

# Big Data Analysis

Lecture 14

2020/1/6

# Time Series Regression

---

- Factors are portfolios returns, Fama and French (1993)
- Stocks are sorted and grouped to form long short portfolios
- With periodic rebalance, portfolio returns are achieved
- Let  $f_t$  be factor returns during t period,
$$R_{it} = \alpha_i + \beta_i f_t + \varepsilon_{it}, t = 1, 2, \dots, T$$
- Taking expectation:  $E_T[R_i] = \beta_i E_T[f_t] + \alpha_i$
- Comparing to  $E[R_i] = \beta_i \lambda + \alpha_i$ , we have  $\lambda = E_T[f_t]$

# Time Series Regression – Momentum

- Monthly Momentum Factor
- Total 6 portfolios formed monthly on Size and Momentum
- [https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data\\_Library/det\\_mom\\_factor.html](https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_mom_factor.html)
- $$\text{Mom} = \frac{1}{2}(\text{Small High} + \text{Big High}) - \frac{1}{2}(\text{Small Low} + \text{Big Low})$$
- Why creating 2x3 portfolios?
- Can we limit the stock to be within a certain industrial sector?

**Monthly Returns:** January 1927 - November 2019

**Annual Returns:** 1927 - 2018

**Construction:** The portfolios, which are constructed monthly, are the intersections of 2 portfolios formed on size (market equity, ME) and 3 portfolios formed on prior (2-12) return. The monthly size breakpoint is the median NYSE market equity. The monthly prior (2-12) return breakpoints are the 30<sup>th</sup> and 70<sup>th</sup> NYSE percentiles.

	Median ME	
70th prior (2-12) percentile	Small Up	Big Up
30th prior (2-12) percentile	Small Medium	Big Medium
	Small Down	Big Down

**Stocks:** The six portfolios constructed each month include NYSE, AMEX, and NASDAQ stocks with prior return data. To be included in a portfolio for month  $t$  (formed at the end of month  $t-1$ ), a stock must have a price for the end of month  $t-13$  and a good return for  $t-2$ . In addition, any missing returns from  $t-12$  to  $t-3$  must be -99.0, CRSP's code for a missing price. Each included stock also must have ME for the end of month  $t-1$ .

# Cross Sectional Regression

---

- Factors don't have to be portfolios returns
- Two step regression estimate
- Beta estimation

$$R_{it} = a_i + \beta_i f_t + \varepsilon_{it}, t = 1, 2, \dots, T$$

- Average stock returns to estimate factor return using GLS

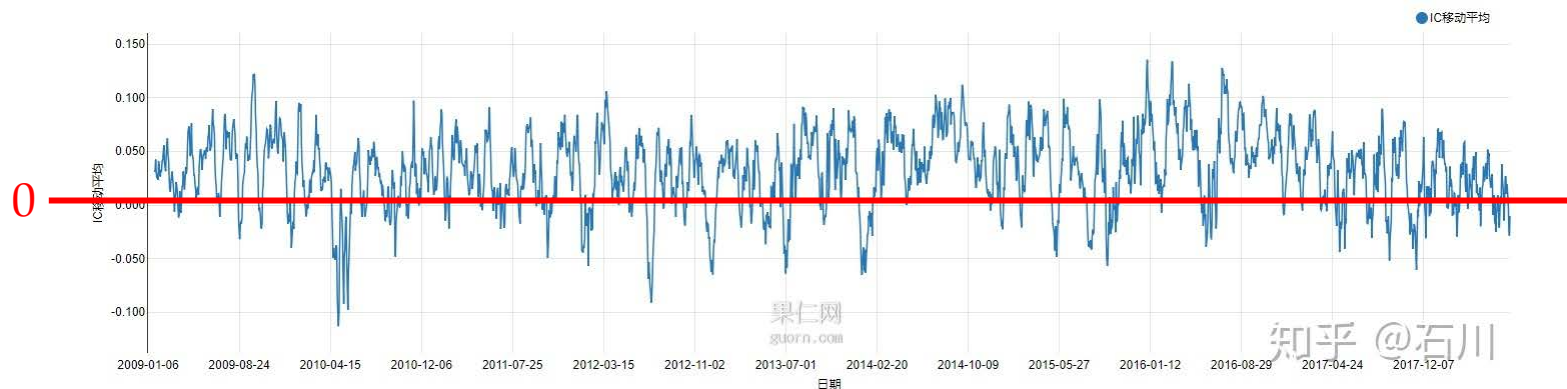
$$E_T[R_i] = \beta_i \lambda + \alpha_i, i = 1, 2, \dots, N$$

