G+1 0 More Next Blog»

LSE Spatial Economics Research Centre

Thursday, 12 September 2013

HS2 Regional Economic Impact: Garbage in ...?

[Posted by Prof Henry G. Overman]

I have finally found time to look at the technical appendix for the HS2 Regional Economic Impact report that emerged yesterday and underpins the wid cited '£15bn a year' benefit claim.

I'm going to preface my comments here by observing that modelling this kind of regional economic impact is *very* difficult. That said, assuming I have understood it correctly, this report does things that are technically wrong (and these things are crucial for the analysis).

To understand the economic impact of HS2 we need to know how changing the connectivity between places affects local economies. The starting poin for figuring this out is to recognise that better connected places have higher productivity (i.e. they can produce more goods and services with less resources). The report uses statistical analysis to make this relationship more precise. To do this, it takes data on wages in 235 local areas and data o the connectivity of those places and looks at how closely those things move together. In technical terms, it runs a regression of a wage based measure productivity on transport connectivity (controlling for some other things that you might think would influence both productivity and connectivity). You mig worry that more productive places end up with better transport links - i.e. there's a chicken and eggs problem here. Let's set that issue aside (as the report does nothing to address it and the literature suggests this is not as important as you might think). Once we know the relationship between productivity and connectivity, we can model how HS2 changes connectivity and hence back out the effect on productivity. So far, so good (this is something I have done in the past for the Northern Way).

The report wants to do something more ambitious by looking at how different types of connectivity affect productivity in different industrial sectors. To d this, it constructs four different measures of connectivity - via the rail network to workers and jobs and via the road network to workers and jobs. The problems start because these four measures of connectivity are very highly correlated (i.e. they tend to move together) so that places with good road connectivity also tend to have good rail connectivity. In fact, they are sufficiently highly correlated that its very difficult to figure out which of them matte To 'solve' this problem the report looks at each of the correlations in turn - asking how does productivity change with rail connectivity to workers, then w rail connectivity to jobs etc. Unfortunately, as we teach our undergraduate students, this doesn't actually solve the problem. It just attributes the effects all the different types of connectivity to the particular type of connectivity you are looking at. It's a little hard to understand what they do next if you've never done any basic statistical analysis (and also because the report is a little unclear). Let me give the technical explanation - I've tried to give a non technical analogy at the end of this post.

From a technical point of view they run four *univariate* regressions of productivity on the four different connectivity measures to give estimated coefficients b1, b2, b3 and b4. They then sum these coefficients to give a 'total' of the estimated coefficients (b1+b2+b3+b4.) Next they take the larges coefficient (say b3) and assume that the effect for each connectivity measure can be calculated using its share in the total coefficients multiplied by b3 (the largest estimated coefficient. So, the coefficient for the first connectivity measure can be calculated as b3 * b1/(b1+b2+b3+b4); for the second connectivity measure it can be calculated as b3 * b2/(b1+b2+b3+b4), etc. The report notes that 'this approach does not have a firm statistical foundatio This is an understatement - this procedure is essentially unfounded and produces estimates of effects that are meaningless. The only world in which th would make sense is one in which the connectivity measures were perfectly correlated so each of them got the same estimated coefficient and you we going to pretend that you could separate out the effects for four different changes (which of course, you can't do).

We had the same problem in work we did for the Northern Way. However, unlike the HS2 report, we present both multivariate and univariate coefficien (for two, rather than four, connectivity measures - one for the car and one for train). In our first set of univariate regressions the coefficient for car connectivity is 0.230, that for train connectivity is 0.344. Adopting the HS2 approach, we would calculate 'corrected' coefficients on car connectivity as 0.137 [=0.344 * 0.230/ (0.230+0.344)] and one for train connectivity of 0.206 [calculated similarly]. Neither of these turn out to be correct. In the multivariate regression, the correct way to calculate the partial effects these coefficients are 0.084 and 0.258 respectively.

In short, it's hard to know how to interpret the coefficients in the report. That's made harder by the fact that the paragraph of the report that explains ho to interpret these coefficients [6.3.38 for those of you that want to take a look] is unfinished - '[this procedure] enables connectivity to other businesses and to labour, by car and by rail, to be reflected in the analysis and captures.' Personally, I think that sentence is completed by the word 'nothing' but I suspect not what the authors intended.

Issues of interpretation aside, it's as worrying that, in our analysis for the Northern Way the coefficient on train connectivity *reduces by a factor of 6 to* once we properly account for other things that might be both driving productivity and accessibility. In the HS2 report, the one thing they do to address t problem is a (somewhat opaque) correction for skills which, worryingly, hardly changes the estimates.

So, on my reading, technically wrong and possibly out by orders of magnitude. I can imagine why the government has rushed this report out, but it wou appear to add very little, if anything, to the debate.

Non technical analogy: I've been trying to come up with an analogy that conveys just how wrong this is to someone who is non-technical. The best I ca do is to think about trying to understand net daily calorie intake. We know that both calories in (through eating) and calories out (through vigorous exercise) affect net daily calorie intake. We know that a 1% increase in the amount eaten will have a much bigger effect (lets say a 10% increase in weight) than a 1% decrease in the amount of vigorous exercise taken (lets say a 1% decrease in weight) because baseline activity requires a couple o thousand calories a day even if we are not taking any vigorous exercise. This simple relationship is the reason why increasing daily exercise by some percentage is not as effective in weight loss as decreasing calories consumed by the same percentage. If we looked at the association between weigh and daily diet, ignoring exercise taken (i.e. did a univariate regression) - we might find that people with 1% higher daily calorie intake have 11% higher weight. If we then looked at the association with exercise (ignoring calories consumed) we might find that those that exercise 1% less have a 9% highe weight. That is, we estimate effects that are larger than the true effects - a little higher for calories, a lot higher for exercise. The problem, of course, is that exercise and calorie intake tend to be quite highly correlated, because people who exercise a lot tend to watch what they eat and vice versa. This causes us to overestimate the effect of calories consumed (because high calorie consumers exercise less) and significantly overestimate the effect of exercise (because low exercisers also eat more). To figure out the effect of exercise on its own we really need to vary the amount of exercise holding t amount of food intake constant - i.e. do a multivariate regression. This would give us the 1% effect that I assumed above and a 10% effect for obesity. Instead, if we followed the approach adopted in this report we would sum the two coefficients to give 20%, take the larger coefficient of 11% and work the coefficient on of exercise on weight loss as 4.95% [11% x (9%/20%)] and that on calorie intake as 6.05% [11% x (11%/20%)].

Posted by Prof Henry G. Overman on Thursday, September 12, 2013 G+1 Recommend this on Google

1 comment:

Anonymous said...

Can the modelling techniques be used in 'reverse'. That is, if it is decided which towns and cities require more economic growth (because they are lagging behind other parts of the country), can the modelling techniques help decide which places should have improved links, and to where, in order to help most to improve the local economy of those places?

21 November 2013 at 11:30

Post a Comment

Newer Post Home Older P

Subscribe to: Post Comments (Atom)

Simple theme. Powered by Blogger.