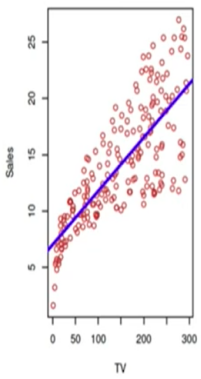
Heteroskedasticity and Multicollinearity

Let's just take stock of what is to happen next. We're going to start by discussing what can go wrong and there are some technical issues and then some more serious issues that have to do with latent variables that are not included in the model.

Then we'll discuss ways of mitigating the various things that can go wrong, which is to add more variables, perhaps by exploiting the data that we have in certain nonlinear ways. On the other hand, we should not overdo it by having too many variables. So we have to address the issue of overfitting and that's often addressed by using so-called regularization methods.

Finally, we're going to talk about performance assessment and that's the topic with which we're going to conclude this session. And performance assessment involves ways of selecting different model types between different algorithms. We want to compare the performance of different models and algorithms and this is done by testing which is something that's done after we train our linear regression models. And it involves methods like cross-validation and the pretty intelligent or clever method that's called the Bootstrap, which is going to come at the end.

- variance σ_i^2 of W_i changes with i (perhaps depends on $\mathbf{X}_i$)



So let's discuss what can go wrong. On the mathematical side, there are certain assumptions that are being used. For example, That the noise is associated with different individuals or different data records. All are generated with the same variance but what is that variance is actually different for different individuals.

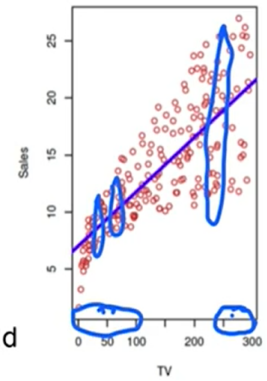
One can see evidence of this, for example, when looking at these data, it appears that individuals or markets in which TV advertisement is low. For those, there's only a little bit of variability in the sales. Whereas in other markets where there is a lot of advertisement, the variability in the sales is much bigger. So this plot here certainly indicates that the variance is not the same for every data point, but depending on the X that we have, we might have different variances. If that's the case, how can we mitigate it? Well, if you know the variances, then we could take that into account by using a weighted least-squares criterion.

What is a weighted least squares criterion?

when optimizing the sum of the squares of the prediction errors, give less weight to those data points that are noisier. So the data points that have a lot of variances that are noisier are less reliable and we should be giving less weight to that. Unfortunately, however, in real life, we rarely know the correct variances, so that doesn't quite apply. The trouble when these variances are not equal is that the assumptions we have made in all of our analyses so far start to fail. So various formulas for standard errors do not hold. So we cannot really trust the confidence intervals or the hypothesis tests that we're doing.

There are methods in the literature for dealing with heteroscedasticity and these range from the simplest one, which is to put down some scatter plots, even like the one that we have hereby eyeballing figuring out sources of travel. And then there are various methods for doing corrections. But this is a big topic, we will not go into it in any more depth, instead, we are going to move to the next thing that can go wrong.

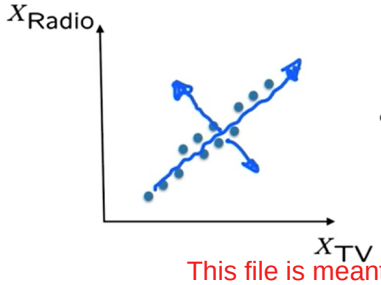
- variance σ_i^2 of W_i changes with i (perhaps depends on $\mathbf{X}_i$)
  - if we knew the σ_i^2, could use a weighted least squares criterion
- formulas for standard errors, etc., do not hold



Suppose that the X vector is associated with one individual. Suppose that when you look at the X vectors of the different individuals, they do not have enough variety, but they tend to be confined in a lower-dimensional set. So if the data were like this, it would mean that TV advertisements and radio advertisements are very closely aligned. If that's the case, then there is nothing in the data that tells us the relative effects of television advertising vs radio advertisement. We might observe that sales keep increasing along with that direction.

can we attribute that to higher TV advertisements or do we attribute it to higher radio advertisements?