- Note that  $\beta_i$  are estimated, therefore the covariance matrix of  $\alpha_i$  are not easy to estimate. GMM is needed to test  $\alpha_i$

# Information Coefficient

- Following Barra models,  $\beta_i$  represents characteristics of individual stocks (normalized)
- Let  $\beta_t$  denote the factor exposure vector of stocks at time  $t$ , and  $y_{t,t+1}$  denote the stock returns from time  $t$  to  $t+1$ , we have

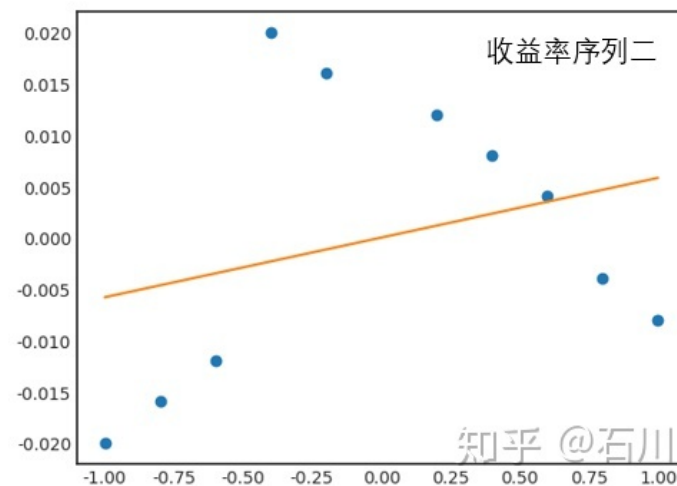
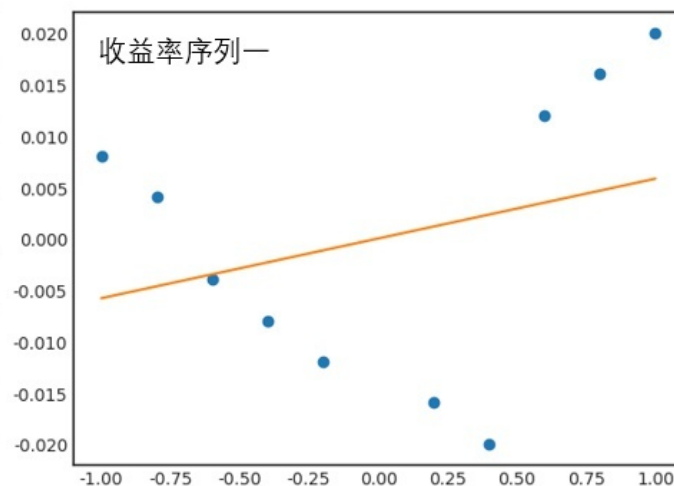
$$IC_t = \text{corr}(\beta_t, y_{t,t+1})$$



