

# Lesson 24: Factor Models

Data Science with Python – BSc Course

Data Science Program

BSc Course

45 Minutes

# Previously on L21–L23...

## L21 – Linear Regression:

- Fit a line through data;  $\beta$  = sensitivity to one variable

## L22 – Regularization:

- Ridge and Lasso prevent overfitting when predictors multiply

## L23 – Regression Metrics:

- MSE,  $R^2$ , and walk-forward validation measure model quality

**Now we apply ALL of this to the biggest question in finance:**

*“What drives stock returns?”*

---

Module 4 arc: one variable → many variables → measuring quality → factor models

# Learning Objectives

**The Problem:** CAPM uses only market beta, but stocks also respond to size, value, and momentum. How do we capture multiple sources of systematic risk?

**After this lesson, you will be able to:**

- Build multi-factor regression models
- Understand Fama-French factors (SMB, HML)
- Interpret factor loadings and alpha
- Create complete ML pipelines with sklearn

---

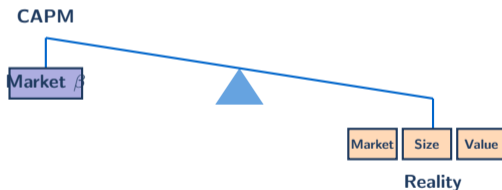
**Finance application: decomposing returns into systematic factors for attribution**

# CAPM's Promise and Problem

CAPM says one number – beta – explains expected returns.  
It won a Nobel Prize. But it's **wrong**.

The anomalies CAPM cannot explain:

- Small-cap stocks outperform large-caps (size effect)
- Cheap “value” stocks outperform expensive “growth” stocks



---

Nobel Prize 1990 (Sharpe). But Fama & French proved CAPM incomplete in 1993

# Excess Returns: The Foundation

## Risk-Free Rate ( $R_f$ ):

- Return with ZERO risk (3-month U.S. Treasury bill,  $\sim 4\text{--}5\%$ )
- Excess Return = Asset Return  $- R_f$
- “How much extra did you earn **for taking risk?**”

## Why excess returns matter:

- Earning 5% when T-bills pay 4% is less impressive than . . .
- . . . earning 5% when T-bills pay 1%
- Excess returns adjust for the “time value of money”

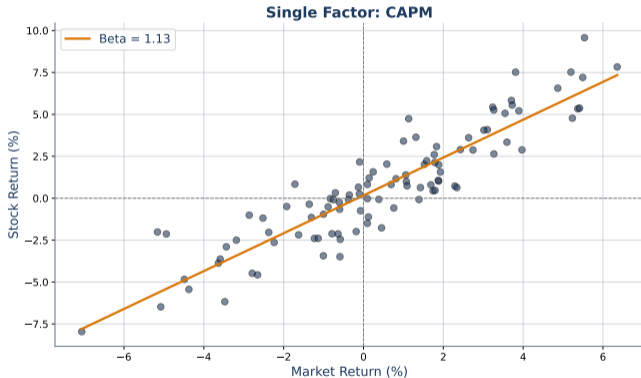
---

Always subtract  $R_f$  when comparing risky investments

# CAPM: One Factor

Market factor alone explains  $\sim 60\%$  of returns.

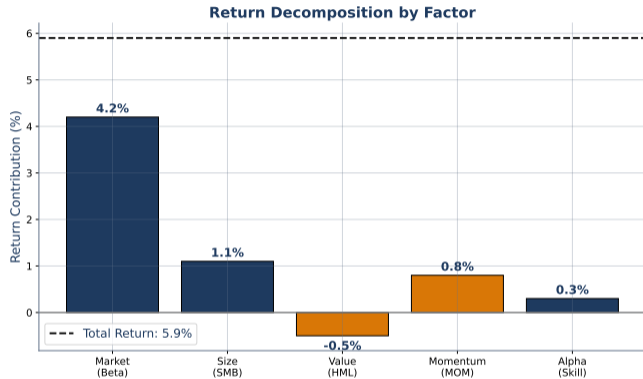
“What about the other 40%?”



