Generalised additive models

## Transcript

Video 3: <https://youtu.be/CoxTTfUGb24>

Hello and welcome to video number 3 in the generalised additive models online material presented by NCRM. In this video I’m going to continue to look at fitting generalised additive models within R, where we’re going to use diagnostic checks to ensure that the models that we have fit are valid and robust and working towards improving our models to ensure that we can make interpretations about our data that are valid.

 So, the model that we had fit from video 1 involved looking at the relationship between the number of bikes rented in Seoul, South Korea, and the temperature for each day. We also included an interaction term for rain as it looked as though the relationship between temperature and bike rental differed between days where there was rain and days where there weren’t. So, we had gotten up to fitting our first model and expanding on this by including an interaction term. And when we plotted the predictions of this, we found that something strange was happening, so particularly the smooth function related to temperature and bike rental on days where it did rain, we had this really, really high estimate for very low temperatures, which suggests that our model is not doing an appropriate job of capturing our data. Now, as with all statistical models, GAMs have a certain number of assumptions that must be checked and must be true for these models to be valid and appropriate. And a lot of these model assumptions are shared with the linear equivalent of that underlying distribution, so for instance in this model we fitted, we haven’t actually specified the underlying distribution family that we’re assuming the residuals or error terms are following. In fact, we’re relying on the GAM function’s default setting, which is normal distribution.

 So, if you’re familiar with linear regression models or a linear model then you’ll be familiar with a lot of these assumptions already which are shared between those models and the GAM equivalent, obviously without the assumption of linearity. Those assumptions are that the residuals or error terms are normally distributed and centred around zero, that these residual terms have a constant variant when plotted against the linear predictors, and that there are no dependencies between the independent variables included in the model

 So, to check these assumptions are valid, we can use the quick GAM dot check function, which produces some useful model diagnostics and plots in order for us to check assumptions of GAMs. I simply put my GAM model that I wish to check, which was this rain model here, into the function and it will produce four diagnostic plots. The first one is a QQ plot, so this is comparing the theoretical quantiles if the underlying assumed distribution were true against the observed residuals from my model. If the underlying distributional assumptions are valid, we would see the residuals lying along this straight line of equivalence, shown here in red, but what we see from this model is a little bit of deviance away from that line, particularly for very small values.

 The next plot shows the residuals plotted against the linear predictor or the intercept of our model. If the model assumptions were valid we would see a constant variance across these residuals, in other words, no funnel shapes, just random scatter around zero. But what we do see is an increase variance and then a decrease, so we do not have this constant variance. This is what’s known as heteroscedasticity and this is showing as that our rather strict assumptions of normality may not be valid for this particular data.

 We have a histogram of the residuals at which we can check again the distributional assumption. In this case we’re checking whether this follows a normal distribution centred around zero, which approximately does with slightly longer tail on the negative side, but we do also have these huge residuals indicating the predictive power of our model is potentially not very good.

 And finally, we have a simple plot showing the predicted outcome, the response of the fitted values based on our model and the actual observed outcome from the dataset and we’re looking to see whether it’s doing a good job and can actually see that we’re getting negative values here for a start off for the fitted values, and we have particularly this overestimation values which is likely coming from those small numbers of rentals on the rainy days, which are still not completely picked up by that interaction term.

 So, all of these are pointing to the fact that maybe our model could be improved, and there’s a couple of ways in which we could do this. I mentioned that the strict assumption of normality, the underlying default in GAM, may not be appropriate, so to change that distributional assumption I use the argument family equals in the GAM function, and there are a number of families that are available to me, the list of which can be found in the help files. So, all of the families available in GLM, so these are all your most common types of regression models, including binomial, gaussian, which is a linear model, gamma, Poisson, as well as some additional values specifically for GAM, such as the Tweedie and negative binomial distributions as well as some others, they are all options for us within this. So, we need a family that’s appropriate for a positive continuous variable but is maybe less restrictive than that normal distribution, so something like the gamma distribution may be appropriate in this case.

 So, if I refit this model, I check the distribution, and that seems to have fixed the strange pattern we were getting from the yes rain days. If we check our model, we still do have these skewed residuals, residuals with non-constant variants and the QQ plot is not exactly showing us what we’d like. We’re also massively overestimating these particular values, but what we are getting is solely positive predictions, which is fixing at least one of those problems.

