Beware fake econometrics

Transportation consultants are regularly misusing mathematical methods to flatter their models' explanatory capabilities. **Robert Bain** explains how and why investors should heed the warning signals

In my reviews of transport demand and revenue projections for investor-financed road and rail projects, I am frequently presented with econometric models. Typical uses are for forecasting traffic or ridership in a simple, brownfield corridor with limited alternatives or to grow demand matrices (trip tables) in the context of broader network models.

The narrative that accompanies these pivotal growth models ranges from near-zero to a deep-dive into the econometric analysis of cross-section and panel data, complete with a battery of t-statistics and p-values in support. The aim of the latter is clearly to demonstrate that serious scientific analysis is in play; credibility through complexity – and woe betide any mere mortal who dares venture there. Yet, a look under the hood suggests that all may not be as it first appears when one reflects on exactly what is being done and why.

DATA TRANSFORMATIONS

Take a simple example. Your consultant claims that the observed growth in car trips in the vicinity of a toll road can be explained entirely by the recent performance of GDP. As the variables (vehicle trips and GDP) employ different scales, they are normalised to a common starting point using indices – typically to a base of 100 – for comparison and analytical purposes. After all, it is the relative (not the absolute) changes in the variables across time that are important.

The next step is to cast the growth model as a log-log formulation by calculating the natural logs of the dependent and independent variables. This is done despite the absence of any text justifying this choice of data transformation or explaining why alternative approaches have been rejected. Log transformations are easy to apply and are certainly popular in economics. They are widely used, particularly when the relationships being examined are non-linear in the parameters. The transformation converts multiplicative relationships into additive ones; allowing for exponential (compound growth) trends to be explained by linear models. If you have a variable, such as tram ridership that increases at a constant or near-constant percentage rate, the log of that variable will grow as a linear function of time.

However, a key reason for the popularity of log transformations is that the estimated coefficients in log regressions have a nice interpretation. The coefficients measure the percentage change in y (trips) that occurs in response to a certain percentage change in x (GDP). In other words, they are elasticities – and elasticities are simple to explain and apply.

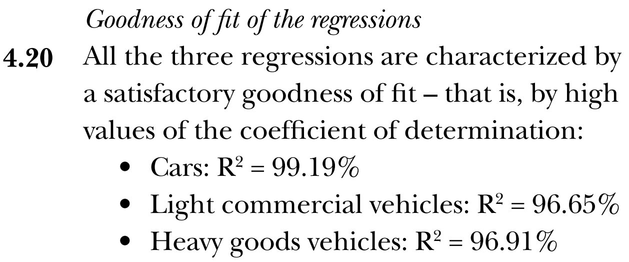
However, such transformations need to be applied cautiously and appropriately. For example, the log-log model assumes a constant elasticity over all values of a data set – whereas in the transportation sector we generally observe elasticities, such as price elasticities, that change at different tariff levels. And the results of standard statistical tests performed on transformed data simply may not apply to the original, non-transformed data.

STATISTICAL MISREPRESENTATION

One impact of adopting the approach outlined above is that it exaggerates the model's goodness-of-fit. The resulting coefficient of determination is often spectacular (R-squared > 0.95) – see

Figure 1, lifted directly from a consultant's report. This is an immediate red flag. Whereas high coefficients of determination may be observed in the physical sciences, anyone dealing with human behaviour (social scientists) will be more familiar with values, as widely reported in the literature, generally sitting around 0.5 – or lower.

FIG.1



Goodness of fit of the regressions

4.20 All the three regressions are characterized by a satisfactory goodness of fit – that is, by high values of the coefficient of determination:

- Cars: R^2 = 99.19%
- Light commercial vehicles: R^2 = 96.65%
- Heavy goods vehicles: R^2 = 96.91%

The objective of the exercise appears to be to get as high an R-squared as possible. Thus, having supposedly demonstrated their model's credibility and outstanding performance, the consultants can move unchallenged from model estimation to forecasting using predicted values for their explanatory variables. I'm not going to dwell here on the uncertainties (and sources of error) introduced at this stage, but I will point out that if a high R-squared is the goal then this is invariably mathematically achievable.

However, as Harvard Professor Gary King points out ('How Not to Lie with Statistics', 1986), this objective is not the same as that for which regression analysis was designed: "The purpose of regression analysis … is to estimate interesting population parameters (regression coefficients in this case). The best regression model usually has an R-squared that is lower than could be obtained otherwise. … strategies to increase your R-squared will add nothing to your analysis … and nothing useful in explaining your results to others. (This) general strategy of analysis will likely destroy most of the desirable properties of regression analysis."

Notwithstanding, transportation consultants use high R-squared values to trumpet their modelling capabilities and give client comfort. Often, unguarded (and unwise) comments follow about predictive ability, but let's ignore that for now. The point is that the R-squared is being misrepresented. The statistical significance of such regressions is used as if it is supportive of correct model specification and strong causal relationships. However, the high significance is typically only because the variables (vehicle trips, GDP or whatever) have upward trends – and that is what is being picked up. Seasonal dummies and/or other binary variables established to account for any data inconveniences simply serve to further flatter the model's fit.

EXAMPLE

To demonstrate this, I set up a simple example which readers can recreate in a spreadsheet. It takes less than 10 minutes.

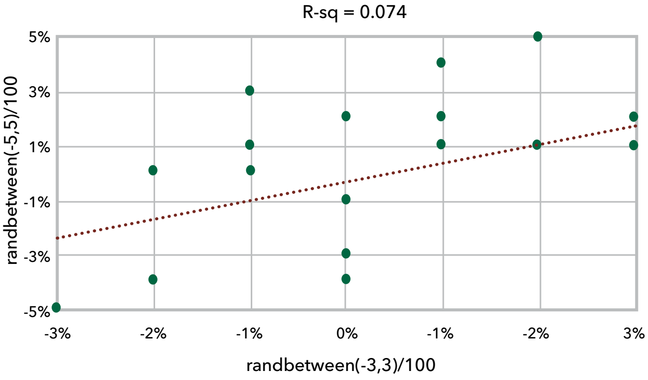
Preparing the Data

Create 20 years of annual random data for two variables (A & B):
A [=RANDBETWEEN(-3,3)/100]; and
B [=RANDBETWEEN(-5,5)/100].

When plotted, these variables (of course) suggest no relationship. They are entirely random. Regress one against the other (see Figure 2) or calculate the correlation coefficient (or Pearson's R) and the results will confirm that the variables are unrelated and uncorrelated.

I selected random numbers in the -5 percent to +5 percent range, as this degree of variation – around a long-term trend – is often observed in real life data.

FIG.2



I then introduce two growth trends with modest percentage growth rates (2 percent and 3 percent), again often observed, and transform the variables into indices starting with a base of 100 and successively applying the resulting growth. An extract from the spreadsheet is shown in Figure 3. The indices are labelled A+ and B+ respectively.

FIG.3

Year	Random Numbers		Growth		Indices	
	A	B	+2%	+3%	A+	B+
0					100	100
1	-1.0%	1.0%	1.0%	4.0%	101	104
2	-2.0%	-4.0%	0.0%	-1.0%	101	103
3	2.0%	1.0%	4.0%	4.0%	105	107
4	0.0%	-4.0%	2.0%	-1.0%	107	106
5	1.0%	1.0%	3.0%	4.0%	110	110

This is effectively the starting point for transportation consultants. They have several variables with growth rates that oscillate from period to period around some longer-term trend, which they transform into indices.

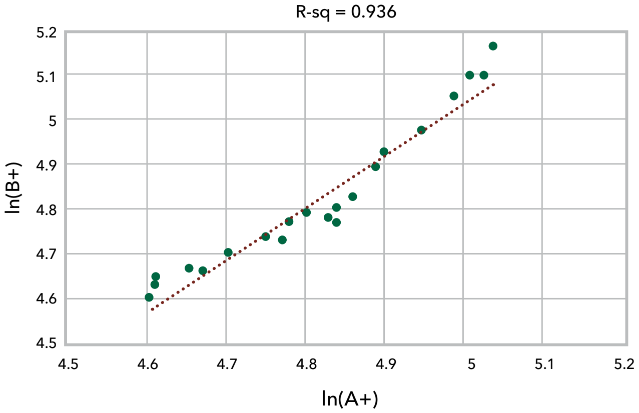
Analysing the data

As mentioned earlier, a common practice – at this stage – is to apply log-log transformations. At first glance, this appears to be reasonable. As I've introduced constant growth rates (as opposed to constant growth) the underlying trends will be multiplicative i.e. exponential.

