

# **The brms Book: Applied Bayesian Regression Modelling Using R and Stan (Early Draft)**

Paul-Christian Bürkner

2024-10-30

# Table of contents

<b>Preface</b>	<b>3</b>
<b>1 Linear Models</b>	<b>4</b>
1.1 Setup . . . . .	4
1.2 Introduction . . . . .	4
1.3 Gaussian linear models . . . . .	5
1.3.1 Why do we need sampling? . . . . .	10
1.3.2 Variations of posterior predictions . . . . .	13
1.3.3 Posterior predictive checks . . . . .	15
1.3.4 Adding more predictors . . . . .	17
1.3.5 Outlier analysis . . . . .	21
1.4 Fat-tailed linear models . . . . .	23
1.4.1 Posterior predictive checks . . . . .	27
1.5 Skewed linear models . . . . .	28
1.6 Gaussian linear models: Analytic vs. MCMC . . . . .	31
1.6.1 Reproducing the analytic results with MCMC . . . . .	33
1.7 Summary . . . . .	36
<b>2 Model Comparison</b>	<b>38</b>
2.1 Setup . . . . .	38
2.2 Introduction . . . . .	38
2.3 Prediction vs. Explanation . . . . .	39
2.4 Measuring Predictive Performance . . . . .	40
2.5 Absolute predictive performance . . . . .	41
2.5.1 Measures of Explained Variance . . . . .	41
2.5.2 Measures of Squared Errors . . . . .	46
2.6 Relative predictive performance . . . . .	49
2.6.1 Likelihood Density Scores . . . . .	52
2.7 Out-of-sample predictions . . . . .	56
2.7.1 Approximate leave-one-out cross-validation . . . . .	59
2.7.2 Pareto smoothed importance sampling . . . . .	61
2.7.3 Correcting the PSIS approximation . . . . .	67
2.7.4 Leave-one-out $R^2$ . . . . .	70
2.8 Prior predictive performance . . . . .	72
2.8.1 Prior predictive checks . . . . .	72

2.8.2	Marginal likelihood-based metrics . . . . .	76
2.9	Model averaging . . . . .	83
2.9.1	Weights from marginal likelihoods . . . . .	84
2.9.2	Weights from ELPD scores . . . . .	85
2.9.3	Weights from stacking of predictive distributions . . . . .	86
2.10	In-distribution vs. out-of-distribution . . . . .	88
2.11	Summary . . . . .	90
<b>3</b>	<b>Generalized Linear Models</b>	<b>91</b>
3.1	Setup . . . . .	91
3.2	Introduction . . . . .	91
3.3	GLMs for lower-bounded responses . . . . .	92
3.3.1	Modeling log counts . . . . .	92
3.3.2	Log transform both response and baseline counts . . . . .	98
3.3.3	Lognormal models . . . . .	101
3.3.4	Poisson models . . . . .	102
3.3.5	Negative binomial models . . . . .	105
3.3.6	More adventures into model comparison . . . . .	111
3.4	GLMs for double-bounded responses . . . . .	114
3.4.1	Adding interactions . . . . .	120
3.4.2	Centering Predictors . . . . .	123
3.4.3	Even More Interactions . . . . .	129
3.4.4	Changing the Link Function . . . . .	134
3.4.5	Binomial Models . . . . .	138
3.5	Summary . . . . .	139
<b>4</b>	<b>Linear multilevel models</b>	<b>141</b>
4.1	Setup . . . . .	141
4.2	Introduction . . . . .	141
4.3	Complete pooling . . . . .	144
4.4	Partial pooling: Varying intercepts . . . . .	146
4.5	Partial pooling: Varying intercepts and slopes . . . . .	157
4.6	Predicting coefficients of new levels . . . . .	163
4.7	Priors on correlation matrices . . . . .	168
4.8	Simulation study of Type 1 errors . . . . .	170
4.9	No pooling . . . . .	175
4.10	Models with more than two levels . . . . .	177
4.11	Different covariance matrices by group . . . . .	180
4.12	Some benefits of multilevel models . . . . .	183
<b>5</b>	<b>Generalized linear multilevel models</b>	<b>184</b>
5.1	Setup . . . . .	184
5.2	Introductions . . . . .	184

5.3	GLMMs for lower-bounded count responses . . . . .	184
5.3.1	Modeling overdispersion . . . . .	187
5.3.2	Adding varying slopes . . . . .	194
5.4	GLMMs for lower-bounded continuous responses . . . . .	196
5.4.1	Basics of time-to-event modeling . . . . .	197
5.4.2	Constant hazard models . . . . .	198
5.4.3	Censoring . . . . .	200
5.4.4	Varying hazards models . . . . .	204
5.4.5	Proportional hazards models . . . . .	206
5.5	GLMMs for double-bounded responses . . . . .	209
5.5.1	Adding item covariates . . . . .	214
5.5.2	Modeling person covariates . . . . .	218
5.5.3	Speeding up sampling using within-chain parallelization . . . . .	220
5.6	Summary . . . . .	222
<b>6</b>	<b>Multivariate Models</b>	<b>223</b>
6.1	Setup . . . . .	223
6.2	Introduction . . . . .	223
6.3	Modeling residual confounding . . . . .	223
6.3.1	Correcting for confounding . . . . .	226
6.3.2	Correcting for residual confounding . . . . .	227
6.3.3	Checking model fit . . . . .	231
6.4	Modeling the phenotype of young birds . . . . .	233
6.4.1	Different formulas across models . . . . .	239
6.4.2	Different families across models . . . . .	239
6.4.3	Modeling residual correlations in the general case . . . . .	241
6.5	Summary . . . . .	244
<b>7</b>	<b>Meta-Analysis</b>	<b>245</b>
7.1	Setup . . . . .	245
7.2	Introduction . . . . .	245
7.2.1	Computation of effect sizes . . . . .	245
7.3	Fixed effects meta-analysis . . . . .	247
7.4	Random effects meta-analysis . . . . .	251
7.4.1	Marginalization of study-specific coefficients . . . . .	255
7.5	Moderator analysis . . . . .	256
7.6	Meta-analysis on count data . . . . .	261
7.6.1	Moderator Analysis . . . . .	264
7.6.2	What to add to nothing? . . . . .	265
7.7	Meta analysis of dependent outcomes . . . . .	266
7.7.1	Multilevel meta-analysis . . . . .	268
7.7.2	Adding known correlations . . . . .	272
7.7.3	Multivariate meta-analysis . . . . .	275

7.8	Raw data meta-analysis . . . . .	277
7.9	Summary . . . . .	284
<b>8</b>	<b>Data-driven non-linear models</b>	<b>285</b>
8.1	Setup . . . . .	285
8.2	Introduction . . . . .	285
8.3	Case Study: Housing Rents in Munich . . . . .	286
8.4	Splines . . . . .	288
8.4.1	Linear Splines . . . . .	288
8.4.2	Cubic Splines . . . . .	292
8.4.3	Smoothing Splines . . . . .	294
8.4.4	Tensor Product Splines . . . . .	300
8.4.5	Further spline features . . . . .	304
8.5	Gaussian processes . . . . .	305
8.5.1	Priors on Gaussian process hyperparameters . . . . .	313
8.5.2	Other Gaussian process kernels . . . . .	316
8.5.3	Multi-dimensional Gaussian processes . . . . .	317
8.5.4	Hilbert space approximate Gaussian processes . . . . .	318
8.6	Summary . . . . .	324
<b>References</b>		<b>325</b>

# Preface

Welcome to this early draft of my brms book. I continue to add new chapters as they get to a sufficiently complete and readable state. I don't necessarily write chapters in the order they will appear in the final version. So please be aware that chapters may change their number or even their order. If you are looking for any updates on the book's status, please check out <https://paulbuerkner.com/software/brms-book>.

The already existing chapters assume the reader to have some basic knowledge about Bayesian statistics. When it's finished, the book will also contain an intro chapter about the basics of Bayes, such that you can easily start learning Bayesian statistics from scratch with this book.

# 1 Linear Models

## 1.1 Setup

```
library(magrittr)
library(posterior)
library(ggplot2)
library(bayesplot)
library(brms)
```

## 1.2 Introduction

Linear models are the most simple regression models we can come up with. Yet, they are already very powerful but also surprisingly complex. For example, most of the discussion around causal modeling takes place in the linear modeling space and few people would claim that topic to be particularly easy.

In this chapter, will analyze the `epilepsy` data set that contains information about seizure counts in a randomized trial of anti-convulsant therapy in epilepsy (see Table 1.1). This data was initially provided by Thall and Vail (1990) and later used as an example in Breslow and Clayton (1993).

```
data("epilepsy", package = "brms")
```

Table 1.1: First six rows of the `epilepsy` data.

count	Trt	Base	Age	patient	visit
5	0	11	31	1	1
3	0	11	30	2	1
2	0	6	25	3	1
4	0	8	36	4	1
7	0	66	22	5	1
5	0	27	29	6	1

Table 1.1: First six rows of the `epilepsy` data.

count	Trt	Base	Age	patient	visit
4	0	17.0	33.0	1	1
12	1	14.0	33.0	1	2
1	0	13.0	29.0	1	3
10	1	13.0	33.0	1	4
1	0	13.0	33.0	2	1

Our response variable to be predicted is the number of seizure counts `count` within a given time interval. As predictors we have treatment (variable `Trt`; dummy-coded), 8-week baseline seizure counts (`Base`), and age of the patients in years (`Age`). Additionally, since each of the 59 `patients` is measured four times over the course of the study, we have an additional `visit` variable running from 1 to 4. We will ignore the `patient` and `visit` information for now but revisit it in later chapters.

As we will see, linear regression is not actually well suited for analyzing the `epilepsy` data. In my view, this coincides with most cases encountered in the real world, where linear regression can be considered only a very crude approximation at best. For a textbook example, this dataset has quite a few “inconvenient” properties, but analyzing it will teach us some important lessons in statistical modeling.

### 1.3 Gaussian linear models

Below, as well as in the upcoming chapters, we will denote response variables as  $y$ , predictor variables (aka covariate or feature) as  $x$ , and the observation index as  $n$ , which runs from 1 to the total number of observations  $N$ . With this notation, we can write the likelihood of a simple (Gaussian) linear model as

$$y_n \sim \text{normal}(\mu_n, \sigma) \quad (1.1)$$

$$\mu_n = b_0 + b_1 x_n. \quad (1.2)$$

In this book, I will write the (univariate) normal distribution in terms of mean  $\mu$  and standard deviation  $\sigma$ . In the literature, you may sometimes see variance  $\sigma^2$  or precision  $1/\sigma^2$  instead of the standard deviation, so don’t be confused if you see other people writing the normal distribution a bit differently. I prefer using the standard deviation because it is on the scale of the variable being normally distributed rather than on a quadratic (or inverse quadratic) scale, whose interpretation I consider harder.

