Linear Regression with Nonlinear Features

We can summarize our takeaways from the previous lectures that when there are latent variables i.e. hidden variables, we can still run regressions and get predictors that give good predictions. But on the other hand, these predictors do not necessarily discover the correct structural relations for the model that we are investigating, and more importantly these predictors can only predict but cannot answer questions of the what-if kind.

What's the difference between these two things?

Well, suppose that the marketing department has a way of operating and it keeps operating the same way, and there's a new market and the marketing department will set the budget, as usual, using the same rules that it used in other towns. Then if you tell me the X for the new market, I will be able to make a good prediction of the Y in that market.

On the other hand, if you ask what if in this new market I operate in a different way, what's going to be the effect on sales? There is no way to answer that question, using prediction when we have those hidden variables and when we do not have those causal relations. On the mathematical side with latent variables, there's quite commonly the case that some basic assumptions are violated. So the mathematical formulas for the standard errors do not hold, so you cannot trust them and you cannot use them for confidence intervals and hypothesis tests.

What can we do, if we have these difficulties?

Well, if the source of the difficulties is that we have omitted certain variables. So let us try to use more variables in our model. So we have this predictor that we learned when we ran our regression and we got coefficients where we tried to predict sales based on TV, radio, newspaper, and advertisement.

As in the example we discussed earlier, the market size is maybe important and maybe it's an important determinant of savings. If that's the case, we should just include the market size as one of the other possible explanatory variables in our model and try to estimate that coefficient **theta 4** as well.

$$\widehat{\text{Sales}} = \theta_0 + \theta_1 \cdot (\text{TV}) + \theta_2 \cdot (\text{Radio}) + \theta_3 \cdot (\text{NewsP})$$

$$+\theta_4 Z$$

Z = market size

What does this amount to?

In our dataset, one column is the TV advertisement and another column is the radio advertisement. Just add another column in which you have the Z, the market size for each particular town. So you're increasing the size of your dataset, but you can still run the regression as before.

$$\widehat{\text{Sales}} = \theta_0 + \theta_1 \cdot (\text{TV}) + \theta_2 \cdot (\text{Radio}) + \theta_3 \cdot (\text{NewsP})$$

$$+\theta_4 Z \quad +\theta_5 U \quad +\theta_6 V$$

Z = market size

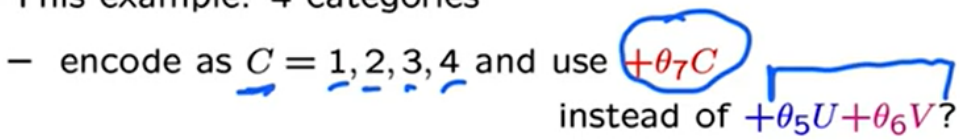
$V = 0$: rural
$V = 1$: urban

$U = 0$: market with competitors
$U = 1$: market without competitors

Suppose now to connect with the other story that we made, that there are certain good markets in which there are no competitors, and there are certain bad markets **U** equal to zero in which we do have competitors. We may hypothesize that the type of the market has an effect and let us try to capture this effect by throwing another term in our regression and try to estimate **theta 5** as well. So we create another column in which we put the **U** variable for each one of the different markets, presumably we know whether we have competitors or not. So we mark that in our dataset and try to learn a coefficient for the effect of that particular variable. Also, we notice that here **U** is a binary variable. There is no difficulty with doing regressions when variables only take a small set of values. We can then think of other possible variables, maybe it makes a difference whether it's a rural or urban area.

If the product you are selling is only useful in rural areas, then you expect to have more sales in rural areas and less sales in urban areas independent of how much you're advertising. So to capture that effect, we could add another term in this regression. So, keep including more and more variables, whatever variables you think might be useful.

So these are, **U** and **V** are so-called categorical variables, and they're very commonly used in regression models. Nothing special about them. However, there's something to notice. In this example, we have two choices for **U** and two choices for **V**. So the possible combinations are four different categories. You might think that you could model such a situation by using a single variable to encode the category as 1, 2, 3, and 4 and then try to use just the single linear term, so that you only have to learn one coefficient, as opposed to having to learn two different coefficients **theta 5** and **theta 6**.



