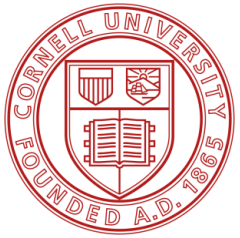


The 7 Reasons Most Econometric Investments Fail

Marcos López de Prado, Ph.D.

Advances in Financial Machine Learning

ORIE 5256

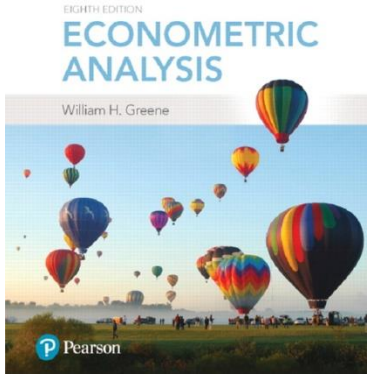
Key Points

- In general, financial problems are beyond the grasp of econometrics:
  - Financial data exhibits complex relationships: non-linear, threshold, hierarchical
  - Most financial data is non-numeric or unstructured: categorical, text, images, recordings, etc.
  - Financial datasets tend to be high-dimensional, with many variables and few observations
- Econometric models generally:
  - rely on *p*-values, or so-called "statistical significance," in violation of ASA protocols
  - are designed to adjudicate variance in-sample (not forecast values out-of-sample)
  - rely on strong assumptions that are not satisfied by financial phenomena
  - do not disentangle the specification search from the variable search
  - pay little or no attention to both forms of overfitting: training set and testing set
- Part of the problem is, most econometric tools originated in scientific fields, and are inadequate for tackling investment problems
- As a result, **most econometric-based investment strategies are likely false**

The Econometric Canon

What is Econometrics?

"[T]he concept of multiple regression and the linear regression model in particular constitutes the underlying platform of most econometric modeling, even if the linear model itself is not ultimately used as the empirical specification."
William Greene, Econometric Analysis (2012, p.7)

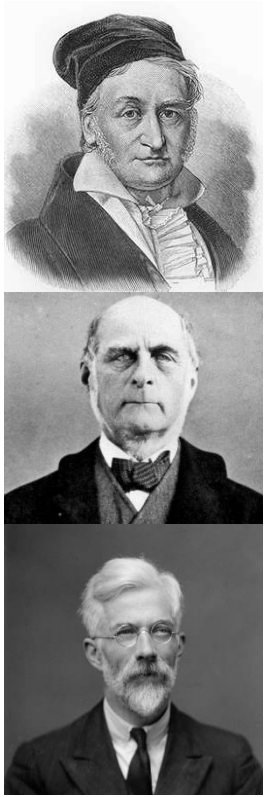


- In a general sense, econometrics encompasses the set of statistical methods applied to economic and financial data, with the purpose of providing empirical support to economic theories
- In practice, however, this set of statistical methods has traditionally concentrated on the multivariate linear regression model

The Origin of Regressions

Multivariate linear regressions are an old technology:

- 1795: Carl Friedrich Gauss applied ordinary least squares (OLS) to geodesic and astronomic datasets
  - Interestingly, Gauss thought that the OLS method was too trivial to merit publication
- 1886: Francis Galton (biostatistician) coined the term “regression”
  - Galton, a proponent of Eugenics, estimated linear equations to argue that hereditary human physical and moral traits exhibit a regression towards the mean
- 1900s: Karl Pearson (biostatistician) introduced
  - the concepts of correlation and regression coefficients
  - the method of moments
- 1920s: Ronald Fisher (biostatistician)
  - developed maximum likelihood estimation
  - popularized the notion of *p*-values, and statistical significance

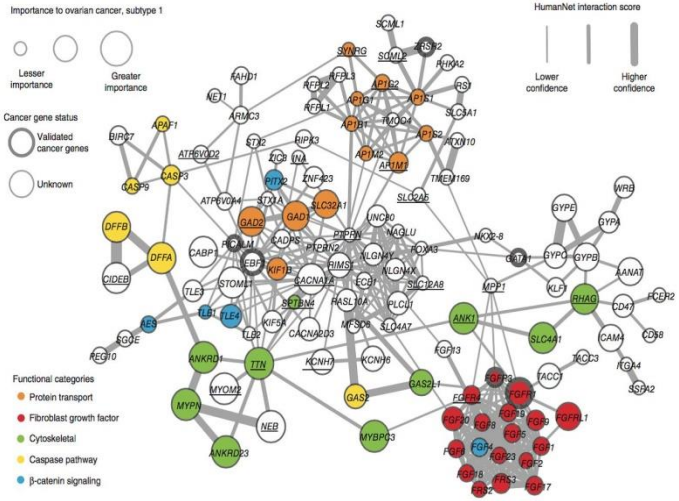


The Econometric Society

- By 1930, Galton, Pearson, Weldon et al. had succeeded at
  - formalizing Biology through the statistical study of biological datasets
  - founding the journal *Biometrika*
- 1930: Inspired by that success, a group of economists founded The Econometric Society, in Cleveland, Ohio
  - Founders: ***"[T]he chief purpose of such an association would be to help in gradually converting economics into a genuine and recognized science."***
- The Econometric Society achieved
  - defining the canon of what was acceptable empirical evidence in economics
  - founding the journal *Econometrica*
- However, it failed to bridge the gap with professional mathematicians or statisticians (unlike its Biology counterpart)
  - **Out of the ~700 Fellows of the Econometric Society, very few have advanced degrees in mathematics or statistics** (contrary to its interdisciplinary goal)

