# Introduction to Bayesian Regression

## Video 1 Transcript

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

Oliver Perra: Hello, I am Oliver Perra and I’m going to give a series of presentations to introduce Bayesian regression. This is the first of a series of three presentations, and together with these presentations there is other material that you can access, including exercises. So, in this first presentation I will introduce the principle of Bayesian regression, firstly by comparing the Bayesian approach with the more usually used frequentist approach, and in the second part of this presentation I will provide more details about the principles of Bayesian analysis.

So, the Bayesian approach basically provides a different approach to hypothesis testing, different from the traditional approach which is often called frequentist. And particularly some of the advantages of a Bayesian approach, so some of the reasons why I’ve been particularly interested in using a Bayesian approach to some analysis is that it allows to include previously knowledge that we have on the phenomena we are interested in, so it allows to capitalise on previous research and previous meta-analysis and so on. And it also allows to incorporate every piece of information, however small, and use that to update our knowledge in light of the new evidence we have provided. And at the end of this short presentation you’ll have a clearer idea of what I mean by this.

So, to appreciate what a Bayesian approach consists of, it is useful to compare it to the conventional methods that are usually learnt in undergraduate courses, the so-called frequentist approach. In both approaches, we are dealing with parameters. Parameters are measurable characteristics of a population, for example, the mean IQ or values in a country. The conventional approach to hypothesis testing assumes that parameters are unknown but are fixed, and it’s for this reason that the conventional methods take a counterfactual approach. Say, for example, if we experimentally tested a training programme and this programme is supposed to make our participants smarter, we can then test the difference in IQ of the trained participants against the average IQ of participants that haven’t been tested. And we will start with a counterfactual scenario where we assume that the difference between this means the average IQ is exactly zero, or as the two groups are supposed to have exactly the same average IQ, so this is the known hypothesis scenario. And based on this counterfactual scenario, the statistical methods that are conventionally used estimate the probability that our observed difference in IQ between the two groups occurred from the null hypothesis. And this scenario is estimated assuming that we can replicate the experiment several times.

So, if the difference in IQ between the two groups that we observe is unlikely to have been generated from the null hypothesis scenario, we can reject the null hypothesis that there is no difference and we can say that there is a difference in the average IQ of the two groups. So, one problem with this approach is that the P value is defined in a scenario where we assume we can repeat the study several times or else the process that generates the data is supposed to be repeatable. But this is not always the case. For example, in meta-analysis, the collection of studies should be considered as a one-off. There is also a problem in interpreting the P value and the results. Assume that our control group had IQ of 100, if the P value of our test of mean differences was 0.01, lay people tend to assume that there is only 1% probability that the mean of our trained participants is also 100. This is incorrect though because the P value represents the probability that we will fail to reject a null hypothesis of no difference between trained and control if, in reality, there was no difference between the average IQ of the two groups. So, the results really do not inform about the probability of the parameter of the interests, for example, what is the probability that our trained participants have average IQ equal to 100? And this is the type of answer that some people may want from the analysis and the conventional approach doesn’t easily provide these type of answers.

Another problem is that without just samples, an increased estimate precision(?) becomes more likely we can reject the null hypothesis, so it’s possible in large studies that we can have significant effects that are relatively unimportant or inconsequential. And again, this problem arises because we cannot easily provide the probability of a parameter assuming certain values in the conventional approach. A further problem is that we can assume that the data from the hypothetical repetition of the study will follow non-distributions if our samples are relatively large. With small samples, the underlying sampling distributions may be unknown and we may not be able to use information collected unless we make some great leaps of faith.

[0:06:30]

So, the Bayesian approach provides a different way of conceptualising parameters, and I will show it in the next slide. So, in the Bayesian approach, the parameters are not considered to be fixed, but instead they are variables, measurable population characteristics that are uncertain. So, if they are uncertain, parameters can be described by a probability distribution, in other words we can estimate the probability that a parameter like a population mean assumes a certain value or else lies between a specific range of values and so on, so we can answer questions about the probability of parameters, characteristics of the population assuming different values and judge how likely or unlikely are some scenarios, different scenarios that we might be interested in, for example, IQ scores being between a certain range rather than another range.