With respect to our linear model, the parameters of are the intercept  $b_0$ , the slope  $b_1$  corresponding to  $x$ , which together form the linear predictor (mean)  $\mu$ , as well as the residual standard deviation  $\sigma$  capturing the errors (noise) that cannot be explained by the linear predictor. An equivalent formulation of Equation 1.1 that some of you may be more familiar with is

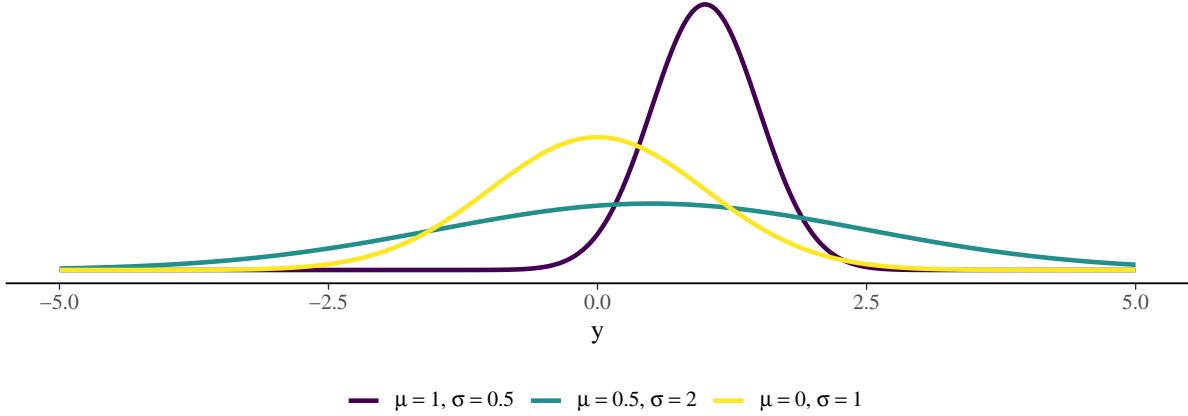


Figure 1.1: Three example densities of the normal distribution varying in both mean  $\mu$  and standard deviation  $\sigma$ .

$$y_n = \mu_n + \varepsilon_n \quad (1.3)$$

$$\varepsilon_n \sim \text{normal}(0, \sigma) \quad (1.4)$$

However, the latter formulation does not generalize well to many non-normal models, so you will need to get comfortable with the former way of writing it.

Since this book is about Bayesian modeling, you may expect a discussion of priors coming up next, but I don't think this is the best place for it here. The simple linear model we are considering is way too simple compared to the richness of the data for prior specification to be practically needed. In my experience, by focusing on priors too early, we (as teachers of Bayesian methods) do more harm than good, because doing so tends to shy away lots of users who would actually benefit from Bayesian methods. To be clear, prior specification can be very useful or may even be required in some cases. But for simple examples as the ones presented here, we will be fine without them. Or rather, we can hopefully trust brms to have made reasonable default choices in that regard.

So let's dive right into the brms specification of models. No matter how complex the model you fit, you will always do it through the `brm` function. Its first argument specifies the linear predictor via a `formula`. It can become quite complex but for our simple model here, it just reads `count ~ 1 + Trt`. This is standard R formula syntax. The response variable is on the left-hand side of the `~` and the predictors are on the right-hand side. The symbol `1` explicitly indicates the presence of an intercept, but we could also leave it out without changing the model, since an intercept is included by default anyway. The second argument `data` needs to be a `data.frame` from which the variables are taken. And that's it. We have everything to complete the specification of our first brms model:

```
fit_epi_gaussian1 <- brm(count ~ 1 + Trt, data = epilepsy)
```

As you run this code yourself, you will see some output popping up in the process that looks something like this:

```
Compiling Stan program...
Start sampling

SAMPLING FOR MODEL 'XYZ' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 3.7e-05 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition ...
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration: 1 / 2000 [ 0%] (Warmup)
Chain 1: Iteration: 200 / 2000 [ 10%] (Warmup)
Chain 1: Iteration: 400 / 2000 [ 20%] (Warmup)
Chain 1: Iteration: 600 / 2000 [ 30%] (Warmup)
Chain 1: Iteration: 800 / 2000 [ 40%] (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%] (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%] (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%] (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%] (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%] (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%] (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%] (Sampling)
Chain 1:
Chain 1: Elapsed Time: 0.023608 seconds (Warm-up)
Chain 1: 0.016017 seconds (Sampling)
Chain 1: 0.039625 seconds (Total)
Chain 1:

<more output from other chains>
```

After model fitting has completed, we can, for example, use trace plots as an initial, graphical convergence check:

```
mcmc_plot(fit_epi_gaussian1, type = "trace")
```

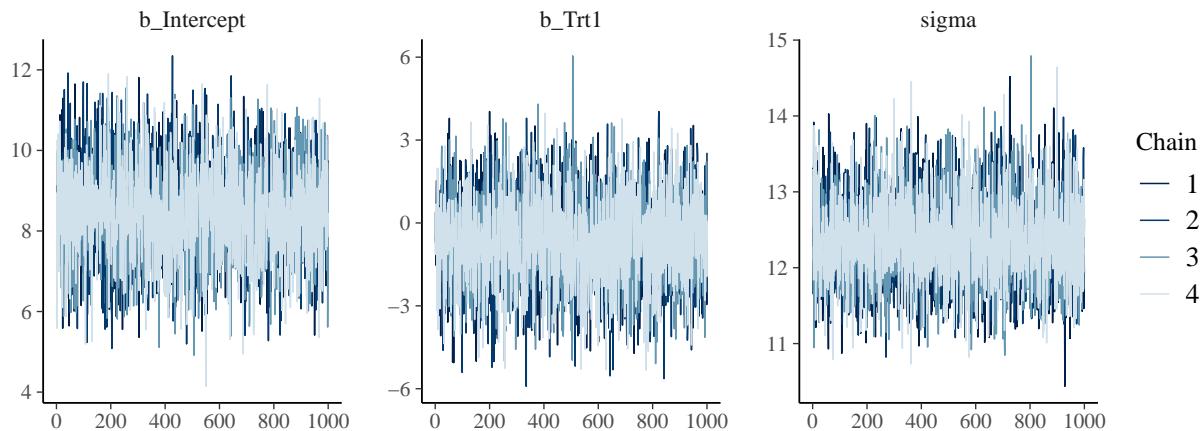


Figure 1.2: MCMC trace plots for model `fit_epi_gaussian1`.

In Figure 1.2, we see, for each parameter separately, traces of the four MCMC chains in different colors with post-warmup iterations on the x-axis and parameter values on the y-axis. These trace plots are showing ideal convergence: All chains are overlaying each other nicely, are stationary (horizontal on average), and show little autocorrelation (moving up and down very quickly).

Having convinced us of convergence, at least graphically for now, we can move on to inspecting the posterior. For example, we can plot histograms of the posterior samples per parameter (see Figure 1.3):

```
mcmc_plot(fit_epi_gaussian1, type = "hist", bins = 30)
```

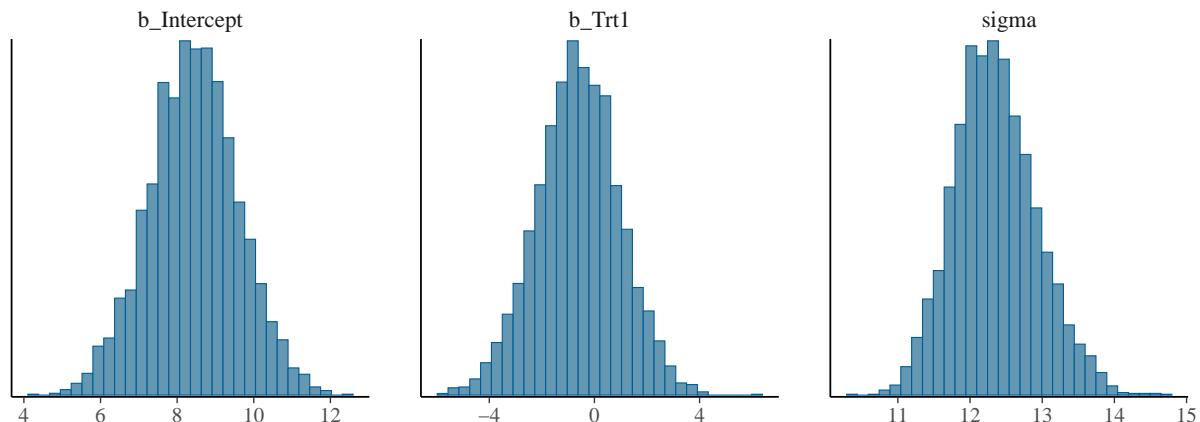


Figure 1.3: Posterior histograms for model `fit_epi_gaussian1`.

Alternatively, we can apply some density estimation algorithm to give us an approximate density on the basis of the posterior samples (see Figure 1.4):

```
mcmc_plot(fit_epi_gaussian1, type = "dens")
```

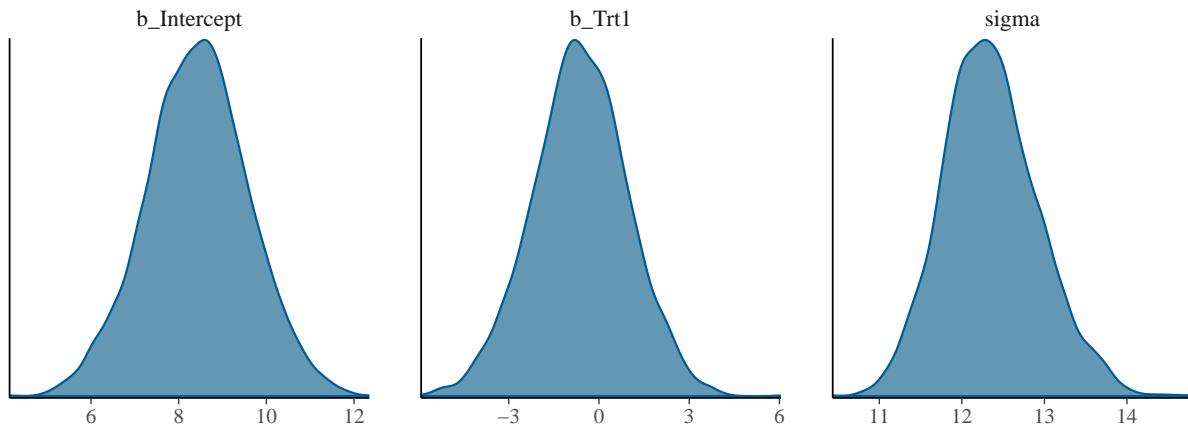


Figure 1.4: Posterior densities for model `fit_epi_gaussian1`.

These densities should be understood only as a pretty visualization of the estimated posterior, but not be used for actual inference (more on that later).

In addition to visualizations, we will also want to summarize the posterior samples and their convergence numerically. This is what the `summary` method is for.

```
summary(fit_epi_gaussian1)
```

```
Family: gaussian
Links: mu = identity
Formula: count ~ 1 + Trt
Data: epilepsy (Number of observations: 236)
Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
      total post-warmup draws = 4000
```

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	8.43	1.16	6.09	10.74	1.00	4064	2708
Trt1	-0.62	1.58	-3.81	2.46	1.00	3844	2879

#### Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	12.35	0.57	11.30	13.58	1.00	4608	3108

Draws were sampled using `sampling(NUTS)`. For each parameter, Bulk\_ESS

and `Tail_ESS` are effective sample size measures, and `Rhat` is the potential scale reduction factor on split chains (at convergence, `Rhat` = 1).

There is a lot to unpack here in the summary output. We start with some aspects and come back to others in later chapters.

At the top of the output, we see a short overview over the most important input, we have given to the model, for example the formula and dataset. Moving to the actual posterior results, `brms` shows two blocks of parameters here, namely **Regression Coefficients** and **Further Distributional Parameters**. The former name is self-explanatory I think. The latter refers to additional parameters of the likelihood that we did not predict with our regression, here the residual standard deviation `sigma`. For each of the displayed parameters (rows), we see seven summary statistics (columns). The `Estimate` column is simply the empirical mean of the posterior draws  $\bar{\theta} = \frac{1}{S} \sum_{s=1}^S \theta^{(s)}$ , which is a sampling-based estimate of the analytical posterior mean. The `Est.Error` columns denotes (the empirical estimate of) the posterior standard deviation, while the lower and upper bounds of the 95% credible interval are computed as the 2.5% and 97.5% quantile of the posterior samples, respectively. The three remaining columns, `Rhat`, `Bulk_ESS`, and `Tail_ESS` give us information about convergence and sampling efficiency, as also shortly explained at the bottom of the summary output. I will go into more details of convergence diagnostics later on in [?@sec-bayesian-workflow](#). At this point, all you need to know is that convergence is basically perfect for this model with `Rhat` being 1.00 and ESS estimates being roughly as large as the total number of post warmup draws  $S = 4000$ .