# Information Coefficient Check

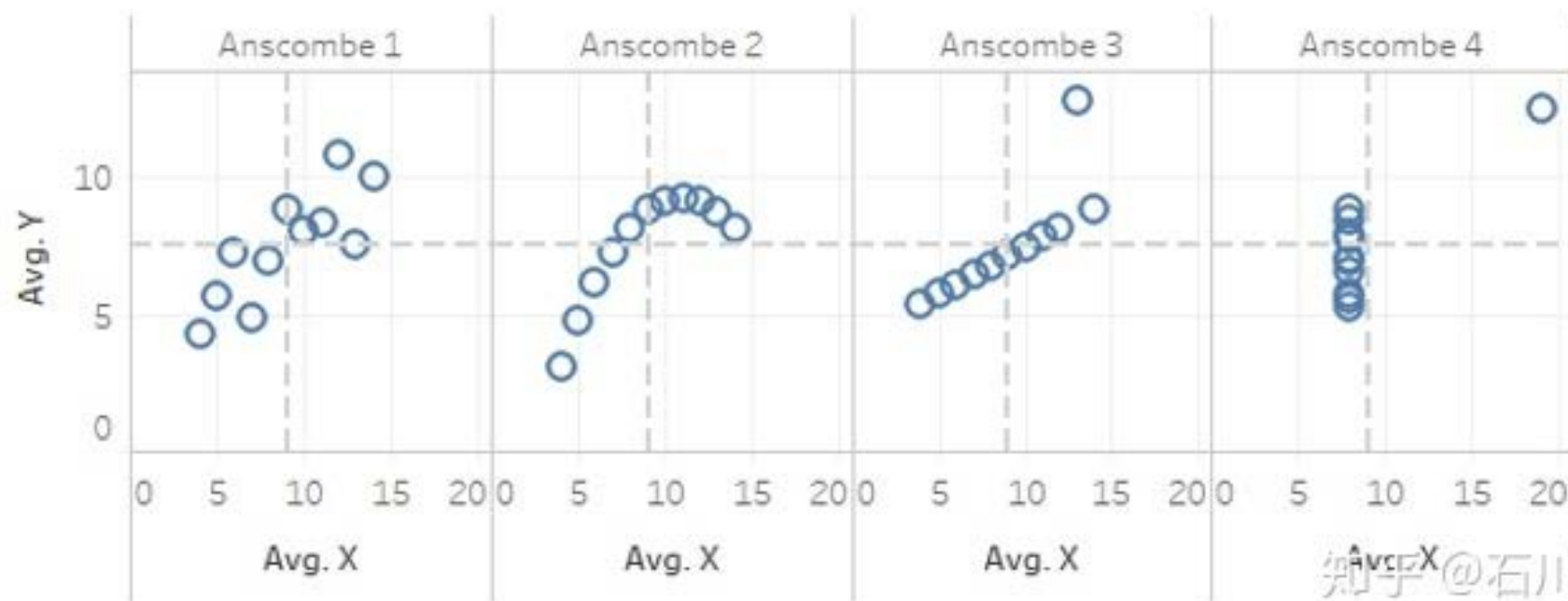
因子取值	收益率序列一	收益率序列二
1.0	2.00 %	-0.80 %
0.8	1.60 %	-0.40 %
0.6	1.20 %	0.40 %
0.4	-2.00 %	0.80 %
0.2	-1.60 %	1.20 %
-0.2	-1.20 %	1.60 %
-0.4	-0.80 %	2.00 %
-0.6	-0.40 %	-1.20 %
-0.8	0.40 %	-1.60 %
-1.0	0.80 %	-2.00 %

Both return time series have the same IC equal to 0.29



# Statistics Characteristics Are Summaries

Anscombe quartet



# Statistics Characteristics Are Summaries

第一组		第二组		第三组		第四组	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

性质	数值
x 的均值	9
x 的方差	11
y 的均值	7.50, 精确到小数点后两位
y 的方差	$4.125 \pm 0.003$
x 与 y 之间的相关系数	0.816, 精确到小数点后三位
线性回归线	$y = 3.00 + 0.500x$ 分别精确到小数点后两位和三位
线性回归 R-squared	0.67, 精确到小数点后两位

Matejka, J. and G. Fitzmaurice (2017). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. CHI 2017 Conference proceedings: ACM SIGCHI Conference on Human Factors in Computing Systems.

