Predictive Anecdotes

Ernst Henle
ErnstHe@UW.edu
Skype: ernst.predixion

Review: Facts and Theories

- "It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

Sir Arthur Conan Doyle as the character of Sherlock Holmes

Predictive Anecdotes (1)

- We were using predictive analytics to look for causes of dropouts in a nursing school.
- At one point we looked for professors who were associated with high dropouts or high retention.
- We found one professor whose students had a 100% retention rate. We thought that this result was significant.
- It turned out that this professor had the final class in this two-year program. In other words, drop-outs occurred prior to this professor's class. In fact her class was a pro-seminar and all the students for this class had essentially already graduated.

Predictive Anecdotes (2)

- In the same nursing school we found that if the students race was "Missing" then the students were more likely to dropout.
- At first we thought that this missing race information indicated that their was an ethnicity that pre-disposed these students to drop out.
- But, we could not find any ethnicity that had a significantly higher retention or dropout rate.
- In fact, further investigation revealed that the proportion of ethnicities was the same for the students. It did not matter whether race was categorized as "Missing" or if the students race had been entered into the database.
- Later, we determined that most of the students who filled out the forms themselves did not enter information on their ethnicity. Only those students who were personally assisted by a (diligent) registrar entered a value for race. Further analysis indicated that personal assistance by a registrar, regardless of race, correlated with high retention rate.

Predictive Anecdotes (3)

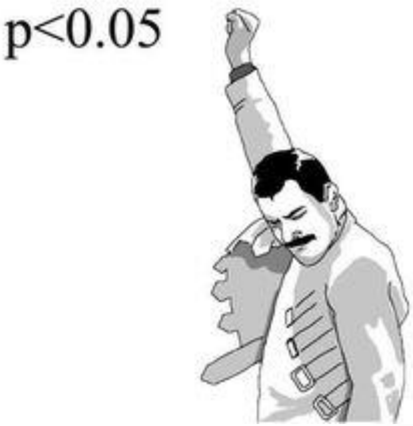
- Many Years ago, a convict in Italy, wrote to 80 stockbrokers from prison. He claimed to have insider information from a fellow convict on a large local manufacturing firm.
- To 40 stockbrokers he wrote that the stock price would rise in the next two days. To the other 40 stockbrokers he wrote that the stock price would fall.
- After two days he followed up letters to the 40 stockbrokers who received the correct prediction. To half of those he wrote that the stock price would rise and to the other half he wrote that the stock price would fall.
- The prisoner repeated this pattern three more times and then requested a fee from the stockbrokers for additional predictions.

Predictive Anecdotes(4)

"Data don't make any sense, we will have to resort to statistics."

- "If you torture the data long enough, it will confess,"
  - Ronald Coase, Professor of Economics,
  - University of Chicago
- Real Story
  - After a failed, very large, epidemiological study, the researchers wanted to justify their grant. They looked for any pattern in their data.
  - When they found a pattern they retrospectively formulated a hypothesis and then they determined if that hypothesis could be verified by their data to a 95% certainty, as is common in such studies. A 95% certainty means that there is a 5% probability that the hypothesis does not account for the patterns. Actually, it means that there is a 5% chance that the pattern is fortuitous.
  - The researchers announced many (50) "verified" hypotheses. Soon colleagues educated them: Constructing a post-facto hypothesis, is similar to re-using training data as testing data.
  - Then the researchers randomly partitioned their data into a pattern search dataset and a pattern corroboration dataset. Although, they corroborated 1 of the patterns, this search was still statistically insignificant because we expect that 5% of these patterns are fortuitous.

Predictive Anecdotes (5)