The Econometric Canon

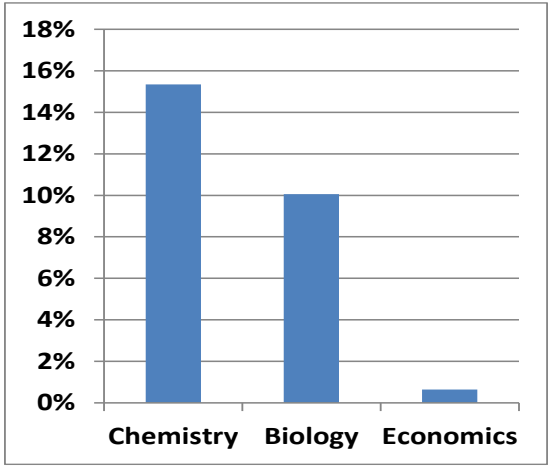
- The same quantitative canon used by economists is called
  - **biostatistics** when applied to biological datasets
  - **chemometrics** when applied to chemical datasets
- However, entry-level biostatistics and chemometrics textbooks often include modern topics, such as
  - Entropy metrics, advanced clustering, classification, graph theory, pattern recognition, computational methods, ...



- [Computational methods](#) have become particularly important in scientific fields, because they can replace some (likely unrealistic) assumptions regarding the data-generating process
- These topics are largely absent in popular econometrics textbooks

Econometric Stagnation

- The Web of Science reports that **13,772 journal articles** have been published on subjects in the intersection of "Economics" and "Statistics & Probability" (as of Nov. 2018)
- Among those publications, **only 89 articles (0.65%)** contained any of the following terms:
  - classifier, clustering, neural network, machine learning

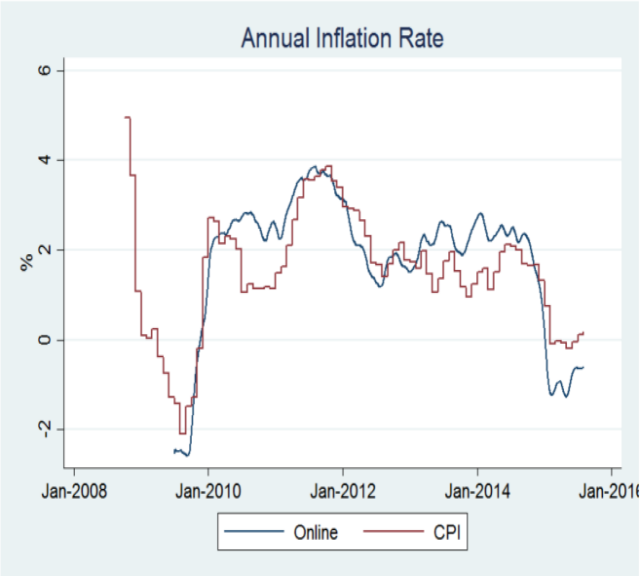


- In contrast, out of the 40,283 articles in the intersection of "Biology" and "Statistics & Probability," a total of **4,049 (10.05%)** contained any of those terms
- Out of the 4,994 articles in the intersection of "Chemistry, Analytical" and "Statistics & Probability," a total of **766 (15.34%)** contained any of those terms

Pitfall #1:
Structured Data

Econometric Data is Relatively Uninteresting

- The most informative datasets are amenable to machine learning, but not to econometrics
  - **Unstructured data**: less than 20% of all data available is structured (IDC [2014])
  - **High-dimensional datasets**: the number of variables often exceeds the number of observations
  - **Sparse and/or noisy datasets**: a large proportion of zeros, or noise, per unit of signal
  - **Hierarchical relations**: economic systems often involve networks of agents, and clustering of dependencies
- Standard econometric transformations achieve stationarity at the expense of wiping out all memory
- **Econometric methods effectively model uninteresting data**



Inflation estimates derived from unstructured online prices produce accurate predictions of annual inflation statistics for the U.S.
Source: Cavallo and Rigobon [2016]

Example of Unstructured Data

In the plot below, an algorithm has identified news articles containing information relevant to Tesla.

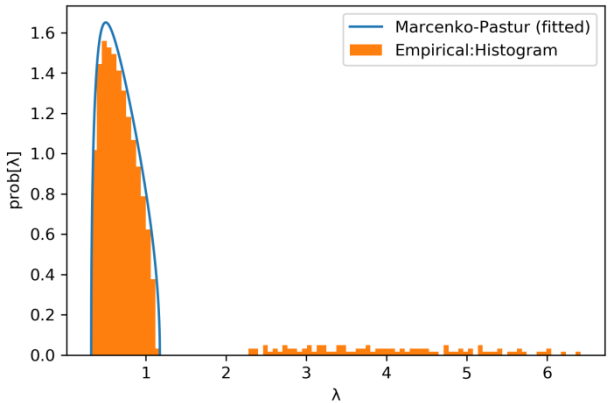
- **Blue bars**: Daily count of the total number of articles. The average is 458 articles/day, with a maximum of ~5000.
- **Green bars**: Daily count of articles expressing a positive sentiment.
- **Red bars**: Daily count of articles expressing a negative sentiment.

