cols = [mkt_col] if mkt_col in factors.columns else []
    if rf_col in factors.columns:
        fac_cols.append(rf_col)

    model_factors = {
        RiskModel.FF3: ['smb', 'hml'],
        RiskModel.CARHART: ['smb', 'hml', 'umd'],
        RiskModel.FF5: ['smb', 'hml', 'rmw', 'cma'],
        RiskModel.CUSTOM: config.custom_factors,
    }
    for f in model_factors.get(model, []):
        if f in factors.columns:
            fac_cols.append(f)

    fac_cols = list(set([date_col] + fac_cols))

    # Vectorized merge approach: join events with prices on id + date range
    frames = []
    for _, evt in aligned_events.iterrows():
        mask = ((prices[id_col] == evt[id_col]) &
                (prices[date_col] >= evt['estper_beg']) &
                (prices[date_col] <= evt['evtwin_end']))
        fd = prices.loc[mask, [id_col, date_col, ret_col]].copy()
        if len(fd) == 0:
            continue
```

```
        fd['evtdate'] = evt['evtdate']
        fd['estper_beg'] = evt['estper_beg']
        fd['estper_end'] = evt['estper_end']
        fd['evtwin_beg'] = evt['evtwin_beg']
        fd['evtwin_end'] = evt['evtwin_end']
        if 'group' in evt.index:
            fd['group'] = evt['group']
        frames.append(fd)

    if not frames:
        raise ValueError("No return data found for any events")

    result = pd.concat(frames, ignore_index=True)
    result = result.merge(factors[fac_cols].drop_duplicates(), on=date_col, how='left')

    # Excess and market-adjusted returns
    if rf_col in result.columns:
        result['ret_excess'] = result[ret_col] - result[rf_col]
    else:
        result['ret_excess'] = result[ret_col]
    if mkt_col in result.columns:
        result['ret_mktadj'] = result['ret_excess'] - result[mkt_col]

    result = result.sort_values([id_col, 'evtdate', date_col]).reset_index(drop=True)
    n_evts = result.groupby([id_col, 'evtdate']).ngroups
    print(f"  Extracted {len(result):,} obs for {n_evts} firm-events")
    return result
```

### 14.3.7 Step 4: Risk Model Estimation

Estimate risk model parameters over the estimation window.

```
def estimate_model(
    event_returns, config, id_col="symbol", date_col="date", ret_col="ret"
):
    """Estimate risk model parameters for each firm-event.

    Runs OLS over the estimation window. Returns alpha, betas, sigma,
    R^2, nobs, and residuals for cross-correlation computation.
    """
    model = config.risk_model
```

```
# Define regression specification
dep_var_map = {
    RiskModel.MARKET_ADJ: "ret_mktadj",
    RiskModel.MARKET_MODEL: ret_col,
    RiskModel.FF3: "ret_excess",
    RiskModel.CARHART: "ret_excess",
    RiskModel.FF5: "ret_excess",
    RiskModel.CUSTOM: "ret_excess",
}
indep_var_map = {
    RiskModel.MARKET_ADJ: [],
    RiskModel.MARKET_MODEL: ["mkt_excess"],
    RiskModel.FF3: ["mkt_excess", "smb", "hml"],
    RiskModel.CARHART: ["mkt_excess", "smb", "hml", "umd"],
    RiskModel.FF5: ["mkt_excess", "smb", "hml", "rmw", "cma"],
    RiskModel.CUSTOM: config.custom_factors,
}

dep_var = dep_var_map[model]
indep_vars = indep_var_map[model]

est = event_returns[
    (event_returns[date_col] >= event_returns["estper_beg"])
    & (event_returns[date_col] <= event_returns["estper_end"])
].copy()

params_list = []

for (firm, evtdate), grp in est.groupby([id_col, "evtdate"]):
    valid = grp.dropna(subset=[dep_var] + indep_vars)
    nobs = len(valid)
    if nobs < config.min_estimation_obs:
        continue

    y = valid[dep_var].values

    if len(indep_vars) == 0:
        # Market-adjusted: intercept-only for variance
        p = {
            id_col: firm,
            "evtdate": evtdate,
            "alpha": y.mean(),
```

```
                "sigma": y.std(ddof=1),
                "variance": y.var(ddof=1),
                "nobs": nobs,
                "r_squared": 0.0,
                "_residuals": y - y.mean(),
            }
        else:
            X = sm.add_constant(valid[indep_vars].values)
            res = sm.OLS(y, X).fit()
            p = {
                id_col: firm,
                "evtdate": evtdate,
                "alpha": res.params[0],
                "sigma": np.sqrt(res.mse_resid),
                "variance": res.mse_resid,
                "nobs": nobs,
                "r_squared": res.rsquared if np.isfinite(res.rsquared) else np.nan,
                "_residuals": res.resid,
            }
            for j, var in enumerate(indep_vars):
                p[f"beta_{var}"] = res.params[j + 1]

        # Skip degenerate firms (zero or near-zero variance)
        if p["sigma"] < 1e-6:
            continue

        params_list.append(p)

    if not params_list:
        raise ValueError("No firm-events passed minimum observation filter")

    params_df = pd.DataFrame(params_list)
    n_total = event_returns.groupby([id_col, "evtdate"]).ngroups
    print(
        f"  Estimated {len(params_df)}/{n_total} firm-events "
        f"(mean R^2 = {params_df['r_squared'].dropna().mean():.4f})"
    )
    return params_df
```

### 14.3.8 Step 5: Abnormal Return Computation

Compute AR, CAR, BHAR, SAR, SCAR for each firm-event-date.

```
def compute_abnormal_returns(
    event_returns, params, config, id_col="symbol", date_col="date", ret_col="ret"
):
    """Compute abnormal returns and aggregate to CARs/BHARs.

    Returns
    -------
    daily_ar : pd.DataFrame - daily AR/SAR/CAR/BHAR per firm-event-date
    event_ar : pd.DataFrame - event-level CAR/BHAR/SCAR per firm-event
    """
    model = config.risk_model

    factor_map = {
        RiskModel.MARKET_ADJ: [],
        RiskModel.MARKET_MODEL: ["mkt_excess"],
        RiskModel.FF3: ["mkt_excess", "smb", "hml"],
        RiskModel.CARHART: ["mkt_excess", "smb", "hml", "umd"],
        RiskModel.FF5: ["mkt_excess", "smb", "hml", "rmw", "cma"],
        RiskModel.CUSTOM: config.custom_factors,
    }
    factor_cols = factor_map[model]

    # Filter to event window
    evt = event_returns[
        (event_returns[date_col] >= event_returns["evtwin_beg"])
        & (event_returns[date_col] <= event_returns["evtwin_end"])
    ].copy()

    # Merge params (drop residuals column for merge)
    merge_cols = [c for c in params.columns if c != "_residuals"]
    evt = evt.merge(params[merge_cols], on=[id_col, "evtdate"], how="inner")

    # Expected returns
    if model == RiskModel.MARKET_ADJ:
        evt["expected_ret"] = evt.get("mkt_excess", 0) + evt.get("risk_free", 0)
        evt["AR"] = evt[ret_col] - evt["expected_ret"]
    else:
        evt["expected_ret"] = evt["alpha"]
        for fc in factor_cols:
```

```
            bcol = f"beta_{fc}"
            if bcol in evt.columns:
                evt["expected_ret"] += evt[bcol] * evt[fc]

        if model == RiskModel.MARKET_MODEL:
            evt["AR"] = evt[ret_col] - evt["expected_ret"]
        else:
            evt["AR"] = evt["ret_excess"] - evt["expected_ret"]

    evt["SAR"] = evt["AR"] / evt["sigma"]
    evt = evt.sort_values([id_col, "evtdate", date_col])

    # Compute event time
    all_dates = sorted(event_returns[date_col].unique())
    d2i = {d: i for i, d in enumerate(all_dates)}
    evt["evttime"] = evt[date_col].map(d2i) - evt["evtdate"].map(d2i)

    # Cumulative measures per firm-event
    daily_recs = []
    event_recs = []

    for (firm, evtdate), g in evt.groupby([id_col, "evtdate"]):
        g = g.sort_values(date_col).copy()
        nd = len(g)

        g["CAR"] = g["AR"].cumsum()
        g["cum_ret"] = (1 + g[ret_col]).cumprod() - 1
        g["cum_expected"] = (1 + g["expected_ret"]).cumprod() - 1
        g["BHAR"] = g["cum_ret"] - g["cum_expected"]
        g["SCAR"] = g["CAR"] / (g["sigma"].iloc[0] * np.sqrt(np.arange(1, nd + 1)))

        daily_recs.append(g)

        last = g.iloc[-1]
        sigma = g["sigma"].iloc[0]
        nobs = g["nobs"].iloc[0]

        rec = {
            id_col: firm,
            "evtdate": evtdate,
            "CAR": last["CAR"],
            "BHAR": last["BHAR"],
```

```
            "cum_ret": last["cum_ret"],
            "SCAR": last["CAR"] / (sigma * np.sqrt(nd)),
            "sigma": sigma,
            "variance": g["variance"].iloc[0],
            "nobs": nobs,
            "n_event_days": nd,
            "alpha": g["alpha"].iloc[0],
            "pat_scale": (nobs - 2) / (nobs - 4) if nobs > 4 else np.nan,
            "pos_car": int(last["CAR"] > 0),
        }

        for fc in factor_cols:
            bcol = f"beta_{fc}"
            if bcol in g.columns:
                rec[bcol] = g[bcol].iloc[0]
        if "group" in g.columns:
            rec["group"] = g["group"].iloc[0]

        event_recs.append(rec)

    daily_ar = pd.concat(daily_recs, ignore_index=True)
    event_ar = pd.DataFrame(event_recs)

    print(
        f"  {len(event_ar)} firm-events | Mean CAR: {event_ar['CAR'].mean():.6f} | "
        f"Mean BHAR: {event_ar['BHAR'].mean():.6f} | "
        f"% positive: {event_ar['pos_car'].mean():.1%}"
    )
    return daily_ar, event_ar
```

### 14.3.9 Step 6: Comprehensive Test Statistics

Eight tests covering parametric, non-parametric, and cross-correlation-robust approaches.

```
def compute_test_statistics(event_ar, params=None, group_col=None):
    """Compute comprehensive test statistics for abnormal returns.

    Implements 8 tests with varying assumptions about variance,
    cross-dependence, and distributional form.
    """
    def _stats(data, label=None):
```

```
        N = len(data)
        if N < 3:
            return None

        cars = data['CAR'].values
        bhars = data['BHAR'].values
        scars = data['SCAR'].values
        pos = data['pos_car'].values

        m_car, s_car = np.mean(cars), np.std(cars, ddof=1)
        m_scar, s_scar = np.mean(scars), np.std(scars, ddof=1)

        r = {'group': label or 'All', 'N': N,
             'mean_CAR': m_car, 'median_CAR': np.median(cars),
             'std_CAR': s_car, 'mean_BHAR': np.mean(bhars),
             'pct_positive': np.mean(pos)}

        # 1. Cross-sectional t
        t1 = m_car / (s_car / np.sqrt(N)) if s_car > 0 else np.nan
        r['t_CS'] = t1
        r['p_CS'] = 2 * (1 - stats.t.cdf(abs(t1), N-1)) if np.isfinite(t1) else np.nan

        # 2. Patell Z
        if 'pat_scale' in data.columns:
            ps = data['pat_scale'].dropna().values
            z2 = np.sum(scars[:len(ps)]) / np.sqrt(np.sum(ps)) if len(ps) > 0 else np.nan
        else:
            z2 = m_scar * np.sqrt(N)
        r['Z_Patell'] = z2
        r['p_Patell'] = 2*(1-stats.norm.cdf(abs(z2))) if np.isfinite(z2) else np.nan

        # 3. BMP
        t3 = m_scar / (s_scar / np.sqrt(N)) if s_scar > 0 else np.nan
        r['t_BMP'] = t3
        r['p_BMP'] = 2*(1-stats.t.cdf(abs(t3), N-1)) if np.isfinite(t3) else np.nan

        # 4. Kolari-Pynnönen
        rbar = 0.0
        if params is not None and '_residuals' in params.columns:
            resids = [row['_residuals'] for _, row in params.iterrows()
                      if isinstance(row.get('_residuals'), np.ndarray)]
            if len(resids) > 1:
```

```
            ml = min(len(x) for x in resids)
            aligned = np.column_stack([x[:ml] for x in resids])
            cm = np.corrcoef(aligned.T)
            np.fill_diagonal(cm, 0)
            rbar = cm.sum() / (len(resids) * (len(resids)-1))

    adj = np.sqrt(1/(1+(N-1)*rbar)) if (1+(N-1)*rbar) > 0 else 1
    t4 = t3 * adj if np.isfinite(t3) else np.nan
    r['t_KP'] = t4
    r['p_KP'] = 2*(1-stats.t.cdf(abs(t4), N-1)) if np.isfinite(t4) else np.nan
    r['r_bar'] = rbar

    # 5. Generalized sign test
    p_hat = np.mean(pos)
    z5 = (p_hat - 0.5) / np.sqrt(0.25 / N)
    r['Z_GSign'] = z5
    r['p_GSign'] = 2*(1-stats.norm.cdf(abs(z5)))

    # 6. Sign test
    r['Z_Sign'] = z5  # Same formula with p0=0.5
    r['p_Sign'] = r['p_GSign']

    # 7. Skewness-adjusted t
    if s_scar > 0:
        zb = m_scar / s_scar
        gam = stats.skew(scars)
        t7 = np.sqrt(N) * (zb + gam*zb**2/3 + gam**2*zb**3/27 + gam/(6*N))
        r['t_SkAdj'] = t7
        r['p_SkAdj'] = 2*(1-stats.t.cdf(abs(t7), N-1)) if np.isfinite(t7) else np.nan

    # 8. Wilcoxon signed-rank
    try:
        w, pw = stats.wilcoxon(cars, alternative='two-sided')
        r['W_Wilcoxon'] = w
        r['p_Wilcoxon'] = pw
    except:
        r['W_Wilcoxon'] = r['p_Wilcoxon'] = np.nan

    return r

results = [_stats(event_ar)]
if group_col and group_col in event_ar.columns:
```

```
        for gv, gd in event_ar.groupby(group_col):
            s = _stats(gd, label=gv)
            if s:
                results.append(s)

    return pd.DataFrame([r for r in results if r is not None])


def compute_daily_stats(daily_ar, id_col='symbol'):
    """Compute test statistics at each event time t."""
    rows = []
    for t, g in daily_ar.groupby('evttime'):
        n = g[id_col].nunique()
        if n < 2:
            continue
        m_ar = g['AR'].mean()
        s_ar = g['AR'].std(ddof=1)
        t_ar = m_ar / (s_ar/np.sqrt(n)) if s_ar > 0 else np.nan
        rows.append({'evttime': t, 'N': n, 'mean_AR': m_ar,
                     'mean_CAR': g['CAR'].mean(), 'mean_BHAR': g['BHAR'].mean(),
                     'mean_cum_ret': g.get('cum_ret', pd.Series()).mean(),
                     't_AR': t_ar})
    return pd.DataFrame(rows).sort_values('evttime')
```

### 14.3.10 Step 7: Publication-Ready Visualization

```
def plot_event_study(daily_stats, title="Cumulative Abnormal Returns Around Event Date",
                     figsize=(12, 7), save_path=None):
    """Publication-ready event study plot with CAR, BHAR, and daily AR panels."""
    fig, axes = plt.subplots(2, 1, figsize=figsize, height_ratios=[3, 1],
                             gridspec_kw={'hspace': 0.05})
    ds = daily_stats.sort_values('evttime')
    t = ds['evttime'].values

    # Top: cumulative returns
    ax = axes[0]
    ax.plot(t, ds['mean_CAR']*100, color='#2166AC', lw=2.5, label='Mean CAR')
    ax.plot(t, ds['mean_BHAR']*100, color='#B2182B', lw=2, ls='--', label='Mean BHAR')
    if 'mean_cum_ret' in ds.columns:
        ax.plot(t, ds['mean_cum_ret']*100, color='#666', lw=1.5, ls=':',
```

```
                label='Mean Cum. Return', alpha=0.7)
    ax.axvline(0, color='k', lw=0.8, alpha=0.5)
    ax.axhline(0, color='k', lw=0.5, alpha=0.3)
    ax.set_ylabel('Cumulative Return (%)', fontsize=12)
    ax.set_title(title, fontsize=14, fontweight='bold')
    ax.legend(loc='upper left', fontsize=10)
    ax.grid(True, alpha=0.2)
    ax.set_xticklabels([])

    # Bottom: daily AR bars
    ax2 = axes[1]
    colors = ['#2166AC' if v >= 0 else '#B2182B' for v in ds['mean_AR']]
    ax2.bar(t, ds['mean_AR']*100, color=colors, alpha=0.7, width=0.8)
    if 't_AR' in ds.columns:
        sig = np.abs(ds['t_AR'].values) > 1.96
        if sig.any():
            ax2.scatter(t[sig], ds['mean_AR'].values[sig]*100,
                        color='gold', s=40, marker='*', zorder=4, label='p<0.05')
            ax2.legend(fontsize=9)
    ax2.axvline(0, color='k', lw=0.8, alpha=0.5)
    ax2.axhline(0, color='k', lw=0.5, alpha=0.3)
    ax2.set_xlabel('Event Time (Trading Periods)', fontsize=12)
    ax2.set_ylabel('Mean AR (%)', fontsize=10)
    ax2.grid(True, alpha=0.2)

    for a in axes:
        a.spines['top'].set_visible(False)
        a.spines['right'].set_visible(False)
    plt.tight_layout()
    if save_path:
        fig.savefig(save_path, dpi=300, bbox_inches='tight')
    return fig


def plot_car_distribution(event_ar, var='CAR', figsize=(12, 5)):
    """Cross-sectional distribution of CARs with histogram and QQ plot."""
    fig, (ax1, ax2) = plt.subplots(1, 2, figsize=figsize)
    data = event_ar[var].dropna() * 100

    ax1.hist(data, bins=50, density=True, alpha=0.6, color='#2166AC', edgecolor='white')
    ax1.axvline(data.mean(), color='k', ls='--', lw=1.5,
                label=f'Mean={data.mean():.2f}%')
```

```
    ax1.axvline(data.median(), color='gray', ls=':', lw=1.5,
                label=f'Median={data.median():.2f}%')
    ax1.set_xlabel(f'{var} (%)')
    ax1.set_ylabel('Density')
    ax1.set_title(f'Distribution of {var}', fontweight='bold')
    ax1.legend()
    ax1.spines['top'].set_visible(False)
    ax1.spines['right'].set_visible(False)

    # QQ plot
    (osm, osr), (slope, intercept, r) = stats.probplot(data, dist='norm')
    ax2.scatter(osm, osr, alpha=0.4, s=10, color='#2166AC')
    ax2.plot(osm, slope*np.array(osm)+intercept, 'r--', lw=1)
    ax2.set_xlabel('Theoretical Quantiles')
    ax2.set_ylabel('Sample Quantiles')
    ax2.set_title('Q-Q Plot (Normal)', fontweight='bold')
    ax2.spines['top'].set_visible(False)
    ax2.spines['right'].set_visible(False)

    plt.tight_layout()
    return fig
```

### 14.3.11 The Master Pipeline

Combine all components into one function:

```
def run_event_study(events, prices, factors, config,
                    id_col='symbol', date_col='date', ret_col='ret',
                    event_date_col='event_date', mkt_col='mkt_excess',
                    rf_col='risk_free', group_col=None, verbose=True):
    """Run a complete event study from raw inputs to test statistics.

    This is the main entry point. Provide your events, price data,
    factor data, and configuration-get back everything you need.

    Parameters
    ----------
    events : pd.DataFrame
        Columns: [id_col, event_date_col], optional 'group'.
    prices : pd.DataFrame
        Daily returns: [id_col, date_col, ret_col or 'ret_excess', rf_col].
```

```
    factors : pd.DataFrame
        Factor returns: [date_col, mkt_col, 'smb', 'hml', ...].
    config : EventStudyConfig

    Returns
    -------
    dict with keys: 'config', 'daily_ar', 'event_ar', 'daily_stats',
        'test_stats', 'params'
    """
    config.validate()

    if verbose:
        print(f"   Event Study: {config.risk_model.value} model   ")
        print(f"  Windows: estimation={config.estimation_window}, "
              f"gap={config.gap}, event=({config.event_window_start},{config.event_window_en
        print(f"  Min obs: {config.min_estimation_obs}\n")

    # 1. Trading calendar
    if verbose: print("Step 1: Building trading calendar...")
    trading_dates = pd.Series(sorted(prices[date_col].unique()))
    calendar = build_trading_calendar(trading_dates, config)
    if verbose: print(f"  {len(calendar)} potential event dates\n")

    # 2. Align events
    if verbose: print("Step 2: Aligning events to trading calendar...")
    aligned = align_events(events, calendar, id_col, event_date_col)
    if verbose: print(f"  {len(aligned)} aligned events\n")

    # 3. Extract returns
    if verbose: print("Step 3: Extracting returns and merging factors...")
    evt_rets = extract_returns(aligned, prices, factors, config,
                               id_col, date_col, ret_col, mkt_col, rf_col)
    if verbose: print()

    # 4. Estimate model
    if verbose: print("Step 4: Estimating risk model parameters...")
    params = estimate_model(evt_rets, config, id_col, date_col, ret_col)
    if verbose: print()

    # 5. Compute abnormal returns
    if verbose: print("Step 5: Computing abnormal returns...")
    daily_ar, event_ar = compute_abnormal_returns(
```

```
        evt_rets, params, config, id_col, date_col, ret_col)
    if verbose: print()

    # 6. Test statistics
    if verbose: print("Step 6: Computing test statistics...")
    test_stats = compute_test_statistics(event_ar, params, group_col)
    daily_stats = compute_daily_stats(daily_ar, id_col)
    if verbose:
        print(f"  Done.\n")
        print("   Results Summary   ")
        cols = ['group', 'N', 'mean_CAR', 'mean_BHAR', 'pct_positive',
                't_CS', 'p_CS', 't_BMP', 'p_BMP', 't_KP', 'p_KP']
        avail = [c for c in cols if c in test_stats.columns]
        print(test_stats[avail].to_string(index=False))

    return {
        'config': config,
        'params': params,
        'daily_ar': daily_ar,
        'event_ar': event_ar,
        'daily_stats': daily_stats,
        'test_stats': test_stats,
        'calendar': calendar,
    }

print("Master pipeline ready.")
```

```
Master pipeline ready.
```

## 14.4 Demonstration with Simulated Data

Since we are building a general-purpose framework (the actual event data will be supplied later), we demonstrate the full pipeline with **realistic simulated data**.

```
np.random.seed(2024)

# --- Simulated trading calendar (Vietnamese market: ~245 days/year) ---
dates = pd.bdate_range('2019-01-01', '2023-12-31', freq='B')
# Remove Tet + national holidays (simplified)
tet_holidays = pd.to_datetime([
```

```
    '2019-02-04','2019-02-05','2019-02-06','2019-02-07','2019-02-08',
    '2020-01-23','2020-01-24','2020-01-27','2020-01-28','2020-01-29',
    '2021-02-10','2021-02-11','2021-02-12','2021-02-15','2021-02-16',
    '2022-01-31','2022-02-01','2022-02-02','2022-02-03','2022-02-04',
    '2023-01-20','2023-01-23','2023-01-24','2023-01-25','2023-01-26',
])
dates = dates.difference(tet_holidays)
T = len(dates)

# --- Simulated factors (realistic Vietnamese market parameters) ---
rf_daily = 0.04 / 252  # ~4% annual risk-free
mkt_excess = np.random.normal(0.0003, 0.012, T)  # ~7.5% annual, ~19% vol
smb = np.random.normal(0.0001, 0.006, T)
hml = np.random.normal(0.0001, 0.005, T)
rmw = np.random.normal(0.00005, 0.004, T)
cma = np.random.normal(0.00005, 0.004, T)

factors_sim = pd.DataFrame({
    'date': dates, 'mkt_excess': mkt_excess, 'smb': smb, 'hml': hml,
    'rmw': rmw, 'cma': cma, 'risk_free': rf_daily
})

# --- 100 simulated stocks ---
n_stocks = 100
symbols = [f'SIM{i:03d}' for i in range(n_stocks)]
betas = np.random.uniform(0.5, 1.5, n_stocks)
alphas = np.random.normal(0, 0.0002, n_stocks)
idio_vols = np.random.uniform(0.015, 0.035, n_stocks)

price_rows = []
for i, sym in enumerate(symbols):
    eps = np.random.normal(0, idio_vols[i], T)
    rets = alphas[i] + betas[i] * mkt_excess + 0.3*smb + 0.2*hml + eps
    for j in range(T):
        price_rows.append({
            'symbol': sym, 'date': dates[j], 'ret': rets[j],
            'ret_excess': rets[j] - rf_daily,
            'risk_free': rf_daily,
            'mktcap': np.random.uniform(100, 5000),
        })

prices_sim = pd.DataFrame(price_rows)
```

```
# --- Simulated events: 50 random firm-dates with KNOWN positive AR ---
event_indices = np.random.choice(range(250, T-50), 50, replace=False)
event_firms = np.random.choice(symbols, 50, replace=True)
event_dates_sim = [dates[i] for i in event_indices]

# Inject abnormal returns on event date (2% positive shock)
for firm, edate in zip(event_firms, event_dates_sim):
    mask = (prices_sim['symbol'] == firm) & (prices_sim['date'] == edate)
    prices_sim.loc[mask, 'ret'] += 0.02
    prices_sim.loc[mask, 'ret_excess'] += 0.02

events_sim = pd.DataFrame({
    'symbol': event_firms,
    'event_date': event_dates_sim,
    'group': np.random.choice([1, 2], 50)
})

print(f"Simulated data: {n_stocks} stocks × {T} days = {len(prices_sim):,} obs")
print(f"Events: {len(events_sim)} firm-event pairs")
print(f"Injected abnormal return: +2% on event date")
```

```
Simulated data: 100 stocks × 1279 days = 127,900 obs
Events: 50 firm-event pairs
Injected abnormal return: +2% on event date
```

### 14.4.1 Running the Full Pipeline

```
config = EventStudyConfig(
    estimation_window=150,
    event_window_start=-10,
    event_window_end=10,
    gap=15,
    min_estimation_obs=120,
    risk_model=RiskModel.FF3
)

results = run_event_study(
    events=events_sim,
    prices=prices_sim,
```

```
    factors=factors_sim,
    config=config,
    group_col='group'
)
```

```
   Event Study: ff3 model
  Windows: estimation=150, gap=15, event=(-10,10)
  Min obs: 120

Step 1: Building trading calendar...
  1094 potential event dates

Step 2: Aligning events to trading calendar...
  50 aligned events

Step 3: Extracting returns and merging factors...
  Extracted 9,300 obs for 50 firm-events

Step 4: Estimating risk model parameters...
  Estimated 50/50 firm-events (mean R^2 = 0.2368)

Step 5: Computing abnormal returns...
  50 firm-events | Mean CAR: 0.033009 | Mean BHAR: 0.032498 | % positive: 60.0%

Step 6: Computing test statistics...
  Done.

   Results Summary
group  N  mean_CAR  mean_BHAR  pct_positive     t_CS     p_CS    t_BMP    p_BMP     t_KP
  All 50  0.033009   0.032498      0.600000 2.288362 0.026468 2.157106 0.035929 2.161291 0.03
    1 20  0.030704   0.028326      0.650000 1.568676 0.133227 1.248382 0.227056 1.249320 0.22
    2 30  0.034545   0.035280      0.566667 1.688848 0.101975 1.734699 0.093413 1.736689 0.09
```

### 14.4.2 Visualizing Results

```
fig1 = plot_event_study(
    results['daily_stats'],
    title="Event Study: FF3 Model - Simulated Vietnamese Market"
)
plt.show()
```

Figure 14.1: Dynamics of cumulative abnormal returns (CARs) and buy-and-hold abnormal returns (BHARs) around the event date. The positive jump at t=0 reflects the injected 2% abnormal return.

```
fig2 = plot_car_distribution(results['event_ar'], 'CAR')
plt.show()
```

Figure 14.2: Cross-sectional distribution of cumulative abnormal returns. The rightward shift from zero and positive skewness are consistent with the injected positive event effect.

### 14.4.3 Complete Test Statistics

### 14.4.4 Running Multiple Models for Robustness

A key best practice is to report results across multiple risk models. If conclusions are robust across models, this strengthens the findings:

```
models_to_run = [
    ("Market-Adjusted", RiskModel.MARKET_ADJ),
    ("Market Model", RiskModel.MARKET_MODEL),
    ("Fama-French 3", RiskModel.FF3),
    ("Fama-French 5", RiskModel.FF5),
]

robustness = []
for name, mdl in models_to_run:
    cfg = EventStudyConfig(
        estimation_window=150, event_window_start=-10, event_window_end=10,
        gap=15, min_estimation_obs=120, risk_model=mdl
    )
    res = run_event_study(events_sim, prices_sim, factors_sim, cfg, verbose=False)
    ts = res['test_stats']
```

Table 14.3: Event study test statistics for the full sample and by subgroup

```
# Format for display
ts = results['test_stats'].copy()

# Select key columns
display_cols = ['group', 'N', 'mean_CAR', 'mean_BHAR', 'pct_positive',
                't_CS', 'p_CS', 'Z_Patell', 'p_Patell',
                't_BMP', 'p_BMP', 't_KP', 'p_KP',
                'Z_GSign', 'p_GSign', 't_SkAdj', 'p_SkAdj']
avail = [c for c in display_cols if c in ts.columns]
display_df = ts[avail].copy()

# Format
for c in display_df.columns:
    if c in ['N']:
        display_df[c] = display_df[c].astype(int)
    elif c.startswith('p_'):
        display_df[c] = display_df[c].map(lambda x: f'{x:.4f}' if pd.notna(x) else '')
    elif c in ['mean_CAR', 'mean_BHAR']:
        display_df[c] = display_df[c].map(lambda x: f'{x:.4%}' if pd.notna(x) else '')
    elif c == 'pct_positive':
        display_df[c] = display_df[c].map(lambda x: f'{x:.1%}' if pd.notna(x) else '')
    elif isinstance(display_df[c].iloc[0], (int, float, np.floating)):
        display_df[c] = display_df[c].map(lambda x: f'{x:.3f}' if pd.notna(x) else '')

print(display_df.to_string(index=False))
```

```
group  N mean_CAR mean_BHAR pct_positive  t_CS   p_CS Z_Patell p_Patell t_BMP  p_BMP  t_KP
  All 50  3.3009%   3.2498%        60.0% 2.288 0.0265    1.832   0.0669 2.157 0.0359 2.161 0
    1 20  3.0704%   2.8326%        65.0% 1.569 0.1332    1.020   0.3075 1.248 0.2271 1.249 0
    2 30  3.4545%   3.5280%        56.7% 1.689 0.1020    1.532   0.1255 1.735 0.0934 1.737 0
```

```
    full = ts[ts['group'] == 'All'].iloc[0]
    robustness.append({
        'Model': name,
        'N': int(full['N']),
        'Mean CAR': f"{full['mean_CAR']:.4%}",
        'Mean BHAR': f"{full['mean_BHAR']:.4%}",
        '% Positive': f"{full['pct_positive']:.1%}",
        't (CS)': f"{full['t_CS']:.2f}",
        't (BMP)': f"{full['t_BMP']:.2f}",
        't (KP)': f"{full.get('t_KP', np.nan):.2f}",
    })

rob_df = pd.DataFrame(robustness)
print("Robustness Across Risk Models:")
print(rob_df.to_string(index=False))
```

```
  Extracted 9,300 obs for 50 firm-events
  Estimated 50/50 firm-events (mean R^2 = 0.0000)
  50 firm-events | Mean CAR: 0.029672 | Mean BHAR: 0.026468 | % positive: 62.0%
  Extracted 9,300 obs for 50 firm-events
  Estimated 50/50 firm-events (mean R^2 = 0.2198)
  50 firm-events | Mean CAR: 0.033785 | Mean BHAR: 0.029974 | % positive: 62.0%
  Extracted 9,300 obs for 50 firm-events
  Estimated 50/50 firm-events (mean R^2 = 0.2368)
  50 firm-events | Mean CAR: 0.033009 | Mean BHAR: 0.032498 | % positive: 60.0%
  Extracted 9,300 obs for 50 firm-events
  Estimated 50/50 firm-events (mean R^2 = 0.2516)
  50 firm-events | Mean CAR: 0.036479 | Mean BHAR: 0.036000 | % positive: 64.0%
Robustness Across Risk Models:
          Model  N Mean CAR Mean BHAR % Positive t (CS) t (BMP) t (KP)
Market-Adjusted 50  2.9672%   2.6468%      62.0%   2.12    2.03   2.01
   Market Model 50  3.3785%   2.9974%      62.0%   2.37    2.26   2.28
  Fama-French 3 50  3.3009%   3.2498%      60.0%   2.29    2.16   2.16
  Fama-French 5 50  3.6479%   3.6000%      64.0%   2.51    2.36   2.41
```

## 14.5 How to Use This Framework with Your Data

### 14.5.1 Required Data Format

To run the event study on real Vietnamese market data, prepare three inputs:

1. **Stock Returns** (`prices` DataFrame):

| Column | Description | Example |
|---|---|---|
| `symbol` | Stock ticker | `'VNM'` |
| `date` | Trading date | `2023-06-15` |
| `ret` or `ret_excess` | Daily return (decimal) | `0.0123` |
| `risk_free` | Daily risk-free rate | `0.000159` |

2. **Factor Returns** (`factors` DataFrame):

| Column | Description |
|---|---|
| `date` | Trading date |
| `mkt_excess` | Market excess return |
| `smb` | Size factor (FF3/FF5) |
| `hml` | Value factor (FF3/FF5) |
| `rmw` | Profitability factor (FF5) |
| `cma` | Investment factor (FF5) |
| `risk_free` | Risk-free rate |

3. **Event File** (`events` DataFrame):

| Column | Description | Example |
|---|---|---|
| `symbol` | Stock ticker | `'VNM'` |
| `event_date` | Event date | `2023-03-15` |
| `group` | (Optional) subgroup | `1` |

### 14.5.2 Minimal Usage Example

```
# Load your data
prices = pd.read_csv('prices_daily.csv', parse_dates=['date'])
factors = pd.read_csv('factors_ff3_daily.csv', parse_dates=['date'])
events = pd.read_csv('my_events.csv', parse_dates=['event_date'])

# Configure
config = EventStudyConfig(
    estimation_window=150,
    event_window_start=-5,
```

```
    event_window_end=5,
    gap=15,
    min_estimation_obs=120,
    risk_model=RiskModel.FF3
)

# Run
results = run_event_study(events, prices, factors, config)

# Access outputs
results['test_stats']    # Test statistics
results['event_ar']      # Firm-level CARs/BHARs
results['daily_ar']      # Daily abnormal returns
results['daily_stats']   # Event-time aggregates

# Plot
plot_event_study(results['daily_stats'], title="My Event Study")
```

## 14.6 Demonstration with Vietnamese Market Data

We now demonstrate the full event study pipeline using actual Vietnamese stock market data. The datasets available are:

- `prices_daily`: `symbol`, `date`, `ret_excess`, `mktcap`, `mktcap_lag`, `risk_free`
- `prices_monthly`: same structure
- `factors_ff3_daily`: `date`, `smb`, `hml`, `mkt_excess`, `risk_free`
- `factors_ff3_monthly` — monthly frequency version
- `factors_ff5_daily`: `date`, `smb`, `hml`, `mkt_excess`, `risk_free`, `rmw`, `cma`
- `factors_ff5_monthly`

Since our data provides `ret_excess` rather than raw returns, we recover raw returns as $R_{it} = R^e_{it} + R_{f,t}$, and the market return as $R_{m,t} = R^e_{m,t} + R_{f,t}$. The `extract_event_returns()` function handles this automatically.

### 14.6.1 Loading the Data

```
# --- Recover raw returns ---
# ret = ret_excess + risk_free
prices_daily['ret'] = prices_daily['ret_excess'] + prices_daily['risk_free']
```

```python
prices_monthly['ret'] = prices_monthly['ret_excess'] + prices_monthly['risk_free']

# --- Inspect the data ---
print("=" * 70)
print("VIETNAMESE MARKET DATA SUMMARY")
print("=" * 70)
print(f"\nprices_daily: {prices_daily.shape[0]:,} rows, "
      f"{prices_daily['symbol'].nunique()} stocks, "
      f"{prices_daily['date'].min().date()} to {prices_daily['date'].max().date()}")
print(f"prices_monthly: {prices_monthly.shape[0]:,} rows, "
      f"{prices_monthly['symbol'].nunique()} stocks")
print(f"\nfactors_ff3_daily: {factors_ff3_daily.shape[0]:,} trading days")
print(f"  Columns: {list(factors_ff3_daily.columns)}")
print(f"factors_ff5_daily: {factors_ff5_daily.shape[0]:,} trading days")
print(f"  Columns: {list(factors_ff5_daily.columns)}")
print(f"\nSample daily returns:")
print(prices_daily[['symbol', 'date', 'ret_excess', 'ret', 'risk_free', 'mktcap']]
      .head(5).to_string(index=False))
print(f"\nSample daily factors:")
print(factors_ff3_daily.head(5).to_string(index=False))
```

```
======================================================================
VIETNAMESE MARKET DATA SUMMARY
======================================================================

prices_daily: 3,462,157 rows, 1459 stocks, 2010-01-05 to 2023-12-29
prices_monthly: 165,499 rows, 1457 stocks

factors_ff3_daily: 3,126 trading days
  Columns: ['date', 'smb', 'hml', 'mkt_excess', 'risk_free']
factors_ff5_daily: 3,126 trading days
  Columns: ['date', 'smb', 'hml', 'rmw', 'cma', 'mkt_excess', 'risk_free']

Sample daily returns:
symbol       date  ret_excess      ret  risk_free     mktcap
   A32 2018-10-24   -0.000159 0.000000   0.000159 176.120000
   A32 2018-10-25   -0.000159 0.000000   0.000159 176.120000
   A32 2018-10-26   -0.000159 0.000000   0.000159 176.120000
   A32 2018-10-29   -0.000159 0.000000   0.000159 176.120000
   A32 2018-10-30   -0.000159 0.000000   0.000159 176.120000

Sample daily factors:
```

```
      date       smb       hml  mkt_excess  risk_free
2011-07-01  0.008587  0.000967   -0.019862   0.000159
2011-07-04  0.005099 -0.001099   -0.000633   0.000159
2011-07-05 -0.009088  0.010152    0.013314   0.000159
2011-07-06  0.004875 -0.003918   -0.008045   0.000159
2011-07-07 -0.011239 -0.000584    0.003391   0.000159
```

### 14.6.2 Creating Sample Events

For this demonstration, we create a sample event file. In practice, events would come from corporate announcements (earnings, M&A, dividends), regulatory changes, or other information shocks. Here we select 50 large-cap Vietnamese stocks and assign random event dates from the most recent two years of data to illustrate the pipeline mechanics.

```python
np.random.seed(2024)

# Select the 50 largest stocks by median market cap
largest = (prices_daily.groupby('symbol')['mktcap']
           .median()
           .nlargest(50)
           .index.tolist())

# Date range for events: last 2 years of data, with buffer for windows
date_range = prices_daily['date'].sort_values().unique()
n_dates = len(date_range)
# Events from the middle portion (need room for estimation + event windows)
event_eligible = date_range[int(n_dates * 0.3):int(n_dates * 0.85)]

# Generate 50 random firm-event pairs
event_firms = np.random.choice(largest, 50, replace=True)
event_dates = np.random.choice(event_eligible, 50, replace=False)

events_demo = pd.DataFrame({
    'symbol': event_firms,
    'event_date': pd.to_datetime(event_dates),
    'group': np.random.choice(['Group_A', 'Group_B'], 50)
})

# Remove any duplicate firm-date pairs
events_demo = events_demo.drop_duplicates(subset=['symbol', 'event_date'])

print(f"Sample event file: {len(events_demo)} firm-event observations")
```

```
print(f"Unique firms: {events_demo['symbol'].nunique()}")
print(f"Date range: {events_demo['event_date'].min().date()} to "
      f"{events_demo['event_date'].max().date()}")
print(f"\nGroup distribution:")
print(events_demo['group'].value_counts().to_string())
print(f"\nFirst 10 events:")
print(events_demo.sort_values('event_date').head(10).to_string(index=False))
```

```
Sample event file: 50 firm-event observations
Unique firms: 35
Date range: 2014-06-25 to 2021-10-29

Group distribution:
group
Group_B    26
Group_A    24

First 10 events:
symbol event_date   group
   MCH 2014-06-25 Group_B
   SIP 2014-10-23 Group_B
   VRE 2014-11-14 Group_B
   QNS 2014-12-25 Group_A
   FOX 2015-01-16 Group_B
   THD 2015-01-26 Group_A
   QNS 2015-02-12 Group_A
   HNG 2015-05-07 Group_B
   MML 2015-08-17 Group_B
   ACV 2015-10-15 Group_A
```

### 14.6.3 Daily Event Study: Fama-French 3-Factor Model

```
config_ff3 = EventStudyConfig(
    estimation_window=150,
    event_window_start=-10,
    event_window_end=10,
    gap=15,
    min_estimation_obs=120,
    risk_model=RiskModel.FF3
)
```

```
results_ff3 = run_event_study(
    events=events_demo,
    prices=prices_daily,
    factors=factors_ff3_daily,
    config=config_ff3,
    group_col='group'
)
```

```
   Event Study: ff3 model
  Windows: estimation=150, gap=15, event=(-10,10)
  Min obs: 120

Step 1: Building trading calendar...
  3308 potential event dates

Step 2: Aligning events to trading calendar...
  50 aligned events

Step 3: Extracting returns and merging factors...
  Extracted 5,421 obs for 30 firm-events

Step 4: Estimating risk model parameters...
  Estimated 26/30 firm-events (mean R^2 = 0.2245)

Step 5: Computing abnormal returns...
  26 firm-events | Mean CAR: 0.021513 | Mean BHAR: 0.024885 | % positive: 61.5%

Step 6: Computing test statistics...
  Done.

   Results Summary
  group  N  mean_CAR  mean_BHAR  pct_positive     t_CS     p_CS    t_BMP    p_BMP     t_KP
    All 26  0.021513   0.024885      0.615385 0.774999 0.445609 0.726797 0.474101 0.699973 0
Group_A 13  0.008601   0.006783      0.538462 0.211606 0.835966 0.577574 0.574229 0.567041 0
Group_B 13  0.034425   0.042987      0.692308 0.879892 0.396197 0.437902 0.669236 0.429917 0
```

### 14.6.4 Visualizing Daily Results

```
fig1 = plot_event_study(
    results_ff3['daily_stats'],
```

```
    title="Event Study: Fama-French 3-Factor Model - Vietnamese Market (Daily)"
)
plt.show()
```

Figure 14.3: Cumulative abnormal returns around event dates for Vietnamese stocks using the Fama-French 3-factor model. The event window spans [-10, +10] trading days.