### 1.3.1 Why do we need sampling?

A common understanding is that we are using sampling-based methods only because we do not have access to the analytic posterior. While this is certainly the primary reason, it turns out that posterior draws are actually way easier to handle than posterior densities as I will illustrate with a simple example. Suppose we are interested in the posterior of the residual variance  $\sigma^2$  while we only have information on the residual standard deviation  $\sigma$  in our model. Fortunately, when the posterior is stored in the form of draws, this is no problem at all. We just have to apply the desired transformation *for each draw separately* and, voilà, we have posterior draws for our transformed quantity of interest. In our examples, this means that the squared posterior draws of  $\sigma$  are in fact valid posterior draws of  $\sigma^2$ :

$$(\sigma^{(s)})^2 = (\sigma^2)^{(s)}$$

But isn't this obvious you may ask? Well, not necessarily, at least once we consider the effort we would have to put into this transformation if we only had posterior densities instead. The key term here is *Jacobian adjustment*. I am not explaining the details here, but you can look them up, for example, in Gelman et al. (2013) to convince yourself that dealing with densities when transforming parameters is quite a lot of work.

The reason we need a Jacobian adjustment above is that the square is a non-linear transformation. But even if our transformation is linear, having access to draws makes our live considerably easier. Suppose we are interested in the posterior of the expected number of seizures in the treatment group (call it  $\mu_{\text{Trt}}$ ), then we don't have this quantity in our model from above. But we have everything we need: Because the treatment variable was contrast coded, we know that  $\mu_{\text{Trt}} = b_0 + b_1$ . Accordingly, all we have to do to get posterior draws of  $\mu_{\text{Trt}}$  (and thus a representation of its posterior) is to apply the above transformation *for each draw separately*:

$$\mu_{\text{Trt}}^{(s)} = b_0^{(s)} + b_1^{(s)}$$

In R, we can easily perform such transformations by first extracting the posterior draws and then applying the transformation per draws in a vectorized manner. For example, using the functionality from the `posterior` package (Bürkner et al. 2024):

```
draws <- as_draws_df(fit_epи_gaussian1) %>%
  mutate_variables(sigma2 = sigma^2,
                  mu_Trt = b_Intercept + b_Trт1)
```

Plotting can then be performed for example via the `bayesplot` package (Gabry and Mahr 2024), which is also implicitly used in the `mcmc_hist` function applied earlier.

```
bayesplot::mcmc_hist(draws, c("sigma2", "mu_Trt"), bins = 30)
```

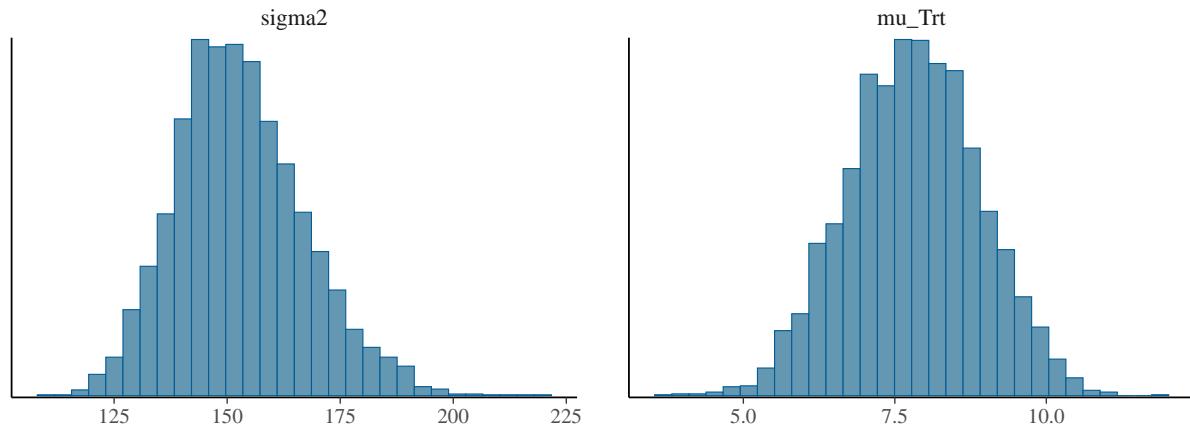


Figure 1.5: Posterior histograms of two transformed quantities for model `fit_epи_gaussian1`.

The output is shown in Figure 1.5. With  $\mu_{\text{Trt}}$  we computed the model-implied predictions of the mean for the treatment group. For the control group, it would just be  $\mu_{\text{Ctrl.}} = b_0$  for

this simple model. For more complex models, computing the predictions for different groups or, more generally, different predictor values manually becomes quite cumbersome. For this reason brms provides you with a convenient method to provide quick graphical summaries of the model-implied predictions per predictor:

```
ce <- conditional_effects(fit_epи_gaussian1)
plot(ce)
```

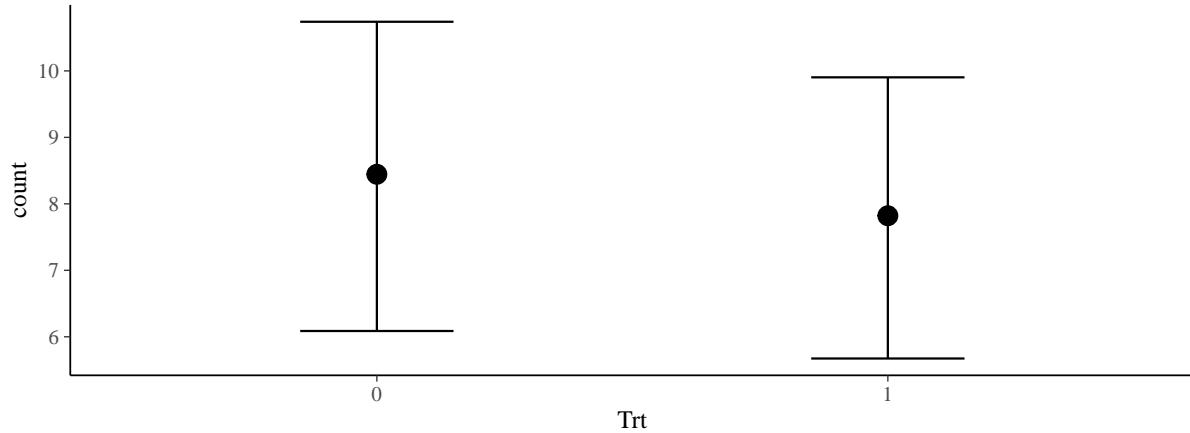


Figure 1.6: Conditional effects of model `fit_epи_gaussian1`.

The error bars in Figure 1.6 are representing 95% credible intervals by default, but we can change that value if we like via the `prob` argument. Comparing the `conditional_effects` plot with our manually computed posterior of  $\mu_{\text{Trt}}$ , we see that they actually are the same.

To put the mean predictions into the context of the observed data, we can also show the data as points in the plot:

```
plot(ce, points = TRUE)
```

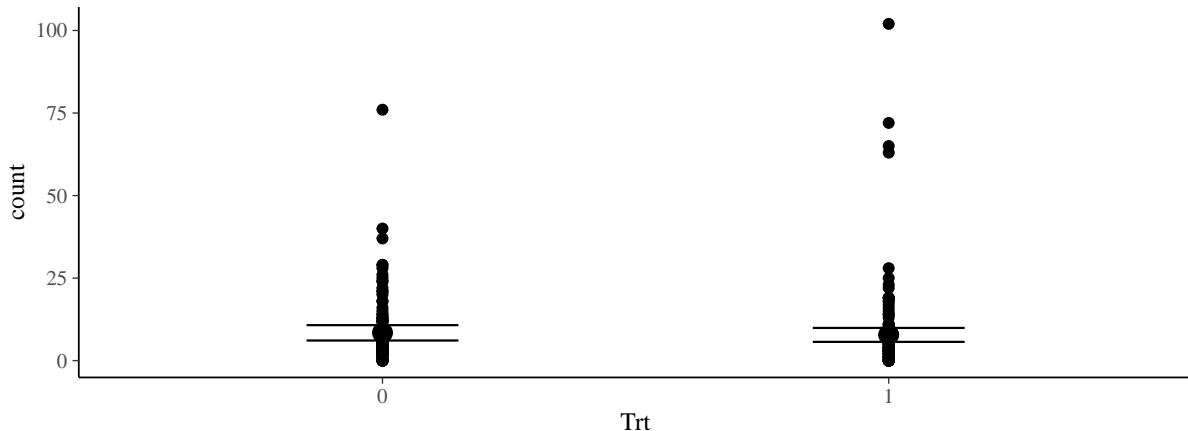


Figure 1.7: Conditional effects model `fit_epi_gaussian1` with observed data points overlaid.

As we see clearly in Figure 1.7, there are some points outlying points in both groups, which makes the differences in group means hard to see. Accordingly, depending on our goal, it may or may not make sense to show the data points.

For the remainder of this chapter, we would like to always see the points in `conditional_effects` so we are setting the corresponding option globally to shorten our code.

```
options(brms.plot_points = TRUE)
```

### 1.3.2 Variations of posterior predictions

What we do in `conditional_effects` by default, and have done above, is visualizing the *expected value* (mean parameter) of the likelihood distribution, conditional on certain predictor values. In brms, this is done via the `posterior_epred` (posterior expected predictions) method. For example, we can run

```
newdata <- data.frame(Trt = c(0, 1))
pe <- posterior_epred(fit_epi_gaussian1, newdata = newdata)
```

to create expected posterior predictions for both groups. The resulting object contains the posterior draws in the rows and the different conditions (here, treatment groups) in the columns

```
str(pe)
```

```
num [1:4000, 1:2] 9.11 8.07 9.13 7.75 8.38 ...
```

We can summarize the draws, for example, via

```
posterior_summary(pe)
```

Table 1.2: Summary of posterior mean predictions for model `fit_epি_gaussian1`. The first row shows the prediction for the control group. The second row shows the predictions for the treatment group.

Estimate	Est.Error	Q2.5	Q97.5
8.43	1.16	6.09	10.74
7.81	1.09	5.67	9.90

In linear models, `posterior_epred` directly coincides with evaluating the linear predictor  $\mu$  as exemplified above. What `posterior_epred` does not include is the residual (aleatoric) uncertainty, which is represented by  $\sigma$  in our linear models.

Let's consider again the task of evaluating predictions for the treatment group. If we are only interested in (the posterior of) the likelihood's mean parameter, we would compute  $\mu_{\text{Trt}}^{(s)} = b_0^{(s)} + b_1^{(s)}$ . In contrast, if we are interested in prediction of hypothetical new data points  $y_{\text{Trt}}^{(s)}$  from the treatment group (i.e., actual posterior predictions), we would sample

$$y_{\text{Trt}}^{(s)} \sim \text{normal}(\mu_{\text{Trt}}^{(s)}, \sigma^{(s)})$$

This is exactly what happens behind the scenes when we execute the code below (see Figure 1.8):

```
conditional_effects(fit_epি_gaussian1, method = "posterior_predict")
```

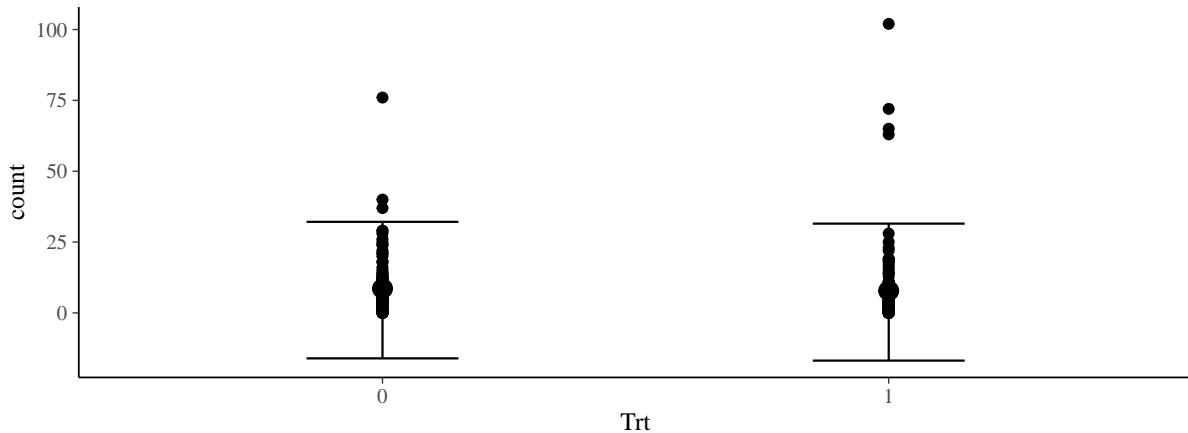


Figure 1.8: Conditional effects using `posterior_predict` of model `fit_epi_gaussian1` with data points overlayed.