An immediate price reaction occurs contemporaneously with strong *sentiment imbalance*. Still, there appears to be a residual price momentum following the initial breaking news. A machine learning algorithm can be trained to identify when prices are most sensitive to sentiment imbalance.

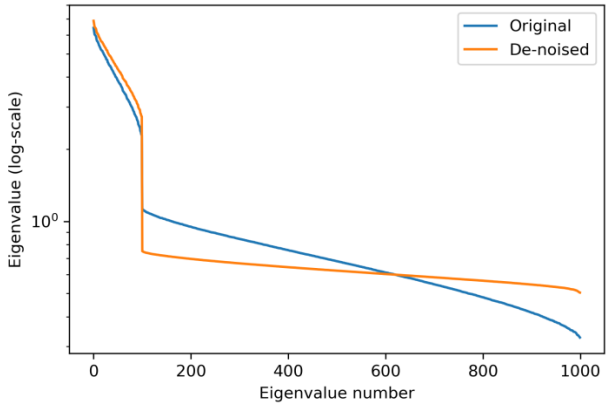
Pitfall #2:
Correlations / Betas

Financial Correlations Are Extremely Noisy

- The econometric canon does not include methods to de-noise and de-tone correlation matrices
- As a result, most econometric studies reach spurious conclusions, supported by noise, not signal



Almost all eigenvalues contained in a financial correlation matrix are associated with noise, not signal. Econometric studies estimate betas that reflect spurious relationships



Mathematical approaches can determine which eigenvalues must be treated numerically to prevent false discoveries, however those approaches are rarely used in econometric studies (N.B.: shrinkage fails to separate signal from noise)

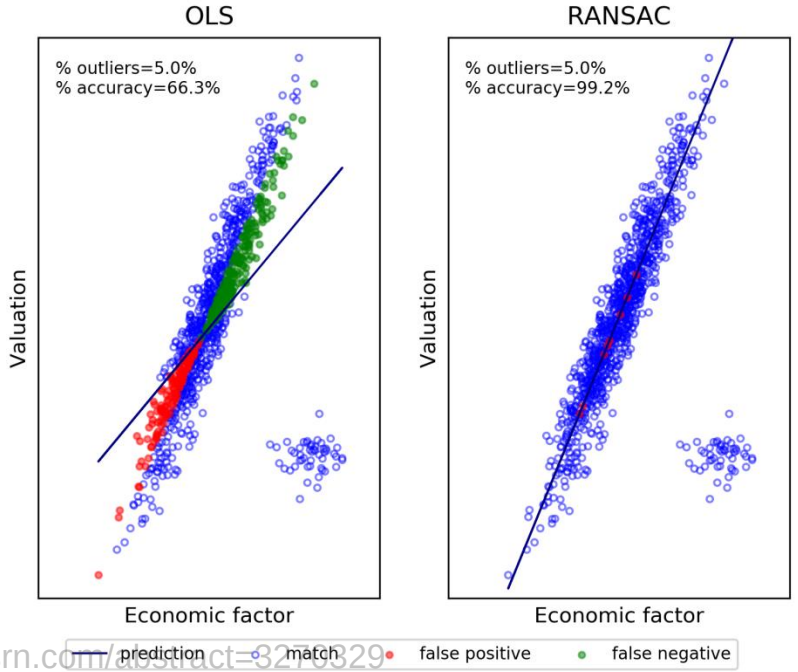
Correlations and Outliers

Cross-sectional studies are particularly sensitive to the presence of outliers. Even a small percentage of outliers can cause a very large percentage of wrong signals: buys that should be sells (false positives), and sells that should be buys (false negatives).

In this plot we run a regression on a cross-section of securities, where a very small percentage (only 5%) are outliers:

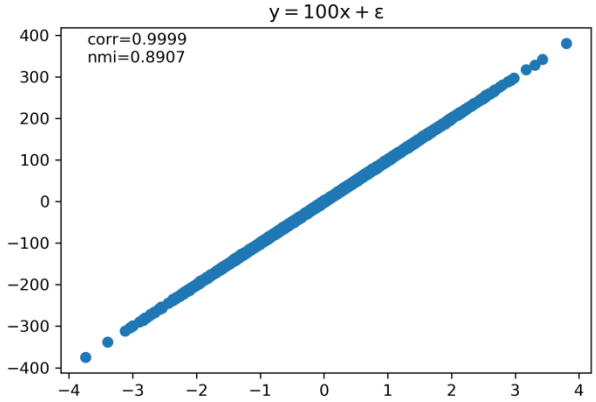
- the **red dots** are securities that are expensive, but the regression wrongly classified as cheap.
- the **green dots** are securities that are cheap, but the regression wrongly classified as expensive.

