8 Varying variances, more about priors, and prior predictive checks

So far our models have been a bit ‘traditional’ in one important way: They have featured a single error term () for all conditions. This means that is the same for all speakers, listeners, and conditions in our experiment. However, we might imagine a situation where one listener’s responses are more widely distributed than another, resulting in a situation where for two listeners and . In this situation, we may not want to use a single value of for all listeners and may instead prefer to use a different value for each listener, e.g.  for listener . In this chapter we’re going to discuss models that allow for more variation in their parameters. In addition, we’re going to go into more detail about setting priors for our models and the use of prior predictive checks.

8.1 Data and Research questions

Below we load the data for our experiment investigating apparent speaker height, in addition to the brms and bmmb packages. We are going to keep focusing on only the ‘actual’, unmodified resonance.

library (brms)  
library (bmmb)  
data (height\_exp)  
options (contrasts = c('contr.sum','contr.sum'))  
height\_exp = height\_exp[height\_exp$R=='a',]

We’re going to use models whose structure is similar to the final model we fit in chapter 7 (model\_interaction). However, in this chapter we’re going to focus on questions related to variation in our standard deviation parameters. We would like to know three things:

Q1) Does our error standard deviation vary as a function of apparent speaker age?

Q2) Does our error standard deviation vary as a function of listener?

Q3) Do our speaker random effects exhibit category-specific standard deviations?

8.2 More about priors

To this point our focus has been on understanding the components that make up our models, and how these are represented using our model parameters. Now that we’ve covered most of the essentials of using categorical predictors, we can focus a bit more on the prior distributions of the parameters in our models. The reason we’ve been able to get away with not talking about priors very much is that we have substantial **domain knowledge** regarding the distribution of human height as a function of age and gender. In addition, our models thus far have been relatively simple, making the consideration of prior distributions relatively straightforward. For example, based on the information in height\_data (@cite CDC), we know that 11 year-old boys and girls are about 150 cm tall, and women and men are 163 and 176 cm tall respectively, on average. This means that there is an difference of 19.5 cm () on average between adults and children between 11 and 12 years of age. A difference of 19.5 cm between age groups suggests about a 10 cm distance (half the group difference) between each age group and the overall intercept. In other words, if apparent height is similar to veridical height, we expect an predictor of about 10 cm in magnitude given sum coding. In contrast, we have an expected height difference of 0 cm across genders for children, and 12 cm across genders for adults for an average difference of 6 cm between males and females. This means that we expect an effect of about 3 cm for gender averaged across ages. Based on this, our prior standard deviation of 12 cm for our b (fixed-effects) parameters seems very reasonable.

However, imagine a situation where there was less certainty about reasonable values for our priors. This can happen for many reasons. For example, imagine you carry out a lexical decisions task where participants listen to a combination of ‘real’ (i.e. ‘map’) and ‘fake’ words (i.e. ‘marp’) and have to decide if the word they heard is ‘real’ or not. You divide your real words into 5 groups based on a numerical measure of their frequency (how often they tend to appear in speech). To complicate matters further, you see that your reaction times are heavily skewed and so decide to model the logarithm of the reaction times. What should your priors be for your frequency groups? The prior distribution of a parameter represents the expected distribution of your parameters a priori. So what do we think are reasonable group differences in this situation? Even in this relatively simple case, setting prior distributions using only your intuitions would require understanding plausible variation in the logarithm of reaction times across word groups. This may not be realistic in this cor nor in many others.

8.2.1 Prior predictive checks

A **prior predictive check** can help understand the consequences of the prior distributions of our parameters, especially in situations where they are otherwise difficult to understand. To carry out a prior predictive check using brms you fit your model in the same way you normally would, except for setting sample\_prior="only". When you do this, your model knows to sample only from the prior, generating parameter estimates and expected values (e.g., ) based only on the prior distributions of the model parameters. The prior predictive check consists of generating fake data (e.g., ) based on the expected values generator by the prior. Conceptually, this is very similar to a posterior predictive check (discussed in section X), save for the fact that prior predictive checks are not influenced by the model likelihood (or the data).

Below we fit model\_interaction from chapter 7 again, except this time we sample only from the prior. Imagine that we did not have much knowledge about speaker heights other than that humans tend to be between 1 and 2 meters tall. Based on this we decided to be cautious and use relatively *uninformative* priors and set all prior standard deviations to 1000 cm (10 meters).

priors = c(brms::set\_prior("student\_t(3,156, 1000)", class = "Intercept"),  
 brms::set\_prior("student\_t(3,0, 1000)", class = "b"),  
 brms::set\_prior("student\_t(3,0, 1000)", class = "sd"),  
 brms::set\_prior("lkj\_corr\_cholesky (1000)", class = "cor"),   
 brms::set\_prior("student\_t(3,0, 1000)", class = "sigma"))  
  