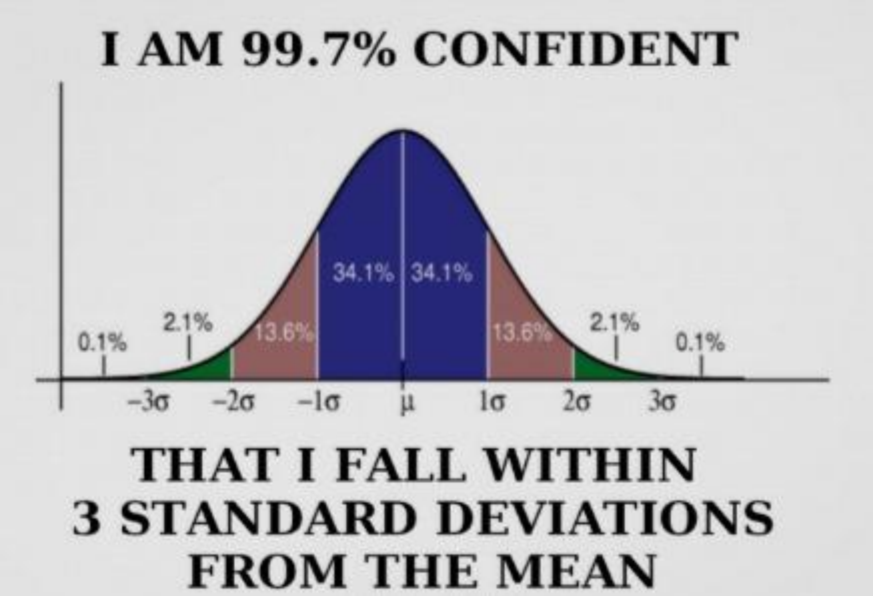


- Do Jelly Beans Cause Acne with $p < 0.05$?
- Green Jelly Beans Cause Acne $p < 0.05$!
- http://xkcd.com/882/

- The null hypothesis states that the observed variations are by chance (aka random). If you choose enough null hypotheses then there is an increasing chance that you will find a null hypothesis that is below the p-value.
- In biology we typically use a p-value of 0.05. That means that there is “only” a 5% chance that the null hypothesis is true.
- If the observed p-value <0.05, then we know that there is only a 5% chance that it is random. In other words there is a 95% chance that it is not random.
- How many hypotheses (n) should we test if we expect (> 50% chance) to find by chance 1 or more p-values (p) at less than 5%?
  - $0.5 < 1 - (1 - p)^n$; for $p = 0.05$ we find: $n \geq 14$

Predictive Anecdotes (6)

- Tautology



Prediction Anecdotes(7)

- Redskins Rule:
  - http://en.wikipedia.org/wiki/Redskins_Rule
  - http://abbottanalytics.blogspot.com/2012/11/why-predictive-modelers-should-be.html

© 2013 Ted Goff

"Our algorithms have linked funny cat videos, UFO reports and searches for tofu pizza. We're now on alert about a suspicious group of cat aliens who infiltrated our pizza industry."

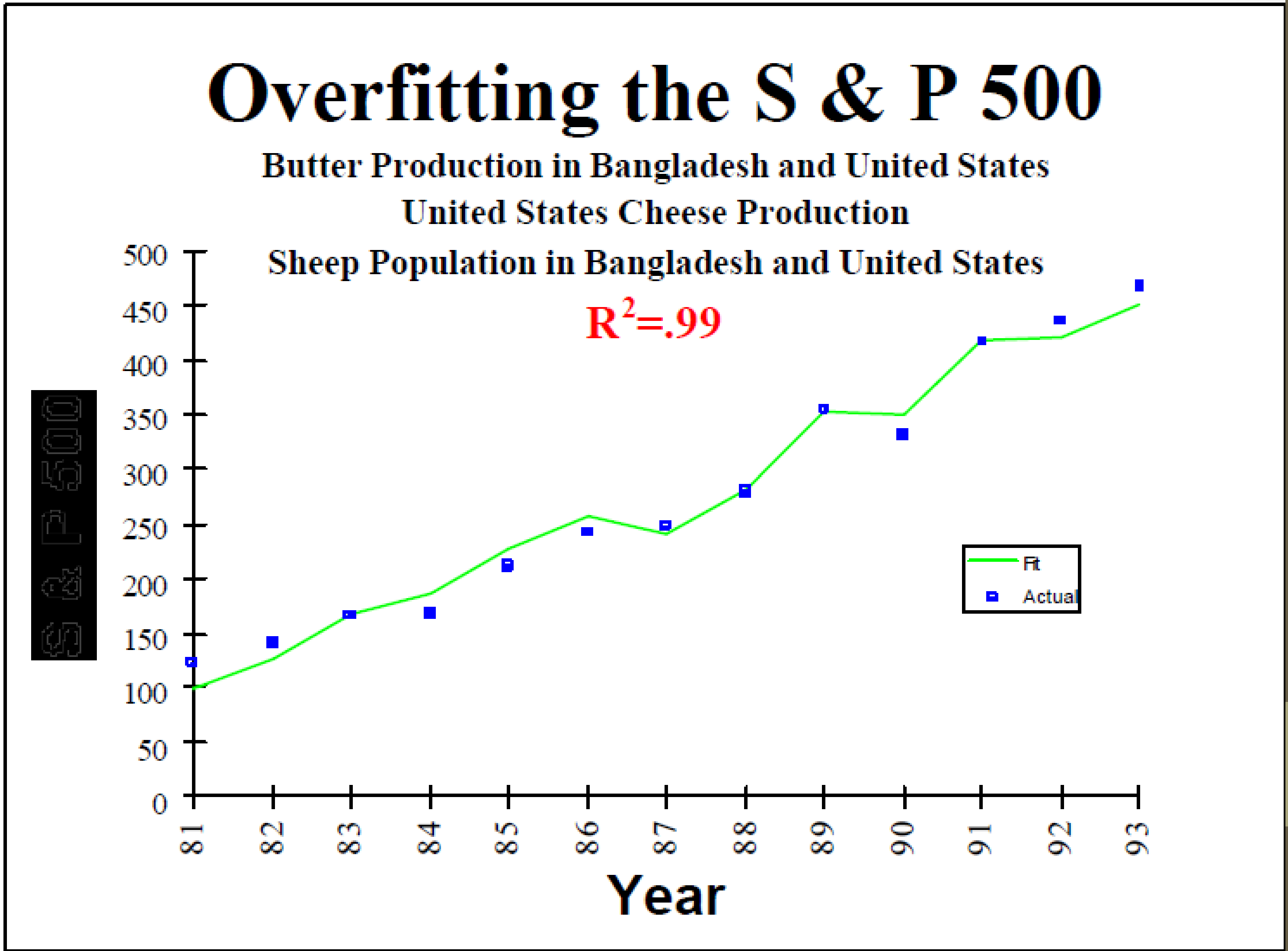
Predictive Anecdotes (8)

