Wharton
University of Pennsylvania
JACOBS LEVY EQUITY MANAGEMENT CENTER
FOR QUANTITATIVE FINANCIAL RESEARCH

Lucky Factors

Campbell R. Harvey
Duke University, NBER and
Man Group plc

Credits

Joint work with

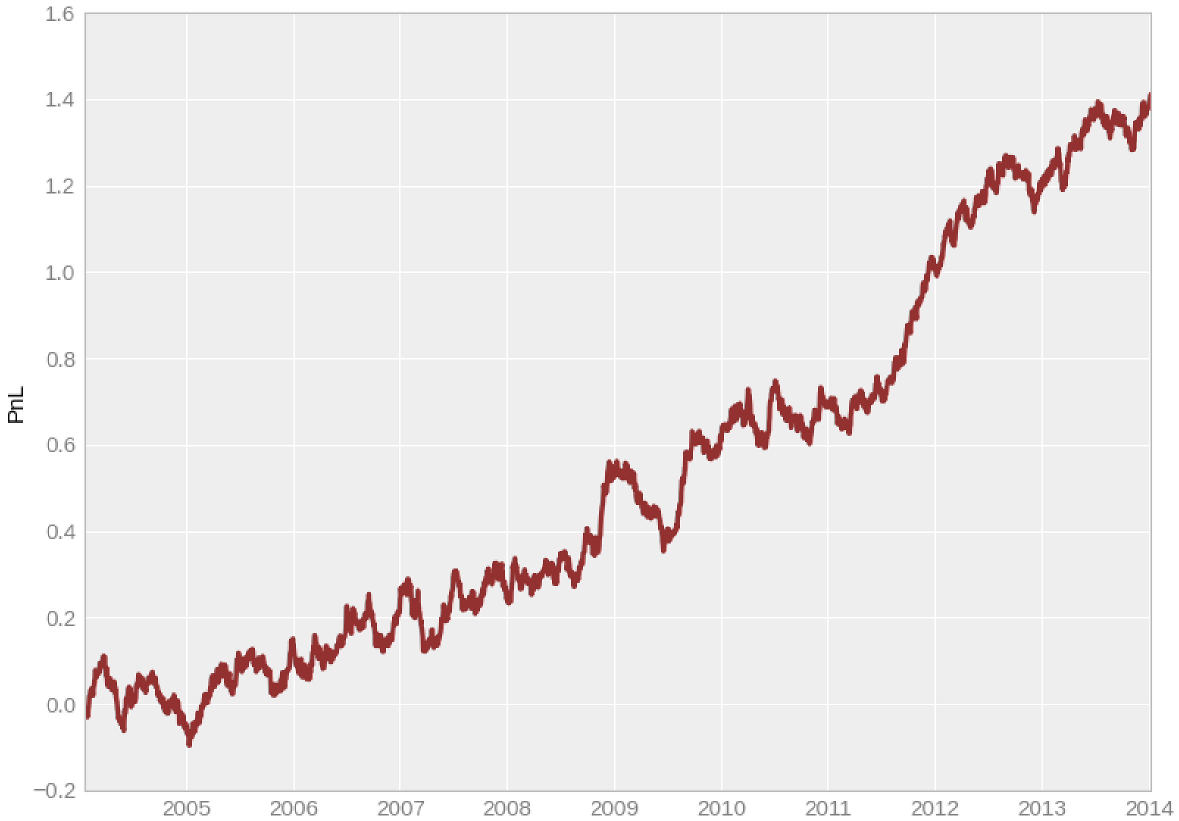
Yan Liu

Texas A&M University

Based on our joint work:

- *"… and the Cross-section of Expected Returns"*
  http://ssrn.com/abstract=2249314 [Best paper in investment, WFA 2014]
- *"Backtesting"*
  http://ssrn.com/abstract=2345489 [1st Prize, INQUIRE Europe/UK]
- *"Evaluating Trading Strategies"* [Jacobs-Levy best paper, JPM 2014]
  http://ssrn.com/abstract=2474755
- *"Lucky Factors"*
  http://ssrn.com/abstract=2528780
- *"A test of the incremental efficiency of a given portfolio"*

The Setting

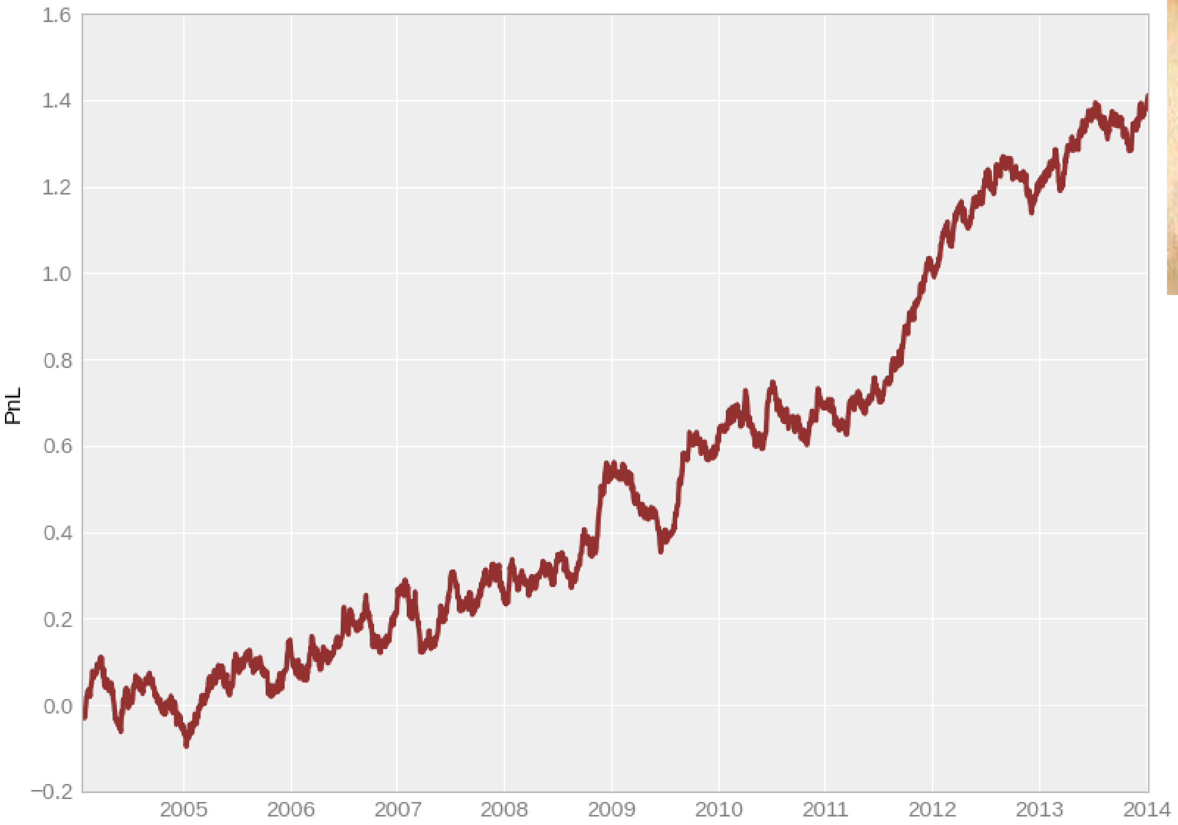


Source: AHL Research

Performance of trading strategy is very impressive.

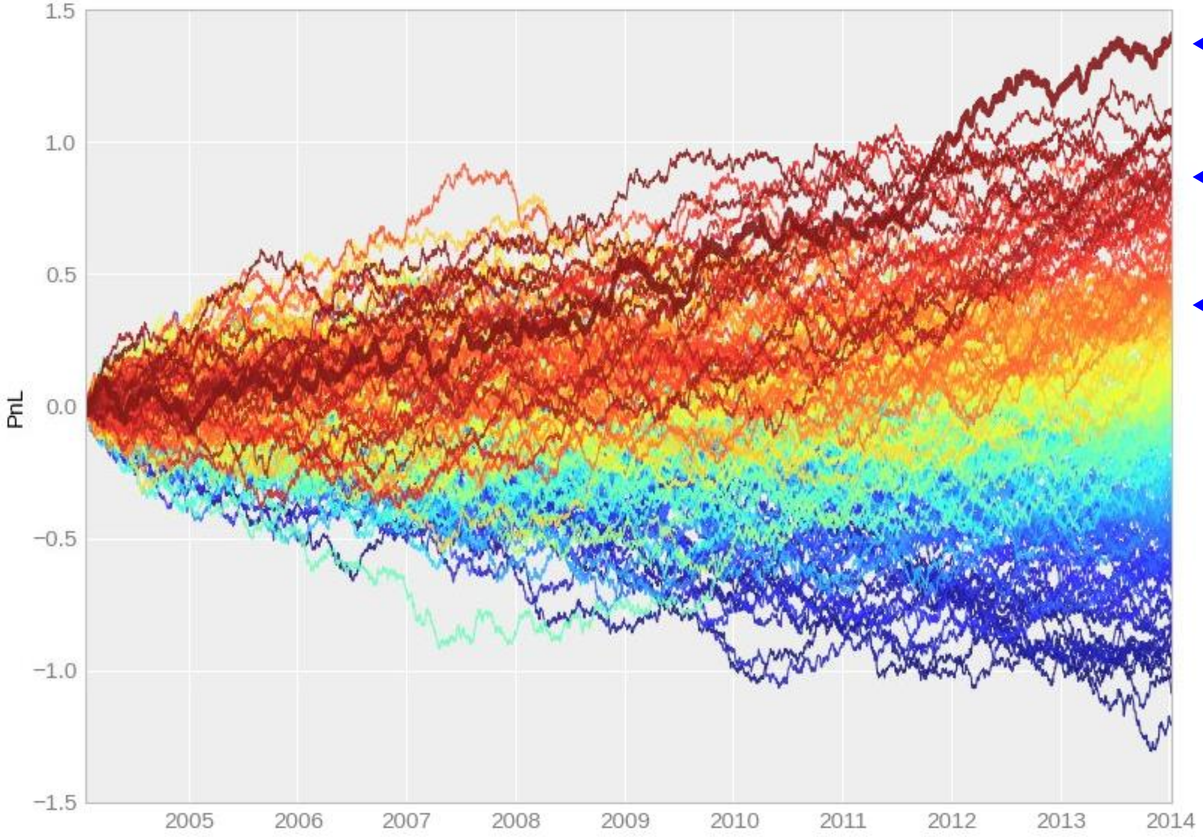
- SR=1
- Consistent
- Drawdowns acceptable

The Setting



Source: AHL Research

The Setting



Sharpe = 1 (t-stat=2.91)

Sharpe = 2/3

Sharpe = 1/3

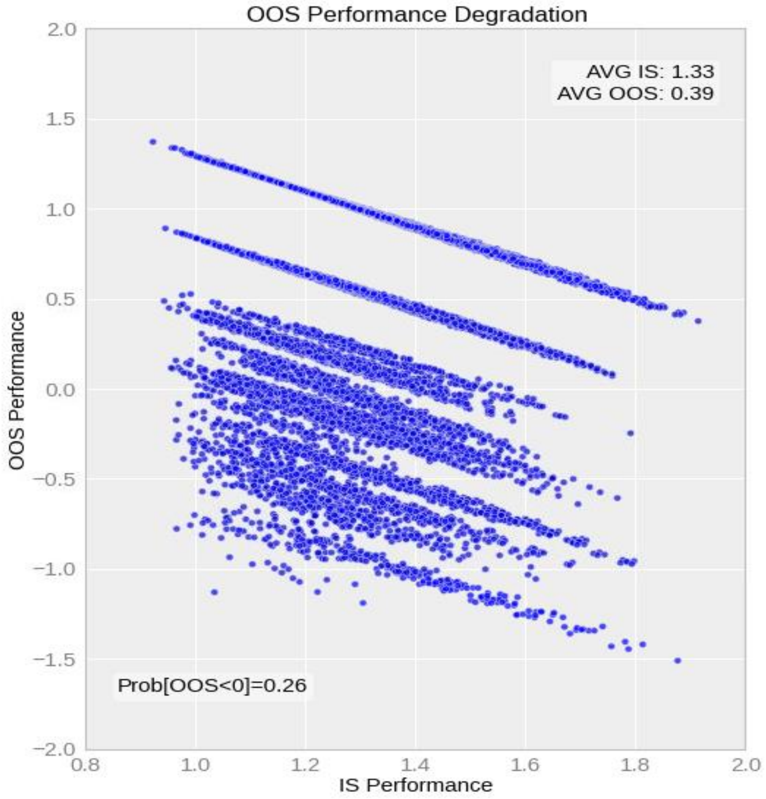
200 random time-series
mean=0; volatility=15%