We could have also done this more manually via

```
newdata <- data.frame(Trt = c(0, 1))
pp <- posterior_predict(fit_epi_gaussian1, newdata = newdata)

posterior_summary(pp)
```

Table 1.3: Summary of posterior predictions for model `fit_epi_gaussian1`. The first row shows the prediction for the control group. The second row shows the predictions for the treatment group.

Estimate	Est.Error	Q2.5	Q97.5
8.518848	12.24587	-15.38176	32.58212
7.951298	12.31463	-16.84521	31.32251

Comparing the visual with the numeric output is a bit difficult here, but be rest assured that both is doing the same thing. One note of caution though: As we sample new  $y_{\text{Trt}}^{(s)}$  as part of `posterior_predict`, we will get slightly different results every time we run it, unless we fix R's random seed via the `set.seed` function.

### 1.3.3 Posterior predictive checks

We are already aware that the considered linear regression model is not ideal for the epilepsy data. But how bad is it? As quick graphical method, we can use posterior predictive (PP)

checks, where we are compare the observed outcome data with the model predicted outcome data, that is, with the posterior predictions. In brms, we can perform PP-checks via

```
pp_check(fit_ephi_gaussian1)
```

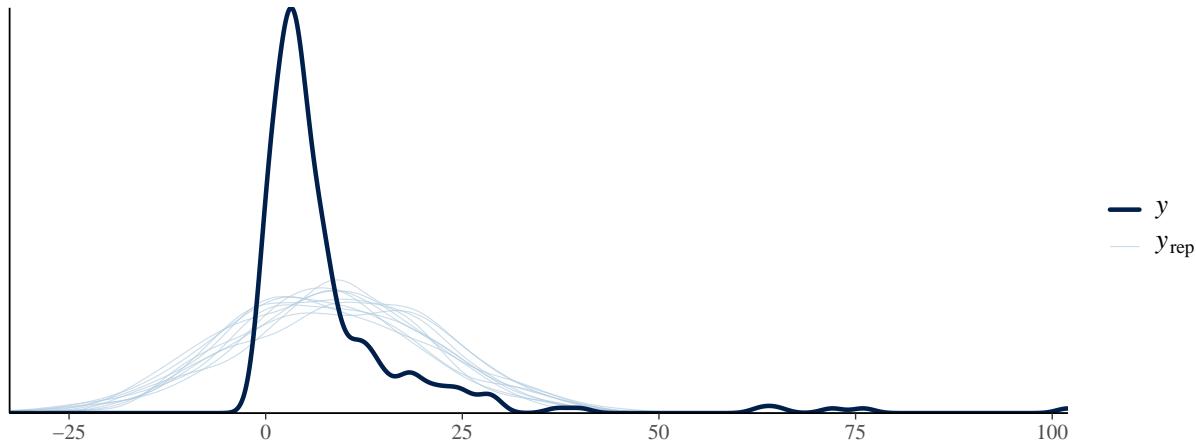


Figure 1.9: Default posterior predictive check (density overlay) for model `fit_ephi_gaussian1`.

By default, the `dens_overlay` (density overlay) type is used, which plots the marginal density of the observed outcomes (shown in dark color) against marginal densities obtained posterior predictions (shown in lighter color), see Figure 1.9. Each posterior draws implies it's own posterior predictive distribution over observations. Hence, in the above plot, we see ten light blue lines, each corresponding to one posterior draw. We could of course show more such lines, but usually the pattern is clear just from these few.

Back to criticizing our linear model, there is a lot to learn even from this simple PP-check shown in Figure 1.9. We see that the model predictions can neither account for the strong spike of observed outcomes close to zero nor for their right-skewness. Instead, the the model also predicts a lot of negative outcomes, which is impossible in reality because we are predicting counts of epileptic seizures. In the plot, it looks as if the observed data also had few negative values (the dark blue density going below zero) but this is just an artifact of estimating a continuous density from counts. While this PP-check type is definitely not ideal to illustrate count outcome data, it still very clearly points to the shortcomings of our linear model.

While the default PP-check was already eye-opening, there are lot of types that can further our understanding of model appropriateness. For example, an often very useful check is obtained by comparing the residuals = observed outcomes - model predictions with the observed outcomes, also known as *residual plot*. In `pp_check` this check type is called `error_scatter_avg`:

```
pp_check(fit_epi_gaussian1, type = "error_scatter_avg")
```

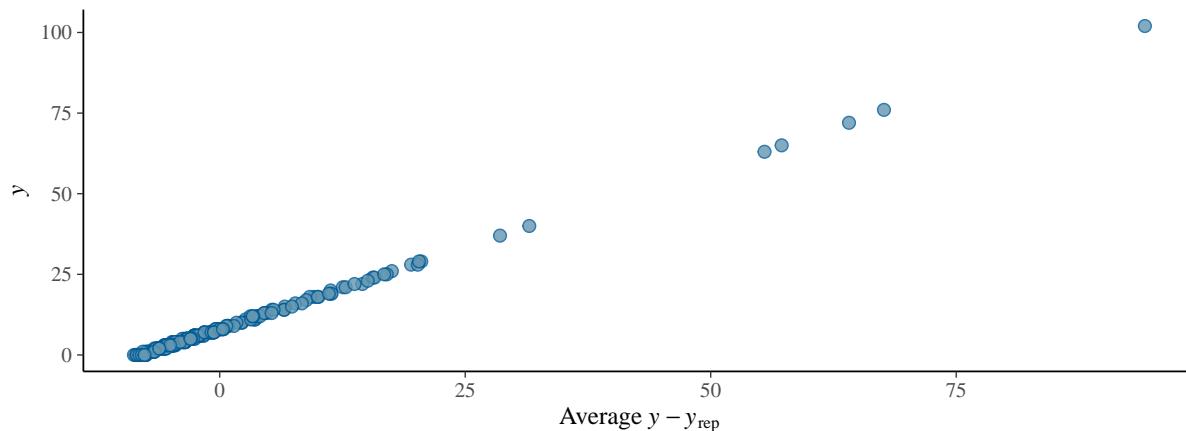


Figure 1.10: Posterior predictive check comparing observed responses (y-axis) with the predictive residuals (x-axis) of model `fit_epi_gaussian1`.

Ideally, as for any residual plot, we want to see a point cloud without a visible relationship between residuals and outcomes. But this is not at all what we see in Figure 1.10. Instead, there is a strongly almost perfectly linear relationship indicating strong problems with the independence assumption of the errors. Essentially, both PP-checks have told us that our initial model is a very bad for the data at hand. In fact, we will spend quite a bit of time in this and later chapters to build better models for these data.

Before we move on, one small tip about the `pp_check` method. If you don't know which check types are available, you can simply pass an arbitrary non-supported `type` name to get a list of all currently supported types:

```
pp_check(fit_epi_gaussian1, type = "help_me")
```

### 1.3.4 Adding more predictors

There are a lot of problems with our initial model. The first problem we will address is the fact that there are other (non-treatment) variables that we want to control for because they have a relevant influence on our outcomes. The treatment was randomized so we need to worry less about confounding than in observational studies, but still it will likely be very beneficial to control for some other variables – even if it was only to reduce the uncertainty in our treatment effects' posterior.

In this example, we will use the number of epileptic seizures observed in a standardized time frame *before* treatment, a variable called `Base`. In `brm` all we have to do is to add it to the right-hand of the model formula:

```
fit_epi_gaussian2 <- brm(count ~ Trt + Base, data = epilepsy)
```

```
summary(fit_epi_gaussian2)
```

```
Family: gaussian
Links: mu = identity
Formula: count ~ Trt + Base
Data: epilepsy (Number of observations: 236)
```

#### Regression Coefficients:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-2.36	0.97	-4.23	-0.49	1.00	4628	3509
Trt1	-0.89	1.06	-2.92	1.16	1.00	4457	3269
Base	0.35	0.02	0.31	0.39	1.00	4612	2888

#### Further Distributional Parameters:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	8.02	0.37	7.33	8.81	1.00	4418	3387

The summary output indicates that, unsurprisingly, `Base` is very strongly and positively related to the outcome, which is confirmed by a graphical summary (see Figure 1.11):

```
conditional_effects(fit_epi_gaussian2, effects = "Base")
```

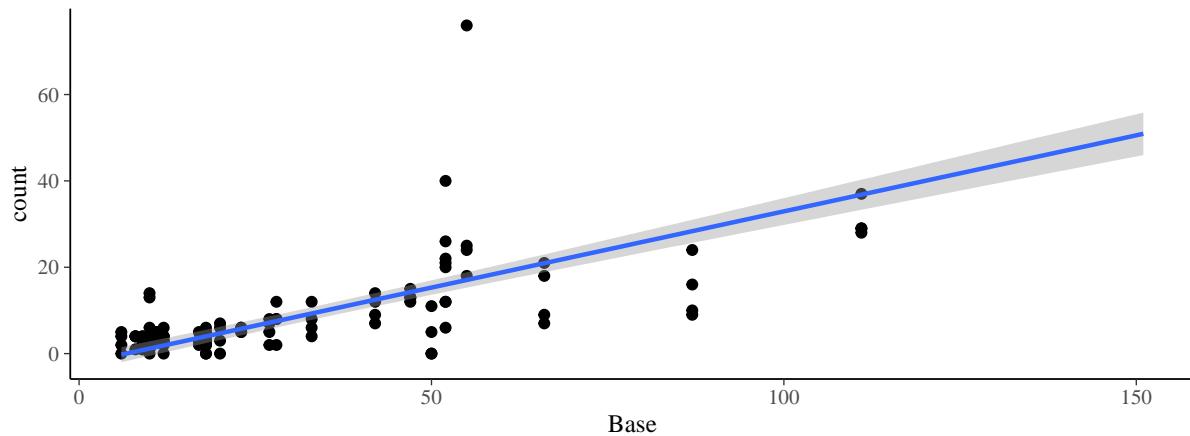


Figure 1.11: Conditional effects of baseline epilepsy score for model `fit_epi_gaussian2`.

What is more, while the posterior mean of `Trt` has changed only a bit, the posterior uncertainty, for example, as quantified by the 95% credible interval bounds has reduced quite a lot as a result of controlling for `Base`.

