Dodgy data
How not to construct a quant model

Quantitatively based investment models are only as good as their inputs, warns PHILLIP DOLAN. Biases can creep in at any stage. Similarly, analysis of investment manager performance can be flawed by the use of inappropriate data or methodologies.

When constructing a quantitatively based model, a number of problems have the potential to make any model appear to work better than is actually the case, with the consequence that the out-of-sample performance of the model when used in the management of “real money” is likely to be disappointing.

When analysing managers, similar problems can indicate investment performance different from that which would have actually been earned, or can lead to other incorrect inferences. Some of the problems are “obvious” errors; others are more subtle.

This article addresses these issues in the construction of quantitative models:

• the need to ensure data integrity;
• the dangers of data mining;
• the perils of backtesting; and
• the need to allow properly for transaction costs.

DATA INTEGRITY
The first task of the serious researcher is to ensure that the data being used are correct. Some of the problems that can arise include mislabelled series, the use of “stale” prices and mis-specified or incorrectly measured values.

Poor-quality data is one of the most basic problems that can occur when modelling. A particularly pernicious example of this is the ASX practice of re-issuing the three-letter codes used to identify firms.

“CBA” is an example of this. In 1981, a bank was listed on the stock exchange with the code CBA (Commercial Bank of Australia). It was subsequently acquired and delisted. At the time, it had a market capitalisation of $599 million. In 1991, there was also a firm (another a bank, in fact) with the code CBA (Commonwealth Bank of Australia). It had a market capitalisation of $1.8 billion. A user of the data who was unaware of the
circumstances, and who relied on the three-letter code, might conclude that it was the same firm, and that it had performed very well over the time it was delisted.

Other (potentially worse) examples exist. Consider the code "ALM". From 1970 to 1979 it was Allied Minerals, from 1983 to 1986 it was Anglo Gold Mines, from 1987 to 1989 it was Associated Liquor, and since 1993 it has been Australis Media.

A solution to this would be the use of a unique numeric identifier which can be attached to a firm and which never changes. One then needs to match the identifier with the current code if the codes change over time.

A more subtle problem arises when two firms merge and one wishes to create a time series that has back history for the merged firm. Which of the two merger partners should be “the firm” prior to the merger? This may be clear if one is much larger than the other, but this is not always the case.

The opposite problem can also occur. This was true in the case of Woolworths. In its previous guise, it was a listed company with the stock exchange code WLW. When it was eventually re-listed, it was assigned the code WOW. Again, for researchers who were not aware of this, it would appear that there were two different firms in the database, whereas the underlying assets were essentially the same.

Misuse of percentages

Economic and financial data are often expressed in percentage form. For example, we have returns (percentage changes in prices), growth rates of various economic variables such as GDP, fractions that are usually expressed as percentages (e.g. the unemployment rate, or the percentage of an overall portfolio represented by a given sector). Calculations involving changes in percentages can be a source of error. An X% increase followed by an X% decrease does not leave us back where we started.

A good (or bad) example of this (reported in *The Economist*) was the case of a local government in Mexico that re-painted the lines on one of its roads to increase the number of lanes from four to six. This resulted in a 50% increase in traffic capacity. As there was a big jump in the number of accidents on the road, the authorities had it re-painted again, reverting to four lanes. This was a 33% decrease. They then reported at the end of the year that there had been a 17% increase in traffic capacity on the road, since it had gone up by 50% and down by only 33%.

Data integrity – thin trading

An assumption often made when looking at prices is that one could have traded at the reported price (at any volume). This is clearly unrealistic.

To get around the problem of “stale” prices, where the last recorded transaction price is out of date due to a firm not trading, some researchers use bid or ask prices to get an indication of the price at which they could have traded. However, since there is no volume, one cannot allow for potential market impact.

An example of this is a bid price for BHP, recorded on the Perth stock exchange after the east coast exchanges had closed, that was significantly above the close at which actual trading had occurred. Given the size of BHP in the index at the time (about 14% of the market), this had the potential to make any strategy that was overweight BHP and that calculated its return on the basis of the Perth price, appear very attractive.

Mis-specification

We need to be sure that we have measured what we thought we measured.