So, how can we do this? The analysis basically starts from a-priori assumptions on the distribution and values of parameters, some prior assumptions on what the data could look like. These assumptions may be based on previous knowledge, previous research meta-analysis and so on, and they may be broad. For example, if I did a study where I trained some people and expected them to become more intelligent, a broad assumption could say that I expect the trained people to have an average IQ between 70 and 230, say. And assume also that any value between those two extreme values is equally likely, is equally probable. The a-priori distribution then, it’s updated considering the new data that I have collected. With new information collected by a study, I can update the probability distribution of the parameter giving more credibility to some values rather than others, giving more credibility to some values conditionally on the evidence I have collected. So, for example, the information that I have collected, make sure that it’s very unlikely that my participants’ average IQ is 70, so I can update the probability distribution of the parameter value. And this updated distribution is called the posterior because it is estimated after seeing the data. It’s the probability distribution of the parameters that is updated considering conditionally on the data I have collected.

Once I have an estimated posterior distribution of the plausible values of the average IQ of my participants, I can sample from this probability distribution the probability distribution of the parameter to describe means and medians of the probability distribution, for example, what is the most plausible value that the parameter may take, or also to describe the range of parameter values that are more plausible, more likely based on the estimated posterior distribution.

So, Bayesian analysis is basically a formal way to update models of the world by using evidence. And Krushke(?), in one of the papers I have referenced with the material for this module, makes an example where if there are two candidates in an election, we can ask our friends who they intend to vote, and based on these answers we can form some expectations. But then we may read some polls and based on these results we update these expectations. And particularly if the polls we’re reading have a large random sample, the data may help us shift our prior beliefs to some more defined, more precise beliefs about who’s going to win the election. Bayesian analysis basically provides a formal way to define how to update our beliefs and how we should use data to update our beliefs. And I will now provide a more formal example of this process with the next slides.

So, the first element in a Bayesian analysis is a prior. A prior is an a-priori assumption of what the data will look like. And I’ll work with this example, so I assume that there is going to be an election and there are two candidates, a blue candidate and a red candidate. And I am assuming that null votes or abstentions are not possible, and indeed no interviewee can say that they don’t know who they’re going to vote so they are forced to give preference or say they are going to vote for one or the other, so that example I’m working with. So, in this scenario I have three variables and the first one is the proportion of people that intend to vote red, which I call P. So, P is the parameter, and I want to get a plausible estimate of what is the proportion of people that intend to vote for the red candidate, but this parameter is uncertain. However, I can infer it by collecting other information, other variables, and the other variables here are the number of people that intend to vote for the red candidate that I observe, and the total number of people I have interviewed.

[0:13:00]

So, I don’t know anything about the constituency and the candidates. I may start by assuming that any type of outcome is equally plausible, so this is my simple prior that you can see represented here, where the horizontal axis represents the possible values of P from zero, which means that everyone intends to vote for candidate blue, for the blue candidate, and P equals one, which means that everyone in the population intends to vote for the red candidate. The flat line between those two extremes, values, means that I am giving equal credibility to every possible outcome. So, this really represents my naïve opinion of not knowing the constituencies, not knowing anything about the candidates and assuming that every possible outcome is equally possible, so the probability that the parameter takes a certain value is completely flat.

So, I can collect some data and I want to use these data to update my beliefs and estimate how plausible different combinations of preferences for red and blue are. So, in this case I have a random sample of five people that I have interviewed and 3 out of 5 say that they intend to vote for red. Now, the function to assign, reassign the plausibility of the parameter P is likelihood function, so this is the second element in a Bayesian approach. The likelihood function is basically a function that represents the most likely values of the parameter given the data, in other words it is a mathematical function that tells me how the data observed may have come about, may have happened. In this example, the likelihood may take the form (inaudible 0:15:30) distribution and this mathematical formula basically states that the number of preferences for the red candidate will be approximately distributed according to binominal distribution, with N representing the number of respondents and P representing the uncertain parameter values that I want to estimate. So, this formula is basically reporting the likelihood that respondents will express a preference for red given N observation and the parameter and thus represents how the variable relates with the model parameters to create the data.