The above model assumes that the effect of the treatment is independent of the baseline number of seizures. But what if the treatment worked better (or worse) for patients with stronger epileptic symptoms, as measured via number of seizures before treatment. This calls for an interaction of `Trt` and `Base`, which we can write in R formula syntax as `Trt * Base`. In fact, as is the case for all standard regression functions in R as well, this will also add the individual coefficients for the two predictors so our interaction model is simply written as:

```
fit_ephi_gaussian3 <- brm(count ~ Trt * Base, data = epilepsy)
```

Again, we first summarize the model both numerically:

```
summary(fit_ephi_gaussian3)
```

```
Family: gaussian
Links: mu = identity
Formula: count ~ Trt * Base
Data: epilepsy (Number of observations: 236)
```

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	0.35	1.13	-1.83	2.59	1.00	2262	2699
Trt1	-5.74	1.53	-8.75	-2.71	1.00	2109	2583
Base	0.27	0.03	0.21	0.32	1.00	2146	2528
Trt1:Base	0.16	0.04	0.08	0.23	1.00	1936	2312

#### Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	7.77	0.35	7.13	8.47	1.00	3468	2974

The negative main effect of the treatment points to a beneficial effect for low number of baseline seizures (close to `Base = 0`). However, we also see a positive interaction of treatment and baseline. That is, if we take this result seriously for a moment, the treatment appears to become worse as the baseline seizure count increases. When summarizing graphically via

```
conditional_effects(fit_ephi_gaussian3, effects = "Base:Trt")
```

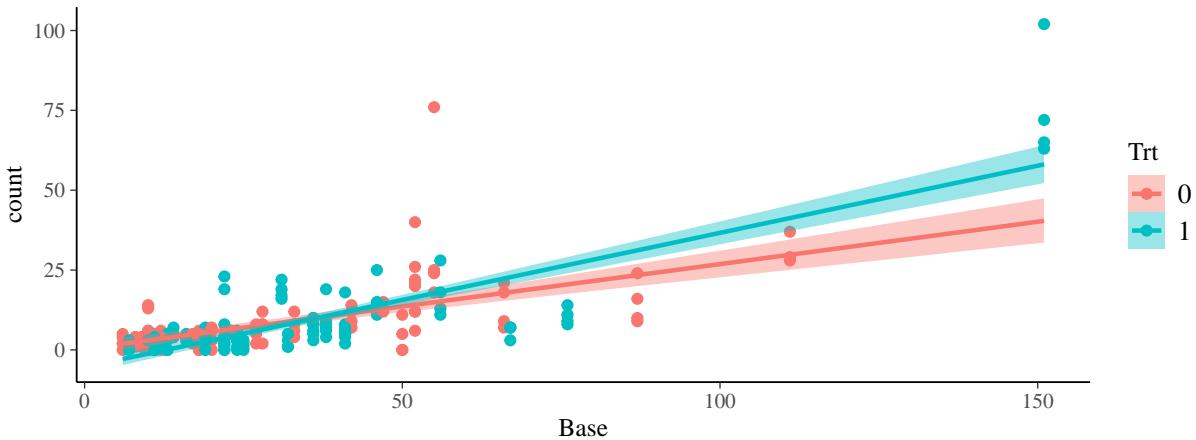


Figure 1.12: Conditional effects of baseline-treatment interaction for model `fit_epi_gaussian3`.

we also see that the model predicts the treatment to be worse than the control for very high baseline values (see Figure 1.12). The resolution of the plot is not great to see what happens at lower baseline values. So let's zoom in a bit (see Figure 1.13):

```
conds <- list(Base = seq(1, 50, 1))
ce <- conditional_effects(
  fit_epi_gaussian3, effects = "Base:Trt",
  int_conditions = conds
)
plot(ce, points = FALSE)
```

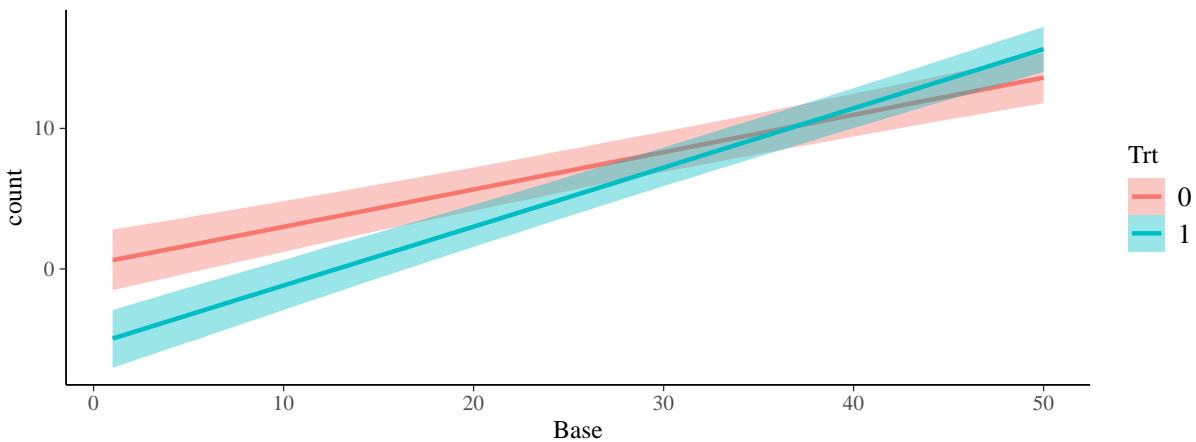


Figure 1.13: Conditional effects of baseline-treatment interaction for model `fit_epi_gaussian3` with reduced range of baseline values.

The argument `int_conditions` (short for *interaction conditions*) can be used to control which predictor values to plot both on the x-axis and in terms of groups differentiated by color. Now, we more clearly see that the summary has told us, namely that the model expects the treatment to be beneficial as long as the baseline does not exceed around 20 seizures.

Let's take a look at the residual PP-check again:

```
pp_check(fit_epi_gaussian3, type = "error_scatter_avg")
```

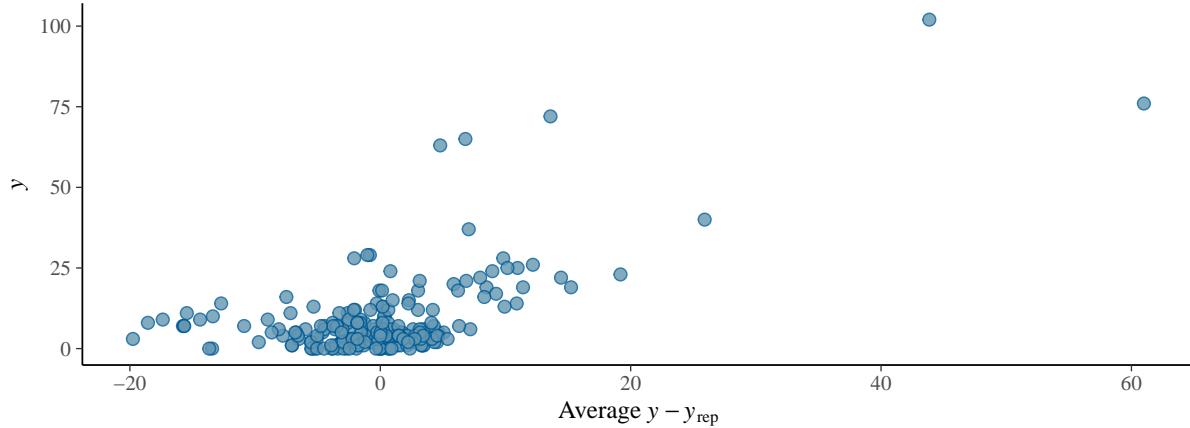


Figure 1.14: Posterior predictive check comparing observed responses (y-axis) with the predictive residuals (x-axis) of model `fit_epi_gaussian3`.

In Figure 1.14, we see that things appear less bad as for our initial models, but still not great. In particular, we clearly see some outliers at the top right of the plot, which we could also see in the conditional plots earlier. Knowing that linear models are susceptible to outliers, it is plausible that they have influenced the model in a qualitative manner that affects our conclusions. In the next section, we will investigate this issue in more detail.

### 1.3.5 Outlier analysis

In Figure 1.15, we see a histogram of the baseline seizure counts we have observed in our data.

There appears to be a group of outliers at `Base = 150` and another one around `Base = 110`. A close investigation reveals that these two groups are belonging to only one person each, whose baseline values have just been repeated four times due the repeated measures of the outcomes. So removing observation with baseline value above, say, `Base = 140` will in fact only remove a single person from the analysis. Let's see what happens to our model results in this case:

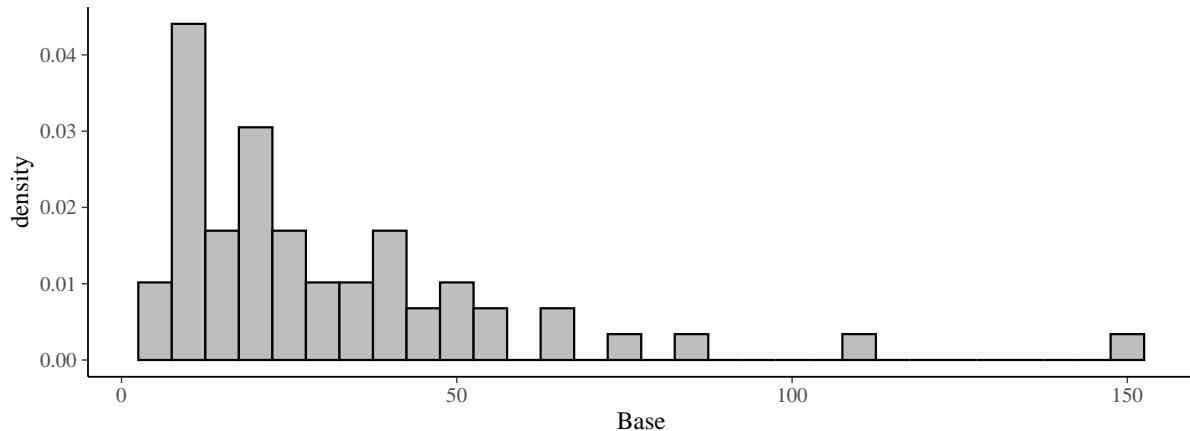


Figure 1.15: Histogram of baseline seizure counts in the epilepsy data set.

```
fit_epi_gaussian_outlier1 <- brm(
  count ~ Trt * Base,
  data = epilepsy %>% filter(Base < 140)
)
```

When taking a selective look at the regression coefficients, we see that the interaction of treatment with baseline has now become negative (Table 1.4). That is, after removing data for that one person, the model thinks that the treatment works *better* for people with higher baseline number of seizures, the complete opposite result of what we had before.

Table 1.4: Posterior summary of regression coefficients for model fit\_epi\_gaussian\_outlier1.

	Estimate	Est.Error	Q2.5	Q97.5
b_Intercept	0.36	0.97	-1.56	2.28
b_Trt1	0.43	1.49	-2.43	3.37
b_Base	0.27	0.02	0.22	0.31
b_Trt1:Base	-0.09	0.04	-0.18	-0.01

Let's now also remove data from the second person with outlying baseline scores by filtering out everything above `Base = 100`:

```
fit_epi_gaussian_outlier2 <- brm(
  count ~ Trt * Base,
  data = epilepsy %>% filter(Base < 100)
)
```

The results shown in Table 1.5 have changed only a little bit from the first outlier model in that the interaction is now slightly less clearly negative.

Table 1.5: Posterior summary of regression coefficients for model `fit_ephi_gaussian_outlier2`.

	Estimate	Est.Error	Q2.5	Q97.5
b_Intercept	0.53	1.05	-1.53	2.63
b_Trt1	0.26	1.55	-2.77	3.24
b_Base	0.26	0.03	0.20	0.32
b_Trt1:Base	-0.08	0.05	-0.17	0.01

However, if someone followed a strict null-hypothesis significance testing approach, they might actually conclude a qualitative change: Using the 95% credible intervals to declare “statistical significance” in a frequentist sense, the interaction would have changed from “significantly different from zero” to “not significantly different from zero”. I think that declaring scientific discovery by such means is not sensible and I do not advocate it. I just want to point it out in case readers used to such interpretation read this book and wonder about my thoughts on that matter.

Summarizing what we have learned about our epilepsy treatment so far is that it may very well have a positive or negative effect, or something in between, depending on how we look at the data. In other words, using only Gaussian linear models and three variables, we have already managed to utterly confuse ourselves. And we are only getting started modeling these data.

## 1.4 Fat-tailed linear models

Selectively removing outliers has value in the contexts of sensitivity analyses. But it also increases the number of more or less arbitrary decisions we have to make and thus increases the danger of selective reporting conditional on the outcome of our analyses. What is more, if we have no reason to believe that the outlying data is in some way invalid, then we should model rather than exclude them. So what if we want to continue using linear models for this data without our analysis being too dependent on very few data points?

