STUPID DATA MINER TRICKS: OVERFITTING THE S&P 500

David J. Leinweber, Ph.D.
Caltech

djl@caltech.edu
dleinweber@post.harvard.edu

(initially written 1995)

"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.

Now, data mining is a small industry, with entire companies devoted to it.[1] There are academic conferences devoted solely to data mining. 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 ten 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 US stock market. Here's a nice example: we often hear that the results of the Superbowl in January will predict whether the stock market will go up or down for that year. If the NFL wins, the market goes up, otherwise, it takes a dive. What's happened over the last thirty years? Well, most of the time, the NFL wins the Superbowl, 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 which category is which.

[1] See www.trivida.com, or www.dataminer.demon.co.uk for examples.

[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, 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 – ie THE NEW YORK STOCK EXCHANGE.

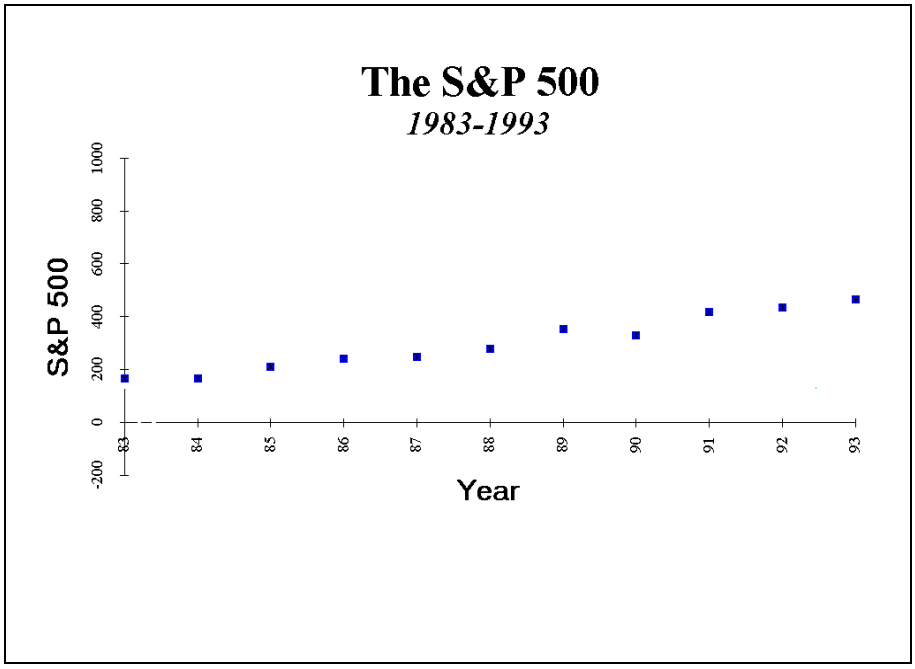
Mr. Bloom's work was brought to my attention by Ron Kahn of BARRA. It is truly a wonder. Bloom himself did not publish in any of the usual channels, but seekers of secondary truth can consult "God and Norman Bloom," by Carl Sagan, in American Scholar, Autumn 1977, p. 462.

The example in this paper is intended as a blatant example of totally bogus application of data mining in finance. We 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 stock index and butter production in Bangladesh, and 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 non-existent 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 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 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%. Strong relationships, like height and weight, would have an R-squared of around 70%. 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 ten years from 1983 to 1993, shown in the chart below.