 So, potentially we’re missing an important variable that might be useful for explaining those strange patterns that maybe isn’t captured and the model isn’t doing a very good predictive job. So, other estimates, other variables that we include within this dataset, things like humidity, and that might play quite an important role, particularly somewhere like South Korea, which experiences quite high levels of humidity in summer sometimes, potentially those days where people were less likely to ride their bike, maybe it was just because it was very, very humid. So, we could again plot our data, I’m going to copy this plot up here, looking at a scatter plot, but rather than colouring the points by rain, I could colour them by humidity percent and then use a gradient, where if it was a high humidity I could set the values to blue, indicating it’s wet, and if it was low humidity, so it was quite dry, maybe I could keep it at that brown to make sense to my brain. I’m going to remove this GAM smooth because I’m just looking at the overall relationship. And that does appear to have quite a strong trend here. These values were that it was the optimal temperature range but still had very low values of bike rental, very high humidity, whereas we don’t get the as high humidity at these very high points, where it did rain, but maybe it wasn’t as unpleasant.

 So, we could take our model that we have up here and we can extend it, so I’m going to give it a different name, and I’m now going to add on a smooth function for humidity. I’m going to keep it as the gamma family, and then I’m going to check this model using the GAM check function. And that has definitely improved the predictive ability of this model. We still have a not exactly perfect prediction for some of them, but it’s very much closer to this line of equivalence. Our residuals are fare less skewed than before, still a little bit skewed but nothing as bad as it was before. It certainly dealt with a lot of the differences in heteroscedasticity, we still have a smaller variance at higher values, but this is also kind of coupled with the fact we have much fewer points, so less of an issue. The QQ plot does still show a bit of an issue here so we may need to change that. But if I plot the predictions, changing the model that I’m generating them from, this is going to show me the expected relationship between temperature and bikes separated by days where it rained and where it didn’t but still accounting for this humidity, so it’s taking out the effect of humidity there, so it’s giving me more of a view of what’s happening solely with the temperature as opposed to those other important factors.

 So, model checks carried out using GAM dot check, model family can be changed using the family argument in GAM. Now, another thing we might want to check is how well these smooth functions are doing. We might want to adapt how they are dealing with the data. And one main issue we want to check is that these are allowed to be wiggly enough, so remember from video 1, we don’t actually specify the number of knots that’s calculated as part of the model fit, but we have to set the maximum value, and if we don’t then R is going to choose the maximum value and the default setting is usually 10. The GAM check function actually gives me some information about this, so this K dash here is the maximum number which is the number specified by the GAM function minus 1 degree of freedom which is taken out for the intercept, so all of them were allowed to have a value of 9. The EDF, effective degrees of freedom, is the number that it actually was estimated to me. And then this P value is related to whether or not this K value has been set high enough, so if the P value was very, very low, that’s indicating that K should be increased. Also, if this EDF is very, very close to the maximum K, that’s also an indication that potentially it’s not set high enough.

 Now, all of these appear to be okay, but if I did see that the maximum should be increased, I can do that within this S smooth function by setting the argument K equals and then setting the maximum value. So, if I change this to K equals 15, you can see that my K dash has changed for the humidity smooth function and actually the EDF has increased even though it was still below the maximum number. So, again, what we want to do here is check that this K is set high enough so that it’s not restricted too far, so I could increase it again, but bearing in mind that every time I increase this it’s taking longer for the model to fit, it’s computationally more costly. Okay? So, if, for instance, I set K equals 20 here, it’s actually 40 because I have two separate smooth functions being fitted based on this interaction, so it actually takes longer. And it’s very hard to see here, because it’s still pretty fast, but imagine if you set that to a much higher value, it slows things down a lot. And it’s not actually really changed anything about my model because K was originally set high enough.

 The other thing I can change within these smooth functions is the basis, so by default GAM, the MGCV package, uses what’s known as a thin plate regression spline. This is a very, very flexible approach, so it works across multiple dimensions as well as a single dimension, and the reason it’s the default is it is a very computationally effective approach, so it’s very, very fast even with a large number of values, a larger number of Ks. If I wanted to change this though, particularly if I’m doing it across multiple dimensions where they’re on different scales, so, for instance, time and space and measured in different ways, then I would need to use a different basis which is given by the BS argument of the S function. And if I wanted to see the options, I can go to the help file and it tells me that, for instance, TP would be thing plate, CR is cubic regression, and if we go to the smooth dot terms help file, it’s going to give me a range of the options available to me. So, if you’re fitting a much more kind of complex approach, this is maybe something that you would have to change as well.

 But the model fitting process, as I said, what we’re trying to do initially is to answer whatever research question, whatever important findings and insights I’m trying to gain from my data, fitting a model that answers those questions whilst also ensuring that it is a valid approach, that any assumptions that are made by that GAM model has been met, and if not, we need to take action in order to improve the model and to make sure that the model chose is valid and represents the data.

 Thank you for listening and for going through this material. I hope it’s been useful and I hope you go ahead and use generalised additive models in your work. Thank you.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK