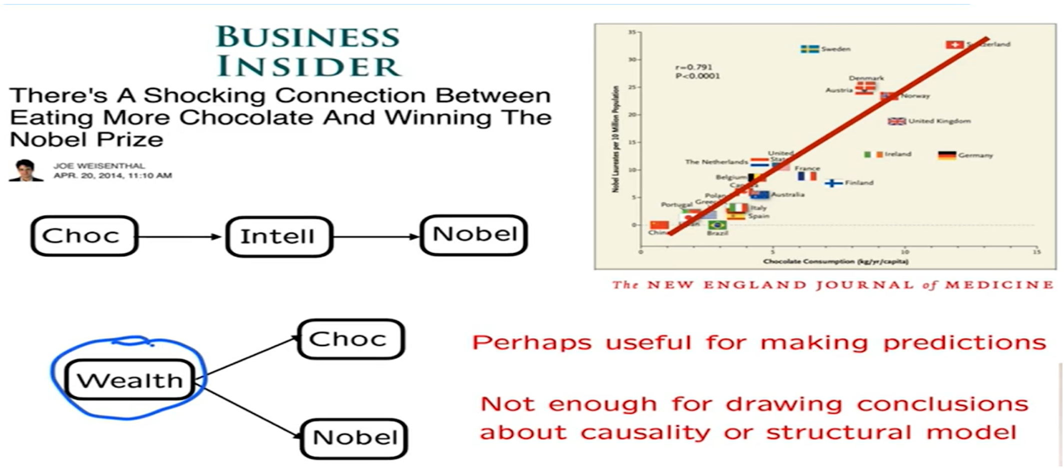
Latent Variables

Let us consider the below example, the source of the difficulty is that there may be a latent variable, a hidden variable sitting in the background that drives the phenomenon of interest but which was not taken into account in our model.



BUSINESS INSIDER

There's A Shocking Connection Between Eating More Chocolate And Winning The Nobel Prize

JOE WEISENTHAL
APR. 20, 2014, 11:10 AM

The NEW ENGLAND JOURNAL of MEDICINE

Perhaps useful for making predictions

Not enough for drawing conclusions about causality or structural model

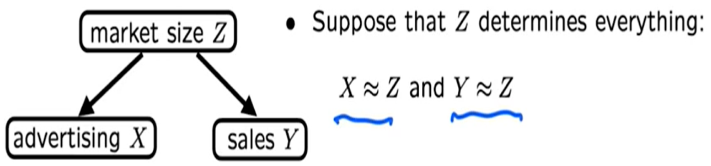
Now, let us consider another example that can arise in the context of advertising. Suppose that X is our advertising budget, just assume that we have a single channel and we're interested in making a decision.

Should we increase our advertising budget or not? Is it going to increase sales or not?

So here we're trying to make an inference about the causal relation: **do changes in X cause changes in Y?**

Now that the world is like this, we work just with advertising data and sales data. But there's a hidden variable in the background. Well, it's not necessarily hidden but we need to not take it into account in our model and that variable is the market science. Maybe the world is like this. Maybe the marketing department looks at the market size and chooses the advertising budget and maybe sales only depend on the market size but are not affected at all by the budget. That's one possible world.

So the market size just determines everything and there's no relation between advertising and sales in the sense of a causal effect. Now we take our data in this model and run the regression.



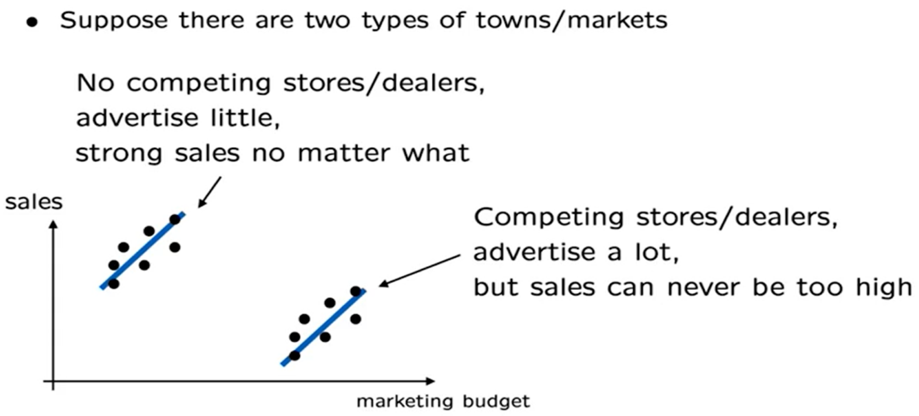
- Suppose that Z determines everything:

