The 7 Reasons Most Econometric Investments Fail

Park Sukjin

Department of Economics, Sogang University

August 31, 2021

The 7 Reasons Most Econometric Investments Fail

- Source: López de Prado, Marcos, The 7 Reasons Most Econometric Investments Fail (April 16, 2019)

Key Points

- Econometric models generally
  - are designed to explain variance in-sample
  - rely on p-values, or so called statistical significance
  - rely on strong assumptions that are not satisfied by financial phenomena
  - do not disentangle the specification search from the variable search
  - pay little attention to overfitting

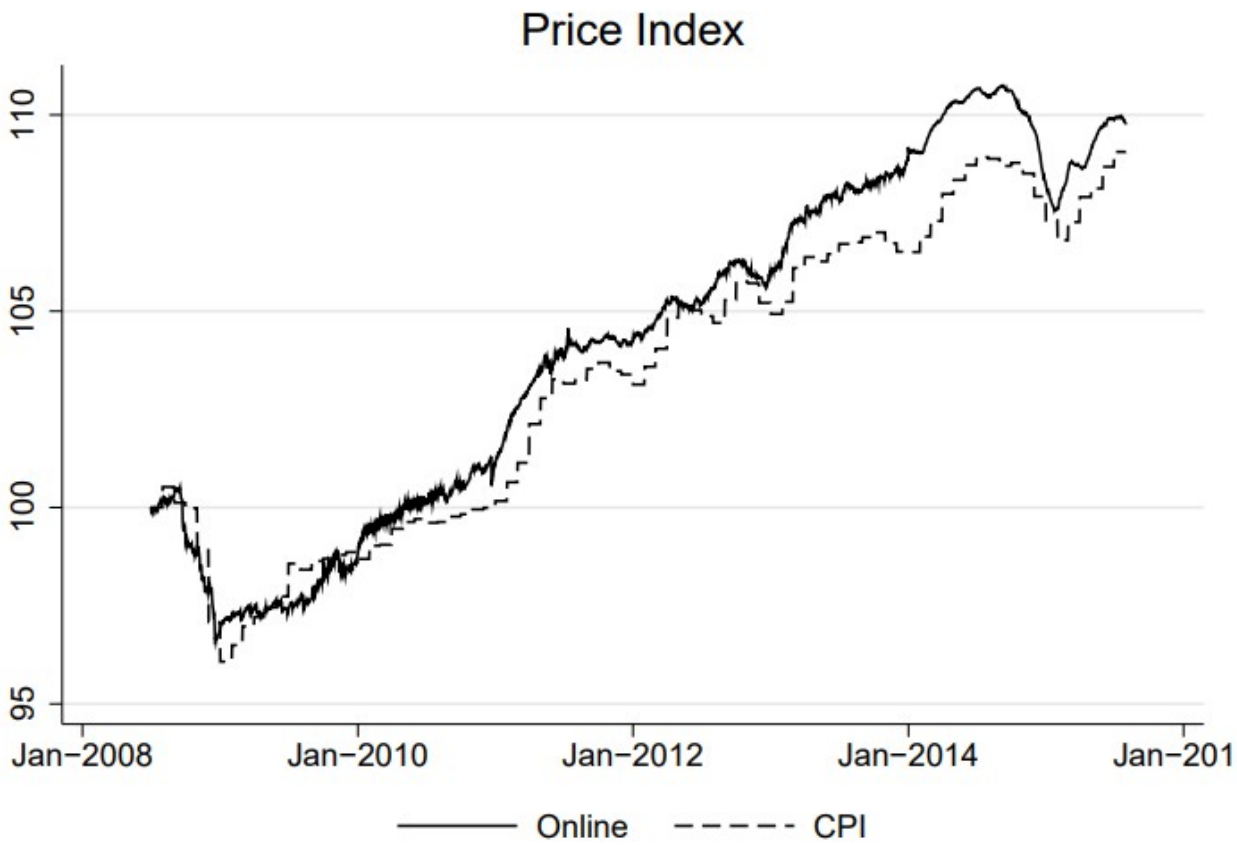
Key Points

- In many cases, financial problems are too complex for econometric models
  - Financial data exhibits complex relationships: non-linear, threshold, hierarchical
  - Most financial data is non-numeric or unstructured: categorical, text, images, recording, etc.
  - Financial datasets tend to be high-dimensional, with many variables and few observations.

(1) Structured data

- The majority of financial data is suited for machine learning models, but not econometric models.
  - Unstructured data
  - High-dimensional datasets: : the number of variables often exceeds the number of observations
  - Low signal-to-noise
  - Hierarchical relations: : economic systems often involve networks of agents, and clustering of dependencies
- Stationary transformation (differentiation), is it optimal?

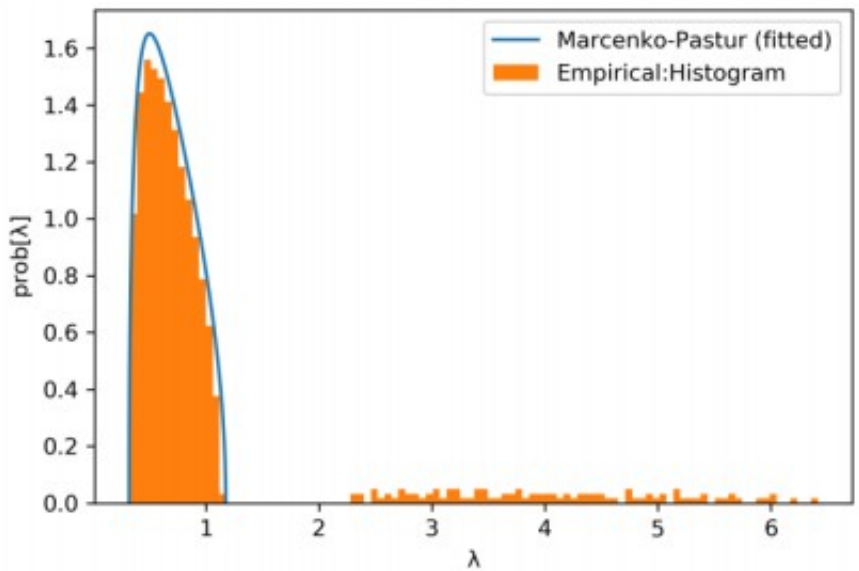
(1) Structured data



source: Cavallo and Rigobon, 2016

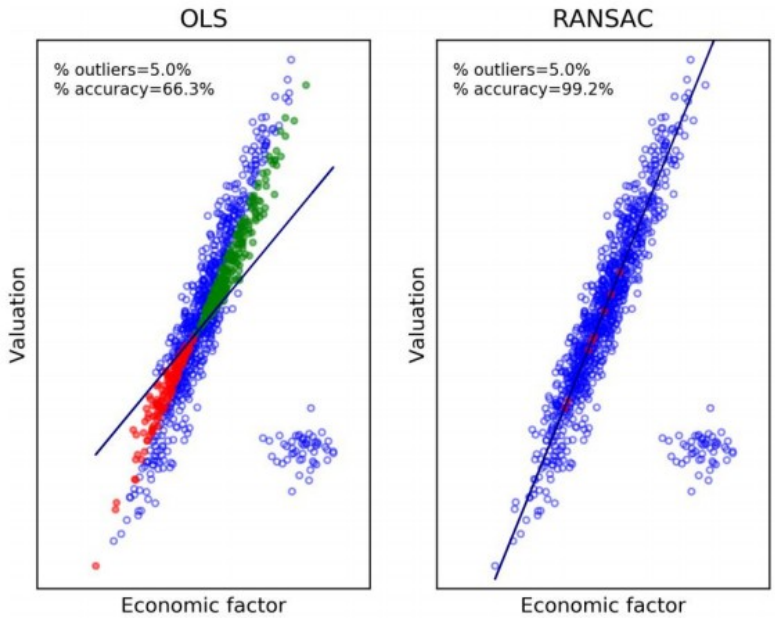
(2) Correlations, Betas

- Problem with correlations
  - While correlations between financial variables are extremly noisy, most econometric models do not include methods to denoise and detone correlation matrices.