- http://www.finanzaonline.com/forum/attachments/econometria-e-modelli-di-trading-operativo/903701d1213616349-variazione-della-vix-e-rendimento-dello-s-p500-dataminejune_2000.pdf (Also available on Catalyst: dataminejune_2000.pdf)

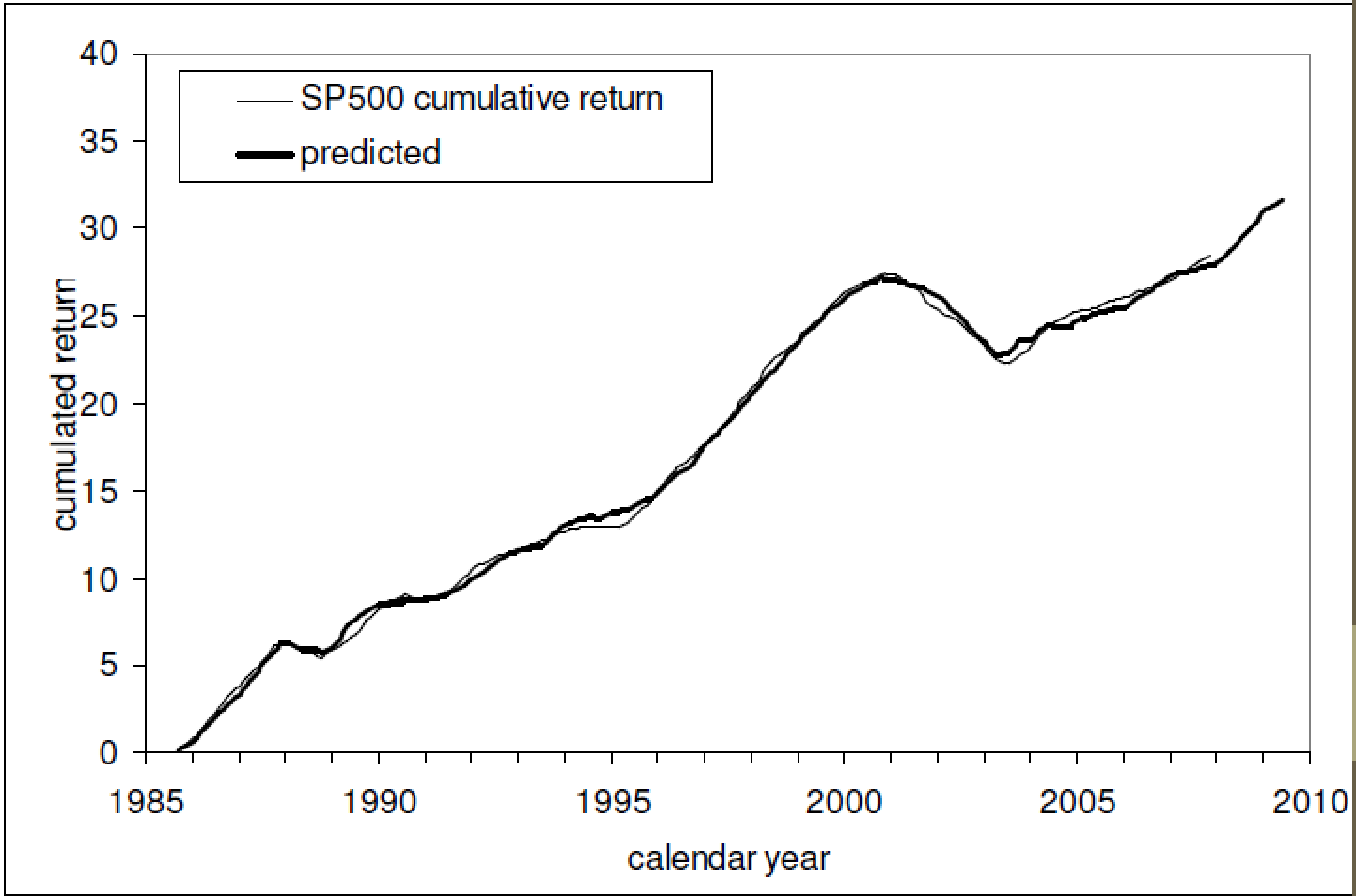
- S&P500 ~ **Butter Production in Bangladesh + Butter Production in United States + United States Cheese Production + Sheep Population in Bangladesh + Sheep Population in United States**
- S&P500 ~ **Butter Production in Bangladesh**
-
- See Also: http://www.forbes.com/sites/davidleinweber/2012/07/24/stupid-data-miner-tricks-quants-fooling-themselves-the-economic-indicator-in-your-pants/

