# Introduction to Bayesian Regression

## Video 2 Transcript

Full resource: <https://www.ncrm.ac.uk/resources/online/all/?id=20843>

Oliver Perra: Hello. I’m Oliver Perra and this is the second part of my introduction to Bayesian regression. In the first presentation I had provided a broad overview of the Bayesian approach. In this second presentation I will apply that to the case of regression analysis. And before that, I just wanted to provide a quick reminder of what we mean by linear regression. Linear regression are models that basically try to learn about the mean variants of measurement variable using an additive combination of other measurements so the basic formula is in fact this one I presented here, where the value of an outcome variable Y for each individual I will be equal to the sum of parameters A and the product of parameters B and by a predictor X. And there is also an element E that represents individual variation from the scores, from the expected scores. Parameter A represents the value that outcome Y takes when predictor X is equal to zero, so it is called the intercept. Parameter B represents the change in outcome Y associated with a one-unit increase in predictor X, and it is also called the slope. The parameter E is supposed to be normally distributed with mean zero and a variance that we can estimate.

 So, one important point is that the linear regression uses normal distributions to describe uncertainty about measurements, and most of you are probably familiar with a normal distribution. It’s a bell-shaped symmetrical distribution, and it’s basically defined by two parameters, the mean indicated by the Greek letter mu as you can see here at the bottom of the slide, and the standard deviation indicated by the Greek letter lower case sigma, as you can see here again in the bottom of the slide. In this graph in the slide I also represented different examples of normal distributions that all have the same mean mu equals zero but with different standard deviations, so just a quick reminder how normal distribution changes in its aspect depending on these parameters, mu and sigma.

 So, I have provided an example of how conventionally regression analysis is formalised, but in running linear regression using Bayesian approaches, the regression model can be described in a slightly different way, as you can see here. This may be bizarre and probably it’s a bit more cryptic to many of you, however there are many advantages in using this alternative notation and description. One is that, for example, many of the assumptions of linear regression, such as (inaudible 0:03:36) and others, are more explicitly read from a similar notation. This notation also provides a more flexible way of describing a linear regression model, which also allows to specify different assumptions in a more straightforward way. And I will provide an example that should make this notation clearer and also show the advantages of using this notation.

 So, for this example I have used some fictional data that I have created, and you can find them with the material for this module together with the script I have used to provide these examples, so look online for the script and the data. So, this fictional dataset represents newborn’s birth weight in grams by maternal weight in kilograms. Studies have indicated that there is an association between maternal weight and a newborn’s birth weight, and this is the regression model that I will develop in the example. So, to start the analysis in Bayesian approach, I need a prior, a model that represents the plausibility of parameters before I see the data, before I collect any data. And before I see the data I can safely assume that the distribution of a newborn’s birth weight is going to be approximately normal, with a mean mu and a standard deviation sigma. So, in the green circle here you can see the notation I have used to say that the birth weight of infants is going to be approximately normally distributed with a mean mu and a standard deviation sigma. And we now move to other parameters in the model.

 So, the purpose of Bayesian analysis is to estimate a posterior distribution that represents the plausibility of different combinations of parameters, conditional on the data I have collected, so learning from the data I have collected, and the model that I had developed before seeing the data. So, here I move to the second line, the second equation circled in orange. In linear regression I can assume that the mean mu of the distribution of birth weight will be equal to the sum of an intercept A and the product and the sum of the product of the slope B by the predictor maternal weight. So, here the slope, B, represents the change in newborn’s birth weight associated with 1kg increase in maternal weight. And it makes it easier to standardise or centre predictors in linear regression and, in this case, I centred maternal weight so that maternal weight is expressed as a deviation from the average maternal weight and the average maternal weight in this sample was 72.01, which is actually the average maternal weight in kilograms according to some statistics. So, here the intercept then represents the expected birth weight of newborns when mother has an average birth weight of about 72kg. Note that by linking the mean mu to this equation where mu is represented as a function of the intercept and the slope, I am making the value of mu dependent on the other parameters A and B. So, A and B are now measurable properties that are uncertain parameters in a Bayesian approach, and I will need some prior assumptions on these to make the Bayesian analysis start.

[0:08:07]

 In most cases the parameters in a model are specified independently. So, here, in yellow, I circled the prior for the standard deviation of birth weight. Here the notation basically means that the standard deviation sigma is expected to be approximately distributed following a uniform distribution that ranges from 0 to 100 grams. So, this just means that I am considering any standard deviation of newborn’s birth weight from 0 to 1kg, 100g, equally probable. And why did I choose 1kg as the upper bound in this uniform distribution of probability? Well, if birth weight is normally distributed then I know that 95% of the values will be between minus 2 standard deviations plus 2 standard deviations. And therefore I thought that it is sensible to assume that 90% of birth weights will be between 2kg above or below the population average. If the average birth weight of children is 4kg, 95% of babies would have birth weight between 2 and 6kg, so I thought that was sensible enough to assume.

 Now, what other prior assumptions should I have for the other parameters in the model? And let’s consider first a model with no predictors. So, assuming that the maternal birth weight is average, parameter A then highlighted in the red circle represents the distribution of birth weight for mothers of average birth weight. So, what priors should I give to this? We know that normal birth weight varies between 2500g and 4000g, so let’s say that I might expect that the average for children of mother’s with average weight may be, say, 3,300g, so 3.3kg. So, this is the value I put here for the normal distribution of the intercept circled in red, and the standard deviation, say, may be 1,500g, so 1.5kg. And again, since I am assuming a normal distribution, I know that 95% of cases fall between 2 standard deviations below or above the mean, so if I take these values as good, I would expect that the average birth weight has 95% probability of being between 300g and 6300g. And note that I am talking about the mean, and I know the average birth weight can hardly be 300g in any typical population, so this seems to be a very broad prior and probably an unrealistic one, and in fact it is always a good idea to plot the priors to have an idea of what we are assuming, what are our assumptions before we run the analysis.