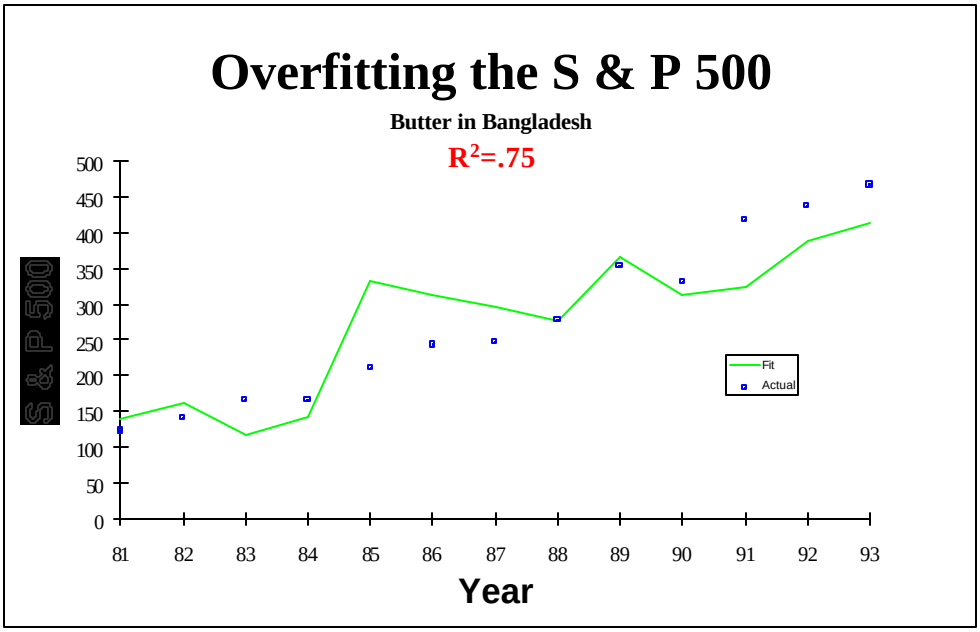
[3] There are many good texts covering the subject. For a less technical explanation, see The Cartoon Guide to Statistics, by Larry Gonnick and Woolcott Smith, Harper Collins, NY, 1993.

[4] History of Statistics, Stephen Stigler, Belkings Harvard Press, Cambridge, MA, © 1986. This invention is also often attributed to Galton

This is the raw data, the S&P 500 for the period. Now, we want to go into the data mine and find the data to use to predict the stock index. If we included other US stock market indices like the Dow Jones Industrials or the Russell 1000, we'd see very good fits, with R-square numbers close to 1.0. But that would be an uninspired choice and useless at making the point about the hazards of data mining.

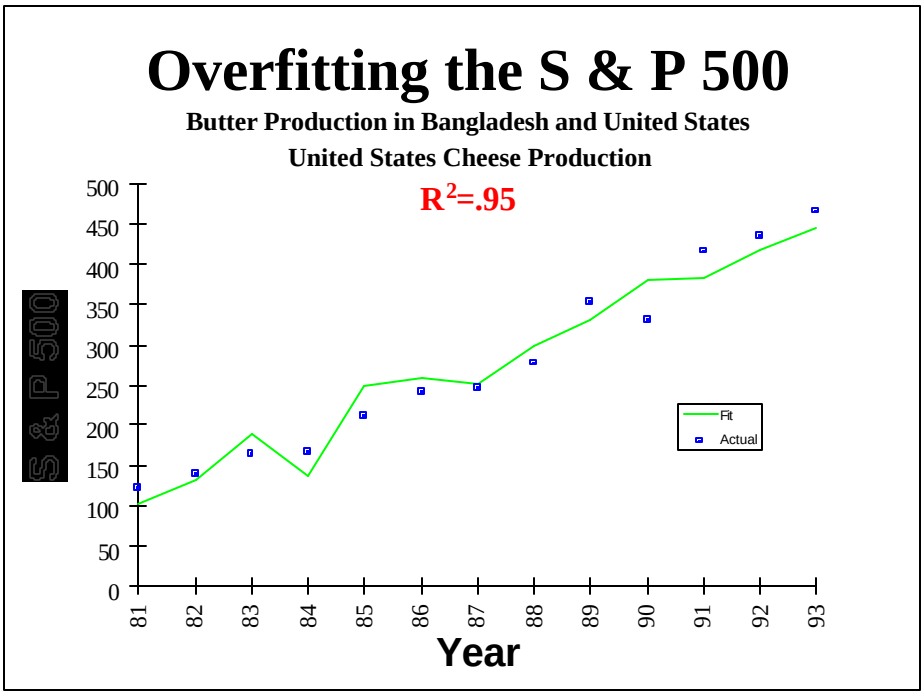
We're going to go find some data to mine in a big CD-ROM database of international data series published by the United Nations[5]. There are all sorts of data series from all 140 member countries in here. If we were trying to do this S&P 500 fit for real, we might look at things like changes interest rates, economic growth, unemployment and the like, but we'll stay away from those.

