# 7 Comparing many groups, interactions, and posterior predictive checks

We’ve been making our models more and more complicated, but we’ve still only fit models with extremely simple fixed effects structures. In the last two chapters we talked about comparing two groups. Although the comparison of two groups is very simple, it also comes up often. In addition, more complicated problems are often broken down into several two-group questions in the course of understanding a data set. However, most real-world experiments don’t usually *begin* as two-group questions. In this chapter, we’re going to talk about the comparison of observations across multiple groups. In terms of model building concepts, in this chapter we’re going to discuss how to build models with multiple fixed effect predictors and how to allow these factors to *interact* in our models.

## 7.1 Comparing four (or any number of) groups

First, we’re going to discuss models that compare observations across any number of groups (i.e., levels of a single factor). A model that can be used to compare multiple groups might have just one predictor factor, but that factor can have any number of levels [@@ NHS: In some sense, I can see how it has “one predictor”, since R can read it in as a factor, but doesn’t it turn it into something like dummy coding (or whatever) under the hood? Maybe it’s not worth making this distinction here…]. Although it’s not very common to have a study design or model that *only* consists of a single multi-group comparison, the inclusion of multi-group comparisons is ubiquitous in many disciplines (including linguistics). Examples of predictors with multiple levels are language (e.g., participants’ first languages), word category (e.g., part of speech, or usage frequency bins), and age groups. In figure 7.1 we see two ways that subjects (or listeners, or the levels of any grouping factor) can be arranged into groups. In the *within-subjects* version of this design, each subject is tested at each level of . In the *between-subjects* version of the design, different subjects are tested at every level of . Here we omit the more complicated design discussed in chapter 5, in which some subjects appear at all levels of and others don’t. Such designs are definitely analyzable with brms, and sometimes they may be unavoidable. However, they are not included in figure 7.1 because there are a large number of possible configurations of this, and because the additional complexity of such designs can make them extremely challenging to use effectively, so we are typically better off not using such designs if at all possible.

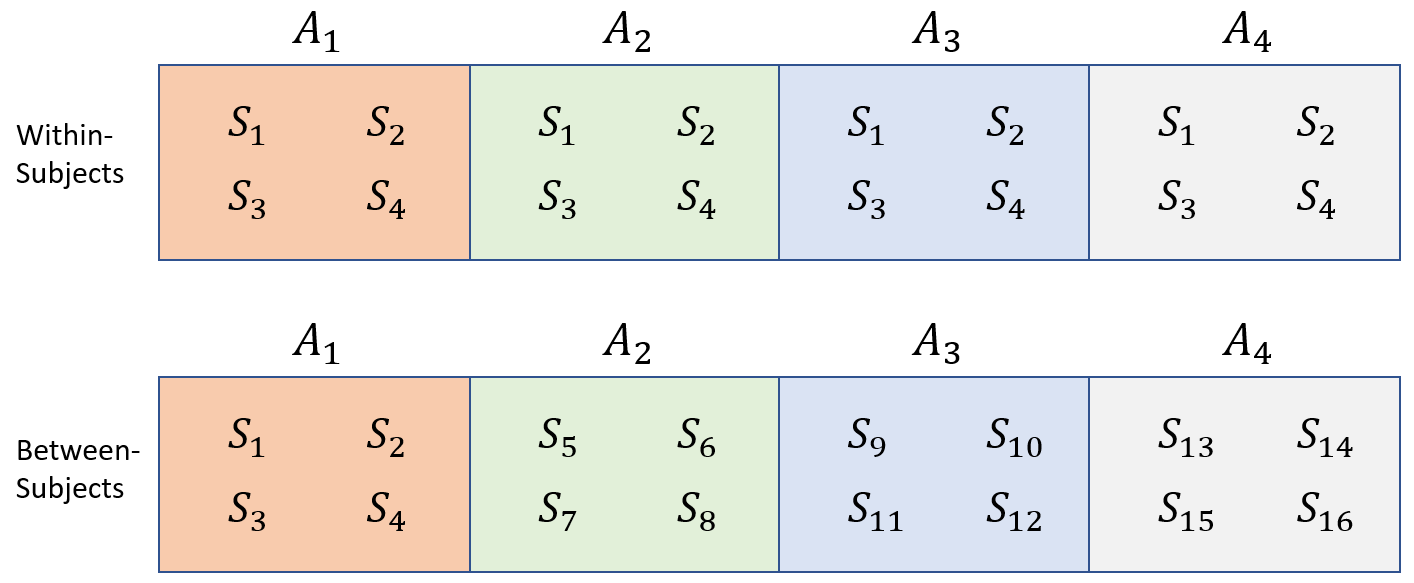


Figure 7.1: Data from a single group of subjects (S) divided according to a single grouping factor A.

### 7.1.1 Data and research questions

Below we load the data for our experiment investigating apparent speaker height, in addition to the brms and bmmb packages.

library (brms)  
library (bmmb)  
options (contrasts = c('contr.sum','contr.sum'))  
data (exp\_data)

In figure 7.2 we see that the distribution of apparent height with respect to apparent speaker category is very similar to its distribution with respect to veridical speaker category. This is not terribly surprising since as we can see below, listeners identified speaker category in 75% of cases. In addition, errors were most common for boys and girls, speakers that are expected to be approximately the same size.

# overall average correct category identification  
mean(exp\_data$C == exp\_data$C\_v)  
## [1] 0.7458  
  
# average correct category identification by category  
tapply(exp\_data$C == exp\_data$C\_v, exp\_data$C\_v, mean)  
## b g m w   
## 0.5778 0.6456 0.9274 0.7097

In this chapter we’re going to focus on the effect of *apparent* speaker category on apparent height. This is because we want to understand how the listener’s assumptions about the speaker affect their height judgments, and listeners did not know the veridical speaker category. We’re first going to treat the four groups as if they had no internal relationships between the groups (e.g., ignoring the fact that we might expect some degree of similarity in the relationships between between women and girls, on the one hand, and men and boys, on the other). In this case it may not be the best approach for this data since we know there are logical ways to subdivide boys, girls, men, and women. However, this is a good starting point since in many cases you will have levels of a factor with no logical internal divisions.

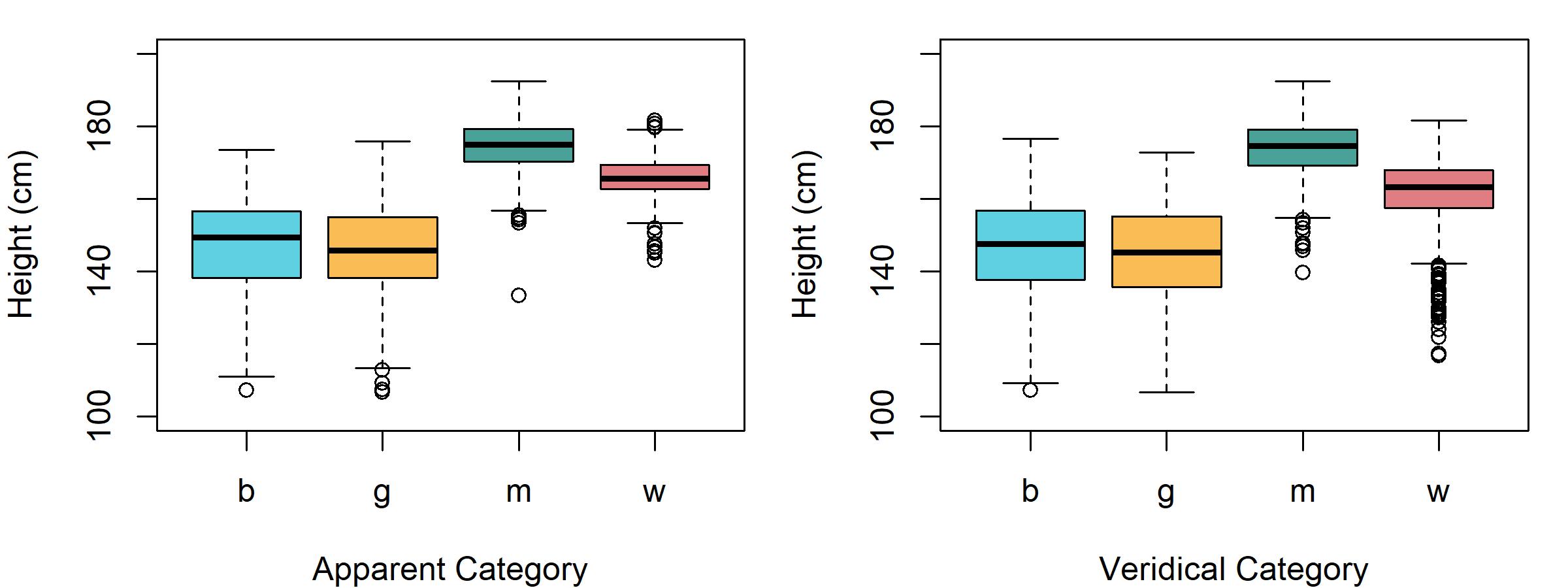


Figure 7.2: Apparent height organized by apparent age (left) and veridical age (right) for boys (b), girls (g), men (m), and women (w).

Even with only four different groups our potential research questions are substantially more complicated than in the two-group case. First, there are four groups now, meaning we could potentially make 6 2-group comparisons. Second, the groups differ along multiple dimensions, making it more difficult to make two-group comparisons that answer a single question. For example, the ‘man’ and ‘girl’ groups differ according to apparent adultness *and* apparent gender. In the presence of a difference between groups, this makes it difficult to isolate the contribution of each characteristic to the difference in apparent height. For now, we’re going to keep things simple and just build a model that can tell us about our average group values. To analyze data like this it should be in a data frame with one row for each observation. Then, one column should contain the dependent variable, the variable whose values and variation you’re trying to predict. Another column should contain information about which group each observation belongs to for the fixed-effects predictor you will be considering. Finally, you need one or more columns indicating the sources of data in your experiment. The model in this section will use the following variables from our data frame:

* L: A number from 1-15 indicating which *listener* responded to the trial, being treated as a character.
* C: A letter representing the speaker *category* (b=boy, g=girl, m=man, w=woman) reported by the listener for each trial.
* height: A number representing the *height* (in centimeters) reported for the speaker on each trial.
* S: A number from 1-139 indicating which *speaker* produced the trial stimulus.

And we are going to ask the following basic question:

Q1) Is average apparent speaker height different based on apparent speaker category?

[@@ NHS: I find that phrases like “based on” kind of bother me here - we don’t really know if people are categorizing the stimuli as b/g/m/w first and then basing their height judgments (in part) on the category, right? maybe they do both simultaneously, or maybe they judge height and that informs the category…]

[@@ SB - I went back over the earlier chapters to remove causal stuff like this]

### 7.1.2 Description of our model

In the last two chapters we discussed the inclusion of a single factor with two levels in our models. Our model formula does not change in structure to include predictors with many levels. The formula to fit our model will be:

height ~ C + (C|L) + (1|S)

Where C is the apparent speaker category indicated by the listener for each trial, a factor with four levels (see above). Since we include C on the left-hand-side of the pipe of L, we know that we are estimating the category by listener interaction using adaptive partial pooling (see section X). At this point we have to make a decision regarding how we represent our models in more formal notation, and striking a balance between readability and models that are directly analogous to the output in R. Each level of the category predictor that we are estimating needs to be represented by a model parameter. Since we are using sum coding for all our factors (see section X), our model will estimate three of the four group effects, omitting the alphabetically last level (in this case woman). As a result, the formula above is analogous to:

height ~ C1 + C2 + C3 + (C1 + C2 + C3|L) + (1|S)

In section 4.X we noted that each parameter in our model needs to be associated with a predictor. If we call the predictors x1, x2, and x3 our model formula could sort of be thought of like this:

height ~ x1\*C1 + x2\*C2 + x3\*C3 + (x1\*C1 + x2\*C2 + x3\*C3|L) + (1|S)

[@@ NHS: I don’t understand what x1 is here or why x2 and x3 aren’t in the formula…] [@@ SB - this was basically a typo]

Of course this is no longer a real or working model formula, we are just trying to show how the might actually be included in our model. When a factor has only two levels, the effect for one level must equal the negative of the effect for the other level. As a result, group membership for observations can be represented using a single predictor equal to either -1 or 1 (for sum coding). When the observation is associated with one level of the factor the predictor is 1, and when the observation is associated with the other level the predictor is -1. In that way, a single predictor and coefficient can reflect two different levels of the factor. When there are more than two levels, each level that is being estimated is associated with a predictor variable, meaning that we expect estimated coefficients for a factor with $J levels in the absence of additional constraints. An example of this is seen below, where we generate three ‘fake’ predictors to represent four groups. Each row is represented by a single 1 save for the final row which is represented by -1 in *all other predictors* for that factor.

contr.sum(1:4)  
## [,1] [,2] [,3]  
## 1 1 0 0  
## 2 0 1 0  
## 3 0 0 1  
## 4 -1 -1 -1

Imagine the tiny matrix above represented the numerical predictors for our C parameters so that the first, second, and third columns represent and respectively. Our tiny data matrix above suggests that the first observation (row) belongs to the first group (boy). This is because our data results in the parameter being multiplied by one and and being zeroed out. The same logic suggests the second row is an observation from the second group and the third from the third. Since is equal to (under sum coding), adding the negative of , , and to the prediction is equivalent to adding the value of even though it is not directly estimated. So, the fourth row of our tiny matrix represents an observation from the fourth group and encodes the effect for even without directly estimating this value. As a result, even though our model only directly estimates three parameters for our predictor, we could think of our model as tacitly representing four parameters for : , , , and . Even though the inclusion of in our model requires the addition of three separate parameters, we’re going to represent these with a single term like in our formal model descriptions. This term selects the value of for trial based on the value of the C (apparent category) predictor. This is consistent with the way we have been representing our listener and speaker-dependent intercepts (as discussed in section 4.X). However, a possible confusion arises from the fact that the subscript of the term references a predictor which has four levels (C), however, our parameter does not actually contain four separate parameters. In other words, in situations where C = w (i.e. ) this would seem to reference the fourth parameter, one that doesn’t actually *exist* in our model. However, in such a situation the model can be thought of as using the negative sum of the other three parameters, effectively using the value of even without directly estimating the parameter. Since we are including adult male speakers in our data we should also reconsider our priors. Based on CDC (cite) we expect that average heights of 176, 162, 150 and 150 cm respectively for men, women, girls, and boys. These four groups have an average of 159 cm and a standard deviation of 12 cm. Based on this, and especially given that we are investigating *apparent* and not *veridical* height, the priors we used in the last chapter still seem appropriate. The formal description of our model is given below:

Structurally, this model is extremely similar to the model described in (??) last chapter with replacing and replacing . Note that since we are estimating three parameters for C, our listener-dependent ‘random effects’ are actually coming from a four-dimensional normal distribution which considers the relationships between the intercepts and category effects for individual listeners.