- vectors X (approximately) confined to a lower-dimensional set



- no way to tell relative effect of TV versus Radio

There is basically no way to tell. There is no information available that tells us how sales would vary in this direction; we simply

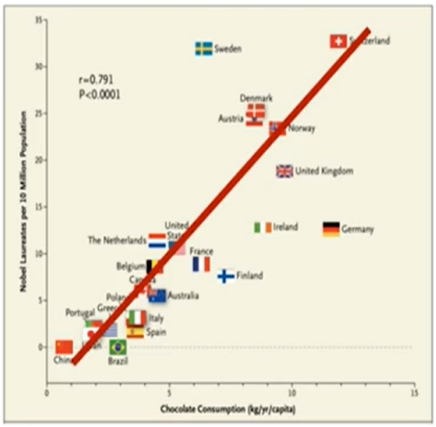
do not have the data for that. In some sense when we have this situation of multicollinearity or linear dependence between the different axes.

In those cases, there's not a lot that you can do. In fact, you're going to see mathematical difficulties. The standard errors will become infinity or huge in certain directions, just reflecting the fact that there is no way to obtain good accuracy in certain directions for which we do not have the right data. What can you do in that case? Well, you just realize that when that happens, some of the variables are redundant. If X_{TV} is always equal to X_{Radio} . Then there's no point in including X_{Radio} in our X Vector. It's redundant. It doesn't give any more information. So the usual remedy is to identify redundant variables and reduce them and remove them from the model.

So, both this difficulty and the previous one Heteroskedasticity that we discussed are more or less of a mathematical type.

There are some more tricky structural issues that can show up, however. here we're going to talk about one example of that and then we'll continue with more examples of this type. And these examples shown are going to highlight that modeling is a very different business from prediction.

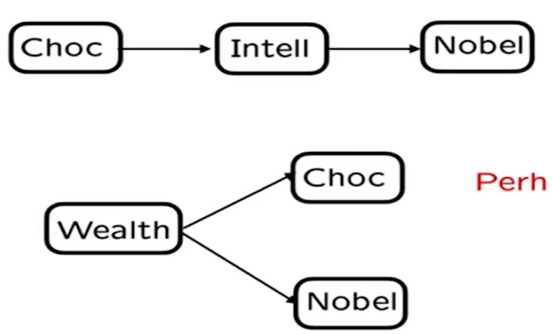
So here's a famous study. It was actually published in the New England Journal of Medicine, a very prestigious journal. What it shows is a data set on the horizontal axis where we have chocolate consumption in different countries, measured in kilograms per year per person. On the vertical axis, we have the number of Nobel Prizes won, again, per year, per person. So it's proportional to the population.



So the different data points are the different countries. For each country, we have the chocolate consumption and the number of Nobel prizes and we see here clearly that there is a positive trend.

So that's news. It got into the popular press, there is a connection between chocolate consumption and Nobel prizes. In fact, if you do the regression, you get a very nice upward moving line and this turns out to be statistically significant that the slope of the light is non-zero and the paper even reports a p-value, which is tiny. So seeing such data just by chance is an extremely low probability. So something is happening.

Nothing wrong with this so far. This is a legitimate straight line that matches the data in the sense that minimizes the sum of the squared errors and that could be useful for predictions. If I tell you that there is another country that has this level of chocolate consumption, you can use that slide to make a prediction that I think that this country is going to have about so many Nobel Prizes per person.



That's quite legitimate. What is not legitimate is to make the extra step to say that there's a causal relation, one might consider a causal relation of this form. Maybe chocolate consumption improves intelligence and then intelligence causes more Nobel prizes. This is a theory that's consistent with the data that we have. But it's not true from the data. That data could have an alternative explanation and an alternative explanation could be that some countries are wealthier than others. Those countries that are wealthier, have more chocolate consumption, and also because they invest more in research and education, they also have more Nobel prizes. The data that we have are consistent with both of these theories or models, and there's no way to decide from these data alone which one of the two theories is correct. So, while the data are useful for making predictions. They do not convey information for drawing conclusions about the causality that's present in this phenomenon, or conclusions about the structure of a true model of the world.

To summarize, data can be used in many cases for making fairly accurate predictions. However, the predictive model that we have is not necessarily a causal model of the phenomenon that we are observing.