# Fit the model yourself  
set.seed (1)  
options (contrasts = c('contr.sum','contr.sum'))  
prior\_uninformative =   
 brms::brm (height ~ A + G + A:G + (A + G + A:G|L) + (1|S), sample\_prior="only",  
 data = height\_exp, chains = 4, cores = 4, warmup = 1000,   
 iter = 5000, thin = 4, prior = priors)

# Or download it from the GitHub page:  
prior\_uninformative = bmmb::get\_model ('8\_prior\_uninformative.RDS')

We can use the predict function to get the prior predictions made by our model.

pp\_uninformative = predict (prior\_uninformative, summary = FALSE)

When we get ‘unsummarized’ predictions we get one for each set of samples in our model. These predictions vary across samples by row and across data points by column. So, to get predictions for a single posterior sample we need to observe a single row of this matrix. This could be done using a single histogram, as below:

hist (pp\_uninformative[1,])

However, the bmmb package has a simple function called p\_check that can be used to consider the density of multiple samples at once. By default this compares 10 random predictions, however the number of predictions can be changed, and the user may also specify specific samples to consider.

bmmb::p\_check (pp\_uninformative)

In the left plot of figure 8.1 we see the result of using p\_check in ten evenly spaced samples from our ‘uninformative’ prior prediction. As can be seen, our ‘cautious’ approach results in a range of simulated heights that make no sense even given our relatively limited prior knowledge, in this case that most humans are between 100 and 200 cm tall. In fact, our simulated data contains not only large negative values, which make no sense in the context of height, but also substantial values above 5000 cm, which is about the size of a 10-story office building.

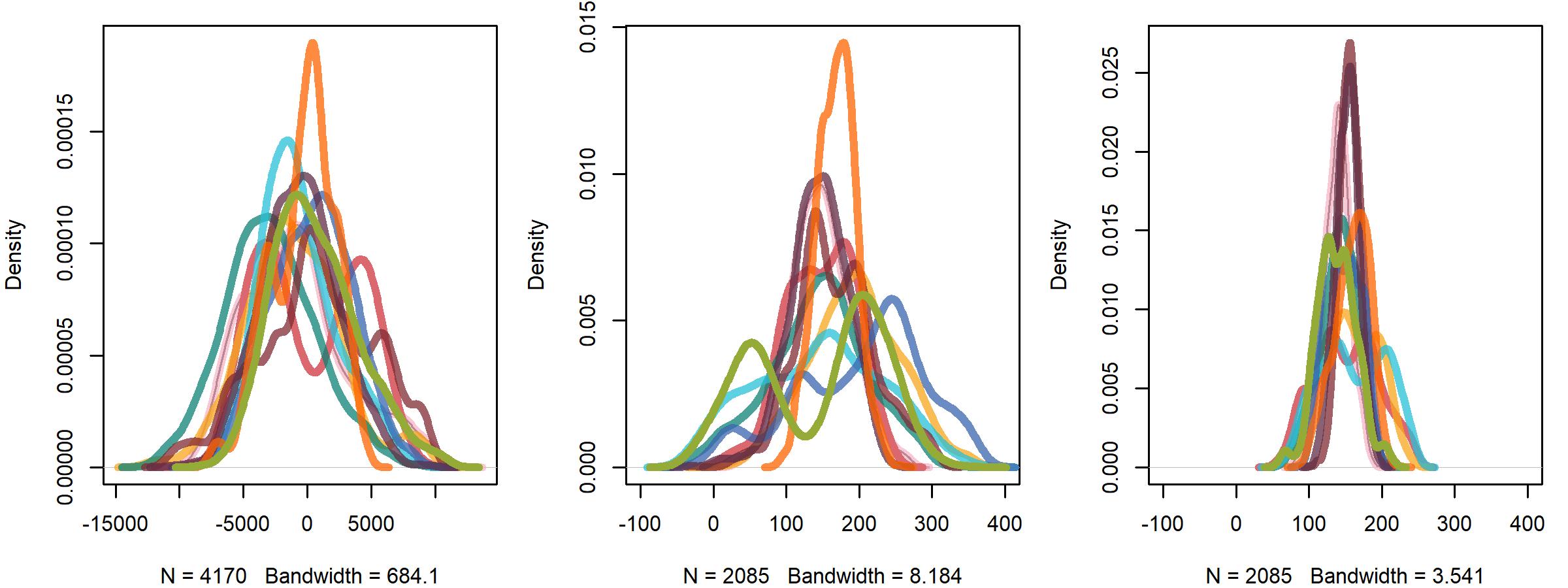


Figure 8.1: (left) Densities of 10 posterior predictions of apparent height for the uninformative prior. (middel) Densities of 10 posterior predictions of apparent height for the mildly informative prior. (right). Densities of 10 posterior predictions of apparent height for the conservative prior.