---

$$\text{CAPM: } R_i - R_f = \alpha + \beta(R_m - R_f) + \varepsilon$$

# Return Decomposition



Multi-factor models split returns into factor contributions + alpha

# Enter Fama and French (1993)

Eugene Fama & Kenneth French – Nobel Prize 2013

**Their insight:** Two additional factors explain the anomalies.

**What they did:**

- Collected decades of US stock return data
- Sorted stocks into portfolios by **size** and **book-to-market**
- Showed CAPM systematically misprices small and value stocks
- Proposed the **3-Factor Model**: Market + SMB + HML

$$R_i - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 \cdot \text{SMB} + \beta_3 \cdot \text{HML} + \varepsilon$$

---

Fama & French (1993): “Common Risk Factors in the Returns on Stocks and Bonds”

# SMB: Small Minus Big

## SMB = Small Minus Big (Size Factor)

- Construction: return of small-cap – return of large-cap stocks
- Size premium: small stocks outperform (riskier, less diversified)
- Positive SMB loading = stock behaves like small caps

## Why does the size premium exist?

- **Risk:** Small firms are more volatile, less liquid
- **Behavioral:** Investors overlook small stocks
- Premium has shrunk since discovery (1981) – markets adapted

---

SMB = Small Minus Big. Positive loading = small-cap-like behavior

# HML: High Minus Low

## HML = High Minus Low (Value Factor)

- Construction: return of high book-to-market – low book-to-market
- Positive HML = value-like behavior (cheap stocks)
- Negative HML = growth-like behavior (expensive stocks)

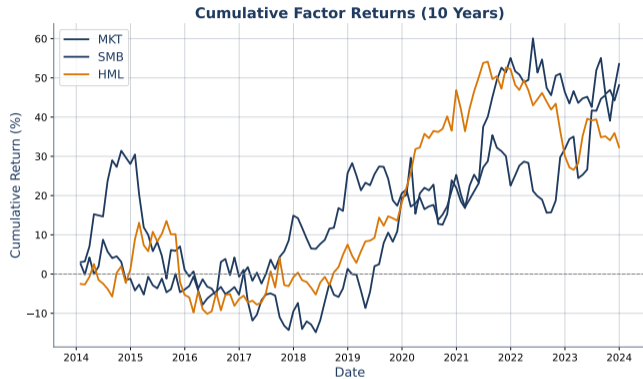
## Why does the value premium exist?

- **Risk:** Value stocks are distressed, may face bankruptcy
- **Behavioral:** Investors overpay for “glamour” growth stocks
- Value struggled 2010–2020 but historically strong

---

HML = High Minus Low (book-to-market). Value stocks have positive HML loading

# Factor Returns Over Time

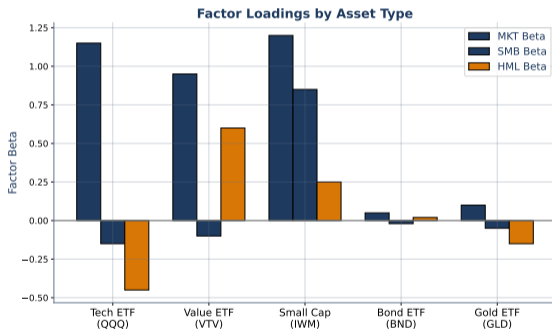


All three premia positive long-term, but with periods of underperformance

# Factor Loadings: Every Stock Has a Fingerprint

Different stocks have different factor exposures:

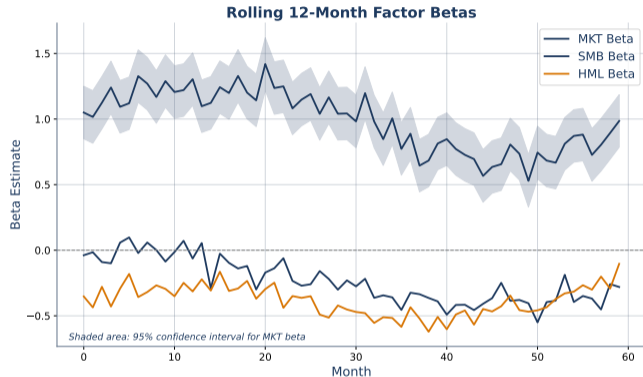
- Tech: high  $\beta$ , negative SMB (large), negative HML (growth)
- Small bank: moderate  $\beta$ , positive SMB, positive HML (value)



