DATA 5600 4/13/2022

- Data mining / overfit : Q: What is it?

- Kuhn & Johnson "highly adaptable models that can easily overemphasize patterns (in the data) that are not reproducible."

  * Chp. 4

- Brock, Lakonishok, and LeBaron, Journal of Finance 1992

  "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns"

  } Academic bias against technical trading

  - Dow Jones returns from 1897 - 1986
  - Null models: Random walk, AR(1), GARCH-M, and Exp. GARCH
  - Treat technical traders as Neyman-Pearson hypothesis testers
  - Generate buy/sell signals, bootstrap the PnL distribution
  - Q: Do the trading returns invalidate the null hypothesis of the Null model? (EMH)
  - I think of this as a sort-of agent-based approach to econometrics
  - I think the lessons of BLL extend into other agent-based approaches in electronic markets and elsewhere where we now apply ML, AI, etc

  * See interview w/ Woodford 121-122

- BLL's strategies enumate to 1000's of "models" (tunable parameters)

- This creates potential for overfit and discovery of spurious models

- In their own words: "... the possibility that various spurious patterns were
  p. 1733
  uncovered by technical analysis cannot be dismissed."

  - They mitigate by:

    1. Reporting results from all strategies

    2. utilizing a very long dataset

    * 3. robustness across non-overlapping subperiods

- More needs to be done to address overfit, but they are asking very deep questions a/b the DGP using agent preditive analysis

- To me, this represents a "sophisticated catallactics" that would make Buchanan smile!

- Sullivan, Timmermann, and White (JF 1999) address the issue in BLL in a very innovative way

- White's Reality Check (RC) White (2000 Ecm)

Hal White proved the universal approx. theorem for neural networks so econometricians have long had an interest in ML

- STW pp. 1647-1648

  " Data-snooping occurs when a given set of data is used more than once for purposes of inference or model selection. When such data reuse occurs, there is always the possibility that any satisfactory results obtained may simply be due to chance rather than to any merit inherent in the method yielding the results. "

- Multiple Hypothesis Testing perspective
    - xkcd Green Jelly Beans
    - 20 "models" with $\alpha = 5\%$
    - P(Type I error) = .05 by definition
- W/ dozens / 100's / 1000's of models w/ tunable parameters (researcher degrees of freedom) the chance of false discovery is very high
- Data dredging / p-hacking is the intentional abuse / gaming of these facts
    - Typically w/ a rent-seeking motive
    - Call BS when you see it
    - This is now your civic responsibility as a Data Scientist!
- Far more dangerous is the possibility of self-deceit (thus the name Reality check)

- Bayes vs. Freq's
  - Notice this is all a consequence of a sampling-based approach to inference
  - Bayesians have a self-defeat test baked into their inference via Bayes' Rule
  - It requires them to be coherent
  - See Bayesian self-defeat test via Dutch Book arguments
  - See also Gellman, Hill, Yajima JREE 2012 on hierarchical models
- See Sheppard notes on WRC, SPA & Stationary bootstrap