Source: AHL Research

The Setting

The good news:

- Harvey and Liu (2014) suggest a multiple testing correction which provides a haircut for the Sharpe Ratios. No strategy would be declared "significant"
- Lopez De Prado et al. (2014) uses an alternative approach, the "probability of overfitting" which in this example is a large 0.26
- Both methods deal with the data mining problem

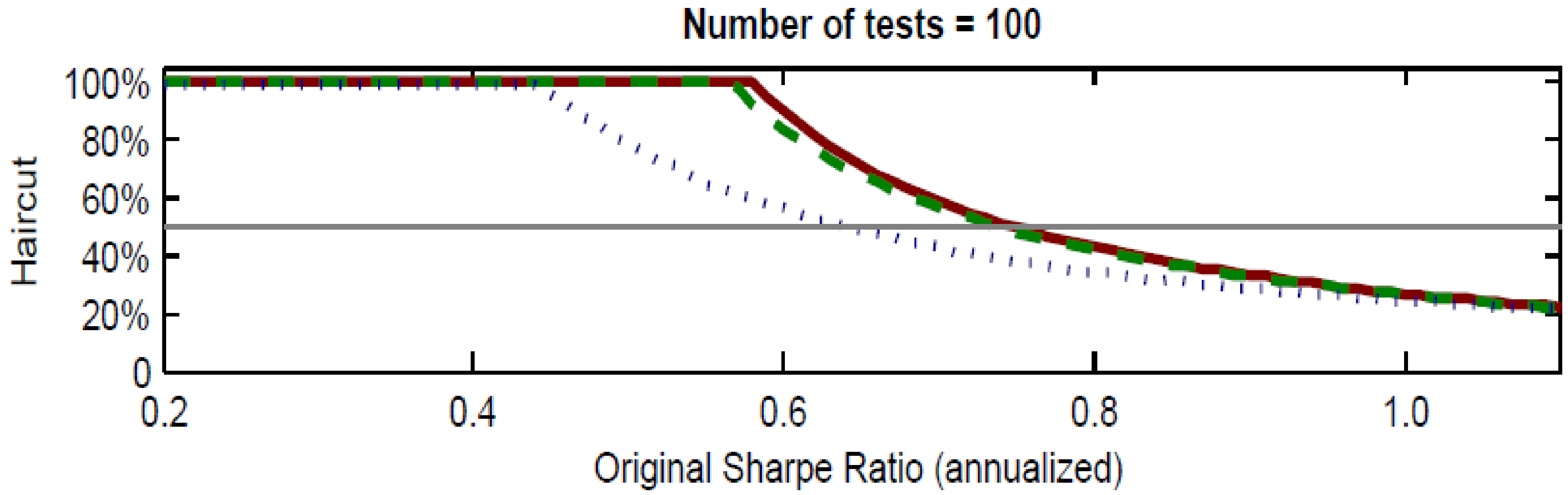


Source: AHL Research

The Setting

The good news:

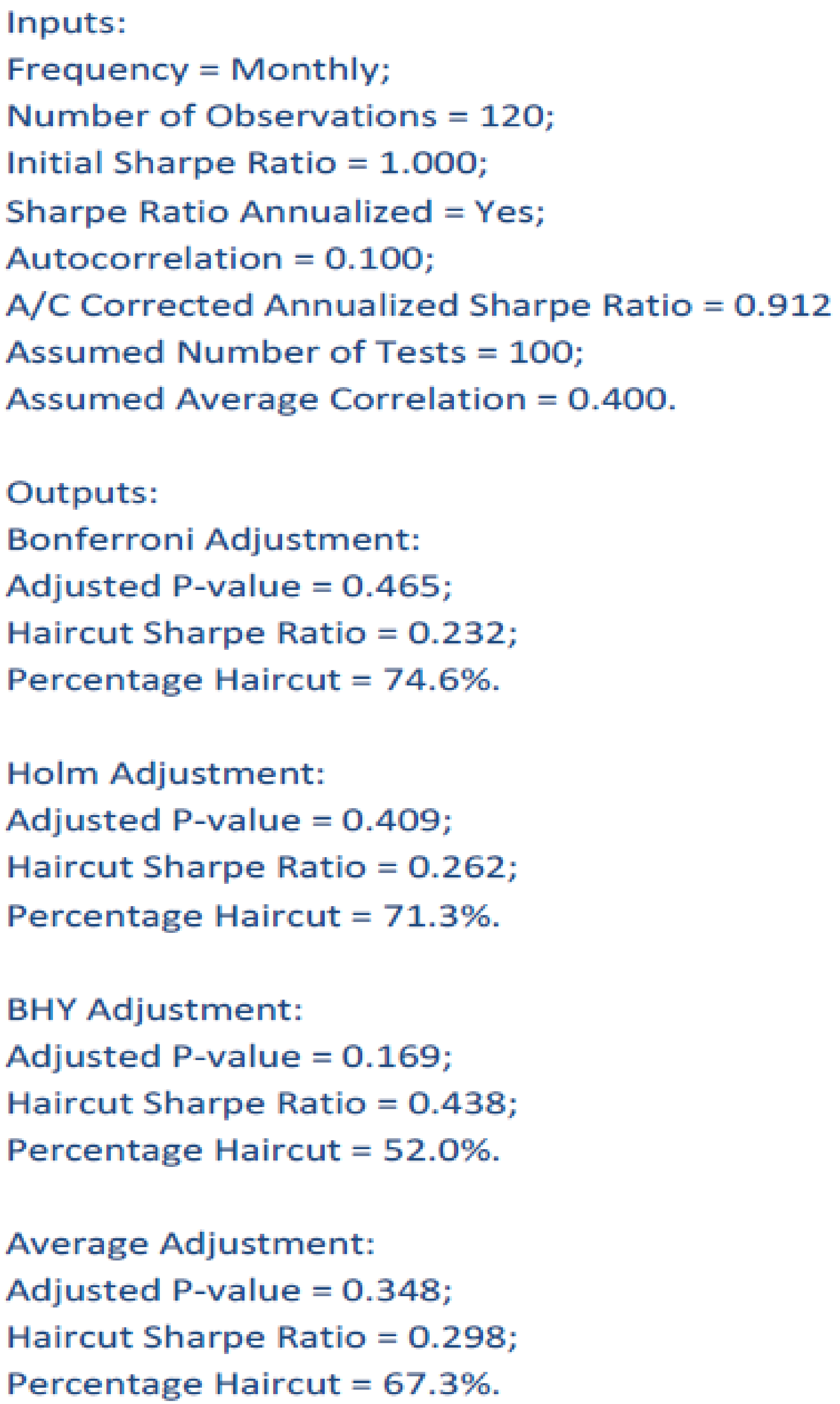
- Harvey and Liu (2014) Haircut Sharpe ratio takes the number of tests into account as well as the size of the sample.



The Setting

The good news:

- Haircut Sharpe Ratio:
  - Sample size

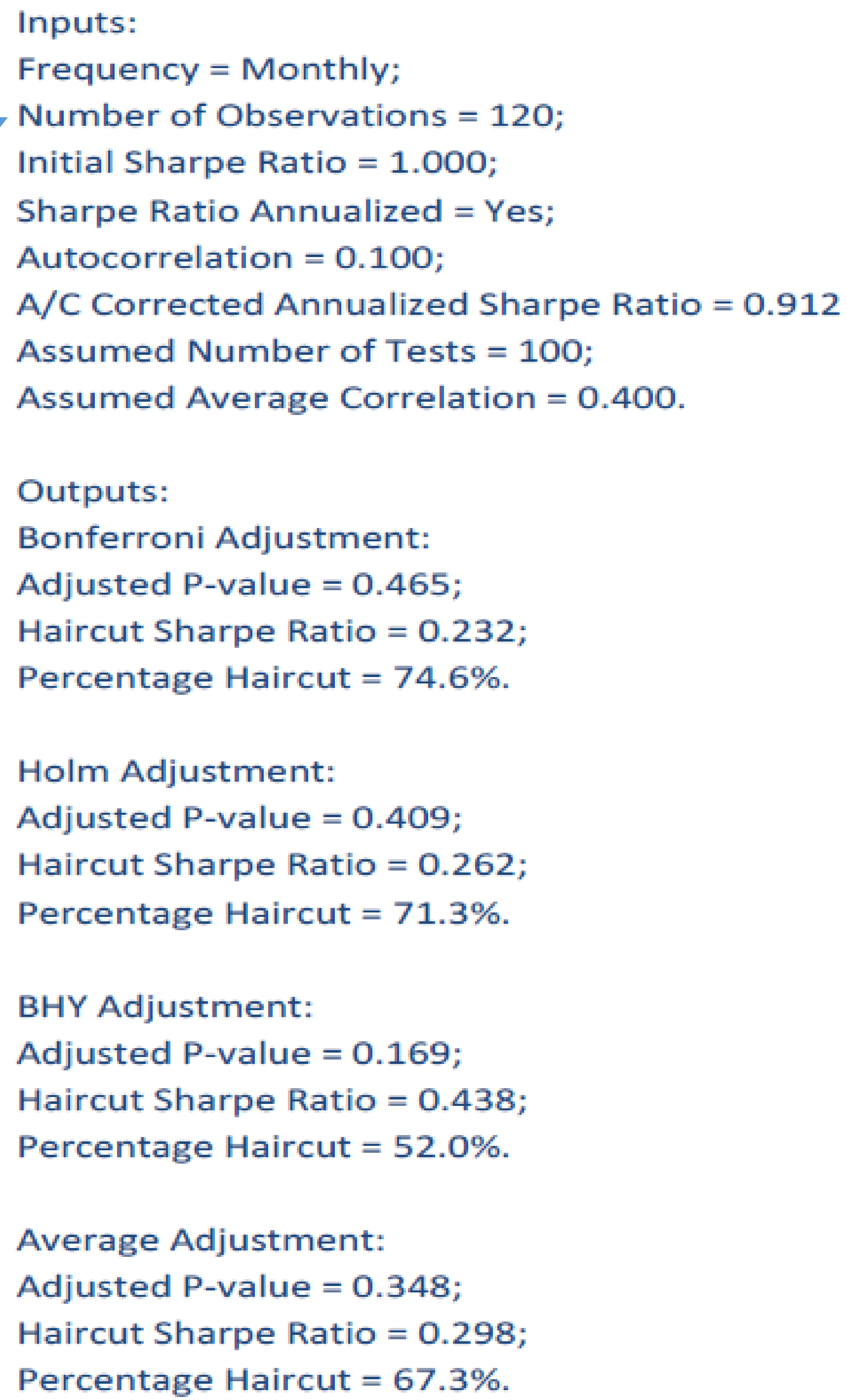


```
Inputs:
Frequency = Monthly;
Number of Observations = 120;
Initial Sharpe Ratio = 1.000;
Sharpe Ratio Annualized = Yes;
Autocorrelation = 0.100;
A/C Corrected Annualized Sharpe Ratio = 0.912
Assumed Number of Tests = 100;
Assumed Average Correlation = 0.400.

Outputs:
Bonferroni Adjustment:
Adjusted P-value = 0.465;
Haircut Sharpe Ratio = 0.232;
Percentage Haircut = 74.6%.

Holm Adjustment:
Adjusted P-value = 0.409;
Haircut Sharpe Ratio = 0.262;
Percentage Haircut = 71.3%.

BHY Adjustment:
Adjusted P-value = 0.169;
Haircut Sharpe Ratio = 0.438;
Percentage Haircut = 52.0%.

Average Adjustment:
Adjusted P-value = 0.348;
Haircut Sharpe Ratio = 0.298;
Percentage Haircut = 67.3%.
```

The Setting

The good news:

- Haircut Sharpe Ratio:
  - Sample size
  - Autocorrelation

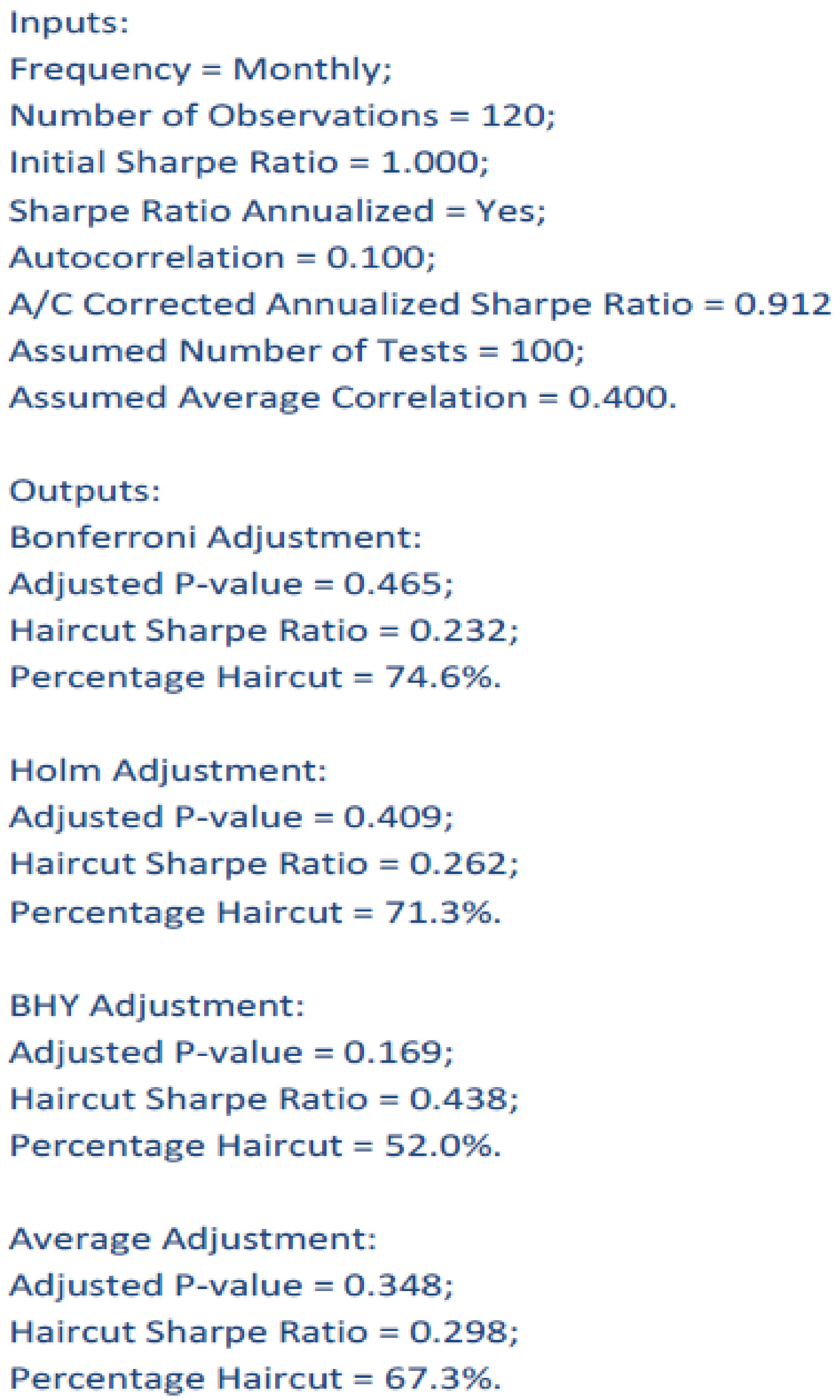


```
Inputs:
Frequency = Monthly;
Number of Observations = 120;
Initial Sharpe Ratio = 1.000;
Sharpe Ratio Annualized = Yes;
Autocorrelation = 0.100;
A/C Corrected Annualized Sharpe Ratio = 0.912
Assumed Number of Tests = 100;
Assumed Average Correlation = 0.400.

Outputs:
Bonferroni Adjustment:
Adjusted P-value = 0.465;
Haircut Sharpe Ratio = 0.232;
Percentage Haircut = 74.6%.

Holm Adjustment:
Adjusted P-value = 0.409;
Haircut Sharpe Ratio = 0.262;
Percentage Haircut = 71.3%.

BHY Adjustment:
Adjusted P-value = 0.169;
Haircut Sharpe Ratio = 0.438;
Percentage Haircut = 52.0%.

Average Adjustment:
Adjusted P-value = 0.348;
Haircut Sharpe Ratio = 0.298;
Percentage Haircut = 67.3%.
```

The Setting

The good news:

- Haircut Sharpe Ratio:
  - Sample size
  - Autocorrelation
  - The number of tests (data mining)

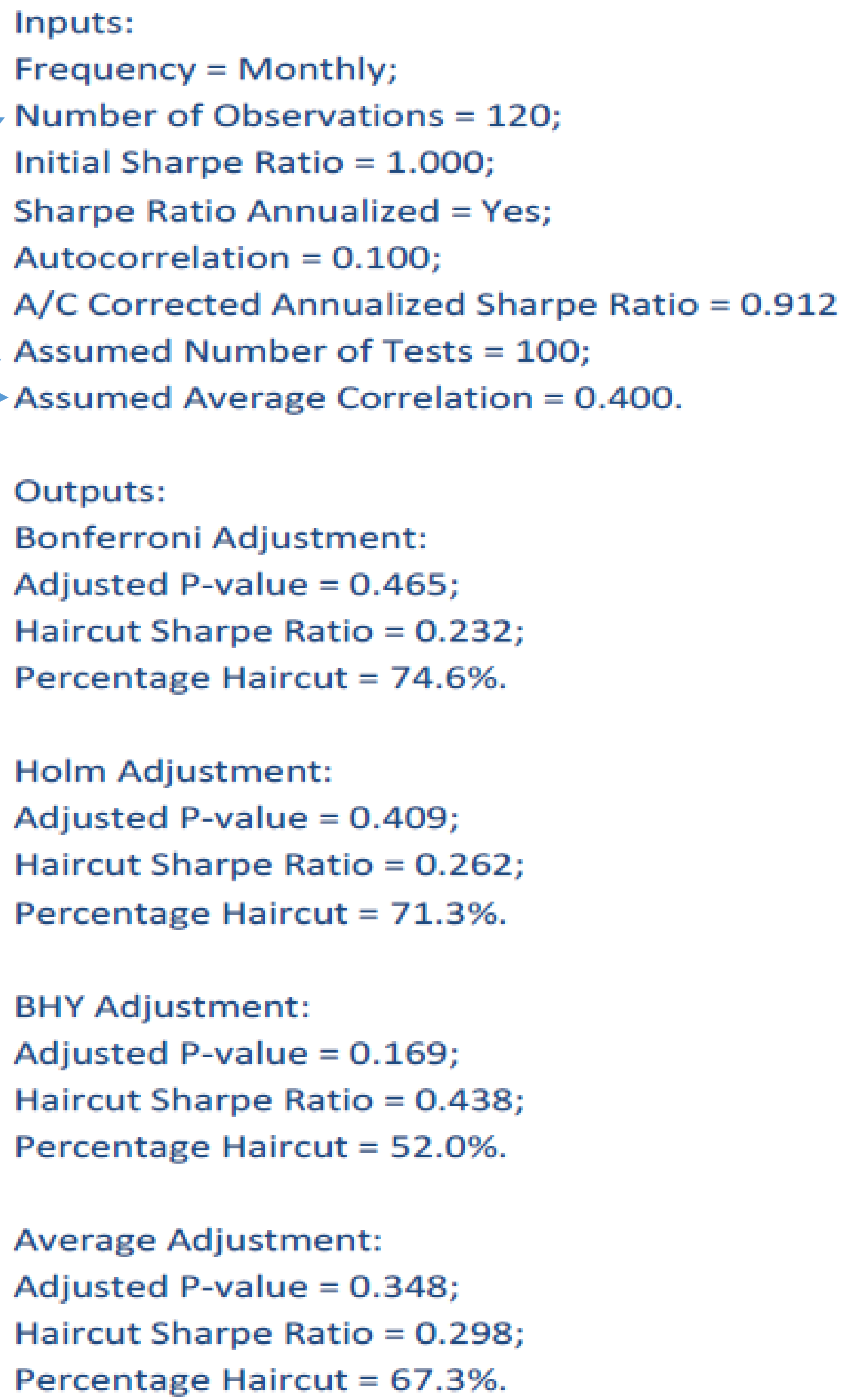


```
Inputs:
Frequency = Monthly;
Number of Observations = 120;
Initial Sharpe Ratio = 1.000;
Sharpe Ratio Annualized = Yes;
Autocorrelation = 0.100;
A/C Corrected Annualized Sharpe Ratio = 0.912
Assumed Number of Tests = 100;
Assumed Average Correlation = 0.400.

Outputs:
Bonferroni Adjustment:
Adjusted P-value = 0.465;
Haircut Sharpe Ratio = 0.232;
Percentage Haircut = 74.6%.

Holm Adjustment:
Adjusted P-value = 0.409;
Haircut Sharpe Ratio = 0.262;
Percentage Haircut = 71.3%.

BHY Adjustment:
Adjusted P-value = 0.169;
Haircut Sharpe Ratio = 0.438;
Percentage Haircut = 52.0%.

Average Adjustment:
Adjusted P-value = 0.348;
Haircut Sharpe Ratio = 0.298;
Percentage Haircut = 67.3%.
```

The Setting

The good news:

- Haircut Sharpe Ratio:
  - Sample size
  - Autocorrelation
  - The number of tests (data mining)
  - Correlation of tests



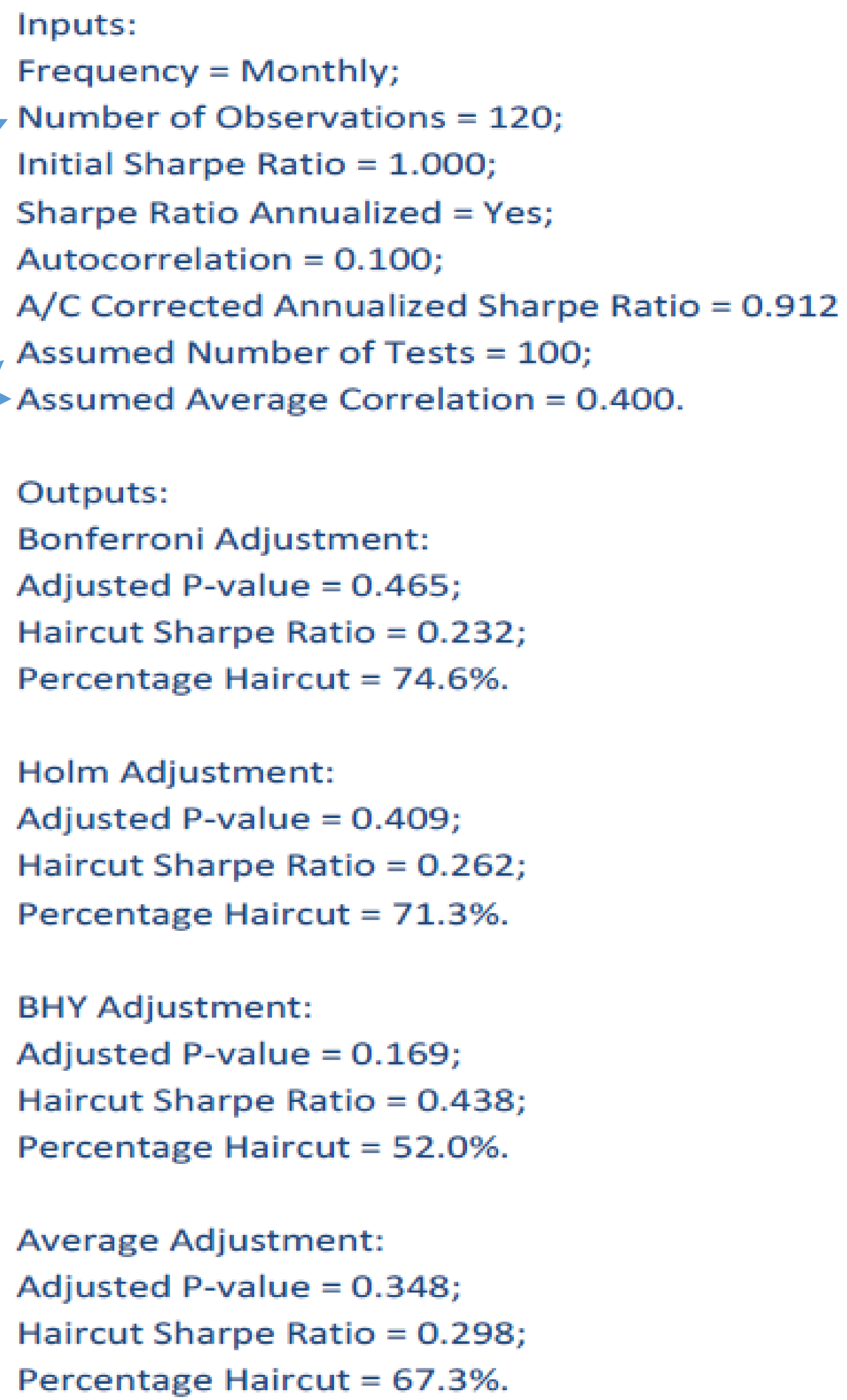
```
Inputs:
Frequency = Monthly;
Number of Observations = 120;
Initial Sharpe Ratio = 1.000;
Sharpe Ratio Annualized = Yes;
Autocorrelation = 0.100;
A/C Corrected Annualized Sharpe Ratio = 0.912
Assumed Number of Tests = 100;
Assumed Average Correlation = 0.400.

Outputs:
Bonferroni Adjustment:
Adjusted P-value = 0.465;
Haircut Sharpe Ratio = 0.232;
Percentage Haircut = 74.6%.

Holm Adjustment:
Adjusted P-value = 0.409;
Haircut Sharpe Ratio = 0.262;
Percentage Haircut = 71.3%.

BHY Adjustment:
Adjusted P-value = 0.169;
Haircut Sharpe Ratio = 0.438;
Percentage Haircut = 52.0%.

Average Adjustment:
Adjusted P-value = 0.348;
Haircut Sharpe Ratio = 0.298;
Percentage Haircut = 67.3%.
```