### 7.1.3 Fitting and interpreting the model

We fit the four-group model below:

# Fit the model yourself  
# set the random seed and contrasts  
set.seed (1); options (contrasts = c('contr.sum','contr.sum'))  
  
# set the priors  
priors = c(brms::set\_prior("student\_t(3,156, 12)", class = "Intercept"),  
 brms::set\_prior("student\_t(3,0, 12)", class = "b"),  
 brms::set\_prior("student\_t(3,0, 12)", class = "sd"),  
 brms::set\_prior("lkj\_corr\_cholesky (2)", class = "cor"),   
 brms::set\_prior("gamma(2, 0.1)", class = "nu"),  
 brms::set\_prior("student\_t(3,0, 12)", class = "sigma"))  
model\_four\_groups =   
 brms::brm (height ~ C + (C|L) + (1|S), data = exp\_data, chains = 4,

cores = 4, warmup = 1000, iter = 5000, thin = 4, prior = priors, family = "student")

# or download it from the GitHub page:  
model\_four\_groups = bmmb::get\_model ('7\_model\_four\_groups.RDS')

listener\_effects\_hat = ranef(model\_sum\_coding, summary = FALSE)$L[,,"Intercept"]  
str (listener\_effects\_hat)  
## num [1:5000, 1:15] 13.17 9.58 12.68 9.24 9.55 ...  
## - attr(\*, "dimnames")=List of 2  
## ..$ : NULL  
## ..$ : chr [1:15] "1" "2" "3" "4" ...

We’ll focus on the fixed effects, seen below:

brms::fixef(model\_four\_groups)  
## Estimate Est.Error Q2.5 Q97.5  
## Intercept 158.581 1.083 156.41 160.685  
## C1 -9.318 1.310 -11.87 -6.682  
## C2 -12.082 1.494 -15.07 -9.110  
## C3 15.286 1.327 12.59 17.950

We can see that the intercept corresponds to the average of the group means, and our C coefficients correspond reasonably to the centered group means.

# group means  
means = tapply (exp\_data$height, exp\_data$C, mean)  
  
# overall mean  
mean (means)  
## [1] 158  
  
# centered means  
means - mean (means)  
## b g m w   
## -11.12 -13.03 16.56 7.59

There’s two things worth noting here. The first is that the coefficients don’t exactly match the group means. There are many possible reasons for this including the fact that the group means don’t account for repeated measures, a lack of balance in the data, or for the way that outliers can unduly influence sample mean estimates. Basically, if they were totally off we should be worried that something may have gone wrong but we should expect that there may be some differences between model coefficients and the sort of ‘careless’ averaging we’ve done above. Second, as discussed above, we’re missing the C4 coefficient. We can use the short\_hypothesis function to easily recover the final group coefficient by finding the negative sum of the coefficients that *were* estimated. An example of this is shown below.

# parameters = centered means  
all\_groups =   
 bmmb::short\_hypothesis (  
 model\_four\_groups,   
 c("Intercept = 0", "C1 = 0", "C2 = 0", "C3 = 0", "-(C1+C2+C3) = 0"))

We can also use short\_hypothesis to compare different combinations of our fixed effects. Using the code below we test for a difference in group means between the first and second groups (boys and girls), and between the third and four groups (men and women).

# parameters = centered means  
comparisons = bmmb::short\_hypothesis (model\_four\_groups,   
 c("C1 = C2", "C3 = -(C1+C2+C3)"))

We can use brmplot from the bmmb package to easily make plots showing means and credible intervals for our fixed effects, including the recovered coefficient. We also plot the two comparisons we made using the hypothesis function above, presented in figure 7.3. The distributions on the left tell us about probable values for our category parameters: Given our data and model structure, there is a 95% probability that the parameter value falls within this interval [@@ NHS: shouldn’t this be (something like) “this is the most probable 95% set of possible parameter values”? the way it’s written sounds like a very frequentist, CI-focused interpretation (and incorrect with respect to what CIs are/mean)]. Our category effects reflect the expected difference between our model intercept, the overall grand mean, and each category mean. So, the intervals on the left of figure 7.3 tell us that there is a 95% probability (given our data and model structure) that the actual difference between the grand mean and each group mean falls within that range [@@ NHS: ditto]. What do the distributions in the right plot in figure 7.2 tell us? They represent the posterior distribution of the difference between category means directly, without involving the intercept. Conceptually, finding the difference between the intercept and a group mean is no different than finding the difference between individual group means. Consider the example of providing directions with respect to your house. If the library is 5 miles west and the park is 3 miles west, you could refer to these as -5 and -3 miles from your house, west to east. However, you could also express the difference between one and the other directly as 2 (or -2). The left plot in figure 7.2 expresses all distances with respect to the intercept (your house) while the right plot expresses differences with respect to individual category means directly (the library and park).

17.4/2  
## [1] 8.7  
163.9 - (17.4/2)  
## [1] 155.2

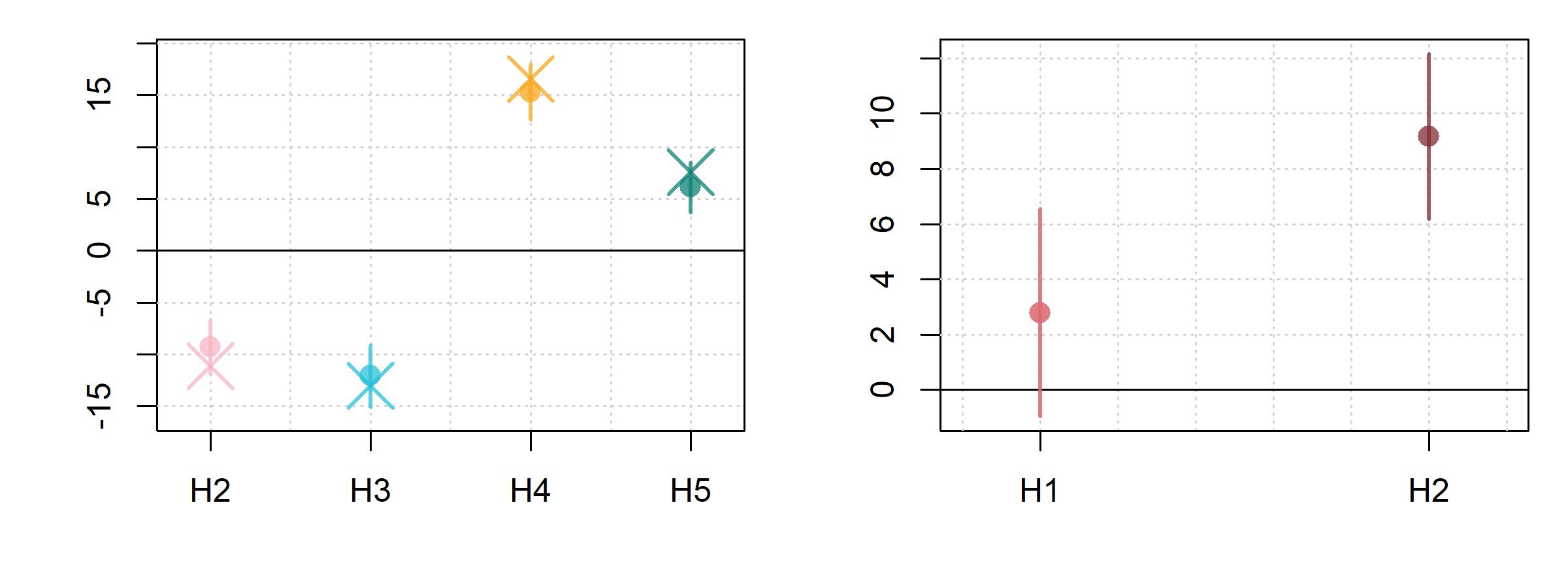


Figure 7.3: (left) Means and 95% credible intervals for boy (H1), girl (H2), man (H3), and woman (H4) effects for our four group model. (right) Means and 95% credible intervals for differences between boys and girls (H1) and men and women (H2).

[@@ NHS: stopping here for now 2022-07-02 11:30 AM]

## 7.2 Investigating multiple factors simultaneously

Our model above had a single factor with four groups/levels. Another way to look at it is that the model predicted variation along a single **dimension**, that is a single piece of information that can be used to distinguish our groups. However, we can usually categorize people (or observations) along multiple dimensions simultaneously. So, a speaker may be an adult *and* female, and not just adult *or* female. This is because adultness and gender are independent characteristics (dimensions) and the value of one does not determine the value of the other. In our height experiment we effectively asked listeners to make two binary classifications: Is the listener an adult or a child? Is the listener female or male? This means we can think of our four groups as varying along two dimensions, each represented by a different factor. However, although we know this our previous model didn’t, it thinks we just have four groups that have no logical internal distinctions. To include multiple predictors in our model and inspect the effect of all combinations of factor levels on our dependent variable, we must *cross* our predictors. This means that we observe each level of one factor at every level of the other factor (and vice versa). For our data this means that we observe both male and female children, and male and female adults. Crossing factors results in individual groups for two factors ( and ) with and number of levels each. Since we have two factors with two levels each, crossing age and gender results in four unique groups: Boys (male child), girls (female child), men (male adult), and women (female adult). Designs where all the factors are crossed are sometimes called **factorial** or **orthogonal** designs. When you have orthogonal designs, that means that the effect of your ‘dimensions’ (i.e., factors) can be estimated independently. For example, if you want to know the difference in apparent height according to apparent gender, it helps to have observed both adults and children for each gender. A comparison of boys and women will conflate gender and age differences, thereby making it difficult to understand the independent effect of apparent gender. By balancing factors across each other, orthogonal designs allow us to *decompose* the variation in our data using different factors more effectively. This allows us to attribute the variation in our observations to the different factors in our design (we will return to this idea in chapter 11). Figure 7.4 presents a few ways that subjects be combined with two two-level factors, and , whose combination can represent four groups. First, we can see that factors and are crossed with each other since there are subjects, and presumably observations, in each little **cell** (i.e. combination of factor levels). What would it look like for these factors to not be crossed? Well, one of the ‘cells’ (boxes) would be completely missing observations. If this were to happen, we would obviously lose the ability to make comparisons across groups involving the values in that cell. If your subjects factor is fully crossed with your fixed effects factors (in this case and ) then you have a fully *within-subjects* design. If you only observe certain subjects for specific combinations of fixed effects, you have a fully *between-subjects* design. Finally, we see a new possibility, a **mixed design**. In a mixed design, one or more factors are between-subjects, and the others are within-subjects. In the example below, is between-subjects but is within-subjects. An example of a mixed design would arise if, for example, we wanted to run our coffee and reading time experiment (discussed in section) on multiple first languages. Each subject could be tested at both coffee and decaf, making that factor within subjects, however, each subject would only be observed within a single first language group, making that factor between subjects.