Logging introduces linearisation. It converts exponential into linear trends, making the transformed data more suitable for fitting with linear regression models (next step).

Figure 4 shows what happens when you now run a linear regression of one of the transformed variables against the other.

FIG.4



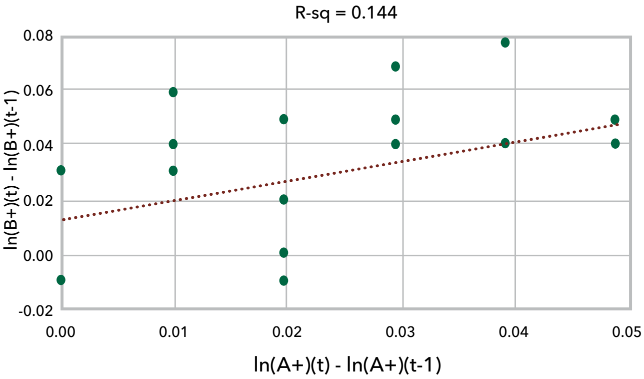
Lo and behold, a very strong correlation appears with a client-pleasing R-squared of 0.94 – yet, this is entirely spurious. The goodness-of-fit statistic is merely capturing the internal trends, both of which are upward. However, in the absence of further information, Figure 4 would be enough to convince anyone that A and B are very strongly correlated indeed.

In truth, one trended time series regressed against another will often reveal a strong, but spurious, relationship – and the use of log transformations does nothing to alter this fact. This is due to a mutual dependency on the passing of time. It is time – known as the confounding variable – that correlates the two series. The internet is awash with examples of this phenomenon, many of which are comical. Correlating per-capita cheese consumption with the number of engineering doctorates awarded, for example, reveals an R-squared of 0.96 (http://www.tylervigen.com/spurious-correlations). But, this is no laughing matter when trying to forecast travel demand, which lies at the very heart of the valuation process used when bidding for transportation concessions.

Recall that we started this illustrated example by creating entirely random variables yet, very quickly, we entered the murky world of spurious correlation and seductive explanatory charts.

If, instead, we examine the growth rates of our two variables – discussed below – by calculating, for example, ln(A+)(t) – ln(A+)(t-1), the spurious relationship simply disappears (Figure 5).

FIG.5



INTEGRITY & TRANSPARENCY

Despite what transportation consultants are doing in practice, the books on my shelf suggest that the appropriate statistical treatment of trending (i.e. non-stationary) variables is to estimate the relationship between the changes or growth rates of the dependent and independent variables. This de-trending removes the underlying impact of time – the confounding variable – allowing for the variations in one series to be contrasted with the variations in others and possible causality to be examined. This is where analysis needs to be focused. Inconveniently, it will show much noisier relationships than the log-log approach (lower R-squared's) however it is much better for model identification (selecting the right macroeconomic variables to use).

The choice of which predictors to use – and the non-trivial challenge of sourcing reliable forecasts for those at the appropriate level of spatial disaggregation – is difficult enough without being distracted by goodness-of-fit statistics which, although presentationally pleasing, are simply an artefact of method.

From the perspective of the transportation investor, an important part of technical due diligence is reproducibility. For cornerstone assumptions (such as growth expectations) we should be able to replicate what our consultants have done. They need to make all their data and econometric models available for cross-examination so that we can clearly understand the process – what and why – and assess output integrity and reliability. This traditionally hasn't happened in the past. We should insist on it in the future. ■

Robert Bain runs RBconsult (www.robbain.com). He reviews transportation demand forecasts for many of the world's leading infrastructure investors and acts as an expert witness in international litigation and arbitration cases. He is a visiting research fellow and part-time lecturer at the University of Leeds.