The Setting

The good news:

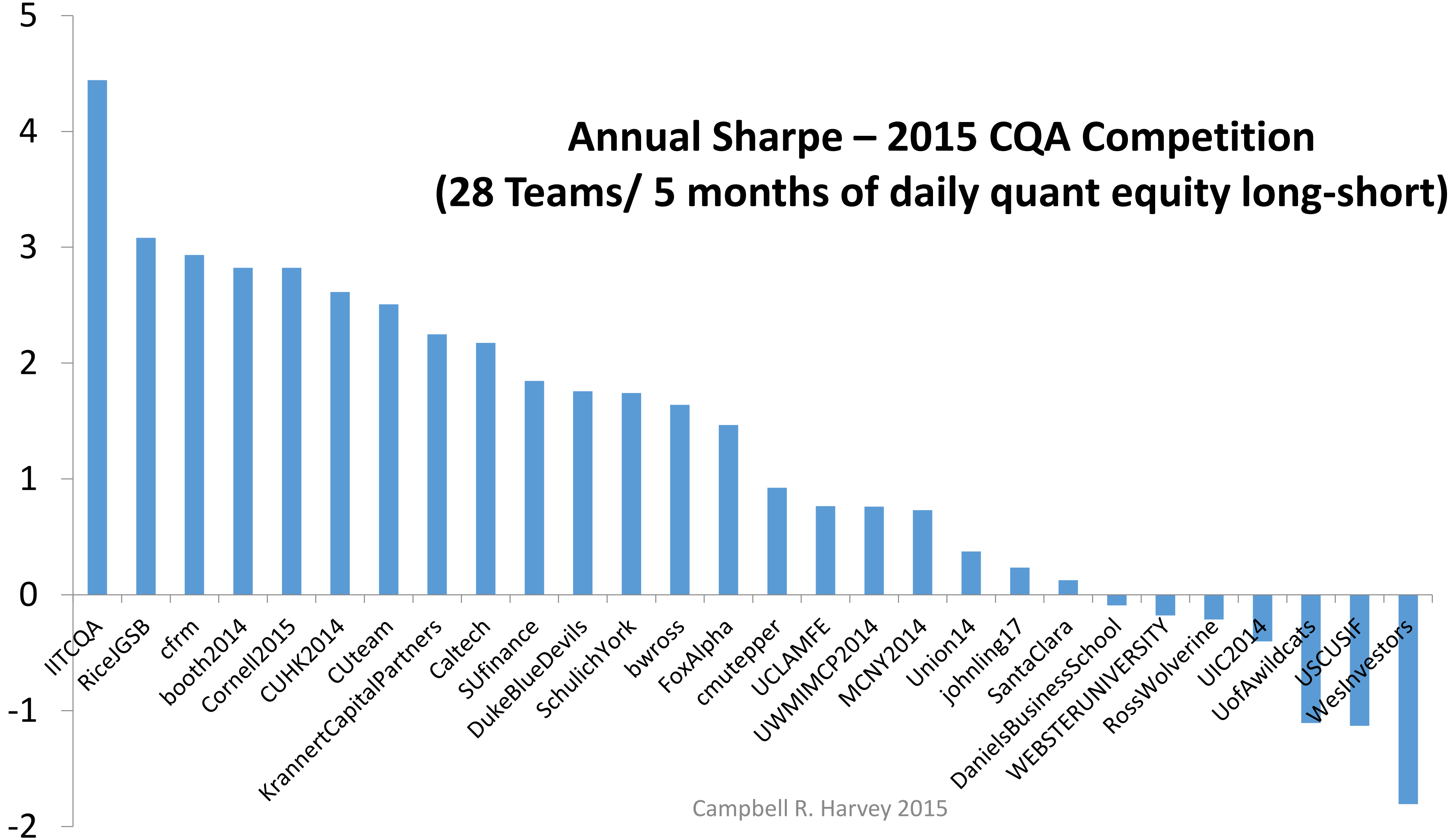
- Haircut Sharpe Ratio:
  - Sample size
  - Autocorrelation
  - The number of tests (data mining)
  - Correlation of tests

Haircut Sharpe Ratio applies to the Maximal Sharpe Ratio



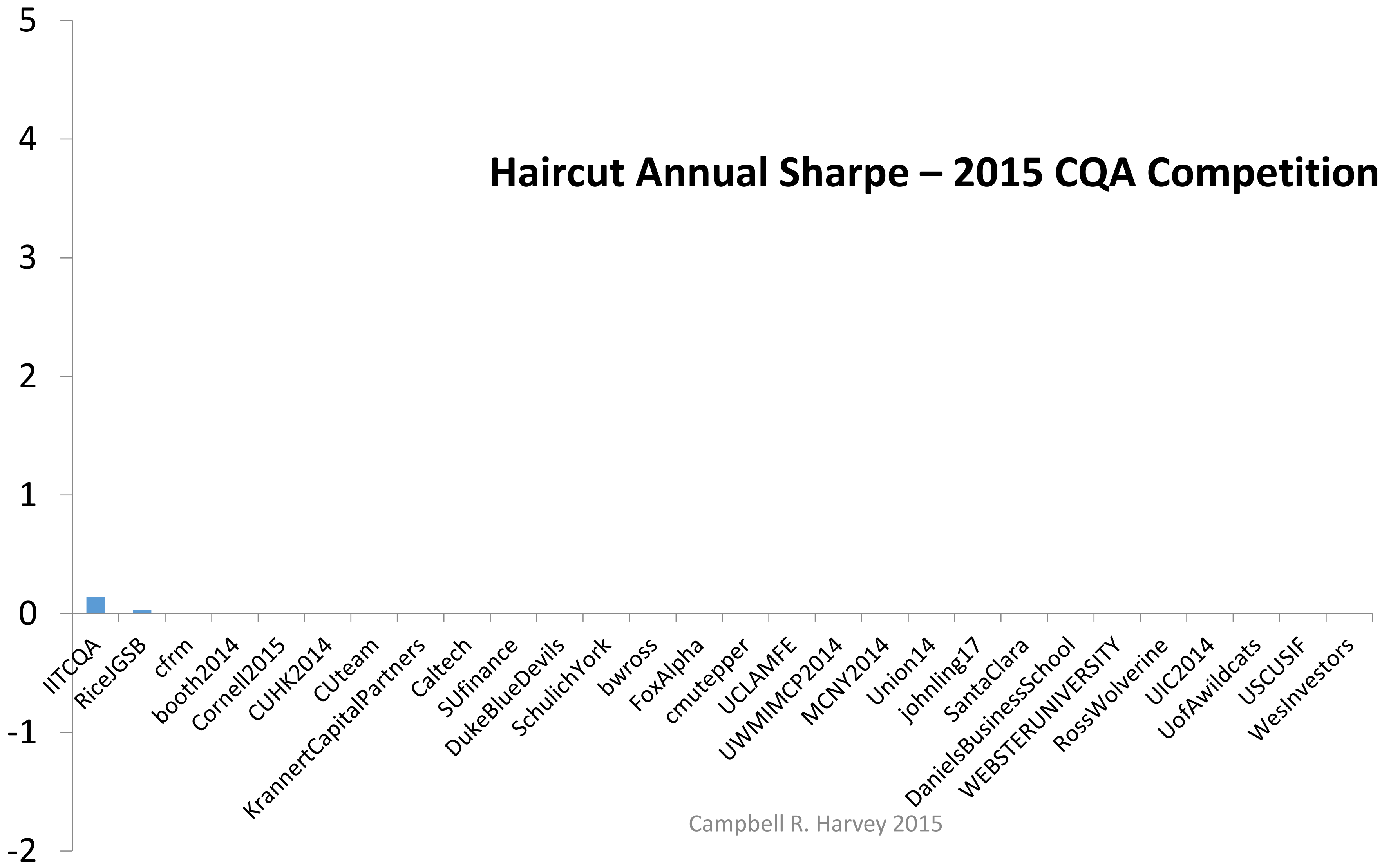
```
Inputs:
Frequency = Monthly;
Number of Observations = 120;
Initial Sharpe Ratio = 1.000;
Sharpe Ratio Annualized = Yes;
Autocorrelation = 0.100;
A/C Corrected Annualized Sharpe Ratio = 0.912
Assumed Number of Tests = 100;
Assumed Average Correlation = 0.400.

Outputs:
Bonferroni Adjustment:
Adjusted P-value = 0.465;
Haircut Sharpe Ratio = 0.232;
Percentage Haircut = 74.6%.

Holm Adjustment:
Adjusted P-value = 0.409;
Haircut Sharpe Ratio = 0.262;
Percentage Haircut = 71.3%.

BHY Adjustment:
Adjusted P-value = 0.169;
Haircut Sharpe Ratio = 0.438;
Percentage Haircut = 52.0%.

Average Adjustment:
Adjusted P-value = 0.348;
Haircut Sharpe Ratio = 0.298;
Percentage Haircut = 67.3%.
```

The Setting



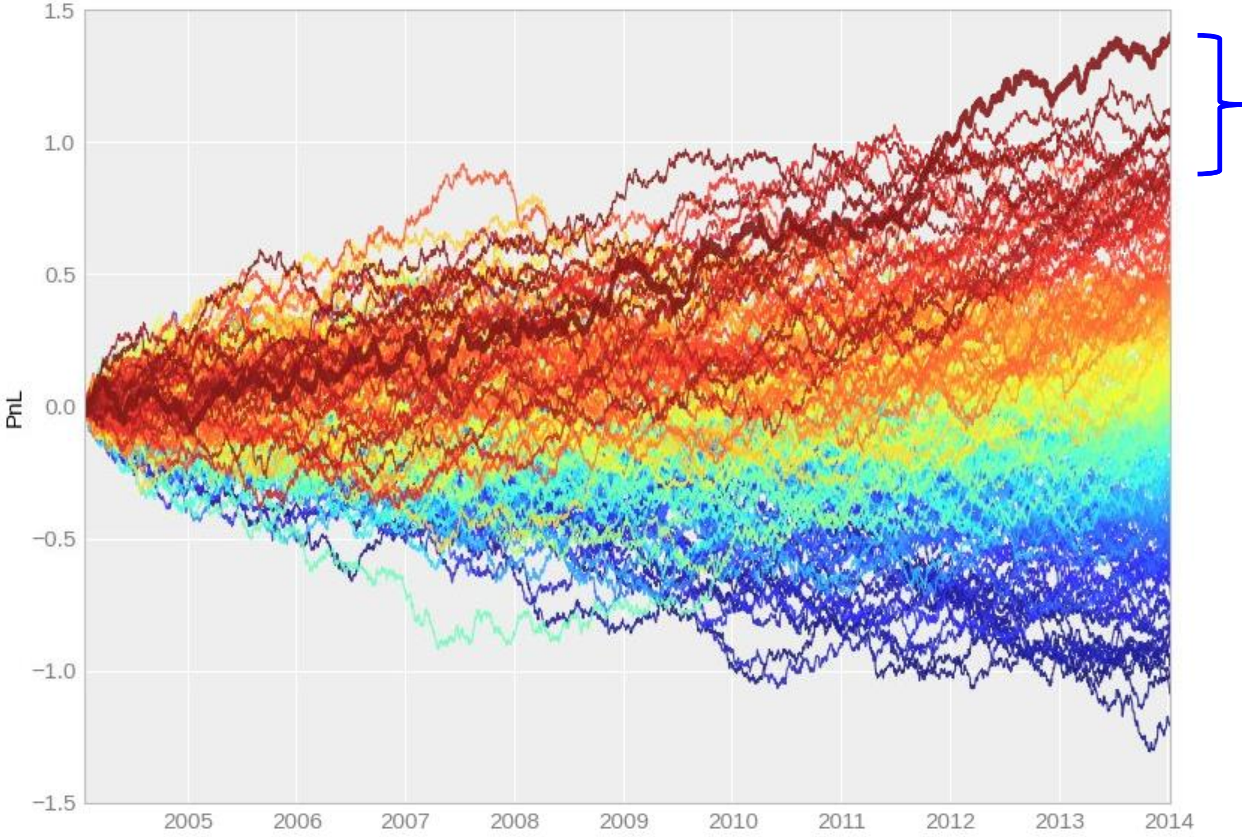
Annual Sharpe – 2015 CQA Competition
(28 Teams/ 5 months of daily quant equity long-short)

The Setting


Haircut Annual Sharpe – 2015 CQA Competition
5
4
3
2
1
0
-1
-2
IITCQA
RiceJGSB
cfrm
booth2014
Cornell2015
CUHK2014
CUteam
KrannertCapitalPartners
Caltech
SUfinance
DukeBlueDevils
SchulichYork
bwross
FoxAlpha
cmutepper
UCLAMFE
UWMIMCP2014
MCNY2014
Union14
johnling17
SantaClara
DanielsBusinessSchool
WEBSTERUNIVERSITY
RossWolverine
UIC2014
UofAwildcats
USCUSIF
WesInvestors

The Setting

The bad news:



Equal weighting of 10 best strategies produces a t-stat=4.5!