The flexibility of full Bayesian inference comes to rescue: We can simply replace the normal likelihood with a likelihood that can have fatter tails if the data suggests so. Intuitively, such a likelihood would “expect” outliers to occur, thus not be surprised by their presence and ultimately less influenced by them. A distributions with potentially fatter tails than normal is the *Student-t* distribution (see Figure 1.16). Its density is quite complicated, but we do

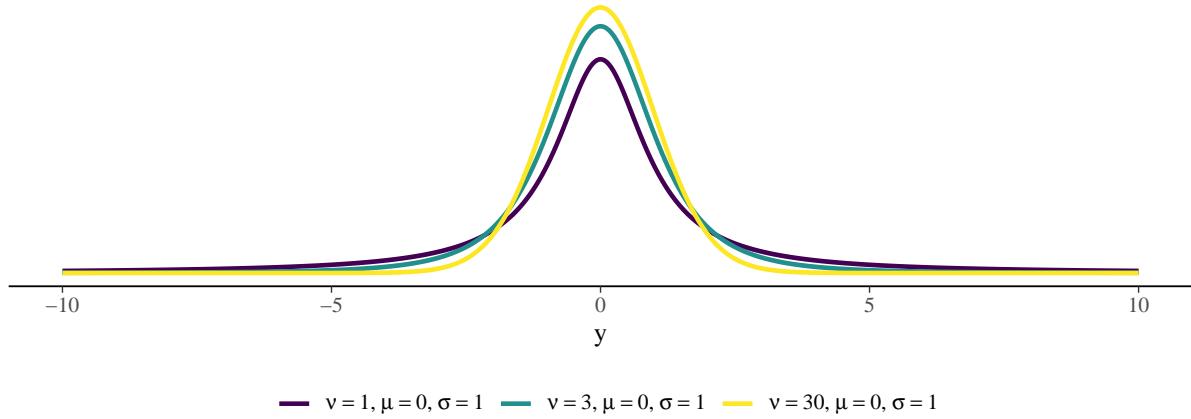


Figure 1.16: Three example densities of the Student-t distribution varying in the degrees of freedom  $\nu$ .

not have to worry about it, since Stan has a stable and efficient implementation that we can use.

What we do have to care about is the distribution's properties, most notably related to the *degrees of freedom* parameter  $\nu$ . When  $\nu$  is small, the Student-t distribution has very fat tails. For  $\nu = 1$ , the tails are so fat that we cannot even compute the mean of the distribution sensibly anymore (mathematically, we say “the mean does not exist”). The  $\nu = 1$  even has a special name: It's called the *Cauchy* distribution. As  $\nu$  becomes larger, the tails become lighter. Mathematically, Student-t converges to normal as  $\nu \rightarrow \infty$  but already for  $\nu \geq 30$ , we can barely see any difference between the two anymore. That is, for practical modeling purposes, it does not matter if  $\nu = 30, 100, 10000$  since all of them behave almost like the normal distribution. Accordingly, we care mostly about the space between  $\nu = 1$  and  $\nu = 30$ .

This implies that, in contrast to the regression coefficients  $b$  or the scale  $\sigma$ , setting an informative prior on  $\nu$  is crucial. In particular, we need to avoid sampling issues if the data suggests residuals close to normal. That is, if we didn't specify an informative prior, samples of  $\nu$  might diverge towards infinity as the Student-t likelihood looks almost the same for any  $\nu \geq 30$ . Following Juárez and Steel (2010), we set the following prior on  $\nu$  by default:

$$\nu \sim \text{Gamma}(2, 0.1).$$

As displayed in Figure 1.17, most of its mass is assigned to values smaller than 50, thus preventing the posterior to move into problematically high areas  $\nu$ .

To use the Student-t likelihood in brms, all we have to do is to add `family = student()` to the `brm` call. The `family` argument controls which distribution to use as likelihood. By default `family` is set to `gaussian` so we didn't have to specify it before in our previous models.

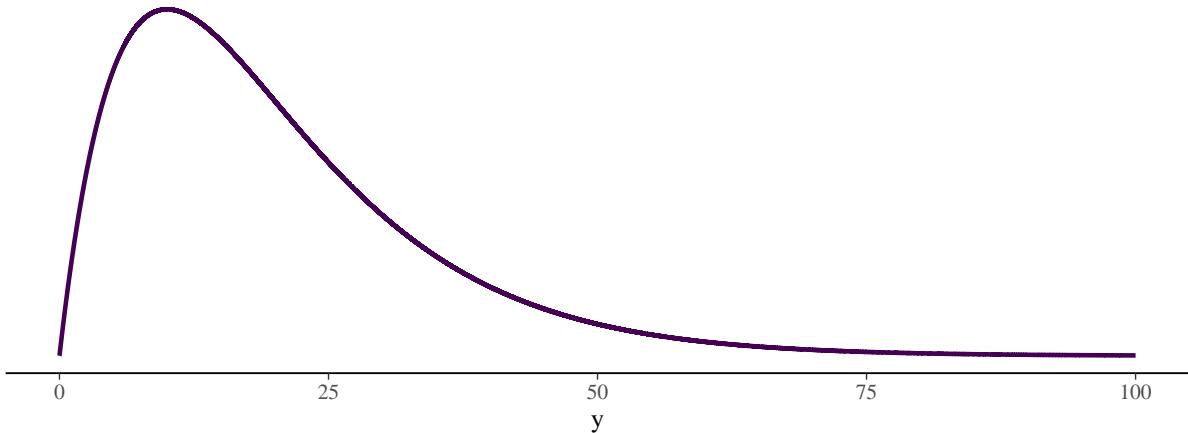


Figure 1.17: Density of the  $\text{Gamma}(2, 0.1)$  distribution. It is used by brms as the default prior for the degrees of freedom parameter  $\nu$  of the Student-t likelihood.

```
fit_epi_student1 <- brm(
  count ~ Trt * Base,
  data = epilepsy,
  family = student()
)
```

```
summary(fit_epi_student1)
```

```
Family: student
Links: mu = identity
Formula: count ~ Trt * Base
Data: epilepsy (Number of observations: 236)
```

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	0.01	0.36	-0.69	0.71	1.00	3126	2665
Trt1	0.41	0.63	-0.85	1.63	1.00	1928	2621
Base	0.26	0.01	0.24	0.28	1.00	2555	2372
Trt1:Base	-0.11	0.02	-0.16	-0.07	1.00	1778	2376

#### Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	2.10	0.20	1.73	2.52	1.00	3131	2578
nu	1.40	0.18	1.09	1.82	1.00	2800	1563

In this summary, two major things stand out. First, both the average effect of treatment and

its interaction with the baseline seizure count are clearly negative. This means that the model thinks this treatment is both effective on average and particularly so for higher number of baseline seizures. We can also clearly see this pattern visually (Figure 1.18):

```
conditional_effects(fit_epi_student1, effects = "Base:Trt")
```

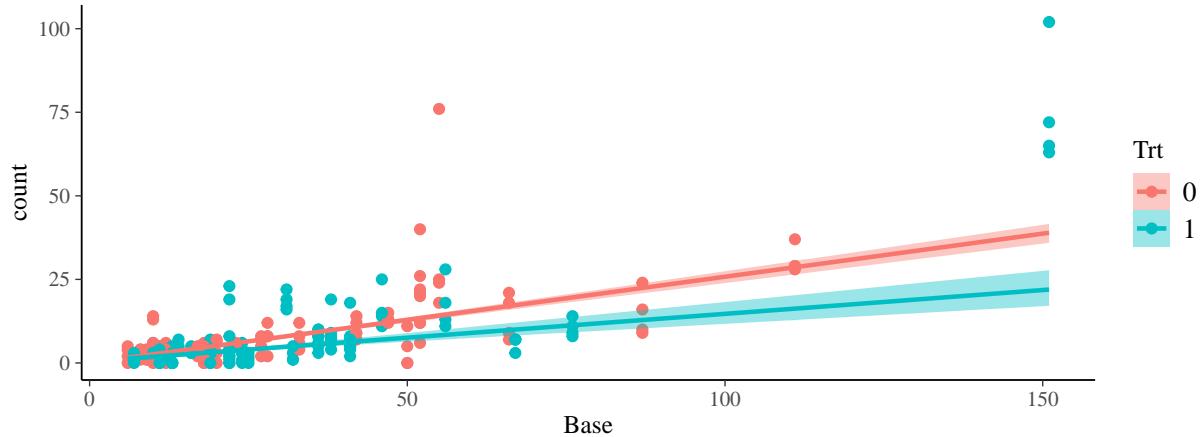


Figure 1.18: Conditional effects of baseline-treatment interaction for model `fit_epi_student1`.

Second, the summary output reveals that the degrees of freedom parameter `nu` is remarkably small, almost at its lower boundary of 1, which we also see via the `plot` method (Figure 1.19):

```
plot(fit_epi_student1, variable = "nu")
```

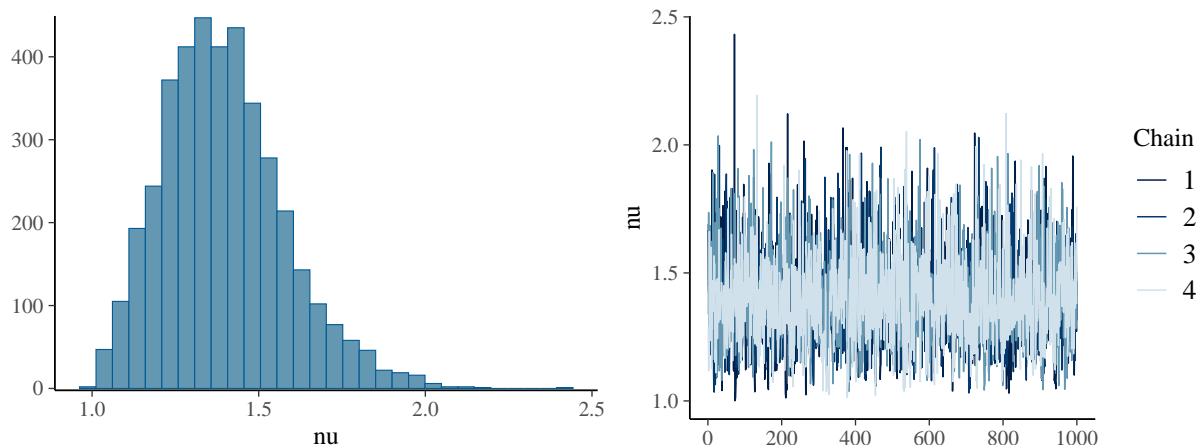


Figure 1.19: Trace plot and posterior histogram of the degrees of freedom parameter `nu` for model `fit_epi_student1`.

This indicates that, in order for this model to fit the data well, its likelihood had to have very fat tails. This is unsurprising given how extreme some of the outliers are. But it also shows how bad of a likelihood the Gaussian distribution is for this data because it has no statistical mechanism to properly account for these outliers.

We also see in the summary that the scale parameter  $\sigma$  is much smaller in comparison to that we saw in the corresponding normal model ( $\sigma \approx 2$  vs.  $\sigma \approx 7.8$ ). This comparison is misleading though. The reason is that, for the Student-t model,  $\sigma$  is no longer directly equal to the residual standard deviation. Instead, we have  $SD = \sigma \sqrt{\frac{\nu}{\nu-2}}$  for  $\nu > 2$  and  $SD = \infty$  for  $\nu < 2$ . Since we have a lot of draws with  $\nu^{(s)} < 2$  in the present case, this means that the residual standard deviation estimated by the Student-t model is estimated as infinite or at least very very large. On a practical level, this is not a realistic estimate for the given data, but the result of the Student-t model's attempt to fit well despite substantial model-misspecification.