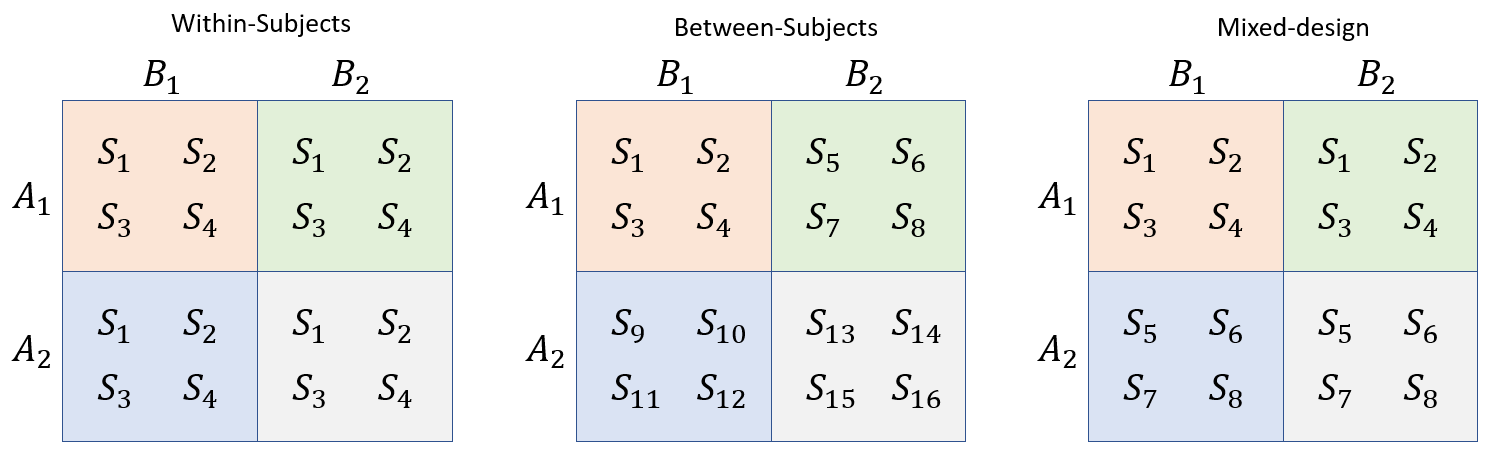


Figure 7.4: Data from four groups of subjects (S) divided according to a two grouping factors A and B.

Our model contains two predictors with two levels each, leading to four groups. Height judgments for these four groups are presented in figure 7.5. These four boxes correspond to the four cells in figure 7.4. The middle plot in 7.5 compares height judgments across levels of gender (male vs. female). Since we have both adults and children within each gender group, each box in the plot is really the average of two of the boxes from the left plot in 7.5 (boxes 1 and 3 vs. 2 and 4). This is equivalent to finding the average of one row in 7.4, and comparing it to the average of the other row. Similarly, the right plot in 7.4 compares height judgments across levels of age (adult vs. child). Since we have both males and females within each gender group, each box in the plot is really the average of two of the boxes from the left plot in 7.5 (boxes 1 and 2 vs. 3 and 4). This is equivalent to finding the average of one column in 7.4, and comparing it to the average of the other column. In this way, our model including two fixed effect predictors can use four groups to answer two questions simultaneously: What is the effect of apparent age on apparent height? What is the effect of apparent gender on apparent height?

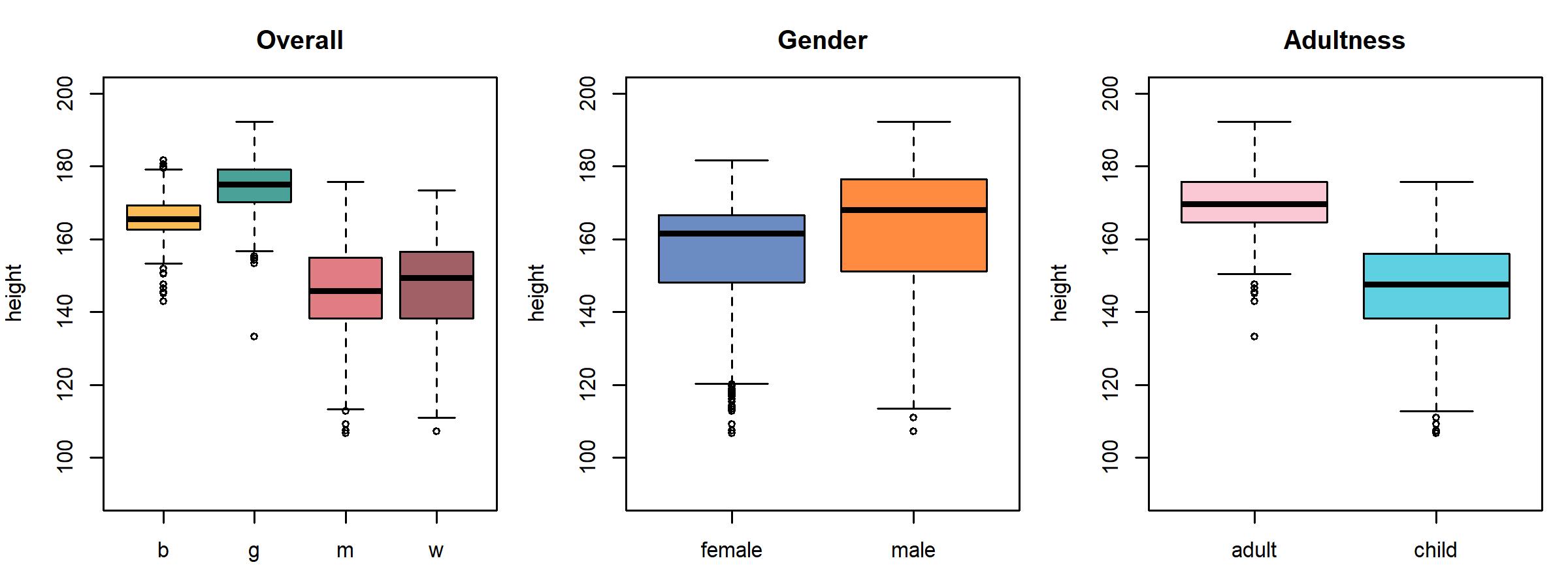


Figure 7.5: (left) Comparison of the four groups (middle) Comparison of productions based on whether the speaker is an adult (right) Comparison of all productions based on whether the speaker is male.

### 7.2.1 Data and research questions

We are going to a use slightly different set of columns from our data this time, this time using A and G rather than C. The model in this section will use the following variables from our data frame:

* L: A number from 1-15 indicating which *listener* responded to the trial, being treated as a character.
* height: A number representing the *height* (in centimeters) reported for the speaker on each trial.
* S: A number from 1-139 indicating which *speaker* produced the trial stimulus.
* G: The *apparent gender* of the speaker indicated by the listener, f (female) or m (male).
* A: The *apparent age* of the speaker indicated by the listener, a (adult) or c (child).

And we are going to split the question we posed above into two:

Q1) Are there differences in average apparent height based on apparent speaker age?

Q2) Are there differences in average apparent height based on apparent speaker gender?

### 7.2.2 Description of the model

In our previous model, we acted like we just had four different groups with no internal structure. Of course, we know that our groups differ systematically from each other in meaningful ways. For example, we might have chosen to fit two separate models investigating age and gender like this:

brm (height ~ A + (A|L) + (1|S)

brm (height ~ G + (G|L) + (1|S)

For several reasons (some of which we’ll see very soon), it’s preferable to fit a single model with both predictors at once, rather than fitting separate models for each one. Our R model formula will now look like this, reflecting the influence of both predictors simultaneously:

height ~ A + G + (A + G|L) + (1|S)

This can be read like “apparent height is distributed according to effects for apparent adultness and gender, with random intercepts, age, and gender effects for each listener, and random intercepts for each speaker”. You may have noticed that our model no longer includes the C predictor. This is because the C label is perfectly predictable on the basis of A and G (i.e., as a female child, a member of the g group must have values of f and c for their G and A predictors). Basically, we have *decomposed* the four categories into two components to help us understand the effect of each. For example, the fixed effect predictors in our model were previously:

We can expand the term in parenthesis as seen below. Since group can be exactly represented by combinations of gender and adult, our model sort of ‘secretly’ contained this more-complicated model inside of it. However, this information was not directly built into the model.

The above can be thought of as a ‘decomposition’ of the variance in your model by breaking it up into parts and assigning them to different sources. We have actually been doing this form the start by making more and more complicated models. Recall that our very first approach (in chapter 3) to understanding variation in apparent height treated all variation as error, and so did not ‘decompose’ anything. In chapter 4 we added between-listener and speaker variation to the model, and in chapter 5 we added apparent age. Now we include apparent age and gender, listener specific effects and intercepts, and speaker specific intercepts. There is a fixed amount of variation in our dependent variable. Every tiny bit of variance that these added components explain takes away from the error variance, and potentially also from the other components that had previously been in the model. Our updated model description is provided in (7.4).

Our models have increasingly large numbers of priors, and many of these are the same. As a result, going forward we will place these on the same line when they have the same prior and represent the same sorts of parameters (in much the same way they are specified by class by the set\_prior function). For example, the line $\sigma\_L, \sigma\_{A \colon L}, \sigma\_{G \colon L}, \sigma\_S \sim N(0,12) \\$ specifies the prior probability of all standard deviation terms individually, but at the same time. In plain English (ignoring the priors) this says:

“We expect apparent height to vary according to a t distribution with a trial-specific mean parameter. That mean varies based on the apparent age and gender of the speaker, listener dependent intercepts, age, and gender effects, and speaker dependent intercepts. Speaker random intercepts were drawn from a univariate normal distribution, and the listener random effects were drawn from a three-dimensional normal distribution.”

### 7.2.3 Fitting and interpreting the model

Below we fit a model with the structure outlined above. We don’t need to change our priors because our new predictors are all covered by the classes of priors we have specified so far, and we think the priors we have been using thus far are still a good fit for our data.

# Fit the model yourself  
# set the random seed and contrasts  
set.seed (1); options (contrasts = c('contr.sum','contr.sum'))  
  
priors = c(brms::set\_prior("student\_t(3,156, 12)", class = "Intercept"),  
 brms::set\_prior("student\_t(3,0, 12)", class = "b"),  
 brms::set\_prior("student\_t(3,0, 12)", class = "sd"),  
 brms::set\_prior("lkj\_corr\_cholesky (2)", class = "cor"),   
 brms::set\_prior("gamma(2, 0.1)", class = "nu"),  
 brms::set\_prior("student\_t(3,0, 12)", class = "sigma"))  
model\_both =   
 brms::brm (height ~ A + G + (A + G|L) + (1|S), data = exp\_data, chains = 4, cores = 4,  
 warmup = 1000, iter = 5000, thin = 4, prior = priors, family = "student")

# Or download it from the GitHub page:  
model\_both = bmmb::get\_model ('7\_model\_both.RDS')

We can inspect the model fixed effects:

# inspect the fixed effects  
brms::fixef (model\_both)  
## Estimate Est.Error Q2.5 Q97.5  
## Intercept 158.654 1.1947 156.319 161.025  
## A1 9.979 1.2107 7.562 12.282  
## G1 -2.205 0.5524 -3.288 -1.124

We now have two non-Intercept ‘Population-Level’ effects: A1 and G1, representing the categories ‘adult’ and ‘female’ respectively. Remember that since we used sum coding and have only two levels, the effects for the groups that are not represented (‘child’, ‘male’) are just the opposite sign of the groups that are represented. Below we calculate the mean apparent height of the four groups and find the grand mean of the four groups. We can see that this is very similar to our intercept. Next we find the average apparent height at the different levels of adultness, find the difference between these, and divide it by negative two (just to flip the sign to match our effects). We also do the same for the levels of age. We can see that these values are reasonably similar to the estimated A1 and G1 predictors, which represent the same information.