With only 5% of outliers, the cross-sectional regression produces a 34% classification error. In contrast, the RANSAC algorithm achieves a classification error of only 1% (mostly borderline cases). For more examples, see this video.

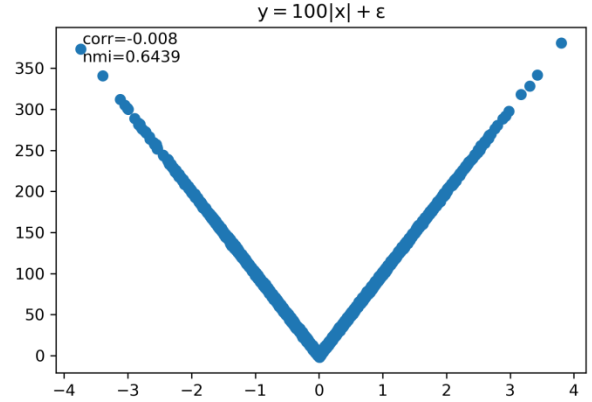


Correlations/Betas Miss Most Dependencies

- Correlation is a useful measure of linear codependence, however it is flawed:
  - Most codependencies in finance are non-linear
  - Correlations are heavily influenced by outliers
  - Correlations have limited use outside multivariate Normal distributions
  - Correlations do not model activation thresholds and regime switches



Example 1: Linear relationship
Correlation is approximately 1, and the normalized mutual information (NMI) is also very high, approximately 0.9



Example 2: Non-linear relationship
Despite of the strong dependence between x and y, correlation is approximately 0. In contrast, the normalized mutual information is still very high

Pitfall #3:
Variance Adjudication and the Causality Fallacy

The Wrong Goal For Investing

- Econometric specifications attempt to *adjudicate* to $\{X_{t,i}\}$ *the variance* of $\{y_t\}$ *in-sample*, while controlling for the variance adjudicated to $\{Z_{t,j}\}$

$$y_t = \alpha + \sum_{i=1}^{I} \beta_i X_{t,i} + \sum_{j=1}^{J} \gamma_j Z_{t,j} + \varepsilon_t$$

- This poses two problems:
  - in-sample adjudication is rarely useful for strategy development
  - the model requires the adjudication of variance to the control variables $\{Z_{t,j}\}$ too!
- In contrast, machine learning methods attempt to use $\{X_{t,i}\}$ to *forecast* $\{y_t\}$ *out-of-sample*, while controlling for $\{Z_{t,j}\}$ (regardless of the specification)
- **Regression is the wrong tool for investing**: Econometrics borrowed its toolkit from Biology, where out-of-sample forecasting is not a critical goal

The Causality Fallacy

- Chen and Pearl [2013] found that six of the most influential Econometrics textbooks make fundamental mathematical and statistical mistakes:
  - confound correlation with causation
  - confound prediction with causation
  - confound causality with Granger-causality (a misnomer)
  - fail to provide coherent mathematical notation that distinguishes causal from statistical concepts
- This general state of confusion leads to spurious claims of causation, which translate into false investment strategies

"The introduction of graphical models and distinct causal notation into elementary econometric textbooks has the potential of revitalizing economics education and bringing next generation economists to par with modern methodologies of modeling and inference."

Pitfall #4:
Specification-Interaction Search

Econometric Entanglement

- Consider the typical Econometric model:

$$y_t = \alpha + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t} x_{2,t} + \varepsilon_t$$

- This requires that the researcher gets two items right at once:
  - The predictive variables: $\{x_1, x_2\}$
  - The functional form: typically a linear specification, with multiplicative interaction effects
- Given how complex financial systems are, these are unrealistic demands
- Even if the researcher can guess what are the relevant variables involved in a phenomenon, often times she will be unable to identify *ex ante* the precise functional form, including all of the interaction effects

False Econometric Conclusions

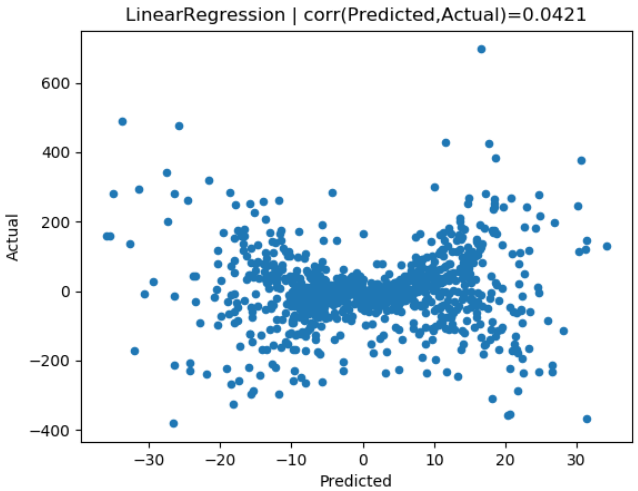
Consider data generated by a simple process with interaction effects, like

$$y_t = x_{1,t} + x_{2,t} + \boldsymbol{x_{1,t}x_{2,t}} + \varepsilon_t$$

Suppose that we get the variables right, however we fail to recognize the interaction effect, testing instead

$$y_t = \alpha + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \varepsilon_t$$

The correlation between predictions and realizations is only **0.04**, *even though we have provided the correct variables to the model* (x_1, x_2).