 In the script I have attached with the course material you can see how I have created this plot that represents the prior specified here for the intercept. And if you look at this probability distribution in the graph, you can see this is not a very good prior. The prior assumes as possible that some average birth weights will be zero or even negative. You can see in the horizontal axis the birth weights of children and the horizontal vertical line here represents zero, so you can see that some, however unlikely, but still this prior is assuming as possible to observe some negative values, which is obviously not a very sensible assumption. And should we care? After all, the model will learn from the data and will update the posterior considering the data, so if we have numerous data, not having a sensible prior will not matter that much. But this is not always the case, so it always makes sense to construct some sensible priors. And many of us highlight that as long as the priors are specified with seed(?) data, there is nothing wrong in using knowledge and substantive knowledge about the issues to provide some priors that are plausible and sensible.

 So, here I changed the prior for the intercept, where I keep the mean to be expected to be 3300g, but the standard deviation is 600g. So, this means that before seeing the data I am assuming that there is a 95% possibility that the average birth weight of babies of mothers of average weight will range between 1kg and 200g more or less the average. And this is the representation in the graph. You can see the representation of this prior, where the expectations are more sensible. So, you can check the script I have provided with the material of this course to see how I have created this graph. So, I’ll keep this prior for the intercept and I’ll move to the slope.

[0:15:06]

 So, here, circled in blue, are the parameters for the slope. And the slope indicates the increase in newborn’s birth weight for a one-unit increase in maternal weight compared to the average maternal weight. So, I may start by assuming that this uncertain parameter is normally distributed and has a mean zero and a standard deviation of 50. Because I am assuming a normal distribution, saying that the mean of this distribution for the slope is zero means that I am assuming there is as much probability of the parameter being below zero as above zero, and if the slope is zero there is no association between child’s birth weight and maternal weight.

 Why choosing this? It does make sense to choose the least informative distribution that is consistent with our knowledge (inaudible 0:16:19). So, in this case you can see I am opting for a sceptical prior because I am not assuming that the association between the predictor and the outcome has a specific sign and I am also considering as plausible that might even be zero, so there is no association between the predictor and the outcome. And the rationale for choosing similar broad and sceptical priors is that since the posterior distribution on which we base our conclusions is estimated by learning from the likelihood of the data, it does make sense to have priors to keep the models learning from the data in check. So, in a sense, we want priors that regularise this process. That just means sceptical priors avoid overfitting while still allowing the model to learn from the regularities in the sample data in the data we observe. And for more discussion about this I refer to the book Statistical Rethinking by McElrath, and particularly chapter 7 where there is more discussion about this. The point really is that to have some sensible priors but priors that, in a way, are more sceptical and keep the learning from the data in check so the posterior distributions are not too skewed or are not overfitting new data.

 It’s also important that similar priors are checked in sensitivity analysis, and some of the references, particularly the reference from Krushke(?) I included, emphasise the importance of some sensitivity analysis when using broad priors or sceptical priors like that, analyses are then run, changing these priors to other broad priors that are possible and check if different priors still produce the same pattern of results. So, it’s important that because there are many different priors we can assume we also check the sensitivity of the posterior distributions to different types of priors. And I refer to the different references I have put with the material of this course.

 So, I assume here that the mean of the slope is zero and the standard deviation is 50, which means that 95% of the time the effect of maternal weight will vary, reducing babies’ weight by 100g or adding 100g for each kilo of maternal weight over the average maternal weight. This is a large effect and you can see this by looking at the graph I report here, where I am plotting the slopes implied by this prior assumption. And again, you can see the script I have used to create this graph, but you can see that this is really creating very implausible slopes, where some maternal weights associated with incredibly high or incredibly low newborn’s weights and that even fall below zero. And the other horizontal line here represents 2500kg, which is the threshold. So, you can see that the slopes implied by this prior assumption about the slope are incredibly implausible and describing incredibly strong relationships that fall aside of the possible range of newborns’ birth weights. So, change in the standard deviation of the slope to 25g here provides assumptions that are more plausible, and here you can see them represented and again you can check the script I have used. So, this seems like a plausible assumption for the slope.

 So, now I have a model ready to analyse all the data I am going to collect, so I have a likelihood function, a function that represents how my data of observed birth weights of newborns may have come about. The likelihood is saying that the underlying distribution that may have generated the birth weights I’m going to observe is normal with mean, it’s normally approximately normally distributed with mean mu and standard deviation sigma. I also have a linear model that says that the average of birth weights is a linear function of additive association with an intercept and a slope that represents the rate of change for different units of maternal weight, changes in maternal weight from the average maternal weight.

[0:22:36]

 And then I have prior assumptions about the distribution of these different parameters, where I am assuming that the slope is normally distributed around an average of 3kg and 300g, and the effect of the slope is normally distributed about around an average of zero, so it may be positive, negative or null, and then I am assuming that the standard deviation of newborns’ birth weight may take any value between zero and 1kg. So, I will use these different elements, priors and likelihood and my linear model to update my assumptions about the distributions of these parameters and rank the plausibility of different combinations of these parameters based conditionally on the data I’m going to observe and conditionally on the model I have just described here. So, I will do this in the next presentation.

 So, thank you very much for your attention and please remember to check the webpage of the National Centre for Research Methods for more material and other resources. 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