# Intercept  
mean (tapply (exp\_data$height, exp\_data$C, mean))   
## [1] 158  
# Age effect  
diff (tapply (exp\_data$height, exp\_data$A, mean) ) / -2  
## c   
## 12.16  
# Gender effect  
diff (tapply (exp\_data$height, exp\_data$G, mean) ) / -2  
## m   
## -3.166

We can recover the four individual group means by adding up individual fixed effect coefficients. This can sometimes be tedious and requires you to be careful and methodical, but isn’t actually difficult. Remember that each of the four groups is uniquely identified by a combination of gender and adultness. This means that to recover the expected group means we need to add the right combination of coefficients to the intercept. For example, the first hypothesis we test below (Intercept + -A1 + -G1 = 0) takes the intercept and adds the effect for ‘child’ (-A1). We use -A1 and not A2 because our model does not contain a parameter called A2. This parameter is not estimated in our model because -A1=A2. So, if we want the value of A2 we use -A1. Similarly, we add the effect for ‘male’ using -G1. Since we started with the overall mean apparent height (the intercept) and add the effects for a ‘male’ and a ‘child’, the hypothesis Intercept + -A1 + -G1 = 0 estimates the group mean for ‘boys’.

means\_pred = bmmb::short\_hypothesis (model\_both,   
 c("Intercept + -A1 + -G1 = 0", # boys  
 "Intercept + -A1 + G1 = 0", # girls   
 "Intercept + A1 + -G1 = 0", # men  
 "Intercept + A1 + G1 = 0")) # women

Below we compare our recreated group means with our average height judgment for each group. There is a reasonably close match, however there also appear to be some important mismatches. For example, we see that adult women and men differ in height by 9 cm in our data, while boys and girls by only 2 cm. So, it seems like in our data the difference between males and females is different for children and adults. However, according to our model men and women and boys and girls both differ from each other by about 4-5 cm each, a value in between the 9 and 1 cm differences between the groups in our data. This discrepancy suggests that our model is potentially missing important information about our data.

means\_pred  
## Estimate Est.Error Q2.5 Q97.5 hypothesis  
## H1 150.9 2.2410 146.6 155.3 (Intercept+-A1+-G1) = 0  
## H2 146.5 2.3680 142.0 151.2 (Intercept+-A1+G1) = 0  
## H3 170.8 1.2988 168.2 173.4 (Intercept+A1+-G1) = 0  
## H4 166.4 0.6901 165.0 167.8 (Intercept+A1+G1) = 0  
  
tapply (exp\_data$height, exp\_data$C, mean)  
## b g m w   
## 146.9 145.0 174.5 165.6

## 7.3 Linear and posterior prediction

So far we’ve been working with very simple models and so we haven’t worried very much about how well they *fit*, or represent our data. Our reconstruction of the group means above suggests our current model may have some issues, and that we should be concerned about its ability to accurately capture the patterns in the data. So far we’ve only been discussing modeling in one direction, going from data to abstract, reduced representations (i.e. regression models). However, it’s also worth considering going in the other direction, from the abstraction representation back towards data. One way to consider how well our model represents our data is to consider what data would be *generated* by the relationships represented in our model. Using your model in this way is called **prediction**, and you can use prediction to evaluate the fit of your model. The reasoning is basically, if your model is a good representations of the data, the ‘fake’ data it generates should look very much like your real data. If the fake data it generates doesn’t look like your real data, something about your model is slightly ‘off’. The discussion of prediction presented here will involve dependent variables that do not require a link function (discussed in chapter 10). Some of the details discussed here, in particular those involving expected values, do not exactly apply to models that require a link function. All our models so far have contained lines like those seen in (7.5). These lines tell our models that: 1) We expect our data to be normally distributed given some mean and standard deviation, and 2) The mean for a given trial can be predicted using some combination of the independent variables.

The value of above is the expected value for a given trial. This is also sometimes called the **linear predictor**. Since the linear predictor is the value you expect for a trial, it’s obvious that there should be a reasonable correspondence between the linear predictor and your dependent variable. If what your model tells you to expect is not like what you actually observe, then what good is the model? One shortcoming of the linear predictor is that this does not incorporate error. As a result, the linear predictor is constant for any single combination of dependent variables, and undersells the amount of variation that can be expected in a given data set. The **posterior predictive distribution** is the distribution of possible data given your parameter estimates and probability model. In other words, the posterior predictive distribution is your linear predictor, plus the random error that you expect in your model. For normally distributed data, our posterior predictive distribution is the distribution of generated values of given estimates of and as in (7.6).

A *posterior predictive check* samples from the posterior predictive distribution with the intention of comparing these samples of to your original data . Posterior prediction can be used to assess how well your model fits your data. If your model really ‘gets’ your data, the fake data it generates ( ) will have similar characteristics to your real data (). If the fake data your model generates looks substantially different from your real data, that suggests a fundamental misalignment between your model and your data. The generic functions fitted and predict will help you get you linear and posterior predictions from your brm model.

# linear predictor  
y\_lin\_pred = fitted (model\_both)  
  
# posterior prediction  
y\_post\_pred = predict (model\_both)

In both cases the functions return four-column data frames with as many rows as the number of observations in the data. These data frames contain columns representing the mean, standard deviation and the upper and lower 95% credible intervals for each one of our data points. The reason we get intervals for our posterior predictions is because our model has a different prediction for each set of posterior samples. This means that if we have 4000 samples we actually have 4000 slightly different models and 4000 slightly different predictions. So, in addition to information about the average prediction for each data point, we get information about expected variation around these estimates. Below we see the (summarized) data frame representing our linear predictors. You can also set summary=FALSE to get the individual (unsummarized) predictions. If you do this you get an individual predictions for every set of posterior samples for your parameters, for every data point. So, if you have 1500 data points and 4000 posterior samples, the unsummarized output of these functions would be a matrix with 4000 rows and 1500 columns.

# linear predictions  
head (y\_lin\_pred)  
## Estimate Est.Error Q2.5 Q97.5  
## [1,] 156.8 2.182 152.5 161.0  
## [2,] 161.7 2.059 157.6 165.6  
## [3,] 160.6 1.983 156.6 164.3  
## [4,] 162.3 1.846 158.5 165.9  
## [5,] 163.0 1.992 159.0 167.0  
## [6,] 155.7 2.371 151.1 160.5

Below we see the data frame corresponding to our posterior predictions. Note that although the means of the linear and posterior predictions are very similar, the intervals around the posterior predictions are noticeably wider. This is because they incorporate error. Another difference between the linear and posterior predictions is that the linear predictors are fixed for a given set of posterior samples. In contrast, every time you run the command predict(model) you will get slightly different results because these incorporate random error.

# posterior predictions  
head (y\_post\_pred)  
## Estimate Est.Error Q2.5 Q97.5  
## [1,] 157.0 7.966 142.0 172.3  
## [2,] 161.7 8.108 146.7 176.8  
## [3,] 160.6 8.407 144.6 176.2  
## [4,] 162.2 7.708 146.8 178.0  
## [5,] 162.7 8.024 147.7 177.5  
## [6,] 155.6 7.589 140.3 170.4

There is a parameter for the predict function, re\_formula, that determines which random effect terms are included in your prediction. By default this parameter is set to NULL meaning that your predictions will be made made using your complete model formula. For example for our last model this means our posterior predictions reflect this model formula:

height ~ A + G + (A+G|S) + (1|L)

We can instead set re\_formula=NA to make predictions using only the fixed effects as seen below.

y\_post\_pred\_no\_re = predict (model\_both, re\_formula = NA)

This corresponds to the predictions made by this model:

height ~ A + G

We could also include a subset of random effects, for example including only the listener random intercepts as seen below:

y\_post\_pred\_some\_re = predict (model\_both, re\_formula = "~(1|L)")

height ~ A + G + (1|L)

Below we sample from the posterior distribution and keep individual, unsummarized samples. We do this for our full model and for a model that makes only fixed effects predictions.

y\_hat = predict (model\_both, summary = FALSE)  
y\_hat\_no\_re = predict (model\_both, re\_formula = NA, summary = FALSE)

In figure 7.6 we compare the distribution of our data to our posterior predictive distributions, without and with the inclusion of random effects. There are some differences in the ‘widths’ of the different boxes, but we will not discuss this for now. Instead, we would like to highlight differences in the relative positions of the boxes representing the four groups across the plots. Our data shows different average heights for men and women but no noticeable height difference between boys and girls. Despite this, both sets of our posterior predictions show a difference in height between boys and girls. This is particularly evident in or fixed effects predictions since the random effects can help represent many of the idiosyncrasies of any given data set. As we will discuss in the following section, the structure of our data suggests that there is an *interaction* between apparent gender and apparent age, meaning that we should include the interaction between these predictors in our model.

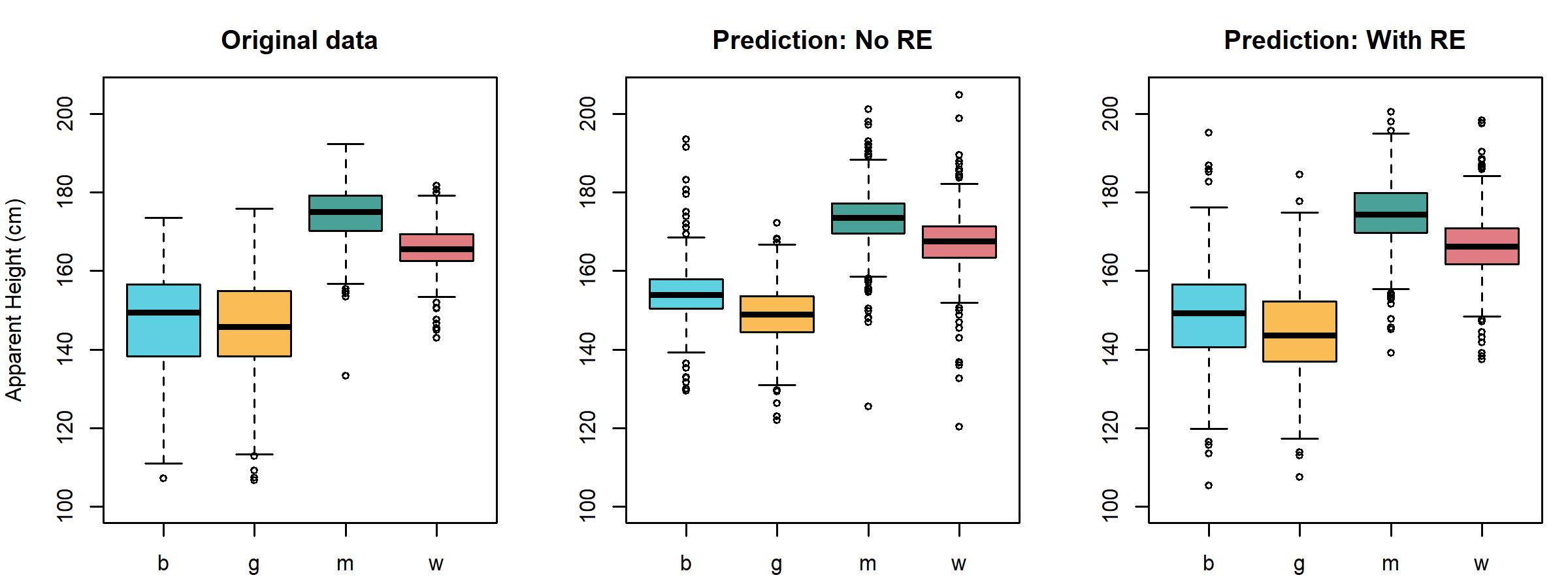


Figure 7.6: (left) Boxplots showing apparent height according to apparent speaker category (boy, girl, man, woman) for our real data. (middle) Posterior predictions made by model\_both without random effects (RE). (right) Posterior predictions made by model\_both with random effects (RE).

## 7.4 Interactions and interaction plots

We can think of a single effect representing a difference between groups/conditions/levels as a slope. For example, in the left panel in figure 7.7 we plot mean apparent height for apparent females and males at arbitrary x-axis locations. The difference in the group means is 6 cm (females 157 cm, males 163 cm). We can use any arbitrary x axis distance to calculate slopes, as long as we are consistent. However, there are obvious practical advantages to choosing to calculate these slopes over the arbitrary ‘distance’ of 1. To do this, we can imagine that we plot the first level (female) at 0 and successive levels at sequential integer values along the x axis (i.e. 1, 2, 3, …). When we do this, the line formed by joining the averages of these groups has a slope of 6 (i.e., it rises 6 cm from 0 to 1, from one group to another, for a slope of 5/1).

# main effect for apparent gender  
tapply (exp\_data$height, exp\_data$G, mean)  
## f m   
## 156.9 163.2

The plot highlighting the effect of apparent gender on apparent height in figure 7.7 is a *main effects* plot. Main effects (discussed in chapter 6) are the effects of one predictor averaged across everything else. Saying ‘averaged across everything else’ basically means we are ignoring everything else in our model. A person looking only at the middle plot in figure 7.5 would have no idea our data also investigates the effect of apparent adultness. We have ‘erased’ the differences in apparent adultness from our data by averaging across all levels of that factor. Another way to think of main effects is that they are *marginal effects*, the overall, average, *unconditional* differences between the levels of the factor. So, if someone asks you “whats the average difference in apparent height between males and females?” you can respond “about 5 cm”. However, as noted in the previous chapter, sometimes the answer is not so simple, and it starts more like: “well… it depends”. Interactions represent situations like these, where the effect of one variable depends on, or is *conditional* on, the value of some other variable. To understand interactions we need to talk about **simple effects** or **simple main effects**, the effects of a factor at one specific level of another factor. Basically, main effects show you the overall effect for a factor and simple effects show you the effects of the factor *depending* on the value of other things. In considering simple effects, we split the singe (main) effect into two simple (main) effects, presented in the second plot in figure 7.7. For example, below we calculate the effect for apparent gender independently for each level of apparent age. We see that the difference across levels of gender is about 9 cm for apparent adults (166 to 175), but only about 2 cm for apparent children (145 to 147). We can plot these simple effects independently at the same x axis positions, as shown in the second plot of figure 7.7.