```
fig2 = plot_car_distribution(results_ff3['event_ar'], 'CAR')
plt.show()
```

Figure 14.4: Cross-sectional distribution of cumulative abnormal returns (CARs) across firm-events. The histogram and Q-Q plot assess normality assumptions underlying parametric tests.

### 14.6.5 Complete Test Statistics (Daily)

### 14.6.6 Robustness: Multiple Risk Models (Daily)

### 14.6.7 Robustness: Multiple Event Windows

A key practice is to examine sensitivity to the event window specification:

### 14.6.8 Monthly Event Study: Fama-French 3-Factor Model

For longer-horizon studies, monthly frequency is appropriate. Note that the estimation window is specified in months rather than days:

```
# Create monthly events aligned to the monthly data
# Map daily event dates to the corresponding month-end
events_monthly = events_demo.copy()
events_monthly['event_date'] = events_monthly['event_date'].dt.to_period('M').dt.to_timestamp

# Use month-end dates from monthly prices
monthly_dates = prices_monthly['date'].sort_values().unique()
```

Table 14.7: Event study test statistics for the full sample and by subgroup — Daily frequency, FF3 model

```
ts = results_ff3['test_stats'].copy()

display_cols = ['group', 'N', 'mean_CAR', 'median_CAR', 'mean_BHAR', 'pct_positive',
                't_CS', 'p_CS', 'Z_Patell', 'p_Patell',
                't_BMP', 'p_BMP', 't_KP', 'p_KP',
                'Z_GSign', 'p_GSign', 't_SkAdj', 'p_SkAdj',
                'W_Wilcoxon', 'p_Wilcoxon']
avail = [c for c in display_cols if c in ts.columns]
display_df = ts[avail].copy()

for c in display_df.columns:
    if c in ['N']:
        display_df[c] = display_df[c].astype(int)
    elif c == 'group':
        continue
    elif c.startswith('p_'):
        display_df[c] = display_df[c].map(lambda x: f'{x:.4f}' if pd.notna(x) else '')
    elif c in ['mean_CAR', 'median_CAR', 'mean_BHAR']:
        display_df[c] = display_df[c].map(lambda x: f'{x:.4%}' if pd.notna(x) else '')
    elif c == 'pct_positive':
        display_df[c] = display_df[c].map(lambda x: f'{x:.1%}' if pd.notna(x) else '')
    elif isinstance(display_df[c].iloc[0], (int, float, np.floating)):
        display_df[c] = display_df[c].map(lambda x: f'{x:.3f}' if pd.notna(x) else '')

print(display_df.to_string(index=False))
```

```
  group  N mean_CAR median_CAR mean_BHAR pct_positive  t_CS   p_CS Z_Patell p_Patell t_BMP  p
    All 26  2.1513%    1.5701%   2.4885%        61.5% 0.775 0.4456    0.961   0.3364 0.727 0
Group_A 13  0.8601%    2.5330%   0.6783%        53.8% 0.212 0.8360    0.739   0.4596 0.578 0
Group_B 13  3.4425%    1.4169%   4.2987%        69.2% 0.880 0.3962    0.620   0.5352 0.438 0
```

Table 14.8: Robustness of event study results across risk models — Daily frequency

```
models_daily = [
    ("Market-Adjusted",  RiskModel.MARKET_ADJ,    factors_ff3_daily),
    ("Market Model",     RiskModel.MARKET_MODEL,   factors_ff3_daily),
    ("Fama-French 3",    RiskModel.FF3,            factors_ff3_daily),
    ("Fama-French 5",    RiskModel.FF5,            factors_ff5_daily),
]

robustness_daily = []
for name, mdl, facs in models_daily:
    cfg = EventStudyConfig(
        estimation_window=150, event_window_start=-10, event_window_end=10,
        gap=15, min_estimation_obs=120, risk_model=mdl
    )
    res = run_event_study(events_demo, prices_daily, facs, cfg, verbose=False)
    ts = res['test_stats']
    full = ts[ts['group'] == 'All'].iloc[0]
    robustness_daily.append({
        'Model': name,
        'N': int(full['N']),
        'Mean CAR': f"{full['mean_CAR']:.4%}",
        'Median CAR': f"{full['median_CAR']:.4%}",
        'Mean BHAR': f"{full['mean_BHAR']:.4%}",
        '% Positive': f"{full['pct_positive']:.1%}",
        't (CS)': f"{full['t_CS']:.3f}",
        'p (CS)': f"{full['p_CS']:.4f}",
        't (BMP)': f"{full['t_BMP']:.3f}",
        'p (BMP)': f"{full['p_BMP']:.4f}",
        't (KP)': f"{full.get('t_KP', np.nan):.3f}",
        'p (KP)': f"{full.get('p_KP', np.nan):.4f}",
    })

rob_daily_df = pd.DataFrame(robustness_daily)
print("Robustness Across Risk Models (Daily Frequency)")
print("=" * 100)
print(rob_daily_df.to_string(index=False))
```

```
  Extracted 5,421 obs for 30 firm-events
  Estimated 28/30 firm-events (mean R^2 = 0.0000)
  28 firm-events | Mean CAR: 0.035338 | Mean BHAR: 0.036936 | % positive: 50.0%
  Extracted 5,421 obs for 30 firm-events
  Estimated 26/30 firm-events (mean R^2 = 0.1960)
  26 firm-events | Mean CAR: 0.032107 | Mean BHAR: 0.033221 | % positive: 61.5%
  Extracted 5,421 obs for 30 firm-events
  Estimated 26/30 firm-events (mean R^2 = 0.2245)
  26 firm-events | Mean CAR: 0.021513 | Mean BHAR: 0.024885 | % positive: 61.5%
  Extracted 5,421 obs for 30 firm-events
  Estimated 26/30 firm-events (mean R^2 = 0.2675)
  26 firm-events | Mean CAR: 0.025684 | Mean BHAR: 0.028399 | % positive: 57.7%
Robustness Across Risk Models (Daily Frequency)
====================================================================================
```

Table 14.9: Sensitivity of results to event window specification

```
windows = [
    ("(-1, +1)",  -1, 1),
    ("(-3, +3)",  -3, 3),
    ("(-5, +5)",  -5, 5),
    ("(-10, +10)", -10, 10),
    ("(-1, +5)",  -1, 5),
    ("(-5, +1)",  -5, 1),
    ("(0, 0)",     0, 0),
]

window_results = []
for label, ws, we in windows:
    cfg = EventStudyConfig(
        estimation_window=150, event_window_start=ws, event_window_end=we,
        gap=15, min_estimation_obs=120, risk_model=RiskModel.FF3
    )
    res = run_event_study(events_demo, prices_daily, factors_ff3_daily, cfg, verbose=False)
    ts = res['test_stats']
    full = ts[ts['group'] == 'All'].iloc[0]
    window_results.append({
        'Window': label,
        'Days': we - ws + 1,
        'N': int(full['N']),
        'Mean CAR': f"{full['mean_CAR']:.4%}",
        'Mean BHAR': f"{full['mean_BHAR']:.4%}",
        '% Positive': f"{full['pct_positive']:.1%}",
        't (CS)': f"{full['t_CS']:.3f}",
        't (BMP)': f"{full['t_BMP']:.3f}",
        'p (BMP)': f"{full['p_BMP']:.4f}",
    })

win_df = pd.DataFrame(window_results)
print("Sensitivity to Event Window Specification (FF3 Model)")
print("=" * 90)
print(win_df.to_string(index=False))
```

```
  Extracted 4,899 obs for 30 firm-events
  Estimated 26/30 firm-events (mean R^2 = 0.2147)
  26 firm-events | Mean CAR: 0.004074 | Mean BHAR: 0.004648 | % positive: 50.0%
  Extracted 5,015 obs for 30 firm-events
  Estimated 26/30 firm-events (mean R^2 = 0.2155)
  26 firm-events | Mean CAR: 0.003761 | Mean BHAR: 0.004327 | % positive: 42.3%
  Extracted 5,131 obs for 30 firm-events
  Estimated 26/30 firm-events (mean R^2 = 0.2217)
  26 firm-events | Mean CAR: -0.001133 | Mean BHAR: 0.001027 | % positive: 42.3%
  Extracted 5,421 obs for 30 firm-events
  Estimated 26/30 firm-events (mean R^2 = 0.2245)
  26 firm-events | Mean CAR: 0.021513 | Mean BHAR: 0.024885 | % positive: 61.5%
  Extracted 5,019 obs for 30 firm-events
  Estimated 26/30 firm-events (mean R^2 = 0.2147)
  26 firm-events | Mean CAR: 0.005006 | Mean BHAR:
```

```python
# Filter events to dates present in monthly data
events_monthly = events_monthly[events_monthly['event_date'].isin(monthly_dates)]
events_monthly = events_monthly.drop_duplicates(subset=['symbol', 'event_date'])

config_monthly = EventStudyConfig(
    estimation_window=36,     # 36 months
    event_window_start=-3,    # 3 months before
    event_window_end=3,       # 3 months after
    gap=3,                    # 3-month gap
    min_estimation_obs=24,    # At least 24 months
    risk_model=RiskModel.FF3
)

if len(events_monthly) > 0:
    results_monthly = run_event_study(
        events=events_monthly,
        prices=prices_monthly,
        factors=factors_ff3_monthly,
        config=config_monthly,
        group_col='group'
    )

    print("\n--- Monthly Test Statistics ---")
    ts_m = results_monthly['test_stats']
    mcols = ['group', 'N', 'mean_CAR', 'mean_BHAR', 'pct_positive',
             't_CS', 'p_CS', 't_BMP', 'p_BMP']
    mavail = [c for c in mcols if c in ts_m.columns]
    print(ts_m[mavail].to_string(index=False))
else:
    print("No monthly events could be aligned. Skipping monthly study.")
```

```
  Event Study: ff3 model
  Windows: estimation=36, gap=3, event=(-3,3)
  Min obs: 24

Step 1: Building trading calendar...
  122 potential event dates

Step 2: Aligning events to trading calendar...
  50 aligned events
```

```
Step 3: Extracting returns and merging factors...
  Extracted 1,036 obs for 33 firm-events

Step 4: Estimating risk model parameters...
  Estimated 18/33 firm-events (mean R^2 = 0.3218)

Step 5: Computing abnormal returns...
  18 firm-events | Mean CAR: -0.005257 | Mean BHAR: -0.014576 | % positive: 55.6%

Step 6: Computing test statistics...
  Done.

   Results Summary
  group  N  mean_CAR  mean_BHAR  pct_positive      t_CS     p_CS     t_BMP    p_BMP      t_KI
    All 18 -0.005257  -0.014576      0.555556 -0.081058 0.936342 -0.249547 0.805928 -
0.320212 0.752709
Group_A  7 -0.141756  -0.135084      0.428571 -1.102110 0.312648 -1.357452 0.223472 -
1.462577 0.193905
Group_B 11  0.081606   0.062111      0.636364  1.390317 0.194599  1.177372 0.266309  1.342593

--- Monthly Test Statistics ---
  group  N  mean_CAR  mean_BHAR  pct_positive      t_CS     p_CS     t_BMP    p_BMP
    All 18 -0.005257  -0.014576      0.555556 -0.081058 0.936342 -0.249547 0.805928
Group_A  7 -0.141756  -0.135084      0.428571 -1.102110 0.312648 -1.357452 0.223472
Group_B 11  0.081606   0.062111      0.636364  1.390317 0.194599  1.177372 0.266309
```

```python
if len(events_monthly) > 0 and 'daily_stats' in results_monthly:
    fig3 = plot_event_study(
        results_monthly['daily_stats'],
        title="Event Study: FF3 Model - Vietnamese Market (Monthly)"
    )
    plt.show()
```

Figure 14.5: Monthly cumulative abnormal returns around event dates. Wider windows capture slower information incorporation typical of emerging markets.

### 14.6.9 Daily Event Study: Fama-French 5-Factor Model

```
config_ff5 = EventStudyConfig(
    estimation_window=150,
    event_window_start=-10,
    event_window_end=10,
    gap=15,
    min_estimation_obs=120,
    risk_model=RiskModel.FF5
)

results_ff5 = run_event_study(
    events=events_demo,
    prices=prices_daily,
    factors=factors_ff5_daily,
    config=config_ff5,
```

```
    group_col='group'
)
```

```
   Event Study: ff5 model
  Windows: estimation=150, gap=15, event=(-10,10)
  Min obs: 120

Step 1: Building trading calendar...
  3308 potential event dates

Step 2: Aligning events to trading calendar...
  50 aligned events

Step 3: Extracting returns and merging factors...
  Extracted 5,421 obs for 30 firm-events

Step 4: Estimating risk model parameters...
  Estimated 26/30 firm-events (mean R^2 = 0.2675)

Step 5: Computing abnormal returns...
  26 firm-events | Mean CAR: 0.025684 | Mean BHAR: 0.028399 | % positive: 57.7%

Step 6: Computing test statistics...
  Done.

   Results Summary
  group  N  mean_CAR  mean_BHAR  pct_positive     t_CS     p_CS    t_BMP    p_BMP     t_KP
    All 26  0.025684   0.028399      0.576923 0.968332 0.342154 0.986788 0.333201 0.962252 0
Group_A 13  0.015352   0.013489      0.461538 0.393655 0.700741 0.786232 0.446980 0.776663 0
Group_B 13  0.036016   0.043309      0.692308 0.965100 0.353542 0.585915 0.568789 0.578785 0
```

### 14.6.10 Comparing FF3 vs FF5 Estimation Quality

### 14.6.11 Event-Level Detail

### 14.6.12 Daily Abnormal Return Dynamics

Table 14.10: Comparison of estimation quality between FF3 and FF5 models

```
params_ff3 = results_ff3['params']
params_ff5 = results_ff5['params']

print("Model Estimation Diagnostics")
print("=" * 60)
print(f"\n{'Metric':<30} {'FF3':>12} {'FF5':>12}")
print("-" * 54)
print(f"{'Firm-events estimated':<30} {len(params_ff3):>12} {len(params_ff5):>12}")
print(f"{'Mean R^2':<30} {params_ff3['r_squared'].mean():>12.4f} {params_ff5['r_squared'].mea
print(f"{'Median R^2':<30} {params_ff3['r_squared'].median():>12.4f} {params_ff5['r_squared']
print(f"{'Mean  ( )':<30} {params_ff3['sigma'].mean():>12.6f} {params_ff5['sigma'].mean():>12
print(f"{'Mean | |':<30} {params_ff3['alpha'].abs().mean():>12.6f} {params_ff5['alpha'].abs()
print(f"{'Mean  (MKT)':<30} {params_ff3['beta_mkt_excess'].mean():>12.4f} {params_ff5['beta_m
if 'beta_smb' in params_ff3.columns:
    print(f"{'Mean  (SMB)':<30} {params_ff3['beta_smb'].mean():>12.4f} {params_ff5['beta_smb'
if 'beta_hml' in params_ff3.columns:
    print(f"{'Mean  (HML)':<30} {params_ff3['beta_hml'].mean():>12.4f} {params_ff5['beta_hml'
if 'beta_rmw' in params_ff5.columns:
    print(f"{'Mean  (RMW)':<30} {'-':>12} {params_ff5['beta_rmw'].mean():>12.4f}")
if 'beta_cma' in params_ff5.columns:
    print(f"{'Mean  (CMA)':<30} {'-':>12} {params_ff5['beta_cma'].mean():>12.4f}")
```

```
Model Estimation Diagnostics
============================================================

Metric                                 FF3          FF5
-------------------------------------------------------
Firm-events estimated                   26           26
Mean R^2                            0.2245       0.2675
Median R^2                          0.1943       0.2692
Mean  ( )                         0.021753     0.021351
Mean | |                          0.001022     0.001130
Mean  (MKT)                         0.8867       0.9721
Mean  (SMB)                        -0.0434       0.0265
Mean  (HML)                         0.2489       0.1493
Mean  (RMW)                              -      -0.0934
Mean  (CMA)                              -       0.1070
```

Table 14.11: Event-level detail: CARs and BHARs for each firm-event (FF3 model)

```
detail = results_ff3['event_ar'].copy()
detail_cols = ['symbol', 'evtdate', 'CAR', 'BHAR', 'SCAR', 'sigma',
               'nobs', 'alpha', 'beta_mkt_excess']
detail_avail = [c for c in detail_cols if c in detail.columns]
detail_show = detail[detail_avail].copy()
detail_show['CAR'] = detail_show['CAR'].map(lambda x: f'{x:.4%}')
detail_show['BHAR'] = detail_show['BHAR'].map(lambda x: f'{x:.4%}')
detail_show['SCAR'] = detail_show['SCAR'].map(lambda x: f'{x:.3f}')

print("Event-Level Results (first 20 firm-events)")
print("=" * 100)
print(detail_show.head(20).to_string(index=False))
```

```
Event-Level Results (first 20 firm-events)
==========================================================================================
symbol    evtdate       CAR      BHAR   SCAR    sigma  nobs     alpha  beta_mkt_excess
   BVH 2016-10-20 -12.6456% -11.6522% -1.603 0.017219   150  0.001193         1.449182
   DHG 2016-01-25  16.1151%  17.1149%  2.273 0.015472   150 -
0.000224         0.754245
   DNH 2019-10-14   1.7233%   1.8068%  0.104 0.036035   150 -
0.000781         1.861996
   DPM 2020-07-30  -5.9576%  -6.2025% -0.545 0.023873   150  0.001279         0.313932
   FOX 2021-01-13   2.7478%   2.9194%  0.329 0.018243   150 -
0.000696         0.148058
   GAS 2020-02-11   0.5371%   0.7801%  0.100 0.011694   150  0.000501         1.851155
   GEX 2020-08-20  11.9985%  13.5843%  1.197 0.021883   150  0.000944         1.583418
   IDC 2018-10-01   1.3889%   0.5651%  0.094 0.032120   150 -
0.000674         0.809305
   MML 2021-10-29 -12.0139% -12.3759% -1.141 0.022985   150  0.002976         0.260588
   MSN 2015-10-23  -5.7205%  -5.5645% -0.718 0.017375   150  0.001177         0.453951
   PGV 2019-06-26  -7.5892%  -9.1366% -0.359 0.046165   150  0.000502         0.151377
   PLX 2020-01-07   2.5330%   2.7847%  0.423 0.013067   150 -
0.000176         0.917578
   PLX 2020-06-01  -6.1517%  -6.0085% -0.739 0.018168   150 -
0.000465         1.815178
   POW 2020-12-23   7.7751%   9.1225%  1.130 0.015014   150 -
0.000575         0.774423
   PVD 2020-05-25   4.8827%   5.6674%  0.510 0.020875   150 -
0.001522         1.316102
   PVS 2017-08-07   2.1360%   2.1918%  0.297 0.015715   150 -
0.000341         1.136273
   QNS 2018-01-18   4.3001%   3.4011%  0.530 0.017703   150 -
0.003716         0.128328
   SAB 2017-08-24   6.0347%   6.7732%  0.748 0.017614   150  0.003284         2.123751
   SNZ 2019-03-12  34.5230%  33.1105%  2.145 0.035121   150  0.000235        -
1.015554
   VCI 2020-01-20  -3.3191%  -2.9072% -0.458 0.015797   150  0.000335        -
0.086268
```

Table 14.12: Daily dynamics of mean abnormal returns and test statistics within the event window

```
ds = results_ff3['daily_stats'].copy()
ds_cols = ['evttime', 'N', 'mean_AR', 'mean_CAR', 'mean_BHAR', 't_AR_CS', 't_AR_BMP']
ds_avail = [c for c in ds_cols if c in ds.columns]
ds_show = ds[ds_avail].copy()

for c in ['mean_AR', 'mean_CAR', 'mean_BHAR']:
    if c in ds_show.columns:
        ds_show[c] = ds_show[c].map(lambda x: f'{x:.4%}')
for c in ['t_AR_CS', 't_AR_BMP']:
    if c in ds_show.columns:
        ds_show[c] = ds_show[c].map(lambda x: f'{x:.3f}' if pd.notna(x) else '')

print("Daily Event-Window Dynamics (FF3 Model)")
print("=" * 80)
print(ds_show.to_string(index=False))
```

```
Daily Event-Window Dynamics (FF3 Model)
================================================================================
 evttime  N  mean_AR mean_CAR mean_BHAR
     -10 23  0.3571%  0.3571%   0.3729%
      -9 23  0.0337%  0.3907%   0.4209%
      -8 23 -0.1581%  0.2326%   0.2766%
      -7 23  0.3691%  0.6018%   0.6581%
      -6 23  1.3416%  1.9433%   2.0278%
      -5 23  0.1509%  2.0942%   2.2313%
      -4 23  0.5512%  2.6454%   2.9187%
      -3 23 -0.4641%  2.1814%   2.4297%
      -2 23  0.2412%  2.4225%   2.8028%
      -1 23  0.5660%  2.9885%   3.3240%
       0 23 -0.0281%  2.9604%   3.4926%
       1 23 -0.2564%  2.7040%   3.1039%
       2 23 -0.4421%  2.2619%   2.5764%
       3 23  0.6337%  2.8956%   3.4659%
       4 23 -0.8432%  2.0524%   2.4738%
       5 23 -0.4174%  1.6350%   2.1339%
       6 23 -0.2272%  1.4078%   1.8085%
       7 23 -0.2085%  1.1993%   1.6819%
       8 23  0.0893%  1.2886%   1.8609%
       9 23  0.3307%  1.6193%   2.1658%
      10 23  0.5320%  2.1513%   2.4885%
```

### 14.6.13 Summary of Key Findings

```
print("=" * 70)
print("EVENT STUDY RESULTS SUMMARY")
print("=" * 70)

ff3_all = results_ff3['test_stats'][results_ff3['test_stats']['group'] == 'All'].iloc[0]

print(f"\nSample: {int(ff3_all['N'])} firm-event observations")
print(f"Frequency: Daily")
print(f"Primary Model: Fama-French 3-Factor")
print(f"Estimation Window: {config_ff3.estimation_window} trading days")
print(f"Event Window: ({config_ff3.event_window_start}, {config_ff3.event_window_end})")
print(f"Gap: {config_ff3.gap} trading days")
print(f"\n--- Abnormal Return Measures ---")
print(f"Mean CAR({config_ff3.event_window_start},{config_ff3.event_window_end}): "
      f"{ff3_all['mean_CAR']:.4%}")
print(f"Median CAR: {ff3_all['median_CAR']:.4%}")
print(f"Mean BHAR: {ff3_all['mean_BHAR']:.4%}")
print(f"Fraction positive CARs: {ff3_all['pct_positive']:.1%}")
print(f"\n--- Statistical Significance ---")
print(f"Cross-Sectional t: {ff3_all['t_CS']:.3f} (p = {ff3_all['p_CS']:.4f})")
print(f"Patell Z: {ff3_all['Z_Patell']:.3f} (p = {ff3_all['p_Patell']:.4f})")
print(f"BMP t: {ff3_all['t_BMP']:.3f} (p = {ff3_all['p_BMP']:.4f})")
print(f"Kolari-Pynnönen t: {ff3_all['t_KP']:.3f} (p = {ff3_all['p_KP']:.4f})")
print(f"Generalized Sign Z: {ff3_all['Z_GSign']:.3f} (p = {ff3_all['p_GSign']:.4f})")

sig_005 = sum(1 for k in ['p_CS','p_Patell','p_BMP','p_KP','p_GSign','p_SkAdj','p_Wilcoxon']
              if k in ff3_all and pd.notna(ff3_all[k]) and ff3_all[k] < 0.05)
total_tests = sum(1 for k in ['p_CS','p_Patell','p_BMP','p_KP','p_GSign','p_SkAdj','p_Wilcox
                  if k in ff3_all and pd.notna(ff3_all[k]))
print(f"\n{sig_005}/{total_tests} tests significant at 5% level")

# Robustness note
print(f"\nRobustness: Results checked across {len(models_daily)} risk models "
      f"and {len(windows)} event windows")
```

```
======================================================================
EVENT STUDY RESULTS SUMMARY
======================================================================
```

```
Sample: 26 firm-event observations
Frequency: Daily
Primary Model: Fama-French 3-Factor
Estimation Window: 150 trading days
Event Window: (-10, 10)
Gap: 15 trading days

--- Abnormal Return Measures ---
Mean CAR(-10,10): 2.1513%
Median CAR: 1.5701%
Mean BHAR: 2.4885%
Fraction positive CARs: 61.5%

--- Statistical Significance ---
Cross-Sectional t: 0.775 (p = 0.4456)
Patell Z: 0.961 (p = 0.3364)
BMP t: 0.727 (p = 0.4741)
Kolari-Pynnönen t: 0.700 (p = 0.4904)
Generalized Sign Z: 1.177 (p = 0.2393)

0/7 tests significant at 5% level

Robustness: Results checked across 4 risk models and 7 event windows
```

## 14.7 Practical Recommendations

Based on the literature and our implementation experience:

1. **Estimation window**: Use 150 trading days (~7 months) for daily studies. This balances parameter precision against structural breaks. For monthly studies, 60 months is standard (Kothari and Warner 2007).
2. **Gap**: 15 trading days is standard. Increase to 30 if information leakage is a concern.
3. **Event window**: Start with (-1, +1) for short-window tests, then expand to (-5, +5) and (-10, +10) for robustness. Report all windows.
4. **Model choice**: Always report market model as the baseline. Add FF3 or FF5 for robustness. For Vietnam, local factors are preferable to global factors.
5. **Test statistics**: Report at minimum: cross-sectional t (for ease of interpretation), BMP (robust to event-induced variance), and one non-parametric test (sign or Wilcoxon). Report Kolari-Pynnönen if events cluster in calendar time.

6. **Thin trading**: For Vietnamese small-caps, consider Dimson (1979) with 1 lead/lag or increase `min_estimation_obs` to filter out illiquid stocks.
7. **Multiple testing**: If testing multiple event windows or subgroups, apply Bonferroni or Holm corrections to control family-wise error rate.

# 15 Standardized Earnings Surprises (SUE)

In the context of the Ho Chi Minh Stock Exchange (HOSE) and the Hanoi Stock Exchange (HNX), earnings announcements represent critical information events. Investors and quantitative analysts continuously monitor the deviation between reported earnings and market expectations. This deviation is quantified as the Standardized Earnings Surprise (SUE).

This chapter details the methodology for calculating SUE using three distinct approaches frequently utilized in academic literature and institutional research. We apply these methods to a dataset of Vietnamese large-cap equities to illustrate the mechanics of the calculation. The goal is to isolate the "surprise" component of earnings, which is a known predictor of post-earnings announcement drift (PEAD) (Bernard and Thomas 1989; Livnat and Mendenhall 2006).

## 15.1 Methodology

We define three primary methods for calculating SUE. Each method differs in how it establishes the "expected" earnings value.

### 15.1.1 Method 1: Seasonal Random Walk

This method assumes that earnings follow a seasonal pattern. The best predictor for the current quarter's earnings per share (EPS) is the EPS from the same quarter in the previous year. This controls for the seasonality often seen in Vietnamese sectors like retail and agriculture.

$$SUE_1 = \frac{EPS_t - EPS_{t-4}}{P_t}$$

Where:

- $EPS_t$ is the current quarterly Earnings Per Share.
- $EPS_{t-4}$ is the Earnings Per Share from the same quarter of the prior fiscal year.
- $P_t$ is the stock price at the end of the quarter (used as a deflator).

### 15.1.2 Method 2: Exclusion of Special Items

Reported earnings often contain non-recurring items (e.g., asset sales, one-time write-offs) that distort the true operating performance. This method adjusts the reported EPS by removing the after-tax impact of special items.

In Vietnam, the standard Corporate Income Tax (CIT) rate is generally 20%. We adjust special items to reflect their impact on net income.

$$Adjusted\ EPS = Reported\ EPS - \frac{Special\ Items \times (1 - CIT)}{Shares\ Outstanding}$$

The SUE calculation then follows the seasonal logic but uses the adjusted EPS figures:

$$SUE_2 = \frac{Adj\ EPS_t - Adj\ EPS_{t-4}}{P_t}$$

### 15.1.3 Method 3: Analyst Consensus

This method relies on market consensus rather than historical time series. It compares the actual reported earnings against the median analyst forecast provided prior to the announcement.

$$SUE_3 = \frac{Actual\ EPS - Median\ Estimate}{P_t}$$

## 15.2 Data Description

For this analysis, we utilize a dataset covering the fiscal years 2023 through 2025. The data includes quarterly financial statements and analyst consensus estimates for a selection of VN30 index constituents.

The dataset, `vietnam_fin_data.csv`, contains the following columns:

- **ticker**: Stock symbol (e.g., VNM, VCB, HPG).
- **fiscal_year**: The financial year.
- **fiscal_qtr**: The financial quarter (1-4).
- **eps_basic**: Basic Earnings Per Share (VND).
- **price_close**: Closing price at quarter end (VND).

- **special_items**: Pre-tax special items value (VND millions). (i.e., `is_other_profit` in DataCore).
- **shares_out**: Shares outstanding (millions).
- **analyst_med**: Median analyst EPS estimate (VND).

### 15.2.1 Visualizing the Core Data

Below is a tabular representation of the raw data we have ingested for the analysis.

| ticker | fiscal_year | fiscal_qtr | eps_basic | price_close | special_items | shares_out | analyst_med |
|---|---|---|---|---|---|---|---|
| VNM | 2023 | 1 | 1200 | 68000 | 0 | 2090 | 1150 |
| VNM | 2023 | 2 | 1350 | 71000 | 50000 | 2090 | 1300 |
| VNM | 2023 | 3 | 1400 | 74000 | 0 | 2090 | 1450 |
| VNM | 2023 | 4 | 1100 | 69000 | -20000 | 2090 | 1150 |
| VNM | 2024 | 1 | 1300 | 72000 | 0 | 2090 | 1250 |
| VNM | 2024 | 2 | 1500 | 75000 | 0 | 2090 | 1400 |
| VCB | 2023 | 1 | 1800 | 85000 | 10000 | 5500 | 1700 |
| VCB | 2024 | 1 | 2100 | 92000 | 0 | 5500 | 2000 |

## 15.3 Implementation

### 15.3.1 Python Setup and Data Loading

First, we establish our environment and load the dataset. We ensure the data is sorted by ticker and time to allow for accurate lag calculations.

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

# Creating the dataset directly for this chapter's demonstration
data = {
    'ticker': ['VNM']*8 + ['VCB']*8 + ['HPG']*8,
    'fiscal_year': [2023, 2023, 2023, 2023, 2024, 2024, 2024, 2024] * 3,
    'fiscal_qtr': [1, 2, 3, 4, 1, 2, 3, 4] * 3,
    'eps_basic': [
        1200, 1350, 1400, 1100, 1300, 1500, 1450, 1250, # VNM
        1800, 1900, 2000, 2200, 2100, 2300, 2400, 2600, # VCB
```

```
        500, 600, 550, 400, 700, 800, 750, 600          # HPG
    ],
    'price_close': [
        68000, 71000, 74000, 69000, 72000, 75000, 73000, 70000, # VNM
        85000, 88000, 90000, 95000, 92000, 96000, 98000, 102000, # VCB
        20000, 22000, 21000, 19000, 25000, 28000, 27000, 24000 # HPG
    ],
    'special_items': [
        0, 50000, 0, -20000, 0, 0, 10000, 0, # VNM (VND Millions)
        10000, 0, 0, 50000, 0, 20000, 0, 0, # VCB
        0, 0, -50000, 0, 100000, 0, 0, 0 # HPG
    ],
    'shares_out': [2090]*8 + [5580]*8 + [5810]*8, # In Millions
    'analyst_med': [
        1150, 1300, 1450, 1150, 1250, 1400, 1480, 1200, # VNM
        1700, 1850, 1950, 2150, 2000, 2250, 2450, 2550, # VCB
        450, 550, 600, 450, 650, 750, 800, 650 # HPG
    ]
}

df = pd.DataFrame(data)

# Sort strictly to ensure shift operations work on correct temporal sequence
df = df.sort_values(by=['ticker', 'fiscal_year', 'fiscal_qtr'])
print(df.head())
```

```
   ticker  fiscal_year  fiscal_qtr  eps_basic  price_close  special_items  \
16    HPG         2023           1        500        20000              0
17    HPG         2023           2        600        22000              0
18    HPG         2023           3        550        21000         -50000
19    HPG         2023           4        400        19000              0
20    HPG         2024           1        700        25000         100000

    shares_out  analyst_med
16        5810          450
17        5810          550
18        5810          600
19        5810          450
20        5810          650
```

### 15.3.2 Calculation Logic

We now apply the functions to calculate the three variations of SUE.

**Step 1: Handling Seasonality (Lags)**

For Methods 1 and 2, we require the data from the same quarter of the previous year (lag 4).

```
# Group by ticker to ensure we don't shift data between companies
df['eps_lag4'] = df.groupby('ticker')['eps_basic'].shift(4)
```

**Step 2: Adjusting for Special Items**

For Method 2, we must strip out non-recurring items. We apply the Vietnamese Corporate Income Tax (CIT) rate of 20%.

The formula for the adjustment per share is:

$$\text{Adjustment} = \frac{\text{Special Items} \times (1 - 0.20)}{\text{Shares Outstanding}}$$

```
# Constants
CIT_RATE_VN = 0.20

# Calculate impact per share
# Note: special_items are in millions, shares_out are in millions
# The units cancel out, leaving the result in VND per share.
df['spi_impact_per_share'] = (df['special_items'] * (1 - CIT_RATE_VN)) / df['shares_out']

# Calculate Adjusted EPS
df['eps_adjusted'] = df['eps_basic'] - df['spi_impact_per_share']

# Create lag for Adjusted EPS
df['eps_adj_lag4'] = df.groupby('ticker')['eps_adjusted'].shift(4)
```

**Step 3: Computing SUE Variants**

We finalize the calculation by computing the difference between actual (or adjusted) and expected values, deflated by the stock price.

```
# Method 1: Seasonal Random Walk (Standard EPS)
df['sue_1'] = (df['eps_basic'] - df['eps_lag4']) / df['price_close']

# Method 2: Seasonal Random Walk (Excluding Special Items)
```

```
df['sue_2'] = (df['eps_adjusted'] - df['eps_adj_lag4']) / df['price_close']

# Method 3: Analyst Forecasts (IBES Equivalent)
df['sue_3'] = (df['eps_basic'] - df['analyst_med']) / df['price_close']

# Scaling for readability (converting to percentage)
df['sue_1_pct'] = df['sue_1'] * 100
df['sue_2_pct'] = df['sue_2'] * 100
df['sue_3_pct'] = df['sue_3'] * 100
```

## 15.4 Results and Analysis

We present the calculated standardized earnings surprises for the fiscal year 2024. Positive values indicate a positive surprise (beating expectations), while negative values indicate a miss.

### 15.4.1 Tabular Results (FY 2024)

```
# Filter for 2024 results where lag data exists
results_2024 = df[df['fiscal_year'] == 2024][['ticker', 'fiscal_qtr', 'sue_1_pct', 'sue_2_pct

# Display formatted table
from IPython.display import display, Markdown
markdown_table = results_2024.to_markdown(index=False, floatfmt=".4f")
display(Markdown(markdown_table))
```

| ticker | fiscal_qtr | sue_1_pct | sue_2_pct | sue_3_pct |
|---|---|---|---|---|
| HPG | 1 | 0.8000 | 0.7449 | 0.2000 |
| HPG | 2 | 0.7143 | 0.7143 | 0.1786 |
| HPG | 3 | 0.7407 | 0.7152 | -0.1852 |
| HPG | 4 | 0.8333 | 0.8333 | -0.2083 |
| VCB | 1 | 0.3261 | 0.3276 | 0.1087 |
| VCB | 2 | 0.4167 | 0.4137 | 0.0521 |
| VCB | 3 | 0.4082 | 0.4082 | -0.0510 |
| VCB | 4 | 0.3922 | 0.3992 | 0.0490 |
| VNM | 1 | 0.1389 | 0.1389 | 0.0694 |
| VNM | 2 | 0.2000 | 0.2255 | 0.1333 |

| ticker | fiscal_qtr | sue_1_pct | sue_2_pct | sue_3_pct |
|---|---|---|---|---|
| VNM | 3 | 0.0685 | 0.0632 | -0.0411 |
| VNM | 4 | 0.2143 | 0.2033 | 0.0714 |

### 15.4.2 Visualization

The following figure plots the Analyst-based SUE (Method 3) for the selected tickers over the 2024 fiscal year.

```
pivot_sue = results_2024.pivot(index='fiscal_qtr', columns='ticker', values='sue_3_pct')

plt.figure(figsize=(10, 6))
for column in pivot_sue.columns:
    plt.plot(pivot_sue.index, pivot_sue[column], marker='o', label=column)

plt.title('Method 3: Analyst Based SUE (FY 2024)')
plt.xlabel('Fiscal Quarter')
plt.ylabel('SUE (%)')
plt.axhline(0, color='black', linestyle='--', linewidth=0.8)
plt.legend(title='Ticker')
plt.grid(True, linestyle=':', alpha=0.6)
plt.xticks([1, 2, 3, 4])
plt.show()
```

Figure 15.1

## 15.5 Conclusion

In this chapter, we have formalized the calculation of Standardized Earnings Surprises for the Vietnamese market. By implementing three distinct methods using Python, we demonstrated that relying solely on raw EPS growth (Method 1) can be misleading in the presence of non-recurring items. Furthermore, analyst-based surprises (Method 3) often provide a cleaner signal of new information reaching the market.

For robust quantitative modeling in Vietnam, we recommend using Method 2 when analyst data is sparse (common in small-cap stocks) and Method 3 for VN30 constituents where analyst coverage is deep and liquid.

# 16 Measuring Divergence of Investor Opinion

A foundational question in financial economics concerns how differences in investor beliefs affect asset prices and trading activity. In markets where investors hold heterogeneous expectations about a firm's future cash flows, the aggregation of these divergent views into a single market price becomes a non-trivial exercise with profound implications for asset valuation, return predictability, and market efficiency. The concept of **divergence of investor opinion** (hereafter DIVOP) has emerged as a central construct in both the accounting and finance literatures, serving as a lens through which researchers examine the information environment of firms, the dynamics of uncertainty resolution, and the nature of market reactions to news.

The theoretical foundations of the DIVOP literature trace back to Miller (1977), who proposed that when investors disagree about the value of a security and short-sale constraints prevent pessimistic investors from fully expressing their views, the market price will reflect the valuation of the most optimistic investors. This leads to systematic overpricing that is increasing in the degree of opinion divergence. The overpricing persists until information events, such as earnings announcements, reduce disagreement and prices converge toward fundamental values (Berkman et al. 2009). Varian (1985) offers an alternative perspective in which divergence of opinion represents an additional risk factor, leading to *higher* rather than lower expected returns, creating a theoretical tension that has motivated extensive empirical investigation.

The empirical literature on DIVOP has expanded considerably since these seminal contributions. Researchers have documented that divergence of opinion helps explain a range of asset pricing anomalies, including post-earnings announcement drift (Garfinkel and Sokobin 2006; K. L. Anderson, Harris, and So 2007), the cross-sectional return difference between value and growth stocks (Doukas, Kim, and Pantzalis 2004), short- and long-run post-IPO returns (Houge et al. 2001), pre- and post-acquisition stock returns (Alexandridis, Antoniou, and Petmezas 2007), takeover premia (Chatterjee, John, and Yan 2012), and the broad cross-section of stock returns (Diether, Malloy, and Scherbina 2002; Doukas, Kim, and Pantzalis 2006). The explanatory power of DIVOP has been demonstrated using a rich set of empirical proxies, ranging from analyst forecast dispersion and abnormal trading volume to bid-ask spreads and idiosyncratic volatility.

Despite the maturity of the DIVOP literature in developed markets, particularly the United States, its application to emerging markets remains remarkably thin. This gap is especially notable given that the theoretical conditions under which divergence of opinion matters most (namely, binding short-sale constraints, information asymmetry, and heterogeneous investor sophistication) are arguably *more* prevalent in emerging markets than in their developed

counterparts. The Vietnamese equity market presents a compelling laboratory for studying investor disagreement. The market is characterized by several features that amplify the relevance of the DIVOP framework:

1. **Binding short-sale constraints.** Short selling was not permitted in Vietnam until January 2025, and even after its introduction, the mechanism remains restricted to a limited set of securities with significant regulatory constraints on execution. This closely mirrors the theoretical setting of Miller (1977) where pessimistic investors are unable to fully express their views through short positions.

2. **Dominance of retail investors.** Individual investors account for approximately 80-85% of daily trading volume on HOSE and HNX, compared to roughly 25% in the United States. Retail investors are more susceptible to behavioral biases, sentiment-driven trading, and information processing limitations that naturally give rise to heterogeneous beliefs (Phan et al. 2023).

3. **Information asymmetry and transparency challenges.** Despite improvements in disclosure standards, Vietnam's regulatory framework for corporate reporting remains less stringent than those in developed markets. Selective disclosure, delayed filing of financial statements, and limited enforcement of insider trading regulations create an environment in which investors operate with substantially different information sets (Vo and Phan 2017).

4. **Foreign ownership limits.** Caps on foreign ownership (currently 49% for most sectors, with exceptions) create a segmented market where domestic and foreign investors may hold systematically different views about firm value, amplifying the divergence of opinion.

5. **Thin analyst coverage.** Whereas a typical S&P 500 firm is followed by 15-25 sell-side analysts, coverage of Vietnamese equities is concentrated among a relatively small number of domestic brokerages and a handful of international research houses. This limits the informativeness of traditional analyst-based DIVOP measures and necessitates greater reliance on market-based proxies.

This chapter provides a methodology for constructing multiple proxies for divergence of investor opinion adapted to the institutional characteristics of the Vietnamese market. We draw on the methodological frameworks established by Garfinkel (2009) and Diether, Malloy, and Scherbina (2002), while introducing modifications that account for the microstructure of Vietnamese exchanges, the $T + 2$ settlement cycle, the absence (until recently) of short selling, and the availability of data through domestic financial platforms. Specifically, we construct and analyze the following DIVOP proxies:

- **Unexplained Volume (DTO):** Market-adjusted turnover detrended by its rolling median, capturing abnormal trading activity attributable to disagreement after controlling for liquidity and market-wide effects.

- **Standardized Unexplained Volume (SUV):** A regression-based measure that explicitly controls for the informedness and liquidity components of volume by modeling turnover as a function of signed returns.
- **Stock Return Volatility (VOLATILITY):** The standard deviation of daily returns over a rolling estimation window, serving as a proxy for the dispersion of investor valuations.
- **Bid-Ask Spread (BASPREAD):** The proportional quoted spread, reflecting the adverse selection component associated with heterogeneous information among market participants.
- **Analyst Forecast Dispersion (DISP):** The cross-sectional standard deviation of individual analyst earnings forecasts, directly measuring disagreement among informed market participants.
- **Idiosyncratic Volatility (IVOL):** The residual volatility from a factor model regression, isolating the firm-specific component of return variation that reflects divergent investor interpretations of firm-level information.
- **Amihud Illiquidity (ILLIQ):** The price impact ratio proposed by Amihud (2002), which captures the information asymmetry dimension of disagreement through the price response to order flow.

For each proxy, we describe the theoretical motivation, the data requirements, the construction methodology adapted for Vietnamese data, the empirical properties observed in the Vietnamese cross-section, and the practical considerations that researchers should bear in mind when employing these measures. We pay particular attention to issues that are specific to emerging markets, including thin trading, corporate action adjustments, exchange-specific microstructure effects, and the interplay between foreign ownership constraints and measures of investor disagreement.

---

# 17 Theoretical Framework

## 17.1 The Miller (1977) Overpricing Hypothesis

The canonical model of divergence of opinion and asset pricing begins with Miller (1977). Miller's central insight is simple: in a market where investors hold heterogeneous beliefs about the future payoffs of a risky asset and short-sale constraints prevent some investors from acting on their pessimistic views, the equilibrium price will be set by the subset of investors who are most optimistic about the asset's value. The severity of overpricing is increasing in both the degree of opinion divergence and the stringency of short-sale constraints. Formally, if investor $i$ assigns a valuation $V_i$ to a security, the market price $P$ satisfies:

$$P = E[V_i \mid V_i \geq V^*]$$

where $V^*$ is the marginal investor's valuation, which exceeds the unconditional mean valuation $E[V_i]$ whenever short-sale constraints bind for some investors. The degree of overpricing is:

$$\text{Overpricing} = P - E[V_i] = E[V_i \mid V_i \geq V^*] - E[V_i]$$

which is positive and increasing in the dispersion of the distribution of $V_i$ (i.e., divergence of opinion) and in $V^*$ (i.e., the severity of short-sale constraints).

Miller's model generates several testable predictions:

- **Cross-sectional prediction:** Stocks with **higher divergence of opinion should have *lower* subsequent returns** as prices gradually correct toward fundamental values.
- **Time-series prediction:** Information events that reduce disagreement (e.g., earnings announcements) should be associated with negative abnormal returns for high-DIVOP stocks, as the "optimism premium" dissipates.
- **Interaction prediction:** The overpricing effect should be strongest among stocks that simultaneously exhibit high divergence of opinion *and* binding short-sale constraints.

## 17.2 Alternative Theoretical Perspectives

Varian (1985) proposes an alternative framework in which divergence of opinion acts as a risk factor. If investors are risk-averse and disagreement represents genuine uncertainty about future payoffs, then **higher dispersion of beliefs should be associated with *higher* expected returns** as compensation for bearing the additional risk. This creates a sharp empirical dichotomy: the Miller hypothesis predicts a negative DIVOP-return relation, whereas the Varian model predicts a positive relation.

The distinction between these theories hinges critically on the market microstructure and institutional setting (@tbl-divop-theories).

Table 17.1: Summary of theoretical predictions for the DIVOP-return relation under different assumptions

| Theoretical Framework | Short-Sale Constraints | DIVOP-Return Relation | Key Mechanism |
|---|---|---|---|
| Miller (1977) | Binding | Negative | Optimistic bias in price |
| Varian (1985) | Non-binding | Positive | Risk premium for uncertainty |
| Hong and Stein (2003) | Binding, gradual info | Negative, time-varying | Slow diffusion of bearish views |
| Scheinkman and Xiong (2003) | Binding, overconfidence | Negative | Speculative bubble premium |

Hong and Stein (2003) extend Miller's framework by incorporating gradual information diffusion. In their model, bearish information is impounded into prices more slowly than bullish information because short-sale constraints raise the cost of acting on negative views. This generates momentum-like patterns in which high-DIVOP stocks exhibit positive short-run returns (as optimists push prices up) followed by negative long-run returns (as bearish information eventually reaches the market).

Scheinkman and Xiong (2003) introduce an additional dimension by noting that when investors are overconfident about their private signals *and* short-sale constraints bind, stock prices contain a "speculative bubble" component that reflects the option value of reselling the asset to a future investor who may be even more optimistic. This model predicts that both high trading volume and high price volatility should be associated with overpricing, providing a theoretical basis for using volume-based and volatility-based DIVOP proxies.

## 17.3 Relevance to the Vietnamese Market

The Vietnamese equity market provides an unusually clean setting for testing the Miller hypothesis. Vietnam's equity market operated without any short-selling mechanism from its inception in 2000 through January 2025, which was a full quarter-century in which the first necessary condition of Miller's model (binding short-sale constraints) was satisfied by regulation rather than by market frictions. Even after the introduction of covered short selling in 2025, the mechanism remains restricted to securities meeting specific liquidity and market capitalization thresholds, and the regulatory environment imposes borrowing requirements that significantly raise the cost of shorting relative to developed markets.

The dominance of retail investors amplifies the second necessary condition (i.e., heterogeneous beliefs). Research on the Vietnamese market has documented significant herding behavior (Vo and Phan 2017; Vo 2015), sentiment-driven trading (Phan et al. 2023; Nguyen and Pham 2018), and information asymmetry between domestic and foreign investors (Vo 2017). These behavioral characteristics naturally generate wider dispersion of investor valuations compared to markets dominated by institutional investors with access to similar analytical frameworks and information sources.

Table 17.2 compares key institutional features relevant to the DIVOP framework between Vietnam and the United States.

Table 17.2: Institutional comparison of Vietnam and the United States relevant to divergence of opinion

| Feature | Vietnam (HOSE/HNX) | United States (NYSE/NASDAQ) |
|---|---|---|
| Short selling | Introduced Jan 2025 (limited) | Permitted (Reg SHO since 2005) |
| Retail investor share of volume | ~80-85% | ~25% |
| Settlement cycle | T+2 (T+1 planned for 2026) | T+1 (since May 2024) |
| Daily price limits | ± 7% (HOSE), ± 10% (HNX) | None |
| Foreign ownership cap | 49% (most sectors) | None |
| Average analyst coverage (VN30) | 5-10 analysts | 15-25 analysts |
| Mandatory quarterly reporting | Yes (since 2012) | Yes |
| Options/derivatives market | VN30 Index Futures (since 2017) | Extensive options/futures |

The presence of daily price limits (± 7% on HOSE and ± 10% on HNX) creates an additional mechanism through which divergence of opinion can be amplified. When a stock hits its price

limit, investors who wish to trade in the direction of the limit are unable to do so, leading to accumulated unfilled orders and delayed price discovery. This institutional feature may create short-term spikes in measured DIVOP that reflect limit-induced friction rather than genuine disagreement. We address this issue in our empirical methodology by flagging limit-hit days and conducting robustness checks that exclude these observations.

# 18 Data Sources and Sample Construction

## 18.1 Data Sources

The construction of DIVOP proxies for the Vietnamese market requires daily stock-level trading data and, for the analyst dispersion measures, individual analyst forecast data. We source all data from DataCore.vn, which provides coverage of all securities listed on HOSE, HNX, and the UPCoM (Unlisted Public Company Market) exchange. Table 18.1 summarizes the datasets and key variables used in this study.

Table 18.1: Data sources and key variables for DIVOP proxy construction

| Dataset | Key Variables | Frequency |
|---|---|---|
| Daily Stock Trading | Close price, high, low, open, volume, shares outstanding, adjusted price, bid, ask | Daily |
| Corporate Actions | Dividends, stock splits, bonus issues, rights offerings | Event-based |
| Company Information | Exchange code, industry classification (ICB), listing date, delisting date | Static/Periodic |
| Analyst Forecasts | Individual analyst EPS forecasts, announcement dates, fiscal period end, analyst ID, broker name | Per estimate |
| Market Index | VN-Index daily returns, VN30 returns, HNX-Index returns | Daily |
| Foreign Ownership | Foreign buy/sell volume, foreign ownership percentage, remaining foreign room | Daily |

## 18.2 Sample Construction

We construct our sample using the following filters, applied sequentially:

```python
import pandas as pd
import numpy as np
from datetime import datetime, timedelta
from sklearn.linear_model import LinearRegression
from scipy import stats as scipy_stats
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')

# ==============================================================================
# Configuration Parameters
# ==============================================================================
# Users can modify these parameters to adjust the methodology
CONFIG = {
    # Sample period
    'beg_date': '2007-01-01',
    'end_date': '2024-12-31',

    # Estimation windows (in trading days)
    'est_window': 60,          # Rolling window for SUV and volatility
    'detrend_window': 180,     # Window for DTO detrending median
    'lag': 7,                  # Lag for DTO detrending
    'gap': 5,                  # Gap between estimation period and event date

    # Filters
    'min_price': 1000,         # Minimum price in VND
    'min_volume_days': 0.8,    # Min fraction of non-zero volume days in window
    'min_analysts': 3,         # Minimum number of analysts for DISP
    'max_spread_pct': 0.50,    # Maximum bid-ask spread as fraction of midpoint
    'forecast_carry_days': 105,# Days to carry forward stale analyst forecasts

    # Exchange identifiers
    'exchanges': ['HOSE', 'HNX'],

    # Price limit thresholds (for flagging)
    'price_limit_hose': 0.07,
    'price_limit_hnx': 0.10,
}

print("Configuration parameters loaded successfully.")
```

```python
print(f"Sample period: {CONFIG['beg_date']} to {CONFIG['end_date']}")
print(f"Estimation window: {CONFIG['est_window']} trading days")
print(f"Detrending window: {CONFIG['detrend_window']} trading days")
```

```
Configuration parameters loaded successfully.
Sample period: 2007-01-01 to 2024-12-31
Estimation window: 60 trading days
Detrending window: 180 trading days
```

The sample universe includes all common stocks (ordinary shares) listed on HOSE and HNX during the period January 2007 through December 2024. We begin in 2007 rather than at market inception (2000 for HOSE, 2005 for HNX) for two reasons. First, the early years of the Vietnamese market were characterized by an extremely small number of listed firms (fewer than 30 on HOSE through 2005), making cross-sectional analysis unreliable. Second, data quality and consistency improve substantially after the market expansion of 2006-2007, during which the number of listed firms on HOSE grew from approximately 40 to over 100.

We apply the following filters to construct the analysis sample:

1. **Security type filter.** We retain only common stocks (ordinary shares), excluding preferred shares, exchange-traded funds (ETFs), covered warrants, and certificates of deposit. This is analogous to the standard filter in the U.S. literature that restricts to CRSP share codes 10 and 11.

2. **Exchange filter.** We include stocks listed on HOSE and HNX but exclude UPCoM securities in our baseline analysis. UPCoM is a registration-based trading venue with less stringent listing requirements and substantially lower liquidity, which may introduce noise into volume-based and spread-based measures. We include UPCoM in robustness checks.

3. **Price filter.** We exclude stock-day observations with closing prices below 1,000 VND. This threshold serves the same purpose as the "penny stock" exclusion common in U.S. studies (typically $1 or $5 thresholds) and helps mitigate the influence of extreme percentage returns and spreads at very low price levels.

4. **Minimum trading activity.** For volume-based measures, we require that a stock has non-zero trading volume on at least 80% of trading days within each estimation window. This filter eliminates the most thinly traded securities for which turnover-based measures would be unreliable.

```python
def load_daily_data(config):
    """
    Load daily stock trading data from DataCore.vn.

    In practice, this function connects to the DataCore API or reads
    from a local database/CSV. Here we document the expected schema.

    Expected columns:
    - ticker: str, stock ticker symbol (e.g., 'VCB', 'HPG', 'VNM')
    - date: datetime, trading date
    - open, high, low, close: float, daily OHLC prices (VND)
    - volume: int, trading volume (shares)
    - shares_outstanding: int, total shares outstanding
    - adjusted_close: float, price adjusted for corporate actions
    - adj_factor: float, cumulative adjustment factor
    - bid, ask: float, best bid/ask at close
    - exchange: str, exchange code ('HOSE', 'HNX', 'UPCOM')
    - industry_icb: str, ICB industry classification code
    - foreign_buy_vol, foreign_sell_vol: int, foreign investor volumes
    - foreign_ownership_pct: float, foreign ownership percentage
    """
    # ==========================================================================
    # Replace with actual DataCore API call:
    # from datacore import Client
    # client = Client(api_key='YOUR_KEY')
    # df = client.daily_stock(
    #     start=config['beg_date'], end=config['end_date'],
    #     exchanges=config['exchanges']
    # )
    # ==========================================================================
    print("Connect to DataCore.vn and load daily stock data.")
    print("Expected schema: ticker, date, open, high, low, close, volume,")
    print("  shares_outstanding, adjusted_close, adj_factor, bid, ask,")
    print("  exchange, industry_icb, foreign_buy_vol, foreign_sell_vol,")
    print("  foreign_ownership_pct")
    return None  # Replace with actual data


def apply_sample_filters(df, config):
    """Apply sequential sample construction filters."""
    print("\n=== Sample Construction ===")
    n0 = len(df)
```

```
    # Date filter
    df = df[(df['date'] >= config['beg_date']) &
            (df['date'] <= config['end_date'])].copy()
    print(f"[1] Date filter: {len(df):,} obs (from {n0:,})")

    # Exchange filter
    df = df[df['exchange'].isin(config['exchanges'])].copy()
    print(f"[2] Exchange filter ({config['exchanges']}): {len(df):,} obs")

    # Price filter
    df = df[df['close'] >= config['min_price']].copy()
    print(f"[3] Price >= {config['min_price']:,} VND: {len(df):,} obs")

    # Compute daily return from adjusted prices
    df = df.sort_values(['ticker', 'date'])
    df['ret'] = df.groupby('ticker')['adjusted_close'].pct_change()

    # Flag price limit hits
    df['limit_hit'] = (
        ((df['exchange'] == 'HOSE') &
         (df['ret'].abs() >= config['price_limit_hose'] - 0.001)) |
        ((df['exchange'] == 'HNX') &
         (df['ret'].abs() >= config['price_limit_hnx'] - 0.001))
    )

    n_tickers = df['ticker'].nunique()
    print(f"\nFinal sample: {len(df):,} stock-day obs, "
          f"{n_tickers} unique tickers")
    print(f"Limit-hit days: {df['limit_hit'].sum():,} "
          f"({100*df['limit_hit'].mean():.2f}%)")
    return df
```

## 18.3 Corporate Action Adjustments

Proper adjustment for corporate actions is critical for volume-based DIVOP measures, as events such as stock splits, bonus share issues, and rights offerings change the number of shares outstanding and can create artificial spikes in measured turnover. We need to use cumulative adjustment factors that account for stock dividends (bonus shares), stock splits, rights offerings, and cash dividends (price adjustment only). We use these to construct adjusted volume and adjusted shares outstanding:

$$\text{AdjVolume}_{i,t} = \text{Volume}_{i,t} \times \text{CumAdjFactor}_{i,t}$$

$$\text{AdjSharesOut}_{i,t} = \text{SharesOut}_{i,t} \times \text{CumAdjFactor}_{i,t}$$

This ensures that the turnover ratio is consistent across corporate action events.

```
def adjust_for_corporate_actions(df):
    """Apply cumulative adjustment factors to volume and shares outstanding."""
    df = df.copy()
    df['adj_volume'] = df['volume'] * df['adj_factor']
    df['adj_shares_out'] = df['shares_outstanding'] * df['adj_factor']

    # Daily turnover ratio
    df['turnover'] = np.where(
        df['adj_shares_out'] > 0,
        df['adj_volume'] / df['adj_shares_out'],
        np.nan
    )

    # Flag extreme turnover (> 50% of float)
    extreme = df['turnover'] > 0.50
    if extreme.any():
        print(f"Warning: {extreme.sum()} obs with turnover > 50%, set to NaN")
        df.loc[extreme, 'turnover'] = np.nan

    return df
```

## 18.4 Trading Calendar Construction

The rolling regression approach for SUV and volatility requires a trading calendar that ensures each estimation window contains exactly the specified number of trading days. We construct this directly from observed trading dates.

```
def build_trading_calendar(df, config):
    """
    Map each trading date to its estimation window [est_start, est_end].

    For date t, the estimation window runs from
    t - gap - est_window to t - gap - 1 (in trading-day terms).
```

```
    """
    trading_dates = sorted(df['date'].unique())
    trading_dates = pd.Series(trading_dates)

    est_window = config['est_window']
    gap = config['gap']
    offset = est_window + gap

    records = []
    for i in range(offset, len(trading_dates)):
        records.append({
            'date': trading_dates.iloc[i],
            'est_start': trading_dates.iloc[i - gap - est_window],
            'est_end': trading_dates.iloc[i - gap - 1]
        })

    calendar = pd.DataFrame(records)
    print(f"Trading calendar: {len(calendar)} dates, "
          f"{calendar['date'].min()} to {calendar['date'].max()}")
    return calendar
```

# 19 Volume-Based DIVOP Proxies

## 19.1 Theoretical Motivation

Trading volume has long been recognized as a natural proxy for divergence of investor opinion. In the rational expectations framework of Milgrom and Stokey (1982), trade occurs only when investors disagree about the value of a security (i.e., a "no-trade theorem" that implies, by contrapositive, that observed trading volume must reflect some form of heterogeneous beliefs). Harris and Raviv (1993) and Kandel and Pearson (1995) formalize this intuition, showing that trading volume is positively related to the dispersion of investors' prior beliefs and to the degree to which public information is differentially interpreted.

The challenge in using raw trading volume as a DIVOP proxy is that volume is also driven by factors unrelated to disagreement, including portfolio rebalancing, liquidity needs, tax-loss selling, and index reconstitution effects. Garfinkel (2009) proposes two approaches to extract the disagreement component from raw volume. The first, **Unexplained Volume (DTO)**, removes market-wide volume effects and secular trends. The second, **Standardized Unexplained Volume (SUV)**, additionally controls for the information content of returns through a cross-sectional regression, isolating the "pure disagreement" component of trading activity.

## 19.2 Unexplained Volume (DTO)

### 19.2.1 Construction Methodology

The construction of the Unexplained Volume measure proceeds in four steps.

**Step 1: Compute firm-level daily turnover.** For each stock $i$ on day $t$:

$$\text{Turn}_{i,t} = \frac{\text{AdjVolume}_{i,t}}{\text{AdjSharesOut}_{i,t}}$$

**Step 2: Compute market-wide turnover.** We calculate aggregate turnover across all common stocks as a value-weighted average:

$$\text{MktTurn}_t = \frac{\sum_i \text{AdjVolume}_{i,t}}{\sum_i \text{AdjSharesOut}_{i,t}}$$

Unlike the U.S. methodology that computes market turnover across NYSE/AMEX stocks only and applies a scaling adjustment for NASDAQ securities (following A.-M. Anderson and Dyl 2005), we compute market turnover across all HOSE and HNX common stocks without any exchange-specific volume scaling. Both Vietnamese exchanges operate as order-driven markets (HOSE uses continuous order matching; HNX uses a combination of continuous matching and periodic call auctions) without the dealer-market double-counting issue that necessitates the NASDAQ volume adjustment in U.S. studies.

**Step 3: Compute market-adjusted turnover.**

$$\text{MATO}_{i,t} = \text{Turn}_{i,t} - \text{MktTurn}_t$$

**Step 4: Detrend by rolling median.** To remove secular trends in firm-specific trading activity:

$$\text{DTO}_{i,t} = \text{MATO}_{i,t} - \text{Median}_{180}(\text{MATO}_{i,t-7})$$

where $\text{Median}_{180}(\text{MATO}_{i,t-7})$ is the median of market-adjusted turnover over the 180-trading-day window ending 7 days before date $t$. The 7-day lag prevents the current day's turnover from influencing its own detrending baseline.

```
def compute_market_turnover(df):
    """Compute daily market-wide turnover across all stocks."""
    mkt_turn = df.groupby('date').apply(
        lambda x: x['adj_volume'].sum() / x['adj_shares_out'].sum()
        if x['adj_shares_out'].sum() > 0 else np.nan
    ).reset_index()
    mkt_turn.columns = ['date', 'market_turnover']
    return mkt_turn


def compute_dto(df, config):
    """
    Construct Unexplained Volume (DTO).

    Steps:
    1. Subtract market turnover -> MATO
    2. Rolling 180-day median of MATO (lagged 7 days) -> trend
```

```
    3. DTO = MATO - trend
    """
    detrend_window = config['detrend_window']
    lag = config['lag']

    # Market turnover
    mkt_turn = compute_market_turnover(df)
    df = df.merge(mkt_turn, on='date', how='left')

    # Market-adjusted turnover
    df['mato'] = df['turnover'] - df['market_turnover']

    # Rolling median with lag, computed per stock
    df = df.sort_values(['ticker', 'date'])

    def _rolling_median_lagged(group):
        mato = group['mato']
        med = mato.rolling(
            window=detrend_window,
            min_periods=int(detrend_window * 0.5)
        ).median()
        return med.shift(lag)

    df['mato_trend'] = (
        df.groupby('ticker', group_keys=False)
          .apply(lambda g: _rolling_median_lagged(g))
    )

    # DTO
    df['dto'] = df['mato'] - df['mato_trend']

    print("DTO construction complete.")
    print(f"  Non-missing: {df['dto'].notna().sum():,}")
    print(f"  Mean: {df['dto'].mean():.6f}, Std: {df['dto'].std():.6f}")
    return df
```

### 19.2.2 Vietnam-Specific Considerations for DTO

Several features of the Vietnamese market require attention when constructing DTO:

1. **No NASDAQ-type volume adjustment needed.** Both HOSE and HNX are order-driven auction markets. The double-counting adjustment applied to NASDAQ securities

in the U.S. literature is not necessary.

2. **Thinly traded stocks.** A substantial fraction of listed Vietnamese stocks, particularly on HNX, may have zero volume on many trading days. For stocks with intermittent trading, the rolling median may be biased toward zero, making DTO less informative. We require at least 80% non-zero volume days in each estimation window.

3. **Price limit effects on volume.** When a stock hits its daily price limit, unfilled orders accumulate and recorded volume may understate true clearing volume. The following day often shows a "catch-up" effect. Researchers should consider flagging limit-hit days.

4. **Foreign investor trading decomposition.** DataCore provides volume by investor type (foreign versus domestic). Researchers may wish to construct separate DTO measures for foreign and domestic volume, or use the foreign-to-domestic volume ratio as an additional dimension of disagreement.

## 19.3 Standardized Unexplained Volume (SUV)

### 19.3.1 Construction Methodology

The Standardized Unexplained Volume measure, proposed by Garfinkel (2009), isolates the disagreement component of volume by explicitly controlling for the information content of returns. The insight is that trading volume has both a **liquidity** component and an **informedness** component correlated with the magnitude and sign of returns. By regressing turnover on signed returns and extracting the standardized residual, SUV captures volume attributable to disagreement after controlling for both liquidity trends and information-driven trading.

For each stock $i$, on each trading date $t$, we estimate using data from the estimation window $[\tau_1, \tau_2]$:

$$\text{Turn}_{i,s} = \alpha_i + \beta_i^+ \cdot \text{RetPos}_{i,s} + \beta_i^- \cdot \text{RetNeg}_{i,s} + \epsilon_{i,s}, \quad s \in [\tau_1, \tau_2] \tag{19.1}$$

where $\text{RetPos}_{i,s} = |r_{i,s}| \cdot \mathbf{1}(r_{i,s} > 0)$ and $\text{RetNeg}_{i,s} = |r_{i,s}| \cdot \mathbf{1}(r_{i,s} < 0)$.

The Standardized Unexplained Volume on date $t$ is:

$$\text{SUV}_{i,t} = \frac{\text{Turn}_{i,t} - \widehat{\text{Turn}}_{i,t}}{\hat{\sigma}_{\epsilon,i}} \tag{19.2}$$

where $\widehat{\text{Turn}}_{i,t}$ is the predicted turnover and $\hat{\sigma}_{\epsilon,i}$ is the RMSE from Equation 19.1.

The asymmetric specification with separate coefficients for positive and negative returns reflects that the volume-return relation differs by return sign. In the U.S., buying pressure tends to

generate more volume than selling pressure due to short-sale frictions. In Vietnam, where short selling was unavailable until 2025, this asymmetry should be even more pronounced because all selling activity was constrained to existing shareholders.

```
def compute_suv(df, calendar, config):
    """
    Compute Standardized Unexplained Volume via rolling regressions.

    For each stock-date, regress Turn on RetPos and RetNeg over the
    estimation window, then compute SUV = (actual - predicted) / RMSE.
    """
    est_window = config['est_window']
    min_obs = int(est_window * config['min_volume_days'])

    # Prepare signed return components
    df = df.copy()
    df['ret_pos'] = np.where(df['ret'] > 0, np.abs(df['ret']), 0.0)
    df['ret_neg'] = np.where(
        (df['ret'] < 0) & df['ret'].notna(), np.abs(df['ret']), 0.0
    )

    results = []
    grouped = {t: g for t, g in df.groupby('ticker')}

    for _, cal_row in calendar.iterrows():
        dt = cal_row['date']
        est_s, est_e = cal_row['est_start'], cal_row['est_end']

        for ticker, tdata in grouped.items():
            # Estimation window
            est = tdata[
                (tdata['date'] >= est_s) & (tdata['date'] <= est_e)
            ].dropna(subset=['turnover', 'ret_pos', 'ret_neg'])

            if len(est) < min_obs:
                continue

            # Event date
            evt = tdata[tdata['date'] == dt]
            if evt.empty or evt['turnover'].isna().all():
                continue

            # OLS: Turn = alpha + beta_pos * RetPos + beta_neg * RetNeg
```

```
            X = est[['ret_pos', 'ret_neg']].values
            y = est['turnover'].values

            reg = LinearRegression().fit(X, y)
            y_hat = reg.predict(X)
            rmse = np.sqrt(np.mean((y - y_hat) ** 2))

            if rmse <= 0:
                continue

            # Predict and standardize for event date
            X_evt = evt[['ret_pos', 'ret_neg']].values
            pred = reg.predict(X_evt)[0]
            actual = evt['turnover'].values[0]
            suv = (actual - pred) / rmse

            results.append({
                'ticker': ticker, 'date': dt,
                'suv': suv,
                'predicted_turnover': pred,
                'rmse_turn': rmse,
                'n_est': len(est),
                'alpha_turn': reg.intercept_,
                'beta_pos': reg.coef_[0],
                'beta_neg': reg.coef_[1],
            })

    suv_df = pd.DataFrame(results)
    print(f"SUV: {len(suv_df):,} stock-date obs")
    print(f"  Mean: {suv_df['suv'].mean():.4f}, "
          f"Median: {suv_df['suv'].median():.4f}")
    return suv_df
```

### 19.3.2 Interpreting the SUV Regression Coefficients

The estimated coefficients from Equation 19.1 are informative about market microstructure. Garfinkel (2009) reports $\hat{\beta}^+ > \hat{\beta}^-$ for most U.S. stocks. In Vietnam, we expect this asymmetry to be even stronger because:

- **No short selling (pre-2025):** All selling is by existing shareholders, limiting volume response to negative returns.

- **T+2 settlement:** Investors cannot immediately reinvest sale proceeds, further dampening sell-side volume.
- **Price limits:** The $\pm$ 7% (HOSE) and $\pm$ 10% (HNX) daily limits truncate the return distribution, compressing the range of both regressors.

Researchers should report summary statistics of $(\hat{\alpha}, \hat{\beta}^+, \hat{\beta}^-, R^2)$ across the cross-section and over time.

```
def suv_diagnostics(suv_df):
    """Report cross-sectional summary of SUV regression parameters."""
    print("\n=== SUV Regression Diagnostics ===")

    params = ['alpha_turn', 'beta_pos', 'beta_neg']
    print(suv_df[params].describe(
        percentiles=[.05, .25, .50, .75, .95]
    ).T.to_string(float_format='{:.6f}'.format))

    # Asymmetry test
    diff = suv_df['beta_pos'] - suv_df['beta_neg']
    print(f"\nbeta_pos - beta_neg: mean = {diff.mean():.6f}, "
          f"frac > 0 = {(diff > 0).mean():.3f}")
```

# 20 Volatility-Based DIVOP Proxies

## 20.1 Total Return Volatility

### 20.1.1 Theoretical Motivation

Stock return volatility serves as a proxy for divergence of opinion through several channels. Shalen (1993) develops a model in which both volume and volatility are increasing in the dispersion of investor beliefs. Scheinkman and Xiong (2003) predict that higher volatility reflects the speculative trading component driven by overconfident investors who disagree about value. Empirically, Boehme, Danielsen, and Sorescu (2006) and Chatterjee, John, and Yan (2012) use idiosyncratic volatility as a DIVOP proxy and find it positively correlated with other disagreement measures and negatively associated with subsequent returns when short-sale constraints bind.

### 20.1.2 Construction

Total return volatility is the standard deviation of daily returns over the rolling estimation window:

$$\text{VOLATILITY}_{i,t} = \sqrt{\frac{1}{N_i - 1} \sum_{s \in [\tau_1, \tau_2]} (r_{i,s} - \bar{r}_i)^2} \tag{20.1}$$

where $N_i$ is the number of non-missing return observations for stock $i$ in the window $[\tau_1, \tau_2]$.

## 20.2 Idiosyncratic Volatility (IVOL)

Idiosyncratic volatility isolates firm-specific return variation by removing the systematic component explained by market movements. We compute IVOL from the residuals of a market model:

$$r_{i,s} = \alpha_i + \beta_i \cdot r_{m,s} + \epsilon_{i,s}, \quad s \in [\tau_1, \tau_2] \tag{20.2}$$

$$\text{IVOL}_{i,t} = \text{Std}(\hat{\epsilon}_{i,s}) \tag{20.3}$$

Researchers may extend this to a Eugene F. Fama and French (1993b) three-factor or five-factor model using Vietnamese factor portfolios constructed elsewhere in this book. A richer factor model yields IVOL estimates that better isolate truly idiosyncratic disagreement, at the cost of requiring factor portfolio construction.

```
def compute_volatility(df, calendar, config):
    """
    Compute total return volatility and idiosyncratic volatility
    via rolling estimation windows.

    Total vol = std(returns) in window.
    IVOL = std(residuals) from market model regression.
    """
    est_window = config['est_window']
    min_obs = int(est_window * config['min_volume_days'])

    # Value-weighted market return
    def _vw_ret(g):
        valid = g.dropna(subset=['ret'])
        if valid.empty:
            return np.nan
        w = valid['adj_shares_out'] * valid['close']
        return np.average(valid['ret'], weights=w)

    mkt_ret = df.groupby('date').apply(_vw_ret).reset_index()
    mkt_ret.columns = ['date', 'mkt_ret']
    df = df.merge(mkt_ret, on='date', how='left')

    results = []
    grouped = {t: g for t, g in df.groupby('ticker')}

    for _, cal_row in calendar.iterrows():
        dt = cal_row['date']
        est_s, est_e = cal_row['est_start'], cal_row['est_end']

        for ticker, tdata in grouped.items():
            est = tdata[
                (tdata['date'] >= est_s) & (tdata['date'] <= est_e)
            ].dropna(subset=['ret', 'mkt_ret'])
```

```
            if len(est) < min_obs:
                continue

            # Total volatility
            total_vol = est['ret'].std()

            # Market model -> IVOL
            X = est[['mkt_ret']].values
            y = est['ret'].values
            reg = LinearRegression().fit(X, y)
            resid = y - reg.predict(X)
            ivol = np.std(resid, ddof=1)

            results.append({
                'ticker': ticker, 'date': dt,
                'total_volatility': total_vol,
                'idio_volatility': ivol,
                'market_beta': reg.coef_[0],
                'market_alpha': reg.intercept_,
                'r_squared_mm': reg.score(X, y),
                'n_vol': len(est),
            })

    vol_df = pd.DataFrame(results)
    print(f"Volatility: {len(vol_df):,} stock-date obs")
    print(f"  Total vol (ann. mean): "
          f"{vol_df['total_volatility'].mean() * np.sqrt(252):.4f}")
    print(f"  IVOL (ann. mean): "
          f"{vol_df['idio_volatility'].mean() * np.sqrt(252):.4f}")
    return vol_df
```

### 20.2.1 Vietnam-Specific Considerations for Volatility

1. **Price limits compress measured volatility.** Daily limits of $\pm$ 7% (HOSE) and $\pm$ 10% (HNX) mechanically truncate the return distribution, leading to underestimation of true volatility. On limit-hit days, the true equilibrium return may exceed the observed return. Researchers should be aware that volatility-based DIVOP measures may be downward-biased for stocks that frequently hit limits.

2. **VN-Index concentration.** The VN-Index is highly concentrated, the top 10 stocks often account for 50-60% of index weight. For small- and mid-cap stocks, an equal-weighted

market return or a composite HOSE+HNX index may provide a better market factor in Equation 20.2.

3. **Thin trading and non-synchronous returns.** For thinly traded stocks, consecutive zero-return days can depress measured volatility. The Dimson (1979) adjustment (including lagged and lead market returns in the market model) may help correct for non-synchronous trading bias in the beta estimate, though its effect on IVOL is typically small.

# 21 Spread-Based and Liquidity DIVOP Proxies

## 21.1 Bid-Ask Spread (BASPREAD)

### 21.1.1 Theoretical Motivation

The bid-ask spread reflects the adverse selection costs faced by limit order providers. When investors hold heterogeneous beliefs, each trade is more likely to convey private information, raising the adverse selection component of the spread. Handa, Schwartz, and Tiwari (2003) show that in order-driven markets the spread widens when divergence of opinion increases because limit order providers face greater risk of being picked off by informed traders. Chung and Zhang (2014) demonstrate that closing bid-ask spreads from daily data provide a reliable approximation to intraday effective spreads.

### 21.1.2 Construction

We compute the proportional bid-ask spread using end-of-day quote data:

$$\text{BASPREAD}_{i,t} = \frac{\text{Ask}_{i,t} - \text{Bid}_{i,t}}{\text{Midpoint}_{i,t}} \tag{21.1}$$

where $\text{Midpoint}_{i,t} = (\text{Ask}_{i,t} + \text{Bid}_{i,t})/2$. When end-of-day bid and ask are unavailable, we use the daily high-low range as a fallback. Following Chung and Zhang (2014), we delete observations where both Bid and Ask are zero, and where the spread exceeds 50% of the midpoint.

## 21.2 Amihud Illiquidity (ILLIQ)

The Amihud (2002) ratio measures the price impact of order flow:

$$\text{ILLIQ}_{i,t} = \frac{|r_{i,t}|}{\text{DolVol}_{i,t}} \tag{21.2}$$

where $\text{DolVol}_{i,t} = \text{Volume}_{i,t} \times \text{Price}_{i,t}$ (in billions VND for scaling). Higher ILLIQ reflects greater information asymmetry. We average daily ratios over monthly horizons and use the log transformation due to heavy right skew.

```
def compute_spread_and_illiq(df, config):
    """Compute bid-ask spread (BASPREAD) and Amihud illiquidity."""
    df = df.copy()

    # --- Bid-Ask Spread ---
    df['midpoint_ba'] = (df['ask'] + df['bid']) / 2
    df['baspread_ba'] = np.where(
        (df['ask'] > 0) & (df['bid'] > 0) & (df['midpoint_ba'] > 0),
        (df['ask'] - df['bid']) / df['midpoint_ba'], np.nan
    )

    # Fallback: high/low range
    df['midpoint_hl'] = (df['high'] + df['low']) / 2
    df['baspread_hl'] = np.where(
        (df['high'] > 0) & (df['low'] > 0) & (df['midpoint_hl'] > 0),
        (df['high'] - df['low']) / df['midpoint_hl'], np.nan
    )

    df['baspread'] = df['baspread_ba'].fillna(df['baspread_hl'])
    df['midpoint'] = df['midpoint_ba'].fillna(df['midpoint_hl'])

    # Chung & Zhang (2009) filters
    bad = (df['baspread'].isna()) | \
          (df['baspread'] > config['max_spread_pct']) | \
          (df['baspread'] < 0)
    df.loc[bad, 'baspread'] = np.nan

    # --- Amihud Illiquidity ---
    df['dollar_vol'] = df['volume'] * df['close'] / 1e9
    df['amihud_daily'] = np.where(
        df['dollar_vol'] > 0,
        np.abs(df['ret']) / df['dollar_vol'], np.nan
    )

    print(f"BASPREAD: {df['baspread'].notna().sum():,} valid obs, "
          f"mean = {df['baspread'].mean():.6f}")
    print(f"AMIHUD: {df['amihud_daily'].notna().sum():,} valid obs, "
          f"mean = {df['amihud_daily'].mean():.6f}")
    return df
```

```python
def compute_amihud_monthly(df):
    """Monthly Amihud = mean daily |ret|/dollar_vol (min 15 days)."""
    df = df.copy()
    df['ym'] = df['date'].dt.to_period('M')
    agg = df.groupby(['ticker', 'ym']).agg(
        illiq_mean=('amihud_daily', 'mean'),
        n_days=('amihud_daily', 'count'),
    ).reset_index()
    agg = agg[agg['n_days'] >= 15].copy()
    agg['log_illiq'] = np.log(agg['illiq_mean'] + 1e-10)
    return agg
```

### 21.2.1 Vietnam-Specific Considerations for Spread and Liquidity

1. **Tick size schedule.** Vietnam uses variable tick sizes: 10 VND (prices $<$ 10,000), 50 VND (10,000–49,950), and 100 VND (  50,000) on HOSE. These impose a floor on quoted spreads for low-priced stocks. Researchers should be cautious interpreting cross-price-decile spread variation as reflecting opinion divergence rather than tick-size mechanics.

2. **Order-driven market structure.** Both HOSE and HNX are pure order-driven markets where public limit orders provide liquidity. This makes the Chung and Zhang (2014) CRSP-based spread approximation appropriate.

3. **Lot size requirements.** HOSE requires 100-share standard lots for continuous trading. For high-priced stocks, the standard lot represents a large capital commitment, potentially inflating quoted spreads relative to effective trading costs.

4. **Call auction effects.** Opening and closing sessions on HOSE use periodic call auctions, which can produce bid-ask quotes that differ substantially from continuous-trading spreads.

# 22 Analyst Forecast Dispersion

## 22.1 Theoretical Motivation

Analyst forecast dispersion, the cross-sectional standard deviation of individual analysts' earnings forecasts, is the most direct measure of divergence of opinion. Unlike market-based proxies that capture disagreement indirectly, forecast dispersion directly measures disagreement among informed market participants. Abarbanell, Lanen, and Verrecchia (1995) establish the theoretical basis, and Diether, Malloy, and Scherbina (2002) demonstrate that stocks with higher analyst forecast dispersion earn lower subsequent returns, consistent with the Miller overpricing hypothesis.

## 22.2 Data Challenges in Vietnam

Constructing analyst forecast dispersion in Vietnam presents substantial challenges relative to the U.S.:

- **Coverage breadth.** While I/B/E/S covers over 4,000 U.S. companies, only 100–150 Vietnamese firms typically have coverage by at least 3 analysts, concentrated among VN30 constituents.
- **Data sources.** Analyst forecasts are available from DataCore.vn, FiinPro, Bloomberg, and Refinitiv. The choice of source affects coverage and timeliness.
- **Forecast staleness.** With limited coverage, forecasts may go unrevised for months. Following I/B/E/S methodology, we carry each forecast forward for a maximum of 105 days.

## 22.3 Construction Methodology

The construction proceeds as follows:

1. **Clean individual forecasts.** Remove observations where the announcement date precedes the review date. Keep only annual EPS forecasts. For each analyst-ticker-fiscal period, retain only the latest forecast per calendar month.

2. **Handle stopped and excluded estimates.** Remove forecasts where the analyst has left the brokerage or the estimate has been excluded from consensus.
3. **Carry forward with staleness control.** Each forecast is valid until the earlier of: (a) the next forecast by the same analyst, (b) 105 days after the announcement, or (c) the actual earnings announcement date.
4. **Expand to monthly frequency.** For each ticker-month, identify all valid outstanding forecasts and compute dispersion.
5. **Compute scaled measures:**

$$\text{DISP1}_{i,m} = \frac{\text{Std}(\hat{\text{EPS}}_{i,m}^{(a)})}{|\text{Mean}(\hat{\text{EPS}}_{i,m}^{(a)})|} \qquad \text{DISP2}_{i,m} = \frac{\text{Std}(\hat{\text{EPS}}_{i,m}^{(a)})}{\bar{P}_{i,m}}$$

```python
def construct_analyst_dispersion(forecasts_df, price_df, config):
    """
    Construct analyst forecast dispersion measures.

    Parameters
    ----------
    forecasts_df : pd.DataFrame
        Individual analyst forecasts with: ticker, analyst_id, broker,
        fpedats, anndats, revdats, value (EPS), anndats_act.
    price_df : pd.DataFrame
        Monthly price: ticker, month, mean_price.
    config : dict
        With min_analysts, forecast_carry_days.
    """
    carry_days = config['forecast_carry_days']
    min_analysts = config['min_analysts']

    df = forecasts_df.copy()
    df = df[df['anndats'] <= df['revdats']].copy()
    df = df.dropna(subset=['fpedats', 'anndats', 'value'])

    # Latest forecast per analyst-month
    df['ym'] = df['anndats'].dt.to_period('M')
    df = df.sort_values(
        ['ticker', 'fpedats', 'analyst_id', 'ym', 'anndats', 'revdats']
    )
    df = df.groupby(['ticker', 'fpedats', 'analyst_id', 'ym']).tail(1)

    # Carry-forward end date
```

```
    df = df.sort_values(
        ['ticker', 'analyst_id', 'fpedats', 'anndats'],
        ascending=[True, True, True, False]
    )
    df['next_ann'] = df.groupby(
        ['ticker', 'analyst_id', 'fpedats']
    )['anndats'].shift(-1)

    def _carry_end(row):
        candidates = [row['anndats'] + pd.Timedelta(days=carry_days)]
        if pd.notna(row.get('next_ann')):
            candidates.append(row['next_ann'])
        if pd.notna(row.get('anndats_act')):
            candidates.append(row['anndats_act'])
        return min(candidates)

    df['carry_end'] = df.apply(_carry_end, axis=1)

    # Monthly expansion
    months = pd.period_range(config['beg_date'], config['end_date'], freq='M')
    records = []
    for month in months:
        me = month.to_timestamp(how='end')
        valid = df[(df['anndats'] <= me) & (df['carry_end'] > me)].copy()
        valid = valid[valid['fpedats'] > me]
        valid = valid.sort_values(['ticker', 'analyst_id', 'anndats'])
        valid = valid.groupby(['ticker', 'analyst_id']).tail(1)

        disp = valid.groupby('ticker').agg(
            n_analysts=('analyst_id', 'nunique'),
            mean_fcst=('value', 'mean'),
            std_fcst=('value', 'std'),
        ).reset_index()
        disp['month'] = month
        records.append(disp)

    if not records:
        return pd.DataFrame()
    disp_df = pd.concat(records, ignore_index=True)

    # Scaled measures
    disp_df['disp1'] = np.where(
```

```
        disp_df['mean_fcst'].abs() > 0,
        disp_df['std_fcst'] / disp_df['mean_fcst'].abs(), np.nan
    )
    disp_df = disp_df.merge(price_df, on=['ticker', 'month'], how='left')
    disp_df['disp2'] = np.where(
        disp_df['mean_price'] > 0,
        disp_df['std_fcst'] / disp_df['mean_price'], np.nan
    )
    disp_df['disp_raw'] = disp_df['std_fcst']

    out = disp_df[disp_df['n_analysts'] >= min_analysts].copy()
    print(f"DISP: {len(out):,} ticker-months (>= {min_analysts} analysts)")
    print(f"  Mean analysts: {out['n_analysts'].mean():.1f}")
    return out
```

## 22.4 Scaling Considerations

Following Cheong and Thomas (2011), we note that each scaling choice has pitfalls. DISP1 (scaled by absolute mean forecast) can produce extreme values when the mean forecast approaches zero—common for Vietnamese firms near breakeven. DISP2 (scaled by price) introduces a mechanical negative correlation between price and scaled dispersion. We recommend reporting all three versions (DISP1, DISP2, and unscaled DISP_RAW with ln(Price) as an additional control), and winsorizing DISP1 at the 1st and 99th percentiles.

> ⚠ Caution on Analyst Dispersion in Thin-Coverage Markets
>
> With typical coverage of 5–10 analysts per firm in Vietnam (versus 15–25 in the U.S.), forecast dispersion is estimated with substantially greater noise. A dispersion measure from 3 analysts has a very different sampling distribution than one from 20. Always include the number of analysts as a control and test robustness with varying minimum-analyst thresholds (3, 5, 7).