---

Factor loadings = the stock's "fingerprint" of systematic risk exposures

# Factor Loadings Change Over Time

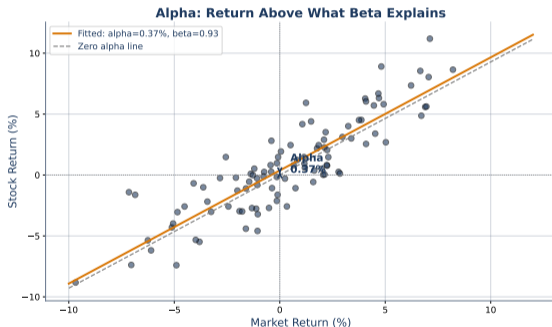


Rolling windows reveal that factor exposures shift as companies evolve

# Alpha: The Holy Grail

Alpha = return unexplained by factors = “skill” (or luck)

- Most alphas cluster near zero after controlling for factors
- True persistent alpha is extremely rare



---

“In search of alpha” – the central quest of active management

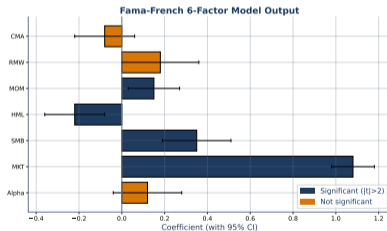
# Multi-Factor Regression in Python

## Code:

- `X = factors[['Mkt-RF', 'SMB', 'HML']]`
- `y = stock_returns - rf` # excess returns!
- `model = LinearRegression().fit(X, y)`

**Interpreting:** each coefficient = effect **holding others constant**.

$\beta_{Mkt}=1.2$ ,  $SMB=0.3$ ,  $HML=-0.1$ : aggressive, small-cap tilt, growth tilt.

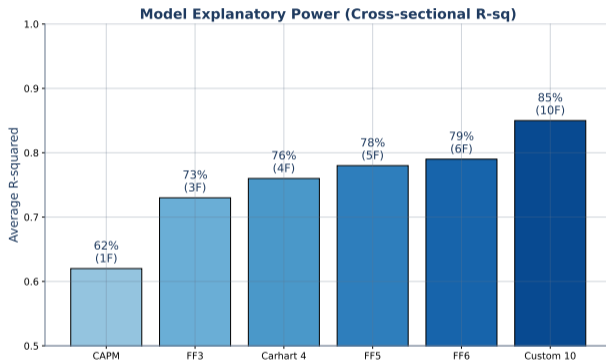


---

Alpha = intercept = return unexplained by factors (“skill”)

# Model Comparison: More Factors = Better?

$R^2$ : CAPM  $\sim 60\%$  vs FF3  $\sim 75\%$ . But check **adjusted**  $R^2$  – diminishing returns from adding factors.



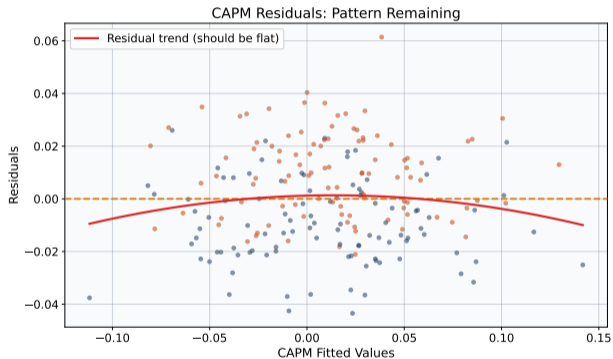
---

Adjusted  $R^2$  penalizes model complexity – more factors is not always better

# CAPM Residuals: The Evidence

CAPM residuals show a systematic pattern:

Small/value stocks consistently under- or over-predicted.



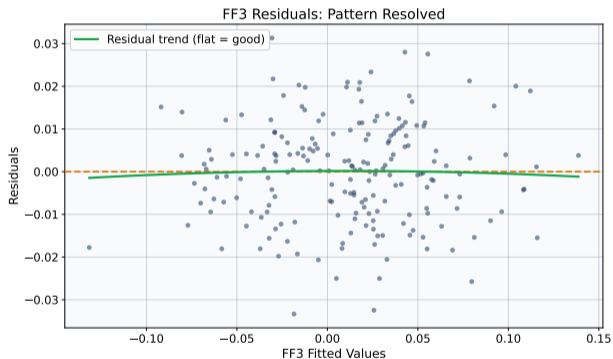
---

Pattern in residuals = model misspecification. CAPM misses size and value effects

# FF3 Residuals: Pattern Resolved

Adding SMB and HML captures what CAPM missed.

Residuals now look random – this is a well-specified model.