We found something even better: **butter production in Bangladesh**. Yes, there it is. A simple single dairy product that explains 75% of the variation in the S&P 500 over ten years. R^2 is 0.75, not bad at all.

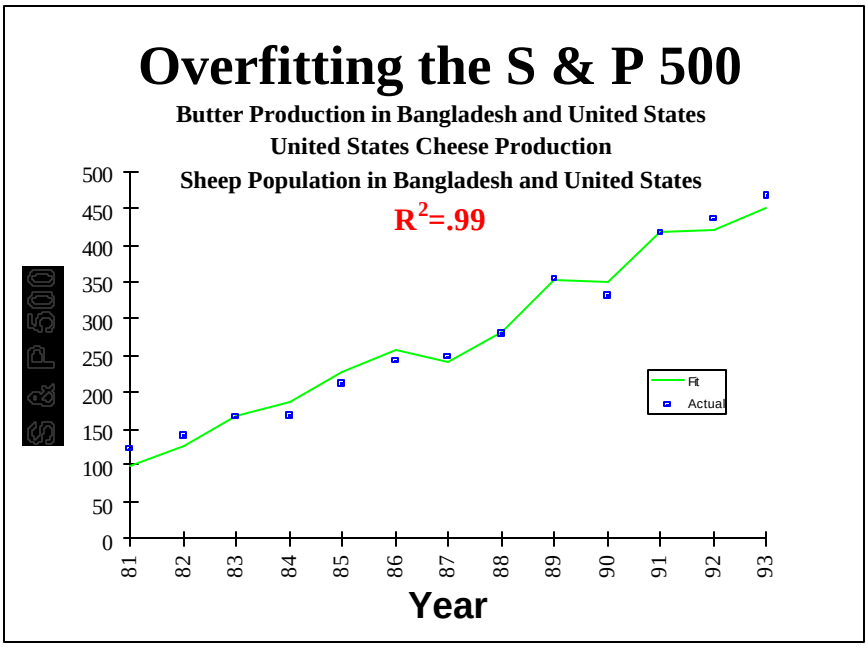


Why stop here? Maybe we can do better. Let's go global on this and expand our selection of dairy products: we'll use put in cheese and include US production as well. This works remarkably well. We're up to 95% accuracy here? How much better can we do?

[5] UN CR_ROM citation



How about 99% with our third variable: **sheep population.** This is an awesome fit. It seems too good to be true, but it is. It is utterly useless for anything outside the fitted period, a total crock before 1983 or after 1993. Just a chance association, which would inevitably show up if you look at enough data series, as we did. 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 like sheep.



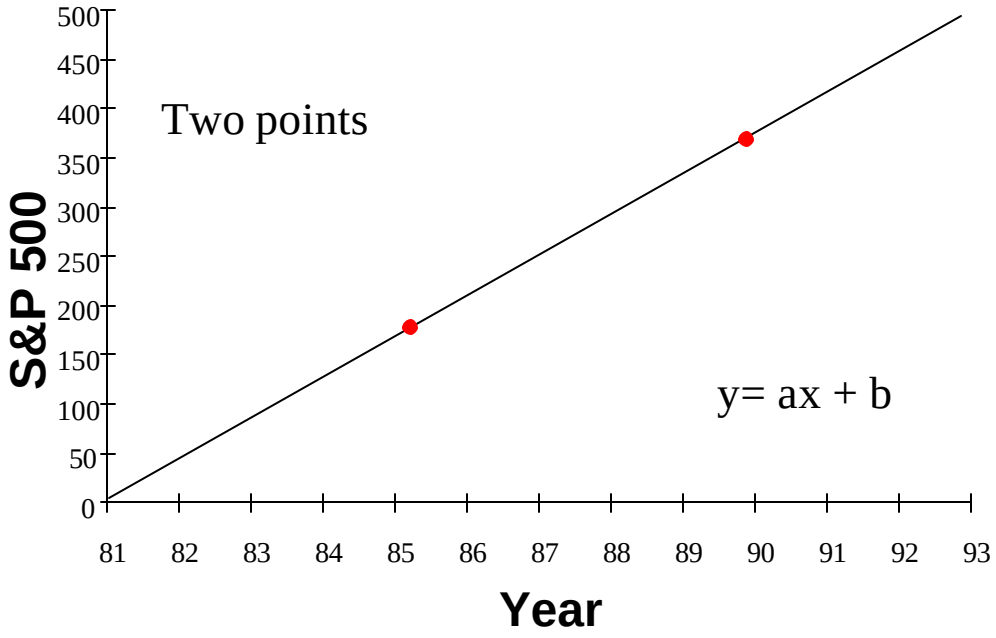
If someone showed up in your office with a model relating stock prices to interest rates, GDP, trade, housing starts and the like, it might have statistics that looked as good as this nonsense, and it might make as much sense, even though it sounded much more plausible.