# 23 Cross-Sectional Correlations Among DIVOP Proxies

An important empirical question is the degree to which the various DIVOP proxies capture the same underlying construct. If divergence of opinion is a well-defined latent variable, we expect positive correlations among all proxies, though correlations need not be high since each captures a different facet of disagreement.

```
def compute_divop_correlations(merged_df, proxies=None):
    """
    Compute and visualize Spearman correlations among DIVOP proxies.
    We use rank correlations because many proxies are right-skewed.
    """
    if proxies is None:
        proxies = [
            'dto', 'suv', 'total_volatility', 'idio_volatility',
            'baspread', 'amihud_daily', 'disp1', 'disp2'
        ]
    available = [p for p in proxies if p in merged_df.columns]
    data = merged_df[available].dropna()

    n = len(available)
    rho_mat = np.eye(n)
    p_mat = np.zeros((n, n))
    for i in range(n):
        for j in range(i + 1, n):
            rho, p = scipy_stats.spearmanr(
                data[available[i]], data[available[j]]
            )
            rho_mat[i, j] = rho_mat[j, i] = rho
            p_mat[i, j] = p_mat[j, i] = p

    labels = {'dto': 'DTO', 'suv': 'SUV',
              'total_volatility': 'VOL', 'idio_volatility': 'IVOL',
              'baspread': 'SPREAD', 'amihud_daily': 'ILLIQ',
              'disp1': 'DISP1', 'disp2': 'DISP2'}
```

```
    pretty = [labels.get(c, c) for c in available]
    corr_df = pd.DataFrame(rho_mat, index=pretty, columns=pretty)

    # Heatmap
    fig, ax = plt.subplots(figsize=(9, 7))
    mask = np.triu(np.ones_like(corr_df, dtype=bool), k=1)
    sns.heatmap(
        corr_df, mask=mask, annot=True, fmt='.3f',
        cmap='RdBu_r', center=0, vmin=-0.4, vmax=0.7,
        square=True, linewidths=0.5,
        cbar_kws={'shrink': 0.8, 'label': 'Spearman  '}, ax=ax
    )
    ax.set_title('Spearman Correlations Among DIVOP Proxies\n'
                 'Vietnamese Equity Market', fontsize=13, fontweight='bold')
    plt.tight_layout()
    plt.savefig('divop_correlations.png', dpi=300, bbox_inches='tight')
    plt.show()

    return corr_df
```

### 23.0.1 Expected Correlation Patterns

Based on U.S. evidence and theory, we expect:

Table 23.1: Expected correlation structure among DIVOP proxies

| Pair | Expected | Rationale |
|---|---|---|
| DTO × SUV | High positive | Both capture abnormal volume; SUV refines DTO |
| VOL × IVOL | High positive | IVOL is a subset of total volatility |
| SPREAD × ILLIQ | Moderate-high positive | Both capture information asymmetry |
| Volume × Volatility | Moderate positive | Shalen (1993) links both to belief dispersion |
| Analyst × Market-based | Weak-moderate positive | Different investor populations |

# 24 Descriptive Statistics and Cross-Sectional Properties

## 24.1 Summary Statistics

```
def descriptive_statistics(merged_df):
    """Comprehensive descriptive statistics for DIVOP proxies."""
    proxies = {
        'dto': 'Unexplained Volume (DTO)',
        'suv': 'Std Unexplained Volume (SUV)',
        'total_volatility': 'Total Return Volatility',
        'idio_volatility': 'Idiosyncratic Volatility',
        'baspread': 'Bid-Ask Spread',
        'amihud_daily': 'Amihud Illiquidity',
        'disp1': 'Analyst Disp (mean-scaled)',
        'disp2': 'Analyst Disp (price-scaled)',
    }
    avail = {k: v for k, v in proxies.items() if k in merged_df.columns}
    rows = []
    for col, label in avail.items():
        s = merged_df[col].dropna()
        rows.append({
            'Proxy': label, 'N': f'{len(s):,}',
            'Mean': f'{s.mean():.6f}', 'Std': f'{s.std():.6f}',
            'P5': f'{s.quantile(.05):.6f}',
            'Median': f'{s.median():.6f}',
            'P95': f'{s.quantile(.95):.6f}',
            'Skew': f'{s.skew():.2f}',
            'Kurt': f'{s.kurtosis():.2f}',
        })
    stats = pd.DataFrame(rows).set_index('Proxy')
    print("\n" + "=" * 90)
    print("Descriptive Statistics of DIVOP Proxies")
    print("Vietnamese Equity Market, HOSE and HNX")
```

```
    print("=" * 90)
    print(stats.to_string())
    return stats
```

## 24.2 DIVOP by Firm Characteristics

```
def divop_by_size(merged_df):
    """Mean DIVOP proxies by market-cap quintile."""
    df = merged_df.copy()
    df['mkt_cap'] = df['close'] * df['shares_outstanding']
    df['size_q'] = df.groupby('date')['mkt_cap'].transform(
        lambda x: pd.qcut(x, 5,
            labels=['Q1 Small','Q2','Q3','Q4','Q5 Large'],
            duplicates='drop')
    )
    proxies = ['dto','suv','total_volatility','idio_volatility',
               'baspread','amihud_daily']
    avail = [p for p in proxies if p in df.columns]
    tab = df.groupby('size_q')[avail].mean()
    print("\n=== Mean DIVOP by Size Quintile ===")
    print(tab.to_string(float_format='{:.6f}'.format))
    return tab

def divop_by_exchange(merged_df):
    """Compare mean DIVOP across HOSE and HNX."""
    proxies = ['dto','suv','total_volatility','idio_volatility',
               'baspread','amihud_daily']
    avail = [p for p in proxies if p in merged_df.columns]
    tab = merged_df.groupby('exchange')[avail].mean()
    print("\n=== Mean DIVOP by Exchange ===")
    print(tab.to_string(float_format='{:.6f}'.format))
    return tab
```

## 24.3 Time-Series Evolution

```
def plot_divop_timeseries(merged_df):
    """Plot monthly cross-sectional median DIVOP with crisis shading."""
```

```
df = merged_df.copy()
df['ym'] = df['date'].dt.to_period('M')
proxies = ['dto','suv','total_volatility','baspread']
avail = [p for p in proxies if p in df.columns]
monthly = df.groupby('ym')[avail].median()
monthly.index = monthly.index.to_timestamp()

fig, axes = plt.subplots(len(avail), 1,
    figsize=(13, 3.5*len(avail)), sharex=True)
if len(avail) == 1: axes = [axes]

labels = {'dto':'DTO','suv':'SUV',
          'total_volatility':'Volatility','baspread':'Spread'}
colors = ['#1976D2','#388E3C','#F57C00','#D32F2F']

for i, (proxy, ax) in enumerate(zip(avail, axes)):
    ax.plot(monthly.index, monthly[proxy],
            color=colors[i], linewidth=1.3)
    ax.set_ylabel(labels.get(proxy, proxy), fontsize=10)
    ax.grid(True, alpha=0.25)
    for s, e, c in [('2008-01','2009-06','red'),
                    ('2020-01','2020-12','orange'),
                    ('2022-09','2023-06','purple')]:
        ax.axvspan(pd.Timestamp(s), pd.Timestamp(e),
                   alpha=0.1, color=c)

axes[0].set_title(
    'Time-Series of DIVOP Proxies\n'
    'Monthly Cross-Sectional Median, HOSE & HNX',
    fontsize=13, fontweight='bold')
from matplotlib.patches import Patch
axes[-1].legend(handles=[
    Patch(facecolor='red', alpha=.2, label='GFC 2008-09'),
    Patch(facecolor='orange', alpha=.2, label='COVID-19'),
    Patch(facecolor='purple', alpha=.2, label='Bond Crisis 2022-23'),
], loc='upper right', fontsize=8)
plt.tight_layout()
plt.savefig('divop_timeseries.png', dpi=300, bbox_inches='tight')
plt.show()
```

# 25 Putting It All Together

```python
def build_divop_dataset(config):
    """
    Master pipeline: load data, construct all DIVOP proxies,
    merge into a single stock-date panel.
    """
    df = load_daily_data(config)
    df = apply_sample_filters(df, config)
    df = adjust_for_corporate_actions(df)
    calendar = build_trading_calendar(df, config)

    df = compute_dto(df, config)
    suv_df = compute_suv(df, calendar, config)
    vol_df = compute_volatility(df, calendar, config)
    df = compute_spread_and_illiq(df, config)

    # Merge
    base = df[['ticker','date','ret','close','volume',
                'shares_outstanding','exchange','industry_icb',
                'foreign_ownership_pct','turnover',
                'mato','dto','baspread','amihud_daily','limit_hit']].copy()

    if not suv_df.empty:
        base = base.merge(
            suv_df[['ticker','date','suv','predicted_turnover']],
            on=['ticker','date'], how='left')
    if not vol_df.empty:
        base = base.merge(
            vol_df[['ticker','date','total_volatility',
                     'idio_volatility','market_beta']],
            on=['ticker','date'], how='left')

    print(f"\n=== Final DIVOP Dataset ===")
    print(f"Shape: {base.shape}")
```

```
    print(f"Tickers: {base['ticker'].nunique()}")
    return base
```

# 26 Empirical Applications

## 26.1 Application 1: DIVOP and the Cross-Section of Returns

The fundamental test of the Miller hypothesis is whether stocks with higher divergence of opinion earn lower subsequent returns. We implement Fama-MacBeth cross-sectional regressions:

$$r_{i,t+1:t+h} = \gamma_{0,t} + \gamma_{1,t} \cdot \text{DIVOP}_{i,t} + \gamma_{2,t}' \mathbf{X}_{i,t} + \varepsilon_{i,t}$$

where $\mathbf{X}_{i,t}$ includes controls for market beta, log market capitalization, and log book-to-market ratio. The Miller hypothesis predicts $\bar{\gamma}_1 < 0$.

```
def fama_macbeth_divop(merged_df, divop_proxy='suv',
                       controls=None, horizon=21):
    """
    Fama-MacBeth cross-sectional regressions.
    Miller predicts gamma_1 < 0; Varian predicts gamma_1 > 0.
    """
    if controls is None:
        controls = ['market_beta', 'log_mktcap']

    df = merged_df.copy()
    df = df.sort_values(['ticker', 'date'])
    df['fwd_ret'] = df.groupby('ticker')['ret'].transform(
        lambda x: x.shift(-1).rolling(horizon).sum().shift(-(horizon-1))
    )
    df['log_mktcap'] = np.log(
        df['close'] * df['shares_outstanding'] + 1
    )

    reg_vars = ['fwd_ret', divop_proxy] + \
               [c for c in controls if c in df.columns]
    df_reg = df[['ticker','date'] + reg_vars].dropna()

    from numpy.linalg import lstsq
```

```
    results = []
    for date, cross in df_reg.groupby('date'):
        if len(cross) < 30: continue
        y = cross['fwd_ret'].values
        X_cols = [divop_proxy] + [c for c in controls if c in cross.columns]
        X = np.column_stack([np.ones(len(cross)), cross[X_cols].values])
        try:
            coefs, _, _, _ = lstsq(X, y, rcond=None)
            results.append({
                'date': date, 'intercept': coefs[0],
                f'gamma_{divop_proxy}': coefs[1], 'n': len(cross),
            })
        except Exception: continue

    fm = pd.DataFrame(results)
    gc = f'gamma_{divop_proxy}'
    mu = fm[gc].mean()
    se = fm[gc].std() / np.sqrt(len(fm))
    t = mu / se

    print(f"\n=== Fama-MacBeth: {divop_proxy} -> "
          f"{horizon}-day fwd returns ===")
    print(f"  Mean gamma: {mu:.6f}, t-stat: {t:.3f}")
    if t < -1.96:   print("  -> Supports Miller (1977)")
    elif t > 1.96:   print("  -> Supports Varian (1985)")
    else:            print("  -> Inconclusive at 5%")
    return fm
```

## 26.2 Application 2: DIVOP and Earnings Announcements

Following Berkman et al. (2009), we test whether high-DIVOP stocks experience negative abnormal returns around earnings announcements, as uncertainty resolution reduces the optimism premium.

```
def divop_earnings_event(merged_df, ea_dates_df,
                         divop_proxy='suv', window=(-1, 3)):
    """
    Sort stocks into DIVOP quintiles pre-EA, compute CAR in window.
    Miller predicts: Q5 (high DIVOP) has lower CAR than Q1 (low DIVOP).
    """
    df = merged_df.copy()
```

```
    ea = ea_dates_df.copy()

    # Pre-EA DIVOP value (5 days before)
    ea['pre_date'] = ea['ea_date'] - pd.Timedelta(days=5)
    ea = ea.merge(
        df[['ticker','date',divop_proxy]].rename(
            columns={'date':'pre_date'}),
        on=['ticker','pre_date'], how='inner'
    )
    ea['divop_q'] = pd.qcut(
        ea[divop_proxy], 5,
        labels=['Q1 Low','Q2','Q3','Q4','Q5 High'],
        duplicates='drop'
    )

    print(f"\n=== EA Event Study by {divop_proxy} quintile ===")
    print(f"  Window: ({window[0]}, {window[1]}) days")
    print(f"  Miller predicts: Q5 has lower CAR than Q1")
    return ea
```

## 26.3 Application 3: Composite DIVOP Index via PCA

When a single summary measure of disagreement is needed, PCA on the battery of standardized proxies extracts the common "disagreement factor."

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

def composite_divop_pca(merged_df, proxies=None):
    """Extract first principal component from standardized DIVOP proxies."""
    if proxies is None:
        proxies = ['dto','suv','total_volatility','idio_volatility',
                   'baspread','amihud_daily']
    avail = [p for p in proxies if p in merged_df.columns]
    data = merged_df[['ticker','date'] + avail].dropna()

    scaler = StandardScaler()
    X = scaler.fit_transform(data[avail])

    pca = PCA(n_components=3)
    factors = pca.fit_transform(X)
```

```
    data['divop_composite'] = factors[:, 0]

    # Ensure positive correlation with inputs
    for col in avail:
        if data['divop_composite'].corr(data[col]) < 0:
            data['divop_composite'] *= -1
            break

    loadings = pd.DataFrame(
        pca.components_.T, index=avail,
        columns=['PC1','PC2','PC3']
    )

    print(f"\n=== PCA Composite DIVOP ===")
    print(f"Variance explained: "
          f"{pca.explained_variance_ratio_[:3].round(3)}")
    print(f"\nLoadings:\n{loadings.to_string(float_format='{:.4f}'.format)}")
    return data[['ticker','date','divop_composite']], loadings
```

# 27 Conclusion and Practical Recommendations

This chapter has provided a comprehensive methodology for constructing seven distinct proxies for divergence of investor opinion adapted to the Vietnamese equity market. We conclude with practical recommendations:

**1. Prefer multiple proxies.** No single DIVOP measure is without limitations. We recommend constructing and reporting results for at least three proxies spanning different economic channels (volume, volatility, spreads or analyst-based).

**2. Account for Vietnam-specific microstructure.** Daily price limits, T+2 settlement, foreign ownership constraints, and the order-driven market structure all affect DIVOP properties. Flag limit-hit days, include exchange fixed effects, and control for foreign ownership.

**3. Vietnam as a natural laboratory for Miller (1977).** The absence of short selling through 2024 and the dominance of retail investors create conditions that closely match Miller's theoretical setting. The introduction of short selling in 2025 creates a natural experiment for examining how relaxation of short-sale constraints affects the DIVOP-return relation.

**4. Control for analyst coverage when using DISP measures.** With typical coverage of 5–10 analysts per firm, forecast dispersion is estimated with greater noise than in developed markets. Always include the number of analysts as a control variable and conduct robustness checks with varying minimum-analyst thresholds.

**5. Consider constructing a composite index.** When researchers need a single summary measure of disagreement, the PCA-based composite index described in Chapter 40 provides a principled approach to aggregating information across the individual proxies. The first principal component typically explains 30-50% of the common variation in the battery of DIVOP measures.

**6. Winsorize aggressively.** Several DIVOP proxies (particularly DISP1, Amihud ILLIQ, and SUV) exhibit extreme outliers in the Vietnamese data. Winsorization at the 1st and 99th percentiles (or even 2nd and 98th for DISP1) is essential for obtaining reliable regression results.

**7. Be cautious about causal inference.** DIVOP proxies are endogenous, they respond to the same firm characteristics (size, leverage, growth) that also affect returns. Researchers should use appropriate controls, consider instrumental variables where feasible, and be explicit about the limitations of their identification strategy.

The DIVOP framework is particularly relevant for the Vietnamese market at this point in its development. As the market matures toward potential FTSE Emerging Market reclassification, as short selling becomes more widely available, and as institutional investor participation grows, the dynamics of opinion divergence and its pricing implications are likely to evolve significantly. The methodology presented in this chapter provides researchers with the tools to document and analyze these changes as they unfold.

# 28 P/E Ratio

The systematic analysis of equity valuations in frontier and emerging markets requires a robust integration of high-fidelity financial data, rigorous mathematical modeling, and an acute understanding of local regulatory frameworks. In the context of the Vietnamese equity market, the Price-Earnings (P/E) ratio serves as a primary instrument for assessing corporate performance and market sentiment. This chapter provides an exploration of P/E valuation methodologies, ranging from firm-specific trailing and forward metrics to cyclically adjusted and unlevered variations.

### 28.0.1 Technical Schema and Variable Engineering

The primary identifier for equities is the symbol, typically a three-letter uppercase code, though special indices such as E1VFVN30 (ETF) or market indices like HNX-INDEX and UPCOM-INDEX follow unique naming conventions. Price data is categorized as OHLC type, encompassing 'open', 'high', 'low', and 'close'. For valuation purposes, the 'adjusted close' is preferred to account for corporate actions such as stock splits and dividend payments.

The accounting variables provided in the Vietnam Fundamentals product follow a standardized classification that reconciles different reporting standards across sectors. This reconciliation is vital because Vietnamese firms may have differing interpretations of revenue and net profit under Vietnamese Accounting Standards (VAS).

## 28.1 Firm-Specific P/E Valuation Metrics

At the individual firm level, the P/E ratio is the most fundamental measure of how much an investor is willing to pay for each unit of current or future profit. However, the calculation of "E" (Earnings) can take several forms, each offering a different insight into the company's value.

### 28.1.1 Trailing P/E and the TTM Methodology

The Trailing P/E ratio is calculated using the reported earnings from the last four quarters, known as Trailing Twelve Months (TTM). In the Vietnamese market, this is often the default metric reported by exchanges and financial news outlets.

The mathematical representation is:

$$P/E_{Trailing} = \frac{P_t}{\sum_{i=0}^{3} EPS_{q-i}}$$

Where $P_t$ is the current market price and $EPS_{q-i}$ represents the earnings per share for each of the four most recent quarters.

### 28.1.2 Forward P/E and Analyst Forecast Reliability

Forward P/E shifts the focus from historical performance to future expectations by using consensus earnings forecasts for the next fiscal year.

$$P/E_{Forward} = \frac{P_t}{EPS_{forecasted,t+1}}$$

In Vietnam, these forecasts are typically generated by research departments within major securities firms such as SSI, VCSC, MBS, and FPTS. These analysts rely on financial statements, projections, models, and subjective evaluations of market sentiment to generate their estimates. However, the accuracy of these forecasts is a significant area of research. However, analysts in Vietnam are often influenced by anchoring and adjustment bias, where they base future predictions heavily on past records and then make insufficient adjustments.

The accuracy of analyst earnings forecasts in Vietnam is significantly impacted by corporate governance characteristics. Specifically, state ownership has been found to have a negative impact on forecast accuracy, while institutional ownership has a positive effect. This implies that for firms with high state-controlled stakes, which are common in Vietnam's strategic industries, researchers should apply a higher discount or "fudge factor" to forward P/E estimates provided by consensus sources.

### 28.1.3 Cyclically Adjusted Price-Earnings (CAPE) Ratio

The Shiller P/E, or CAPE ratio, is designed to smooth out the volatility of the business cycle by using a 10-year average of inflation-adjusted earnings. For a market like Vietnam, which has seen periods of extreme growth followed by stabilization, the CAPE ratio provides a more tempered view of valuation than trailing metrics.

The calculation requires adjusting historical earnings by the Consumer Price Index (CPI). Vietnam's CPI data is updated monthly and is available from January 1996 through early 2026, with an average year-on-year growth rate of approximately 12.2%.

$$EPS_{real,t} = EPS_{nominal,t} \times \frac{CPI_{current}}{CPI_t}$$

$$CAPE = \frac{P_t}{\frac{1}{10}\sum_{i=0}^{9} EPS_{real,t-i}}$$

Historically, Vietnam's CPI has reached an all-time high of 28.3% YoY in August 2008 and a record low of -2.6% in July 2000. These extreme swings mean that nominal earnings in the mid-2000s are not comparable to earnings today without the Shiller adjustment. Researchers must be mindful of the different base years used by the National Statistics Office (GSO), such as 2024=100, 2019=100, and 2014=100.

### 28.1.4 Unlevered P/E and the Leibowitz Framework

The concept of the "Unlevered P/E" ratio attempts to view a company's valuation independent of its capital structure. This is particularly relevant in the Vietnamese market, where debt loads vary significantly between state-owned enterprises (SOEs) and private firms. Research by Leibowitz (2002) highlights that for the investment analyst, a company is already levered, and the task is to estimate its theoretical value by inferring the underlying structure of returns.

According to Leibowitz, leverage always moves the P/E toward a lower value than what would be obtained from a standard Gordon growth formula. As a company adds debt, the market discount rate for equity increases to compensate for the additional risk, which in turn compresses the P/E multiple. For example, a debt-free company with a theoretical P/E of 30 might see its P/E drop to 23 with a 40% debt ratio, and down to 20 with a 50% debt ratio.

The Unlevered P/E (often proxied by Enterprise Value to EBIT or EBITDA) can be calculated using fundamental variables:

$$EV = (\text{Shares Outstanding} \times \text{Price}) + \text{Total Debt} - \text{Cash}$$

$$P/E_{\text{Unlevered}} \approx \frac{EV}{EBIT \times (1-\tau)}$$

Where $\tau$ is the corporate income tax rate. The standard CIT rate in Vietnam is 20%, which has been stable since the mid-2010s but is subject to new revisions under the Law on CIT ratified in June 2025.

## 28.2 Market-Level Aggregation Techniques

Aggregating firm-specific metrics into a single market-level index P/E is a complex task that requires careful consideration of weighting and the inclusion of loss-making entities. In the Vietnamese market, two primary indices (i.e., the VN-Index and the VN30), provide the benchmarks for performance and valuation.

### 28.2.1 The VN-Index: Capitalization-Weighted Aggregation

The VN-Index is a capitalization-weighted index of all companies listed on HOSE. It serves as a broad indicator of market health, where an increase in the index reflects a general rise in the stock prices of listed companies.

The index value is calculated as:

$$\text{VN-Index} = \frac{\sum(Price_i \times Shares_i)}{\text{Base Factor}} \times \text{Adjustment Factor}$$

Similarly, the index-level P/E is typically calculated as the sum of all market capitalizations divided by the sum of all net incomes (earnings) of the constituents.

$$P/E_{Market} = \frac{\sum MarketCap_i}{\sum Earnings_i}$$

As of early 2026, the total market capitalization for 701 tracked companies was approximately 8,629.8 trillion VND, with total earnings of 588.9 trillion VND, yielding a median P/E of approximately 14.7x.

### 28.2.2 The VN30 and Free-Float Adjustments

The VN30 Index represents a basket of the 30 largest and most liquid stocks on HOSE, screened through layers of capitalization, free-float ratio, and transactional volume. Unlike the broad VN-Index, the VN30 applies a free-float adjustment to its capitalization weighting to ensure that only shares available for public trading affect the index movement.

The free-float ratio ($f$) is calculated as:

$f = \frac{N - N_l}{N}$

Where $N$ is the total number of outstanding shares and $N_l$ is the number of limited trading shares (held by founders, state, or strategic partners).

When aggregating P/E for the VN30, researchers must account for this adjustment, as it more accurately reflects the valuation of the investable universe. In the Vietnamese market, a handful of large-cap stocks can significantly skew the broad index. For instance, Vingroup (VIC) and its related companies have at times represented nearly a quarter of the VN-Index's total weighting, creating a "closet indexing" dilemma for fund managers. Removing these high-impact stocks can reveal a much different valuation profile; while the VN-Index might trade at 13x, the market excluding VIC-related stocks could be closer to 11x.

### 28.2.3 Median vs. Mean Aggregation

Quantitative researchers often prefer median P/E ratios over mean P/E ratios to mitigate the influence of extreme outliers (i.e., companies with either exceptionally high P/E ratios due to temporarily low earnings or those that are technically "expensive" because of high growth expectations). Table 28.1 shows different aggregation methods.

Table 28.1: Median vs. Mean Aggregation

| **Aggregation Method** | **Mathematical Definition** | **Resilience to Outliers** |
|---|---|---|
| Simple Mean | $\frac{1}{n}\sum(P/E)_i$ | Low (highly skewed by extreme values) |
| Weighted Mean | $\frac{\sum(Cap_i \times (P/E)_i)}{\sum Cap_i}$ | Medium (skewed by large-cap valuation) |
| Median | $\text{Median}((P/E)_1, \dots, (P/E)_n)$ | High (most robust for market sentiment) |
| Total-to-Total | $\frac{\sum MarketCap_i}{\sum Earnings_i}$ | High (standard for index providers) |

The Vietnamese market currently trades near its 3-year average P/E of 14.8x, suggesting that investors are relatively neutral on current valuations, expecting earnings to grow in line with historical rates.

## 28.3 Implementation

### 28.3.1 Data Ingestion and Processing

### 28.3.2 Shiller CAPE

The Shiller CAPE requires a more sophisticated approach, involving the merging of earnings data with monthly CPI series.

## 28.4 Macroeconomic and Regulatory Factors in Valuation

The P/E ratio does not exist in a vacuum; it is deeply influenced by the macroeconomic environment and the regulatory framework within which companies operate.

### 28.4.1 The Role of Corporate Income Tax (CIT)

The earnings component of the P/E ratio is a post-tax metric. In Vietnam, the standard CIT rate is 20%. However, there is a significant shift occurring with the ratification of the new Law on CIT in June 2025, taking effect on October 1, 2025. This law introduces:

- A 15% rate for enterprises with annual revenue not exceeding 3 billion VND.
- A 17% rate for enterprises with revenue between 3 and 50 billion VND.
- The maintenance of higher rates (up to 50%) for the oil, gas, and mineral resource sectors.

Furthermore, Vietnam has adopted the Global Minimum Tax (GMT) rules under Pillar Two, effective January 1, 2024, to protect its tax revenue from inbound and outbound investments. These changes mean that historical P/E ratios may not be directly comparable to future ratios for companies that previously benefited from generous tax incentives, as the effective tax rate is likely to rise for large multinational subsidiaries operating in Vietnam.

### 28.4.2 Macroeconomic Determinants: Inflation and Interest Rates

Stock price indices, and by extension P/E ratios, are heavily influenced by fundamental macroeconomic factors such as inflation, exchange rates, and interest rates. The interest rate serves as the cost of capital for enterprises. When interest rates fall, the cost of borrowing drops, increasing corporate profits and potentially raising the P/E that investors are willing to pay.

The Vietnamese dong (VND) exchange rate also plays a crucial role. In early 2025, the VND depreciated by 1.33% YTD, reflecting volatility in the timeline for Federal Reserve rate cuts. Exchange rate depreciation can be inflationary by default, which may prompt the State Bank of Vietnam to tighten monetary policy, thereby putting downward pressure on equity valuations.

## 28.5 Synthesis of Valuation Dynamics

The Vietnamese equity market is currently in a state of transition. While the market has seen robust growth—the VN-Index rose 38% in the year leading up to early 2026—investors remain neutral on overall valuations, as reflected by the market's alignment with its 3-year average P/E of 14.8x. This neutrality suggests that investors expect earnings growth (forecast at 14% annually) to keep pace with price appreciation.

The reliability of these valuations, however, remains dependent on the quality of non-financial disclosure. While companies are increasingly required by law to provide non-financial information, the level of transparency remains at a medium level (approximately 58.5% of required disclosures). Researchers must therefore balance quantitative P/E analysis with qualitative assessments of corporate governance and sustainability strategies. Companies that effectively integrate ESG principles have been shown to experience fewer negative impacts on abnormal stock returns during uncertain periods, such as the COVID-19 pandemic.

## 28.6 Conclusions

The exploration of P/E ratio valuation and aggregation within the Vietnamese equity market reveals a sophisticated landscape where traditional metrics must be adapted to local realities. The reliance on trailing metrics, while common, fails to account for the cyclicality and high inflation history of the Vietnamese economy, a gap that the Shiller CAPE ratio effectively bridges through CPI-adjusted normalization. Furthermore, the sensitivity of P/E to capital structure, as defined in the Leibowitz framework, highlights the necessity of unlevered valuation metrics in a market with diverse corporate funding models.

The aggregation of these metrics into market-level indices like the VN-Index and VN30 requires an awareness of concentration risks and the importance of free-float adjustments. The "Vingroup Effect" demonstrates how a single corporate ecosystem can skew index-wide valuations, necessitating a "closet indexing" awareness for fund managers and researchers.

Looking forward, the evolution of the regulatory environment—specifically the CIT law of 2025 and the Global Minimum Tax—will introduce new variables into the valuation equation. The shift toward higher-quality non-financial disclosures and the increasing accuracy of analyst forecasts in institutionalized firms suggest that the Vietnamese equity market is steadily maturing, offering a more transparent and predictable environment for the application of advanced quantitative valuation techniques.

# 29 Institutional Ownership Analytics in Vietnam

## 29.1 Institutional Ownership in Vietnam: A Distinct Landscape

Vietnam's equity market presents a fundamentally different institutional ownership landscape from the mature markets of the US, Europe, or Japan. Since the Ho Chi Minh City Securities Trading Center (now HOSE) opened on July 28, 2000 with just two listed stocks, the market has grown to over 1,700 listed companies across three exchanges (HOSE, HNX, and UPCOM) with a combined market capitalization exceeding 200 billion USD. Yet the ownership structure remains distinctive in several critical ways:

- **Retail dominance.** Individual investors account for approximately 85% of trading value on Vietnamese exchanges, far exceeding the institutional share. This contrasts sharply with the US, where institutional investors dominate both ownership and trading (Bao Dinh and Tran 2024). The implications for market efficiency, price discovery, and volatility are profound.
- **State ownership legacy.** Vietnam's equitization (privatization) program, initiated under Đổi Mới reforms in 1986, means that the state remains a significant or controlling shareholder in many listed companies. As of 2022, SOEs (firms with state ownership > 50%) account for approximately 30% of total market capitalization despite representing less than 10% of listed firms (Huang, Liu, and Shu 2023). State ownership introduces unique agency problems, governance dynamics, and liquidity constraints.
- **Foreign Ownership Limits (FOLs).** Vietnam imposes sector-specific caps on aggregate foreign ownership, typically 49% for most sectors, 30% for banking, and varying limits for aviation, media, and telecommunications. When a stock reaches its FOL, foreign investors can only buy from other foreign sellers, creating a segmented market with distinct pricing dynamics and a well-documented "FOL premium" (Vo 2015).
- **Disclosure regime.** Unlike the US quarterly 13F filing system, Vietnam's ownership disclosure is event-driven and periodic. Major shareholders ( 5%) must disclose within 7 business days of crossing thresholds. Annual reports contain detailed shareholder registers. Semi-annual fund reports provide portfolio snapshots. This creates a patchwork of disclosure frequencies that require careful handling.

## 29.2 Data Infrastructure: DataCore.vn

**DataCore.vn** is a comprehensive Vietnamese financial data platform that provides academic-grade datasets for the Vietnamese market. Throughout this chapter, we assume all data is sourced exclusively from DataCore.vn, which provides:

Table 29.1: DataCore.vn Data Tables Used in This Chapter

| DataCore.vn Dataset | Content | Key Variables |
|---|---|---|
| **Stock Prices** | Daily/monthly OHLCV for HOSE, HNX, UPCOM | `ticker`, `date`, `close`, `adjusted_close`, `volume`, `shares_outstanding` |
| **Ownership Structure** | Shareholder composition snapshots | `ticker`, `date`, `shareholder_name`, `shares_held`, `ownership_pct`, `shareholder_type` |
| **Major Shareholders** | Detailed  5% holders | `ticker`, `date`, `shareholder_name`, `shares_held`, `is_foreign`, `is_state`, `is_institution` |
| **Corporate Actions** | Dividends, stock splits, bonus shares, rights issues | `ticker`, `ex_date`, `action_type`, `ratio`, `record_date` |
| **Company Profile** | Sector, exchange, listing date, charter capital | `ticker`, `exchange`, `industry_code`, `listing_date`, `fol_limit` |
| **Financial Statements** | Quarterly/annual financials | `ticker`, `period`, `revenue`, `net_income`, `total_assets`, `equity` |
| **Foreign Ownership** | Daily foreign ownership tracking | `ticker`, `date`, `foreign_shares`, `foreign_pct`, `fol_limit`, `foreign_room` |
| **Fund Holdings** | Semi-annual fund portfolio disclosures | `fund_name`, `report_date`, `ticker`, `shares_held`, `market_value` |

```
class DataCoreReader:
    """
    Unified data reader for DataCore.vn datasets.

    Assumes data has been downloaded from DataCore.vn and stored locally.
    Supports both Parquet (recommended for performance) and CSV formats.

    Parameters
```

```
    ----------
    data_dir : str or Path
        Root directory containing DataCore.vn data files
    file_format : str
        'parquet' or 'csv' (default: 'parquet')
    """

    # Expected file names in the data directory
    FILE_MAP = {
        'prices': 'stock_prices',
        'ownership': 'ownership_structure',
        'major_shareholders': 'major_shareholders',
        'corporate_actions': 'corporate_actions',
        'company_profile': 'company_profile',
        'financials': 'financial_statements',
        'foreign_ownership': 'foreign_ownership_daily',
        'fund_holdings': 'fund_holdings',
    }

    def __init__(self, data_dir: Union[str, Path], file_format: str = 'parquet'):
        self.data_dir = Path(data_dir)
        self.fmt = file_format
        self._cache = {}

        # Verify data directory exists
        if not self.data_dir.exists():
            raise FileNotFoundError(
                f"Data directory not found: {self.data_dir}\n"
                f"Please download data from DataCore.vn and place it in this directory."
            )

        print(f"DataCore.vn reader initialized: {self.data_dir}")
        available = [f.stem for f in self.data_dir.glob(f'*.{self.fmt}')]
        print(f"Available datasets: {available}")

    def _read(self, key: str) -> pd.DataFrame:
        """Read and cache a dataset."""
        if key in self._cache:
            return self._cache[key]

        fname = self.FILE_MAP.get(key, key)
        filepath = self.data_dir / f"{fname}.{self.fmt}"
```

```
        if not filepath.exists():
            raise FileNotFoundError(
                f"Dataset not found: {filepath}\n"
                f"Expected file: {fname}.{self.fmt} in {self.data_dir}"
            )

        if self.fmt == 'parquet':
            df = pd.read_parquet(filepath)
        else:
            df = pd.read_csv(filepath, parse_dates=True)

        # Auto-detect and parse date columns
        for col in df.columns:
            if 'date' in col.lower() or col.lower() in ['period', 'ex_date', 'record_date']:
                try:
                    df[col] = pd.to_datetime(df[col])
                except (ValueError, TypeError):
                    pass

        self._cache[key] = df
        print(f"Loaded {key}: {len(df):,} rows, {len(df.columns)} columns")
        return df

    @property
    def prices(self) -> pd.DataFrame:
        return self._read('prices')

    @property
    def ownership(self) -> pd.DataFrame:
        return self._read('ownership')

    @property
    def major_shareholders(self) -> pd.DataFrame:
        return self._read('major_shareholders')

    @property
    def corporate_actions(self) -> pd.DataFrame:
        return self._read('corporate_actions')

    @property
    def company_profile(self) -> pd.DataFrame:
        return self._read('company_profile')
```

```python
    @property
    def financials(self) -> pd.DataFrame:
        return self._read('financials')

    @property
    def foreign_ownership(self) -> pd.DataFrame:
        return self._read('foreign_ownership')

    @property
    def fund_holdings(self) -> pd.DataFrame:
        return self._read('fund_holdings')

    def clear_cache(self):
        """Clear all cached datasets to free memory."""
        self._cache.clear()

# Initialize reader - adjust path to your local DataCore.vn data
# dc = DataCoreReader('/path/to/datacore_data', file_format='parquet')
```

This chapter proceeds as follows. Section 29.3 builds the complete data pipeline from raw DataCore.vn extracts to clean, analysis-ready datasets, with particular attention to corporate action adjustments. Section 29.4 defines Vietnam's unique ownership taxonomy. Section 29.5 computes institutional ownership ratios, concentration, and breadth for the Vietnamese market. Section 29.6 develops specialized foreign ownership analytics including FOL utilization and room premium. Section 29.7 derives institutional trades from ownership disclosure snapshots. Section 29.8 computes fund-level flows and turnover. Section 29.9 analyzes state ownership dynamics. Section 29.10 introduces network analysis, ML classification, and event-study frameworks. Section 29.11 presents complete empirical applications, and Section 29.12 concludes.

## 29.3 Data Pipeline

### 29.3.1 Stock Price Data and Corporate Action Adjustments

Vietnam's equity market is notorious for frequent corporate actions, particularly stock dividends and bonus share issuances, that dramatically alter share counts. A company issuing a 30% stock dividend means every 100 shares become 130 shares, and the reference price adjusts downward proportionally. Failure to properly adjust historical shares and prices for these events is the single most common source of error in Vietnamese equity research.

```python
# ==============================================================================
# Step 1: Corporate Action Adjustment Factors
# ==============================================================================

def build_adjustment_factors(corporate_actions: pd.DataFrame) -> pd.DataFrame:
    """
    Build cumulative adjustment factors from the corporate actions history.

    In Vietnam, the most common share-altering corporate actions are:
    1. Stock dividends (cổ tức bằng cổ phiếu): e.g., 30% → ratio = 0.30
       Effect: shares × (1 + 0.30), price × (1 / 1.30)
    2. Bonus shares (thưởng cổ phiếu): mechanically identical to stock dividends
    3. Stock splits (chia tách): e.g., 2:1 → ratio = 2.0
       Effect: shares × 2, price × 0.5
    4. Rights issues (phát hành thêm): dilutive, but not all shareholders exercise
       We approximate with the subscription ratio
    5. Reverse splits (gộp cổ phiếu): rare in Vietnam
       Effect: shares ÷ ratio, price × ratio

    We construct a FORWARD-LOOKING cumulative adjustment factor such that:
       adjusted_shares = raw_shares × cum_adj_factor(from_date, to_date)
       adjusted_price = raw_price / cum_adj_factor(from_date, to_date)

    This is analogous to CRSP's cfacshr in the US context.

    Parameters
    ----------
    corporate_actions : pd.DataFrame
        DataCore.vn corporate actions with columns:
        ticker, ex_date, action_type, ratio

        action_type values:
        - 'stock_dividend': ratio = dividend rate (e.g., 0.30 for 30%)
        - 'bonus_shares': ratio = bonus rate (e.g., 0.20 for 20%)
        - 'stock_split': ratio = split factor (e.g., 2.0 for 2:1)
        - 'reverse_split': ratio = merge factor (e.g., 5.0 for 5:1 merge)
        - 'rights_issue': ratio = subscription rate (e.g., 0.10 for 10:1)
        - 'cash_dividend': ratio = VND per share (no share adjustment needed)

    Returns
    -------
    pd.DataFrame
```

```
        Adjustment factors: ticker, ex_date, point_factor, cum_factor
    """
    # Filter to share-altering events only
    share_events = ['stock_dividend', 'bonus_shares', 'stock_split',
                    'reverse_split', 'rights_issue']
    ca = corporate_actions[
        corporate_actions['action_type'].isin(share_events)
    ].copy()

    if len(ca) == 0:
        print("No share-altering corporate actions found.")
        return pd.DataFrame(columns=['ticker', 'ex_date', 'point_factor', 'cum_factor'])

    # Compute point adjustment factor for each event
    def compute_point_factor(row):
        atype = row['action_type']
        ratio = row['ratio']

        if atype in ['stock_dividend', 'bonus_shares']:
            # 30% stock dividend: 100 shares → 130 shares
            return 1 + ratio
        elif atype == 'stock_split':
            # 2:1 split: 100 shares → 200 shares
            return ratio
        elif atype == 'reverse_split':
            # 5:1 reverse: 500 shares → 100 shares
            return 1.0 / ratio
        elif atype == 'rights_issue':
            # Approximate: assume all rights exercised
            # In practice, this overestimates the adjustment
            return 1 + ratio
        else:
            return 1.0

    ca['point_factor'] = ca.apply(compute_point_factor, axis=1)

    # Sort chronologically within each ticker
    ca = ca.sort_values(['ticker', 'ex_date']).reset_index(drop=True)

    # Cumulative factor: product of all point factors from listing to date
    # This gives us a running "total adjustment" for each ticker
    ca['cum_factor'] = ca.groupby('ticker')['point_factor'].cumprod()
```

```python
    # Summary statistics
    n_tickers = ca['ticker'].nunique()
    n_events = len(ca)
    avg_events = n_events / n_tickers if n_tickers > 0 else 0

    print(f"Corporate action adjustment factors built:")
    print(f"  Tickers with adjustments: {n_tickers:,}")
    print(f"  Total share-altering events: {n_events:,}")
    print(f"  Average events per ticker: {avg_events:.1f}")
    print(f"\nEvent type distribution:")
    print(ca['action_type'].value_counts().to_string())

    return ca[['ticker', 'ex_date', 'action_type', 'ratio',
               'point_factor', 'cum_factor']]


def adjust_shares(shares: float, ticker: str, from_date, to_date,
                  adj_factors: pd.DataFrame) -> float:
    """
    Adjust a share count from one date to another for corporate actions.

    Example: If a company had a 30% stock dividend with ex_date between
    from_date and to_date, then 1000 shares at from_date = 1300 shares
    at to_date.

    Parameters
    ----------
    shares : float
        Number of shares at from_date
    ticker : str
        Stock ticker
    from_date, to_date : pd.Timestamp
        Period for adjustment
    adj_factors : pd.DataFrame
        Output of build_adjustment_factors()

    Returns
    -------
    float
        Adjusted shares at to_date
    """
    events = adj_factors[
```

```
        (adj_factors['ticker'] == ticker) &
        (adj_factors['ex_date'] > pd.Timestamp(from_date)) &
        (adj_factors['ex_date'] <= pd.Timestamp(to_date))
    ]

    if len(events) == 0:
        return shares

    total_factor = events['point_factor'].prod()
    return shares * total_factor


# Example usage:
# adj_factors = build_adjustment_factors(dc.corporate_actions)
```

**❗ The Stock Dividend Problem in Vietnam**

Vietnamese companies issue stock dividends with remarkable frequency, many growth companies do so 2-3 times per year. Consider **Vinhomes (VHM)** or **FPT Corporation**: their share counts may double or triple over a 5-year period purely from stock dividends. If you compare raw ownership shares from 2019 to 2024 without adjustment, you will obtain nonsensical ownership ratios. **Every time-series analysis of Vietnamese ownership data must use adjusted shares.** This is the Vietnamese equivalent of the CRSP cfacshr adjustment factor problem in US data, but more severe because the events are more frequent and larger in magnitude.