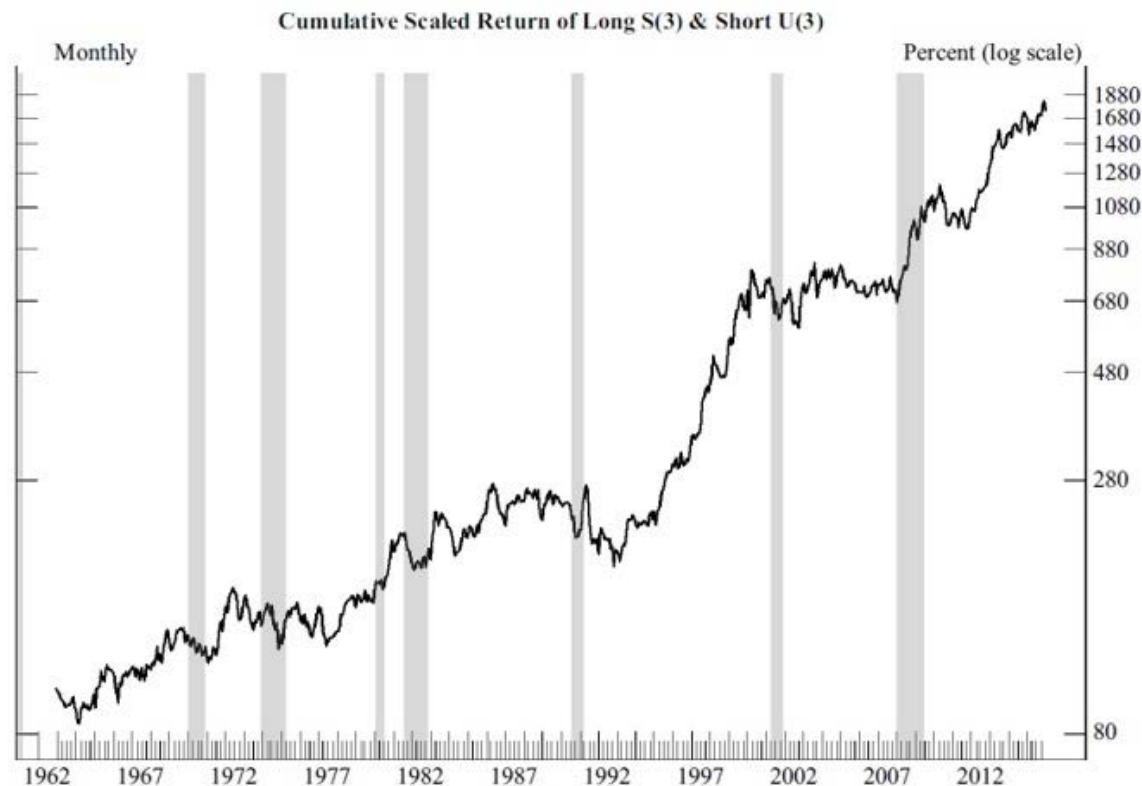


# Backtesting is Everything?

## EXHIBIT 1

Long-Short Market-Neutral Strategy Based on NYSE Stocks, January 1963 to December 2015

1. The same simple factor construction method:  
 $S(3) - U(3)$
2. Very low correlation with most well known factors
3. Only 10% turnover annually



Notes: Gray areas denote NBER recessions. Strategy returns scaled to match S&P 500 T-bill volatility during this period.

Source: Campbell Harvey, using data from CRSP.

出处: Arnott, Harvey, and Markowitz (2019)

知乎 @石川

# Research Protocol – Part I

## Arnott, Harvey, and Markowitz (2019)

---

### EXHIBIT 2

#### Seven-Point Protocol for Research in Quantitative Finance

---

##### 1. Research Motivation

- a. Does the model have a solid economic foundation?
- b. Did the economic foundation or hypothesis exist before the research was conducted?

##### 2. Multiple Testing and Statistical Methods

- a. Did the researcher keep track of all models and variables that were tried (both successful and unsuccessful), and are the researchers aware of the multiple-testing issue?
- b. Is there a full accounting of all possible interaction variables if interaction variables are used?
- c. Did the researchers investigate all variables set out in the research agenda, or did they cut the research as soon as they found a good model?

##### 3. Data and Sample Choice

- a. Do the data chosen for examination make sense? And, if other data are available, is it reasonable to exclude these data?
- b. Did the researchers take steps to ensure the integrity of the data?
- c. Do the data transformations, such as scaling, make sense? Were they selected in advance? And are the results robust to minor changes in these transformations?
- d. If outliers are excluded, are the exclusion rules reasonable?
- e. If the data are winsorized, was there a good reason to do it? Was the winsorization rule chosen before the research was started? Was only one winsorization rule tried (as opposed to many)?

##### 4. Cross-Validation

- a. Are the researchers aware that true out-of-sample tests are only possible in live trading?
- b. Are steps in place to eliminate the risk of out-of-sample iterations (i.e., an in-sample model that is later modified to fit out-of-sample data)?
- c. Is the out-of-sample analysis representative of live trading? For example, are trading costs and data revisions taken into account?

# Research Protocol – Part II

## Arnott, Harvey, and Markowitz (2019)

---

### 4. Cross-Validation

- a. Are the researchers aware that true out-of-sample tests are only possible in live trading?
- b. Are steps in place to eliminate the risk of out-of-sample iterations (i.e., an in-sample model that is later modified to fit out-of-sample data)?
- c. Is the out-of-sample analysis representative of live trading? For example, are trading costs and data revisions taken into account?

### 5. Model Dynamics

- a. Is the model resilient to structural change, and have the researchers taken steps to minimize the overfitting of the model dynamics?
- b. Does the analysis take into account the risk/likelihood of overcrowding in live trading?
- c. Do researchers take steps to minimize the tweaking of a live model?

### 6. Complexity

- a. Is the model designed to minimize the curse of dimensionality?
- b. Have the researchers taken steps to produce the simplest practicable model specification?
- c. Has an attempt been made to interpret the predictions of the machine learning model rather than using it as a black box?

### 7. Research Culture

- a. Does the research culture reward the quality of the science rather than the finding of a winning strategy?
- b. Do the researchers and management understand that most tests will fail?
- c. Are expectations clear (that researchers should seek the truth, not just something that works) when research is delegated?

# Further Readings – Journals

---

- Academic: JF, JFE, RFS
- Journal of Portfolio Management
- Journal of Financial Data Science

EDITORS		
FRANK J. FABOZZI EDHEC Business School	MARCOS LÓPEZ DE PRADO AQR Capital Management	JOSEPH SIMONIAN Natixis Investment Managers
MANAGING EDITOR FRANCESCO A. FABOZZI Frank J. Fabozzi Associates		
ADVISORY BOARD		
IRENE ALDRIDGE Cornell University and AbleMarkets	MICHAEL IMERMAN Peter F. Drucker and Masatoshi Ito Graduate School of Management Claremont Graduate University	PETTER KOLM Courant Institute, New York University
ROBERT ARNOTT Research Affiliates	BRUCE JACOBS Jacobs Levy Equity Management	HAN LIU Northwestern University
JOSEPH CERNIGLIA University of Pennsylvania	RON KAHN BlackRock Financial Management	ANTHONY LEDFORD Man-AHL
RAMA CONT Oxford University	KATHRYN KAMINSKI AlphaSimplex Group LLC	JIM KYUNG-SOO LIEW Johns Hopkins University
GERMÁN G. CREAMER Stevens Institute of Technology	HOSSEIN KAZEMI CISDM/Isenberg School of Management and University of Massachusetts at Amherst	PAUL F. MENDE MIT Sloan School of Management
SANJIV RANJAN DAS Santa Clara University		WARREN PENNINGTON Vanguard
CAMPBELL HARVEY Fuqua School of Business Duke University		SIDNEY C. PORTER FEV Analytics



# Further Readings – Online


- <https://www.zhihu.com/people/mitshi/activities>
- <https://www.zhihu.com/people/JoinQuant/activities>
- <https://www.zhihu.com/people/llanglli/activities>
- Youtube
  - Asset pricing john cochrane
  - Statquest





## 石川 量化投资


居住地 现居北京，曾在硅谷、波士顿住过

所在行业 证券投资

职业经历  北京量信投资管理有限公司 · 创始合伙人

 甲骨文 (Oracle)

 花旗集团 (Citi)

教育经历  麻省理工学院 (MIT)

 清华大学

个人简介 北京量信投资管理有限公司创始合伙人，毕业于清华大学，获工学学士和硕士学位，后于美国麻省理工学院获得博士学位。

原创不易，请保护版权，请在获得授权后转载，并注明出处，谢谢。  
已委托“维权骑士” (rightknights.com) 为进行维权行动。

微信公众号：川总写量化

# Thank you!

---