Nerds ON Wall street

Math, Machines, and Wired Markets

David J. Leinweber

WILEY

John Wiley & Sons, Inc.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Chapter opener images courtesy of www.wordle.net

Library of Congress Cataloging-in-Publication Data:

Leinweber, David, 1952-
Nerds on Wall Street : math, machines, and wired markets / David J. Leinweber.
p. cm.
Includes index.
ISBN 978-0-471-36946-2 (cloth)
1. Investments—Computer network resources. 2. Wall Street (New York, N.Y.) I. Title.
HG4515.95.L43 2009
332.64'273—dc22

2009008848

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Chapter 6

Stupid Data Miner Tricks

To Err Is Human. To Really Screw Up,
You Need a Computer.

—POPULAR CAMPUS T-SHIRT, CA. 1980

This chapter started out over 10 years ago as a set of joke slides showing silly, spurious correlations. Originally, my quant equity research group planned on deliberately abusing the genetic algorithm (see Chapter 8 on evolutionary computation) to find the wackiest relationships, but as it turned out, we didn't need to get that fancy. Just looking at enough data using plain-vanilla regression would more than suffice.

We uncovered utterly meaningless but statistically appealing relationships between the stock market and third world dairy products and livestock populations which have been cited often—in *BusinessWeek*, the *Wall Street Journal*, the book *A Mathematician Plays the Stock Market*,[1] and many others. Students from Bill Sharpe's classes at Stanford seem to be familiar with them. This was expanded, to have some actual content about data mining, and reissued as an academic working paper in 2001. Occasional requests for this arrive from distant corners of the world. An updated version appeared in the *Journal of Investing* in 2007.*

*This article originally appeared in the Spring 2007 issue of the *Journal of Investing* ("Stupid Data Miner Tricks: Overfitting the S&P 500"). It is reprinted with permission. To view the original article, please go to iijoi.com.

Without taking too much of a hatchet to the original, the advice here is still valuable—perhaps more so now that there is so much more data to mine. Monthly data arrives as one data point, once a month. It's hard to avoid data mining sins if you look twice. Ticks, quotes, and executions arrive in millions per minute, and many of the practices that fail the statistical sniff tests for low-frequency data can now be used responsibly. New frontiers in data mining have been opened up by the availability of vast amounts of textual information. Whatever raw material you choose, fooling yourself remains an occupational hazard in quantitative investing. The market has only one past, and constantly revisiting it until you find that magic formula for untold wealth will eventually produce something that looks great, in the past. A fine longer exposition of these ideas is found in Nassim Taleb's book, *Fooled by Randomness: The Hidden Role of Chance in Markets and Life* (W.W. Norton, 2001).

"Your Mama Is a Data Miner"

It wasn't too long ago that calling someone a data miner was a very bad thing. You could start a fistfight at a convention of statisticians with this kind of talk. It meant that you were finding the analytical equivalent of the bunnies in the clouds, poring over data until you found something. Everyone knew that if you did enough poring, you were bound to find that bunny sooner or later, but it was no more real than the one that blows over the horizon.

Data mining is a small industry, with entire companies and academic conferences devoted to it. The phrase no longer elicits as many invitations to step into the parking lot as it used to. What's going on? These new data mining people are not fools. Sometimes data mining makes sense, and sometimes it doesn't.

The new data miners pore over large, diffuse sets of raw data trying to discern patterns that would otherwise go undetected. This can be a good thing. Suppose a big copier company has thousands of service locations all over the world. It wouldn't be unusual for any one of them to see a particular broken component from any particular copier. These gadgets do fail. But if all of a sudden the same type of part starts showing up in the repair shops at 10 times its usual rate, that would be an indication of a manufacturing problem that could be corrected at the factory. This is a good (and real) example of how data mining

can work well, when it is applied to extracting a simple pattern from a large data set. That's the positive side of data mining. But there's an evil twin.