$X \approx Z$ and $Y \approx Z$

With this model, X is always approximately equal to Y. So we find a very good linear relation between X and Y. If we know X then we know Z and then we know Y. So X allows us to make very good predictions. So regression is going to do great. We are going to get a great R-square. We have a very small prediction error. On the other hand, if we are to use these results, to make statements about the effect of X on Y, we will be very wrong even though Y runs along X, if we do change X, and this is the true model of the world. X is not going to have any effects on Y. In other words, we can make good predictive statements if you know X you can predict Y. On the other hand, there is no basis for making statements about the effects of one variable on another.

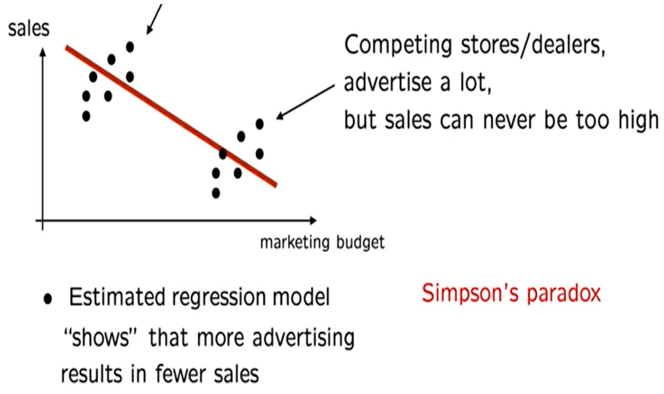
A situation like this, unfortunately, is very common and it's a big source of difficulty when trying to do statistics in fields in which there are many possible hidden variables or latent variables that affect what we see. And in particular in the social sciences or in the Biomedical Sciences whenever you deal with something complicated. This issue is always bound to come up. Can we tell whether we're discovering a causal relation? It is actually much, much more difficult than just running a linear regression model.

Here's another example, again from the advertising and sales story that we have been using. Let's again, assume that the marketing budget is one-dimensional, that we have only one channel, and suppose that there are two types of towns or markets. There are some towns in which you have a monopoly. There are no competing stores. So even if you advertise very little, you're going to capture that market and you're going to have strong sales and if you market a little bit more, if you have a little more advertisement, then sales will also go up and you run a regression on those types of towns and you find the positive relation, which makes sense.



Now suppose that there are some other towns and in those towns, there are many competing stores, many competing dealers. So when you advertise, your advertisement does help your sales and you have an upwards relation, but because of this competition, your sales will never be as high as they were in the first group of towns.

Now, suppose that you take this data set and give it to a regression algorithm. What is the regression going to do? It's going to do a least-squares fit and it's going to discover a line that goes this way and actually slopes downwards. So the estimated regression model seems to be saying that the more you advertise, the fewer sales there are. And yes, it is true in the towns that you advertise more, you actually have smaller sales. That's a true fact. But these smaller sales are not caused by the increased advertisement. It's not a causal relation. It's rather because there is this hidden variable that there are two types of markets and we have not taken that into account.



This situation where regressions within groups move in one direction, but when you do the regression for the entire population, the regression line moves the other direction, shows up in many other contexts and it has a name, it's called The Simpson's paradox. And sometimes it can create interesting paradoxical stories.

But once more the moral is that if you're not properly taking into account the hidden variables, the latent variables, and all the variables that might matter, then you might get regression results that give you the wrong type of qualitative conclusions.