However, this type of encoding is not right. It's not the thing to do, because by encoding them that way, we're assuming that there is a sort of steady progression from one to four and according to the coefficient **theta 7**, the effects of **C** going from 1 to 4 would move gradually. However, the four categories we have here, they're not naturally ordered in any way. So we should not impose an artificial order on them. We should keep the four different categorical variables separate, a separate variable **U**, and the separate variable **V** and it is an important detail of how to use categorical variables.

So in this example, we have the story that maybe market size and maybe the presence of competitors is something that's useful for making predictions. So we include them, but maybe there are other types of variables that we might include.

We have the marketing example and we saw that when we just predict sales based on newspapers, the results appear to be positive in the sense that newspapers seem to have an effect on sales, on the other hand, once we included TV and radio advertisement and run the regression once more, then we find that newspaper is actually not significant. The value of this coefficient is close to zero within the standard error.

Marketing example:

$$\widehat{\text{Sales}} = 12.35 + 0.055 \cdot (\text{NewsP}) \qquad \text{“significant”}$$

standard error $= 0.01$

$$\widehat{\text{Sales}} = 2.94 + 0.046 \cdot (\text{TV}) + 0.19 \cdot (\text{Radio}) - 0.001 \cdot (\text{NewsP})$$

standard error $= 0.006$

supports the null hypothesis that $\theta_{\text{NewsP}} = 0$

once we take into account TV and Radio:
NewsP is “inconsequential”

So while in this model newspapers seem to be significant, once we include more variables newspaper stops being significant. So if we go through that exercise, we start with the newspaper but then we add the other variables, once we add the other variables, we would remove the newspaper variable that we have in the model because it appears to be insignificant. So the data supports the null hypothesis that newspaper has no effect and so we remove it from the model.

It gives us a way, at least for removing variables. Now, you can think of a combination, you keep adding variables, but then as you add variables, some of them end up having very small coefficients theta, and then you can remove some of these variables, and we can keep going with this exercise.

There are even more ways of creating new variables and here is the trick. You can create new variables by just working with the variables that you already have. Take the original variables, X1 and X2, and create some new variables. Let's say the **logarithm of X2**, or maybe the **product of X1 and X2**, and think of these as new variables. Again, what that corresponds to is that you have your dataset with data and there's a column for X1, a column for X2, you create another column for the log of X2, and you create another column that has its entries the product of X1 and X2. These are new explanatory variables and they are created in nonlinear ways from the original explanatory variables. Then we can run a regression and try to predict Y from these new variables.

$$\hat{Y} = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 \log X_2 + \theta_4 X_1 X_2$$

When would that make sense?

If we have some reason to believe from the physics of the model, that maybe the log of X2 is what determines Y as opposed to X2 itself, then we might want to include a term of this kind. The interesting thing is that even this problem can be thought of as a linear regression problem.

What we have done is that we took a typical data record, which is a bunch of X's and we added some more components to it and we just created a bigger, longer, so-called augmented data vector. And then the predictor of Y is a linear function of the augmented vector, linear combination with the linear coefficients being the thetas, and so the predictor is also linear in theta. So we still have an ordinary regression problem, except that our data records are longer, and we have more thetas. This idea in general would correspond to predictors of this kind.

For a typical data record, we construct a bunch of features or possibly interesting non-linear transformations of our data records, and we predict Y's by forming a linear combination of these features. The way to determine the thetas in this model is again, the same least-squares minimization formula. That is, we look into predictors that are linear combinations of the features.

$$\hat{Y} = \sum_{j=1}^{k} \theta_j \phi_j(\mathbf{X}) = \theta^T \phi(\mathbf{X})$$

features

feature vector

$$\min_{\theta} \sum_{i=1}^{n} \left(Y_i - \theta^T \phi(\mathbf{X}_i) \right)^2$$

This is our predictor, a linear combination of features. We look at the squared error between the actual and predicted value. We look at the squared errors or the residuals

for all the points in our data sets and we try to find the parameter theta that minimizes this sum of squares. This is no different from the regression problems we have been studying before. The expression we're minimizing is still a quadratic in theta. So this problem can be solved very efficiently using the very same regression software that we have been using all along.