The dark side of data mining is to pick and choose from a large set of data to try to explain a small one. Evil data miners often specialized in "explaining" financial data, especially the U.S. stock market. Here's a nice example: We often hear that the results of the Super Bowl in January will predict whether the stock market will go up or down for that year. If the National Football Conference (NFC) wins, the market goes up; otherwise, it takes a dive. What has happened over the past 30 years? Most of the time, the NFC has won the Super Bowl and the market has gone up. Does it mean anything? Nope. We see similar claims for hemlines, and even the phases of the moon.[2]

When data mining techniques are used to scour a vast selection of data to explain a small piece of financial market history, the results are often ridiculous. These ridiculous results fall into two categories: those that are taken seriously, and those that are regarded as totally bogus. Human nature being what it is, people often differ on what falls into which category.

The example in this paper is intended as a blatant instance of totally bogus application of data mining in finance. My quant equity research group first did this several years ago to make the point about the need to be aware of the risks of data mining in quantitative investing. In total disregard of common sense, we showed the strong statistical association between the annual changes in the S&P 500 index and butter production in Bangladesh, along with other farm products. Reporters picked up on it, and it has found its way into the curriculum at the Stanford Business School and elsewhere. We never published it, since it was supposed to be a joke. With all the requests for the nonexistent publication, and the graying out of many generations of copies of copies of the charts, it seemed to be time to write it up for real. So here it is. Mark Twain (or Disraeli, or both) spoke of "lies, damn lies, and statistics." In this paper, we offer all three.

Strip Mining the S&P 500

Regression is the main statistical technique used to quantify the relationship between two or more variables.[3] It was invented by Adrien-Marie

Legendre in 1805.[4] A regression analysis would show a positive relationship between height and weight, for example. If we threw in waistline along with height, we'd get an even better regression to predict weight.

The measure of the accuracy of a regression is called *R*-squared. A perfect relationship, with no error, would have an *R*-squared of 1.00 or 100 percent. Strong relationships, like height and weight, would have an *R*-squared of around 70 percent. A meaningless relationship, like zip code and weight, would have an *R*-squared of zero.

With this background, we can get down to some serious data mining. First, we need some data to mine. We'll use the annual closing price of the S&P 500 index for the 10 years from 1983 to 1993, shown in Figure 6.1.

This is the raw data, the S&P 500 for the period, what we are going to predict in terms of the idea of "maximizing predictability" discussed at the end of the previous chapter. Now, we want to go into the data mine and find some data to use to predict the stock index. If we included other U.S. stock market indexes such as the Dow Jones Industrial Average or the Russell 1000, we would see very good fits, with *R*-squared numbers close to 1.00. That would be an uninspired choice, though—and useless at making the point about the hazards of data mining.

Now we need some more data to mine in which to fit the S&P data; that is, make a correlation. Let's go find some on a CD-ROM published by the United Nations. There are all kinds of data series from 140 member countries. If we were trying to do this S&P 500 fit for real, we might look at things like changes in interest rates, economic growth,

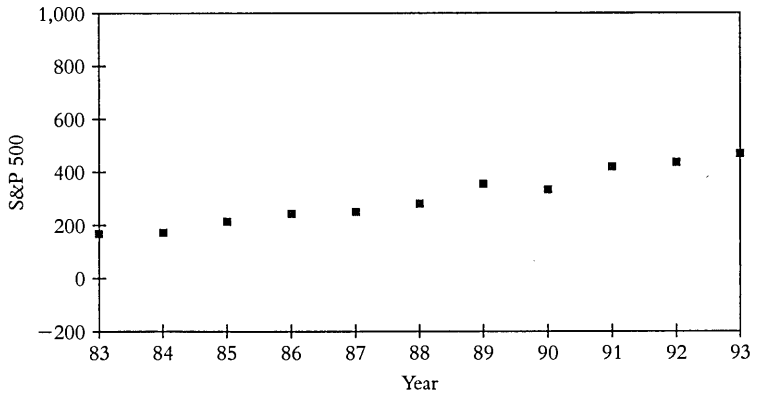

Figure 6.1 The S&P 500, 1983–1993: The Unmined Raw Data

unemployment, and the like, but we'll stay away from those. Let's use something even better: butter production in Bangladesh. Yes, there it is: a simple, single dairy product that explains 75 percent of the variation in the S&P 500 over 10 years. R^2 is 0.75; not bad at all. (See Figure 6.2.)