---

Flat residual trend = model captures the relevant risk factors

# Checkpoint: Reading Factor Loadings

AAPL has  $\beta_{Mkt}=1.2$ ,  $SMB=-0.3$ ,  $HML=-0.4$

## Quiz yourself:

1. Is Apple value or growth?  
→ Growth (negative HML = growth-like behavior)
2. Large-cap or small-cap behavior?  
→ Large-cap (negative SMB = large-cap-like)
3. Aggressive or defensive?  
→ Aggressive ( $\beta_{Mkt} > 1$ , amplifies market moves)

**Summary:** Apple is a large-cap growth stock with above-average market sensitivity.

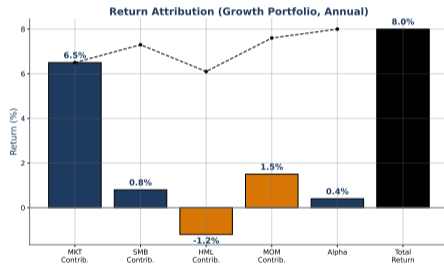
---

Factor loadings tell you what kind of risk a stock carries

# Portfolio Attribution: Worked Example

**Given:**  $\alpha=0.2\%$ ,  $\beta_{Mkt}=1.1$ ,  $\beta_{SMB}=0.3$ ,  $\beta_{HML}=-0.2$ . **Month:** Mkt-RF=2%, SMB=1%, HML=-0.5%.

**Attribution:** Market  $1.1 \times 2\% = 2.2\%$  + Size 0.3% + Value 0.1% + Alpha 0.2% = **2.8%**.  
Most return (2.2%) came from market exposure, not skill.



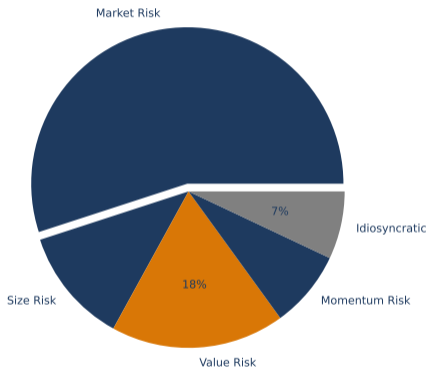
---

**Attribution: Return** =  $\alpha + \sum \beta_j \times F_j$

# Risk Decomposition

Factor risk vs idiosyncratic risk: where does volatility come from?

**Risk Decomposition (Variance Attribution)**



---

Factor risk contribution shows where volatility comes from

# Pipelines: Preventing Data Leakage

## The problem:

- Preprocessing (scaling, imputation) must fit ONLY on training data
- If you scale on ALL data, then split → test info leaks into training

## Pipeline bundles preprocessing + model:

- `Pipeline([('scaler', StandardScaler()),`
- `('model', LinearRegression())])`
- Cross-validation respects this automatically

---

**Pipeline = preprocessing + model in one object. Prevents data leakage**

# Common Mistakes in Factor Models

- ✘ Using **raw returns** instead of excess returns ( $R_i - R_f$ )
- ✘ Not checking **statistical significance** of factor loadings
- ✘ Assuming factor loadings are **constant over time**
- ✘ Confusing alpha with **luck** – need a statistical test!

---

Avoid these pitfalls when building and interpreting factor models

# Hands-On Exercise (25 min)

## Task: Build a Fama-French 3-Factor Model

1. Download FF3 data from Kenneth French's website
2. Merge with your stock returns (AAPL or similar)
3. Fit multi-factor regression: stock excess returns vs [Mkt-RF, SMB, HML]
4. Interpret: What are the factor loadings? Is there significant alpha?
5. Compare  $R^2$  to single-factor CAPM model

**Deliverable:** Table of factor loadings +  $R^2$  comparison.

**Extension:** Add momentum (UMD) as a 4th factor – does  $R^2$  improve?

*Reference:* `inclass/L24_inclass_factor_attribution.ipynb`

---

Use `pandas_datareader` or download CSV from [Kenneth French Data Library](#)

# Module Wrap-Up: The Regression Journey

## Module 4 arc (L21–L24):

- **L21:** One variable → best-fit line, slope = beta
- **L22:** Many variables → regularization prevents overfitting
- **L23:** Measuring quality → MSE,  $R^2$ , walk-forward
- **L24:** Multi-factor finance models → decompose risk and return

*“From Galton’s sweet peas to Fama-French factor models – 150 years of the best-fit line.”*

## Next module preview:

What if  $Y$  is not a number but a **category**?

Welcome to **Classification** (L25–L28).

---

Module 4 complete. Next: L25 Logistic Regression – predicting categories