```
# ============================================================================
# Step 2: Process Stock Price Data
# ============================================================================

def process_price_data(prices: pd.DataFrame,
                       adj_factors: pd.DataFrame,
                       company_profile: pd.DataFrame) -> pd.DataFrame:
    """
    Process DataCore.vn stock price data:
    1. Align dates to month-end and quarter-end
    2. Merge company metadata (exchange, sector, FOL limit)
    3. Compute adjusted prices and shares outstanding
    4. Compute market capitalization
    5. Create quarter-end snapshots

    Parameters
```

```
    ----------
    prices : pd.DataFrame
        Daily/monthly price data from DataCore.vn
    adj_factors : pd.DataFrame
        Corporate action adjustment factors
    company_profile : pd.DataFrame
        Company metadata including exchange, sector, FOL

    Returns
    -------
    pd.DataFrame
        Quarter-end processed stock data
    """
    df = prices.copy()

    # Standardize date
    df['date'] = pd.to_datetime(df['date'])
    df['month_end'] = df['date'] + pd.offsets.MonthEnd(0)
    df['quarter_end'] = df['date'] + pd.offsets.QuarterEnd(0)

    # Merge company profile
    profile_cols = ['ticker', 'exchange', 'industry_code', 'fol_limit',
                    'listing_date', 'company_name']
    profile_cols = [c for c in profile_cols if c in company_profile.columns]
    df = df.merge(company_profile[profile_cols], on='ticker', how='left')

    # Build cumulative adjustment factor for each ticker-date
    # For each observation, compute the total adjustment from listing to that date
    df = df.sort_values(['ticker', 'date'])

    # Merge adjustment events
    # For each ticker-date, find the cumulative factor as of that date
    def get_cum_factor_at_date(group):
        ticker = group.name
        ticker_adj = adj_factors[adj_factors['ticker'] == ticker].copy()

        if len(ticker_adj) == 0:
            group['cum_adj_factor'] = 1.0
            return group

        # For each date, find cumulative factor (product of all events up to that date)
        group = group.sort_values('date')
```

```
        group['cum_adj_factor'] = 1.0

        for _, event in ticker_adj.iterrows():
            mask = group['date'] >= event['ex_date']
            group.loc[mask, 'cum_adj_factor'] *= event['point_factor']

        return group

    df = df.groupby('ticker', group_keys=False).apply(get_cum_factor_at_date)

    # Adjusted price and shares
    # adjusted_close should already be provided by DataCore.vn
    # But we compute our own for consistency
    if 'adjusted_close' not in df.columns:
        df['adjusted_close'] = df['close'] / df['cum_adj_factor']

    # Adjusted shares outstanding
    df['adjusted_shares'] = df['shares_outstanding'] * df['cum_adj_factor']

    # Market capitalization (in billion VND)
    df['market_cap'] = df['close'] * df['shares_outstanding'] / 1e9

    # Monthly returns
    df = df.sort_values(['ticker', 'date'])
    df['ret'] = df.groupby('ticker')['adjusted_close'].pct_change()

    # Keep quarter-end observations
    # For daily data: keep last trading day of each quarter
    df_quarterly = (df.sort_values(['ticker', 'quarter_end', 'date'])
                      .groupby(['ticker', 'quarter_end'])
                      .last()
                      .reset_index())

    print(f"Processed price data:")
    print(f"  Total records (daily): {len(df):,}")
    print(f"  Quarter-end records: {len(df_quarterly):,}")
    print(f"  Unique tickers: {df_quarterly['ticker'].nunique():,}")
    print(f"  Date range: {df_quarterly['quarter_end'].min()} to "
          f"{df_quarterly['quarter_end'].max()}")
    print(f"\nExchange distribution:")
    print(df_quarterly.groupby('exchange')['ticker'].nunique().to_string())
```

```
    return df_quarterly

# prices_q = process_price_data(dc.prices, adj_factors, dc.company_profile)
```

### 29.3.2 Ownership Structure Data

Vietnamese ownership data captures the composition of shareholders as disclosed in annual reports, semi-annual reports, and event-driven disclosures. The key distinction from US 13F data is that Vietnamese disclosures provide a **complete ownership decomposition**, not just institutional long positions, but the full breakdown into state, institutional, foreign, and individual ownership.

```
# ============================================================================
# Step 3: Process Ownership Structure Data
# ============================================================================

class OwnershipType:
    """
    Vietnam's ownership taxonomy.

    Unlike the US where 13F captures only institutional long positions,
    Vietnamese disclosure provides a complete ownership decomposition.
    We classify shareholders into five mutually exclusive categories.
    """
    STATE = 'state'                     # Nhà nước (government entities, SOE parents)
    FOREIGN_INST = 'foreign_inst'       # Tổ chức nước ngoài
    DOMESTIC_INST = 'domestic_inst'     # Tổ chức trong nước (non-state)
    INDIVIDUAL = 'individual'           # Cá nhân
    TREASURY = 'treasury'               # Cổ phiếu quỹ

    ALL_TYPES = [STATE, FOREIGN_INST, DOMESTIC_INST, INDIVIDUAL, TREASURY]
    INSTITUTIONAL = [STATE, FOREIGN_INST, DOMESTIC_INST]
    FOREIGN = [FOREIGN_INST]  # Can be expanded if foreign individuals are tracked


def classify_shareholders(ownership: pd.DataFrame) -> pd.DataFrame:
    """
    Classify shareholders into Vietnam's ownership taxonomy.

    DataCore.vn may provide a `shareholder_type` field, but naming
    conventions vary. This function standardizes the classification
```

```
    using a combination of provided flags and name-based heuristics.

    The classification challenge in Vietnam (noted by @huang2023factors):
    DataCore.vn may not always cleanly separate institution types, so we
    use a cascading approach:
    1. Use explicit flags (is_state, is_foreign, is_institution) if available
    2. Apply name-based heuristics for Vietnamese entity names
    3. Default to 'individual' for unclassified shareholders

    Parameters
    ----------
    ownership : pd.DataFrame
        Raw ownership data from DataCore.vn

    Returns
    -------
    pd.DataFrame
        Ownership data with standardized `owner_type` column
    """
    df = ownership.copy()

    # --- Method 1: Use explicit flags if available ---
    if all(col in df.columns for col in ['is_state', 'is_foreign', 'is_institution']):
        conditions = [
            (df['is_state'] == True),
            (df['is_foreign'] == True) & (df['is_institution'] == True),
            (df['is_foreign'] == True) & (df['is_institution'] != True),
            (df['is_institution'] == True) & (df['is_state'] != True) &
                (df['is_foreign'] != True),
        ]
        choices = [
            OwnershipType.STATE,
            OwnershipType.FOREIGN_INST,
            OwnershipType.FOREIGN_INST,  # Foreign individuals often grouped
            OwnershipType.DOMESTIC_INST,
        ]
        df['owner_type'] = np.select(conditions, choices,
                                     default=OwnershipType.INDIVIDUAL)

    # --- Method 2: Name-based heuristics ---
    elif 'shareholder_name' in df.columns:
        name = df['shareholder_name'].str.lower().fillna('')
```

```
    # State entities: government ministries, SCIC, state corporations
    state_keywords = [
        'bộ tài chính', 'tổng công ty đầu tư', 'scic',
        'ủy ban nhân dân', 'nhà nước', 'state capital',
        'tổng công ty', 'vốn nhà nước', 'bộ công thương',
        'bộ quốc phòng', 'bộ giao thông', 'vinashin',
    ]
    is_state = name.apply(
        lambda x: any(kw in x for kw in state_keywords)
    )

    # Foreign entities: common fund names, foreign company patterns
    foreign_keywords = [
        'fund', 'investment', 'capital', 'limited', 'ltd', 'inc',
        'corporation', 'holdings', 'asset management', 'pte',
        'gmbh', 'management', 'partners', 'advisors',
        'dragon capital', 'vinacapital', 'templeton',
        'blackrock', 'jpmorgan', 'samsung', 'mirae',
    ]
    # Also check for non-Vietnamese characters as a heuristic
    is_foreign_name = name.apply(
        lambda x: any(kw in x for kw in foreign_keywords)
    )

    # Domestic institutions: Vietnamese bank, securities, insurance names
    domestic_inst_keywords = [
        'ngân hàng', 'chứng khoán', 'bảo hiểm', 'quỹ đầu tư',
        'công ty quản lý', 'bảo việt', 'techcombank', 'vietcombank',
        'bidv', 'vietinbank', 'vpbank', 'mb bank', 'ssi', 'hsc',
        'vcsc', 'vndirect', 'fpt capital', 'manulife',
    ]
    is_domestic_inst = name.apply(
        lambda x: any(kw in x for kw in domestic_inst_keywords)
    )

    # Treasury shares
    is_treasury = name.str.contains('cổ phiếu quỹ|treasury', case=False)

    # Apply classification cascade
    df['owner_type'] = OwnershipType.INDIVIDUAL  # Default
    df.loc[is_domestic_inst, 'owner_type'] = OwnershipType.DOMESTIC_INST
    df.loc[is_foreign_name, 'owner_type'] = OwnershipType.FOREIGN_INST
```

```
        df.loc[is_state, 'owner_type'] = OwnershipType.STATE
        df.loc[is_treasury, 'owner_type'] = OwnershipType.TREASURY

    # --- Method 3: Use shareholder_type directly ---
    elif 'shareholder_type' in df.columns:
        type_map = {
            'state': OwnershipType.STATE,
            'foreign_institution': OwnershipType.FOREIGN_INST,
            'foreign_individual': OwnershipType.FOREIGN_INST,
            'domestic_institution': OwnershipType.DOMESTIC_INST,
            'individual': OwnershipType.INDIVIDUAL,
            'treasury': OwnershipType.TREASURY,
        }
        df['owner_type'] = df['shareholder_type'].str.lower().map(type_map)
        df['owner_type'] = df['owner_type'].fillna(OwnershipType.INDIVIDUAL)

    else:
        raise ValueError(
            "Cannot classify shareholders. Expected one of:\n"
            "  1. Columns: is_state, is_foreign, is_institution\n"
            "  2. Column: shareholder_name (for heuristic classification)\n"
            "  3. Column: shareholder_type (pre-classified)"
        )

    # Summary
    print("Ownership classification results:")
    print(df['owner_type'].value_counts().to_string())

    return df

# ownership_classified = classify_shareholders(dc.ownership)
```

---

## 29.4 Vietnam's Ownership Taxonomy

### 29.4.1 The Five Ownership Categories

Vietnam's ownership structure is decomposed into five mutually exclusive categories that together sum to 100% of shares outstanding:

Table 29.2: Vietnam's Ownership Taxonomy

| Category | Vietnamese Term | Description | Typical Share (2020s) |
|---|---|---|---|
| **State** | Sở hữu Nhà nước | Government entities, SCIC, SOE parent companies | ~15-25% of market cap |
| **Foreign Institutional** | Tổ chức nước ngoài | Foreign funds, banks, corporations | ~15-20% |
| **Domestic Institutional** | Tổ chức trong nước | Vietnamese funds, banks, insurance, securities firms | ~5-10% |
| **Individual** | Cá nhân | Retail investors (both Vietnamese and foreign individuals) | ~55-65% |
| **Treasury** | Cổ phiếu quỹ | Company's own repurchased shares | ~0-2% |

This taxonomy differs fundamentally from the US 13F framework in several ways:

1. **Completeness:** We observe 100% of ownership, not just institutional long positions above $100 million AUM.
2. **State as a category:** State ownership is a first-class analytical category, not subsumed under "All Others" as in the LSEG type code system.
3. **Individual visibility:** We observe aggregate individual ownership directly, whereas in the US, individual ownership is merely the residual (100% − institutional ownership).
4. **No short position ambiguity:** Vietnam's market has very limited short-selling infrastructure, so ownership data genuinely represents long positions.

```
# ============================================================================
# Step 4: Compute Ownership Decomposition
# ============================================================================

def compute_ownership_decomposition(ownership: pd.DataFrame,
                                    prices_q: pd.DataFrame) -> pd.DataFrame:
    """
    Compute the full ownership decomposition for each stock at each
    disclosure date.

    For each stock-date combination, aggregates shares held by each
    ownership category and computes ownership ratios relative to
```

```
    total shares outstanding.

    Parameters
    ----------
    ownership : pd.DataFrame
        Classified ownership data (output of classify_shareholders)
    prices_q : pd.DataFrame
        Quarter-end price data with shares_outstanding

    Returns
    -------
    pd.DataFrame
        Stock-period level ownership decomposition with columns for
        each ownership type's share count and percentage
    """
    # Aggregate shares by ticker, date, and owner type
    agg = (ownership.groupby(['ticker', 'date', 'owner_type'])['shares_held']
                    .sum()
                    .reset_index())

    # Pivot to wide format: one column per ownership type
    wide = agg.pivot_table(
        index=['ticker', 'date'],
        columns='owner_type',
        values='shares_held',
        fill_value=0
    ).reset_index()

    # Rename columns
    type_cols = [c for c in wide.columns if c in OwnershipType.ALL_TYPES]
    rename_map = {t: f'shares_{t}' for t in type_cols}
    wide = wide.rename(columns=rename_map)

    # Total institutional shares
    inst_cols = [f'shares_{t}' for t in OwnershipType.INSTITUTIONAL
                 if f'shares_{t}' in wide.columns]
    wide['shares_institutional'] = wide[inst_cols].sum(axis=1)

    # Total foreign shares (for FOL tracking)
    foreign_cols = [f'shares_{t}' for t in OwnershipType.FOREIGN
                    if f'shares_{t}' in wide.columns]
    wide['shares_foreign_total'] = wide[foreign_cols].sum(axis=1)
```

```
# Align with quarter-end dates for merging with price data
wide['quarter_end'] = wide['date'] + pd.offsets.QuarterEnd(0)

# Merge with price data to get shares outstanding
merged = wide.merge(
    prices_q[['ticker', 'quarter_end', 'shares_outstanding',
              'adjusted_shares', 'market_cap', 'exchange',
              'industry_code', 'fol_limit', 'close']],
    on=['ticker', 'quarter_end'],
    how='left'
)

# Compute ownership ratios
tso = merged['shares_outstanding']
for col in merged.columns:
    if col.startswith('shares_') and col != 'shares_outstanding':
        ratio_col = col.replace('shares_', 'pct_')
        merged[ratio_col] = merged[col] / tso
        merged.loc[tso <= 0, ratio_col] = np.nan

# Derived measures
merged['pct_free_float'] = 1 - merged.get('pct_state', 0) - merged.get('pct_treasury', 0

# SOE flag: state ownership > 50%
merged['is_soe'] = (merged.get('pct_state', 0) > 0.50).astype(int)

# FOL utilization
if 'fol_limit' in merged.columns and 'pct_foreign_total' in merged.columns:
    merged['fol_utilization'] = merged['pct_foreign_total'] / merged['fol_limit']
    merged['foreign_room'] = merged['fol_limit'] - merged['pct_foreign_total']
    merged.loc[merged['fol_limit'] <= 0, ['fol_utilization', 'foreign_room']] = np.nan

# Number of institutional owners (breadth)
n_owners = (ownership[ownership['owner_type'].isin(OwnershipType.INSTITUTIONAL)]
            .groupby(['ticker', 'date'])['shareholder_name']
            .nunique()
            .reset_index()
            .rename(columns={'shareholder_name': 'n_inst_owners'}))

n_foreign_owners = (ownership[ownership['owner_type'] == OwnershipType.FOREIGN_INST]
                    .groupby(['ticker', 'date'])['shareholder_name']
                    .nunique()
```

```
                        .reset_index()
                        .rename(columns={'shareholder_name': 'n_foreign_owners'}))

    merged = merged.merge(n_owners, on=['ticker', 'date'], how='left')
    merged = merged.merge(n_foreign_owners, on=['ticker', 'date'], how='left')
    merged[['n_inst_owners', 'n_foreign_owners']] = (
        merged[['n_inst_owners', 'n_foreign_owners']].fillna(0)
    )

    print(f"Ownership decomposition computed:")
    print(f"  Stock-period observations: {len(merged):,}")
    print(f"  Unique tickers: {merged['ticker'].nunique():,}")
    print(f"\nMean ownership structure:")
    pct_cols = [c for c in merged.columns if c.startswith('pct_')]
    print(merged[pct_cols].mean().round(4).to_string())

    return merged

# ownership_decomp = compute_ownership_decomposition(
#     ownership_classified, prices_q
# )
```

---

## 29.5 Institutional Ownership Measures

### 29.5.1 Ownership Ratio

The **Institutional Ownership Ratio (IOR)** for stock $i$ at time $t$ in Vietnam is:

$$IOR_{i,t} = \frac{S_{i,t}^{state} + S_{i,t}^{foreign_inst} + S_{i,t}^{domestic_inst}}{TSO_{i,t}} \tag{29.1}$$

where $S_{i,t}^{type}$ denotes adjusted shares held by each ownership category and $TSO_{i,t}$ is total shares outstanding. Unlike the US where the IOR can exceed 100% due to long-only reporting and short selling, the Vietnamese IOR is bounded by construction in $[0, 1]$ because we observe the complete ownership decomposition.

We also compute category-specific ownership ratios:

$$
\begin{aligned}
IOR_{i,t}^{foreign} &= \frac{S_{i,t}^{foreign_inst}}{TSO_{i,t}}, \\
IOR_{i,t}^{state} &= \frac{S_{i,t}^{state}}{TSO_{i,t}}, \\
IOR_{i,t}^{domestic} &= \frac{S_{i,t}^{domestic_inst}}{TSO_{i,t}}
\end{aligned}
\tag{29.2}
$$

### 29.5.2 Concentration: Herfindahl-Hirschman Index

The **Institutional Ownership Concentration** via the Herfindahl-Hirschman Index is:

$$
IOC_{i,t}^{HHI} = \sum_{j=1}^{N_{i,t}} \left( \frac{S_{i,j,t}}{\sum_{k=1}^{N_{i,t}} S_{i,k,t}} \right)^2 \tag{29.3}
$$

In Vietnam, the HHI is particularly informative because it captures the dominance of state shareholders. A company where the government holds 65% will have a mechanically high HHI even if the remaining 35% is diversely held.

We therefore compute **separate HHI measures** for different ownership categories:

$$
HHI_{i,t}^{total} = \sum_{j} w_{i,j,t}^2, \quad HHI_{i,t}^{non-state} = \sum_{j \notin state} \left( \frac{S_{i,j,t}}{\sum_{k \notin state} S_{i,k,t}} \right)^2 \tag{29.4}
$$

The non-state HHI is more comparable to the US institutional HHI, as it captures concentration among market-driven investors.

### 29.5.3 Breadth of Ownership

Following J. Chen, Hong, and Stein (2002), **Institutional Breadth** ($N_{i,t}$) is the number of institutional investors holding stock $i$ in period $t$. The **Change in Breadth** is:

$$
\Delta Breadth_{i,t} = \frac{N_{i,t}^{cont} - N_{i,t-1}^{cont}}{TotalInstitutions_{t-1}} \tag{29.5}
$$

where $N_{i,t}^{cont}$ counts only institutions that appear in the disclosure universe in both periods $t$ and $t-1$, following the Lehavy and Sloan (2008) algorithm. This adjustment is particularly important in Vietnam where:

- New funds launch frequently (especially ETFs tracking VN30)
- Foreign funds enter and exit the market
- Domestic securities firms consolidate or spin off asset management divisions

```python
# ============================================================================
# Step 5: Compute All IO Metrics
# ============================================================================

def compute_io_metrics_vietnam(ownership: pd.DataFrame,
                               ownership_decomp: pd.DataFrame,
                               adj_factors: pd.DataFrame) -> pd.DataFrame:
    """
    Compute security-level institutional ownership metrics adapted for Vietnam.

    Computes:
    1. Ownership ratios by category (state, foreign, domestic inst, individual)
    2. HHI concentration (total, non-state, foreign-only)
    3. Number of institutional owners (total, foreign, domestic)
    4. Change in breadth (Lehavy-Sloan adjusted)
    5. FOL-related metrics (utilization, room, near-cap indicator)

    Parameters
    ----------
    ownership : pd.DataFrame
        Classified ownership data with individual shareholder records
    ownership_decomp : pd.DataFrame
        Aggregated ownership decomposition (output of compute_ownership_decomposition)
    adj_factors : pd.DataFrame
        Corporate action adjustment factors

    Returns
    -------
    pd.DataFrame
        Stock-period level metrics
    """
    # Start with the ownership decomposition
    metrics = ownership_decomp.copy()

    # --- HHI Concentration ---
    # Total HHI: across all institutional shareholders
    inst_ownership = ownership[
        ownership['owner_type'].isin(OwnershipType.INSTITUTIONAL)
```

```
].copy()

def compute_hhi_group(group):
    """Compute HHI for a group of shareholders."""
    total = group['shares_held'].sum()
    if total <= 0:
        return np.nan
    weights = group['shares_held'] / total
    return (weights ** 2).sum()

# Total institutional HHI
hhi_total = (inst_ownership.groupby(['ticker', 'date'])
                           .apply(compute_hhi_group)
                           .reset_index(name='hhi_institutional'))
metrics = metrics.merge(hhi_total, on=['ticker', 'date'], how='left')

# Non-state HHI (exclude state shareholders)
non_state = ownership[
    ownership['owner_type'].isin([OwnershipType.FOREIGN_INST,
                                  OwnershipType.DOMESTIC_INST])
]
hhi_nonstate = (non_state.groupby(['ticker', 'date'])
                         .apply(compute_hhi_group)
                         .reset_index(name='hhi_non_state'))
metrics = metrics.merge(hhi_nonstate, on=['ticker', 'date'], how='left')

# Foreign-only HHI
foreign_only = ownership[ownership['owner_type'] == OwnershipType.FOREIGN_INST]
hhi_foreign = (foreign_only.groupby(['ticker', 'date'])
                           .apply(compute_hhi_group)
                           .reset_index(name='hhi_foreign'))
metrics = metrics.merge(hhi_foreign, on=['ticker', 'date'], how='left')

# --- Change in Breadth (Lehavy-Sloan Algorithm) ---
metrics = metrics.sort_values(['ticker', 'date'])

# Get list of all institutions filing in each period
inst_by_period = (inst_ownership.groupby('date')['shareholder_name']
                                .apply(set)
                                .to_dict())

# For each stock-period: count continuing institutions
```

```python
    def compute_breadth_change(group):
        group = group.sort_values('date').reset_index(drop=True)
        group['dbreadth'] = np.nan

        for i in range(1, len(group)):
            current_date = group.loc[i, 'date']
            prev_date = group.loc[i-1, 'date']

            # Institutions in universe for both periods
            current_universe = inst_by_period.get(current_date, set())
            prev_universe = inst_by_period.get(prev_date, set())
            continuing_universe = current_universe & prev_universe

            if len(prev_universe) == 0:
                continue

            # Count continuing institutions holding this stock in each period
            ticker = group.loc[i, 'ticker']

            current_holders = set(
                inst_ownership[
                    (inst_ownership['ticker'] == ticker) &
                    (inst_ownership['date'] == current_date)
                ]['shareholder_name']
            )
            prev_holders = set(
                inst_ownership[
                    (inst_ownership['ticker'] == ticker) &
                    (inst_ownership['date'] == prev_date)
                ]['shareholder_name']
            )

            # Count only continuing institutions
            n_current_cont = len(current_holders & continuing_universe)
            n_prev_cont = len(prev_holders & continuing_universe)

            group.loc[i, 'dbreadth'] = (
                (n_current_cont - n_prev_cont) / len(prev_universe)
            )

        return group
```

```
    metrics = metrics.groupby('ticker', group_keys=False).apply(compute_breadth_change)

    # --- FOL Indicators ---
    if 'fol_utilization' in metrics.columns:
        metrics['near_fol_cap'] = (metrics['fol_utilization'] > 0.90).astype(int)
        metrics['at_fol_cap'] = (metrics['fol_utilization'] > 0.98).astype(int)

    print(f"IO metrics computed for Vietnam:")
    print(f"  Observations: {len(metrics):,}")
    print(f"\nKey metric distributions:")
    summary_cols = ['pct_institutional', 'pct_state', 'pct_foreign_total',
                    'hhi_institutional', 'n_inst_owners', 'dbreadth']
    summary_cols = [c for c in summary_cols if c in metrics.columns]
    print(metrics[summary_cols].describe().round(4).to_string())

    return metrics

# io_metrics = compute_io_metrics_vietnam(
#     ownership_classified, ownership_decomp, adj_factors
# )
```

### 29.5.4 Time Series Visualization

---

## 29.6 Foreign Ownership Dynamics

### 29.6.1 Foreign Ownership Limits and the FOL Premium

Vietnam's Foreign Ownership Limits create a unique market segmentation. When a stock reaches its FOL, the only way for a new foreign investor to buy is if an existing foreign holder sells. This creates a de facto "foreign-only" market for FOL-constrained stocks, with documented price premiums (Vo 2015).

The **FOL Utilization Ratio** for stock $i$ at time $t$ is:

$$FOL_Util_{i,t} = \frac{ForeignOwnership_{i,t}}{FOL_Limit_i} \tag{29.6}$$

Stocks are classified by FOL proximity (Table 29.4).

```python
def plot_ownership_timeseries_vietnam(metrics: pd.DataFrame):
    """
    Create publication-quality time series plots of Vietnamese
    ownership structure evolution.
    """
    fig, axes = plt.subplots(3, 1, figsize=(12, 14))

    # Aggregate across all stocks (market-cap weighted)
    ts = metrics.groupby('quarter_end').apply(
        lambda g: pd.Series({
            'pct_state': np.average(g['pct_state'].fillna(0),
                                    weights=g['market_cap'].fillna(1)),
            'pct_foreign': np.average(g['pct_foreign_total'].fillna(0),
                                      weights=g['market_cap'].fillna(1)),
            'pct_domestic_inst': np.average(g['pct_domestic_inst'].fillna(0),
                                            weights=g['market_cap'].fillna(1)),
            'pct_individual': np.average(g['pct_individual'].fillna(0),
                                         weights=g['market_cap'].fillna(1)),
            'n_stocks': g['ticker'].nunique(),
            'total_mktcap': g['market_cap'].sum(),
            'median_n_inst': g['n_inst_owners'].median(),
            'median_hhi': g['hhi_institutional'].median(),
            'pct_soe': g['is_soe'].mean(),
        })
    ).reset_index()

    # ---- Panel A: Ownership Composition (Stacked Area) ----
    ax = axes[0]
    dates = ts['quarter_end']
    ax.stackplot(dates,
                 ts['pct_state'] * 100,
                 ts['pct_foreign'] * 100,
                 ts['pct_domestic_inst'] * 100,
                 ts['pct_individual'] * 100,
                 labels=['State', 'Foreign Institutional',
                         'Domestic Institutional', 'Individual'],
                 colors=[OWNER_COLORS['State'], OWNER_COLORS['Foreign Institutional'],
                         OWNER_COLORS['Domestic Institutional'], OWNER_COLORS['Individual']]
                 alpha=0.8)
    ax.set_ylabel('Ownership Share (%)')
    ax.set_title('Panel A: Ownership Composition of Vietnamese Listed Companies '
                 '(Market-Cap Weighted)')
    ax.legend(loc='upper right', frameon=True, framealpha=0.9)
    ax.set_ylim(0, 100)

    # ---- Panel B: Institutional Ownership by Component ----
    ax = axes[1]
    ax.plot(dates, ts['pct_state'] * 100, label='State',
            color=OWNER_COLORS['State'], linewidth=2)
    ax.plot(dates, ts['pct_foreign'] * 100, label='Foreign Institutional',
            color=OWNER_COLORS['Foreign Institutional'], linewidth=2)
    ax.plot(dates, ts['pct_domestic_inst'] * 100, label='Domestic Institutional',
            color=OWNER_COLORS['Domestic Institutional'], linewidth=2)
```

```python
def plot_io_by_exchange_size(metrics: pd.DataFrame):
    """Plot IO ratios by exchange and size quintile."""
    df = metrics[metrics['market_cap'].notna() & (metrics['market_cap'] > 0)].copy()

    # Size quintiles within each quarter
    df['size_quintile'] = df.groupby('quarter_end')['market_cap'].transform(
        lambda x: pd.qcut(x, 5, labels=['Q1\n(Small)', 'Q2', 'Q3', 'Q4', 'Q5\n(Large)'],
                          duplicates='drop')
    )

    fig, axes = plt.subplots(1, 3, figsize=(15, 5), sharey=True)

    metrics_to_plot = [
        ('pct_institutional', 'Total Institutional'),
        ('pct_foreign_total', 'Foreign Institutional'),
        ('pct_state', 'State'),
    ]

    for ax, (col, title) in zip(axes, metrics_to_plot):
        for exchange, color in EXCHANGE_COLORS.items():
            data = df[df['exchange'] == exchange]
            if len(data) == 0:
                continue
            means = data.groupby('size_quintile')[col].mean() * 100
            ax.bar(np.arange(len(means)) + list(EXCHANGE_COLORS.keys()).index(exchange) * 0.2
                   means, width=0.25, label=exchange, color=color, alpha=0.8)

        ax.set_title(title)
        ax.set_xlabel('Size Quintile')
        if ax == axes[0]:
            ax.set_ylabel('Mean Ownership (%)')
        ax.legend()
        ax.set_xticks(np.arange(5) + 0.25)
        ax.set_xticklabels(['Q1\n(Small)', 'Q2', 'Q3', 'Q4', 'Q5\n(Large)'])

    plt.tight_layout()
    plt.savefig('fig_io_by_exchange_size.png', dpi=300, bbox_inches='tight')
    plt.show()

# plot_io_by_exchange_size(io_metrics)
```

Figure 29.2

Table 29.3: Summary Statistics of Ownership Structure in Vietnam by Size Quintile and Exchange (Pooled 2010-2024)

```
def tabulate_io_summary(metrics: pd.DataFrame, start_year: int = 2010) -> pd.DataFrame:
    """
    Create publication-quality summary table of Vietnamese ownership
    structure by firm size.
    """
    df = metrics[
        (metrics['quarter_end'].dt.year >= start_year) &
        (metrics['market_cap'].notna()) & (metrics['market_cap'] > 0)
    ].copy()

    df['size_quintile'] = df.groupby('quarter_end')['market_cap'].transform(
        lambda x: pd.qcut(x, 5, labels=['Q1 (Small)', 'Q2', 'Q3', 'Q4', 'Q5 (Large)'],
                          duplicates='drop')
    )

    table = df.groupby('size_quintile').agg(
        N=('ticker', 'count'),
        Mean_MktCap=('market_cap', 'mean'),
        Mean_IO_Total=('pct_institutional', 'mean'),
        Mean_State=('pct_state', 'mean'),
        Mean_Foreign=('pct_foreign_total', 'mean'),
        Mean_Domestic_Inst=('pct_domestic_inst', 'mean'),
        Mean_Individual=('pct_individual', 'mean'),
        Median_N_Owners=('n_inst_owners', 'median'),
        Median_HHI=('hhi_institutional', 'median'),
        Pct_SOE=('is_soe', 'mean'),
        Mean_FOL_Util=('fol_utilization', 'mean'),
    ).round(4)

    # Format
    table['N'] = table['N'].apply(lambda x: f"{x:,.0f}")
    table['Mean_MktCap'] = table['Mean_MktCap'].apply(lambda x: f"{x:,.0f}B VND")
    for col in ['Mean_IO_Total', 'Mean_State', 'Mean_Foreign',
                'Mean_Domestic_Inst', 'Mean_Individual', 'Pct_SOE', 'Mean_FOL_Util']:
        table[col] = table[col].apply(lambda x: f"{x:.1%}" if pd.notna(x) else "-")
    table['Median_N_Owners'] = table['Median_N_Owners'].apply(lambda x: f"{x:.0f}")
    table['Median_HHI'] = table['Median_HHI'].apply(lambda x: f"{x:.3f}" if pd.notna(x) else

    table.columns = ['N', 'Mean Mkt Cap', 'IO Total', 'State', 'Foreign',
                     'Dom. Inst.', 'Individual', 'Med. # Owners',
                     'Med. HHI', '% SOE', 'FOL Util.']

    return table

# io_summary = tabulate_io_summary(io_metrics)
# print(io_summary.to_string())
```

Table 29.4: FOL Proximity Zones

| FOL Zone | Utilization Range | Market Implication |
|---|---|---|
| **Green** | < 50% | Ample foreign room; normal trading |
| **Yellow** | 50-80% | Moderate room; some foreign interest pressure |
| **Orange** | 80-95% | Limited room; foreign premium emerging |
| **Red** | 95-100% | Near cap; significant foreign premium |
| **Capped** | 100% | At limit; foreign-only secondary market |

```python
# ============================================================================
# Step 6: Foreign Ownership Limit Analysis
# ============================================================================

class FOLAnalyzer:
    """
    Analyze Foreign Ownership Limit dynamics in the Vietnamese market.

    Key analyses:
    1. FOL utilization tracking and classification
    2. FOL premium estimation (price impact of being near cap)
    3. Foreign room dynamics (opening/closing events)
    4. Cross-sectional determinants of foreign ownership
    """

    FOL_ZONES = {
        'Green': (0, 0.50),
        'Yellow': (0.50, 0.80),
        'Orange': (0.80, 0.95),
        'Red': (0.95, 1.00),
        'Capped': (1.00, 1.50),
    }

    def __init__(self, io_metrics: pd.DataFrame,
                 foreign_daily: Optional[pd.DataFrame] = None):
        """
        Parameters
        ----------
        io_metrics : pd.DataFrame
            Full ownership metrics from compute_io_metrics_vietnam()
        foreign_daily : pd.DataFrame, optional
```

```
        Daily foreign ownership tracking from DataCore.vn
    """
    self.metrics = io_metrics.copy()
    self.foreign_daily = foreign_daily

def classify_fol_zones(self) -> pd.DataFrame:
    """Classify stocks into FOL proximity zones."""
    df = self.metrics.copy()

    if 'fol_utilization' not in df.columns:
        print("FOL utilization not available in metrics.")
        return df

    conditions = []
    choices = []
    for zone, (lo, hi) in self.FOL_ZONES.items():
        conditions.append(
            (df['fol_utilization'] >= lo) & (df['fol_utilization'] < hi)
        )
        choices.append(zone)

    df['fol_zone'] = np.select(conditions, choices, default='Unknown')

    # Summary
    zone_dist = df.groupby('fol_zone')['ticker'].nunique()
    print("FOL Zone Distribution (unique stocks):")
    print(zone_dist.to_string())

    return df

def estimate_fol_premium(self) -> pd.DataFrame:
    """
    Estimate the FOL premium using a cross-sectional approach.

    For each period, regress stock valuations (P/B or P/E) on FOL
    utilization, controlling for fundamentals. The coefficient on
    FOL utilization captures the premium investors pay for stocks
    near their foreign ownership cap.

    Alternative: Compare returns of stocks transitioning between
    FOL zones as a natural experiment.
    """
```

```
        df = self.metrics.copy()
        df = df[df['fol_utilization'].notna() & df['market_cap'].notna()].copy()

        # FOL zone dummies
        df['near_cap'] = (df['fol_utilization'] > 0.90).astype(int)
        df['at_cap'] = (df['fol_utilization'] > 0.98).astype(int)

        # Price-to-book as valuation measure
        # (Assumes 'equity' is available from financial data)
        if 'equity' in df.columns:
            df['pb_ratio'] = df['market_cap'] * 1e9 / df['equity']
        else:
            # Use market cap as proxy for cross-sectional analysis
            df['log_mktcap'] = np.log(df['market_cap'])

        # Fama-MacBeth style: run cross-sectional regressions each period
        results = []
        for quarter, group in df.groupby('quarter_end'):
            group = group.dropna(subset=['fol_utilization', 'log_mktcap'])
            if len(group) < 50:
                continue

            y = group['log_mktcap']
            X = sm.add_constant(group[['fol_utilization', 'pct_state',
                                       'n_inst_owners']])
            try:
                model = sm.OLS(y, X).fit()
                results.append({
                    'quarter': quarter,
                    'beta_fol': model.params.get('fol_utilization', np.nan),
                    'tstat_fol': model.tvalues.get('fol_utilization', np.nan),
                    'r2': model.rsquared,
                    'n': len(group),
                })
            except Exception:
                continue

        if results:
            results_df = pd.DataFrame(results)
            print("FOL Premium (Fama-MacBeth Regression):")
            print(f"  Mean  (FOL_util): {results_df['beta_fol'].mean():.4f}")
            print(f"  t-statistic: {results_df['beta_fol'].mean() / "
```

```
                f"(results_df['beta_fol'].std() / np.sqrt(len(results_df))):.2f}")
        return results_df

    return pd.DataFrame()

def analyze_foreign_room_events(self) -> pd.DataFrame:
    """
    Analyze events where foreign room opens or closes.

    Room-opening events (FOL cap raised, foreign seller exits) can
    trigger significant price movements as pent-up foreign demand
    is released. Room-closing events (approaching cap) can create
    selling pressure as foreign investors anticipate illiquidity.
    """
    if self.foreign_daily is None:
        print("Daily foreign ownership data required for event analysis.")
        return pd.DataFrame()

    df = self.foreign_daily.copy()
    df = df.sort_values(['ticker', 'date'])

    # Compute daily change in foreign room
    df['foreign_room_change'] = df.groupby('ticker')['foreign_room'].diff()

    # Identify room-opening events (room increases by > 1 percentage point)
    df['room_open_event'] = (df['foreign_room_change'] > 0.01).astype(int)

    # Identify room-closing events (room decreases to < 2%)
    df['room_close_event'] = (
        (df['foreign_room'] < 0.02) &
        (df.groupby('ticker')['foreign_room'].shift(1) >= 0.02)
    ).astype(int)

    events = df[
        (df['room_open_event'] == 1) | (df['room_close_event'] == 1)
    ].copy()

    print(f"Foreign room events identified:")
    print(f"  Room-opening events: {df['room_open_event'].sum():,}")
    print(f"  Room-closing events: {df['room_close_event'].sum():,}")

    return events
```

```
# fol_analyzer = FOLAnalyzer(io_metrics, dc.foreign_ownership)
# fol_classified = fol_analyzer.classify_fol_zones()
# fol_premium = fol_analyzer.estimate_fol_premium()
```

## 29.7 Institutional Trades

### 29.7.1 Trade Inference in Vietnam

In the US, institutional trades are inferred from quarterly 13F holding snapshots. In Vietnam, the challenge is more acute because disclosure frequency varies:

- **Major shareholders ($\geq$ 5%)**: Must disclose within 7 business days of crossing ownership thresholds (5%, 10%, 15%, 20%, 25%, 50%, 65%, 75%)
- **Fund portfolio reports:** Semi-annual disclosure required; some funds report quarterly
- **Annual reports:** Provide complete shareholder register but only once per year
- **Daily foreign ownership:** HOSE/HNX publish aggregate daily foreign buy/sell data

We derive trades from the **change in ownership between consecutive disclosure dates**, applying the same logic as the US Ben-David et al. (2013) algorithm but adapted for Vietnam's irregular disclosure intervals.

```
# ============================================================================
# Step 7: Derive Institutional Trades
# ============================================================================

def derive_trades_vietnam(ownership: pd.DataFrame,
                          adj_factors: pd.DataFrame) -> pd.DataFrame:
    """
    Derive institutional trades from changes in ownership disclosures.

    Adapted from Ben-David, Franzoni, and Moussawi (2012) for
    Vietnam's irregular disclosure frequency.

    Key differences from US approach:
    1. Disclosure intervals are irregular (not always quarterly)
    2. We observe ALL institutional types, not just 13F filers
    3. No $100M AUM threshold (we see all institutional holders)
    4. Must adjust for corporate actions between disclosure dates

    Trade types:
```

```python
def plot_fol_utilization(metrics: pd.DataFrame):
    """Plot FOL utilization distribution by sector."""
    df = metrics[metrics['fol_utilization'].notna()].copy()

    # Assign broad sectors
    sector_map = {
        'Banking': ['VCB', 'BID', 'CTG', 'TCB', 'VPB', 'MBB', 'ACB', 'HDB', 'STB', 'TPB'],
        'Real Estate': ['VHM', 'VIC', 'NVL', 'KDH', 'DXG', 'HDG', 'VRE'],
        'Technology': ['FPT', 'CMG', 'FOX'],
        'Consumer': ['VNM', 'MSN', 'SAB', 'MWG', 'PNJ'],
    }

    fig, ax = plt.subplots(figsize=(10, 6))

    for sector, tickers in sector_map.items():
        data = df[df['ticker'].isin(tickers)]['fol_utilization']
        if len(data) > 0:
            ax.hist(data * 100, bins=30, alpha=0.4, label=sector, density=True)

    ax.axvline(x=30, color='red', linestyle='--', alpha=0.7, label='Banking FOL (30%)')
    ax.axvline(x=49, color='blue', linestyle='--', alpha=0.7, label='Standard FOL (49%)')
    ax.set_xlabel('FOL Utilization (%)')
    ax.set_ylabel('Density')
    ax.set_title('Foreign Ownership Limit Utilization Distribution')
    ax.legend()

    plt.tight_layout()
    plt.savefig('fig_fol_utilization.png', dpi=300, bbox_inches='tight')
    plt.show()

# plot_fol_utilization(io_metrics)
```

Figure 29.3

```
    +1: Initiating Buy (new position)
    +2: Incremental Buy (increased existing position)
    -1: Terminating Sale (fully exited position)
    -2: Incremental Sale (reduced existing position)

    Parameters
    ----------
    ownership : pd.DataFrame
        Classified ownership with: ticker, date, shareholder_name,
        shares_held, owner_type
    adj_factors : pd.DataFrame
        Corporate action adjustment factors

    Returns
    -------
    pd.DataFrame
        Trade-level data: date, shareholder_name, ticker, trade,
        buysale, owner_type
    """
    # Focus on institutional shareholders only
    inst = ownership[
        ownership['owner_type'].isin(OwnershipType.INSTITUTIONAL)
    ].copy()

    inst = inst.sort_values(['shareholder_name', 'ticker', 'date']).reset_index(drop=True)

    trades_list = []

    for (shareholder, ticker), group in inst.groupby(['shareholder_name', 'ticker']):
        group = group.reset_index(drop=True)

        for i in range(len(group)):
            current = group.iloc[i]
            current_date = current['date']
            current_shares = current['shares_held']
            owner_type = current['owner_type']

            if i == 0:
                # First observation: if institution appears, it's an initiating buy
                # (we don't know if they held before our data starts)
                # Skip the very first observation to avoid false initiating buys
                continue
```

```
        prev = group.iloc[i - 1]
        prev_date = prev['date']
        prev_shares = prev['shares_held']

        # Adjust previous shares for corporate actions between dates
        prev_shares_adj = adjust_shares(
            prev_shares, ticker, prev_date, current_date, adj_factors
        )

        # Compute trade (in adjusted shares)
        trade = current_shares - prev_shares_adj

        # Classify trade type
        if abs(trade) < 1:  # De minimis threshold
            continue

        if prev_shares_adj <= 0 and current_shares > 0:
            buysale = 1  # Initiating buy
        elif prev_shares_adj > 0 and current_shares <= 0:
            buysale = -1  # Terminating sale
        elif trade > 0:
            buysale = 2  # Incremental buy
        else:
            buysale = -2  # Incremental sale

        trades_list.append({
            'date': current_date,
            'shareholder_name': shareholder,
            'ticker': ticker,
            'trade': trade,
            'prev_shares_adj': prev_shares_adj,
            'current_shares': current_shares,
            'buysale': buysale,
            'owner_type': owner_type,
            'days_between': (current_date - prev_date).days,
        })

trades = pd.DataFrame(trades_list)

if len(trades) > 0:
    print(f"Trades derived: {len(trades):,}")
    print(f"\nTrade type distribution:")
```

```
        labels = {1: 'Initiating Buy', 2: 'Incremental Buy',
                  -1: 'Terminating Sale', -2: 'Incremental Sale'}
        for bs, label in sorted(labels.items()):
            n = (trades['buysale'] == bs).sum()
            print(f"  {label}: {n:,} ({n/len(trades):.1%})")

        print(f"\nBy owner type:")
        print(trades.groupby('owner_type')['trade'].agg(['count', 'mean', 'median'])
              .round(0).to_string())

    return trades

# trades = derive_trades_vietnam(ownership_classified, adj_factors)
```

**Corporate Action Adjustment in Trade Derivation**

When computing trades as $\Delta Shares = Shares_t - Shares_{t-1}$, the previous period's shares **must** be adjusted for any corporate actions between $t-1$ and $t$. If VNM issued a 20% stock dividend between the two disclosure dates, then 1,000 shares at $t-1$ should be compared to 1,200 adjusted shares, not 1,000 raw shares. Failing to make this adjustment would create a phantom "buy" of 200 shares that never actually occurred.

```
def derive_trades_vectorized_vietnam(ownership: pd.DataFrame,
                                     adj_factors: pd.DataFrame) -> pd.DataFrame:
    """
    Vectorized version of Vietnamese trade derivation.

    Uses pandas groupby and vectorized operations instead of Python loops.
    Approximately 20-50x faster for large datasets.