200 random time-series
mean=0; volatility=15%

Source: AHL Research

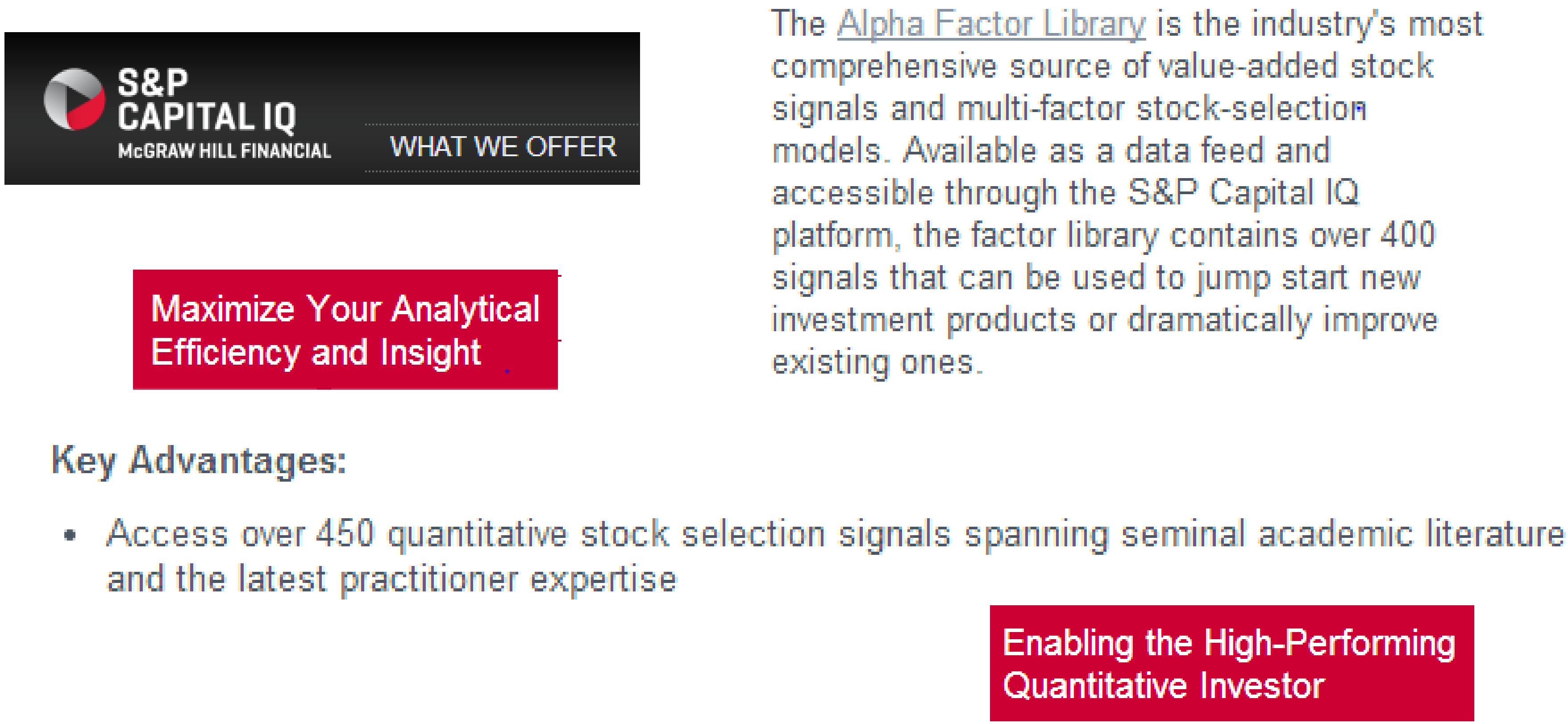
A Common Thread

A common thread connecting many important problems in finance

- Not just the in-house evaluation of trading strategies.
- There are thousands of fund managers. How to distinguish skill from luck?
- Dozens of variables have been found to forecast stock returns. Which ones are true?
- More than 300 factors have been published and thousands have been tried to explain the cross-section of expected returns. Which ones are true?

A Common Thread

Even more in the practice of finance. 400 factors!



S&P CAPITAL IQ
McGRAW HILL FINANCIAL

WHAT WE OFFER

Maximize Your Analytical Efficiency and Insight

The Alpha Factor Library is the industry's most comprehensive source of value-added stock signals and multi-factor stock-selection models. Available as a data feed and accessible through the S&P Capital IQ platform, the factor library contains over 400 signals that can be used to jump start new investment products or dramatically improve existing ones.

Key Advantages:

- Access over 450 quantitative stock selection signals spanning seminal academic literature and the latest practitioner expertise

Enabling the High-Performing Quantitative Investor

Source: https://www.capitaliq.com/home/who-we-help/investment-management/quantitative-investors.aspx

The Question

- The common thread is *multiple testing* or *data mining*
- Our research question:

How do we adjust standard models for data mining and how do we handle multiple factors?

A Motivating Example

Suppose we have 100 “X” variables to explain a single “Y” variable. The problems we face are:

I. Which regression model do we use?
   - E.g., for factor tests, panel regression vs. Fama-MacBeth

II. Are any of the 100 variables significant?
   - Due to data mining, significance at the conventional level is not enough
   - Need to take into account dependency among the Xs and between X and Y

A Motivating Example

III. Suppose we find one explanatory variable to be significant. How do we find the next?
   - The next needs to explain Y in addition to what the first one can explain
   - There is again multiple testing since 99 variables have been tried

IV. When do we stop? How many factors?

Our Approach

We propose a new framework that addresses multiple testing in regression models. Features of our framework include:

- It takes multiple testing into account
  - Our method allows for both time-series and cross-sectional dependence
- It sequentially identifies the group of “true” factors
- The general idea applies to different regression models
  - In the paper, we show how our model applies to predictive regression, panel regression, and the Fama-MacBeth procedure

Related Literature

Our framework leans heavily on Foster, Smith and Whaley (FSW, *Journal of Finance*, 1997) and White (*Econometrica*, 2000)

- FSW (1997) use simulations to show how regression R-squares are inflated when a few variables are selected from a large set of variables
  - We bootstrap from the real data (rather than simulate artificial data)
  - Our method accommodates a wide range of test statistics
- White (2000) suggests the use of the max statistics to adjust for data mining
  - We show how to create the max statistic within standard regression models

A Predictive Regression

Let’s return to the example of a Y variable and 100 possible X (predictor) variables. Suppose 500 observations.

- Step 1. Orthogonalize each of the X variables with respect to Y. Hence, a regression of Y on any X produces exactly zero R^2. This is the null hypothesis – no predictability.
- Step 2. Bootstrap the data, that is, the original Y and the orthogonalized Xs (produces a new data matrix 500x101)

A Predictive Regression

- Step 3. Run 100 regressions and save the max statistic of your choice (could be R^2, t-statistic, F-statistic, MAE, etc.), e.g. save the highest t-statistic from the 100 regressions. Note, in the unbootstrapped data, every t-statistic is exactly zero.
- Step 4. Repeat steps 2 and 3 10,000 times.
- Step 5. Now that we have the empirical distribution of the max t-statistic under the null of no predictability, compare to the max t-statistic in real data.

A Predictive Regression

- Step 5a. If the max t-stat in the real data fails to exceed the threshold (95th percentile of the null distribution), stop (no variable is significant).
- Step 5b. If the max t-stat in the real data exceeds the threshold, declare the variable, say, X_7, “true”
- Step 6. Orthogonalize Y with respect to X_7 and call it Y^e. This new variable is the part of Y that cannot be explained by X_7.
- Step 7. Reorthogonalize the remaining X variables (99 of them) with respect to Y^e.

A Predictive Regression

- Step 8. Repeat Steps 3-7 (except there are 99 regressions to run because one variable is declared true).
- Step 9. Continue until the max t-statistic in the data fails to exceed the max from the bootstrap

Advantages

- Addresses data mining directly
- Allows for cross-correlation of the X-variables because we are bootstrapping rows of data
- Allows for non-normality in the data (no distributional assumptions imposed – we are resampling the original data)
- Potentially allows for time-dependence in the data by changing to a block bootstrap technique.
- Answers the question: How many factors?

Fund Evaluation

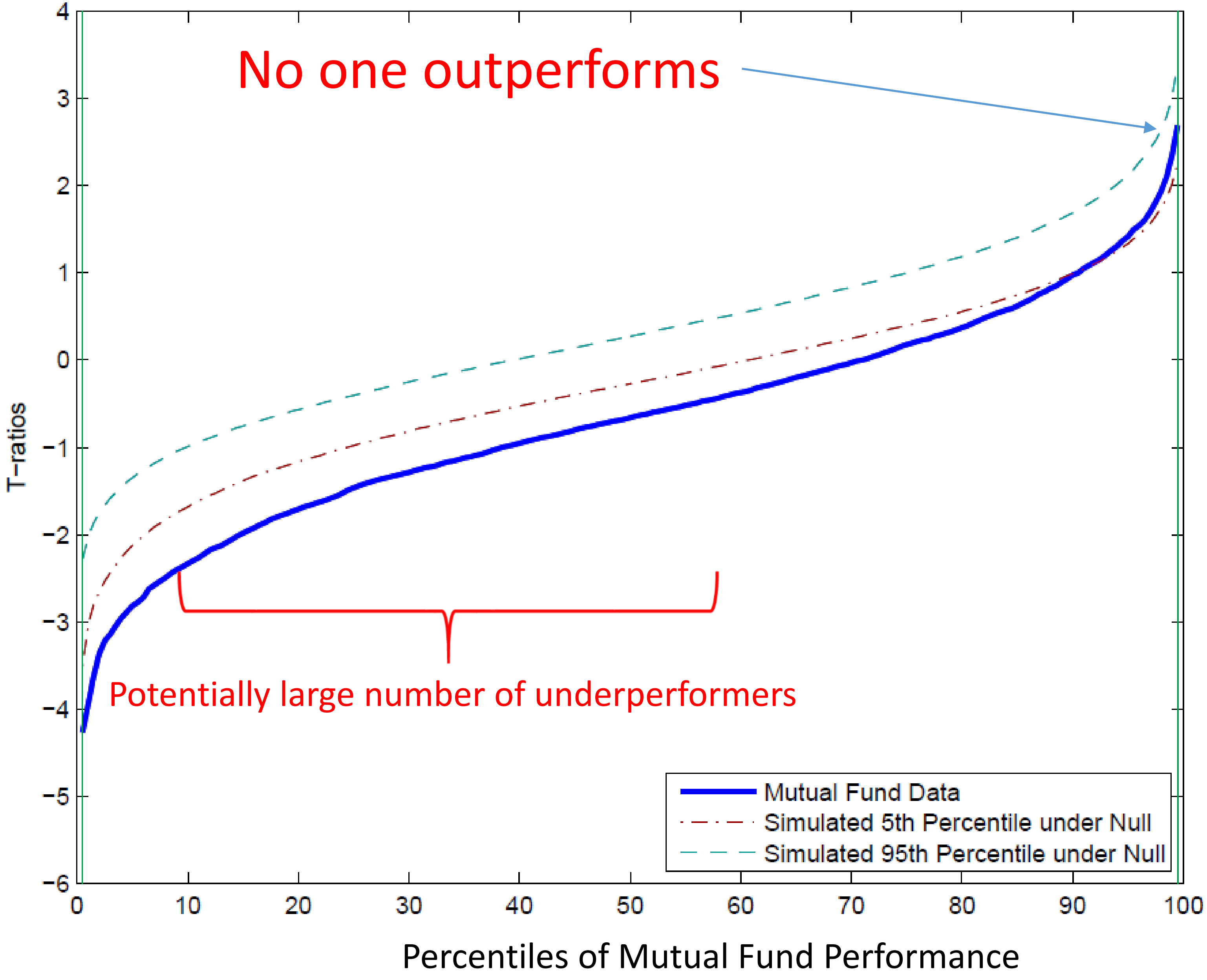
- Our technique similar (but has important differences) with Fama and French (2010)
- In FF 2010, each mutual fund is stripped of its “alpha”. So in the null (of no skill), each fund has exactly zero alpha and zero t-statistic.
- FF 2010 then bootstrap the null (and this has all of the desirable properties, i.e. preserves cross-correlation, non-normalities).

Fund Evaluation

- We depart from FF 2010 in the following way. Once, we declare a fund “true”, we replace it in the null data with its actual data.
- To be clear, suppose we had 5,000 funds. In the null, each fund has exactly zero alpha. We do the max and find Fund 7 has skill. The new null distribution replaces the “de-alphaed” Fund 7 with the actual Fund 7 data. That is, 4,999 funds will have a zero alpha and one, Fund 7, has alpha>0.
- We repeat the bootstrap

Fund Evaluation

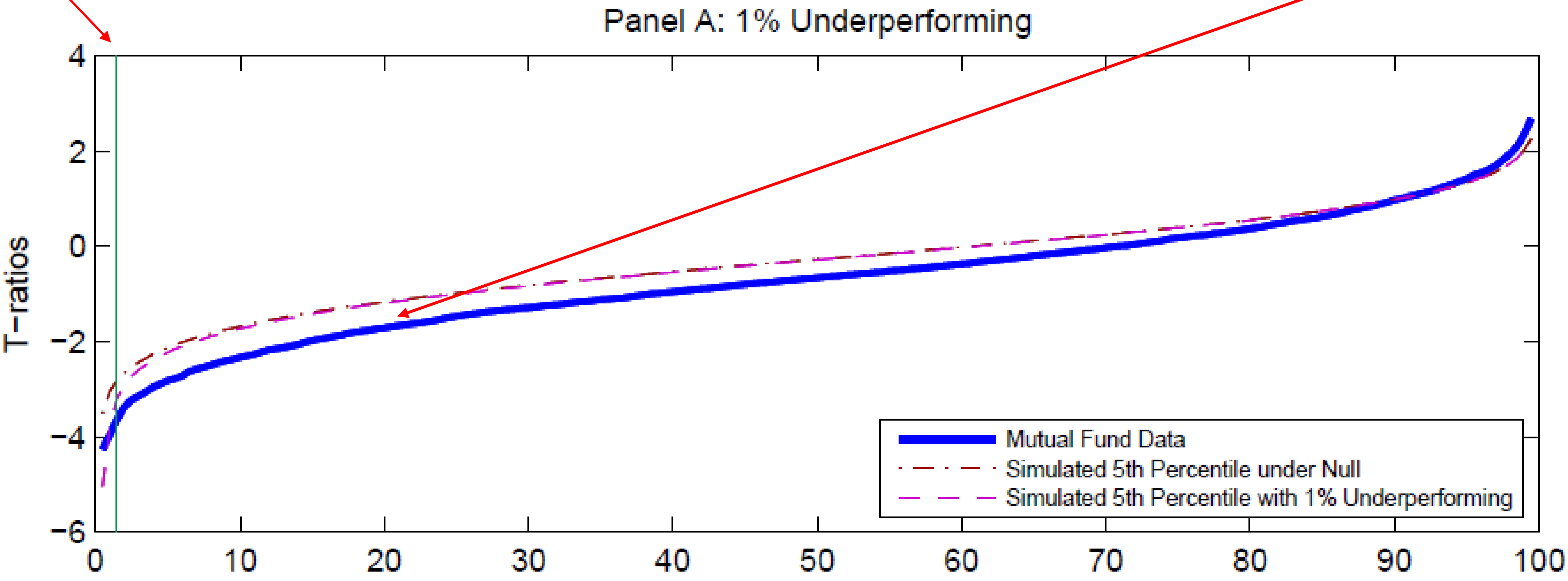
Null = No outperformers or underperformers



1% “True” underperformers added back to null

Fund Evaluation

Still there are more that appear to underperform

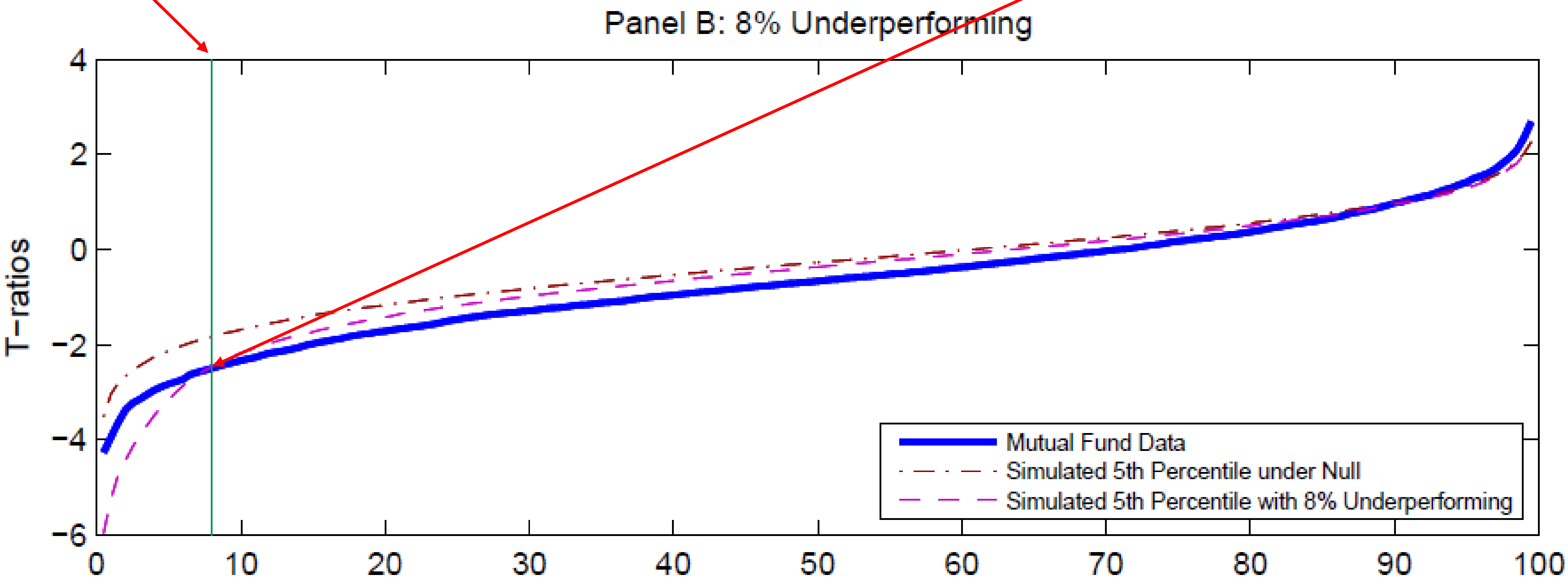


Percentiles of Mutual Fund Performance

8% “True” underperformers added back to null

Fund Evaluation

Cross-over point: Simulated and real data



Percentiles of Mutual Fund Performance

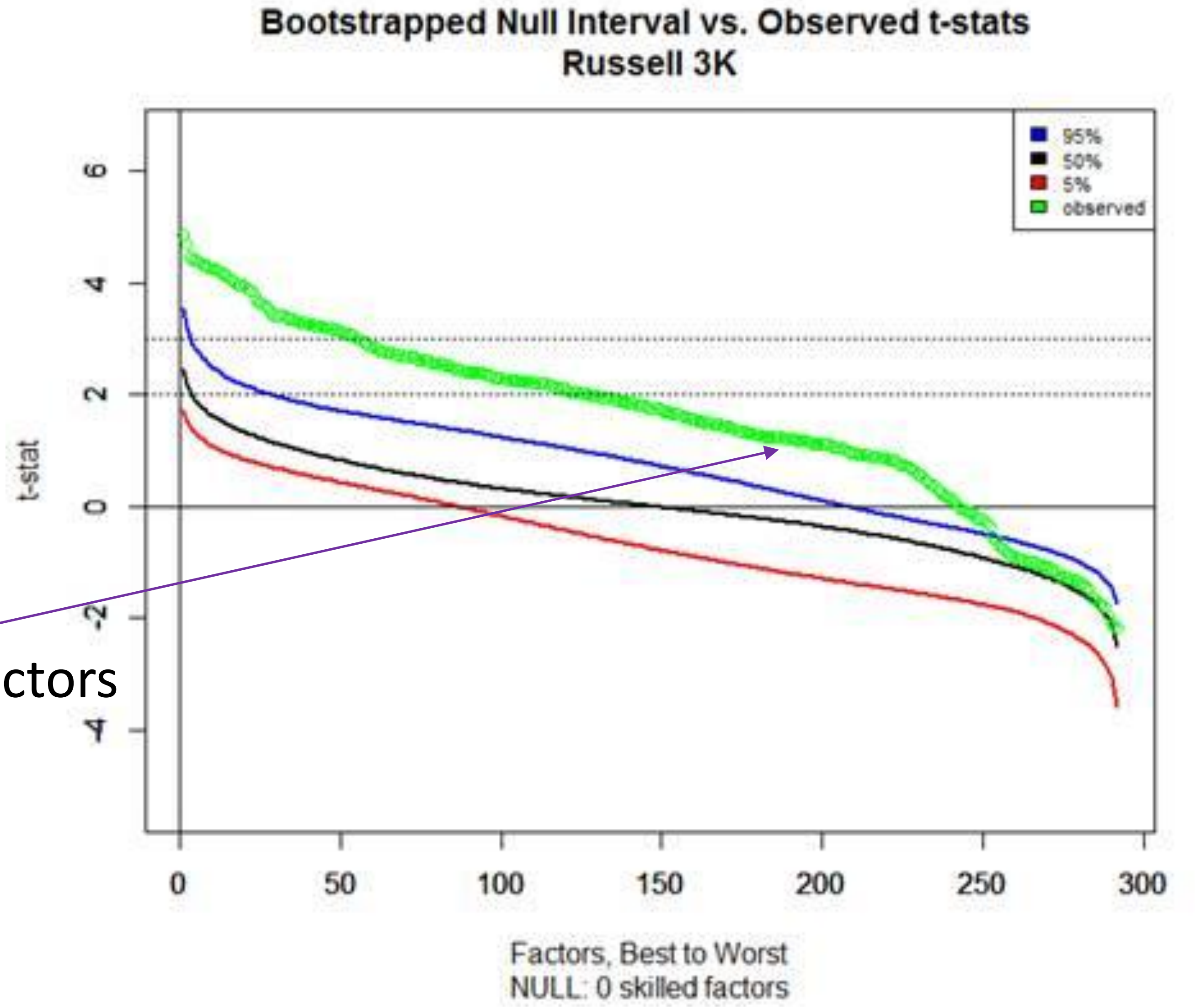
Factor Evaluation

- Easy to apply to standard factor models
- Think of each factor as a fund return
- Return of the S&P Capital IQ data* (thanks to Kirk Wang, Paul Fruin and Dave Pope). Application of Harvey-Liu done last week!
- 293 factors examined

*Note: Data from 2010, sector-neutralized, equal weighted, Q1-Q5 spread

Factor Evaluation

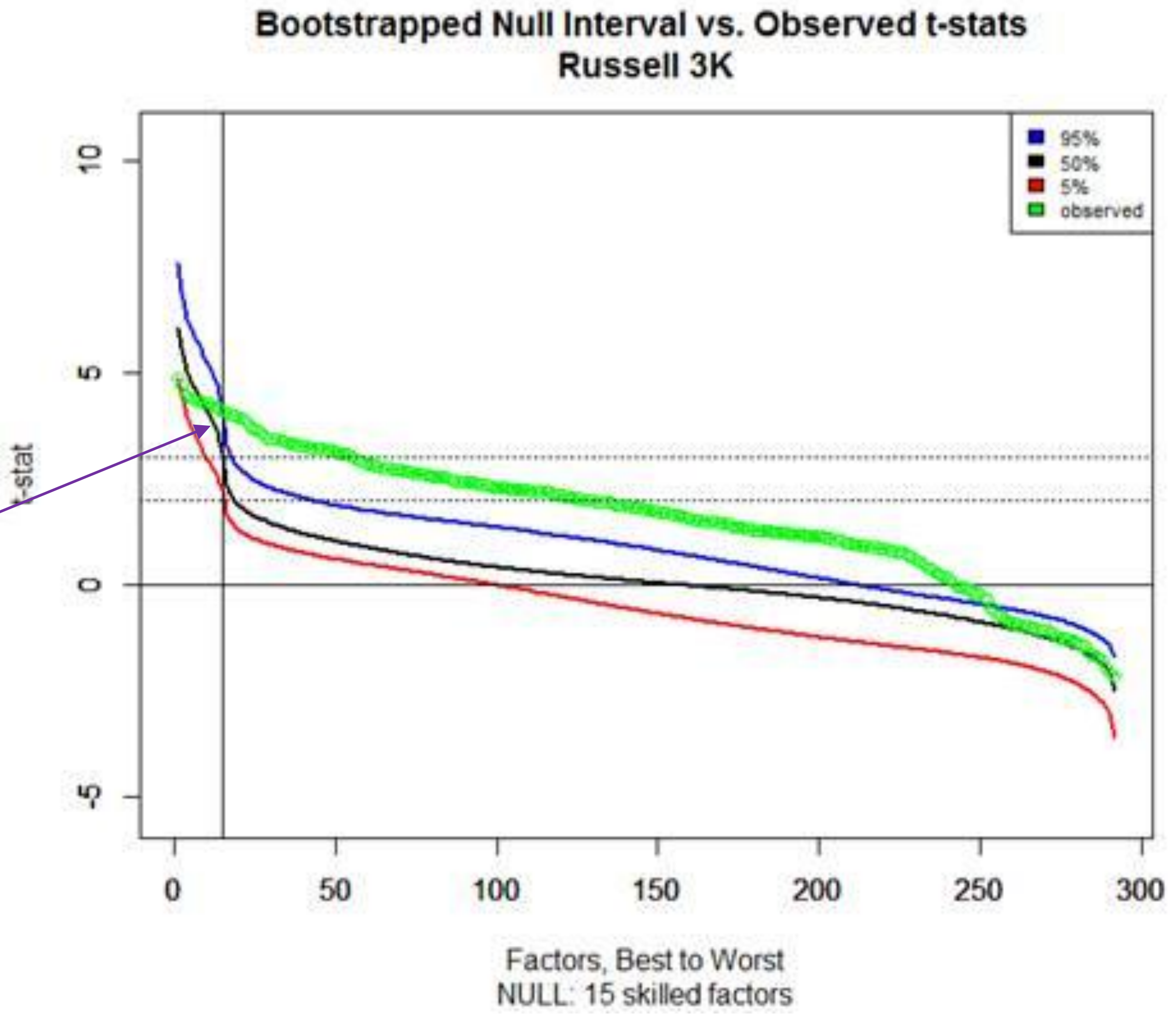
126 factors pass typical threshold of t-stat > 2
54 factors pass modified threshold of t-stat > 3



Large number of potentially "significant" factors

Factor Evaluation

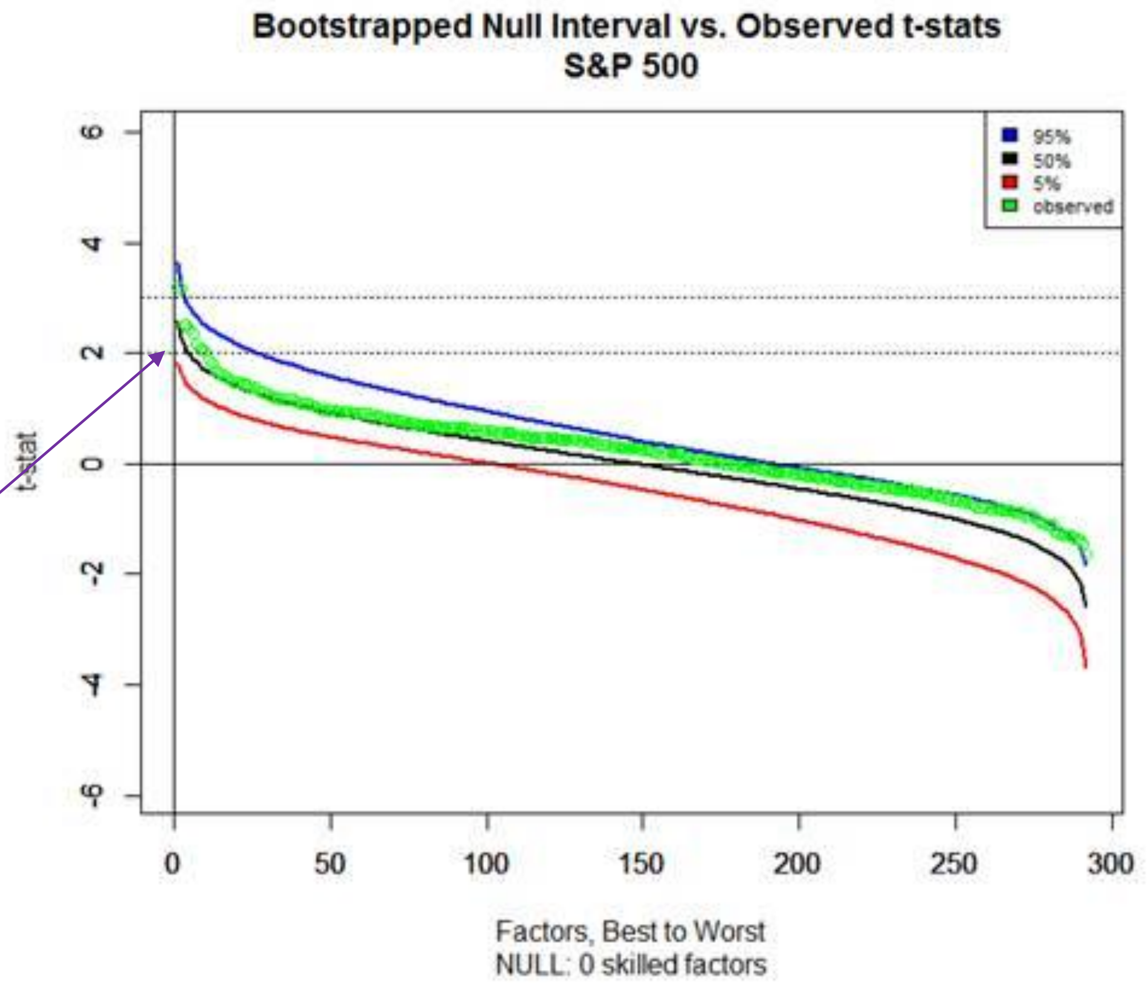
Only 15 declared "significant factors"



Factor Evaluation

Redo with S&P 500 universe.

Nothing significant.



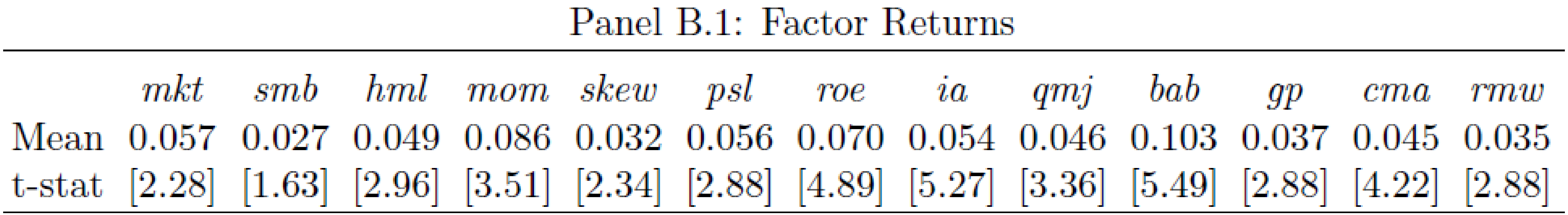
Factor Evaluation

- What about published factors?
- 13 widely cited factors:
  - MKT, SMB, HML
  - MOM
  - SKEW
  - PSL
  - ROE, IA
  - QMJ
  - BAB
  - GP
  - CMA, RMW

Factor Evaluation

- Use panel regression approach
- Illustrative example only
- One weakness is you need to specify a set of portfolios
- Choice of portfolio formation will influence the factor selection
- Illustration uses FF Size/Book to Market sorted 25 portfolios

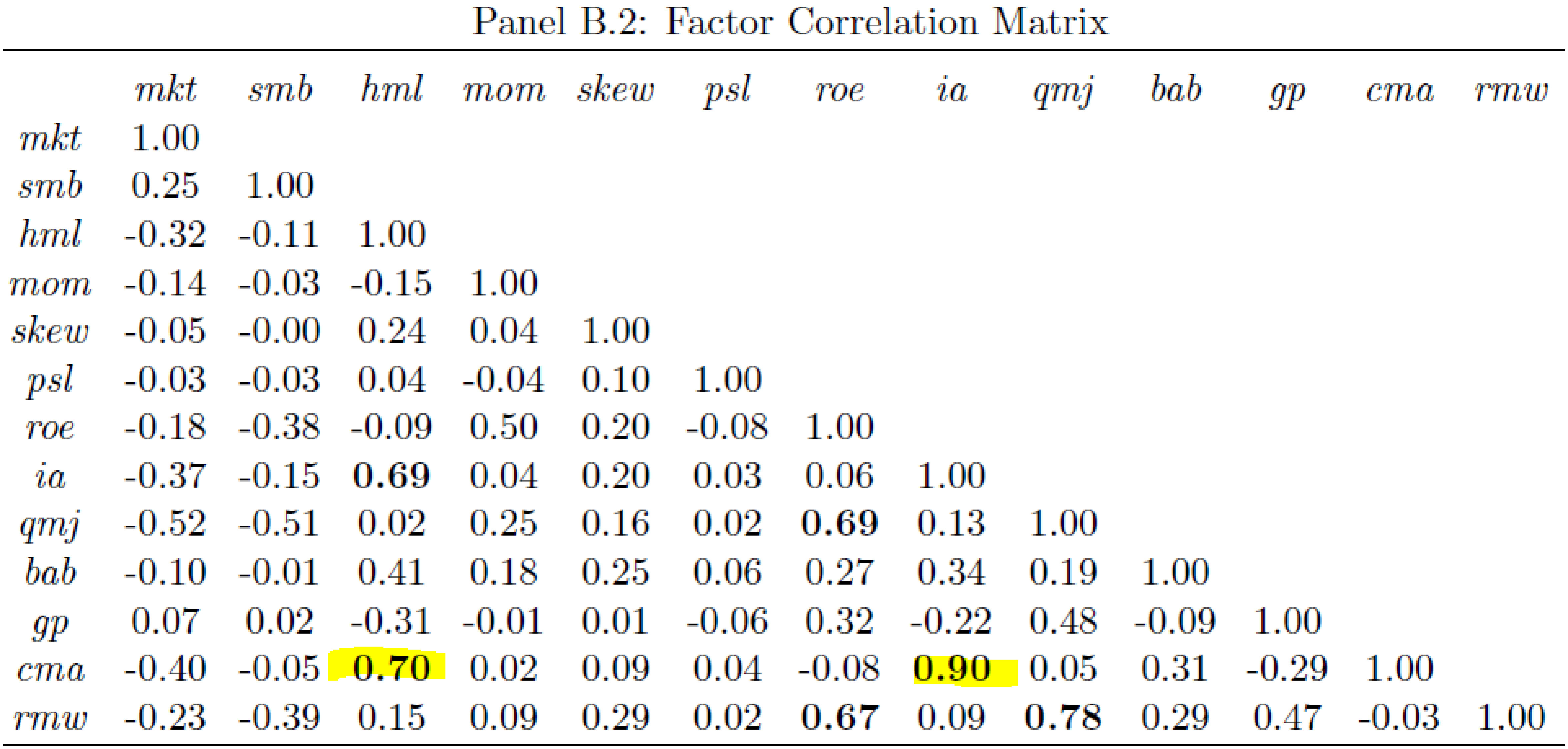
Factor Evaluation



Panel B.1: Factor Returns

	mkt	*smb*	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*
Mean	0.057	0.027	0.049	0.086	0.032	0.056	0.070	0.054	0.046	0.103	0.037	0.045	0.035
t-stat	[2.28]	[1.63]	[2.96]	[3.51]	[2.34]	[2.88]	[4.89]	[5.27]	[3.36]	[5.49]	[2.88]	[4.22]	[2.88]

Factor Evaluation



Panel B.2: Factor Correlation Matrix

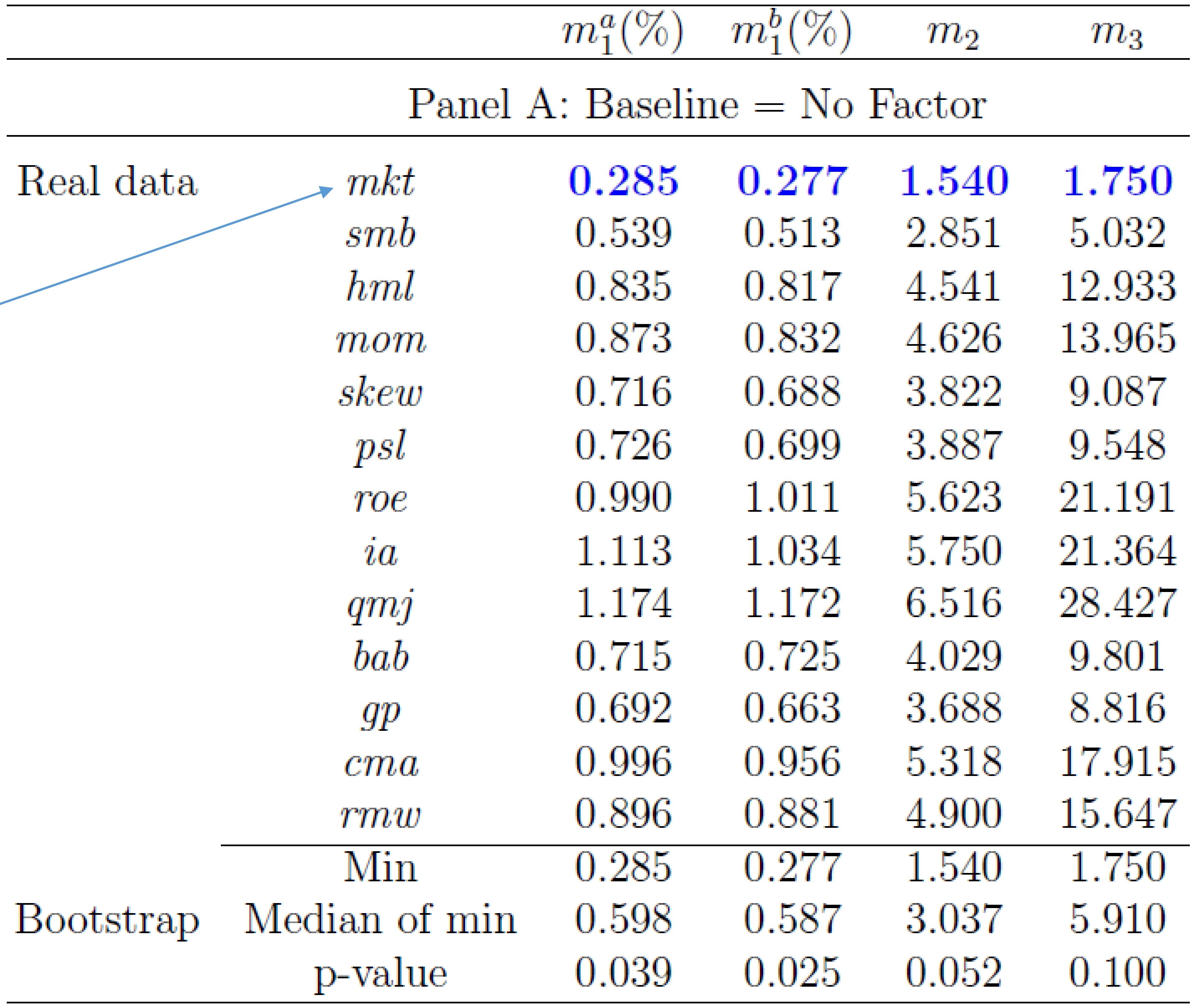
	mkt	*smb*	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*
mkt	1.00												
smb	0.25	1.00											
hml	-0.32	-0.11	1.00										
mom	-0.14	-0.03	-0.15	1.00									
skew	-0.05	-0.00	0.24	0.04	1.00								
psl	-0.03	-0.03	0.04	-0.04	0.10	1.00							
roe	-0.18	-0.38	-0.09	0.50	0.20	-0.08	1.00						
ia	-0.37	-0.15	**0.69**	0.04	0.20	0.03	0.06	1.00					
qmj	-0.52	-0.51	0.02	0.25	0.16	0.02	**0.69**	0.13	1.00				
bab	-0.10	-0.01	0.41	0.18	0.25	0.06	0.27	0.34	0.19	1.00			
gp	0.07	0.02	-0.31	-0.01	0.01	-0.06	0.32	-0.22	0.48	-0.09	1.00		
cma	-0.40	-0.05	**0.70**	0.02	0.09	0.04	-0.08	**0.90**	0.05	0.31	-0.29	1.00	
rmw	-0.23	-0.39	0.15	0.09	0.29	0.02	**0.67**	0.09	**0.78**	0.29	0.47	-0.03	1.00

Factor Evaluation

- Evaluation metrics
  - m_{1a} = median absolute intercept
  - m_1 = mean absolute intercept
  - m_2 = m_1/average absolute value of demeaned portfolio return
  - m_3 =mean squared intercept/average squared value of demeaned portfolio returns
  - GRS (not used)

Factor Evaluation

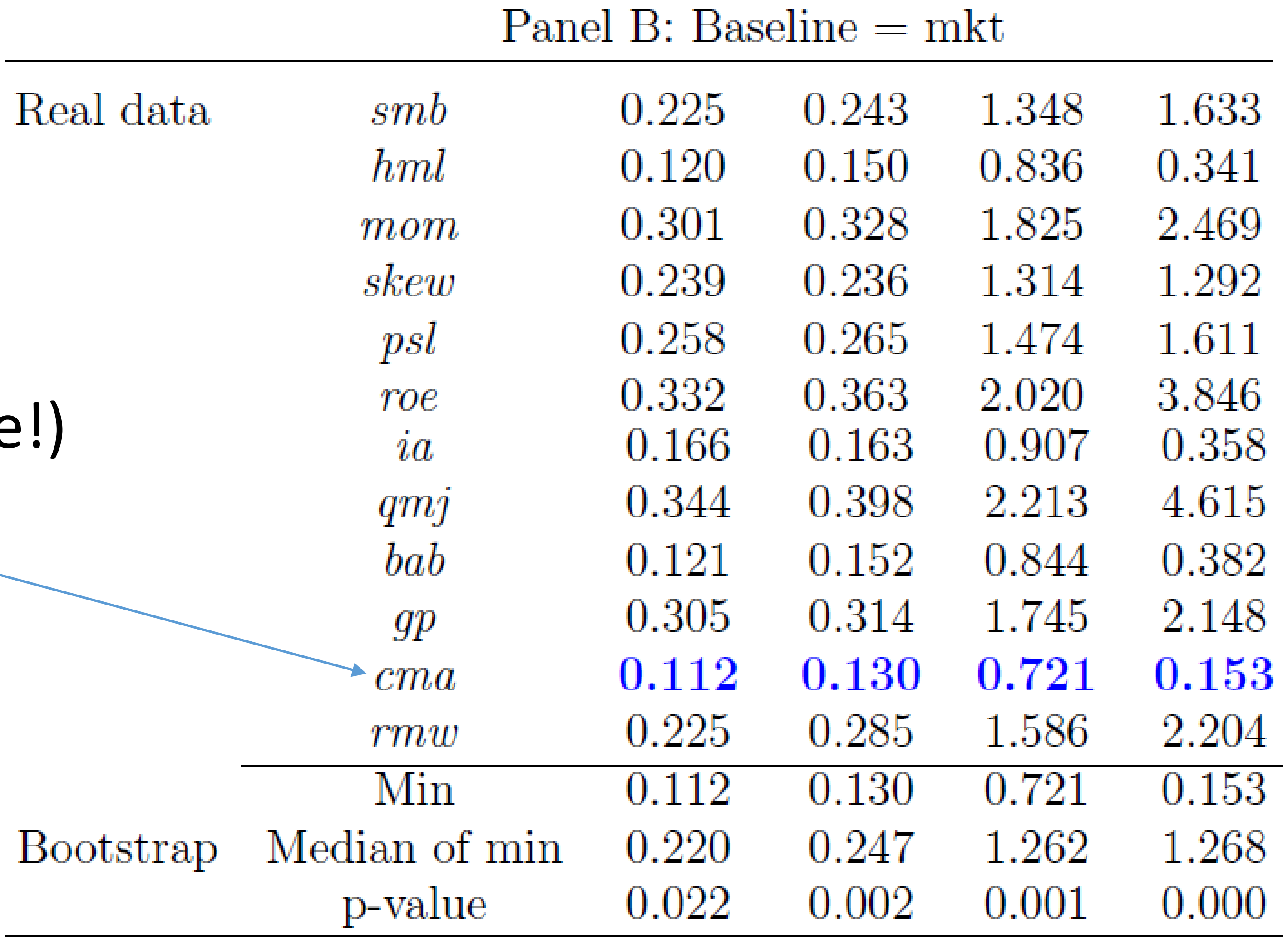
Select market factor first



		$m_1^a(\%)$	$m_1^b(\%)$	m_2	m_3
Panel A: Baseline = No Factor					
Real data	*mkt*	**0.285**	**0.277**	**1.540**	**1.750**
	smb	0.539	0.513	2.851	5.032
	hml	0.835	0.817	4.541	12.933
	mom	0.873	0.832	4.626	13.965
	skew	0.716	0.688	3.822	9.087
	psl	0.726	0.699	3.887	9.548
	roe	0.990	1.011	5.623	21.191
	ia	1.113	1.034	5.750	21.364
	qmj	1.174	1.172	6.516	28.427
	bab	0.715	0.725	4.029	9.801
	gp	0.692	0.663	3.688	8.816
	cma	0.996	0.956	5.318	17.915
	rmw	0.896	0.881	4.900	15.647
	Min	0.285	0.277	1.540	1.750
Bootstrap	Median of min	0.598	0.587	3.037	5.910
	p-value	0.039	0.025	0.052	0.100

Factor Evaluation

Next cma chosen (hml, bab close!)



Panel B: Baseline = mkt

Real data	*smb*	0.225	0.243	1.348	1.633
	hml	0.120	0.150	0.836	0.341
	mom	0.301	0.328	1.825	2.469
	skew	0.239	0.236	1.314	1.292
	psl	0.258	0.265	1.474	1.611
	roe	0.332	0.363	2.020	3.846
	ia	0.166	0.163	0.907	0.358
	qmj	0.344	0.398	2.213	4.615
	bab	0.121	0.152	0.844	0.382
	gp	0.305	0.314	1.745	2.148
	cma	0.112	0.130	0.721	0.153
	rmw	0.225	0.285	1.586	2.204
	Min	0.112	0.130	0.721	0.153
Bootstrap	Median of min	0.220	0.247	1.262	1.268
	p-value	0.022	0.002	0.001	0.000

Factor Evaluation

- This implementation assumes a single panel estimation
- Harvey and Liu (2015) *"Lucky Factors"* shows how to implement this in Fama-MacBeth regressions (cross-sectional regressions estimated at each point in time)

Factor Evaluation

- But…. the technique is only as good as the inputs
- Different results are obtained for different portfolio sorts

Factor Evaluation Using Individual Stocks

- Logic of using portfolios:
  - Reduces noise
  - Increases power (create a large range of expected returns)
  - Manageable covariance matrix

Factor Evaluation Using Individual Stocks

- Harvey and Liu (2015) "*A test of the incremental efficiency of a given portfolio*"
  - Yes, individual stocks noisier
  - No arbitrary portfolio sorts – input data is the same for every test
  - Avoid estimating the covariance matrix and rely on measures linked to average pricing errors (intercepts)

American Statistical Association

Ethical Guidelines for Statistical Practice, August 7, 1999.

II.A.8

- “Recognize that any frequentist statistical test has a random chance of indicating significance when it is not really present. Running multiple tests on the same data set at the same stage of an analysis increases the chance of obtaining at least one invalid result. Selecting the one "significant" result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading.”

Conclusions

- "More than half of the reported empirical findings in financial economics are likely false."

  Harvey, Liu & Zhu (2015) "*...and the Cross-Section of Expected Returns*"
- New guidelines to reduce the Type I errors
- Applies not just in finance but to any situation where many "X" variables are proposed to explain "Y"