#### 1.4.1 Posterior predictive checks

When we run posterior predictive checks with the default `dens_overlay` type, we don't really get something useful at first sight (see Figure 1.20).

```
pp_check(fit_epi_student1)
```

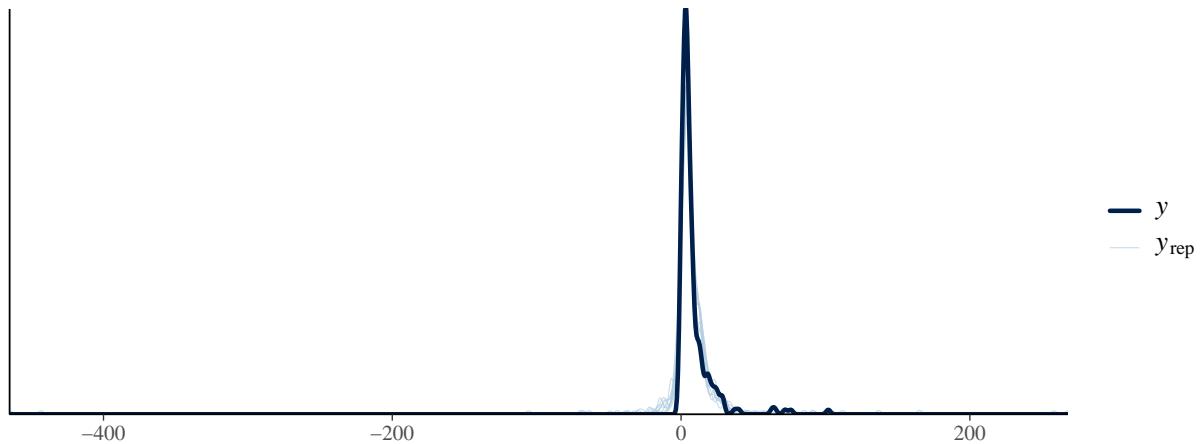


Figure 1.20: Default posterior predictive check for model `fit_epi_student1`.

This is an unfortunate side-effect of our likelihood being estimated as so fat-tailed (small  $\nu$ ). When we generate new data from posterior predictive distribution, some of this data will be sampled as very far away from the bulk of the distribution. While this is to be expected, it effectively destroys the scale of the x-axis of our plot. So let's redo the plot this time truncating the x-axis (see Figure 1.20). Since the `pp_check` method returns ggplot objects, amending the figure is simple:

```
pp_check(fit_epi_student1) + xlim(-50, 150)
```

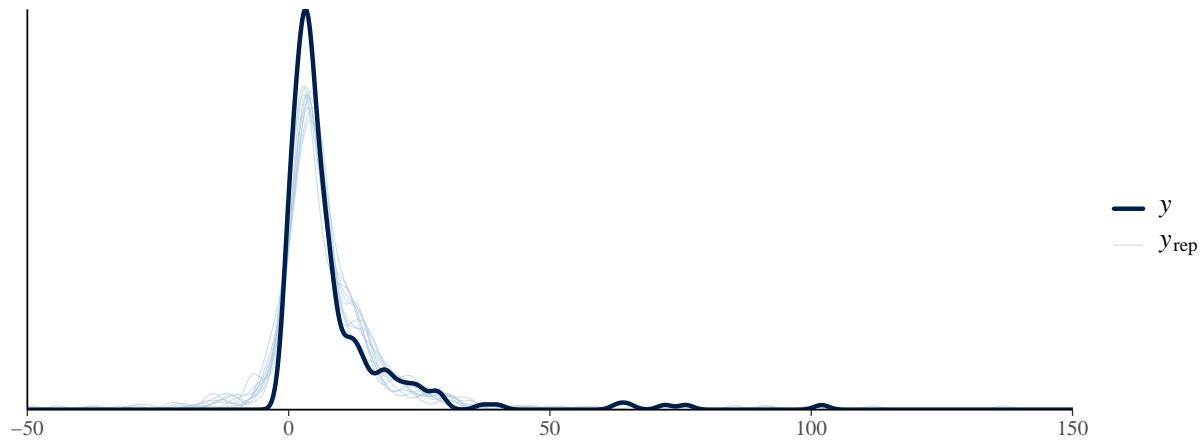


Figure 1.21: Default posterior predictive check for model `fit_epi_student1` with a truncated x-axis.

Looks much better. And we have only excluded very few predictive draws in the process, so the plot as we see it is still valid. It reveals a surprisingly good marginal in-sample fit of the Student-t model, which can apparently even account for most of the right-skewness of the data distribution. This is only indirectly related to the use of a Student-t likelihood though, because the likelihood itself remains symmetric. Rather, it is a combination of our linear predictor including the highly right-skewed (and strongly predictive) `Base` variable in combination with a residual scale  $\sigma$  that is now allowed to be small since the outliers are taken into account separately via  $\nu$ . The good marginal in-sample fit doesn't necessarily mean our Student-t model is particularly good either. It is apparently just better than the normal models we have tried before. We will dive deeper into the topic of model comparison in Chapter 2.

## 1.5 Skewed linear models

Above, we did come up with the Student-t model to better handle outliers and, as an unexpected side effect, it also turned out to be helping with the skewness. But what if we address the skewness directly via a skewed likelihood family? Here, we will use the skew-normal distribution. Similar to Student-t, it generalizes the normal distribution by adding a third distributional parameter only that, for the skew-normal, the third parameter controls the skewness instead of the tail-heaviness.

If you check out text books for details on the skew normal distribution, or Wikipedia, you will usually find it being parameterized in terms of location  $\zeta$ , scale  $\omega$ , and skewness  $\alpha$ . If  $\alpha$  is

negative the distribution is left skewed, if  $\alpha$  is positive the distribution is right skewed, and if  $\alpha = 0$  the skew normal is just equal to the normal distribution.

For non-zero  $\alpha$ , location  $\zeta$  and scale  $\omega$  are related to the mean and standard deviation only indirectly. Specifically, we can compute the mean  $\mu$  and the standard deviation  $\sigma$  as

$$\mu = \xi + \omega \sqrt{\frac{2\alpha^2}{\pi(1+\alpha^2)}}, \quad \sigma = \omega \sqrt{\left(1 - \frac{2\alpha^2}{\pi(1+\alpha^2)}\right)}.$$

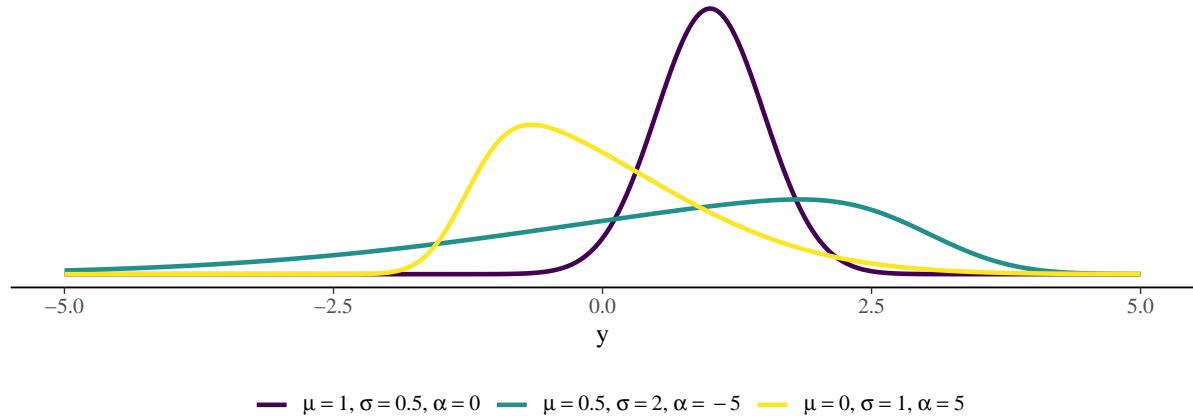


Figure 1.22: Three example densities of the skew-normal distribution varying in mean  $\mu$ , standard deviation  $\sigma$ , and skewness  $\alpha$ .

Especially the fact that  $\xi \neq \mu$  (for  $\alpha \neq 0$  and  $\omega > 0$ ) makes regression modeling based on this parameterization a bit awkward. We would very much prefer to predict  $\mu$  because we understand it much better than a scale and skewness affected location. For this reason, `brms` reparameterizes the skew normal in terms of  $\mu$ ,  $\sigma$ , and  $\alpha$ , some example densities of which you can see in Figure 1.22. From this mean-standard deviation parameterization, we can recover  $\xi$  and  $\omega$  as follows:

$$\xi = \mu - \omega \sqrt{\frac{2\alpha^2}{\pi(1+\alpha^2)}}, \quad \omega = \sigma \sqrt{\left(1 - \frac{2\alpha^2}{\pi(1+\alpha^2)}\right)}.$$

As a user of `brms`, you don't have to worry about these details. All you need to do is to set `family = skew_normal()`, which ensures that  $\mu$ , the main parameter to be predicted, is the mean of the likelihood and  $\sigma$  is the corresponding residual standard deviation. For the purpose of analysing the epilepsy data, our skew normal models looks as follows:

```
fit_epi_skew_normal1 <- brm(
  count ~ Trt * Base,
  data = epilepsy,
  family = skew_normal()
)
```

```
summary(fit_epi_skew_normal1)
```

Family: skew\_normal  
 Links: mu = identity  
 Formula: count ~ Trt \* Base  
 Data: epilepsy (Number of observations: 236)

Regression Coefficients:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	5.34	1.13	2.98	7.49	1.00	2175	2434		
Trt1	-3.22	1.36	-5.87	-0.60	1.00	2237	2470		
Base	0.15	0.03	0.10	0.21	1.00	2334	2239		
Trt1:Base	0.07	0.03	0.01	0.13	1.00	2214	2193		

Further Distributional Parameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sigma	7.39	0.37	6.72	8.16	1.00	2244	2357		
alpha	5.36	1.20	3.31	8.01	1.00	1824	2065		

From the summary output, we see (a) that the skewness **alpha** is clearly positive, indicating right skewness in the residuals and (b) the interaction of treatment and baseline being positive again. The latter points to the fact that, apparently, the skew normal model is still strongly influenced by the outliers, similar to the normal models, despite trying to account for the right skewness in the data. Let's look at the posterior predictive distribution:

```
pp_check(fit_epi_skew_normal1)
```

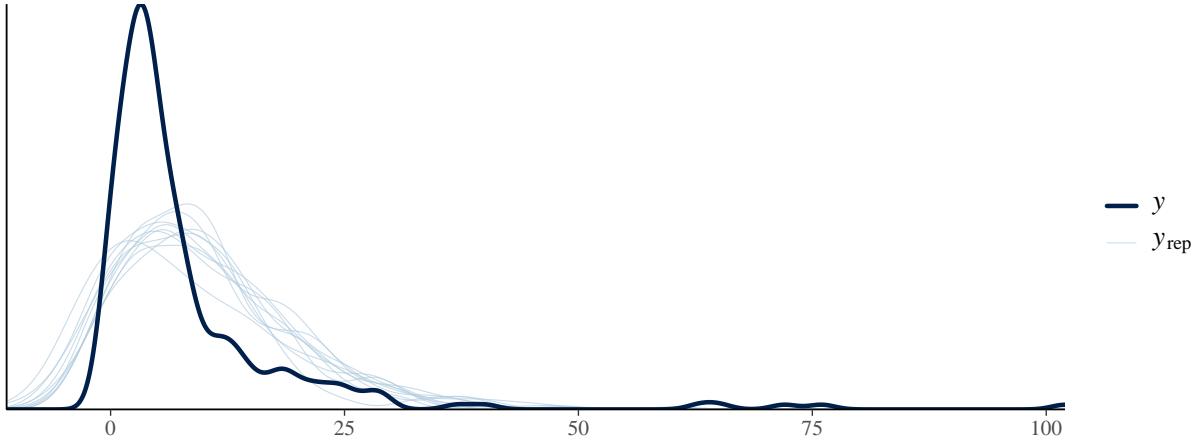


Figure 1.23: Default posterior predictive check for model `fit_epi_skew_normal1`.