Why stop here? Maybe we can do better. Let's go global on this and expand our selection of dairy products. We'll put in cheese and include U.S. production as well. This works remarkably well. We're up to 95 percent accuracy here. (See Figure 6.3.) How much better can we do?

How about 99 percent with our third variable: sheep population? This is an awesome fit. (See Figure 6.4.) It seems too good to be true, and it is. That is the point. It is utterly useless for anything outside the

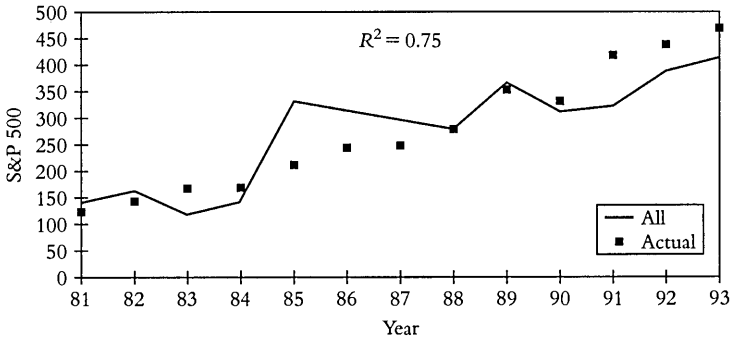

Figure 6.2 Overfitting the S&P 500: butter production in Bangladesh—a single variable that "explains" 75 percent of the S&P's returns.

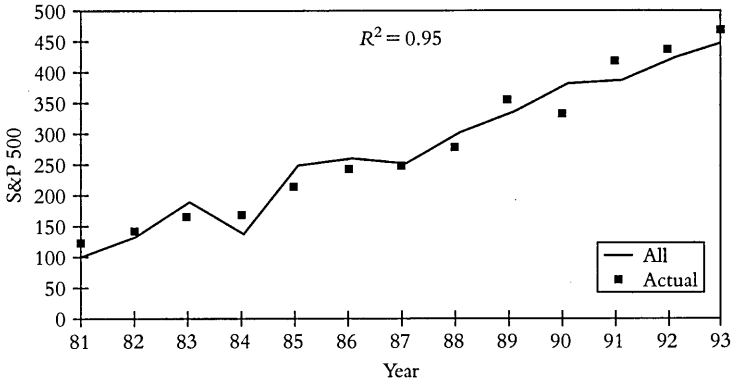

Figure 6.3 Overfitting the S&P 500: butter in Bangladesh and United States, plus U.S. cheese production—two more dairy variables that take us to 95 percent.

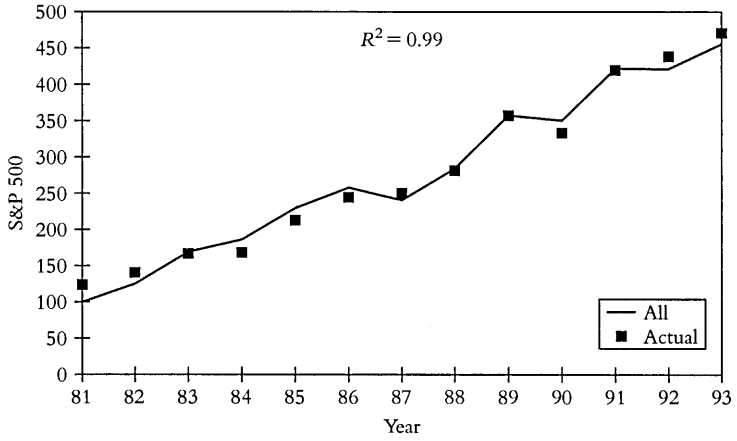

Figure 6.4 Overfitting the S&P 500: butter in Bangladesh and United States, plus U.S. cheese production, as well as sheep population in Bangladesh and United States. Now we're at 99 percent. You can do this as long as you can find data not perfectly correlated with butter, cheese, sheep, and so on. There is no shortage of that.

fitted period, a total crock before 1983 or after 1993. This is just a chance association that would inevitably show up if you look at enough data series. The butter fit was the result of a lucky fishing expedition. The rest comes from throwing in a few other series that were uncorrelated to the first one. Pretty much anything would have worked, but we liked sheep. They are more photogenic than dairy products, and make for a great slide when this stuff is shown to an audience at one of those open-bar financial conference dinners.