It is difficult to prove that such a wide prior is ‘bad’, and in fact some authors recommend the use of very uninformative priors (@cite krushcke). However, it is difficult to deny that a prior that results in prior predictions that are more in line with our domain knowledge is generally better for at least two reasons. First, more informative priors can help the model converge on credible parameter values, especially for complicated models with many parameters. Second, more informative priors can help improve out-of-sample prediction by providing actual information about plausible parameter values for our model. Basically, no fancy statistical model is going to convince anyone that 50 meters tall is a plausible apparent height for human speakers because human speakers are simply not nearly that tall. As a result, a model that acts as if 50 meters is a plausible apparent height for human speakers is likely to offer worse out-of-sample prediction that a model that correctly assigns little to no prior belief to this range of apparent heights. For a more in-depth discussion on the role and function of prior probabilities in multilevel Bayesian models see (@cite gelman prior). Below we again sample from the prior using the setting we’ve been using so far, and that we used in the previous chapter.

model we fit in the previous chapter (model\_interaction).

loo\_compare (model\_interaction, model\_A\_sigma, model\_A\_L\_sigma)  
## elpd\_diff se\_diff  
## model\_A\_L\_sigma 0.0 0.0   
## model\_A\_sigma -122.8 15.7   
## model\_interaction -171.7 20.2

This time we find a very large difference in , 7.8 times greater than the standard error of the difference. Again, it seems like the more complex is justified and arguably ‘better’.

## 8.6 Answering our research questions

The research questions we posed above were:

Q1) Does our error standard deviation vary as a function of apparent speaker age?

Q2) Does our error standard deviation vary as a function of listener?

Based on the information we’ve already provided, we can answer these questions: Yes, yes, and yes. We think these results are reasonable based on what we know about the heights of adults and children and about the average speaker’s average knowledge of the heights of adults and children. For example, we did think it was reasonable to expect that listener’s height judgments would be more variable for children than for adults. We think most people have a good handle on how tall adults tend to be. Does the average person know how tall 10-12 year-old children tend to be? We don’t think so. In addition, 10-12 year old can show a lot of variation in height based on differences in growth rates, which might also lead people to provide more variable estimates for the height of children. We think it also ‘makes’ sense that different listeners could be more or less systematic (i.e. predictable) than others, for many reasons. As for the final question, based on our discussion so far we also think its reasonable that the average rated heights for different children vary more with respect to those of adults.

We’re going to add another, more meta, question: Which should we report? We can approach this from the perspective that the best model according to is the ‘real’ model. Based on the comparison of above, it seems that the last model we fit (model\_A\_L\_sigma) is the ‘best’ model. Does this mean it is the ‘real’ model? And does it mean that this is the one we ought to report? The short answers to these questions are no and no. The fit between a model and data can’t ‘prove’ that the model represents the ‘real’ relationship underlying the data. The reason for this is that there are potentially a large number of other, slightly different, models that provide just as good a fit to the data, or perhaps a better one. If the best model is the real one, how can you ever know you have the best model? Further, even if you did somehow find the best model for the data you have, how could you guarantee that it will also be the best model for the data you don’t have? It’s impossible to know if some new data might come along that will not fit your model as well as one of the other model, making the ‘best’ model contingent on the data you have seen so far.

Ok so we can’t prove that our best model is the real one, fine. But its still the best one we have. Do we necessarily need to report the ‘best’ model? This sort of reasoning can lead to serious problems because there is almost always a better ‘truer’ model out there, and we often don’t try to find it. For example, many people would never think to fit a heteroscedastic model and so would not even worry about reporting it. Does the existence of a hypothetical better model mean that the model they report is invalid? We don’t think so. If it doesn’t, then why should this be the case for the researcher that *did* think to try the heteroscedastic model? In general, we can imagine that 10 people might approach any given research question in 10 different ways, a concept sometimes referred to as researcher degrees of freedom (cite). Slight differences in model structure and data processing results in slight differences in model outputs, resulting in a sort of ‘distribution’ for any given result. How can a fixed underlying reality result in a distribution or results? When they are all slightly wrong.

Rather than think of our models as representations of *reality*, we can think of them as mathematical implementations of our research questions. The model we report should include the information and structure that we think are necessary to represent and investigate our research questions. Using a different model can result in different results given the same data, but asking a different question can also lead to different results given then same data. When interpreting a statistical analysis, it is very useful to keep in mind that results are contingent on:

1. The data you collected. Given other data you may have come to different conclusions.
2. The model you chose. Given another model you may have come to different conclusions.

pp\_uninformative = predict (prior\_uninformative, summary = FALSE)

# Or download it from the GitHub page:  
prior\_conservative = bmmb::get\_model ('8\_prior\_conservative.RDS')

Consider the residuals for the random effects model we just fit to our data, which we can get from the model using the residuals function:

ztop (c(0,3.4,6.8,10.2,13.6))  
## [1] 0.5000000 0.9677045 0.9988875 0.9999628 0.9999988

# changes probabilities (p) to logits  
ptoz = function (p){  
 p[p==1] = .99 # if p=1, change to 0.99  
 p[p==0] = .01 # if p=0, change to 0.01 (i.e. 1-0.99)  
 log (p) - log(1-p)  
}