For example, if we are examining manager performance, are the returns with which we are working before or after fees (or, even worse, are they pre-fees for some managers and post-fees for others)? This can be a problem when some managers charge their fees by reducing the unit price while others do it by maintaining the price and cancelling units.

Similarly, are all the returns pre-tax or post-tax?

Are the averages with which we are working arithmetic or geometric? The “correct” method depends on the use to which the analysis will be put. Use of arithmetic returns that possibly ignore the effect of compounding can give wrong results.

Data obtained from external sources need to be examined carefully. For example, are the estimates of earnings per share obtained from brokers before or after abnormals?

Do all brokers formulate their estimates on a similar basis?

Extreme values

When modelling, we are often faced with deciding how to treat extreme values. For example, should we include or exclude the 1987 stockmarket crash? While it may seem reasonable to exclude what are regarded as “unusual” events, there can often be considerable value in examining behaviour at such times (e.g. perhaps the best time to get an estimate of elasticity of demand for petrol is during an oil shock when prices jump).

The question of what constitutes an “outlier” is non-trivial. For example, most people would regard the 1987 crash as an extreme value, but what about the very strong market performance in the lead-up to it? A possible reason for the poor out-of-sample performance of tactical asset allocation models is that, to look any good at all, they must have had investors out for the crash, but the models are fitted to ensure that the investors were in the market until just before the crash. A handful of data points end up driving the whole model. If one decides to exclude the crash, a case can be made for also leaving out the run-up.

Just how bad the crash was is also a non-trivial question. If one estimates how many standard deviations from the mean was the monthly return for October 1987, using only data to that point, it appears as a much “rarer” event than if the estimation period includes October 1987 (it was a “9 sigma” event using data to September 1987, and only a 4.7 sigma one if October is included).

DATA MINING

Researchers seldom, if ever, approach a problem with no existing beliefs. They may have read or heard about the research of others, attended the same conferences, or be examining the same data sets.

The publishing process tends to ensure that only “successful” or “interesting” research is widely distributed. In such cases, classical statistical inference techniques (which, for
example, assume that only one regression has been run before the tests are performed) are not valid.

If a top-line journal publishes, say, only one paper out of every 20 that it receives, and if researchers are using 5% significance levels in their tests, then sheer chance could be driving many of the “statistically significant” results we see reported.

**Spurious correlations**

The advent of large on-line and easily accessible databases, combined with the dramatic reduction in the cost of computing power, has made it easy to “mine” data.

As one observer has commented, the torture of the data, especially by electronic means, can eventually make it confess to anything. In many cases, this can result in models that lack economic intuition.

Suppose we are seeking a relationship that will explain the level of inflation in the UK. There are a number of variables we could choose. We test a number of them, and one that seems promising is the growth in the money supply two years prior to the period for which we are trying to forecast inflation.

Fitting the data in a linear regression gives us a very good fit (see Figure 1). Can we conclude from this that there is actually a causal relationship?

Figure 1 seems to suggest a strongly positive relationship between increases in the money supply and subsequent (two-year lagged) inflation. The correlation coefficient is very high at 0.848, and any of the usual tests of statistical significance suggests that one could expect to do a good job of forecasting the future price level based on current changes to the money supply.

Consider an alternative way of predicting the UK inflation level. One of the variables tested by researchers (along with the growth in money supply) when looking for a model was the incidence of dysentery in Scotland. Why did they test this? Why not? The data were available, after all.

Fitting a linear relationship to dysentery cases and one-year lagged inflation produced the result shown in Figure 2. The correlation of –0.863 was slightly higher than that obtained for the money supply explanatory variable. If we were ranking the models solely on the basis of highest explanatory power, we would prefer this model over the previous one.

Does an increase in the incidence of Scottish dysentery result in a fall in the price level in the UK? If so, it’s one more reason I’d rather be “short” dysentery than “long” it.

(The above two examples were developed by David Lynch and reported to the Q-Group as an entry in the group’s Spurious Correlations contest.)

**Econometric modelling**

Suppose we are seeking a relationship between some financial variable and one or more explanatory variables. We can select from hundreds, even thousands, of such variables. For each of the “standard” ones, we can use transformations such as leads, lags, logs, moving averages, etc.

For example, we might have 100 variables we could use, and we might decide to select...
the five that give the best fit. How confident can we be if our model has been developed in this way?

A standard “cut-off” level for tests of statistical significance is 5%. By this, we mean that the chance of seeing a relationship as strong as that obtained, when there is in fact no relationship, is only 5 in 100. But, suppose we tested 100 variables in the first place. We would expect that 5 of those 100 would appear significant at the 5% level (by definition) even if this were not true.

Causality
It is usually assumed that if there is any causality present in a model, then a variable that precedes another is the one responsible for it.

This is not always so. Consider the question of whether Christmas cards “cause” Christmas. For as far back as we can look, the sales of cards have preceded Christmas. One can predict the arrival of Christmas quite accurately and with great consistency by observing whether or not there were sales of cards. This is a trivial example, but less obvious ones have similar characteristics.

Often there can be an underlying source of causality that can generate a seemingly spurious relationship. Consider research showing a strong positive relationship between the rate of algal growth in ponds and sick days taken by bus drivers. The initial reaction might be that this fits into the “inflation and Scottish dysentery” category. As it happens, both the effects have a quite plausible underlying cause, namely the temperature on any given day. Very hot days are good for algal growth, and also seem to result in above-average numbers of sick days for bus drivers.

Lookahead bias
Suppose I am developing a model that relates changes in GDP to some financial variable (e.g. interest rates). I get data for each of the series and fit a model.

But have I assumed, say, that I knew the June quarter GDP value on 30 June? It is not released until about eight weeks after the end of the quarter. Any model that assumes I knew on 30 June information that the rest of the market did not get for some weeks is bound to perform well.

Also, the number initially released may subsequently have been revised. The market would have reacted to the original number, but some sources for the historic data would have only the revised values.

Similar comments apply to company announcements, such as earnings. Assuming a company’s information was available at its balance-sheet date introduces a large potential bias, since it pre-supposes an ability at the balance-sheet date to forecast earnings surprises, which we know are a strong driver of stock returns.

Survivorship bias
In many cases, a particular data series membership of a larger set (such as that of an individual stock in the All-Ordinaries index) depends on how that series has behaved relative to the other members of the set.

Thus, poor performing stocks may exit the index, fund managers performing below average may be sacked by their clients and so disappear from the published surveys, etc.

Figure 3 shows the All-Ords accumulation index (solid line) over the past 20 years. The dotted line shows a “biased” version of the index, in which I selected all the stocks in the All-Ords as at 31 December 1997 and took their performance back through time.

We see that, overall, the biased series has outperformed the “real” All-Ords. This is because it excludes all those stocks that were in the index previously but had exited before 31 December 1997. Thus, a guaranteed way to beat the index over the period 1977–97 is to invest only in stocks that are still in the index in 1997. Easier said than done.

Any quantitative model that selected only from stocks in the index at December 1997 will be biased in its reported performance during any backtesting over earlier periods. It is worth noting that, depending on the period, the bias can go either way.
Figure 4 compares the performance of the “true” All-Ords and our biased series in the period 1985–96. Here we see that the unbiased index has done better. This is because, in the lead-up to the 1987 crash, the main reason firms exited the index was not underperformance, but because they were taken over. Such firms typically did better than the index as their prices were bid up during the takeover. Omitting them from the index has the effect of biasing the performance downwards.

TRANSACTION COSTS

Any well-conducted research must allow for the impact of transaction costs on the performance of a proposed investment strategy. Such costs always serve to reduce returns, so ignoring them will bias reported returns upwards.

One can estimate some costs (such as brokerage and stamp duty) reasonably accurately, but market impact is much harder to gauge when, for example, running a paper portfolio or undertaking a backtest.

Additionally, use of mid-rates, rather than the bid when selling or the ask when buying, can introduce a bias that will make results appear better than they really are.

CONCLUSION

There are potential traps at every stage of the research process. This article has touched on some of the most egregious; any good statistics text will have many more.

Unfortunately, many of the biases that can creep do not “cancel out” but rather have the effect of helping to produce the result that the researcher wants to see. This suggests that care should be taken in interpreting test statistics, and that one should be extremely conservative when quoting what appear to be significant results. The end result is likely to be a more robust model.

REFERENCES