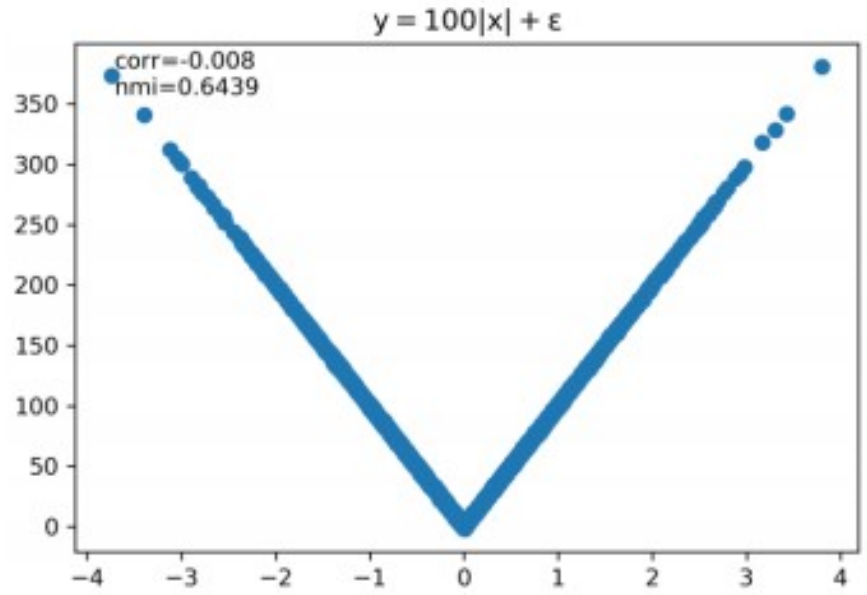
(2) Correlations, Betas

- Problem with correlations
  - Correlations are particularly sensitive to the presence of outliers.



(2) Correlations, Betas

- Problem with correlations
  - Correlation measures only linear codependence.



(3) Variance Adjudication and the Causality Fallacy

- Econometric models aim to adjudicate to $X_{i,t}$ the variance of y_t in-sample, while controlling for the variance adjudicated to $Z_{j,t}$.

$$y_t = \alpha + \sum \beta_i X_{i,t} + \sum \gamma_j Z_{j,t} + u_t$$

- On the other hand, machine learning models attempt to use $X_{i,t}$ to forecast y_t out-of-sample, while controlling for $Z_{j,t}$.
- In terms of strategy development, regression is the wrong tool for investing.

(3) Variance Adjudication and the Causality Fallacy

- Chen and Pearl (2013) found that six of the most influential econometrics textbooks make fundamental mistakes
  - confound correlation with causation
  - confound prediction with causation
  - confound causality with Granger-causality
  - 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

(3) Variance Adjudication and the Causality Fallacy

(4) Specification-Interaction Search

- Consider the typical econometric model

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

- This requires the researcher to get two items right at once:
  - (1) The predictive variables $x_{1,t}$, $x_{2,t}$ (variable search)
  - (2) The functional form (model specification)
- Given how complex financial systems are, these are unrealistic demands.

(4) Specification-Interaction Search

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

$$y_t = x_{1,t} + x_{2,t} + x_{1,t}x_{2,t} + u_t$$

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

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

- The correlation between predictions and realizations is only 0.04, even though we have provided the correct variables to the model.
- Traditional econometric models do not "learn" the structure of the data.

(5) p-values

- Most findings in financial economics rely on a $p < 0.05$ argument. However,
  - p-values require strong (unrealistic) assumptions, such as correct specification, mutually uncorrelated regressors, iid normal residuals, etc.
  - In the common case of multicollinear regressors, p-values cannot be robustly estimated.
  - p-values assess significance in-sample, not out-of-sample.
  - p-values evaluate an irrelevant probability ($P[X > x|H_0]$). What we really care about is $P[H_1|X > x]$
- This casts doubt over decades of econometric studies (the factor zoo)
- Statistically significant factors discovered through p-values include:
  - Value, Momentum, Size, Liquidity, etc.

Overcoming the Limitations of Econometrics

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