# simple effect for apparent gender across apparent age  
tapply (exp\_data$height, exp\_data[,c('A','G')], mean)  
## G  
## A f m  
## a 165.6 174.5  
## c 145.0 146.9

The second and fourth plots in figure 7.7 are sometimes called **interaction plots**. Knowing how to interpret interaction plots is extremely useful because they help you quickly understand the nature of any interactions in your data. Interaction plots show you the simple main effects of one predictor across the levels of the other. They do this by having one factor vary across the x-axis and using different lines for the levels of the other factor. For example in the second plot gender varies across the x axis and different lines are used for the levels of age, and in the fourth plot age varies across the x axis and different lines are used for gender. Note however that the same information is presented in the two interaction plots in figure 7.7, it is just presented differently in each case. In fact, we can see that the y-axis values of the four line endpoints is equal across the two plots. The slopes of the lines in the second plot correspond to the differences between line endpoints in the fourth plot; one is small and one is zero. The differences between the line endpoints in the second plot are the slopes of the fourth plot, both are large but one is larger.

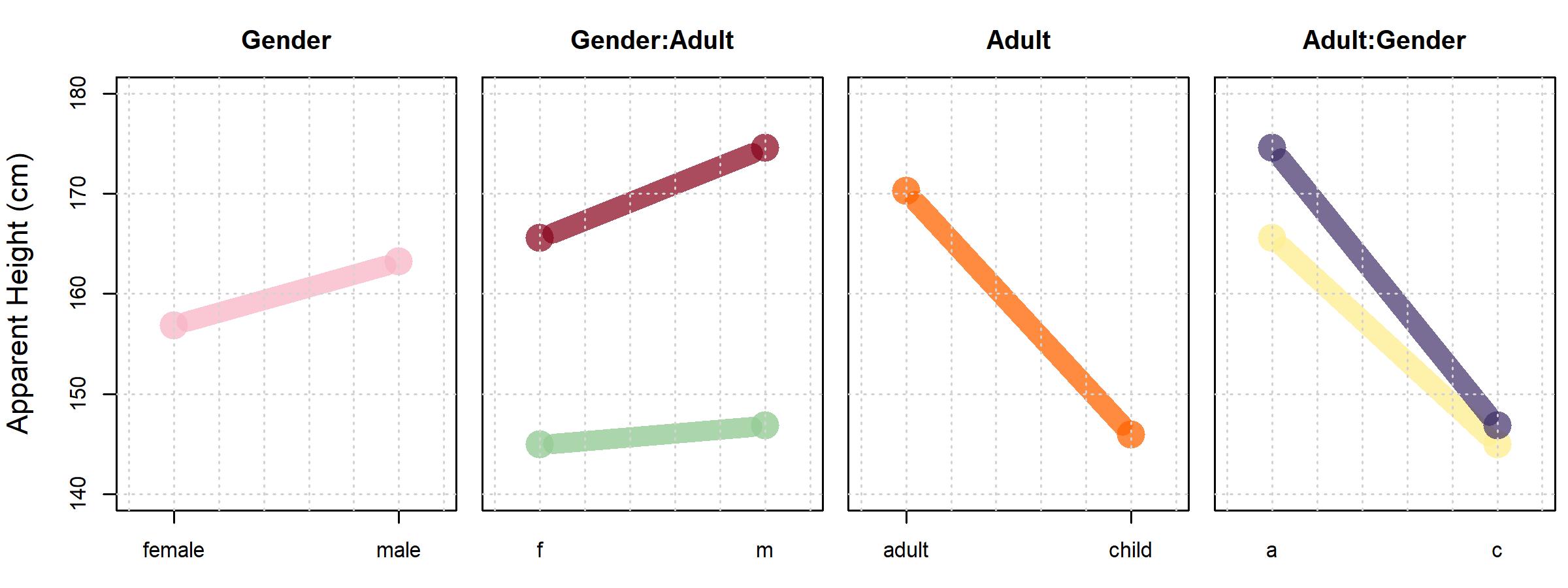


Figure 7.7: Plots showing different ways to consider our f0 data.

Imagine what would happen if we began with a main effect plot like the first plot in figure 7.7, and added/subtracted a single value to each end of the line segment. This single value could represent the main effect of some other factor, for example the main effect for apparent age. Since a single number would be added to both points on the line, the line representing the effect of apparent gender would *slide* up/down the y-axis, however, the line would not change its slope. We might imagine trying to slide the line in the first plot in figure 7.7 up and down the y axis to try to get it to match the simple effects presented in the second plot of the same figure. However, we will not be successful because these two lines do not share a slope. In order to make the lower and upper lines *not* be parallel, we need to add different values to each side of the blue line segment. In other words, we don’t slide the lines up and down in a balanced manner but move one side up more than the other. The varying effect for apparent gender at different levels of apparent age indicates that these effects *interact* in our data. Interactions between our factors are detectable visually in the form of a lack of parallel lines in interaction plots (i.e. varying slopes in the simple effects). In the absence of an interaction, all lines in such a plot should be parallel, save for variation due to random error. We can see that the lines in the second and fourth plots in figure 7.7 are *not* parallel, suggesting an interaction between apparent age and gender in our data. In the absence of an interaction, we could just answer the question “whats the average difference between apparent height between males and females?” with a number like 6 cm. In the presence of an interaction we need to consider the *conditional effects* of each predictor at the levels of the other predictor (i.e., the simple effects). So, the most important thing to remember about interpreting main effects is:

* If there are no interactions between factors, focus on main effects.
* If there are interactions between factors, you must inspect the simple effects to understand the nature of the interaction.

Let’s now consider the main effect for perceived adultness presented in the third plot in figure 7.7). We’re going to rehash some of what we just said, but we are going to adopt the *other* perspective. We can calculate the difference in apparent age based on apparent adultness and see that there is a 24 cm decrease in average apparent height based on the perceived adultness of the speaker.

# main effect for apparent age  
tapply (exp\_data$height, exp\_data$A, mean)  
## a c   
## 170.3 145.9

However, in the fourth plot in figure 7.7 we see that the simple effect for apparent adultness across levels of apparent gender are not parallel. This tells us there is an interaction between age and gender. We calculate the simple effects for apparent age at the different levels of apparent gender below:

# simple effect for apparent age across apparent gender  
tapply (exp\_data$height, exp\_data[,c('G','A')], mean)  
## A  
## G a c  
## f 165.6 145.0  
## m 174.5 146.9

And see a 21 cm in apparent height for apparent females and a 28 cm difference for apparent females. Note that the simple effects involve the same four values were present in these simple effects calculated above. The only difference is the row wise differences in the first are column wise differences in the second. For example, relative to the second plot in figure 7.7 we can say “the larger slope for adults reflects a stronger effect for apparent gender on apparent height”. Relative to the fourth plot we can say “the larger difference between the line endpoints for adults compared to children reflects a stronger effect for apparent gender on apparent height for adults”. So, we see that interactions represent conditional effects, situations when the effect of a variables varies across levels of another variable. The idea of conditional effects (interactions) *feels* complicated, but it is something we all understand intuitively. For example, how much progress will a person learning a second language make in a year? What if I told you that one group of speakers is 3 and the other is 65. You *know* that makes a difference, which is to say, that you *know* there is an interaction between the effect of time spent learning a language, and the age at which the learner begins. Anyone who understands these varying effects *understands* interactions, whether or not they know how to relate this concept to the mathematical formalisms used to implement these in regression models.

## 7.5 Investigating interactions with a model

In figure 7.6 we made posterior predictions of our data with and without the inclusion of the speaker random effects, and presented these as boxplots. We presented these same predictions as interaction plots in figure 7.8. We calculate these just as we did above save for the inclusion of our posterior predictions rather than our real data, as in:

# simple effect for apparent age across apparent gender for posterior predictions  
tapply (y\_post\_pred[,1], exp\_data[,c('G','A')], mean)  
## A  
## G a c  
## f 166.2 144.5  
## m 174.3 148.4

We can see in figure 7.8 that the predictions made using only fixed effects (middle plot) are systematically wrong, and that the problem with our predictions is that our predicted lines are parallel, while those for our real data are not. As we’ve just discussed, in the absence of interactions, interaction plots contain only parallel lines. Since our model (model\_both) does not include interaction terms it cannot represent interactions, and so is only capable of making predictions along parallel lines. This means our model is not capable of representing the pattern in our data using only fixed effects. However, we can see that when random effects are included, prediction is better. This is not surprising since the listener and speaker-specific intercept adjustments allow for deviations from the fixed effects predictions to still be captured by the model. However, if accurate prediction is only possible using random effects, that limits the generalizability of our model to new levels of the grouping variable, e.g. subject or listener. In other words, the absence of good fixed-effects prediction limits the generalizability of our model and may be a cause for concern.

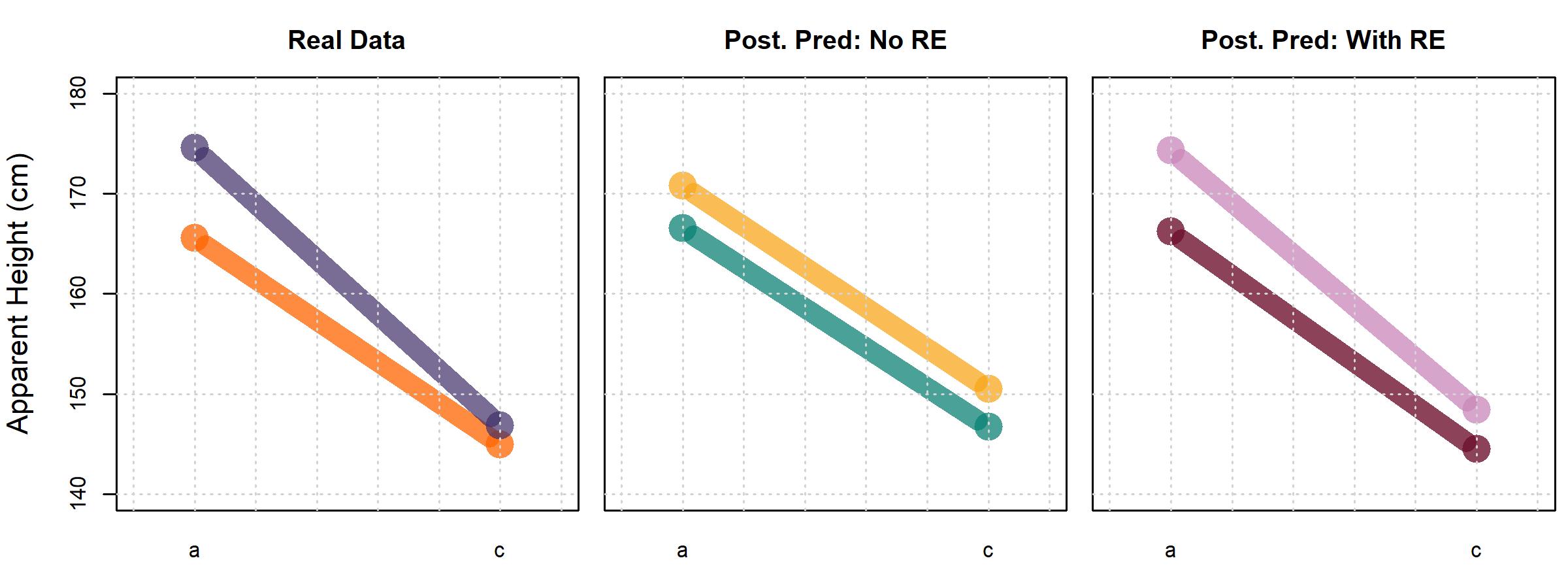


Figure 7.8: Interaction plots showing comparing our real data to posterior predictions, with and without RE (random effects). In each case the top line represents height judgments for apparent males.

Our model is a little universe we made up, and it only includes the information we included in it. This ‘universe’ only contains parallel lines because we only gave it that capability. So, the fact that our model generates parallel lines does not in any way ‘prove’ that the lines are parallel, because they were bound to be. In order to properly investigate whether the lines are parallel, the model must be built in a way that allows it to (at least potentially) represent the interactions in our data. Before fitting a model with an interaction term we want to remind you that you cannot estimate interactions between factors are not crossed: If you did not observe all combinations of factor levels, you are not in a position to talk about effects vary for those combinations. So, it is extremely important to be sure to only include interactions in your model when these predictors are *crossed*, or your model may run into all sorts of problems.

### 7.5.1 Data and research questions

We’re going to a use the very same variables from our data frame that we used in the previous model. These are:

* L: A number from 1-15 indicating which *listener* responded to the trial, being treated as a character.
* height: A number representing the *height* (in centimeters) reported for the speaker on each trial.
* S: A number from 1-139 indicating which *speaker* produced the trial stimulus.
* G: The *apparent gender* of the speaker indicated by the listener, f (female) or m (male).
* A: The *apparent age* of the speaker indicated by the listener, a (adult) or c (child).

And we’re going to add one more research question:

Q1) Are there differences in average apparent height based on apparent speaker age?

Q2) Are there differences in average apparent height based on apparent speaker gender?

Q3) Is there an interaction between apparent gender and age on apparent height? If so, what is it’s effect?