ENOUGH REGRESSION TRICKS.

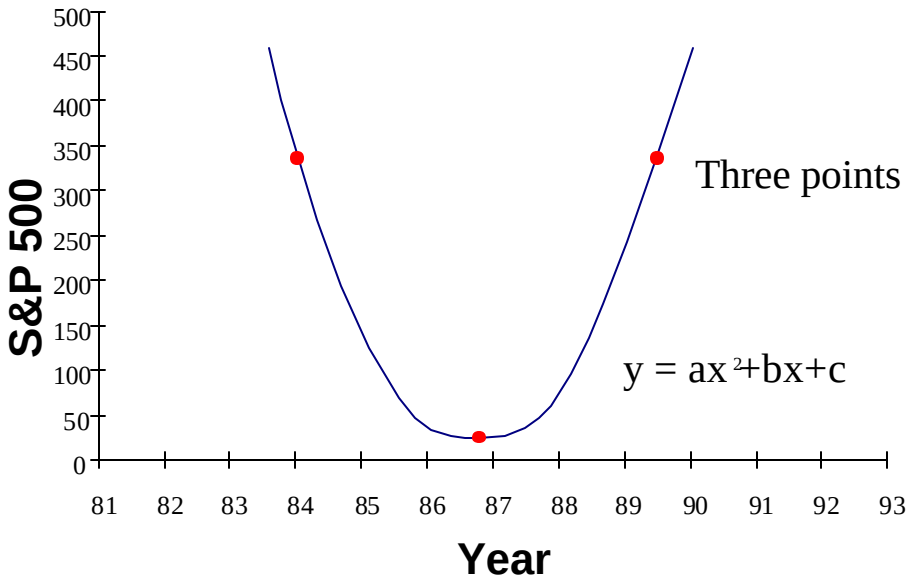
To hammer on this point about the dangers of data mining a little harder, lets show another equally bogus example. Who wants to go count pregnant sheep in Bangladesh to figure out next year's sheep population? It's dirty work and the food is lousy. We'll get away from ordinary linear regressions and show how we can fit a *perfect* model, with $R^2 = 100\%$ 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.

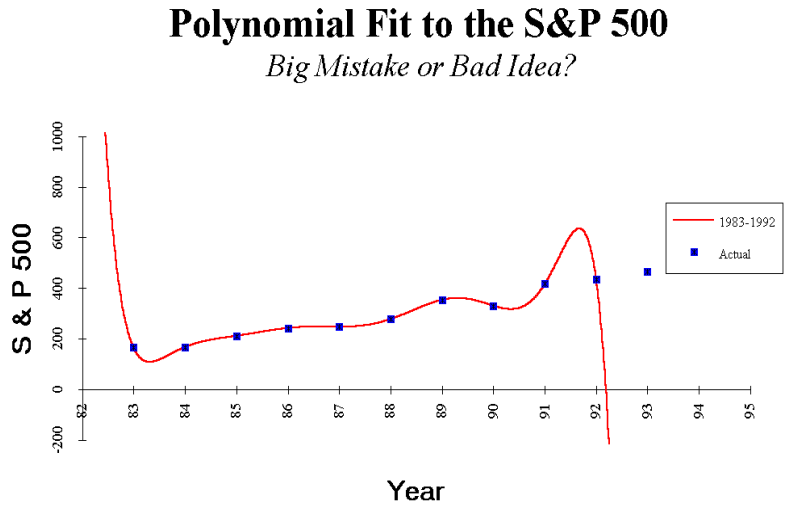
Instead of regression, we'll use a different prediction method to make this work, a polynomial fit. Everyone with any recollection of junior high school math knows that there's a line (a first degree polynomial) through any two points. Like this:



Put in third point and you can fit a parabola, or second degree polynomial through all three points. Like this:



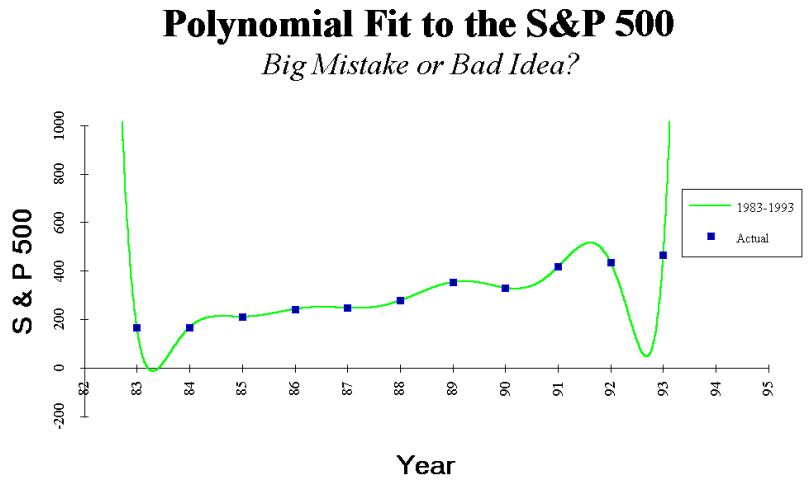
We have 10 points in the S&P 500 annual series from 1983 to 1992, so we fit a 9th degree polynomial[6]. It hits every annual close *exactly*. We've got 100% in-sample accuracy with only one variable.



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

[6] As Mr. Wizard says, "Don't try this at home, kids." 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.

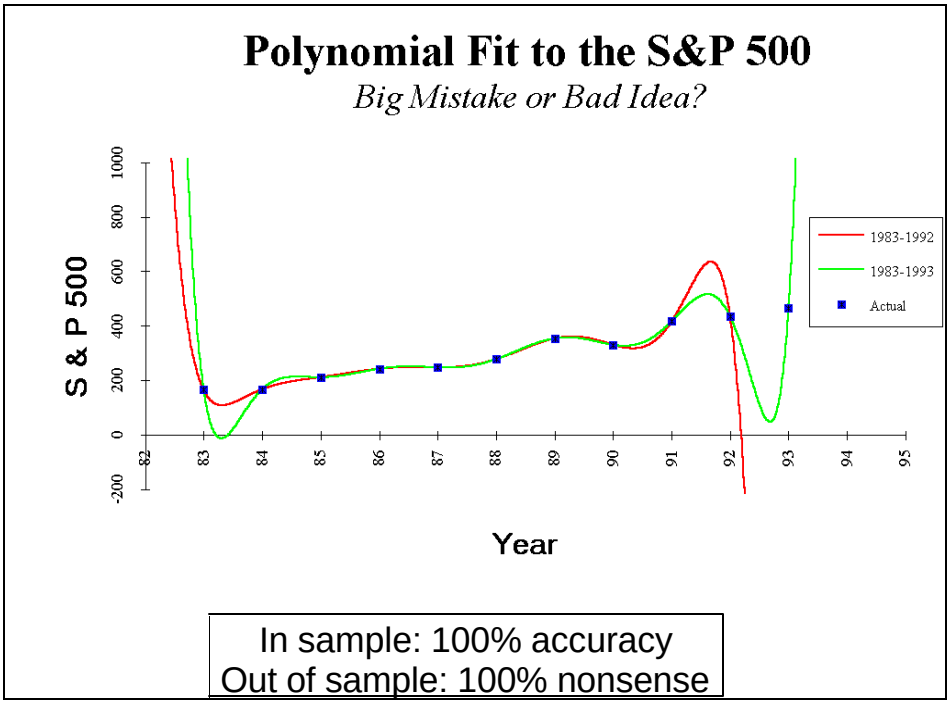
Notice that the fitted curve in the chart above is heading south very rapidly. What closing value the S&P did this method predict for the end of 1993? Minus 64,311. Fortunately for the global economy, it actually turned out to be positve, +445. We seem to have a problem with our model's out of sample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



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

Don't panic! 1993 ends, we get another data point, and restore our 100% in-sample accuracy, this time with a 10^{th} 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!.

So here we have two models, each 100% accurate in-sample, and each 100% nonsense out-of-sample.



IS THERE ANY HOPE FOR DATA MINERS?

The central problem is that the market has only one past. This will not go away.

Some forecasters just ignore this, 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:

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 datamine. An excellent article on the subject is "Behind the Smoke and Mirrors: Gauging the Integrity of Investment Simulations", by John Freeman (Financial Analysts Journal, Nov/Dec 1992, p 26)

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 (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 CUSIPS, 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. But how can we decide if the apparent improvements in a new model are significant? This question is address in "A Reality Check for Data Snooping" [Hal White, UCSD, working paper, May 1997], and implemented in software in the similarly named software [www.quantmetrics.com]

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 be brought to bear on forecasting problems. One neural net product advertised in 'Technical Analysis of Stocks and Commodities" claims to be able to "forecast any market, using any data". 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 any good? A useful indicator in answering this question is to take the same model building process, use to build a test model, for the same forecast target, but using a completely random[7] set of data. 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.

SERMONETTE

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

Just because something appears plausible, it doesn't mean that it is. The wide availability of machine readable data, and the tools to analyze it easily mean 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%, five of them are there just by chance. Look at 100,000 models at 95% 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's important to be very careful what you ask for, because you'll get it. Leaving some of the data out of the sample used to build the model, as we did with the polynomial example, is a good idea to hold back part of the data to use in testing the model out of the sample used to develop it. This holdback sample can be a period of time, as we did, or a cross section of data. The cross sectional hold back 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's 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.

Unlimited computational resources like dynamite, used properly, you can move mountains, used improperly, you can blow up your garage, or your portfolio. The easy access to data and tools to mine it has given new meaning to Mark Twain's admonition about "lies, damn lies and statistics." The old adage of Caveat Emptor, Buyer Beware, is still excellent advice. ***If it seems too good to be true, it is.***

[7] There are several alternatives in forming this random data to be used for forecasting. A shuffling in time of the real data will preserve 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.

ACKNOWLEDGEMENTS

I'd like to acknowledge the valuable contributions of David Krider and the quantitative investment firm First Quadrant. This paper grew out of examples that we developed to illustrate the risk of data mining in quantitative investment strategies.

I am forever indebted to Ron Kahn, of Barclay's Global Investing, for sharing his extensive collection of the work of world champion data-miner Norman Bloom.

BIBLIOGRAPHY

Freeman, David A., "A Note on Screening Regression Equations," *The American Statistician*, 1983.

Iyengar, S., and J. Greenhouse, "Selection Models and the File Drawer Problem," Statistical Science; 1988, 3: 109-135.

Judge, George, and M. Bock, *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*; Amsterdam, North-Holland; 1978.

Judge, George, W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl, Tsoung-Chao Lee, *The Theory and Practice of Econometrics*, John Wiley and Sons; 1985.

Leamer, E., *Specification Searches*, John Wiley and Sons, New York; 1978.

Lo, Andrew W., and A. Craig MacKinlay, "Data-Snooping Biases in Tests of Financial Asset Pricing Models," *The Review of Financial Studies*; 1990, vol. 3, #3; pp. 431-467.

Lo, Andrew W., and A. Craig MacKinlay, "Maximizing Predictability in the Stock and Bond Markets," mimeo; August 1992.

Markowitz, Harry M., and Gan Lin Xu, "Data Mining Corrections," *Journal of Portfolio Management*; Fall 1994

Ross, Stephen A., "Regression to the Max," mimeo Yale University; December 1987.

Hand, D.J., "Data Mining, Statistics and More," *American Statistician*; May 1998, vol. 52, no. 2, pp. 112.