If someone showed up in your office with a model relating stock prices to interest rates, gross domestic product (GDP), trade, housing starts, and the like, it might have statistics that looked as good as this nonsense, and it might make as much sense (i.e., none), even though it sounded much more plausible.

Enough Regression Tricks

To hammer a little harder on this point about the dangers of data mining, look at another equally bogus example. Who wants to go count pregnant

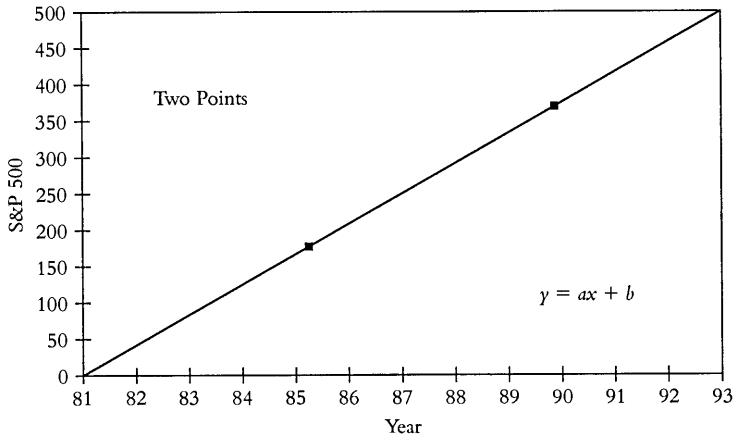

Figure 6.5 First-degree polynomial fit: just a plain old line.

sheep in Bangladesh to figure out next year's sheep population? We'll get away from ordinary linear regressions and show how we can fit a *perfect* model, with R^2 = 100 percent, using only one variable: the year's digits.

This has to be about the most accessible data on the planet. There is no need to go counting sheep. Instead of regression, we'll use a different prediction method to make this work, a polynomial fit. Everyone with a recollection of junior high school math knows that there is a line (a first-degree polynomial) through any two points, as shown in Figure 6.5.

Put in a third point and you can fit a parabola, or second-degree polynomial, through all three points, as shown in Figure 6.6.

We have 10 points in the S&P 500 annual series from 1983 to 1992, so we fit a ninth-degree polynomial. However, as Mr. Wizard says, "Don't try this at home," unless you have some sort of infinite precision math tool like Mathematica or Maple. The ordinary floating point arithmetic in a spreadsheet or regular programming language isn't accurate enough for this to work. That said, our ninth-degree polynomial hits every annual close *exactly*. We have a 100 percent in-sample accuracy with only one variable, as shown in Figure 6.7.

$$.25 * 10^{16} - .26 * 10^{13}y + .12 * 10^{10}y^2 - 320000.y^3$$
$$+ 56.y^4 - .0064y^5 + .49 * 10^{-6}y^6 - .24 * 10^{-10}y^7$$
$$+ .69 * 10^{-15}y^8 - .88 * 10^{-20}y^9$$

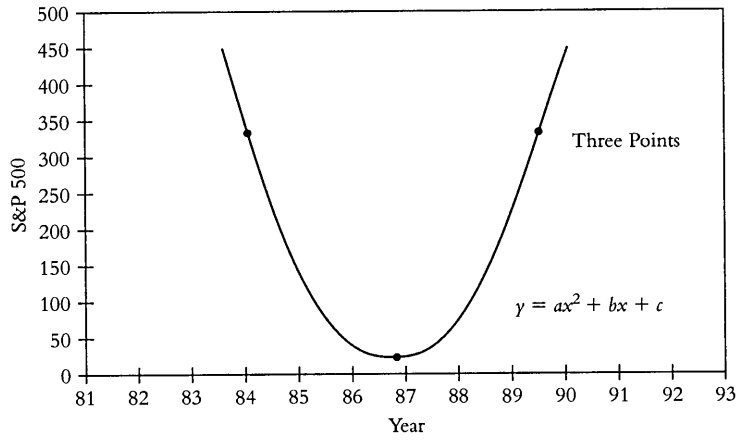

Figure 6.6 Second-degree polynomial fit: a plain old parabola.