### 7.5.2 Description of our model

The model presented above (model\_both) requires only a slight tweak to include the interaction in our data. There are two ways to include interactions in R model formulas, either:

height ~ A + G + A:G + (A + G + A:G|L) + (1|S)

Or:

height ~ A \* G + (A \* G|L) + (1|S)

The first way includes an explicit interaction term, A:G. The syntax for these is X:Z for an interaction between effects X and Z, W:X:Z for a three-way interaction, and so on. The second way uses \* between our two predictors. This tells R to include those predictors, and the interactions between them. This can be much faster than specifying all interactions, but you lose control over which ones you include. For example this formula:

y ~ Z \* X \* W

Implies this one:

y ~ Z + X + W + Z:X + Z:W + X:W + Z:X:W

But cannot represent this one (since it omits the X:W interaction):

y ~ Z + X + W + Z:X + X:W + Z:X:W

Our full model specification is presented in (7.7). The model now includes an *interaction* term () that can help explain variation that cannot be explained by the independent effects of adultness and gender. This interaction term helps us model the *conditional* effect of one predictor given the other. Our model also actually contains a three-way interaction as well (), though it may not seem that way to us superficially. This is the ‘random’ effect of the interaction between and for each level of listener, (A:G|L) in our model. This is a three-way interaction because it represents the interaction between the two-way interaction of and with our listener factor: The listener dependent two-way interaction between and .

Our updated plain English description (still ignoring much of the priors) now says:

“We expect apparent height to vary according to a t distribution with a trial-specific mean parameter. That mean varies based on the apparent age and gender of speaker and the interaction between these predictors. The model includes listener dependent intercepts, age and gender effects, and their interaction, and listener dependent intercepts. Speaker random intercepts were drawn from a univariate normal distribution, and the listener random effects were drawn from a four-dimensional normal distribution.”

Before continuing, we want to address the fact that we’ve included the same structure in our ‘fixed effects’ as for our listener ‘random effects’. We will do this for all the models we fit in this book whenever possible. There are a couple of reasons for this, and we will discuss these with respect to the ‘random effect’ of apparent age by listener (i.e. the interaction). Not including this predictor is equivalent to setting it to zero for all listeners (and setting ), meaning we expect the effect for apparent age to be the same across all levels of listener. However, we saw in the previous chapter that this is not the case and that there is substantial variation in the effect for apparent age across listeners. So, omitting random effects more or less involves assuming they equal zero, which is an assumptions that may not be founded a priori. Furthermore, in chapter 6 we fit models with and without random effects and saw that the model with random effects produced a substantially wider credible intervals for the apparent age ‘fixed’ effect, reflecting increased uncertainty in the parameter estimate. In that case, assuming listener-variation in the parameter was zero when it was not resulted in an overly precise estimate of the apparent age parameter. So, we see that omitting random effects means assuming they are equa to zero *and* can affect the amount of uncertainty we get in our resulting parameter estimates. The only way to know whether a listener-dependent parameter shows substantial variation or not is to include the possibility for listener-dependent variation in your model, at least initially. Of course, the inclusion of any ‘random effect’ in your model assumes that: 1) Your model is designed in a way that the ‘random’ effect can be estimated, and 2) You have enough data to estimate those parameters. These issues will be discussed further in chapter 11.

### 7.5.3 Fitting and interpreting the model

Below we fit the our model including an interaction term. Remember that the line set\_prior("student\_t(3, 0, 12)", class = "b") sets the prior for all non-intercept ‘Population-Level’ predictors (i.e. fixed effects). This includes all of our main effects predictors and our interaction terms.

# Fit the model yourself  
# set the random seed and contrasts  
set.seed (1); options (contrasts = c('contr.sum','contr.sum'))  
  
priors = c(brms::set\_prior("student\_t(3,156, 12)", class = "Intercept"),  
 brms::set\_prior("student\_t(3,0, 12)", class = "b"),  
 brms::set\_prior("student\_t(3,0, 12)", class = "sd"),  
 brms::set\_prior("lkj\_corr\_cholesky (2)", class = "cor"),   
 brms::set\_prior("gamma(2, 0.1)", class = "nu"),  
 brms::set\_prior("student\_t(3,0, 12)", class = "sigma"))  
  
model\_interaction =   
 brms::brm (height ~ A + G + A:G + (A + G + A:G|L) + (1|S), data = exp\_data,   
 chains = 4, cores = 4, warmup = 1000, iter = 5000, thin = 4,   
 prior = priors, family = "student")

# Or download it from the GitHub page:  
model\_interaction = bmmb::get\_model ('7\_model\_interaction.RDS')

We are primarily interested in discussing interactions, but we want to spend a moment looking at the listener random effects in our model, since there are now several.

# inspect fixed effects  
bmmb::short\_summary (model\_interaction)  
## Formula: height ~ A + G + A:G + (A + G + A:G | L) + (1 | S)  
##   
## Group-Level Effects:  
## ~L (Number of levels: 15)  
## Estimate Est.Error l-95% CI u-95% CI  
## sd(Intercept) 4.24 0.79 2.96 6.03  
## sd(A1) 4.47 0.85 3.13 6.42  
## sd(G1) 2.10 0.49 1.37 3.22  
## sd(A1:G1) 1.34 0.36 0.76 2.13  
## cor(Intercept,A1) -0.71 0.13 -0.90 -0.38  
## cor(Intercept,G1) -0.20 0.22 -0.59 0.24  
## cor(A1,G1) -0.24 0.21 -0.62 0.20  
## cor(Intercept,A1:G1) 0.17 0.23 -0.29 0.58  
## cor(A1,A1:G1) -0.02 0.23 -0.46 0.43  
## cor(G1,A1:G1) -0.34 0.24 -0.74 0.18  
##   
## ~S (Number of levels: 139)  
## Estimate Est.Error l-95% CI u-95% CI  
## sd(Intercept) 2.36 0.31 1.79 2.99  
##   
## Population-Level Effects:  
## Estimate Est.Error l-95% CI u-95% CI  
## Intercept 158.46 1.12 156.22 160.62  
## A1 10.78 1.21 8.39 13.15  
## G1 -2.87 0.60 -4.07 -1.72  
## A1:G1 -1.64 0.41 -2.44 -0.81  
##   
## Family Specific Parameters:  
## Estimate Est.Error l-95% CI u-95% CI  
## sigma 5.01 0.16 4.70 5.34  
## nu 3.44 0.33 2.87 4.15

We can see in the short model summary above that we’ve estimated four standard deviation terms for our listener-dependent predictors. These correspond to (sd(Intercept)), (sd(A1)), (sd(G1)), and (sd(A1:G1)). We’ve also estimated six correlation terms. Recall that our random effects are drawn from a multivariate normal distribution. We have four terms which means this four-dimensional distribution is associate with a 4x4 correlation matrix and 16 correlation terms. Of the correlations, four (on the main diagonal) equal one and can be ignored. This leaves 12 terms. Of these only half are unique because the correlation of and equals the correlation of and . This is how we end up with six estimated correlations for our four listener-dependent predictors.

We can now focus on the fixed effects (Population-Level Effects), of which there are four: An intercept, two main effects terms and an interaction. The interaction term is just another element of your prediction equation intended to help explain variation that can’t be predicted by the independent effects of the other predictors in the model. If there is no interaction in your data, then the value of the interaction terms will be zero and your model will look just as if you had not even included the interaction at all. A look at the model output above indicates that our interaction is much smaller than the age effect, but not much smaller than the gender effect. This suggests it has the potential to influence our conclusions about gender more than our conclusions about age.

Why is there only a single interaction? Basically for the same reason that we can only estimate a single effect for apparent age and gender (i.e., linear dependence, discussed in section X). The number of levels of a factor that you can estimate is generally one fewer than the number of levels in the factor. For interaction terms, the number of parameters you can estimate is equal to (number of levels of factor A - 1)x(number of levels of factor B - 1) for two factors. Since each of our factors have two levels, we can only estimate one parameter, .

The A1:G1 term can be read aloud as “the effect of A1 given (conditional on) the level of G1”. For individual factors, our models cannot estimate A2 because -A1 = A2. For interactions between factors with two levels, the model cannot estimate A2:G1 because -A1:G1 = A2:G1. For interactions between factors with two levels, every time you want to change A1 or G1 to A2 or G2 you flip the sign of the A1:G1 term instead. As a result, going left to right and changing one level at a time we get A1:G1 = -A2:G1 = A2:G2 = -A1:G2, which is precisely why we can’t estimate more than one interaction for this model. Why does this happen? Looking at how interaction plots are built up out of main effects and interactions may help illustrate the reason for this. Figure 7.9 presents an example of how our interaction plot can be built up from the components in our fixed effects above, and is very useful for understanding the geometry of main effects and interactions. In (a) we see that the intercept lifts up a line from 0 to the level of the overall mean apparent height of 158 cm. Then, the effect for A1 (adult, +11 cm) and -A1 (children, -11 cm) are added to this value, causing the separation between the lines indicating a ‘main effect’ for adultness. Notice that the lines are parallel with respect to each other, and also parallel to the x-axis. This is because we have not added an effect for gender, which will be indicated as a slope along the x axis. In (b) we see the addition of the gender effect: G1 (female, -2.8 cm), and -G1 (male, 2.8 cm). Notice that the effect is added to one end of each line segment and subtracted from the other end of the line segment. Panel (c) shows the addition of all of the effects shown in the top row, the result of a model that includes both main effects but the same non-zero slope for both lines (i.e. no interaction yet). In (e) we add the gender interaction A1:G1, which was -1.6 cm.

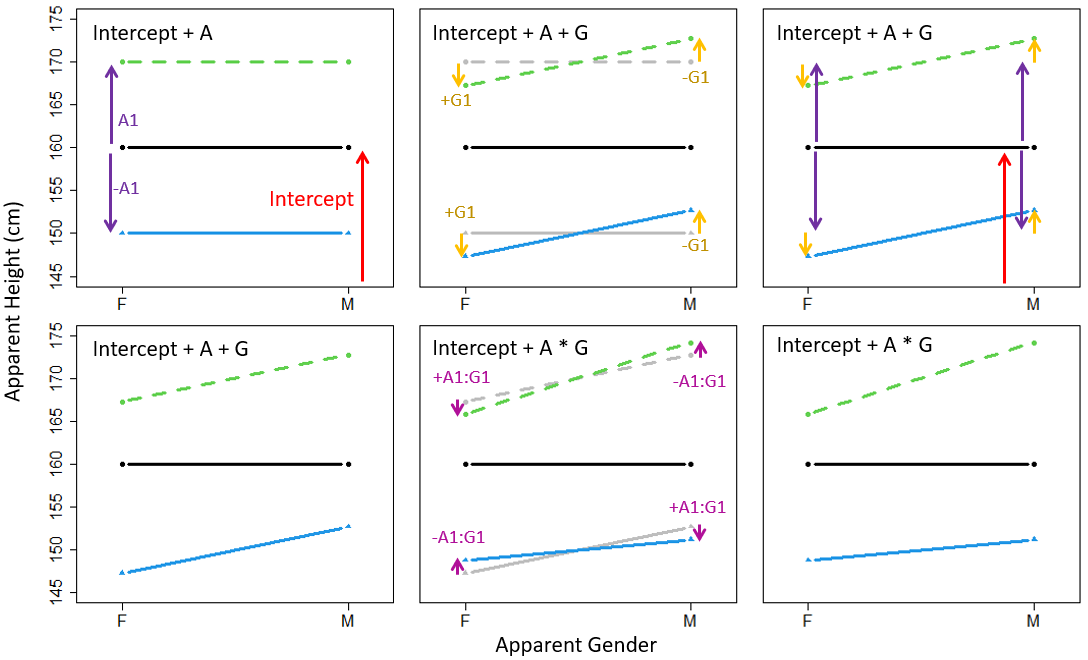


Figure 7.9: An interaction plot built up from its individual components. Need to add letters to plots.

Notice that the interaction effect has different signs when added to the same line, and also for each end of each line segment. This means that the interaction terms *cancel out* at each end of the line segments across lines, but also for different endpoints of the same line. We can see that beginning with the parallel lines in (d), the interaction in (e) constitutes an equal ‘pinch’ of one end of the lines and a spreading of the other side of the lines. This equal pinch is necessary in order to maintain the same average slope across both lines (the gender main effect) and the same average spacing between both ends of the lines (the age main effect). This is effectively why we get only a single interaction term in a model with this fixed effects structure. Recovering the predicted group means using the model coefficients is straightforward, but can be tedious for larger models: We must now either add or subtract the value of the interaction term (A1:G1) from each group as determined by the combination of main effects. We can easily determine which to do for this model because the sign on the interaction term is the product of the signs on the relevant ‘main effects’ terms. So, two positive or negative main effects result in a positive interaction term, while mismatched positive and negative main effects result in a negative interaction. Below, we use our wrapper the hypothesis function provided in brms, short\_hypothesis, to reconstruct expected group means given the model fixed effects. The fourth hypothesis we are testing below is the simplest to understand, so we will start there. This hypothesis asks whether the sum of the Intercept, A1 (the effect for adultness), G1 (the effect for femaleness), and A1:G1 (the interaction of adult and female), is equal to zero. Because of the specific parameters we combined, this results in the comparison of the ‘woman’ mean height to zero. The first hypothesis we are testing below says Intercept + -A1 + -G1 + A1:G1 = 0. Since we have flipped the sign on A1 and G1, these now correspond to the effects of ‘child’ and ‘male’ (i.e. boy). Since both A1 and G1 are negative, we flip the sign on the interaction term twice and it remains positive. Note that for all of our hypotheses below, the sign on the interaction terms always depends on the signs of the corresponding main effects terms.