- **Exact prediction of S&P 500 returns** by Ivan O. Kitov, Oleg I. Kitov (See: SSRN-id1045281.pdf on Catalyst)

Predictive Anecdotes (9)



Predictive Anecdotes (10)

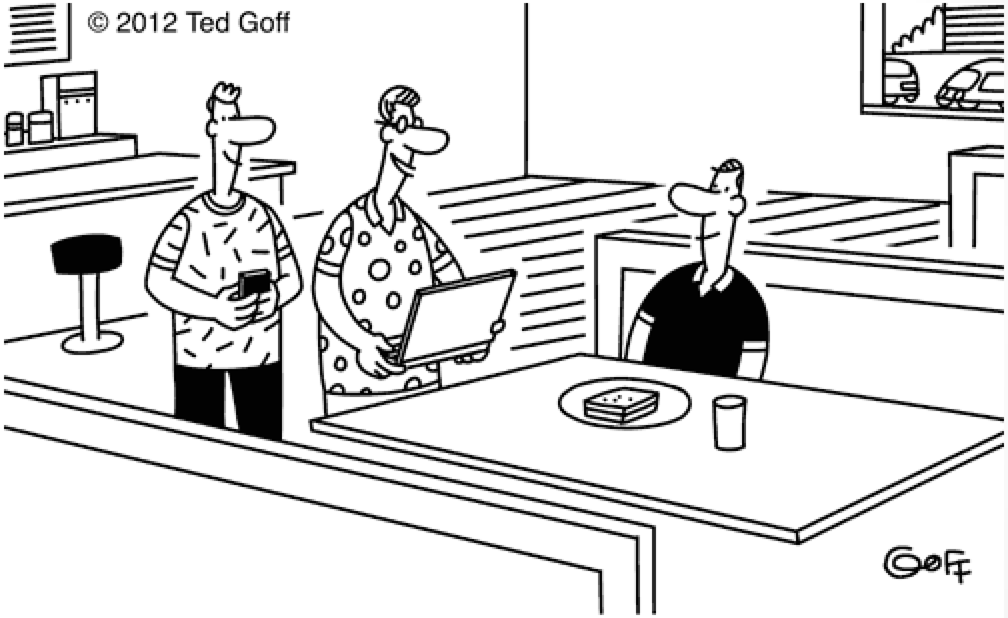


Predictive Anecdotes (11)

- Orange Cars Predicted to be better!
- "An orange used car is least likely to be a lemon" (http://seattletimes.com/html/businesstechnology/2017983961_apusscienceassport.html): *Of the 72,983 used cars, 8,976 were bad buys (12.3%). Yet, of the 415 orange cars in the dataset, only 34 were bad (8.2%)*
-
- *Debunked in this paper:* "Are Orange Cars Really not Lemons?" by Ben Bullard and John Elder. The paper is on Catalyst: orange cars.pdf.

Predictive Anecdotes (12)

- Miscellaneous
  - Confusing Correlation with Causation:
    - http://en.wikipedia.org/wiki/Spurious_relationship
  - Proxy Columns and Audience Gullibility:
    - Scam artists use proxy attributes in their “predictions”
    - A true story from about 20 years ago. A fortune teller went on a radio talk show on KGO in the Bay Area. He demonstrated how he could mimic psychic abilities and get people to divulge information without their knowledge. After the show, this confessed scam artist was flooded with requests for psychic readings. The audience preferred to believe in his psychic powers and not his confessions.



© 2012 Ted Goff

“Twitter and Facebook can’t predict the election, but they did predict what you’re going to have for lunch: a tuna salad sandwich. You’re having the wrong sandwich.”

Predictive Anecdotes