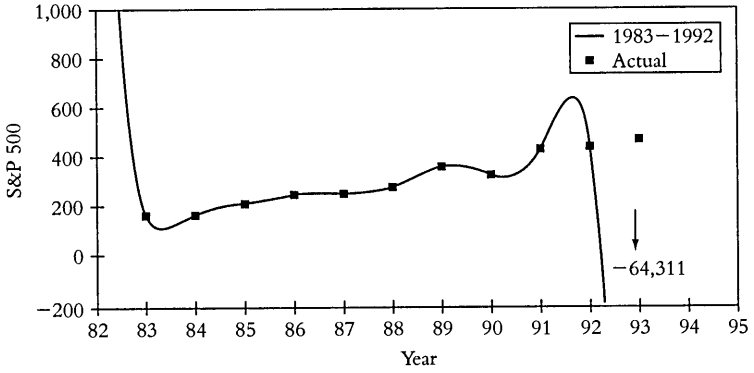

Figure 6.7 Polynomial fit to the S&P 500: big mistake or bad Idea? That minus 64,311 looks like trouble.

Notice that the fitted curve in the chart is suddenly heading south very rapidly. What closing value for the S&P did this method predict for the end of 1993? Minus 64,311. Fortunately for the global economy, it actually turned out to be positive, +445. We seem to have a problem with our model's out-of-sample performance, as shown in Figure 6.8.

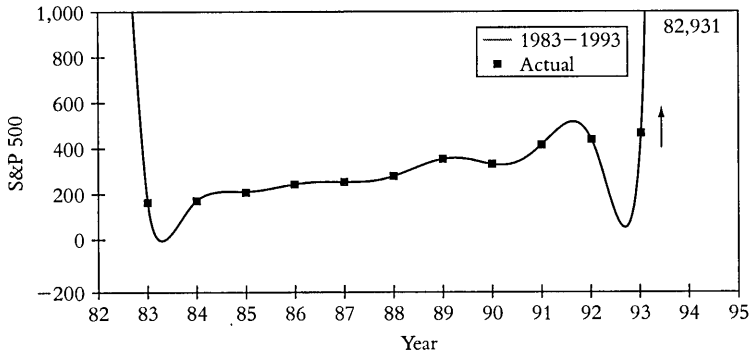

Figure 6.8 Polynomial fit to the S&P 500: big mistake or bad idea? Add a new data point. An S&P 500 at 82,931 is nice think about, but ridiculous.

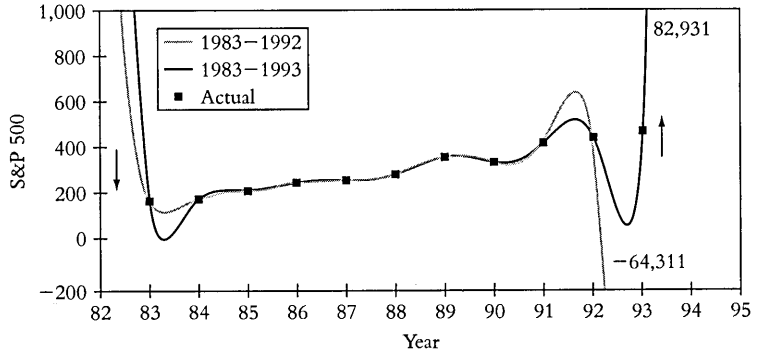

Figure 6.9 Polynomial fit to the S&P 500: big mistake or bad idea? It is both. If only all bad investment ideas were as easy to spot.

Don't panic! The year 1993 ends, we get another data point, and we restore our 100 percent in-sample accuracy, this time with a tenth-degree polynomial using the 11 data points in the sample. What did this new fit predict for the S&P close in 1994? Plus 82,931!

$$.77 * 10^{17} - .88 * 10^{14}y + .45 * 10^{11}y^2 - .14 * 10^{8}y^3 + 2700.y^4 - .37y^5 + .000035y^6 - .23 * 10^{-8}y^7 + .99 * 10^{-13}y^8 - .25 * 10^{-17}y^9 + .28 * 10^{-22}y^{10}$$

So in Figure 6.9, we have two models—each 100 percent accurate in-sample, and each 100 percent nonsense out-of-sample.

Is There Any Hope for Data Miners?

The central problem in mining financial data like this is that the market has only one past. This will not go away. Some forecasters just ignore this fact, dive in, and hope for the best. This makes about as much sense as the "butter in Bangladesh" story. It would be a healthy idea to take measures to mitigate the risk of data mining. Here are a few suggestions:

- *Avoid the other pitfalls of investment simulations.* These include survivor bias, look-ahead bias, use of revised data not available at the time of the forecasts, ignoring transaction costs, and liquidity constraints. There are many ways to fool yourself, even before you set foot in the data mine.[5]
- *Use holdback samples, temporal and cross-sectional.* Reserve some of the data for out-of-sample testing. This can be hard to do when the history is short, or the frequency is low as is the case for monthly data. Be cautious about going to the holdback set, since with each new visit, you are mining that as well. This approach to temporal holdback samples is easier with higher-frequency data, such as daily information or ticks. In these cases, a three-level holdback protocol using in-sample, out-of-sample, and in-the-vault out-of-sample can be (and is) used.

  When there are multiple securities to analyze, you can also hold back a cross-sectional sample. As an example, if you were developing a model to forecast individual stock returns, keeping back all the stocks with symbols in the second half of the alphabet, or even Committee on Uniform Securities Identification Procedures (CUSIP) numbers, would retain half of the data for out-of-sample testing. Temporal and cross-sectional holdbacks can be combined in data-rich situations.
- *Apply statistical measures of data mining and snooping.* Econometricians have performed explicit analyses of the problem of data mining in forecasting models. These techniques are based on the idea of testing the hypothesis that a new model has predictive superiority over a previous benchmark model. The new model is clearly data-mined to some extent, given that the benchmark model was developed beforehand.[6]
- *Use truly bogus test models.* You can calibrate the model-building process using a model based on random data. There is a ferocious amount of technology that can be brought to bear on forecasting problems. One neural net product advertised in *Technical Analysis of Stocks & Commodities* magazine claims to be able to "forecast any market, using any data."[7] This is no doubt true, subject to the usual caveats. Throw in enough genetic algorithms, wavelets, and the like, and you are certain to come up with a model. But is it any good? A useful indicator in answering this question is to take the same model-building process and use it to build a test model for the same forecast target, but using a completely random set of data.[8] This test model has to be truly bogus. If your actual model has performance statistics similar to the bogus test version, you know it's time to visit the data miner's rehabilitation clinic.

Summary (and Sermonette)

These dairy product and calendar examples are obviously contrived. They are not far removed from many ill-conceived quantitative investment and trading ideas. It is just as easy to fool yourself with ideas that are plausible-sounding and no more valid.

Just because something appears plausible, that doesn't mean that it is. The wide availability of machine-readable data, and the tools to analyze it, easily means that there are a lot more regressions going on than Legendre could ever have imagined back in 1805. If you look at 100 regressions that are significant at a level of 95 percent, five of them are there just by chance. Look at 100,000 models at 95 percent significance, and 5,000 are false positives. Data mining, good or bad, is next to impossible to do without a computer.

When doing this kind of analysis, it is important to be very careful of what you ask for, because you will get it. Holding back part of your data is the first line of a defense against data mining. Leaving some of the data out of the sample used to build the model is a good idea as is holding back some data to use in testing the model. This holdback sample can be a period of time or a cross section of data. The cross-sectional holdback works where there is enough data to do this, as in the analysis of individual stocks. You can use stocks with symbols starting with A through L for model building and save M through Z for verification purposes.

It is possible to mine these holdback samples as well. Every time you visit the out-of-sample period for verification purposes, you do a little more data mining. Testing the process to see if you can produce models of similar quality using purely random data is often a sobering experience. An unlimited amount of computational resources is like dynamite: If used properly, it can move mountains. Used improperly, it can blow up your garage or your portfolio.

If someone knocks on your door with a new strategy that beat the market by 5 percent a year for the past 50 years in the back tests, wait. Someone else will put some money into it. If it tanks by 10 percent in the year after the real cash showed up, you just met a data miner. There are plenty of them, and some have really nice suits.

Counting the Kiddies

Despite warnings like this one, the temptation to data mine remains strong. Fidelity's Steve Snider sent me a 2007 paper called "Exact Prediction of S&P 500 Returns."[9] The authors demonstrate a near-perfect method of predicting index returns based on the U.S. population of nine-year-olds—not eight, not ten, just nine.

> A linear link between *S&P 500* return and the change rate of the number of nine-year-olds in the USA has been found. The return is represented by a sum of monthly returns during previous twelve months. The change rate of the specific age population is represented by moving averages. The period between January 1990 and December 2003 is described by monthly population intercensal estimates as provided by the US Census Bureau. Four years before 1990 are described using the estimates of the number of 17-year-olds shifted 8 years back. The *prediction* of *S&P 500* returns for the months after 2003, including those beyond 2007, are obtained using the number of 3-year-olds between 1990 and 2003 shifted by 6 years ahead and quarterly estimates of real GDP per capita.

The authors do some pretty fancy econometrics, but find that the near-exact fit breaks down after 2003. Their conclusion? "Therefore, it is reasonable to assume that the 9-year-old population was not well estimated by the U.S. Census Bureau after 2003." Children are our future and all that, but, please, think before testing cointegration.

A computer lets you make more mistakes faster than any invention in human history—with the possible exceptions of handguns and tequila. The easy access to data and tools to mine it gives new meaning to the admonition about "lies, damn lies, and statistics." The old adage *caveat emptor*, buyer beware, is still excellent advice. If it seems too good to be true, it is.

Notes

1. John Allen Paulos, *A Mathematician Plays the Stock Market* (New York: Basic Books, 2003).
2. It gets much wackier than this. A man named Norman Bloom, no doubt a champion of all data miners, went beyond trying to predict the stock market. Instead, he used the stock market, along with baseball scores, particularly those involving the New York Yankees, to "read the mind of God." I offer a small sample of Bloom, in the original punctuation and spelling, here: "The instrument God has shaped to brig proof he has the power to shape the physical actions of mankind—is organized athletics, and particularly baseball. the second instrument shaped by the one God, as the means to bring proof he is the one God concerned with the mental and business aspects of mankind and his civilization is the stock market—and particularly the greatest and most famous of all these—i.e., the New York Stock Exchange." Mr. Bloom's work was brought to my attention by Ron Kahn of Barclays Global Investing. Bloom himself did not publish in any of the usual channels, but seekers of secondary truth can consult "God and Norman Bloom" by Carl Sagan, in *American Scholar* (Autumn 1977), p. 462.
3. There are many good texts covering the subject. For a less technical explanation, see *The Cartoon Guide to Statistics* by Larry Gonnick and Woolcott Smith (New York: HarperCollins, 1993).
4. Stephen M. Stigler, *The History of Statistics: The Measure of Uncertainty before 1900* (Cambridge, MA: Belknap Press, 1986). This invention is also often attributed to Francis Galton, never to Disraeli or Mark Twain.
5. See John Freeman, "Behind the Smoke and Mirrors: Gauging the Integrity of Investment Simulations," *Financial Analysts Journal* 48, no. 6 (November–December 1992): 26–31.
6. This question is addressed in "A Reality Check for Data Snooping" by Hal White, UCSD Econometrics working paper, University of California at San Diego, May 1997.
7. This was the actual ad copy for a neural net system, InvestN-32, from Race Com., which was promoted heavily in *Technical Analysis of Stocks & Commodities* magazine, often a hotbed of data mining.
8. There are several alternatives in forming random data to be used for forecasting. A shuffling in time of the real data preserves the distribution of the original data, but loses many time series properties. A series of good old machine-generated random numbers,

matched to the mean, standard deviation, and higher moments of the original data will do the same thing. A more elaborate random data generator is needed if you want to preserve time series properties such as serial correlation and mean reversion.

9. Ivan Kitov and Oleg Kitov, "Exact Prediction of S&P 500 Returns," Russian Academy of Sciences—Institute for the Geospheres Dynamics and University of Warwick, December 2007, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1045281.