# intercept, boys, girls, men, women  
means\_pred\_interaction =   
 bmmb::short\_hypothesis (model\_interaction,   
 c("Intercept + -A1 + -G1 + A1:G1 = 0",  
 "Intercept + -A1 + G1 + -A1:G1 = 0",  
 "Intercept + A1 + -G1 + -A1:G1 = 0",  
 "Intercept + A1 + G1 + A1:G1 = 0"))  
  
# actual data means  
tapply (exp\_data$height, exp\_data$C, mean)  
## b g m w   
## 146.9 145.0 174.5 165.6  
  
# predictions with no interaction term  
means\_pred[,1]  
## [1] 150.9 146.5 170.8 166.4  
  
# predictions with interaction term  
means\_pred\_interaction[,1]  
## [1] 148.9 146.4 173.8 164.7

By comparing our recovered group means to we can see that the inclusion of an interaction term allows our model to represent the differing effect for gender across levels of age (and age across levels of gender). Below we print the estimates of the ‘fixed’ effects in the model so we can focus on those. If you fit a model like this and are having trouble interpreting it, we would encourage you to write the coefficient values down and try to recreate expected effects using pen and paper, focusing on the decomposition of values provided by the regression model. We’ve previously warned against averaging parameters first and combining them second, however, this approach can be useful for a quick investigation of our model parameters. An example of this process will be given below using rounded parameter estimates to make the discussion easier to follow.

round (brms::fixef (model\_interaction))  
## Estimate Est.Error Q2.5 Q97.5  
## Intercept 158 1 156 161  
## A1 11 1 8 13  
## G1 -3 1 -4 -2  
## A1:G1 -2 0 -2 -1

First, we will recover the simple effects for apparent age at the levels of apparent gender. The average apparent height across all groups is 158 cm. There is an age-based 11 cm deviation from the intercept, meaning the difference between apparent adults and children is 22 cm (11 \* 2). This means that, overall, the averages for apparent adults and children are about 169 and 147 cm (158 ± 11). However, the A1:G1 interaction is -2 cm. This means that when the speaker was female (G1), the affect for adultness decreased from 11 cm to 9 cm (11 - 2, A1+A1:G1), meaning the difference across women and girls was 18 cm (2\*9). In contrast, when the speaker was a male (-G1), the affect for adultness increased from 11 cm to 13 cm (11+2, A1-A1:G1), meaning the difference between boys and men was about 26 cm. We could instead consider the simple effects for apparent gender given apparent age. There is a gender-based 3 cm deviation from the intercept, meaning the difference between apparent adults and children is -6 cm (-3 \* 2). This means that, overall, the female and male averages are about 155 and 161 cm respectively (158 ± -3). However, the A1:G1 interaction is -2 cm. This means that when the speaker was an adult (A1) the effect for gender increased in magnitude from -3 cm to -5 cm (-3 + -2, G1+A1:G1), resulting in a difference between groups of 10 cm. Finally, when the speaker was a child (-A1), the effect for gender decreased in magnitude from -3 cm to 1 cm (-3+2, G1-A1:G1), resulting in a difference in groups of 2 cm based on apparent gender. Below we carry out these same calculations outlined above using our hypothesis function, with slightly different results due to the lack of rounding. In each case, we take the main effect and add or subtract the interaction in order to estimate the simple effects.

# intercept, boys, girls, men, women  
simple\_effects =   
 bmmb::short\_hypothesis (model\_interaction,   
 c("A1 + A1:G1 = 0",  
 "A1 - A1:G1 = 0",  
 "G1 + A1:G1 = 0",  
 "G1 - A1:G1 = 0"))  
  
# predictions with interaction term  
simple\_effects  
## Estimate Est.Error Q2.5 Q97.5 hypothesis  
## H1 9.142 1.2792 6.639 11.5977 (A1+A1:G1) = 0  
## H2 12.424 1.2707 9.938 14.9913 (A1-A1:G1) = 0  
## H3 -4.514 0.6174 -5.731 -3.3177 (G1+A1:G1) = 0  
## H4 -1.232 0.8192 -2.860 0.3919 (G1-A1:G1) = 0

### 7.5.4 Assessing model fit: Bayesian

We can asses the fit for our model including interaction terms by making more posterior predictions with our new model. We will focus on the average predictions for each data point since our objective is to see to what extent the fixed effects can represent the patterns in the data.

y\_post\_pred\_int = predict (model\_interaction)  
y\_post\_pred\_no\_re\_int = predict (model\_interaction, re\_formula = NA)

In figure 7.10 we compare our data, the predictions of our original model, and the predictions of our model that includes interactions (both models with and without random effects). Whereas the model with no interactions enforced parallelism on the simple effects, our new model is able to capture the interaction in our data by representing different effects for gender at different levels of age (and vice versa). Although both of our models are able to capture the pattern in the data when random effects are included, only one of these is able to do so with its fixed effect structure.

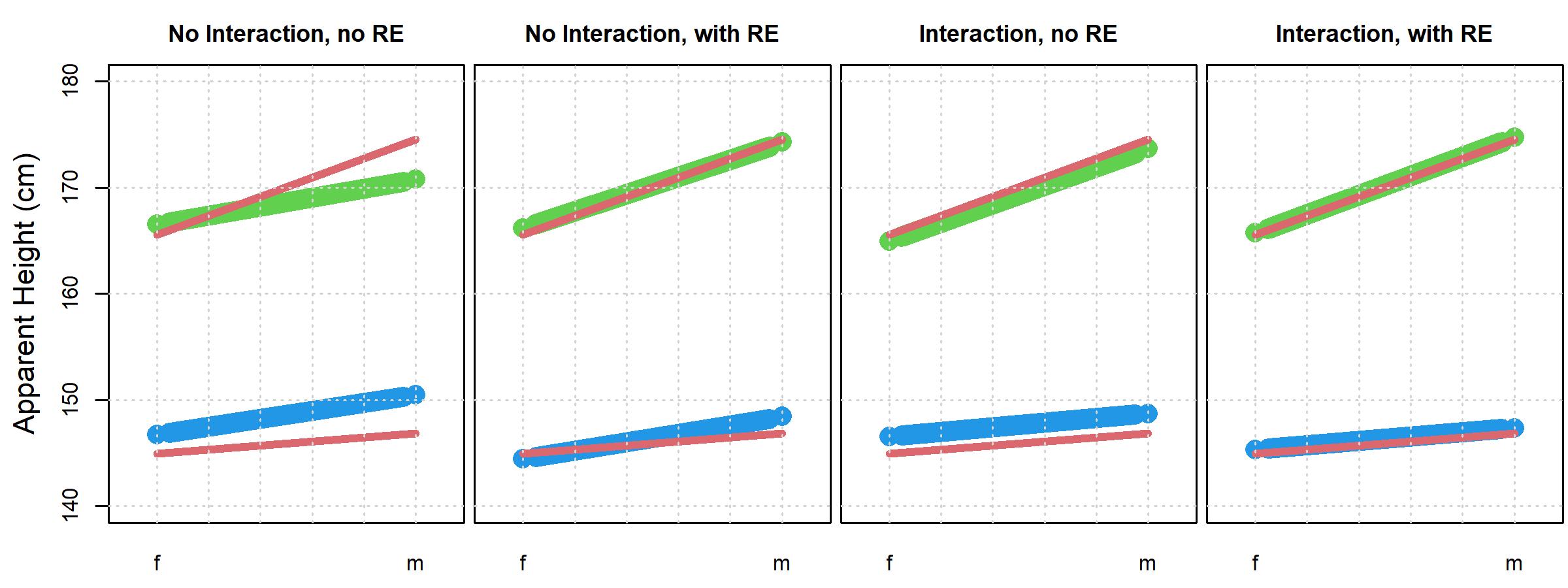


Figure 7.10: Interaction plots comparing veridical apparent heights (red) to different posterior predictions (blue and red lines). Upper lines are apparent adults and lower lines are apparent children for apparent females (f) and males (m). Posterior predictions with no interaction come from model\_both and predictions with interactions come from model\_interaction. With and no RE (random effects) refer to whether the predictions were made using only fixed effects or fixed and random effects.

We can compare the models with and without the intercept using leave one out cross-validation, discussed in section 6.X.

model\_both = brms::add\_criterion (model\_both, criterion="loo")  
model\_interaction = brms::add\_criterion (model\_interaction, criterion="loo")

The comparison suggests a large difference between the models, but also substantial uncertainty regarding the difference.

brms::loo\_compare (model\_both, model\_interaction)  
## elpd\_diff se\_diff  
## model\_interaction 0.0 0.0   
## model\_both -29.1 11.2

To some extent the difference may appear to be smaller than it is because our random effects do a good job of explaining listener and speaker-dependent deviations from the expectations set out by the fixed effects. Explaining things with ‘random’ effects is less useful for understanding the general properties in your data. If we were to replicate this experiment with new speakers and listeners we would expect all the random effects to change. However, if the relations between apparent age and gender and apparent height expressed in our model are ‘true’, we would expect the ‘fixed’ effects in the replication to be about the same. For this reason it is extremely useful to consider fixed effects prediction in addition to the predictive power of a complete model including both fixed and random effects. To quantify how well our models fit our data we can also use **Bayesian** . There are potentially several ways to define in a Bayesian context. The account of Bayesian presented here is a summary of the proposal outlined in Gelman (cite). In addition, the discussion of to follow assumes that your model involves either Gaussian or t-distributed errors and a continuous dependent variable. is a measure of model fit that breaks down the total variance in your dependent variable into two parts: 1) The variance that your model can explain, and 2) the variance that your model *cannot* explain, the residual error. This is shown in (7.8).

is the ratio of the explained variance to the total variance, as seen in (7.9). Since the explained variance cannot be less than zero nor can it be greater than the total variance, values of range from zero to one. As a result, can be interpreted as the proportion of variance in the dependent variable that is explained by the model

To make a *Bayesian* , we need to incorporate the posterior distribution of parameter values into our estimation. This is done, in part, by calculating a value of for every individual set of posterior samples, as shown in (7.10). For every posterior sample of parameter values (from total samples) we calculate: 1) The variance in the predicted values () of the dependent variable for our data points (), and 2) the variance of the residual error for our data points ().

The residual error is the difference between predicted () and observed () values for each data point, shown in (7.10).

This results in a *distribution* of values, unlike more ‘traditional’ estimates of that result in point estimates. The brms package includes a function called bayes\_R2 that helps you calculate Bayesian for brm models. However, we are going to use the r2\_bayes function from the bmmb package instead because this function allows you to modify (or omit) the fixed effects included in the prediction. Below we calculate for the models we fit this chapter, including all random effects.

r2\_both = r2\_bayes(model\_both)  
r2\_interaction = r2\_bayes(model\_interaction)

We can inspect the output and see that they have very similar values. This is not surprising given their similarities seen in figure 7.10.

r2\_both  
## Estimate Est.Error Q2.5 Q97.5  
## [1,] 0.7755 0.005566 0.7641 0.7858  
  
r2\_interaction  
## Estimate Est.Error Q2.5 Q97.5  
## [1,] 0.7802 0.005347 0.7694 0.7901

Just as we can omit the random effects when we use predict, we can omit random effects from our calculations of by setting re\_formula = NA.

r2\_both\_no\_re = r2\_bayes(model\_both, re\_formula = NA)  
r2\_interaction\_no\_re = r2\_bayes(model\_interaction, re\_formula = NA)

Now a comparison reveals a more substantial difference in explanatory power between the models.

r2\_both\_no\_re  
## Estimate Est.Error Q2.5 Q97.5  
## [1,] 0.503 0.07258 0.3411 0.6216  
  
r2\_interaction\_no\_re  
## Estimate Est.Error Q2.5 Q97.5  
## [1,] 0.5732 0.05909 0.4389 0.6678

We would rather explain as much variation as possible using fixed as opposed to random effects. In addition, the interactions ‘makes sense’ given our domain knowledge. What we mean by the is that we know that the difference in veridical height between boys and girls 10-12 years old is inconsistent and may be about zero on average. So, we should not be too surprised if the effect for apparent gender on apparent height is smaller for children. However, there is a larger and more consistent difference in average veridical height between adult males and females. As a result, we do expect a meaningful effect for apparent gender on the apparent heights of adults. For these reasons, it would be surprising if there were *not* an interaction, and we prefer the model with the interaction despite the only middling support provided by our cross-validation model comparison.