Traditional econometric models do not "learn" the structure of the data.

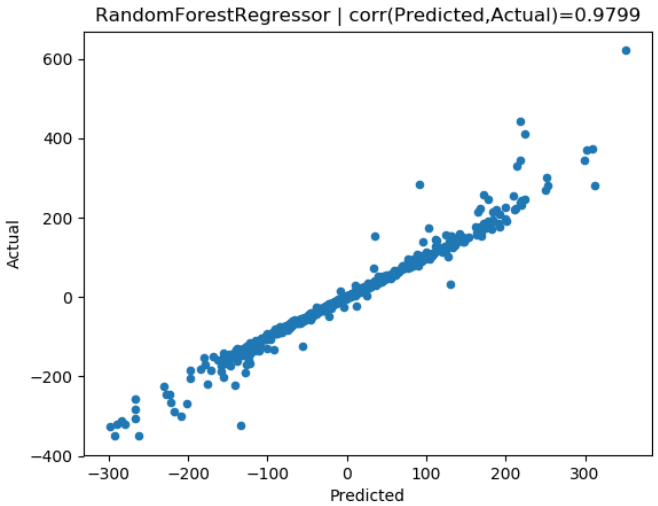
Unless we provide the exact (unknown) specification, we will reach false conclusions.

Machine Learning Disentanglement

Alternatively, we could follow a machine learning approach: estimate 1,000 Decision Trees by bootstrap, and form an ensemble forecast.

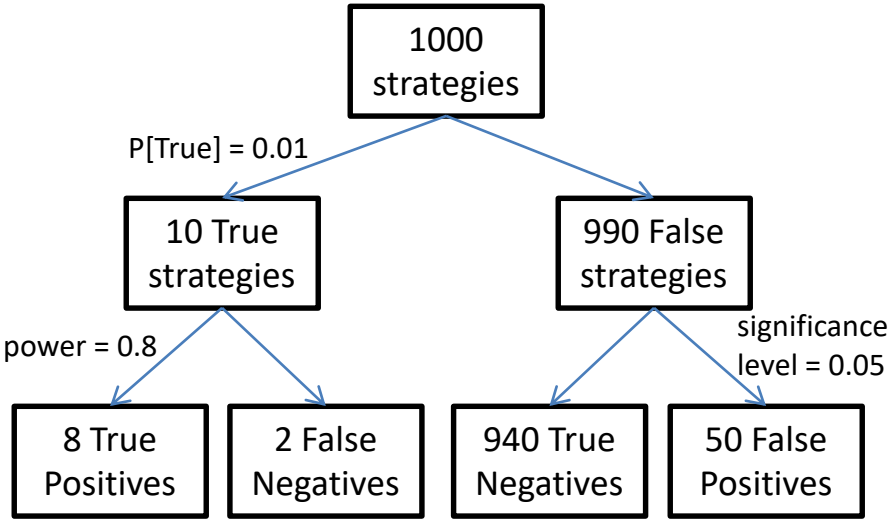
Like before, we do *not* inform the algorithm about the presence of interactions effects. Unlike before, the algorithm has "learned" the correct model specification for the provided variables.

The out-of-sample correlation between the predicted y_t and the actual y_t rises to **0.98**, thanks to the algorithm's ability to "learn" the structure of the data, without us directing that search.



Pitfall #5:
p-Values

At p=0.05, Most Strategies are False



Suppose that the probability of a backtested strategy being profitable is 1%.

Then, at the standard thresholds of 5% significance and 80% power, researchers are expected to make 58 discoveries out of 1000 trials, where 8 are true positives and 50 are false positives.

Under these circumstances, **a p-value of 5% implies that at least 86% of the discoveries are false!**

In practice, the false discovery rate in Finance is much greater than 86%, because:

- the familywise significance level in Finance is much greater than 5%, due to multiple testing.
- the probability of discovering a new investment strategy is lower than 1%, as a result of competition.
- the shelf-life of a strategy is short. Strategies do not remain "true" for more than a few months or years.
- specification errors and other assumption violations further increase the false discovery rate.

Econometrics is Stuck in the "p<0.05" Era

- Most findings in financial economics rely on a "p<0.05" argument, even though
  - *p*-values require strong (unrealistic) assumptions, such as correct specification, mutually-uncorrelated regressors, white noise residuals that follow a Normal distribution, etc.
  - in the common case of multicollinear regressors, *p*-values cannot be robustly estimated
  - *p*-values evaluate an irrelevant probability, $\mathrm{p}[X > x|H_0]$. What we really care about is $\mathrm{p}[H_1|X > x]$
  - *p*-values assess significance in-sample, not out-of-sample
- "Statistically significant" factors discovered through *p*-values include:
  - Value, Momentum, Quality, Size, Carry, Liquidity, Defensiveness, etc.
- **The misuse of *p*-values is so widespread that the American Statistical Association has discouraged their application going forward as a measure of statistical significance** (Wasserstein et al. [2019])
- This casts doubt over decades of econometric studies (the "factor zoo")