So, now that I have a prior assumption, data and the likelihood function, I want to use the information, the data, to update my beliefs or rather the plausibility, the estimated plausibility of different combinations of the parameter. And how can I do that? For this purpose I used the Bayesian theorem, so this basically states that the probability, and you can see it formally described here, and says that the probability of any particular value of the parameter P considering the data is equal to the product of the relative plausibility of the data conditional on the parameter P and the prior value of the parameter P. So, this product then is divided by the average probability of the data. So, what this basically means is that the posterior distribution of probability is proportional to the product of the prior assumptions and the likelihood of the data. And in the material attached with this module I also have provided a script that shows these examples and how I created these prior and posterior probabilities. And you can see how that effectively the posterior is created by firstly calculating the likelihood of my observed data considering the range of values of the parameter P from 0 to 1, and this is multiplied by the prior probability of P, the probability that I had assumed different values of different combinations of values that P will have. And this is divided by the sum of all the products of the likelihood by the prior and will create this updated distribution of probability that you can see here on the right side, where the probability, the plausibility of the different combinations of the parameter P has been reallocated. Here you can see that the mean in this posterior distribution of the parameter P shows that the most plausible parameter value is very close to 0.060, which is the observed value in the small sample I had. But you can also see that other values are quite likely, for example, a P value of 0.50 is also quite plausible according to this posterior distribution. And this is because obviously with only five participants I shouldn’t reallocate the credibility of the parameters to narrowly and observed result of three intentions to vote for red out of five could have easily been generated by the parameter being 0.050 or 0.070 and so on.

But when I collect more interviews from a new random sample of 50 participants here, I can work from my previous model what the plausibility of the combinations of parameter values had been reallocated, so this means my prior now is more precise based on my previous updating, and indeed why shouldn’t it be the case? Why should I have worked from the assumption that zero votes for red was as likely as other combinations, when my previous data shows that this was already a very unlikely event since some of the people in the sample, in the small sample I had earlier already expressed an intention to vote for red. So, using the same methods I now can update the previous assumptions that I had updated to reallocate the plausibility of the P values considering my new data. And you can see how I did this in the script attached with this module, and you can see that considering new evidence allows to narrow the posterior distribution, as you can see here.

[0:20:52]

So, the analysis has allowed me to update the relative credibility of the parameter values in ways that are consistent with the new data. And this updated attribution of the credibility to a range of parameter values is useful if my data are not biased, for example if my sample is representative of the population of interest. But this is really a type of bias that is not a statistical problem; it’s a problem of research. So, as long as I have good data, I can update my belief, my expected plausibilities of the parameters in ways that are providing informative results.

So, I want to emphasise again that the posterior probability distribution describes the credibility of different values of the parameter P conditionally on the data collected and condition on the previous assumptions, conditionally on the previous model. And the prior plays an important role in avoiding that our models get too skewed or too narrowed by new data, and I will talk more about this in the second presentation after this one. But once I have a posterior distribution, I can use it to create large random samples from this distribution that allows to describe model information. In other words, I can draw random samples from the posterior distribution and because those samples are random, they are unbiased and provide a reliable description of the underlying distribution of plausible values of the parameters. And you can follow the script I have used to (inaudible 0:22:54) samples. The script again is provided with the material for this module. And in this script I drew 10,000 random samples from the posterior distribution and in the left plot here you can see I represented them here. In the vertical axis you can see represented different probabilities of P, so a preference for candidate red, and you can see that most of the samples from the posterior distribution cluster around values that are 0.040 to 0.75, and indeed looking at the posterior distribution, I can say that 90% of the samples lie between values that indicate a 49.40% probability of voting for red and 66.26% of preference for red, so I can say that there is 90% probability that the rate of votes for the red candidate will be between 49% and 66%.

And using the samples from the posterior, you can look at different others (inaudible 0:24:31) to report the results, so you can say, for example, that the median value is 57.83, so one of the most plausible values of preference for red is around 58% and so on. So, the key characteristic of the posterior distribution is that all possible values of the parameter are ranked by the logical plausibility. And in the second presentation after this one I will illustrate how these methods can be applied to regression analysis.

So, just to conclude this presentation, I wanted to use these quotes from McElreath and his (inaudible 0:25:32) Statistical Rethinking, and it’s a very good quote because it describes the gist of Bayesian analysis, saying that Bayesian analysis takes a question in the form of a model and uses logic to produce an answer in the form of probability distributions, and I think this is really a way to describe what Bayesian analysis can do. In the second presentation I will then talk about Bayesian approach to regression analysis.

Thank you very much. Bye now. Please remember to also check the webpage of the National Centre for Research Methods for more presentations and more material. 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