## 7.6 Answering our research questions

The results (based on model\_interaction) indicate that the average apparent height across all speaker groups (i.e. the intercept) was 158.5 (s.d. = 1.12, 95% C.I = [156.22, 160.62]). We also found an average effect of 10.8 cm for apparent speaker age (s.d. = 1.21, 95% C.I = [8.39, 13.15]) and -2.9 cm for apparent speaker gender (s.d. = 0.6, 95% C.I = [-4.07, -1.72]). In addition, we found an interaction between the effects of apparent age and apparent gender on apparent heights (mean = -1.64, s.d. = 0.41, 95% C.I = [-2.44, -0.81]). The result of these effects are that adults were perceived as taller than children and males were perceived as taller than males. However, the difference in apparent height due to gender was larger for adults than for children (and the effect for age was larger for males than for females). Figure 7.11 presents the model fixed effects other than the intercept (whose value is too large to plot in this range). The fixed effects all have 95% credible intervals that do not overlap with zero. Historically, this sort of result has been thought of as indicating that the effect is ‘real’, as in not equal to zero. In contrast, when the 95% (or some other) credible interval of an effect overlaps with zero, researchers will often conclude that this effect is not ‘real’, i.e. that it is exactly equal to zero. Gelman (cite) notes that effects are rarely exactly equal to zero, and that an inability to conclusively distinguish the value of an effect from zero is not the same thing as thinking that the most probable value of an effect is exactly zero. As a result, rather than focusing on whether any given interval crosses zero or not, Gelman (cite) recommends thinking about data analysis in terms of avoiding **type S** and **type M** errors. Type S (sign) errors occur when you think an effect is negative but its actually positive (or vice versa). A type M (magnitude) error is when you think an effect is large but its actually small (or vice versa). Focusing on type M and S errors rather than whether an interval crosses zero forces researchers to focus on the magnitude of the predictor, its practical importance, and the uncertainty involved in the estimate. That being said, if a parameter has a ‘large’ magnitude (in the context of the dependent variable) and a ‘small’ amount of uncertainty (relative to the magnitude), this predictor is likely to be important for understanding the dependent variable in the context of your model.

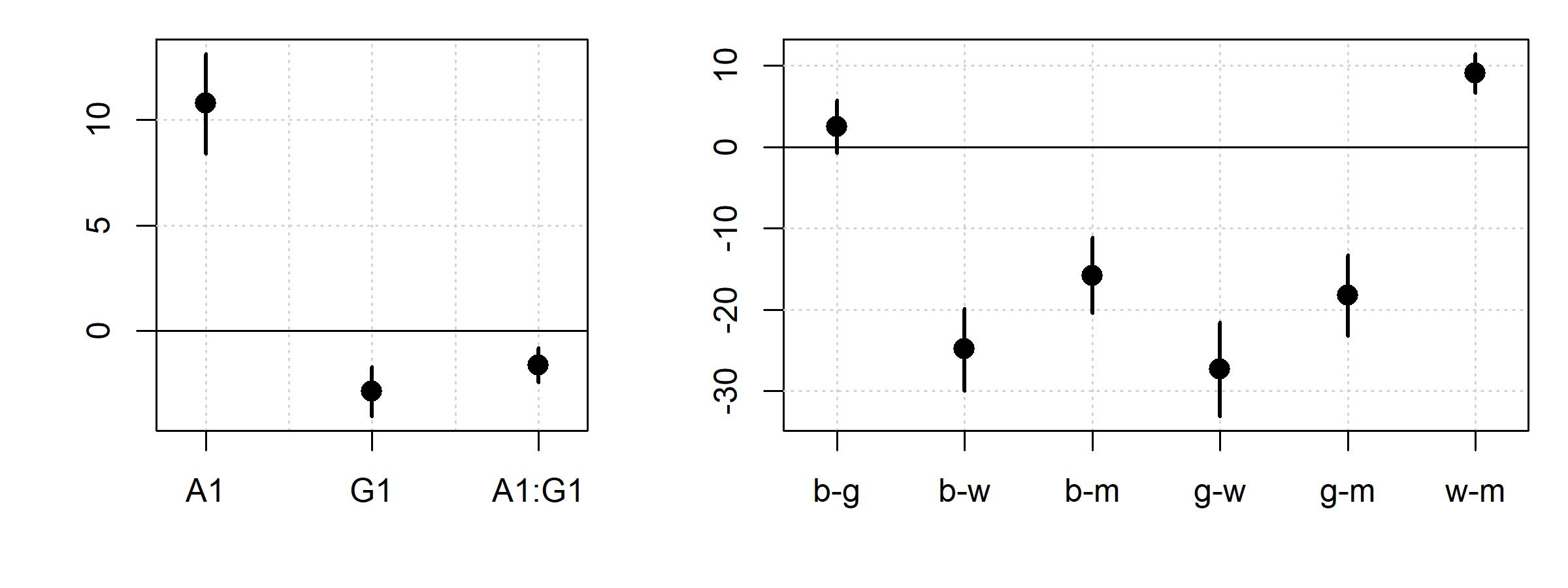


Figure 7.11: (left) Fixed effect means and 95% credible intercals (minus the intercept) for model\_interaction. (right) Means and 95% credible intercals for the pairwise differences between predicted means for different apparent speaker categories (boys, girls, men and women).

* The reference category mean is the ‘Intercept’ in the model.
* The value of the coefficients of any non-intercept group are equal to group mean - Intercept (reference group mean).
* To recover the mean estimate for any non-intercept group, we add group effect + Intercept (reference group mean).

Figure 7.11 also presents the pairwise difference between all four apparent speaker groups (boys, girls, women, and men). The difference between pairs of group means is calculated by extracting the samples from the group means we reconstructed in section 7.X (means\_pred\_interaction). The result of this is a matrix with four columns, each representing a group effect, with individual posterior samples differing along rows. To find the difference between group means between each pairs of groups, we subtract the columns representing the group effects and inspect the distribution of the differences. This process is shown below and the distribution of pairwise group differences in presented in the right plot of figure 7.11.

C = attributes(means\_pred\_interaction)$samples  
  
pairwise\_diffs = cbind("b-g"=C[,1]-C[,2],"b-w"=C[,1]-C[,3],"b-m"=C[,1]-C[,4],  
 "g-w"=C[,2]-C[,3],"g-m"=C[,2]-C[,4],"w-m"=C[,3]-C[,4])  
  
pairwise\_diffs\_summary = posterior\_summary(pairwise\_diffs)

We can think about these differences in the context of our type M and S errors. All of the group differences (save for boy-girl) have values of about 10 cm (or more), which is a meaningful (and noticeable) difference in human height. In addition, the credible intervals around these differences are small relative to their magnitudes. From this we can conclude that these differences are unlikely to *really* by the opposite sign (a type S error). In addition, since the group differences (save for boy-girl) are mostly large and the credible intervals do not go anywhere near zero, these differences seem unlikely to actually be some small and insignificant value near zero (a type S error). But what about the boy-girl difference? The 95% credible interval of the difference between groups overlaps with zero. Does this mean there is no difference between the groups, i.e.  and ? No, it just means that 0 is a value that is not so implausible for the difference between groups. The most likely difference is still 2.5. Consider a replication that found a group difference with a mean and 95% credible interval of 2.6 and [0.1, 5.1] respectively, rather than our values of 2.5 and [-0.8, 5.7]. Would such a difference constitute a successful replication or an opposite finding? If we focus on whether intervals cross zero to decide if they are ‘real’, we might decide that these experiments do not agree: We found no difference in our first experiment but the replication *did* find a difference. In contrast, if we focus on the sign and magnitude of the effects and the uncertainty around their estimates, the two experiments have extremely compatible results and constitute a successful replication.

# summary of difference between boy and girl means  
pairwise\_diffs\_summary[1,]  
## Estimate Est.Error Q2.5 Q97.5   
## 2.4643 1.6384 -0.7837 5.7199

We might wonder, under what conditions can we accept that an effect or group difference *is* actually equal to zero. If the 95% (or some other high number) credible interval is relatively narrow and (more or less) centered around zero, the parameter value may very well be zero. It may also be a small number that has no practical effect on your dependent variable so that it doesn’t really matter what it is even if its not exactly equal to . In either case, a coefficient with a narrow interval (roughly) centered at zero may not have an important role in the way you interpret your model in order to tell the ‘story’ of your data. However, it is extremely important to keep in mind that under a different parametrization of your model, or given a slightly different experimental design, it is possible that an equivalent parameter may exhibit a clearly non-zero value. In our case, we will conclude that this difference in apparent height between boys and girls is likely to be ‘real’ and likely reflect stereotypes that male speakers are generally taller than female speakers, whether or not this is actually true for speakers of this age. We might also suggest that if we want to know more about the precise value of this difference, future research might investigate this question with a larger number of speakers, listeners, or repetitions. By increasing the sample size, researchers would narrow the parameter likelihoods (discussed in section 2.X) thereby narrowing their posterior credible intervals (given new data that’s roughly the same as the data we have).

## 7.7 ‘Traditionalists’ corner

In traditionalists corner, we’re going to compare the output of brms to some more ‘traditional’ approaches. We’re not going to talk about the traditional models in any detail, the focus of this section is simply to highlight the similarities between different approaches, and to point out where to find equivalent information in the different models. If you are already familiar with these approaches, these sections may be helpful. If not, some of the information provided here may not make much sense, although it may still be helpful. If you want to know more about the statistical methods being discussed here, please see section X for a list of suggested background reading in statistics.

### 7.7.1 Bayesian multilevel models vs. lmer

This is going to be a short one. The main shortcoming when it comes to using lmer to fit models including factors with many levels is that there is no easy way to compare group effects within the model. For example, we can fit the model below which encodes the difference between each group mean and the overall mean.

lmer\_four\_groups = lme4::lmer (height ~ C + (C|L) + (1|S), data = exp\_data)  
summary (lmer\_four\_groups)$coefficients  
## Estimate Std. Error t value  
## (Intercept) 158.344 1.040 152.184  
## C1 -9.543 1.301 -7.338  
## C2 -11.126 1.332 -8.354  
## C3 14.604 1.371 10.654

The results provided by lmer are very similar to those provided by brm:

brms::fixef (model\_four\_groups)  
## Estimate Est.Error Q2.5 Q97.5  
## Intercept 158.581 1.083 156.41 160.685  
## C1 -9.318 1.310 -11.87 -6.682  
## C2 -12.082 1.494 -15.07 -9.110  
## C3 15.286 1.327 12.59 17.950

However, it is not as straightforward to test for differences between parameter values in our lmer models. The same could be said for the combination of parameters as is necessary for the investigation of group means (section 7.X) and simple effects (7.X). The investigation of differences and combinations of parameters is definitely possible in lmer for many cases, however this will involve either the re-parametrization of the model or the use of additional packages that allow for these sorts of questions to be asked. This can be contrasted with the ease and consistency of inspecting posterior distributions in our brms models.

* Is the variable on a ratio or interval scale? This is a prerequisite for a quantitative value to be used as a dependent variable. An interval scale means that distances are meaningful, and a ratio scale means that 0 is meaningful.
* Is the underlying value continuous? Many variables are discrete in practice due to limitations in measurement. However, if the underlying value is continuous (e.g., height, time) then this can motivate treating the measurement as a quantitative dependent variable since fractional values ‘make sense’. For example, even if you measure time only down to the nearest millisecond, a value of 0.5 milliseconds is possible and interpretable. In contrast, a value of 0.5 people is not.
* Are there a large number (>50) of possible values the measured variable can take? For example a die can only take on 6 quantitative values, which is not enough.
* Are most/all of the observed values far from their bounds? Human height does not really get much smaller than about 50 cm and longer than about 220 cm, so it is technically bounded. However, in most cases our observations are expected to not be clustered at the boundaries.

rbernoulli = function (n,p) rbinom(n,1,p)

# a single trial, probability of 0.5  
rbernoulli (1,.5)  
## [1] 0  
  
# ten single trials, probability of 0.5  
rbernoulli (10,.5)  
## [1] 1 0 1 0 1 1 1 0 1 0

|  | Estimate | Est.Error | Q2.5 | Q97.5 | hypothesis |
| --- | --- | --- | --- | --- | --- |
| H1 | 12.54 | 2.26 | 8.16 | 16.94 | (2\*(A1+A1:G1)) = 0 |
| H2 | 18.52 | 2.46 | 13.56 | 23.44 | (2\*(A1-A1:G1)) = 0 |
| H3 | -4.03 | 1.56 | -7.18 | -0.88 | (2\*(G1+A1:G1)) = 0 |
| H4 | 1.95 | 1.74 | -1.53 | 5.29 | (2\*(G1-A1:G1)) = 0 |
| H5 | -4.21 | 1.21 | -6.54 | -1.82 | (2\*(vtl:A1+vtl:A1:G1)) = 0 |
| H6 | -1.47 | 1.05 | -3.60 | 0.66 | (2\*(vtl:A1-vtl:A1:G1)) = 0 |
| H7 | -0.97 | 1.09 | -3.13 | 1.17 | (2\*(vtl:G1+vtl:A1:G1)) = 0 |
| H8 | 1.77 | 1.04 | -0.27 | 3.91 | (2\*(vtl:G1-vtl:A1:G1)) = 0 |

### 9.9.1 A word on causality

* causal but still observational
* causality is a scientific claim not statistical

We’ve specifically tried to avoid causal language so far, saying things like ‘is associated with a change’ rather than ‘causes a change’. In earlier chapters, this was because we were