An Experiment with *p*-values

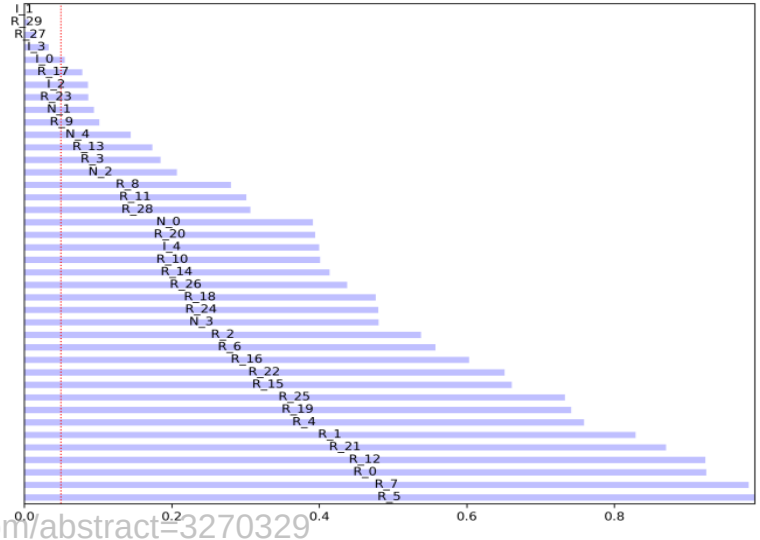
Consider a binary random classification problem composed of 40 features, where 5 are informative, 30 are redundant, and 5 are noise.

- **Informative features** (marked with the "**I_**" prefix) are those used to generate labels
- **Redundant features** (marked with the "**R_**" prefix) are those that are formed by adding Gaussian noise to a randomly chosen informative feature
- **Noise features** (marked with the "**N_**" prefix) are those that are not used to generate labels

The horizontal bars report the *p*-values from a Logit regression, and the vertical dash line marks the 5% significance level.

These *p*-values misrepresent the ground truth:

- only 4 out of the 35 non-noise features are deemed significant
- noise features are ranked as relatively important
- 14 of the features ranked as least important are not noise



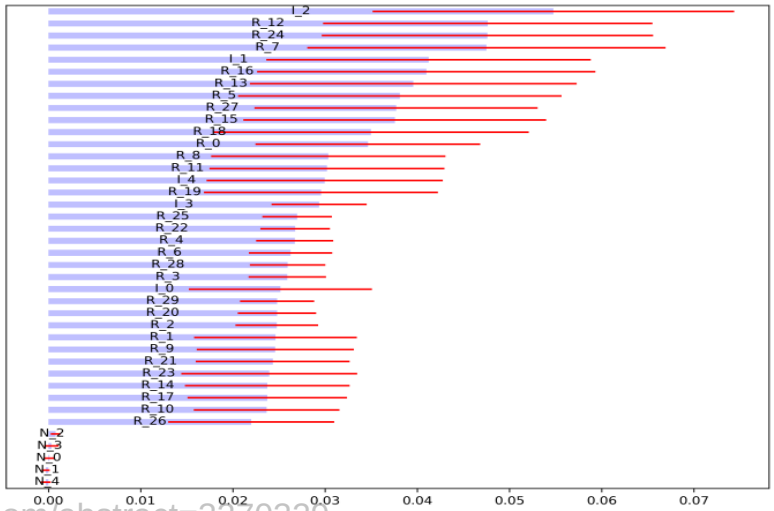
A Modern Approach to Feature Importance

We can repeat the same experiment where, instead of computing *p*-values, we apply the Mean Decrease Accuracy method (MDA):

1. fit a machine learning algorithm on the training set, and estimate the out-of-sample accuracy
2. shuffle one variable at a time, and re-estimate the out-of-sample accuracy
3. evaluate the decay in out-of-sample accuracy from shuffling each variable

MDA results are consistent with the ground truth:

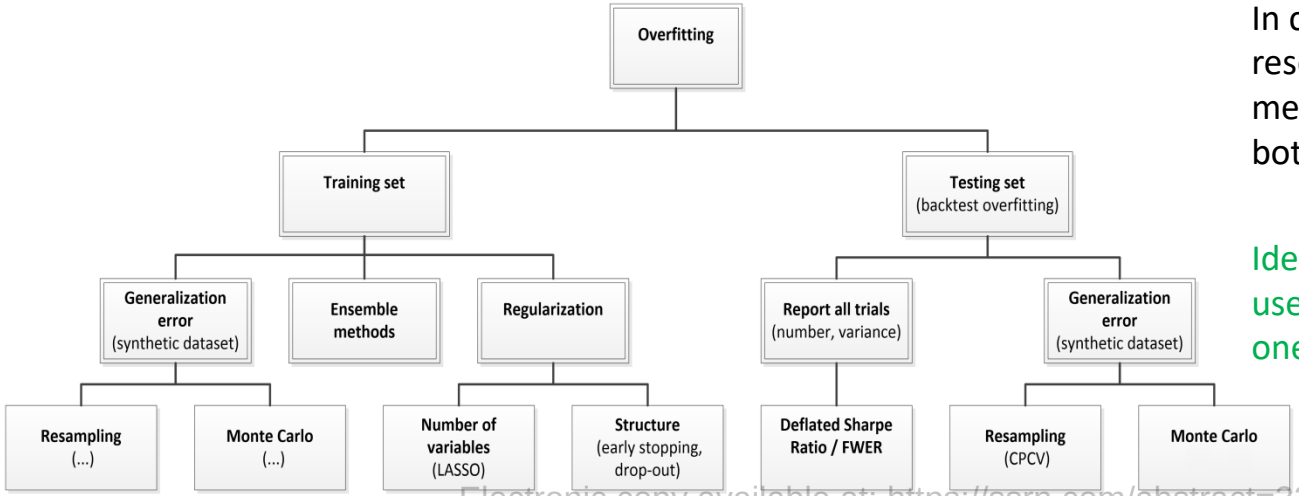
- MDA does a good job overall at separating noise features from the rest. Noise features are ranked last
- noise features are also deemed unimportant in magnitude, with MDA values of essentially zero
- results can be further improved by clustering together similar variables (an unsupervised learning approach)



Pitfall #6:
Training-Set Overfitting

The Two Types of Overfitting

- A statistical model can be overfit in two ways:
  - **Training-set Overfitting**: the specification is so flexible that it explains the noise (instead of signal)
  - **Testing-set Overfitting**: the model is chosen based on testing set performance (while concealing that alternative models performed worse)
- The econometric canon largely fails to address and quantify both problems



In contrast, machine learning researchers have developed methods to address and quantify both forms of overfitting.

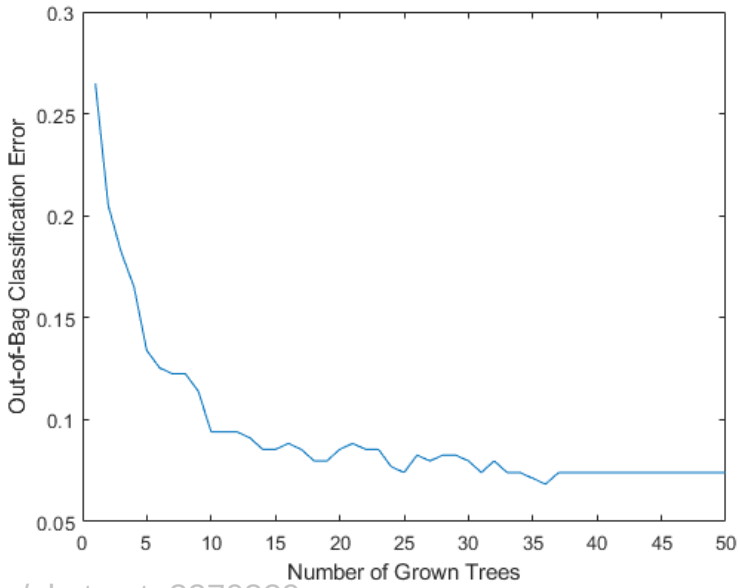
Ideally, *all* methods should be used concurrently, rather than one of each type.

Ensemble Example: Bootstrap Aggregation

1. Select a subset of the data, by random sampling with replacement
2. Fit a machine learning algorithm on (1)
3. Use (2) to make a prediction on data points not included in (1)
   - the error in this prediction is called “out-of-bag error”
4. Repeat (1)-(3) many times

Bootstrap aggregation (bagging) achieves two goals:

- it reduces the variance of the forecasting errors (see plot)
- if the individual estimators are minimally accurate, the accuracy of the *combined forecast* exceeds the average accuracy of the individual classifiers



Pitfall #7:
Testing-Set Overfitting

How Financial Firms Conduct Research

- Suppose that you are looking for an investment strategy. You run multiple econometric regressions, and find results that achieve high Sharpe ratios, some of them above 3
- You show these results to your boss, who decides to paper-trade the strategy for a few weeks. Luckily, paper trading performance seems consistent with the backtest, so the investment committee approves its deployment
- The strategy receives a $100 million allocation, but unfortunately a 20% loss follows shortly after
- The strategy never performs as advertised, and it is eventually decommissioned
- **What happened?**

The Most Important Plot in Finance

The y-axis displays the distribution of the maximum Sharpe ratios (max{SR}) for a given number of trials (x-axis).

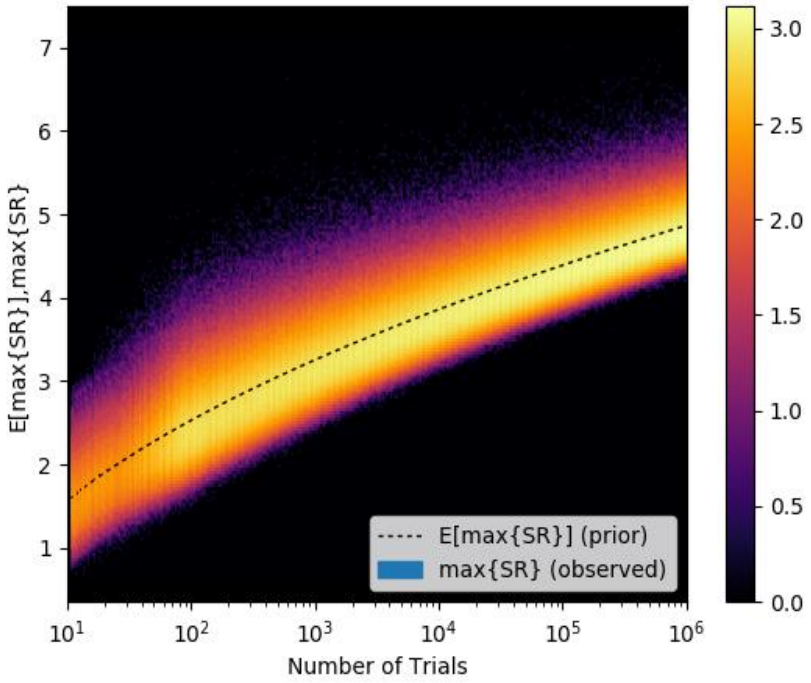
A lighter color indicates a higher probability of obtaining that result, and the dash-line indicates the expected value.

For example, after only 1,000 backtests (trials), the expected maximum Sharpe ratio (E[max{SR}]) is 3.26, **even if the true Sharpe ratio of the strategy is 0!**

The best backtested outcome is not representative of the ground truth (**selection bias**).

Limiting backtests to a few specifications suggested by theory is not a solution. It is a form of **confirmation bias**.

Most econometric investments fail because asset managers and academic journals do not control for selection bias *and* confirmation bias (see p.29).



A Solution

Overcoming the Limitations of Econometrics

Financial firms and academic authors could modernize their statistical methods.

Such modernization is within reach: For every step in econometric analysis, there is a counterpart in the machine learning toolkit.

STEP	ECONOMETRICS	ML
Goal Setting	Variance adjudication (in-sample)	Out-of-sample prediction
Visualization	Time plots, scatter plots, histograms	t-SNE, networks, treemaps, etc.
Outlier detection	Winsorizing, trimming, Dixon's Q test, etc.	Anomaly detection methods, RANSAC
Feature extraction	PCA	Kernel-PCA, LDA, biclustering
Regression	Algebraic models	Neural networks, SVR, GA, regression trees, etc.
Classification	Logit, probit	RF, SVC, k-NN, etc.
Feature importance	*p*-values	MDI, MDA per cluster
Model selection / overfitting prevention	Forward selection, backward elimination, stepwise	Regularization, bagging, boosting, early stopping, drop-out, pruning, bandwith, etc.
Goodness of fit	Adjusted R-squared (in-sample)	Out-of-sample (cross-validated): Explained variance, accuracy, F1, cross-entropy

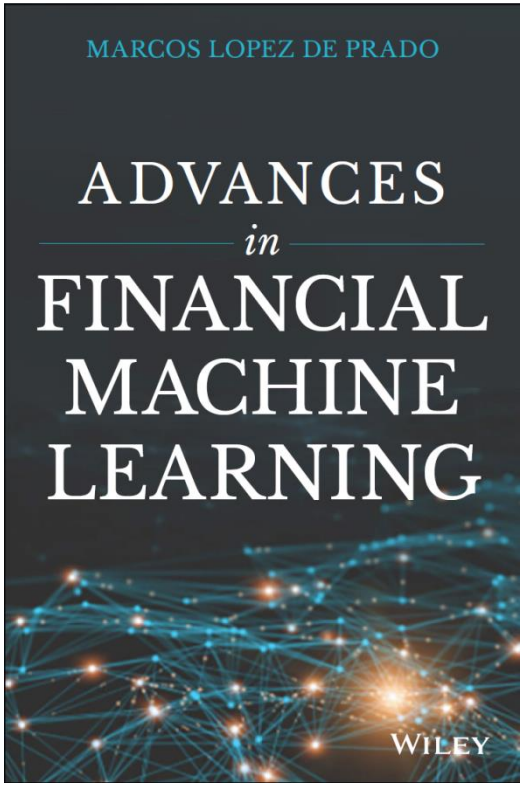
The Journal of Financial Data Science

"Until recently, the depth and breadth of datasets available to financial researchers was, to put it mildly, extremely shallow [...] This state of data paucity set a hard limit on the sophistication of the techniques that financial researchers could use. In that financial paleo-data age, the linear regression method was a reasonable choice, even though most of us suspected that the linearity assumption may not provide a realistic representation of a system as complex and dynamic as modern financial markets.

Today, we live in a different era, the age of financial Big Data. Researchers have at their disposal datasets that only a few years ago were unimaginable […] The size, quality, and variety of these sources of information, combined with the power of modern computers, allow us to apply more sophisticated mathematical techniques."

The Journal of Financial Data Science, Winter 2019, 1(1): 1-3

For Additional Details



The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.
— Prof. **Frank Fabozzi**, EDHEC Business School. Editor of The Journal of Portfolio Management.

THANKS FOR YOUR ATTENTION!

Disclaimer

- The views expressed in this document are the author's and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-