    Note: Corporate action adjustment is applied per-group, which still
    requires some iteration but is much faster than row-by-row.
    """
    inst = ownership[
        ownership['owner_type'].isin(OwnershipType.INSTITUTIONAL) &
        (ownership['shares_held'] > 0)
    ].copy()

    inst = inst.sort_values(['shareholder_name', 'ticker', 'date']).reset_index(drop=True)

    # Lagged values
```

```
inst['prev_date'] = inst.groupby(['shareholder_name', 'ticker'])['date'].shift(1)
inst['prev_shares'] = inst.groupby(['shareholder_name', 'ticker'])['shares_held'].shift(1
inst['is_first'] = inst['prev_date'].isna()

# Remove first observations (no prior to compare)
inst = inst[~inst['is_first']].copy()

# Adjust previous shares for corporate actions
# Vectorized: for each row, apply adjustment between prev_date and date
def adjust_row(row):
    return adjust_shares(
        row['prev_shares'], row['ticker'],
        row['prev_date'], row['date'], adj_factors
    )

inst['prev_shares_adj'] = inst.apply(adjust_row, axis=1)

# Compute trade
inst['trade'] = inst['shares_held'] - inst['prev_shares_adj']
inst['days_between'] = (inst['date'] - inst['prev_date']).dt.days

# Classify trade type
inst['buysale'] = np.select(
    [
        (inst['prev_shares_adj'] <= 0) & (inst['shares_held'] > 0),
        (inst['prev_shares_adj'] > 0) & (inst['shares_held'] <= 0),
        inst['trade'] > 0,
        inst['trade'] < 0,
    ],
    [1, -1, 2, -2],
    default=0
)

# Remove zero trades
trades = inst[inst['buysale'] != 0].copy()

trades = trades[['date', 'shareholder_name', 'ticker', 'trade',
                 'buysale', 'owner_type', 'days_between',
                 'prev_shares_adj', 'shares_held']].copy()
trades = trades.rename(columns={'shares_held': 'current_shares'})

print(f"Vectorized trades: {len(trades):,}")
```

```
    return trades

# trades = derive_trades_vectorized_vietnam(ownership_classified, adj_factors)
```

## 29.8 Fund-Level Flows and Turnover

### 29.8.1 Portfolio Assets and Returns from Fund Holdings

Using DataCore.vn's fund holdings data, we compute fund-level portfolio analytics analogous to the US 13F approach:

$$Assets_{j,t} = \sum_{i=1}^{N_{j,t}} S_{i,j,t} \times P_{i,t} \tag{29.7}$$

$$R^{holdings}_{j,t\to t+1} = \frac{\sum_i S_{i,j,t} \times P_{i,t} \times R_{i,t\to t+1}}{\sum_i S_{i,j,t} \times P_{i,t}} \tag{29.8}$$

$$NetFlows_{j,t} = Assets_{j,t} - Assets_{j,t-1} \times (1 + R^{holdings}_{j,t-1\to t}) \tag{29.9}$$

### 29.8.2 Turnover Measures

Following Mark M. Carhart (1997a), adapted for Vietnam's fund reporting:

$$Turnover^{Carhart}_{j,t} = \frac{\min(TotalBuys_{j,t}, TotalSales_{j,t})}{\overline{Assets}_{j,t}} \tag{29.10}$$

```
# ============================================================================
# Step 8: Fund-Level Portfolio Analytics
# ============================================================================

def compute_fund_analytics(fund_holdings: pd.DataFrame,
                           prices_q: pd.DataFrame,
                           adj_factors: pd.DataFrame) -> Dict:
    """
    Compute fund-level portfolio analytics from DataCore.vn fund holdings.

    Vietnamese fund disclosure is typically semi-annual (some quarterly),
    which limits the frequency of these analytics compared to the US
```

```
    quarterly approach.

    Returns
    -------
    dict with keys:
        'fund_assets': pd.DataFrame of fund-level assets and returns
        'fund_trades': pd.DataFrame of fund-level derived trades
        'fund_aggregates': pd.DataFrame of flows and turnover
    """
    fh = fund_holdings.copy()
    fh = fh[fh['shares_held'] > 0].copy()

    # Merge with prices
    fh = fh.merge(
        prices_q[['ticker', 'quarter_end', 'close', 'adjusted_close', 'ret']],
        left_on=['ticker', 'report_date'],
        right_on=['ticker', 'quarter_end'],
        how='inner'
    )

    # Portfolio value
    fh['holding_value'] = fh['shares_held'] * fh['close']

    # --- Fund-Level Assets ---
    fund_assets = fh.groupby(['fund_name', 'report_date']).agg(
        total_assets=('holding_value', lambda x: x.sum() / 1e9),  # Billion VND
        n_stocks=('ticker', 'nunique'),
    ).reset_index()

    # Holdings return (value-weighted)
    fh['weight'] = fh.groupby(['fund_name', 'report_date'])['holding_value'].transform(
        lambda x: x / x.sum()
    )
    fund_hret = (fh.groupby(['fund_name', 'report_date'])
                   .apply(lambda g: np.average(g['ret'].fillna(0), weights=g['weight']))
                   .reset_index(name='holdings_return'))

    fund_assets = fund_assets.merge(fund_hret, on=['fund_name', 'report_date'])

    # --- Fund-Level Trades ---
    # Derive trades from changes in holdings
    fh_sorted = fh.sort_values(['fund_name', 'ticker', 'report_date'])
```

```
fh_sorted['prev_shares'] = fh_sorted.groupby(['fund_name', 'ticker'])['shares_held'].shi
fh_sorted['prev_date'] = fh_sorted.groupby(['fund_name', 'ticker'])['report_date'].shift

# Adjust for corporate actions
fh_sorted['prev_shares_adj'] = fh_sorted.apply(
    lambda r: adjust_shares(r['prev_shares'], r['ticker'],
                            r['prev_date'], r['report_date'], adj_factors)
    if pd.notna(r['prev_shares']) else np.nan,
    axis=1
)

fh_sorted['trade'] = fh_sorted['shares_held'] - fh_sorted['prev_shares_adj']
fh_sorted['trade_value'] = fh_sorted['trade'] * fh_sorted['close'] / 1e9  # Billion VND

# Aggregate buys and sells per fund-period
fund_trades = fh_sorted[fh_sorted['trade'].notna()].copy()
fund_flows = fund_trades.groupby(['fund_name', 'report_date']).agg(
    total_buys=('trade_value', lambda x: x[x > 0].sum()),
    total_sales=('trade_value', lambda x: -x[x < 0].sum()),
).reset_index()

# --- Fund-Level Aggregates ---
fund_agg = fund_assets.merge(fund_flows, on=['fund_name', 'report_date'], how='left')
fund_agg[['total_buys', 'total_sales']] = fund_agg[['total_buys', 'total_sales']].fillna

fund_agg = fund_agg.sort_values(['fund_name', 'report_date'])
fund_agg['lag_assets'] = fund_agg.groupby('fund_name')['total_assets'].shift(1)
fund_agg['lag_hret'] = fund_agg.groupby('fund_name')['holdings_return'].shift(1)

# Net flows
fund_agg['net_flows'] = (fund_agg['total_assets'] -
                         fund_agg['lag_assets'] * (1 + fund_agg['holdings_return']))

# Turnover (Carhart definition)
fund_agg['avg_assets'] = (fund_agg['total_assets'] + fund_agg['lag_assets']) / 2
fund_agg['turnover'] = (
    fund_agg[['total_buys', 'total_sales']].min(axis=1) / fund_agg['avg_assets']
)

# Annualize (approximate, since disclosure may be semi-annual)
fund_agg['periods_per_year'] = 365 / fund_agg.groupby('fund_name')['report_date'].diff()
fund_agg['turnover_annual'] = fund_agg['turnover'] * fund_agg['periods_per_year'].fillna
```

```
    print(f"Fund analytics computed:")
    print(f"  Unique funds: {fund_agg['fund_name'].nunique():,}")
    print(f"  Fund-period observations: {len(fund_agg):,}")
    print(f"\nTurnover statistics:")
    print(fund_agg[['turnover', 'turnover_annual']].describe().round(4))

    return {
        'fund_assets': fund_assets,
        'fund_trades': fund_trades,
        'fund_aggregates': fund_agg,
    }

# fund_analytics = compute_fund_analytics(dc.fund_holdings, prices_q, adj_factors)
```

---

## 29.9 State Ownership Analysis

### 29.9.1 Equitization and the Decline of State Ownership

Vietnam's equitization (cổ phần hóa) program has been a defining feature of the market since the early 2000s. The program converts state-owned enterprises into joint-stock companies, typically with the state retaining a controlling or significant minority stake that is then gradually reduced through secondary offerings.

```
# ============================================================================
# Step 9: State Ownership Analysis
# ============================================================================

def analyze_state_ownership(metrics: pd.DataFrame) -> Dict:
    """
    Comprehensive analysis of state ownership in Vietnam.

    Computes:
    1. Aggregate state ownership trends
    2. SOE population dynamics (entry/exit from SOE classification)
    3. Equitization event detection (large drops in state ownership)
    4. State ownership by sector and size
    5. Governance implications (state as blockholder)
    """
```

```python
    df = metrics.copy()

    # --- 1. Aggregate Trends ---
    ts = df.groupby('quarter_end').agg(
        n_soe=('is_soe', 'sum'),
        n_total=('ticker', 'nunique'),
        pct_soe=('is_soe', 'mean'),
        mean_state_pct=('pct_state', 'mean'),
        median_state_pct=('pct_state', 'median'),
        # Market cap share of SOEs
        soe_mktcap=('market_cap', lambda x: x[df.loc[x.index, 'is_soe'] == 1].sum()),
        total_mktcap=('market_cap', 'sum'),
    ).reset_index()
    ts['soe_mktcap_share'] = ts['soe_mktcap'] / ts['total_mktcap']

    # --- 2. Equitization Events ---
    # Detect large drops in state ownership (>10 percentage points)
    df_sorted = df.sort_values(['ticker', 'quarter_end'])
    df_sorted['state_change'] = df_sorted.groupby('ticker')['pct_state'].diff()

    equitization_events = df_sorted[
        df_sorted['state_change'] < -0.10  # > 10pp drop
    ][['ticker', 'quarter_end', 'pct_state', 'state_change', 'market_cap']].copy()

    # --- 3. By Sector ---
    if 'industry_code' in df.columns:
        by_sector = df.groupby('industry_code').agg(
            mean_state=('pct_state', 'mean'),
            pct_soe=('is_soe', 'mean'),
            n_firms=('ticker', 'nunique'),
        ).sort_values('mean_state', ascending=False)
    else:
        by_sector = None

    print(f"State Ownership Analysis:")
    print(f"  Current SOE count: {ts.iloc[-1]['n_soe']:.0f} / {ts.iloc[-1]['n_total']:.0f}")
    print(f"  SOE market cap share: {ts.iloc[-1]['soe_mktcap_share']:.1%}")
    print(f"  Mean state ownership: {ts.iloc[-1]['mean_state_pct']:.1%}")
    print(f"\nEquitization events detected: {len(equitization_events):,}")

    return {
        'trends': ts,
```

```
        'equitization_events': equitization_events,
        'by_sector': by_sector,
    }

# state_analysis = analyze_state_ownership(io_metrics)
```

---

## 29.10 Modern Extensions

### 29.10.1 Network Analysis of Co-Ownership

Institutional co-ownership networks capture how stocks are connected through shared investors. In Vietnam, these networks reveal the influence structure of major domestic conglomerates (e.g., Vingroup, Masan, FPT) and the overlap between foreign fund portfolios.

```
def construct_stock_coownership_network(ownership: pd.DataFrame,
                                        period: str,
                                        min_overlap: int = 3) -> Dict:
    """
    Construct a stock-level co-ownership network.

    Two stocks are connected if they share institutional investors.
    Edge weight = number of shared institutional investors.

    This is particularly informative in Vietnam where:
    - Foreign fund portfolios concentrate on the same blue-chips
    - Conglomerate cross-holdings create explicit linkages
    - State ownership creates implicit connections (SCIC holds multiple stocks)

    Parameters
    ----------
    ownership : pd.DataFrame
        Classified ownership data
    period : str
        Analysis date
    min_overlap : int
        Minimum shared investors to create an edge
```

```python
def plot_state_ownership(state_analysis: Dict, metrics: pd.DataFrame):
    """Plot state ownership dynamics."""
    fig, axes = plt.subplots(2, 1, figsize=(12, 10))
    ts = state_analysis['trends']

    # Panel A: SOE trends
    ax = axes[0]
    ax.plot(ts['quarter_end'], ts['pct_soe'] * 100,
            label='% of Firms that are SOEs', linewidth=2, color='#d62728')
    ax.plot(ts['quarter_end'], ts['soe_mktcap_share'] * 100,
            label='SOE Market Cap Share (%)', linewidth=2, color='#1f77b4')
    ax.plot(ts['quarter_end'], ts['mean_state_pct'] * 100,
            label='Mean State Ownership (%)', linewidth=2, color='#2ca02c', linestyle='--')
    ax.set_ylabel('Percentage')
    ax.set_title('Panel A: State Ownership and SOE Prevalence Over Time')
    ax.legend(frameon=True, framealpha=0.9)

    # Panel B: Distribution
    ax = axes[1]
    # Use most recent period
    latest = metrics[metrics['quarter_end'] == metrics['quarter_end'].max()]
    state_pct = latest['pct_state'].dropna() * 100

    ax.hist(state_pct, bins=50, color='#d62728', alpha=0.7, edgecolor='black')
    ax.axvline(x=50, color='black', linestyle='--', alpha=0.7, label='50% (SOE threshold)')
    ax.set_xlabel('State Ownership (%)')
    ax.set_ylabel('Number of Companies')
    ax.set_title('Panel B: Distribution of State Ownership (Most Recent Quarter)')
    ax.legend()

    plt.tight_layout()
    plt.savefig('fig_state_ownership.png', dpi=300, bbox_inches='tight')
    plt.show()

# plot_state_ownership(state_analysis, io_metrics)
```

Figure 29.4

```
    Returns
    -------
    dict with network statistics and adjacency data
    """
    import networkx as nx

    date = pd.Timestamp(period)

    # Get institutional holders for this period
    inst = ownership[
        (ownership['date'] == date) &
        (ownership['owner_type'].isin(OwnershipType.INSTITUTIONAL))
    ][['ticker', 'shareholder_name', 'owner_type']].copy()

    # Create bipartite mapping: institution → set of stocks held
    inst_to_stocks = inst.groupby('shareholder_name')['ticker'].apply(set).to_dict()

    # Stock → set of institutions
    stock_to_inst = inst.groupby('ticker')['shareholder_name'].apply(set).to_dict()

    # Build stock-level network
    stocks = list(stock_to_inst.keys())
    G = nx.Graph()

    for i in range(len(stocks)):
        for j in range(i + 1, len(stocks)):
            shared = stock_to_inst[stocks[i]] & stock_to_inst[stocks[j]]
            if len(shared) >= min_overlap:
                G.add_edge(stocks[i], stocks[j], weight=len(shared),
                           shared_investors=list(shared)[:5])  # Store sample

    # Add node attributes
    for stock in stocks:
        if stock in G.nodes:
            G.nodes[stock]['n_inst_holders'] = len(stock_to_inst[stock])

    # Network statistics
    stats = {
        'n_nodes': G.number_of_nodes(),
        'n_edges': G.number_of_edges(),
        'density': nx.density(G) if G.number_of_nodes() > 1 else 0,
        'avg_clustering': nx.average_clustering(G, weight='weight') if G.number_of_nodes() >
```

```
        'n_components': nx.number_connected_components(G),
    }

    # Centrality measures
    if G.number_of_nodes() > 0:
        degree_cent = nx.degree_centrality(G)
        stats['most_connected'] = sorted(degree_cent.items(),
                                         key=lambda x: x[1], reverse=True)[:10]

        if G.number_of_nodes() > 2:
            try:
                eigen_cent = nx.eigenvector_centrality_numpy(G, weight='weight')
                stats['most_central'] = sorted(eigen_cent.items(),
                                               key=lambda x: x[1], reverse=True)[:10]
            except Exception:
                stats['most_central'] = []

    print(f"Co-Ownership Network ({period}):")
    for k, v in stats.items():
        if k not in ['most_connected', 'most_central']:
            print(f"  {k}: {v}")

    if 'most_connected' in stats:
        print(f"\nMost connected stocks:")
        for stock, cent in stats['most_connected'][:5]:
            print(f"  {stock}: {cent:.3f}")

    return {'graph': G, 'stats': stats}

# network = construct_stock_coownership_network(
#     ownership_classified, '2024-06-30'
# )
```

### 29.10.2 ML-Enhanced Investor Classification

Vietnam's investor classification challenge is distinct from the US. While the US has the Bushee typology based on portfolio turnover and concentration, Vietnam requires classification of both investor **type** (when not explicitly labeled) and investor **behavior** (active vs passive, short-term vs long-term).

```python
def classify_investors_vietnam(ownership: pd.DataFrame,
                               prices_q: pd.DataFrame,
                               n_clusters: int = 4) -> pd.DataFrame:
    """
    ML-based classification of Vietnamese institutional investors.

    Features adapted for Vietnam's market:
    1. Portfolio concentration (HHI of holdings)
    2. Holding duration (average time in positions)
    3. Size preference (average market cap of holdings)
    4. Sector concentration
    5. Foreign/domestic indicator
    6. Trading frequency (inverse of average days between disclosures)

    Expected clusters for Vietnam:
    - Passive State Holders: SOE parents, SCIC - low turnover, concentrated
    - Active Foreign Funds: Dragon Capital, VinaCapital - moderate turnover
    - Domestic Securities Firms: SSI, VNDirect - high turnover, diversified
    - Long-Term Foreign: Pension funds, sovereign wealth - low turnover
    """
    from sklearn.cluster import KMeans
    from sklearn.preprocessing import StandardScaler

    inst = ownership[
        ownership['owner_type'].isin(OwnershipType.INSTITUTIONAL)
    ].copy()

    # Merge with price data
    inst = inst.merge(
        prices_q[['ticker', 'quarter_end', 'close', 'market_cap']],
        left_on=['ticker', 'date'],
        right_on=['ticker', 'quarter_end'],
        how='left'
    )

    inst['holding_value'] = inst['shares_held'] * inst['close'].fillna(0)

    # Compute features per investor-period
    features = inst.groupby(['shareholder_name', 'date']).agg(
        n_stocks=('ticker', 'nunique'),
        total_value=('holding_value', 'sum'),
        hhi_portfolio=('holding_value',
```

```
                        lambda x: ((x/x.sum())**2).sum() if x.sum() > 0 else np.nan),
    avg_mktcap=('market_cap', 'mean'),
    is_foreign=('owner_type',
                lambda x: (x == OwnershipType.FOREIGN_INST).any().astype(int)),
    is_state=('owner_type',
              lambda x: (x == OwnershipType.STATE).any().astype(int)),
).reset_index()

# Average across all periods per investor
investor_features = features.groupby('shareholder_name').agg(
    avg_n_stocks=('n_stocks', 'mean'),
    avg_hhi=('hhi_portfolio', 'mean'),
    avg_mktcap=('avg_mktcap', 'mean'),
    avg_total_value=('total_value', 'mean'),
    is_foreign=('is_foreign', 'max'),
    is_state=('is_state', 'max'),
    n_periods=('date', 'nunique'),
).dropna()

# Feature matrix
feature_cols = ['avg_n_stocks', 'avg_hhi', 'avg_mktcap', 'avg_total_value']
X = investor_features[feature_cols].copy()

# Log-transform
for col in feature_cols:
    X[col] = np.log1p(X[col].clip(lower=0))

# Add binary features
X['is_foreign'] = investor_features['is_foreign']
X['is_state'] = investor_features['is_state']

# Standardize
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# K-means
kmeans = KMeans(n_clusters=n_clusters, random_state=42, n_init=20)
investor_features['cluster'] = kmeans.fit_predict(X_scaled)

# Label clusters
cluster_profiles = investor_features.groupby('cluster').agg({
    'avg_n_stocks': 'mean',
```

```
        'avg_hhi': 'mean',
        'avg_total_value': 'mean',
        'is_foreign': 'mean',
        'is_state': 'mean',
        'shareholder_name': 'count',
    }).rename(columns={'shareholder_name': 'n_investors'})

    print("Investor Clusters:")
    print(cluster_profiles.round(3).to_string())

    return investor_features

# investor_classes = classify_investors_vietnam(ownership_classified, prices_q)
```

### 29.10.3 Event Study: Ownership Disclosure Shocks

Vietnam's threshold-based major shareholder disclosure creates natural events for studying the price impact of ownership changes.

```
def ownership_event_study(major_shareholders: pd.DataFrame,
                          prices: pd.DataFrame,
                          event_window: Tuple[int, int] = (-5, 20),
                          estimation_window: int = 120) -> pd.DataFrame:
    """
    Event study of ownership disclosure announcements.

    Vietnam requires major shareholders ( 5%) to disclose within 7
    business days of crossing ownership thresholds. These disclosures
    can be informationally significant, especially:
    1. Foreign fund accumulation (signal of quality)
    2. State divestiture (equitization signal)
    3. Insider purchases (management confidence signal)

    Uses market model for expected returns:
    E[R_i,t] =  _i +  _i × R_m,t

    Parameters
    ----------
    major_shareholders : pd.DataFrame
        Disclosure events from DataCore.vn
    prices : pd.DataFrame
```

```
        Daily stock prices
    event_window : tuple
        (pre_event_days, post_event_days)
    estimation_window : int
        Days before event window for market model estimation
    """
    events = major_shareholders.copy()
    events = events.sort_values(['ticker', 'date'])

    # Identify significant ownership changes
    events['ownership_change'] = events.groupby(
        ['ticker', 'shareholder_name']
    )['ownership_pct'].diff()

    significant_events = events[
        events['ownership_change'].abs() > 0.01  # > 1 percentage point
    ].copy()

    significant_events['event_type'] = np.where(
        significant_events['ownership_change'] > 0, 'accumulation', 'divestiture'
    )

    # Merge with daily prices
    prices_daily = prices[['ticker', 'date', 'ret']].copy()
    prices_daily = prices_daily.sort_values(['ticker', 'date'])

    # VN-Index as market return (ticker code depends on data provider)
    if 'VNINDEX' in prices_daily['ticker'].values:
        market_ret = prices_daily[prices_daily['ticker'] == 'VNINDEX'][['date', 'ret']].copy
        market_ret = market_ret.rename(columns={'ret': 'mkt_ret'})
    else:
        # Use equal-weighted market return as proxy
        market_ret = (prices_daily.groupby('date')['ret']
                                  .mean()
                                  .reset_index()
                                  .rename(columns={'ret': 'mkt_ret'}))

    # For each event, compute abnormal returns
    results = []
    pre, post = event_window

    for _, event in significant_events.iterrows():
```

```
        ticker = event['ticker']
        event_date = event['date']

        # Get stock returns around the event
        stock_ret = prices_daily[prices_daily['ticker'] == ticker].copy()
        stock_ret = stock_ret.merge(market_ret, on='date', how='left')
        stock_ret = stock_ret.sort_values('date').reset_index(drop=True)

        # Find event date index
        event_idx = stock_ret[stock_ret['date'] >= event_date].index
        if len(event_idx) == 0:
            continue
        event_idx = event_idx[0]

        # Estimation window
        est_start = max(0, event_idx - estimation_window + pre)
        est_end = event_idx + pre
        est_data = stock_ret.iloc[est_start:est_end].dropna(subset=['ret', 'mkt_ret'])

        if len(est_data) < 30:
            continue

        # Market model
        X = sm.add_constant(est_data['mkt_ret'])
        y = est_data['ret']
        try:
            model = sm.OLS(y, X).fit()
        except Exception:
            continue

        # Event window abnormal returns
        ew_start = event_idx + pre
        ew_end = min(event_idx + post + 1, len(stock_ret))
        event_data = stock_ret.iloc[ew_start:ew_end].copy()

        if len(event_data) == 0:
            continue

        event_data['expected_ret'] = (model.params['const'] +
                                      model.params['mkt_ret'] * event_data['mkt_ret'])
        event_data['abnormal_ret'] = event_data['ret'] - event_data['expected_ret']
        event_data['car'] = event_data['abnormal_ret'].cumsum()
```

```
        event_data['event_day'] = range(pre, pre + len(event_data))
        event_data['ticker'] = ticker
        event_data['event_date'] = event_date
        event_data['event_type'] = event['event_type']
        event_data['ownership_change'] = event['ownership_change']
        event_data['shareholder_name'] = event['shareholder_name']

        results.append(event_data)

    if results:
        all_results = pd.concat(results, ignore_index=True)

        # Average CARs by event type
        avg_car = (all_results.groupby(['event_type', 'event_day'])['car']
                              .agg(['mean', 'std', 'count'])
                              .reset_index())
        avg_car['t_stat'] = avg_car['mean'] / (avg_car['std'] / np.sqrt(avg_car['count']))

        print(f"Event Study Results:")
        print(f"  Total events: {significant_events['event_type'].value_counts().to_string()

        # CAR at event day 0, +5, +10, +20
        for et in ['accumulation', 'divestiture']:
            print(f"\n  {et.title()} Events:")
            subset = avg_car[avg_car['event_type'] == et]
            for day in [0, 5, 10, 20]:
                row = subset[subset['event_day'] == day]
                if len(row) > 0:
                    print(f"    CAR({day:+d}): {row.iloc[0]['mean']:.4f} "
                          f"(t={row.iloc[0]['t_stat']:.2f})")

        return all_results

    return pd.DataFrame()

# event_results = ownership_event_study(dc.major_shareholders, dc.prices)
```

---

## 29.11 Empirical Applications

### 29.11.1 Application 1: Foreign Ownership and Stock Returns in Vietnam

Does foreign institutional ownership predict returns in Vietnam? Huang, Liu, and Shu (2023) find evidence consistent with the information advantage hypothesis.

```python
def test_foreign_io_returns(metrics: pd.DataFrame) -> pd.DataFrame:
    """
    Test whether changes in foreign institutional ownership predict
    future stock returns in Vietnam.

    Methodology:
    1. Sort stocks into quintiles by change in foreign IO
    2. Compute equal-weighted and VN-Index-adjusted returns
    3. Report portfolio returns and long-short spread

    This adapts the Chen, Hong, and Stein (2002) breadth test
    specifically for Vietnam's foreign ownership component.
    """
    df = metrics.copy()
    df = df.sort_values(['ticker', 'quarter_end'])

    # Change in foreign IO
    df['delta_foreign'] = df.groupby('ticker')['pct_foreign_total'].diff()

    # Forward quarterly return
    df['fwd_ret'] = df.groupby('ticker')['ret'].shift(-1)

    # Drop missing
    df = df.dropna(subset=['delta_foreign', 'fwd_ret'])

    # Quintile portfolios each quarter
    df['foreign_quintile'] = df.groupby('quarter_end')['delta_foreign'].transform(
        lambda x: pd.qcut(x, 5, labels=[1, 2, 3, 4, 5], duplicates='drop')
    )

    # Portfolio returns
    port_ret = (df.groupby(['quarter_end', 'foreign_quintile'])['fwd_ret']
                  .mean()
                  .reset_index())
```

```
    port_wide = port_ret.pivot(index='quarter_end', columns='foreign_quintile',
                               values='fwd_ret')
    port_wide['LS'] = port_wide[5] - port_wide[1]

    # Test significance
    results = {}
    for q in [1, 2, 3, 4, 5, 'LS']:
        data = port_wide[q].dropna()
        mean_ret = data.mean()
        t_stat = mean_ret / (data.std() / np.sqrt(len(data)))
        results[q] = {
            'Mean Return (%)': mean_ret * 100,
            't-statistic': t_stat,
            'N quarters': len(data),
        }

    results_df = pd.DataFrame(results).T
    results_df.index.name = 'ΔForeign IO Quintile'

    print("Foreign Ownership Change and Future Returns (Vietnam)")
    print("=" * 60)
    print(results_df.round(3).to_string())

    return results_df

# foreign_return_results = test_foreign_io_returns(io_metrics)
```

### 29.11.2 Application 2: State Divestiture and Value Creation

```
def analyze_equitization_value(metrics: pd.DataFrame,
                               state_analysis: Dict) -> pd.DataFrame:
    """
    Test whether reductions in state ownership are associated with
    subsequent value creation (higher returns, improved governance).

    Hypothesis: State divestiture reduces agency costs, improves
    operational efficiency, and attracts institutional investors,
    leading to positive abnormal returns.

    Uses a difference-in-differences approach:
```

```
    Treatment: Firms experiencing >10pp drop in state ownership
    Control: Matched firms with stable state ownership
    """
    df = metrics.copy()
    events = state_analysis['equitization_events']

    if len(events) == 0:
        print("No equitization events detected.")
        return pd.DataFrame()

    # Get treated firms and their event quarters
    treated = events[['ticker', 'quarter_end']].drop_duplicates()
    treated['treated'] = 1

    # Merge with metrics
    df = df.merge(treated, on=['ticker', 'quarter_end'], how='left')
    df['treated'] = df['treated'].fillna(0)

    # Pre/post comparison for treated firms
    treated_tickers = treated['ticker'].unique()

    results = []
    for ticker in treated_tickers:
        firm = df[df['ticker'] == ticker].sort_values('quarter_end')
        event_row = firm[firm['treated'] == 1]
        if len(event_row) == 0:
            continue

        event_q = event_row.iloc[0]['quarter_end']

        # Pre-event (4 quarters before)
        pre = firm[firm['quarter_end'] < event_q].tail(4)
        # Post-event (4 quarters after)
        post = firm[firm['quarter_end'] > event_q].head(4)

        if len(pre) < 2 or len(post) < 2:
            continue

        results.append({
            'ticker': ticker,
            'event_quarter': event_q,
            'state_pct_pre': pre['pct_state'].mean(),
```

```
            'state_pct_post': post['pct_state'].mean(),
            'foreign_pct_pre': pre['pct_foreign_total'].mean(),
            'foreign_pct_post': post['pct_foreign_total'].mean(),
            'n_inst_pre': pre['n_inst_owners'].mean(),
            'n_inst_post': post['n_inst_owners'].mean(),
            'ret_pre': pre['ret'].mean(),
            'ret_post': post['ret'].mean(),
        })

    if results:
        results_df = pd.DataFrame(results)

        # Paired t-tests
        print("Equitization Value Analysis")
        print("=" * 60)
        for metric in ['state_pct', 'foreign_pct', 'n_inst', 'ret']:
            pre_col = f'{metric}_pre'
            post_col = f'{metric}_post'
            diff = results_df[post_col] - results_df[pre_col]
            t_stat, p_val = stats.ttest_1samp(diff.dropna(), 0)
            print(f"  Δ{metric}: {diff.mean():.4f} (t={t_stat:.2f}, p={p_val:.3f})")

        return results_df

    return pd.DataFrame()

# equitization_results = analyze_equitization_value(io_metrics, state_analysis)
```

### 29.11.3 Application 3: Institutional Herding in Vietnam

```
def compute_herding_vietnam(trades: pd.DataFrame,
                            owner_types: Optional[List[str]] = None) -> pd.DataFrame:
    """
    Compute the Lakonishok, Shleifer, and Vishny (1992) herding measure
    adapted for the Vietnamese market.

    Can be computed separately for:
    - All institutional investors
    - Foreign institutions only
    - Domestic institutions only
```

```
    The herding measure captures whether institutions systematically
    trade in the same direction beyond what chance would predict.
    """
    from scipy.stats import binom

    t = trades.copy()

    if owner_types:
        t = t[t['owner_type'].isin(owner_types)]

    t['is_buy'] = (t['trade'] > 0).astype(int)

    # For each stock-period
    stock_trades = t.groupby(['ticker', 'date']).agg(
        n_traders=('shareholder_name', 'nunique'),
        n_buyers=('is_buy', 'sum'),
    ).reset_index()

    # Minimum traders threshold
    stock_trades = stock_trades[stock_trades['n_traders'] >= 3]
    stock_trades['p_buy'] = stock_trades['n_buyers'] / stock_trades['n_traders']

    # Expected proportion per period
    E_p = stock_trades.groupby('date').apply(
        lambda g: g['n_buyers'].sum() / g['n_traders'].sum()
    ).reset_index(name='E_p')

    stock_trades = stock_trades.merge(E_p, on='date')

    # Adjustment factor
    def expected_abs_dev(n, p):
        k = np.arange(0, n + 1)
        probs = binom.pmf(k, n, p)
        return np.sum(probs * np.abs(k / n - p))

    stock_trades['adj_factor'] = stock_trades.apply(
        lambda r: expected_abs_dev(int(r['n_traders']), r['E_p']), axis=1
    )

    stock_trades['hm'] = (np.abs(stock_trades['p_buy'] - stock_trades['E_p']) -
                          stock_trades['adj_factor'])
```

```
    stock_trades['buy_herd'] = np.where(
        stock_trades['p_buy'] > stock_trades['E_p'], stock_trades['hm'], np.nan
    )
    stock_trades['sell_herd'] = np.where(
        stock_trades['p_buy'] < stock_trades['E_p'], stock_trades['hm'], np.nan
    )

    # Time series of herding
    ts_herding = stock_trades.groupby('date').agg(
        mean_hm=('hm', 'mean'),
        mean_buy_herd=('buy_herd', 'mean'),
        mean_sell_herd=('sell_herd', 'mean'),
        pct_herding=('hm', lambda x: (x > 0).mean()),
        n_stocks=('ticker', 'nunique'),
    ).reset_index()

    print(f"Herding Analysis ({owner_types or 'All Institutions'}):")
    print(f"  Mean HM: {stock_trades['hm'].mean():.4f}")
    print(f"  Mean Buy Herding: {stock_trades['buy_herd'].mean():.4f}")
    print(f"  Mean Sell Herding: {stock_trades['sell_herd'].mean():.4f}")
    print(f"  % stocks with herding: {(stock_trades['hm'] > 0).mean():.1%}")

    return stock_trades, ts_herding

# herding_all, herding_ts = compute_herding_vietnam(trades)
# herding_foreign, _ = compute_herding_vietnam(
#     trades, owner_types=[OwnershipType.FOREIGN_INST]
# )
```

## 29.12 Conclusion and Practical Recommendations

### 29.12.1 Summary of Measures

Table 29.5 summarizes all institutional ownership measures developed in this chapter for the Vietnamese market.

Table 29.5: Summary of All Ownership Measures for Vietnam

| Measure | Definition | Key Adaptation for Vietnam | Python Function |
|---|---|---|---|
| IO Ratio | Inst. shares / TSO | Decomposed into state, foreign, domestic | `compute_ownership_decomposition(` |
| HHI Concentration | $\sum w_j^2$ | Separate HHI for total, non-state, foreign | `compute_io_metrics_vietnam()` |
| ΔBreadth | Lehavy-Sloan adjusted | Applied to irregular disclosure intervals | `compute_io_metrics_vietnam()` |
| FOL Utilization | Foreign % / FOL limit | Vietnam-specific; no US equivalent | `FOLAnalyzer` |
| FOL Premium | Price impact of FOL proximity | Cross-sectional regression approach | `FOLAnalyzer.estimate_fol_premium` |
| Trades | ΔShares (corp-action adjusted) | Critical: adjust for stock dividends | `derive_trades_vectorized_vietnam` |
| Fund Turnover | min(B,S)/avg(A) | Semi-annual frequency; annualized | `compute_fund_analytics()` |
| SOE Status | State ownership > 50% | Tracks equitization program | `analyze_state_ownership()` |
| LSV Herding | $\lvert p - E[p]\rvert - E[\lvert p - E[p]\rvert]$ | Separate foreign vs domestic herding | `compute_herding_vietnam()` |
| Co-Ownership Network | Shared institutional holders | Reveals conglomerate linkages | `construct_stock_coownership_netw` |

### 29.12.2 Data Quality Checklist for Vietnam

Vietnam Data Quality Checklist

1. **Corporate actions:** Have you built and applied adjustment factors for ALL stock dividends, bonus shares, splits, and rights issues?
2. **Shareholder classification:** Have you verified the owner type classification (state vs foreign vs domestic institutional vs individual)?
3. **FOL limits:** Are sector-specific FOL limits correctly assigned (30% for banks, 49% standard, unlimited for some sectors)?
4. **Disclosure dates:** Are you using the actual disclosure date (not the record date or ex-date) for ownership snapshots?
5. **Treasury shares:** Are treasury shares excluded from ownership ratio denominators?

6. **UPCOM coverage:** Does your sample include or exclude UPCOM stocks (which have weaker disclosure requirements)?
7. **Cross-listings:** Are you handling NVDR (Non-Voting Depository Receipts) if applicable after market reforms?
8. **Name consistency:** Are shareholder names standardized across disclosure periods (Vietnamese names can have multiple romanization forms)?
9. **Trade adjustment:** When deriving trades between periods, have you adjusted previous shares for ALL intervening corporate actions?
10. **Fund mandate changes:** For fund analytics, have you accounted for fund mergers, closures, and mandate changes that affect time-series continuity?

### 29.12.3 Comparison with US Framework

Table 29.6: US vs Vietnam Institutional Ownership Framework Comparison

| Dimension | US (WRDS/13F) | Vietnam (DataCore.vn) |
|---|---|---|
| **Disclosure** | Quarterly 13F (mandatory) | Annual reports + event-driven |
| **Coverage** | Institutions > $100M AUM | All shareholders in annual reports |
| **Ownership observed** | Long positions only | Complete decomposition |
| **IO can exceed 100%** | Yes (short selling) | No (by construction) |
| **Permanent ID** | CRSP PERMNO | Ticker (with manual tracking of changes) |
| **Adjustment factors** | CRSP cfacshr | Must build from corporate actions |
| **Investor classification** | LSEG typecode / Bushee | State/Foreign/Domestic/Individual |
| **Short selling** | Not in 13F; exists in market | Very limited; not a concern |
| **Unique features** | — | FOL, SOE ownership, stock dividend frequency |

# 30 Institutional Trades, Flows, and Turnover Ratios

Institutional investors play a pivotal role in price discovery, corporate governance, and market liquidity. Understanding *how* institutions trade and *how much* they trade provides insights into both asset pricing dynamics and the real effects of institutional monitoring. The seminal work of Grinblatt, Titman, and Wermers (1995) on mutual fund momentum trading, Wermers (2000) on fund performance decomposition, and Yan (2008) on the relationship between turnover and future returns all rely on accurately measured institutional trades, flows, and turnover.

In the United States, this research is enabled by the mandatory quarterly 13F filing system administered by the Securities and Exchange Commission (SEC). Every institutional investment manager with at least $100 million in qualifying assets must disclose their equity holdings within 45 days of each calendar quarter end. The Thomson-Reuters (now Refinitiv) 13F database, accessible through WRDS, provides the canonical data infrastructure for this literature.

Vietnam's equity market presents a fundamentally different institutional landscape. This chapter adapts the core methodology for the Vietnamese context, addressing five critical differences:

1. **Disclosure regime.** Vietnam has no 13F-equivalent mandatory quarterly filing. Ownership disclosure is a patchwork of event-driven reports (threshold crossings at 5%, 10%, etc.), annual/semi-annual reports with shareholder registers, and daily foreign ownership tracking by exchanges.

2. **Corporate actions.** Vietnamese firms issue stock dividends and bonus shares at extremely high rates compared to US firms. A firm might issue 20-30% bonus shares in a single year, fundamentally altering the share count. Share adjustment is therefore critical and nontrivial.

3. **Foreign ownership limits (FOLs).** Binding foreign ownership ceiling, typically 49% for most sectors, 30% for banking, and 0% for certain restricted sectors, create a unique institutional constraint. When a stock approaches its FOL, foreign buying becomes mechanically restricted, distorting standard trade inference.

4. **State ownership.** The Vietnamese government retains significant ownership in many listed firms through the State Capital Investment Corporation (SCIC) and other state entities. This creates a distinct ownership category not present in the US 13F data.

5. **Market microstructure.** Daily price limits (±7% on HOSE, ±10% on HNX, ±15% on UPCOM), T+2 settlement, and the absence of short-selling all affect how institutional trades translate into market outcomes.

## 30.1 Measuring Institutional Ownership and Trading

The measurement of institutional ownership and trading activity has been a central concern in empirical finance since Gompers, Ishii, and Metrick (2003) documented the rise of institutional investors. The approach relies on comparing holdings snapshots across consecutive reporting periods to infer trades. If manager $j$ holds $h_{j,i,t}$ shares of stock $i$ at time $t$, then the inferred trade is:

$$\Delta h_{j,i,t} = h_{j,i,t} - h_{j,i,t-1} \tag{30.1}$$

where $\Delta h_{j,i,t} > 0$ indicates a buy and $\Delta h_{j,i,t} < 0$ indicates a sale. This simple differencing approach requires that holdings are observed at regular intervals (e.g., quarterly), share counts are adjusted for corporate actions between reporting dates, and entry and exit from the dataset are handled appropriately.

H.-L. Chen, Jegadeesh, and Wermers (2000) introduced the concept of ownership *breadth* (i.e., the number of institutions holding a stock) and showed that changes in breadth predict future returns. Sias (2004) decomposed institutional demand into a herding component and an information component. Yan (2008) linked fund turnover to information-based trading and documented that high-turnover funds outperform, challenging the view that turnover reflects noise trading.

## 30.2 Trade Classification

Table 30.1 shows four categories of trades:

Table 30.1: Trade Classification Taxonomy

| Code | Type | Description |
|---|---|---|
| +1 | Initiating Buy | Manager enters a new position |
| +2 | Incremental Buy | Manager increases an existing position |
| −1 | Terminating Sale | Manager completely exits a position |
| −2 | Regular Sale | Manager reduces an existing position |

This classification is informative because initiating buys and terminating sales represent discrete portfolio decisions with different information content from marginal position adjustments (Alexander, Cici, and Gibson 2007).

## 30.3 Turnover Measures

Three standard turnover definitions have been used in the literature:

**Carhart (1997) Turnover.** The minimum of aggregate buys and sales, normalized by average assets:

$$\text{Turnover}^{C}_{j,t} = \frac{\min\left(\sum_i B_{j,i,t},\, \sum_i S_{j,i,t}\right)}{\frac{1}{2}\left(A_{j,t} + A_{j,t-1}\right)} \tag{30.2}$$

where $B_{j,i,t}$ and $S_{j,i,t}$ are the dollar values of buys and sales of stock $i$ by manager $j$ in quarter $t$, and $A_{j,t}$ is total portfolio assets (Mark M. Carhart 1997b).

**Flow-Adjusted Turnover.** Adds back the absolute value of net flows to account for flow-driven trading:

$$\text{Turnover}^{F}_{j,t} = \frac{\min\left(\sum_i B_{j,i,t},\, \sum_i S_{j,i,t}\right) + |\text{NetFlow}_{j,t}|}{A_{j,t-1}} \tag{30.3}$$

**Symmetric Turnover.** Uses the sum of buys and sales minus the absolute net flow:

$$\text{Turnover}^{S}_{j,t} = \frac{\sum_i B_{j,i,t} + \sum_i S_{j,i,t} - |\text{NetFlow}_{j,t}|}{A_{j,t-1}} \tag{30.4}$$

The relationship between these measures depends on the correlation between discretionary trading and flow-induced trading (Pástor and Stambaugh 2003).

## 30.4 Institutional Ownership in Emerging Markets

The emerging markets literature has documented several stylized facts about institutional ownership that differ from developed market findings. Aggarwal et al. (2011) documented that foreign institutional ownership improves corporate governance in emerging markets. For Vietnam specifically, Phung and Mishra (2016) examined the relationship between ownership structure and firm performance, while Vo (2015) studied the impact of foreign ownership on stock market liquidity.

## 30.5 Net Flows and Performance Attribution

Net flows measure the dollar amount of new money entering or leaving a fund:

$$\text{NetFlow}_{j,t} = A_{j,t} - A_{j,t-1}(1 + R^p_{j,t}) \tag{30.5}$$

where $R^p_{j,t}$ is the portfolio return. This decomposition, due to Sirri and Tufano (1998), separates changes in fund assets into investment returns and investor capital allocation decisions. Coval and Stafford (2007) showed that flow-driven trades create price pressure, with fire sales by funds experiencing redemptions generating significant negative abnormal returns.

# 31 Data Infrastructure

Table 31.1 summarizes the datasets used in this chapter.

Table 31.1: DataCore.vn Datasets Used in This Chapter

| Dataset | Content | Frequency | Key Variables |
|---|---|---|---|
| Stock Prices | Daily/monthly OHLCV | Daily | `ticker`, `date`, `close`, `adjusted_close`, `volume`, `shares_outstanding` |
| Ownership Structure | Shareholder composition | Quarterly/Annual | `ticker`, `date`, `shareholder_name`, `shares_held`, `pct`, `type` |
| Major Shareholders | Holders $\geq 5\%$ | Event-driven | `ticker`, `date`, `shareholder_name`, `shares`, `is_foreign`, `is_state` |
| Corporate Actions | Splits, dividends, bonus | Event | `ticker`, `ex_date`, `action_type`, `ratio` |
| Company Profile | Sector, exchange, FOL | Static/Annual | `ticker`, `exchange`, `industry`, `listing_date`, `fol_limit` |
| Foreign Ownership | Daily foreign tracking | Daily | `ticker`, `date`, `foreign_shares`, `foreign_pct`, `fol_limit` |
| Fund Holdings | Fund portfolio snapshots | Semi-annual | `fund_name`, `report_date`, `ticker`, `shares_held`, `market_value` |

## 31.1 Data Reader Class

We begin by defining a unified data reader that handles file loading, date parsing, and basic validation:

```
@dataclass
class DataCoreReader:
    """
    Unified reader for DataCore.vn datasets stored locally.

    Supports Parquet (recommended) and CSV formats. Implements
    lazy loading with caching to minimize memory footprint.

    Parameters
    ----------
    data_dir : str or Path
        Directory containing DataCore.vn data files.
    file_format : str
        File format: 'parquet' or 'csv'.

    Examples
    --------
    >>> dc = DataCoreReader('/data/datacore', file_format='parquet')
    >>> prices = dc.prices
    >>> ownership = dc.ownership
    """
    data_dir: Path
    file_format: str = 'parquet'
    _cache: Dict[str, pd.DataFrame] = field(
        default_factory=dict, repr=False
    )

    FILE_MAP: Dict[str, str] = field(default_factory=lambda: {
        'prices': 'stock_prices',
        'ownership': 'ownership_structure',
        'major_shareholders': 'major_shareholders',
        'corporate_actions': 'corporate_actions',
        'company_profile': 'company_profile',
        'financials': 'financial_statements',
        'foreign_ownership': 'foreign_ownership',
        'fund_holdings': 'fund_holdings',
    }, repr=False)
```

```python
    def __post_init__(self):
        self.data_dir = Path(self.data_dir)
        if not self.data_dir.exists():
            raise FileNotFoundError(
                f"Data directory not found: {self.data_dir}"
            )

    def _read(self, key: str) -> pd.DataFrame:
        """Read and cache a dataset with automatic date parsing."""
        if key in self._cache:
            return self._cache[key]

        fname = self.FILE_MAP.get(key, key)
        filepath = self.data_dir / f"{fname}.{self.file_format}"

        if not filepath.exists():
            raise FileNotFoundError(
                f"Dataset not found: {filepath}\n"
                f"Available: "
                f"{list(self.data_dir.glob(f'*.{self.file_format}'))}"
            )

        if self.file_format == 'parquet':
            df = pd.read_parquet(filepath)
        else:
            df = pd.read_csv(filepath, parse_dates=True)

        # Auto-detect and parse date columns
        date_cols = [
            'date', 'ex_date', 'record_date', 'period',
            'report_date', 'listing_date'
        ]
        for col in df.columns:
            if col.lower() in date_cols or 'date' in col.lower():
                try:
                    df[col] = pd.to_datetime(df[col])
                except (ValueError, TypeError):
                    pass

        self._cache[key] = df
        print(f"  Loaded {key}: {len(df):,} rows x {len(df.columns)} cols")
        return df
```

```python
    @property
    def prices(self) -> pd.DataFrame:
        return self._read('prices')

    @property
    def ownership(self) -> pd.DataFrame:
        return self._read('ownership')

    @property
    def major_shareholders(self) -> pd.DataFrame:
        return self._read('major_shareholders')

    @property
    def corporate_actions(self) -> pd.DataFrame:
        return self._read('corporate_actions')

    @property
    def company_profile(self) -> pd.DataFrame:
        return self._read('company_profile')

    @property
    def foreign_ownership(self) -> pd.DataFrame:
        return self._read('foreign_ownership')

    @property
    def fund_holdings(self) -> pd.DataFrame:
        return self._read('fund_holdings')

    def clear_cache(self):
        n = len(self._cache)
        self._cache.clear()
        print(f"  Cleared {n} cached datasets")

# Initialize:
# dc = DataCoreReader('/path/to/datacore_data', file_format='parquet')
```

# 32 Stock Price and Return Processing

The first step processes stock data to obtain adjusted prices, shares outstanding, and quarterly returns.

## 32.1 Price Data Extraction and Adjustment

Vietnamese stock data requires careful adjustment for frequent corporate actions. Unlike the US where CRSP provides a cumulative adjustment factor (`cfacpr`, `cfacshr`), in Vietnam we must construct adjustment factors from the corporate actions history.

> **Vietnamese Corporate Actions**
>
> Vietnamese firms commonly execute the following corporate actions, each requiring share count and/or price adjustment:
>
> - **Stock dividend** (*co tuc bang co phieu*): e.g., 20% stock dividend means 100 shares become 120 shares
> - **Bonus shares** (*co phieu thuong*): free shares distributed from retained earnings
> - **Rights issue** (*phat hanh quyen mua*): right to buy new shares at a discount
> - **Stock split/reverse split** (*chia/gop co phieu*): rare but occasionally used

```python
def build_adjustment_factors(
    corporate_actions: pd.DataFrame,
) -> pd.DataFrame:
    """
    Construct cumulative share adjustment factors from corporate actions.

    This is the Vietnamese equivalent of CRSP's cfacshr factor. For each
    ticker, we compute a cumulative product of adjustment ratios from
    corporate actions, working forward in time.

    The adjustment factor at date t converts historical share counts to
    be comparable with current (post-action) share counts:
```

```
        shares_adjusted_t = shares_raw_t * cfacshr_t

    Parameters
    ----------
    corporate_actions : pd.DataFrame
        Corporate actions with columns: ticker, ex_date, action_type,
        ratio. The ratio field represents:
        - Stock dividend 20%: ratio = 1.20
        - 2:1 stock split: ratio = 2.00
        - Bonus shares 10%: ratio = 1.10

    Returns
    -------
    pd.DataFrame
        Adjustment factors: ticker, ex_date, cfacshr (cumulative).
    """
    share_actions = corporate_actions[
        corporate_actions['action_type'].isin([
            'stock_dividend', 'bonus_shares', 'stock_split',
            'reverse_split', 'rights_issue'
        ])
    ].copy()

    if share_actions.empty:
        return pd.DataFrame(columns=['ticker', 'ex_date', 'cfacshr'])

    share_actions = share_actions.sort_values(['ticker', 'ex_date'])

    share_actions['cfacshr'] = (
        share_actions
        .groupby('ticker')['ratio']
        .cumprod()
    )

    return share_actions[['ticker', 'ex_date', 'cfacshr']].reset_index(
        drop=True
    )


def get_cfacshr_at_date(
    ticker: str,
    date: pd.Timestamp,
```

```
    adj_factors: pd.DataFrame,
) -> float:
    """
    Look up the cumulative share adjustment factor for a given
    ticker and date. Returns 1.0 if no corporate actions occurred.
    """
    mask = (
        (adj_factors['ticker'] == ticker) &
        (adj_factors['ex_date'] <= date)
    )
    subset = adj_factors.loc[mask]

    if subset.empty:
        return 1.0
    return subset.iloc[-1]['cfacshr']


def adjust_shares_between_dates(
    shares: float,
    ticker: str,
    date_from: pd.Timestamp,
    date_to: pd.Timestamp,
    adj_factors: pd.DataFrame,
) -> float:
    """
    Adjust a share count observed at date_from to be comparable
    with shares observed at date_to, accounting for all intervening
    corporate actions.

    Example
    -------
    >>> # Investor held 1000 shares on 2023-01-01
    >>> # A 20% stock dividend occurred on 2023-03-15
    >>> adjust_shares_between_dates(
    ...     1000, 'VNM',
    ...     pd.Timestamp('2023-01-01'),
    ...     pd.Timestamp('2023-06-30'), adj_factors
    ... )
    1200.0
    """
    factor_from = get_cfacshr_at_date(ticker, date_from, adj_factors)
    factor_to = get_cfacshr_at_date(ticker, date_to, adj_factors)
```

```
    relative_factor = factor_to / factor_from
    return shares * relative_factor
```

## 32.2 Monthly and Quarterly Price Processing

```
def process_prices(
    prices: pd.DataFrame,
    adj_factors: pd.DataFrame,
    begdate: str = '2010-01-01',
    enddate: str = '2024-12-31',
) -> Tuple[pd.DataFrame, pd.DataFrame]:
    """
    Process raw DataCore.vn price data into analysis-ready format.

    Block logic:
    1. Filter to date range
    2. Compute adjusted prices and shares outstanding
    3. Compute quarterly compounded returns
    4. Create forward quarterly returns (shifted one quarter)

    Parameters
    ----------
    prices : pd.DataFrame
        Raw price data with: ticker, date, close, adjusted_close,
        volume, shares_outstanding.
    adj_factors : pd.DataFrame
        Corporate action adjustment factors.
    begdate, enddate : str
        Sample period boundaries.

    Returns
    -------
    Tuple[pd.DataFrame, pd.DataFrame]
        (price_quarterly, qret): quarter-end observations with
        adjusted price, total shares, and forward quarterly return.
    """
    price = prices[
        (prices['date'] >= begdate) & (prices['date'] <= enddate)
    ].copy()
```

```python
# Month-end and quarter-end dates
price['mdate'] = price['date'] + pd.offsets.MonthEnd(0)
price['qdate'] = price['date'] + pd.offsets.QuarterEnd(0)

# Adjusted price
if 'adjusted_close' in price.columns:
    price['p'] = price['adjusted_close']
else:
    price['p'] = price['close']

# Total shares outstanding
price['tso'] = price['shares_outstanding']

# Market capitalization (millions VND)
price['mcap'] = price['p'] * price['tso'] / 1e6

# Filter out zero shares
price = price[price['tso'] > 0].copy()

# Compute daily returns if not present
if 'ret' not in price.columns:
    price = price.sort_values(['ticker', 'date'])
    price['ret'] = price.groupby('ticker')['p'].pct_change()

price['ret'] = price['ret'].fillna(0)
price['logret'] = np.log(1 + price['ret'])

# ---- Quarterly compounded returns ----
qret = (
    price
    .groupby(['ticker', 'qdate'])['logret']
    .sum()
    .reset_index()
)
qret['qret'] = np.exp(qret['logret']) - 1

# Shift qdate back one quarter: make qret a *forward* return
qret['qdate'] = qret['qdate'] + pd.offsets.QuarterEnd(-1)
qret = qret.drop(columns=['logret'])

# ---- Quarter-end observations ----
price_q = price[price['qdate'] == price['mdate']].copy()
```

```
    price_q = price_q[['qdate', 'ticker', 'p', 'tso', 'mcap']].copy()

    # Merge forward quarterly return
    price_q = price_q.merge(qret, on=['ticker', 'qdate'], how='left')

    # Build cfacshr lookup at each quarter-end
    price_q['cfacshr'] = price_q.apply(
        lambda row: get_cfacshr_at_date(
            row['ticker'], row['qdate'], adj_factors
        ),
        axis=1
    )

    return price_q, qret
```

**Performance Optimization**

The `get_cfacshr_at_date` function uses a row-wise lookup which can be slow for large datasets. For production use with millions of rows, vectorize using `pd.merge_asof()`:

```
price_q = pd.merge_asof(
    price_q.sort_values('qdate'),
    adj_factors.sort_values('ex_date'),
    by='ticker',
    left_on='qdate',
    right_on='ex_date',
    direction='backward'
).fillna({'cfacshr': 1.0})
```

The output is a quarterly panel of stock-level observations (@tbl-institutional-price-vars)

Table 32.1: Quarter-End Price Panel Variables

| Variable | Description |
|---|---|
| `ticker` | Stock ticker (e.g., VNM, VCB, FPT) |
| `qdate` | Quarter-end date |
| `p` | Adjusted closing price (VND) |
| `tso` | Total shares outstanding |
| `mcap` | Market capitalization (millions VND) |
| `qret` | Forward quarterly compounded return |
| `cfacshr` | Cumulative share adjustment factor |

# 33 Ownership Data Processing

## 33.1 Ownership Taxonomy

We define a classification system for Vietnamese shareholders that maps to the categories available in DataCore.vn:

```
class OwnershipType:
    """
    Vietnamese ownership type classification.

    Vietnam's ownership structure is fundamentally different from the US:

    - **State** (Nha nuoc): SCIC, ministries, state-owned parents
    - **Foreign Institutional** (To chuc nuoc ngoai): foreign funds,
      ETFs, pension funds, insurance, sovereign wealth funds
    - **Domestic Institutional** (To chuc trong nuoc): Vietnamese
      securities companies, fund managers, banks, insurance
    - **Individual** (Ca nhan): retail investors (domestic + foreign)
    - **Treasury** (Co phieu quy): company repurchases
    """

    STATE = 'State'
    FOREIGN_INST = 'Foreign Institutional'
    DOMESTIC_INST = 'Domestic Institutional'
    INDIVIDUAL = 'Individual'
    TREASURY = 'Treasury'

    INSTITUTIONAL = [FOREIGN_INST, DOMESTIC_INST]
    ALL_INSTITUTIONAL = [STATE, FOREIGN_INST, DOMESTIC_INST]
    ALL_TYPES = [STATE, FOREIGN_INST, DOMESTIC_INST, INDIVIDUAL, TREASURY]

    STATE_KEYWORDS = [
        'scic', 'state capital', 'bo', 'ubnd', 'tong cong ty',
        'nha nuoc', 'state', 'government', "people's committee",
        'ministry', 'vietnam national', 'vnpt', 'evn', 'pvn',
```

```
    ]

    FOREIGN_KEYWORDS = [
        'fund', 'investment', 'capital', 'asset management',
        'securities', 'gic', 'templeton', 'dragon capital',
        'vinacapital', 'mekong capital', 'kb securities',
        'mirae asset', 'samsung', 'jp morgan', 'goldman',
        'blackrock', 'vanguard', 'aberdeen', 'hsbc',
    ]

    @classmethod
    def classify(cls, row: pd.Series) -> str:
        """Classify based on explicit flags, then keyword fallback."""
        if pd.notna(row.get('is_state')) and row['is_state']:
            return cls.STATE
        if pd.notna(row.get('is_foreign')) and row['is_foreign']:
            if pd.notna(row.get('is_institution')) and row['is_institution']:
                return cls.FOREIGN_INST
            return cls.INDIVIDUAL
        if pd.notna(row.get('is_institution')) and row['is_institution']:
            return cls.DOMESTIC_INST

        name = str(row.get('shareholder_name', '')).lower()
        if any(kw in name for kw in cls.STATE_KEYWORDS):
            return cls.STATE
        if any(kw in name for kw in cls.FOREIGN_KEYWORDS):
            return cls.FOREIGN_INST

        return cls.INDIVIDUAL
```

## 33.2 Building the Holdings Panel

We construct the holdings panel (i.e., the Vietnamese equivalent of merging the 13F Type 1 and Type 3 datasets). The key steps are:

1. Identify the first available vintage for each shareholder-stock-report date combination.
2. Compute reporting gaps to flag first and last reports.
3. Classify shareholders.
4. Adjust shares for corporate actions.

```python
def build_holdings_panel(
    ownership: pd.DataFrame,
    adj_factors: pd.DataFrame,
    price_q: pd.DataFrame,
    company_profile: pd.DataFrame,
    begdate: str = '2010-01-01',
    enddate: str = '2024-12-31',
) -> pd.DataFrame:
    """
    Construct the institutional holdings panel from DataCore.vn
    ownership data.
    """
    own = ownership.copy()

    # Align to quarter-end
    own['rdate'] = own['date'] + pd.offsets.QuarterEnd(0)
    own['fdate'] = own['date']

    own = own[
        (own['rdate'] >= begdate) & (own['rdate'] <= enddate)
    ].copy()

    # Keep earliest vintage per shareholder-ticker-rdate
    own = own.sort_values(
        ['shareholder_name', 'ticker', 'rdate', 'fdate']
    )
    fst_vint = (
        own
        .groupby(['shareholder_name', 'ticker', 'rdate'])
        .first()
        .reset_index()
    )

    # ---- Reporting gaps for first/last flags ----
    fst_vint = fst_vint.sort_values(
        ['shareholder_name', 'ticker', 'rdate']
    )

    grp = fst_vint.groupby(['shareholder_name', 'ticker'])
    fst_vint['lag_rdate'] = grp['rdate'].shift(1)

    fst_vint['qtr_gap'] = fst_vint.apply(
```

```
        lambda r: (
            (r['rdate'].to_period('Q')
             - r['lag_rdate'].to_period('Q')).n
            if pd.notna(r['lag_rdate']) else np.nan
        ),
        axis=1
    )

    fst_vint['first_report'] = (
        fst_vint['qtr_gap'].isna() | (fst_vint['qtr_gap'] >= 2)
    )

    # Last report flag (forward gap)
    fst_vint = fst_vint.sort_values(
        ['shareholder_name', 'ticker', 'rdate'],
        ascending=[True, True, False]
    )
    fst_vint['lead_rdate'] = grp['rdate'].shift(1)

    fst_vint['lead_gap'] = fst_vint.apply(
        lambda r: (
            (r['lead_rdate'].to_period('Q')
             - r['rdate'].to_period('Q')).n
            if pd.notna(r['lead_rdate']) else np.nan
        ),
        axis=1
    )

    fst_vint['last_report'] = (
        fst_vint['lead_gap'].isna() | (fst_vint['lead_gap'] >= 2)
    )

    fst_vint = fst_vint.drop(
        columns=['lag_rdate', 'qtr_gap', 'lead_rdate', 'lead_gap'],
        errors='ignore'
    )

    # ---- Classify shareholders ----
    fst_vint['owner_type'] = fst_vint.apply(
        OwnershipType.classify, axis=1
    )
```

```python
    # ---- Adjust shares for corporate actions ----
    fst_vint = fst_vint.merge(
        price_q[['ticker', 'qdate', 'cfacshr']],
        left_on=['ticker', 'rdate'],
        right_on=['ticker', 'qdate'],
        how='inner'
    )

    fst_vint['shares_adj'] = (
        fst_vint['shares_held'] * fst_vint['cfacshr']
    )
    fst_vint = fst_vint[fst_vint['shares_adj'] > 0].copy()

    fst_vint = fst_vint.drop_duplicates(
        subset=['shareholder_name', 'ticker', 'rdate']
    )

    # Merge company profile
    if company_profile is not None:
        fst_vint = fst_vint.merge(
            company_profile[['ticker', 'exchange', 'fol_limit']]
            .drop_duplicates(),
            on='ticker',
            how='left'
        )

    cols = [
        'shareholder_name', 'ticker', 'rdate', 'fdate',
        'shares_held', 'shares_adj', 'owner_type',
        'first_report', 'last_report'
    ]
    if 'exchange' in fst_vint.columns:
        cols.extend(['exchange', 'fol_limit'])

    holdings = fst_vint[cols].copy()

    print(f"Holdings panel: {len(holdings):,} observations")
    print(f"  Shareholders: {holdings['shareholder_name'].nunique():,}")
    print(f"  Stocks: {holdings['ticker'].nunique():,}")
    print(f"  Quarters: {holdings['rdate'].nunique()}")

    return holdings
```

# 34 Institutional Ownership Metrics

Before computing trades, we establish the standard institutional ownership metrics that serve as both outputs and inputs to the trading analysis.

## 34.1 Institutional Ownership Ratio

The institutional ownership ratio (IO) for stock $i$ at time $t$ is:

$$IO_{i,t} = \frac{\sum_{j \in \mathcal{J}} h_{j,i,t}}{TSO_{i,t}} \tag{34.1}$$

where $\mathcal{J}$ is the set of institutional investors and $TSO_{i,t}$ is total shares outstanding. In Vietnam, we compute separate ratios for each ownership type:

$$IO_{i,t}^{\text{type}} = \frac{\sum_{j \in \mathcal{J}^{\text{type}}} h_{j,i,t}}{TSO_{i,t}}, \quad \text{type} \in \{\text{State}, \text{Foreign}, \text{Domestic}, \text{Individual}\} \tag{34.2}$$

```
def compute_io_ratios(
    holdings: pd.DataFrame,
    price_q: pd.DataFrame,
) -> pd.DataFrame:
    """Compute IO ratios by type for each stock-quarter."""
    agg = (
        holdings
        .groupby(['ticker', 'rdate', 'owner_type'])['shares_adj']
        .sum()
        .reset_index()
    )

    io_wide = agg.pivot_table(
        index=['ticker', 'rdate'],
        columns='owner_type',
        values='shares_adj',
```

```
        fill_value=0
    ).reset_index()

    io_wide.columns = [
        c if c in ['ticker', 'rdate']
        else f'shares_{c.lower().replace(" ", "_")}'
        for c in io_wide.columns
    ]

    io_wide = io_wide.merge(
        price_q[['ticker', 'qdate', 'tso']],
        left_on=['ticker', 'rdate'],
        right_on=['ticker', 'qdate'],
        how='inner'
    )

    share_cols = [c for c in io_wide.columns if c.startswith('shares_')]
    for col in share_cols:
        ratio_name = col.replace('shares_', 'io_')
        io_wide[ratio_name] = io_wide[col] / io_wide['tso']

    inst_cols = [
        c for c in io_wide.columns
        if c.startswith('shares_')
        and 'individual' not in c
        and 'treasury' not in c
    ]
    io_wide['io_total_inst'] = (
        io_wide[inst_cols].sum(axis=1) / io_wide['tso']
    )

    return io_wide
```

## 34.2 Ownership Concentration: Herfindahl-Hirschman Index

The HHI measures ownership concentration:

$$HHI_{i,t} = \sum_{j=1}^{N_{i,t}} \left( \frac{h_{j,i,t}}{\sum_{k=1}^{N_{i,t}} h_{k,i,t}} \right)^2 \tag{34.3}$$

where $N_{i,t}$ is the number of shareholders. HHI ranges from $1/N_{i,t}$ (equal) to 1 (single shareholder). In Vietnam, ownership tends to be highly concentrated due to large state and founding-family blocks.

```
def compute_hhi(holdings: pd.DataFrame) -> pd.DataFrame:
    """Compute HHI for each stock-quarter, overall and institutional."""
    def _hhi(shares: pd.Series) -> float:
        total = shares.sum()
        if total <= 0:
            return np.nan
        weights = shares / total
        return (weights ** 2).sum()

    hhi_overall = (
        holdings.groupby(['ticker', 'rdate'])['shares_adj']
        .apply(_hhi).reset_index()
        .rename(columns={'shares_adj': 'hhi_overall'})
    )

    inst = holdings[
        holdings['owner_type'].isin(OwnershipType.ALL_INSTITUTIONAL)
    ]
    hhi_inst = (
        inst.groupby(['ticker', 'rdate'])['shares_adj']
        .apply(_hhi).reset_index()
        .rename(columns={'shares_adj': 'hhi_institutional'})
    )

    return hhi_overall.merge(hhi_inst, on=['ticker', 'rdate'], how='left')
```

## 34.3 Ownership Breadth

Following H.-L. Chen, Jegadeesh, and Wermers (2000), ownership breadth is the number of institutional holders:

$$\text{Breadth}_{i,t} = \#\{j : h_{j,i,t} > 0,\ j \in \mathcal{J}\} \tag{34.4}$$

The *change* in breadth predicts future returns:

$$\Delta\text{Breadth}_{i,t} = \text{Breadth}_{i,t} - \text{Breadth}_{i,t-1} \tag{34.5}$$

```python
def compute_breadth(holdings: pd.DataFrame) -> pd.DataFrame:
    """Compute ownership breadth and changes by type."""
    breadth = (
        holdings[
            holdings['owner_type'].isin(OwnershipType.ALL_INSTITUTIONAL)
        ]
        .groupby(['ticker', 'rdate', 'owner_type'])['shareholder_name']
        .nunique()
        .reset_index()
        .rename(columns={'shareholder_name': 'n_holders'})
    )

    breadth_wide = breadth.pivot_table(
        index=['ticker', 'rdate'],
        columns='owner_type',
        values='n_holders',
        fill_value=0
    ).reset_index()

    breadth_wide.columns = [
        c if c in ['ticker', 'rdate']
        else f'n_{c.lower().replace(" ", "_")}'
        for c in breadth_wide.columns
    ]

    n_cols = [c for c in breadth_wide.columns if c.startswith('n_')]
    breadth_wide['n_total_inst'] = breadth_wide[n_cols].sum(axis=1)

    breadth_wide = breadth_wide.sort_values(['ticker', 'rdate'])
    for col in n_cols + ['n_total_inst']:
        breadth_wide[f'd_{col}'] = (
            breadth_wide.groupby('ticker')[col].diff()
        )

    return breadth_wide
```

($\mathrm{BS} = -1$) is generated for the prior position, dated to the quarter after the last report.

For intermediate gaps (reports at $t-2$ and $t$ but not $t-1$), we split into:

- A terminating sale at $t-1$ of $-h^{\mathrm{adj}}_{j,i,t-2}$;
- An initiating buy at $t$ of $h_{j,i,t}$.

## 34.4 Implementation

```
def compute_trades(
    holdings: pd.DataFrame,
    adj_factors: pd.DataFrame,
) -> pd.DataFrame:
    """
    Compute institutional trades from holdings panel.

    Uses vectorized conditional logic (NOT apply()) for performance.

    Algorithm:
    1. Sort holdings by shareholder, ticker, quarter
    2. Compute lagged holdings and reporting gaps
    3. Apply modified trade logic based on first_report, gap
    4. Handle terminating sales and intermediate gaps
    5. Append all trade records
    """
    t1 = holdings.sort_values(
        ['shareholder_name', 'ticker', 'rdate']
    ).copy()

    # Previous holding quarter and shares
    grp = t1.groupby(['shareholder_name', 'ticker'])
    t1['phrdate'] = grp['rdate'].shift(1)
    t1['pshares_adj'] = grp['shares_adj'].shift(1)

    # Raw trade
    t1['trade'] = t1['shares_adj'] - t1['pshares_adj']

    # Quarter gap
    t1['qtrgap'] = t1.apply(
        lambda r: (
            (r['rdate'].to_period('Q')
             - r['phrdate'].to_period('Q')).n
            if pd.notna(r['phrdate']) else np.nan
        ),
        axis=1
    )

    # Boundary detection keys
```

```
t1['l_key'] = (
    t1['shareholder_name'] + '_' + t1['ticker']
).shift(1)
t1['n_key'] = (
    t1['shareholder_name'] + '_' + t1['ticker']
).shift(-1)
t1['curr_key'] = t1['shareholder_name'] + '_' + t1['ticker']

# ---- Vectorized trade classification ----
is_new = (t1['curr_key'] != t1['l_key'])
not_first = ~t1['first_report']
consec = (t1['qtrgap'] == 1)
gap = (t1['qtrgap'] != 1) & t1['qtrgap'].notna()

cond1   = is_new
cond1_1 = is_new & not_first
cond2_1 = (~is_new) & not_first & consec
cond2_2 = (~is_new) & not_first & gap

# Modified trade amounts
t1['modtrade'] = t1['trade']
t1.loc[cond1, 'modtrade'] = np.nan
t1.loc[cond1_1, 'modtrade'] = t1.loc[cond1_1, 'shares_adj']
t1.loc[cond2_1, 'modtrade'] = t1.loc[cond2_1, 'trade']
t1.loc[cond2_2, 'modtrade'] = t1.loc[cond2_2, 'shares_adj']

# Buy/sale classification
t1['buysale'] = np.nan
t1.loc[cond1_1, 'buysale'] = 1
t1.loc[cond2_1, 'buysale'] = (
    2 * np.sign(t1.loc[cond2_1, 'trade'])
)
t1.loc[cond2_2, 'buysale'] = 1.5  # placeholder for split

# ---- Handle intermediate gaps (buysale == 1.5) ----
t2 = t1[t1['buysale'] == 1.5].copy()
t2['rdate'] = t2['phrdate'] + pd.offsets.QuarterEnd(1)
t2['buysale'] = -1
t2['modtrade'] = -t2['pshares_adj']

t1.loc[t1['buysale'] == 1.5, 'buysale'] = 1
```

```
    # ---- Terminating sales ----
    is_last_combo = (t1['curr_key'] != t1['n_key'])
    not_last_rpt = ~t1['last_report']

    t3 = t1[is_last_combo & not_last_rpt].copy()
    t3['rdate'] = t3['rdate'] + pd.offsets.QuarterEnd(1)
    t3['modtrade'] = -t3['shares_adj']
    t3['buysale'] = -1

    # ---- Combine ----
    trades = pd.concat([t1, t2, t3], ignore_index=True)
    trades = trades[
        (trades['modtrade'] != 0) &
        trades['modtrade'].notna() &
        trades['buysale'].notna()
    ].copy()

    trades = trades[[
        'rdate', 'shareholder_name', 'ticker', 'modtrade',
        'buysale', 'owner_type', 'first_report', 'last_report'
    ]].rename(columns={'modtrade': 'trade'})

    print(f"\nTrade computation complete:")
    print(f"  Total records: {len(trades):,}")
    print(f"  Initiating buys:  {(trades['buysale'] == 1).sum():,}")
    print(f"  Incremental buys: {(trades['buysale'] == 2).sum():,}")
    print(f"  Terminating sales:{(trades['buysale'] == -1).sum():,}")
    print(f"  Regular sales:    {(trades['buysale'] == -2).sum():,}")

    return trades
```

### 34.4.1 Trade Visualization

```python
def plot_trade_distribution(trades: pd.DataFrame):
    """Plot time series of trade types by quarter."""
    bs_labels = {
        1: 'Initiating Buy', 2: 'Incremental Buy',
        -1: 'Terminating Sale', -2: 'Regular Sale'
    }
    trades = trades.copy()
    trades['trade_type'] = trades['buysale'].map(bs_labels)

    counts = (
        trades
        .groupby([pd.Grouper(key='rdate', freq='QE'), 'trade_type'])
        .size()
        .unstack(fill_value=0)
    )

    fig, axes = plt.subplots(2, 1, figsize=(12, 8), sharex=True)

    buy_cols = [c for c in counts.columns if 'Buy' in c]
    counts[buy_cols].plot(
        kind='bar', stacked=True, ax=axes[0],
        color=['#1f77b4', '#aec7e8'], width=0.8
    )
    axes[0].set_title('Panel A: Institutional Purchases', fontweight='bold')
    axes[0].set_ylabel('Number of Trades')

    sale_cols = [c for c in counts.columns if 'Sale' in c]
    counts[sale_cols].plot(
        kind='bar', stacked=True, ax=axes[1],
        color=['#d62728', '#ff9896'], width=0.8
    )
    axes[1].set_title('Panel B: Institutional Sales', fontweight='bold')
    axes[1].set_ylabel('Number of Trades')

    for ax in axes:
        ax.tick_params(axis='x', rotation=45)
        for i, label in enumerate(ax.get_xticklabels()):
            if i % 4 != 0:
                label.set_visible(False)

    plt.tight_layout()
    plt.show()

# plot_trade_distribution(trades)
```

Figure 134.1

```python
def plot_net_trading_by_type(trades: pd.DataFrame, price_q: pd.DataFrame):
    """Plot net trading volume by owner type over time."""
    _t = trades.merge(
        price_q[['ticker', 'qdate', 'p']],
        left_on=['ticker', 'rdate'],
        right_on=['ticker', 'qdate'],
        how='inner'
    )
    _t['trade_vnd'] = _t['trade'] * _t['p'] / 1e9

    net = (
        _t
        .groupby([pd.Grouper(key='rdate', freq='QE'), 'owner_type'])
        ['trade_vnd'].sum()
        .unstack(fill_value=0)
    )

    fig, ax = plt.subplots(figsize=(12, 6))
    for col in net.columns:
        ax.plot(net.index, net[col], label=col,
                color=OWNER_COLORS.get(col, '#333'), linewidth=1.5)
    ax.axhline(y=0, color='black', linewidth=0.5)
    ax.set_title('Net Institutional Trading by Ownership Type',
                 fontweight='bold')
    ax.set_ylabel('Net Trading (Billions VND)')
    ax.legend(loc='best')
    plt.tight_layout()
    plt.show()

# plot_net_trading_by_type(trades, price_q)
```

Figure 34.2

# 35 Portfolio Assets, Flows, and Returns

This section computes total portfolio assets, aggregates buys and sales, and portfolio-level returns for each institutional investor.

## 35.1 Total Assets and Portfolio Returns

For each manager $j$ and quarter $t$, portfolio assets are:

$$A_{j,t} = \sum_{i=1}^{N_{j,t}} h_{j,i,t} \cdot P_{i,t} \tag{35.1}$$

The portfolio return assuming buy-and-hold is:

$$R^p_{j,t} = \frac{\sum_{i=1}^{N_{j,t}} h_{j,i,t} \cdot P_{i,t} \cdot r_{i,t+1}}{\sum_{i=1}^{N_{j,t}} h_{j,i,t} \cdot P_{i,t}} \tag{35.2}$$

```
def compute_assets_and_returns(
    holdings: pd.DataFrame,
    price_q: pd.DataFrame,
) -> pd.DataFrame:
    """Compute total portfolio assets and buy-and-hold returns."""
    _assets = holdings[
        ['shareholder_name', 'ticker', 'rdate', 'shares_adj']
    ].merge(
        price_q[['ticker', 'qdate', 'p', 'qret']],
        left_on=['ticker', 'rdate'],
        right_on=['ticker', 'qdate'],
        how='inner'
    )

    _assets['hold_per_stock'] = _assets['shares_adj'] * _assets['p'] / 1e6
    _assets['next_value'] = (
        _assets['shares_adj'] * _assets['p'] * _assets['qret']
```

```
    )
    _assets['curr_value'] = _assets['shares_adj'] * _assets['p']

    assets = (
        _assets
        .groupby(['shareholder_name', 'rdate'])
        .agg(
            assets=('hold_per_stock', 'sum'),
            total_next=('next_value', 'sum'),
            total_curr=('curr_value', 'sum'),
        )
        .reset_index()
    )

    assets['pret'] = assets['total_next'] / assets['total_curr']
    assets = assets.drop(columns=['total_next', 'total_curr'])
    return assets
```

## 35.2 Aggregate Buys and Sales

Total buys and sales for manager $j$ in quarter $t$:

$$B_{j,t} = \sum_{i:\Delta h>0} \Delta h_{j,i,t} \cdot P_{i,t}, \qquad S_{j,t} = \sum_{i:\Delta h<0} |\Delta h_{j,i,t}| \cdot P_{i,t} \tag{35.3}$$

The trade gain is:

$$G_{j,t} = \sum_{i=1}^{N_{j,t}} \Delta h_{j,i,t} \cdot P_{i,t} \cdot r_{i,t+1} \tag{35.4}$$

```
def compute_buys_sales(
    trades: pd.DataFrame,
    price_q: pd.DataFrame,
) -> pd.DataFrame:
    """Compute aggregate buys, sales, trade gains per manager-quarter."""
    _flows = trades.merge(
        price_q[['ticker', 'qdate', 'p', 'qret']],
        left_on=['ticker', 'rdate'],
        right_on=['ticker', 'qdate'],
        how='inner'
```

```
    )

    _flows['tbuys'] = (
        _flows['trade'] * (_flows['trade'] > 0).astype(float)
        * _flows['p'] / 1e6
    )
    _flows['tsales'] = (
        (-1) * _flows['trade'] * (_flows['trade'] < 0).astype(float)
        * _flows['p'] / 1e6
    )
    _flows['tgain'] = (
        _flows['trade'] * _flows['p'] * _flows['qret'] / 1e6
    )

    flows = (
        _flows
        .groupby(['shareholder_name', 'rdate'])
        .agg(
            tbuys=('tbuys', 'sum'),
            tsales=('tsales', 'sum'),
            tgain=('tgain', 'sum'),
        )
        .reset_index()
    )
    return flows
```

# 36 Net Flows and Turnover Ratios

## 36.1 Net Flows

Net flows separate capital allocation decisions from investment returns:

$$\text{NetFlow}_{j,t} = A_{j,t} - A_{j,t-1}(1 + R^p_{j,t}) \tag{36.1}$$

> **Interpreting Net Flows in Vietnam**
>
> For state entities or corporate cross-holders, "net flows" do not necessarily reflect investment decisions. State ownership changes often result from government policy (equitization, divestment programs). Interpretation should account for institutional context.

## 36.2 Three Turnover Measures

```
def compute_aggregates(
    holdings: pd.DataFrame,
    assets: pd.DataFrame,
    flows: pd.DataFrame,
) -> pd.DataFrame:
    """
    Compute net flows and three turnover measures.

    1. Carhart (1997): min(buys, sales) / avg(assets)
    2. Flow-adjusted: [min(buys, sales) + |net flows|] / lag assets
    3. Symmetric: [buys + sales - |net flows|] / lag assets
    """
    report_flags = (
        holdings
        .groupby(['shareholder_name', 'rdate'])
        .agg(first_report=('first_report', 'any'),
```

```
            last_report=('last_report', 'any'))
        .reset_index()
    )

    agg = report_flags.merge(
        assets, on=['shareholder_name', 'rdate'], how='inner'
    )
    agg = agg.merge(
        flows, on=['shareholder_name', 'rdate'], how='left'
    )

    agg = agg.sort_values(['shareholder_name', 'rdate'])

    agg['assets_comp'] = agg['assets'] * (1 + agg['pret'].fillna(0))

    grp = agg.groupby('shareholder_name')
    agg['lassets_comp'] = grp['assets_comp'].shift(1)
    agg['lassets'] = grp['assets'].shift(1)

    # Trade gain return
    agg['tgainret'] = agg['tgain'] / (agg['tbuys'] + agg['tsales'])

    # Net flows
    agg['netflows'] = agg['assets'] - agg['lassets_comp']

    # Turnover 1: Carhart (1997)
    agg['turnover1'] = (
        agg[['tbuys', 'tsales']].min(axis=1) /
        agg[['assets', 'lassets']].mean(axis=1)
    )

    # Turnover 2: Flow-adjusted
    agg['turnover2'] = (
        (agg[['tbuys', 'tsales']].min(axis=1)
         + agg['netflows'].abs().fillna(0))
        / agg['lassets']
    )

    # Turnover 3: Symmetric
    agg['turnover3'] = (
        (agg['tbuys'].fillna(0) + agg['tsales'].fillna(0)
         - agg['netflows'].abs().fillna(0))
```

```
        / agg['lassets']
    )

    # Missing for first report
    first_mask = agg['first_report']
    for col in ['netflows', 'tgainret',
                'turnover1', 'turnover2', 'turnover3']:
        agg.loc[first_mask, col] = np.nan

    agg = agg.drop(columns=['assets_comp', 'lassets_comp', 'lassets'])

    print(f"\nAggregates: {len(agg):,} manager-quarters")
    print(f"  Turnover1 mean: {agg['turnover1'].mean():.4f}")
    print(f"  Turnover2 mean: {agg['turnover2'].mean():.4f}")
    print(f"  Turnover3 mean: {agg['turnover3'].mean():.4f}")

    return agg
```

### 36.2.1 Turnover Summary Statistics

Table 36.1: Summary statistics for three turnover measures across institutional investor types in Vietnam. Turnover 1 follows Mark M. Carhart (1997b), Turnover 2 adds back absolute net flows, and Turnover 3 uses the symmetric definition.

```python
def turnover_summary_table(
    aggregates: pd.DataFrame,
    holdings: pd.DataFrame,
) -> pd.DataFrame:
    """Publication-quality turnover summary statistics table."""
    owner_map = (
        holdings.groupby('shareholder_name')['owner_type']
        .first().reset_index()
    )
    agg = aggregates.merge(owner_map, on='shareholder_name', how='left')

    turnover_cols = ['turnover1', 'turnover2', 'turnover3']
    results = []

    for otype in ['All'] + OwnershipType.ALL_TYPES:
        subset = agg if otype == 'All' else agg[agg['owner_type'] == otype]
        row = {'Owner Type': otype, 'N': len(subset)}
        for col in turnover_cols:
            s = subset[col].dropna()
            row[f'{col}_mean'] = s.mean()
            row[f'{col}_median'] = s.median()
            row[f'{col}_std'] = s.std()
        results.append(row)

    return pd.DataFrame(results).round(4)

# turnover_summary_table(aggregates, holdings)
```

```python
def plot_turnover_timeseries(
    aggregates: pd.DataFrame, holdings: pd.DataFrame
):
    """Plot turnover time series by ownership type."""
    owner_map = (
        holdings.groupby('shareholder_name')['owner_type']
        .first().reset_index()
    )
    agg = aggregates.merge(owner_map, on='shareholder_name', how='left')

    fig, ax = plt.subplots(figsize=(12, 6))
    for otype in OwnershipType.ALL_INSTITUTIONAL:
        subset = agg[agg['owner_type'] == otype]
        qtr_mean = (
            subset
            .groupby(pd.Grouper(key='rdate', freq='QE'))['turnover1']
            .mean()
        )
        ax.plot(qtr_mean.index, qtr_mean.values, label=otype,
                color=OWNER_COLORS.get(otype, '#333'), linewidth=1.5)

    ax.set_title('Quarterly Average Turnover (Carhart)',
                 fontweight='bold')
    ax.set_ylabel('Turnover Ratio')
    ax.legend(loc='best')
    ax.yaxis.set_major_formatter(mticker.PercentFormatter(1.0))
    plt.tight_layout()
    plt.show()

# plot_turnover_timeseries(aggregates, holdings)
```

Figure 36.1

# 37 Foreign Ownership Analytics

Vietnam's foreign ownership limits create unique analytical dimensions absent from developed market studies.

## 37.1 FOL Utilization

$$\text{FOL_Util}_{i,t} = \frac{FO_{i,t}}{FOL_i} \tag{37.1}$$

Stocks with $\text{FOL_Util}_{i,t} \to 1$ face mechanical foreign buying restrictions.

```
def compute_fol_analytics(
    foreign_ownership: pd.DataFrame,
    company_profile: pd.DataFrame,
) -> pd.DataFrame:
    """Compute FOL utilization and related metrics."""
    fo = foreign_ownership.copy()
    fo = fo.merge(
        company_profile[['ticker', 'fol_limit']].drop_duplicates(),
        on='ticker', how='left'
    )

    fo['fol_utilization'] = fo['foreign_pct'] / fo['fol_limit']
    fo['foreign_room'] = fo['fol_limit'] - fo['foreign_pct']
    fo['fol_binding'] = (fo['fol_utilization'] >= 0.98)
    fo['fol_category'] = pd.cut(
        fo['fol_utilization'],
        bins=[0, 0.25, 0.50, 0.75, 0.95, 1.0, float('inf')],
        labels=['<25%', '25-50%', '50-75%', '75-95%',
                '95-100%', '>100%']
    )
    return fo
```

## 37.2 Room Premium Regression

When foreign ownership approaches the FOL, remaining "room" becomes scarce. We model:

$$r_{i,t+1} = \alpha + \beta_1 \cdot \text{FOL_Util}_{i,t} + \beta_2 \cdot \text{FOL_Util}^2_{i,t} + \gamma \cdot X_{i,t} + \varepsilon_{i,t} \tag{37.2}$$

The quadratic term captures nonlinear acceleration of the premium as ownership approaches the limit.

```
def estimate_room_premium(
    fol_analytics: pd.DataFrame,
    price_q: pd.DataFrame,
) -> dict:
    """Estimate foreign ownership room premium via panel regression."""
    fol_q = (
        fol_analytics
        .assign(qdate=lambda x: x['date'] + pd.offsets.QuarterEnd(0))
        .groupby(['ticker', 'qdate'])
        .agg(fol_utilization=('fol_utilization', 'last'),
             foreign_room=('foreign_room', 'last'))
        .reset_index()
    )

    panel = fol_q.merge(
        price_q[['ticker', 'qdate', 'mcap', 'qret']],
        on=['ticker', 'qdate'], how='inner'
    )

    panel['log_mcap'] = np.log(panel['mcap'] + 1)
    panel['fol_util_sq'] = panel['fol_utilization'] ** 2
    panel = panel.dropna(subset=['qret', 'fol_utilization', 'log_mcap'])

    X = panel[['fol_utilization', 'fol_util_sq', 'log_mcap']]
    X = sm.add_constant(X)
    y = panel['qret']

    model = sm.OLS(y, X).fit(
        cov_type='cluster', cov_kwds={'groups': panel['ticker']}
    )
    return {'model': model, 'n_obs': len(panel)}

# results = estimate_room_premium(fol_analytics, price_q)
```

```python
def plot_fol_utilization(fol_analytics: pd.DataFrame):
    """Plot FOL utilization distribution."""
    latest = (
        fol_analytics.sort_values(['ticker', 'date'])
        .groupby('ticker').last().reset_index()
    )

    fig, axes = plt.subplots(1, 2, figsize=(14, 5))

    axes[0].hist(latest['fol_utilization'].dropna(), bins=50,
                 color='#1f77b4', alpha=0.7, edgecolor='white')
    axes[0].axvline(x=0.95, color='red', linestyle='--',
                    label='95% threshold')
    axes[0].set_title('Panel A: FOL Utilization Distribution',
                      fontweight='bold')
    axes[0].set_xlabel('FOL Utilization Ratio')
    axes[0].set_ylabel('Number of Stocks')
    axes[0].legend()

    for exch in ['HOSE', 'HNX', 'UPCOM']:
        sub = latest[latest.get('exchange') == exch]
        if len(sub) > 0:
            axes[1].hist(sub['fol_utilization'].dropna(), bins=30,
                         alpha=0.5, label=exch,
                         color=EXCHANGE_COLORS.get(exch, '#333'))
    axes[1].set_title('Panel B: By Exchange', fontweight='bold')
    axes[1].set_xlabel('FOL Utilization Ratio')
    axes[1].legend()

    plt.tight_layout()
    plt.show()

# plot_fol_utilization(fol_analytics)
```

Figure 37.1

# 38 Complete Pipeline

We integrate all steps into a single end-to-end function:

```
def run_complete_pipeline(
    dc: 'DataCoreReader',
    begdate: str = '2010-01-01',
    enddate: str = '2024-12-31',
) -> Dict[str, pd.DataFrame]:
    """
    Execute the complete institutional ownership analytics pipeline.

    Steps:
    1. Build corporate action adjustment factors
    2. Process stock prices
    3. Construct holdings panel (Steps 2-4)
    4. Compute IO metrics
    5. Compute institutional trades (Step 5)
    6. Compute portfolio assets and returns (Step 6a)
    7. Compute aggregate buys, sales, trade gains (Step 6b)
    8. Compute net flows and turnover (Step 7)
    9. Compute foreign ownership analytics

    Returns dict of all output DataFrames.
    """
    print("=" * 60)
    print("INSTITUTIONAL TRADES, FLOWS, AND TURNOVER PIPELINE")
    print(f"Sample: {begdate} to {enddate}")
    print("=" * 60)

    print("\n[1/9] Building adjustment factors...")
    adj_factors = build_adjustment_factors(dc.corporate_actions)

    print("\n[2/9] Processing stock prices...")
    price_q, qret = process_prices(
        dc.prices, adj_factors, begdate, enddate
    )
```

```
    print("\n[3/9] Building holdings panel...")
    holdings = build_holdings_panel(
        dc.ownership, adj_factors, price_q,
        dc.company_profile, begdate, enddate
    )

    print("\n[4/9] Computing ownership metrics...")
    io_ratios = compute_io_ratios(holdings, price_q)
    hhi = compute_hhi(holdings)
    breadth = compute_breadth(holdings)

    print("\n[5/9] Computing institutional trades...")
    trades = compute_trades(holdings, adj_factors)

    print("\n[6/9] Computing portfolio assets...")
    assets = compute_assets_and_returns(holdings, price_q)

    print("\n[7/9] Computing aggregate buys and sales...")
    flows = compute_buys_sales(trades, price_q)

    print("\n[8/9] Computing net flows and turnover...")
    aggregates = compute_aggregates(holdings, assets, flows)

    print("\n[9/9] Computing foreign ownership analytics...")
    fol_analytics = compute_fol_analytics(
        dc.foreign_ownership, dc.company_profile
    )

    print("\n" + "=" * 60)
    print("PIPELINE COMPLETE")
    print("=" * 60)

    return {
        'price_q': price_q, 'holdings': holdings,
        'io_ratios': io_ratios, 'hhi': hhi,
        'breadth': breadth, 'trades': trades,
        'assets': assets, 'flows': flows,
        'aggregates': aggregates, 'fol_analytics': fol_analytics,
    }

# dc = DataCoreReader('/path/to/datacore_data', file_format='parquet')
# results = run_complete_pipeline(dc, '2010-01-01', '2024-12-31')
```

# 39 Advanced Extensions

## 39.1 Herding Measures

Following Sias (2004), the Lakonishok-Shleifer-Vishny herding measure is:

$$HM_{i,t} = \left| \frac{B_{i,t}}{B_{i,t} + S_{i,t}} - p_t \right| - E\left[ \left| \frac{B_{i,t}}{B_{i,t} + S_{i,t}} - p_t \right| \right] \tag{39.1}$$

where $B_{i,t}$ is the number of managers buying stock $i$ in quarter $t$, $S_{i,t}$ the number selling, and $p_t$ the expected buyer proportion under independent trading.

```
def compute_lsv_herding(
    trades: pd.DataFrame,
    min_traders: int = 5,
) -> pd.DataFrame:
    """Compute LSV herding measure for each stock-quarter."""
    tc = (
        trades.groupby(['ticker', 'rdate'])
        .apply(lambda g: pd.Series({
            'n_buyers': (g['trade'] > 0).sum(),
            'n_sellers': (g['trade'] < 0).sum(),
            'n_traders': len(g),
        }))
        .reset_index()
    )

    tc = tc[tc['n_traders'] >= min_traders].copy()
    tc['buy_prop'] = tc['n_buyers'] / tc['n_traders']
    tc['p_t'] = tc.groupby('rdate')['buy_prop'].transform('mean')
    tc['raw_hm'] = (tc['buy_prop'] - tc['p_t']).abs()

    def expected_abs_deviation(row):
        n = int(row['n_traders'])
        p = row['p_t']
        if n == 0 or p == 0 or p == 1:
```

```
            return 0
        from scipy.stats import binom
        k = np.arange(0, n + 1)
        probs = binom.pmf(k, n, p)
        return np.sum(np.abs(k / n - p) * probs)

    tc['expected_hm'] = tc.apply(expected_abs_deviation, axis=1)
    tc['herding'] = tc['raw_hm'] - tc['expected_hm']

    tc['buy_herding'] = np.where(
        tc['buy_prop'] > tc['p_t'], tc['herding'], np.nan
    )
    tc['sell_herding'] = np.where(
        tc['buy_prop'] < tc['p_t'], tc['herding'], np.nan
    )

    return tc[['ticker', 'rdate', 'n_buyers', 'n_sellers',
               'n_traders', 'herding', 'buy_herding', 'sell_herding']]
```

## 39.2 Demand Persistence

Sias (2004) showed institutional demand is persistent:

$$\rho_t = \text{Corr}\left(\Delta IO_{i,t},\, \Delta IO_{i,t-1}\right) \tag{39.2}$$

```
def compute_demand_persistence(io_ratios: pd.DataFrame) -> pd.DataFrame:
    """Rolling cross-sectional correlation of IO changes."""
    io = io_ratios[['ticker', 'rdate', 'io_total_inst']].copy()
    io = io.sort_values(['ticker', 'rdate'])
    io['dio'] = io.groupby('ticker')['io_total_inst'].diff()
    io['lag_dio'] = io.groupby('ticker')['dio'].shift(1)

    persistence = (
        io.dropna(subset=['dio', 'lag_dio'])
        .groupby('rdate')
        .apply(lambda g: g['dio'].corr(g['lag_dio']))
        .reset_index()
        .rename(columns={0: 'persistence'})
    )
```

```
    persistence = persistence.sort_values('rdate')
    persistence['persistence_ma'] = (
        persistence['persistence'].rolling(window=20, min_periods=4).mean()
    )
    return persistence
```

```
def plot_demand_persistence(persistence: pd.DataFrame):
    fig, ax = plt.subplots(figsize=(12, 5))
    ax.bar(persistence['rdate'], persistence['persistence'],
           width=80, alpha=0.3, color='#1f77b4', label='Quarterly')
    ax.plot(persistence['rdate'], persistence['persistence_ma'],
            color='#d62728', linewidth=2, label='Rolling Average')
    ax.axhline(y=0, color='black', linewidth=0.5)
    ax.set_title('Persistence of Institutional Demand', fontweight='bold')
    ax.set_ylabel('Cross-Sectional Correlation')
    ax.legend()
    plt.tight_layout()
    plt.show()
```

Figure 39.1

## 39.3 Information Content of Trades

Following Alexander, Cici, and Gibson (2007), the InfoTrade ratio measures the proportion of dollar trading from entry/exit decisions vs. position adjustments:

$$\text{InfoTrade}_{i,t} = \frac{\sum_{j:BS\in\{+1,-1\}} |\Delta h_{j,i,t}| \cdot P_{i,t}}{\sum_j |\Delta h_{j,i,t}| \cdot P_{i,t}} \tag{39.3}$$

```
def compute_info_trade_ratio(
    trades: pd.DataFrame, price_q: pd.DataFrame
) -> pd.DataFrame:
    """Compute info trade ratio for each stock-quarter."""
    _t = trades.merge(
        price_q[['ticker', 'qdate', 'p']],
        left_on=['ticker', 'rdate'],
        right_on=['ticker', 'qdate'],
        how='inner'
```

```
    )
    _t['dollar_trade'] = _t['trade'].abs() * _t['p'] / 1e6
    _t['is_discrete'] = _t['buysale'].isin([1, -1])

    info = _t.groupby(['ticker', 'rdate']).apply(
        lambda g: pd.Series({
            'discrete_vol': g.loc[g['is_discrete'], 'dollar_trade'].sum(),
            'total_vol': g['dollar_trade'].sum(),
        })
    ).reset_index()

    info['info_trade_ratio'] = (
        info['discrete_vol'] / info['total_vol']
    ).clip(0, 1)
    return info
```

# 40 Empirical Applications

## 40.1 Application 1: Institutional Ownership Changes and Future Returns

We test whether changes in institutional ownership predict future stock returns (H.-L. Chen, Jegadeesh, and Wermers 2000) via Fama-MacBeth regressions:

$$r_{i,t+1} = \alpha_t + \beta_{1,t} \cdot \Delta IO_{i,t} + \beta_{2,t} \cdot \Delta \text{Breadth}_{i,t} + \gamma_t \cdot X_{i,t} + \varepsilon_{i,t} \tag{40.1}$$

```python
def fama_macbeth_io_returns(
    io_ratios: pd.DataFrame,
    breadth: pd.DataFrame,
    price_q: pd.DataFrame,
) -> pd.DataFrame:
    """Run Fama-MacBeth regressions of future returns on IO changes."""
    panel = io_ratios[['ticker', 'rdate', 'io_total_inst']].merge(
        breadth[['ticker', 'rdate', 'n_total_inst', 'd_n_total_inst']],
        on=['ticker', 'rdate'], how='inner'
    ).merge(
        price_q[['ticker', 'qdate', 'mcap', 'qret']],
        left_on=['ticker', 'rdate'],
        right_on=['ticker', 'qdate'],
        how='inner'
    )

    panel = panel.sort_values(['ticker', 'rdate'])
    panel['dio'] = panel.groupby('ticker')['io_total_inst'].diff()
    panel['log_mcap'] = np.log(panel['mcap'] + 1)
    panel['mom'] = panel.groupby('ticker')['qret'].shift(1)

    reg_vars = ['qret', 'dio', 'd_n_total_inst', 'log_mcap', 'mom']
    panel = panel.dropna(subset=reg_vars)

    quarters = sorted(panel['rdate'].unique())
```

```
    results = []

    for q in quarters:
        qdata = panel[panel['rdate'] == q]
        if len(qdata) < 30:
            continue
        X = sm.add_constant(
            qdata[['dio', 'd_n_total_inst', 'log_mcap', 'mom']]
        )
        try:
            model = sm.OLS(qdata['qret'], X).fit()
            coefs = model.params.to_dict()
            coefs['rdate'] = q
            coefs['n_obs'] = len(qdata)
            results.append(coefs)
        except Exception:
            continue

    fm = pd.DataFrame(results)

    # Time-series averages with Newey-West t-statistics
    print("\nFama-MacBeth Results:")
    print("=" * 50)
    for var in ['const', 'dio', 'd_n_total_inst', 'log_mcap', 'mom']:
        coefs = fm[var].dropna()
        mean_c = coefs.mean()
        nw_se = sm.OLS(
            coefs - mean_c, np.ones(len(coefs))
        ).fit(cov_type='HAC', cov_kwds={'maxlags': 4}).bse[0]
        t = mean_c / nw_se if nw_se > 0 else np.nan
        print(f"  {var:20s}: coef={mean_c:8.4f}, t={t:6.2f}")

    return fm
```

## 40.2 Application 2: Turnover and Performance

Yan (2008) documented a positive turnover-performance relationship. We test in Vietnam:

$$\alpha_{j,t} = a + b \cdot \text{Turnover}_{j,t-1} + c \cdot \log(A_{j,t-1}) + d \cdot \text{Flow}_{j,t} + \varepsilon_{j,t} \tag{40.2}$$

```python
def turnover_performance_regression(
    aggregates: pd.DataFrame,
) -> dict:
    """Test turnover-performance relationship."""
    agg = aggregates.sort_values(['shareholder_name', 'rdate']).copy()
    agg['lag_turnover1'] = (
        agg.groupby('shareholder_name')['turnover1'].shift(1)
    )
    agg['log_assets'] = np.log(agg['assets'] + 1)
    agg['flow_ratio'] = agg['netflows'] / agg['assets'].shift(1)

    panel = agg.dropna(
        subset=['pret', 'lag_turnover1', 'log_assets', 'flow_ratio']
    )

    for col in ['pret', 'lag_turnover1', 'flow_ratio']:
        lo, hi = panel[col].quantile([0.01, 0.99])
        panel[col] = panel[col].clip(lo, hi)

    X = sm.add_constant(
        panel[['lag_turnover1', 'log_assets', 'flow_ratio']]
    )
    model = sm.OLS(panel['pret'], X).fit(
        cov_type='cluster',
        cov_kwds={'groups': panel['shareholder_name']}
    )
    return {'model': model, 'n': len(panel)}
```

## 40.3 Application 3: Foreign vs. Domestic Trading

```python
def compare_foreign_domestic(
    trades: pd.DataFrame, price_q: pd.DataFrame,
) -> pd.DataFrame:
    """Compare trading patterns between foreign and domestic institutions."""
    _t = trades.merge(
        price_q[['ticker', 'qdate', 'p']],
        left_on=['ticker', 'rdate'],
        right_on=['ticker', 'qdate'],
        how='inner'
```

```
    )
    _t['dollar_trade'] = _t['trade'] * _t['p'] / 1e6
    _t['is_buy'] = _t['trade'] > 0

    return (
        _t[_t['owner_type'].isin(OwnershipType.INSTITUTIONAL)]
        .groupby('owner_type')
        .agg(
            n_trades=('trade', 'count'),
            n_buys=('is_buy', 'sum'),
            avg_dollar=('dollar_trade', lambda x: x.abs().mean()),
            net_buying=('dollar_trade', 'sum'),
            pct_initiating=('buysale', lambda x: (x.abs() == 1).mean()),
        )
        .reset_index()
    )
```

```python
def plot_cumulative_net_buying(
    trades: pd.DataFrame, price_q: pd.DataFrame
):
    _t = trades.merge(
        price_q[['ticker', 'qdate', 'p']],
        left_on=['ticker', 'rdate'],
        right_on=['ticker', 'qdate'],
        how='inner'
    )
    _t['trade_vnd'] = _t['trade'] * _t['p'] / 1e9

    inst = _t[_t['owner_type'].isin(OwnershipType.INSTITUTIONAL)]
    net = (
        inst.groupby(
            [pd.Grouper(key='rdate', freq='QE'), 'owner_type']
        )['trade_vnd'].sum().unstack(fill_value=0)
    )
    cum = net.cumsum()

    fig, ax = plt.subplots(figsize=(12, 6))
    for col in cum.columns:
        ax.plot(cum.index, cum[col], label=col,
                color=OWNER_COLORS.get(col, '#333'), linewidth=2)
    ax.axhline(y=0, color='black', linewidth=0.5)
    ax.set_title('Cumulative Net Institutional Buying', fontweight='bold')
    ax.set_ylabel('Billions VND')
    ax.legend(loc='best')
    plt.tight_layout()
    plt.show()
```

Figure 40.1

# 41 Data Quality and Robustness

## 41.1 Common Pitfalls

### 41.1.1 Corporate Action Misadjustment

Example: Phantom Trade from Unadjusted Stock Dividend

Vinamilk (VNM) issues a 20% stock dividend with ex-date March 15, 2023.

- Q4 2022: Fund X holds 1,000,000 shares of VNM
- Q1 2023: Fund X holds 1,200,000 shares of VNM

**Without adjustment:** Inferred buy of +200,000 shares (BS = +2) **With adjustment:** Prior holdings become 1,200,000 adjusted shares, trade = 0
This phantom trade inflates measured turnover and creates spurious buying signals.

### 41.1.2 Disclosure Timing Mismatches

Vietnamese ownership disclosure dates may not align with calendar quarter ends. Our pipeline addresses this by aligning all disclosures to the nearest quarter-end.

### 41.1.3 Name Changes and Entity Mergers

Vietnamese institutions frequently rename. Without a stable identifier, the same entity may appear as two different shareholders, creating phantom entries/exits. We recommend maintaining a master entity mapping table.

## 41.2 Validation Checks

```python
def validate_pipeline_outputs(
    results: Dict[str, pd.DataFrame],
) -> pd.DataFrame:
    """Run comprehensive validation on pipeline outputs."""
    checks = []
    h = results['holdings']
    t = results['trades']
    a = results['aggregates']

    checks.append({
        'Check': 'No negative adjusted shares',
        'Result': 'PASS' if (h['shares_adj'] < 0).sum() == 0 else 'FAIL',
        'Detail': f'{(h["shares_adj"] < 0).sum()} negative obs'
    })

    checks.append({
        'Check': 'No duplicate holdings',
        'Result': 'PASS' if h.duplicated(
            subset=['shareholder_name', 'ticker', 'rdate']
        ).sum() == 0 else 'FAIL',
    })

    checks.append({
        'Check': 'Valid buysale codes only',
        'Result': 'PASS' if t['buysale'].isin([1, 2, -1, -2]).all()
        else 'FAIL',
    })

    checks.append({
        'Check': 'No zero trades',
        'Result': 'PASS' if (t['trade'] == 0).sum() == 0 else 'FAIL',
    })

    t1 = a['turnover1'].dropna()
    checks.append({
        'Check': 'Turnover1 in [0, 10]',
        'Result': 'PASS' if ((t1 < 0) | (t1 > 10)).sum() == 0
        else 'WARNING',
        'Detail': f'{((t1<0)|(t1>10)).sum()} extreme values'
    })

    first_rpt = a[a['first_report']]
```

```
    checks.append({
        'Check': 'First report -> missing netflows',
        'Result': 'PASS' if first_rpt['netflows'].isna().all()
        else 'FAIL',
    })

    return pd.DataFrame(checks)

# validate_pipeline_outputs(results)
```

# 42 Summary

This chapter developed a framework for computing institutional trades, flows, and turnover ratios in the Vietnamese equity market. The key contributions include:

1. **Corporate action adjustment** for Vietnam's frequent stock dividends and bonus shares, preventing phantom trades that contaminate standard differencing.
2. **Four-way ownership taxonomy** (state, foreign institutional, domestic institutional, individual) capturing Vietnam's unique ownership landscape.
3. **FOL utilization analytics** for studying foreign ownership constraints absent from developed markets.
4. **Irregular disclosure handling** with correct gap splitting into terminating sales and initiating buys.
5. **Advanced extensions** including herding, demand persistence, and information content decomposition.

The pipeline produces several output datasets (Table 42.1)

Table 42.1: Summary of Pipeline Output Datasets

| Output | Grain | Key Variables | Use Cases |
|---|---|---|---|
| `holdings` | Shareholder x Ticker x Quarter | `shares_adj`, `owner_type` | Cross-sectional ownership |
| `io_ratios` | Ticker x Quarter | `io_state`, `io_foreign`, etc. | Governance, liquidity |
| `trades` | Shareholder x Ticker x Quarter | `trade`, `buysale` | Informed trading, herding |
| `aggregates` | Shareholder x Quarter | `assets`, `turnover`, `netflows` | Fund performance, flows |
| `fol_analytics` | Ticker x Date | `fol_utilization`, `foreign_room` | FOL premium, foreign investment |

# 43 Conclusion

Empirical finance in emerging and frontier markets is often judged less by the elegance of an estimator than by the credibility of its inputs and the transparency of its decisions. Vietnam makes this point vividly: trading venues and regulatory regimes have evolved quickly, firm coverage can be uneven across time, corporate actions need careful treatment, and accounting conventions require close attention to timing and comparability. Those features do not prevent high-quality research; they simply shift the center of gravity toward *reproducible data engineering*, *auditable transformations*, and *clear identification of assumptions*.

## 43.1 What you should take away

### 43.1.1 Reproducibility is an identification strategy

In textbook settings, identification focuses on variation and exogeneity. In real-world market data, identification also depends on whether your dataset is *the same dataset* when you rerun the work next month or next year. The practical discipline of versioned inputs, deterministic transformations, and documented filters reduces the scope for accidental $p$-hacking and silent sample drift (e.g., survivorship bias from symbol changes or late-arriving delistings). Reproducible workflows are not administrative overhead; they are a commitment device that makes results more trustworthy and easier to challenge constructively (Peng 2011; Sandve et al. 2013).

### 43.1.2 Vietnam rewards "microstructure humility"

The chapters on returns, beta estimation, and factor construction emphasized that naïve carryover of developed-market defaults can be costly. Thin trading, price limits, lot-size rules, and regime changes mean that decisions like (i) return interval, (ii) stale-price handling, (iii) corporate-action adjustment, and (iv) portfolio formation frequency can materially change inference. This is not a Vietnam-only phenomenon, but it is more visible there, and therefore a useful laboratory for best practices in emerging markets.

## 43.2 A reproducibility checklist you can actually use

The list below is designed to be operational: each item can be verified in a repository review.

Table 43.1: Reproducibility deliverables for research

| Deliverable | What "done" looks like | Where it lives |
|---|---|---|
| Deterministic transforms | Same raw inputs yield identical normalized outputs | `R/transform_*.R` (or `python/transform_*.py`) |
| Test suite | Coverage, identity, and corporate-action tests run in CI | `tests/` + CI config |
| Data dictionary | Tables/fields documented with units, timing, and keys | `docs/dictionary.qmd` |
| Research log | All key design choices recorded (filters, winsorization, periods) | `notes/research_log.md` |
| Artifact registry | Every figure/table has a script and a checksum | `artifacts/manifest.json` |

# 44 Closing perspective

Vietnam is not "hard mode" finance; it is *real mode* finance. The market's growth, institutional evolution, and data idiosyncrasies force the habits that modern empirical finance increasingly requires everywhere: transparent datasets, careful treatment of identities and corporate actions, and codebases that can be rerun and audited.

# References


Abarbanell, Jeffery S, William N Lanen, and Robert E Verrecchia. 1995. "Analysts' Forecasts as Proxies for Investor Beliefs in Empirical Research." *Journal of Accounting and Economics* 20 (1): 31–60.

Aggarwal, Reena, Isil Erel, Miguel Ferreira, and Pedro Matos. 2011. "Does Governance Travel Around the World? Evidence from Institutional Investors." *Journal of Financial Economics* 100 (1): 154–81.

Aharony, Joseph, and Itzhak Swary. 1980. "Quarterly Dividend and Earnings Announcements and Stockholders' Returns: An Empirical Analysis." *The Journal of Finance* 35 (1): 1–12.

Alexander, Gordon J, Gjergji Cici, and Scott Gibson. 2007. "Does Motivation Matter When Assessing Trade Performance? An Analysis of Mutual Funds." *The Review of Financial Studies* 20 (1): 125–50.

Alexandridis, George, Antonios Antoniou, and Dimitris Petmezas. 2007. "Divergence of Opinion and Post-Acquisition Performance." *Journal of Business Finance & Accounting* 34 (3-4): 439–60.

Amihud, Yakov. 2002. "Illiquidity and Stock Returns: Cross-Section and Time-Series Effects." *Journal of Financial Markets* 5 (1): 31–56.

Anderson, Anne-Marie, and Edward A Dyl. 2005. "Market Structure and Trading Volume." *Journal of Financial Research* 28 (1): 115–31.

Anderson, Kirsten L, Jeffrey H Harris, and Eric C So. 2007. "Opinion Divergence and Post-Earnings Announcement Drift." *Available at SSRN 969736*.

Andrade, Gregor, Mark Mitchell, and Erik Stafford. 2001. "New Evidence and Perspectives on Mergers." *Journal of Economic Perspectives* 15 (2): 103–20.

Bali, Turan G, Robert F Engle, and Scott Murray. 2016. *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons. https://doi.org/10.1002/9781118445112.stat07954.

Ball, Ray, and Philip Brown. 2013. "An Empirical Evaluation of Accounting Income Numbers." In *Financial Accounting and Equity Markets*, 27–46. Routledge.

Bao Dinh, Ngoc, and Van Nguyen Hong Tran. 2024. "Institutional Ownership and Stock Liquidity: Evidence from an Emerging Market." *SAGE Open* 14 (1): 21582440241239116.

Barber, Brad M, and John D Lyon. 1997. "Detecting Long-Run Abnormal Stock Returns: The Empirical Power and Specification of Test Statistics." *Journal of Financial Economics* 43 (3): 341–72.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny. 1998. "A Model of Investor Sentiment." *Journal of Financial Economics* 49 (3): 307–43.

Ben-David, ITZHAK, Francesco Franzoni, Augustin Landier, and Rabih Moussawi. 2013. "Do Hedge Funds Manipulate Stock Prices?" *The Journal of Finance* 68 (6): 2383–2434.

Berkman, Henk, Valentin Dimitrov, Prem C Jain, Paul D Koch, and Sheri Tice. 2009. "Sell on the News: Differences of Opinion, Short-Sales Constraints, and Returns Around Earnings Announcements." *Journal of Financial Economics* 92 (3): 376–99.

Bernard, Victor L, and Jacob K Thomas. 1989. "Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium?" *Journal of Accounting Research* 27: 1–36.

Bhattacharya, Utpal, Hazem Daouk, Brian Jorgenson, and Carl-Heinrich Kehr. 2000. "When an Event Is Not an Event: The Curious Case of an Emerging Market." *Journal of Financial Economics* 55 (1): 69–101.

Binder, John. 1998. "The Event Study Methodology Since 1969." *Review of Quantitative Finance and Accounting* 11 (2): 111–37.

Boehme, Rodney D, Bartley R Danielsen, and Sorin M Sorescu. 2006. "Short-Sale Constraints, Differences of Opinion, and Overvaluation." *Journal of Financial and Quantitative Analysis* 41 (2): 455–87.

Boehmer, Ekkehart, Jim Masumeci, and Annette B Poulsen. 1991. "Event-Study Methodology Under Conditions of Event-Induced Variance." *Journal of Financial Economics* 30 (2): 253–72.

Brown, Stephen J, and Jerold B Warner. 1980. "Measuring Security Price Performance." *Journal of Financial Economics* 8 (3): 205–58.

———. 1985. "Using Daily Stock Returns: The Case of Event Studies." *Journal of Financial Economics* 14 (1): 3–31.

Campbell, John Y, Andrew W Lo, A Craig MacKinlay, and Robert F Whitelaw. 1998. "The Econometrics of Financial Markets." *Macroeconomic Dynamics* 2 (4): 559–62.

Carhart, Mark M. 1997a. "On Persistence in Mutual Fund Performance." *The Journal of Finance* 52 (1): 57–82.

Carhart, Mark M. 1997b. "On persistence in mutual fund performance." *The Journal of Finance* 52 (1): 57–82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x.

Chan, Kalok, Allaudeen Hameed, and Wilson Tong. 2000. "Profitability of Momentum Strategies in the International Equity Markets." *Journal of Financial and Quantitative Analysis*, 153–72.

Chatterjee, Sris, Kose John, and An Yan. 2012. "Takeovers and Divergence of Investor Opinion." *The Review of Financial Studies* 25 (1): 227–77.

Chen, Hsiu-Lang, Narasimhan Jegadeesh, and Russ Wermers. 2000. "The Value of Active Mutual Fund Management: An Examination of the Stockholdings and Trades of Fund Managers." *Journal of Financial and Quantitative Analysis* 35 (3): 343–68.

Chen, Joseph, Harrison Hong, and Jeremy C Stein. 2002. "Breadth of Ownership and Stock Returns." *Journal of Financial Economics* 66 (2-3): 171–205.

Cheong, Foong Soon, and Jacob Thomas. 2011. "Why Do EPS Forecast Error and Dispersion Not Vary with Scale? Implications for Analyst and Managerial Behavior." *Journal of Accounting Research* 49 (2): 359–401.

Chui, Andy CW, Sheridan Titman, and KC John Wei. 2010. "Individualism and Momentum Around the World." *The Journal of Finance* 65 (1): 361–92.

Chung, Kee H, and Hao Zhang. 2014. "A Simple Approximation of Intraday Spreads Using Daily Data." *Journal of Financial Markets* 17: 94–120.

Cooper, Michael J, Roberto C Gutierrez Jr, and Allaudeen Hameed. 2004. "Market States and Momentum." *The Journal of Finance* 59 (3): 1345–65.

Corrado, Charles J. 1989. "A Nonparametric Test for Abnormal Security-Price Performance in Event Studies." *Journal of Financial Economics* 23 (2): 385–95.

Coval, Joshua, and Erik Stafford. 2007. "Asset Fire Sales (and Purchases) in Equity Markets." *Journal of Financial Economics* 86 (2): 479–512.

Cowan, Arnold Richard. 1992. "Nonparametric Event Study Tests." *Review of Quantitative Finance and Accounting* 2 (4): 343–58.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam. 1998. "Investor Psychology and Security Market Under-and Overreactions." *The Journal of Finance* 53 (6): 1839–85.

Daniel, Kent, and Tobias J Moskowitz. 2016. "Momentum Crashes." *Journal of Financial Economics* 122 (2): 221–47.

Diether, Karl B, Christopher J Malloy, and Anna Scherbina. 2002. "Differences of Opinion and the Cross Section of Stock Returns." *The Journal of Finance* 57 (5): 2113–41.

Dimson, Elroy. 1979. "Risk Measurement When Shares Are Subject to Infrequent Trading." *Journal of Financial Economics* 7 (2): 197–226.

Doukas, John A, Chansog Francis Kim, and Christos Pantzalis. 2006. "Divergence of Opinion and Equity Returns." *Journal of Financial and Quantitative Analysis* 41 (3): 573–606.

Doukas, John A, Chansog Kim, and Christos Pantzalis. 2004. "Divergent Opinions and the Performance of Value Stocks." *Financial Analysts Journal* 60 (6): 55–64.

Fama, Eugene F. 1998. "Market Efficiency, Long-Term Returns, and Behavioral Finance." *Journal of Financial Economics* 49 (3): 283–306.

Fama, Eugene F, Lawrence Fisher, Michael C Jensen, and Richard Roll. 1969. "The Adjustment of Stock Prices to New Information." *International Economic Review* 10 (1): 1–21.

Fama, Eugene F., and Kenneth R. French. 1992. "The cross-section of expected stock returns." *The Journal of Finance* 47 (2): 427–65. https://doi.org/2329112.

———. 1993a. "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics* 33 (1): 3–56. https://doi.org/10.1016/0304-405X(93)90023-5.

———. 1993b. "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics* 33 (1): 3–56. https://doi.org/10.1016/0304-405X(93)90023-5.

———. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116 (1): 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010.

Fama, Eugene F, and Kenneth R French. 1996. "Multifactor Explanations of Asset Pricing Anomalies." *The Journal of Finance* 51 (1): 55–84.

Fama, Eugene F., and James D. MacBeth. 1973. "Risk, return, and equilibrium: Empirical tests." *Journal of Political Economy* 81 (3): 607–36. https://doi.org/10.1086/260061.

Flannery, Mark J, and Aris A Protopapadakis. 2002. "Macroeconomic Factors Do Influence Aggregate Stock Returns." *The Review of Financial Studies* 15 (3): 751–82.

Frazzini, Andrea, and Lasse Heje Pedersen. 2014. "Betting against beta." *Journal of Financial Economics* 111 (1): 1–25. https://doi.org/10.1016/j.jfineco.2013.10.005.

Garfinkel, Jon A. 2009. "Measuring Investors' Opinion Divergence." *Journal of Accounting*

*Research* 47 (5): 1317–48.

Garfinkel, Jon A, and Jonathan Sokobin. 2006. "Volume, Opinion Divergence, and Returns: A Study of Post–Earnings Announcement Drift." *Journal of Accounting Research* 44 (1): 85–112.

Gentzkow, Matthew, and Jesse M Shapiro. 2014. "Code and Data for the Social Sciences: A Practitioner's Guide." Working Paper, University of Chicago.

Gompers, Paul, Joy Ishii, and Andrew Metrick. 2003. "Corporate Governance and Equity Prices." *The Quarterly Journal of Economics* 118 (1): 107–56.

Griffin, John M, Patrick J Kelly, and Federico Nardari. 2010. "Do Market Efficiency Measures Yield Correct Inferences? A Comparison of Developed and Emerging Markets." *The Review of Financial Studies* 23 (8): 3225–77.

Grinblatt, Mark, Sheridan Titman, and Russ Wermers. 1995. "Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior." *American Economic Review* 85 (5): 1088–1105.

Grundy, Bruce D, and J Spencer Martin Martin. 2001. "Understanding the Nature of the Risks and the Source of the Rewards to Momentum Investing." *The Review of Financial Studies* 14 (1): 29–78.

Hall, Peter. 1992. "On the Removal of Skewness by Transformation." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 54 (1): 221–28.

Handa, Puneet, Robert Schwartz, and Ashish Tiwari. 2003. "Quote Setting and Price Formation in an Order Driven Market." *Journal of Financial Markets* 6 (4): 461–89.

Harris, Milton, and Artur Raviv. 1993. "Differences of Opinion Make a Horse Race." *The Review of Financial Studies* 6 (3): 473–506.

Hofstede, Geert. 2001. "Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations." *International Educational and Professional*.

Hong, Harrison, and Jeremy C Stein. 1999. "A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets." *The Journal of Finance* 54 (6): 2143–84.

———. 2003. "Differences of Opinion, Short-Sales Constraints, and Market Crashes." *The Review of Financial Studies* 16 (2): 487–525.

Hou, Kewei, Chen Xue, and Lu Zhang. 2014. "Digesting anomalies: An investment approach." *Review of Financial Studies* 28 (3): 650–705. https://doi.org/10.1093/rfs/hhu068.

———. 2020. "Replicating anomalies." *Review of Financial Studies* 33 (5): 2019–2133. https://doi.org/10.1093/rfs/hhy131.

Houge, Todd, Tim Loughran, Gerry Suchanek, and Xuemin Yan. 2001. "Divergence of Opinion, Uncertainty, and the Quality of Initial Public Offerings." *Financial Management*, 5–23.

Huang, Xiangqian, Clark Liu, and Tao Shu. 2023. "Factors and Anomalies in the Vietnamese Stock Market." *Pacific-Basin Finance Journal* 82: 102176.

Jagannathan, Ravi, and Zhenyu Wang. 1996. "The conditional CAPM and the cross-section of expected returns." *The Journal of Finance* 51 (1): 3–53. https://doi.org/10.2307/2329301.

Jegadeesh, Narasimhan. 1990. "Evidence of Predictable Behavior of Security Returns." *The Journal of Finance* 45 (3): 881–98.

Jegadeesh, Narasimhan, and Sheridan Titman. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *The Journal of Finance* 48 (1): 65–91.

Jensen, Michael C, and Richard S Ruback. 1983. "The Market for Corporate Control: The Scientific Evidence." *Journal of Financial Economics* 11 (1-4): 5–50.

Johnson, Timothy C. 2002. "Rational Momentum Effects." *The Journal of Finance* 57 (2): 585–608.

Kandel, Eugene, and Neil D Pearson. 1995. "Differential Interpretation of Public Signals and Trade in Speculative Markets." *Journal of Political Economy* 103 (4): 831–72.

Kolari, James W, and Seppo Pynnönen. 2010. "Event Study Testing with Cross-Sectional Correlation of Abnormal Returns." *The Review of Financial Studies* 23 (11): 3996–4025.

Korajczyk, Robert A, and Ronnie Sadka. 2004. "Are Momentum Profits Robust to Trading Costs?" *The Journal of Finance* 59 (3): 1039–82.

Kothari, Sagar P, and Jerold B Warner. 2007. "Econometrics of Event Studies." In *Handbook of Empirical Corporate Finance*, 3–36. Elsevier.

Lehavy, Reuven, and Richard G Sloan. 2008. "Investor Recognition and Stock Returns." *Review of Accounting Studies* 13 (2): 327–61.

Leibowitz, Martin L. 2002. "The Levered p/e Ratio." *Financial Analysts Journal* 58 (6): 68–77.

Lesmond, David A, Michael J Schill, and Chunsheng Zhou. 2004. "The Illusory Nature of Momentum Profits." *Journal of Financial Economics* 71 (2): 349–80.

Lintner, John. 1965. "Security prices, risk, and maximal gains from diversification." *The Journal of Finance* 20 (4): 587–615. https://doi.org/10.1111/j.1540-6261.1965.tb02930.x.

Livnat, Joshua, and Richard R Mendenhall. 2006. "Comparing the Post–Earnings Announcement Drift for Surprises Calculated from Analyst and Time Series Forecasts." *Journal of Accounting Research* 44 (1): 177–205.

MacKinlay, A Craig. 1997. "Event Studies in Economics and Finance." *Journal of Economic Literature* 35 (1): 13–39.

Markowitz, Harry. 1952. "Portfolio selection." *The Journal of Finance* 7 (1): 77–91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x.

Milgrom, Paul, and Nancy Stokey. 1982. "Information, Trade and Common Knowledge." *Journal of Economic Theory* 26 (1): 17–27.

Miller, Edward M. 1977. "Risk, Uncertainty, and Divergence of Opinion." *The Journal of Finance* 32 (4): 1151–68.

Mitchell, Mark L, and Jeffry M Netter. 1993. "The Role of Financial Economics in Securities Fraud Cases: Applications at the Securities and Exchange Commission." *Bus. Law.* 49: 545.

Mitchell, Mark L, and Erik Stafford. 2000. "Managerial Decisions and Long-Term Stock Price Performance." *The Journal of Business* 73 (3): 287–329.

Mossin, Jan. 1966. "Equilibrium in a capital asset market." *Econometrica* 34 (4): 768–83. https://doi.org/10.2307/1910098.

Newey, Whitney K., and Kenneth D. West. 1987. "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance Matrix." *Econometrica* 55 (3): 703–8. http://www.jstor.org/stable/1913610.

Nguyen, Du D, and Minh C Pham. 2018. "Search-Based Sentiment and Stock Market Reactions: An Empirical Evidence in Vietnam." *The Journal of Asian Finance, Economics and Business* 5 (4): 45–56.

Pástor, L'uboš, and Robert F Stambaugh. 2003. "Liquidity Risk and Expected Stock Returns." *Journal of Political Economy* 111 (3): 642–85.

Patell, James M. 1976. "Corporate Forecasts of Earnings Per Share and Stock Price Behavior: Empirical Test." *Journal of Accounting Research*, 246–76.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science* 334 (6060): 1226–27.

Phan, Thi Nha Truc, Philippe Bertrand, Hong Hai Phan, and Xuan Vinh Vo. 2023. "The Role of Investor Behavior in Emerging Stock Markets: Evidence from Vietnam." *The Quarterly Review of Economics and Finance* 87: 367–76.

Phung, Duc Nam, and Anil V Mishra. 2016. "Ownership Structure and Firm Performance: Evidence from Vietnamese Listed Firms." *Australian Economic Papers* 55 (1): 63–98.

Rouwenhorst, K Geert. 1998. "International Momentum Strategies." *The Journal of Finance* 53 (1): 267–84.

Sandve, Geir Kjetil, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. "Ten Simple Rules for Reproducible Computational Research." *PLoS Computational Biology* 9 (10): e1003285.

Scheinkman, Jose A, and Wei Xiong. 2003. "Overconfidence and Speculative Bubbles." *Journal of Political Economy* 111 (6): 1183–1220.

Scheuch, Christoph, Stefan Voigt, and Patrick Weiss. 2023. *Tidy Finance with r*. CRC Press.

Scheuch, Christoph, Stefan Voigt, Patrick Weiss, and Christoph Frey. 2024. *Tidy Finance with Python*. Chapman; Hall/CRC.

Scholes, Myron, and Joseph Williams. 1977. "Estimating Betas from Nonsynchronous Data." *Journal of Financial Economics* 5 (3): 309–27.

Schwert, G William. 1981. "Using Financial Data to Measure Effects of Regulation." *The Journal of Law and Economics* 24 (1): 121–58.

Shalen, Catherine T. 1993. "Volume, Volatility, and the Dispersion of Beliefs." *The Review of Financial Studies* 6 (2): 405–34.

Sharpe, William F. 1964. "Capital asset prices: A theory of market equilibrium under conditions of risk ." *The Journal of Finance* 19 (3): 425–42. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x.

Sias, Richard W. 2004. "Institutional Herding." *The Review of Financial Studies* 17 (1): 165–206.

Sirri, Erik R, and Peter Tufano. 1998. "Costly Search and Mutual Fund Flows." *The Journal of Finance* 53 (5): 1589–1622.

Varian, Hal R. 1985. "Divergence of Opinion in Complete Markets: A Note." *The Journal of Finance* 40 (1): 309–17.

Vilhuber, Lars. 2020. "Reproducibility and Replicability in Economics." *Harvard Data Science Review* 2 (4): 1–39.

Vo, Xuan Vinh. 2015. "Foreign Ownership and Stock Return Volatility–Evidence from Vietnam." *Journal of Multinational Financial Management* 30: 101–9.

———. 2017. "Do Foreign Investors Improve Stock Price Informativeness in Emerging Equity Markets? Evidence from Vietnam." *Research in International Business and Finance* 42: 986–91.

Vo, Xuan Vinh, and Dang Bao Anh Phan. 2017. "Further Evidence on the Herd Behavior in Vietnam Stock Market." *Journal of Behavioral and Experimental Finance* 13: 33–41.

Warner, Jerold B, Ross L Watts, and Karen H Wruck. 1988. "Stock Prices and Top Management Changes." *Journal of Financial Economics* 20: 461–92.

Wermers, Russ. 2000. "Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transactions Costs, and Expenses." *The Journal of Finance* 55 (4): 1655–95.

Yan, Xuemin Sterling. 2008. "Liquidity, Investment Style, and the Relation Between Fund Size and Fund Performance." *Journal of Financial and Quantitative Analysis* 43 (3): 741–67.