

August 8, 2009

Data Mining Isn't a Good Bet For Stock-Market Predictions

By JASON ZWEIG

Slicing and dicing data to predict the future can get dicey.



The Super Bowl market indicator holds that stocks will do well after a team from the old National Football League wins the Super Bowl. The Pittsburgh Steelers, an original NFL team, won this year, and the market is up as well. Unfortunately, the losing Arizona Cardinals also are an old NFL team.

The "Sell in May and go away" rule advises investors to get out of the market after April and get back in after October. With the market up 17% since April 30, that rule isn't looking so good at this point.

Meanwhile, dozens -- probably hundreds -- of Web sites hawk "proprietary trading tools" and analytical "models" based on factors with cryptic names like McMillan oscillators or floors and ceilings.

There is no end to such rules. But there isn't much sense to most of them either. An entertaining new book, "Nerds on Wall Street," by the veteran quantitative money manager David Leinweber, dissects the shoddy thinking that underlies most of these techniques.

The stock market generates such vast quantities of information that, if you plow through enough of it for long enough, you can always find some relationship that appears to generate spectacular returns -- by coincidence alone. This sham is known as "data mining."

Every year, billions of dollars pour into data-mined investing strategies. No one knows if these

techniques will work in the real world. Their results are hypothetical -- based on "back-testing," or a simulation of what would have happened if the manager had actually used these techniques in the past, typically without incurring any fees, trading costs or taxes.

Those assumptions are completely unrealistic, of course. But data-mined numbers can be so irresistible that, as Mr. Leinweber puts it, "they are one of the leading causes of the evaporation of money, especially in quantitative strategies."

Mr. Leinweber got so frustrated by "irresponsible" data mining that he decided to satirize it. After casting about to find a statistic so absurd that no sensible person could possibly believe it could forecast U.S. stock prices, Mr. Leinweber settled on annual butter production in Bangladesh. Over an 13-year period, he found, this statistic "explained" 75% of the variation in the annual returns of the Standard & Poor's 500-stock index.

By tossing in U.S. cheese production and the total population of sheep in both Bangladesh and the U.S., Mr. Leinweber was able to "predict" past U.S. stock returns with 99% accuracy.



But the entire exercise, he says, is a total crock. There is no conceivable reason

why U.S. stock returns would be determined by Bangladeshi livestock returns.

Mr. Leinweber's exercise isn't much more absurd than some actual examples of data mining. One recent scholarly paper purported to show that you can predict stock returns by tracking the number of nine-year-olds in the U.S. Another academic study asserts that stocks are more likely to go up on days when smog goes down.

That points to the first rule for keeping yourself from falling into a data mine: The results have to make sense. Correlation isn't causation, so there needs to be a logical reason why a particular factor should predict market returns. No matter

how appealing the numbers may look, if the cause isn't plausible, the returns probably won't last.

The second rule is to break the data into pieces. Divide the measurement period into thirds, for example, to see whether the strategy did well only part of the time. Ask to see the results only for stocks whose names begin with A through J, or R through Z, to see whether the claims hold up when you hold back some of the data.

Next, ask what the results would look like once trading costs, management fees and applicable taxes are subtracted.

Finally, wait. Hypothetical results usually crumple after they collide with the real-world costs of investing. "If a strategy's worthwhile," Mr. Leinweber says, "then it'll still be worthwhile in six months or a year."

Mr. Leinweber still gets inquiries from money managers who want him to share his data on Bangladeshi butter production so they can get the latest numbers and build a trading strategy around them. "A distressing number of people don't get that it was a joke," Mr. Leinweber sighs.

Don't let the joke be on you.

Write to Jason Zweig at intelligentinvestor@wsj.com

Copyright 2009 Dow Jones & Company, Inc. All Rights Reserved

Questions:

1. What is data mining?
2. What is the difference between correlation and causation?
3. What percentage of variation in annual S&P500 returns was David Leinweber able to explain over a 13-year period? What variables did he use in explaining the returns?
4. What was Mr. Leinweber's point?
5. Assume you (or someone else) find a relationship between stock returns and some set of variables. How can you differentiate between relationships that occur by random chance and relationships that really exist?
6. Assume markets are efficient. Why would following various trading rules lead to lower returns than holding an index fund?