From Figure 1.23, we clearly see the skew normal model struggles to model the outliers satisfactorily. It clearly predicts right-skewed data, but to a degree that is insufficient for the data at hand. Having seen this, it is not so surprising anymore that the skew normal model's posterior is much more similar to the normal model's than to the Student-t model's posterior. Within the set of models considered above, we would probably trust the Student-t model most. But keep in mind that none of these models is particularly good yet. We will come back to this later to now focus on more theoretical aspect of Bayesian linear models.

## 1.6 Gaussian linear models: Analytic vs. MCMC

So far in this chapter, we have looked at the models' posteriors from the approximating lens of MCMC sampling. However, it turns out that Bayesian linear regression is one of the few cases where we can actually compute the posterior analytically. Below, we will use this fact to validate our MCMC approximation. Readers who are only interested in the applied side of Bayes can safely skip this section without missing anything super important for the following chapters.

For the analytic case, we need a bit of math but I will try to keep it minimal since the purpose is to check the accuracy of the MCMC approximation not the computation of analytic posteriors. We continue to predict the response variable  $y$  based on a set of predictors  $x_1, x_2, \dots$ , which we gather as columns in the *design matrix*  $X$ . The design matrix also includes an “intercept predictor”  $x_0 = (1, 1, \dots)$  as first column so that the intercept  $b_0$  can be written as part of  $b = (b_0, b_1, \dots)$ . With this notation, we can write the (vectorized) likelihood of linear regression as

$$y \sim \text{normal}(X b, \sigma).$$

To make the example below simpler, we will fix  $\sigma = 12$ , which is a reasonable value given the data as we have seen above. Then, the only goal becomes to obtain the posterior of the regression coefficients. For the posterior of linear regression to be analytic, our prior on the coefficient vector  $b$  needs to be (multivariate) normal as well. Here, we use a multivariate normal prior with mean  $\mu_0$  and covariance matrix  $\sigma^2 \Sigma_0$ :

$$b \sim \text{multi\_normal}(b | \mu_{\text{prior}}, \sigma^2 \Sigma_{\text{prior}})$$

The multiplication of  $\Sigma_{\text{prior}}$  with the residual variance  $\sigma^2$  is not strictly necessary, but it simplifies the analytic expression of the implied posterior. Perhaps confusingly at first glance, but consistent with common statistics notation, I use mean-standard deviation parameterization for the univariate normal distribution but mean-(co)variance parameterization for the multivariate normal distribution. I will dive deeper into in the context of multilevel models in Chapter 4.

After doing the math to combine likelihood and prior (Gelman et al. 2013), we can see that the analytic posterior of Bayesian linear regression is

$$b | y \sim \text{multi\_normal}(b | \mu_{\text{post}}, \sigma^2 \Sigma_{\text{post}})$$

with

$$\Sigma_{\text{post}} = (X^T X + \Sigma_{\text{prior}}^{-1})^{-1} \quad \mu_{\text{post}} = \Sigma_{\text{post}} (\Sigma_{\text{prior}}^{-1} \mu_{\text{prior}} + X^T \hat{b}),$$

where  $\hat{b}$  is the maximum likelihood estimate computed as

$$\hat{b} = (X^T X)^{-1} X^T y.$$

Here,  $-1$  implies the matrix inverse and  $T$  the matrix transpose (i.e., switching rows and columns). The formula for  $\mu_{\text{post}}$  illustrates the close relationship of maximum likelihood and fully Bayesian estimation: In linear regression, the posterior mean is essentially a weighted average of the maximum likelihood estimate and the prior mean.

Let's apply these formulas concretely to the model `count ~ 1 + Trt + Base` on the epilepsy data. As priors on the coefficients, we choose

$$b_0 \sim \text{normal}(0, 3), \quad b_1 \sim \text{normal}(0, 5), \quad b_2 \sim \text{normal}(0, 1).$$

That is, before seeing the data, we assume all coefficients to be roughly around zero with varying amounts of (un)certainty. For the intercept, this reflects the assumption that without treatment and with zero previous epileptic seizures, the number of new seizures should be roughly zero. Of course, our prior also allows this number to be negative, something that is impossible in reality. But this is primarily due to the inappropriateness of the normal likelihood

for this response variable and cannot (easily) be fixed by the prior distribution. Accordingly, we have to accept for now that our prior implies partially impossible responses. For  $b_1$  and  $b_2$  the prior mean of zero contains some kind of “conservative” assumption that we first want to be convinced by the data before believing that the corresponding predictor is actually relevant. The corresponding standard deviations (5 and 1) reflect our understanding of the predictors’ scales. For example, it is unlikely, that the treatment decreases or increases the number of seizures by more than 10 (i.e., twice the standard deviation) on average and it is unlikely that with each baseline seizure, we will see an increase (or decrease) of than 2 seizures during the study. These are all rough considerations though. Prior specification is hard and continues to be a highly active area of research. We will dive deeper into the details of prior specification in later chapters.

In terms of the multivariate normal notation above, these prior choices imply  $\mu_0 = (0, 0, 0)$  and  $\Sigma_0 = (3, 5, 1)^2/\sigma^2 = (9, 25, 1)/144$  since we decided to fix  $\sigma = 12$ . With this choice, and after rounding to three digits, the posterior mean and covariance evaluate to

$$\mu_{\text{post}} = (-1.880, -1.071, 0.347) \quad \sigma^2 \Sigma_{\text{post}} = \begin{pmatrix} 1.664 & -0.938 & -0.022 \\ -0.938 & 2.109 & -0.003 \\ -0.022 & -0.003 & 0.001 \end{pmatrix}$$

In terms of the *marginal posteriors* of the coefficients (now rounded to two digits), this implies

$$b_0 \sim \text{normal}(-1.88, 1.29), \quad b_1 \sim \text{normal}(-1.07, 1.45), \quad b_2 \sim \text{normal}(0.35, 0.03).$$

### 1.6.1 Reproducing the analytic results with MCMC

MCMC is an approximate procedure in the sense that MCMC estimates are based on (dependent) random draws from the posterior and will thus vary slightly depending on which draws we have obtained exactly. How much the estimates are expected to vary is quantified by the *MCMC error*, that is, the (estimated) standard deviation of the distribution of MCMC estimates around their true value. For each parameter, we don’t just have a single MCMC error but as many as we have summary statistics. That is, for each parameter, we have a separate MCMC error for the posterior mean, the posterior standard deviation, as well as for every posterior quantile. Independently of whether or not we can obtain the analytic posterior, the MCMC error is a highly useful tool to understand the variation in posterior estimates attributable purely due to algorithmic (“random”) fluctuations. The computational details of MCMC errors are quite complex (Vehtari et al. 2021) but for our purposes, the above intuition should be sufficient.

Okay, let’s try to reproduce the analytic posteriors with MCMC in brms. For the first time, we will have to specify priors for our model explicitly. An immensely helpful function for this

purpose is `get_prior` which you can feed with the regular model input and it will give you all the information you need in order to specify your own priors:

```
get_prior(count ~ 1 + Trt + Base, data = epilepsy)
```

Table 1.6: Default priors of a Gaussian model with formula `count ~ 1 + Trt + Base` to be fitted on the `epilepsy` data.

prior	class	coef	group	resp	dpar	nlpar	lb	ub	tag	source
(hidden)	b									default
(hidden)	b		Base							default
(hidden)	b		Trt1							default
(hidden)	Intercept									default
(hidden)	sigma						0			default

The output shown in Table 1.6 is a data frame with one prior specification option per row. In the first column, we can find the default priors that brms uses but I have hidden them here since there are not in focus right now. In the second column, we see the parameter `class` where, for example, `b` stands for regression coefficients. We also see that the intercept  $b_0$  seems to have its own class (`Intercept`) for reasons that will become apparent soon. In the third column (`coef`), we see the specific coefficient name if we want to set specific priors on specific coefficients. The data frame has a bunch of other columns but we will only make use of them in later chapters. For now, the first few columns are all we need to specify our priors from above:

```
prior_epic_gaussian5 <-
  prior(normal(0, 3), class = "Intercept") +
  prior(normal(0, 5), class = "b", coef = "Trt1") +
  prior(normal(0, 1), class = "b", coef = "Base") +
  prior(constant(12), class = "sigma")
```

We can now pass this object to the `prior` argument for `brm` and run the model:

```
fit_epic_gaussian5 <- brm(count ~ 1 + Trt + Base, data = epilepsy,
                            prior = prior_epic_gaussian5)
```

To investigate whether the MCMC results match their analytic counterparts up to MCMC error, we need to see both the posterior summaries (here mean and SD) and their corresponding MCMC error estimates. These can be obtained with the help of functions from the *posterior* package:

```
draws <- as_draws(fit_epi_gaussian5, variable = "^\b_",
summarise_draws(draws, "mean", "sd", "mcse_mean", "mcse_sd")
```

Table 1.7: Posterior summaries of model `fit_epi_gaussian5`.

variable	mean	sd	mcse_mean	mcse_sd
b_Intercept	-2.80	1.40	0.02	0.02
b_Trt1	-0.84	1.49	0.02	0.02
b_Base	0.35	0.03	0.00	0.00

When comparing the analytic results with the posterior estimates shown in Table 1.7, we quickly see that something is not quite right, in particular with the posterior of  $b_0$ : Our MCMC estimate of the posterior mean together with its MCMC error are  $-2.8 \pm 0.021$  but the analytic posterior mean is  $-1.88$ . This difference is way to large than could be explained by the very small MCMC error. So what has gone wrong?

The answer lies in how brms parameterizes the intercept  $b_0$ . By default, the intercept parameter on which we specified the prior above is understood as the intercept when all predictors are centered, that is, have a mean of zero. This is done mainly for reasons of computational efficiency and automatically corrected after model fitting so that the model output contains the regular intercept as if the predictors had not been centered. Accordingly, as users, we don't have to worry about it unless we want to specify priors on the intercept.

Arguably, specifying priors on the “centered” intercepts is actually simpler because it is immediately related to the mean of the response variable, at least in linear models. In contrast, before centering the predictors, the intercept is often non-sensible, since predictor values of zero may be far away from any actually observed data. For example, if we used participants' age as predictor in a study with adults, an age of zero would not be a sensible value and we would struggle specifying a prior on the resulting intercept when an hypothetical adult had just been born.

Anyway, for the purpose of reproducing the analytical results, we need the standard intercept parameterization without automatic predictor centering. There are multiple equivalent ways to achieve this. First, we can use the protected predictor name `Intercept` in the model formula as follows:

```
count ~ 0 + Intercept + Trt + Base
```

Second, we can wrap our regular formula inside the `bf` function (short for `brmsformula`) and explicitly disable the auto-centering:

```
bf(count ~ 1 + Trt + Base, center = FALSE)
```

In both cases, the intercept will now behave like any other regression coefficient also when it comes to prior specification, so we can specify priors on it via `class = "b"` and `coef = "Intercept"`:

```
prior_epi_gaussian6 <-
  prior(normal(0, 3), class = "b", coef = "Intercept") +
  prior(normal(0, 5), class = "b", coef = "Trt1") +
  prior(normal(0, 1), class = "b", coef = "Base") +
  prior(constant(12), class = "sigma")
```

```
fit_epi_gaussian6 <- brm(bf(count ~ 1 + Trt + Base, center = FALSE),
                           data = epilepsy, prior = prior_epi_gaussian6)
```

Again, we extract the posterior summaries together with their MCMC errors:

```
draws <- as_draws(fit_epi_gaussian6, variable = "^b_", regex = TRUE)
summarise_draws(draws, "mean", "sd", "mcse_mean", "mcse_sd")
```

Table 1.8: Posterior summaries of model `fit_epi_gaussian6`.

variable	mean	sd	mcse_mean	mcse_sd
b_Intercept	-1.90	1.31	0.03	0.02
b_Trt1	-1.07	1.43	0.03	0.02
b_Base	0.35	0.03	0.00	0.00

The MCMC estimates shown in Table 1.8 are now very close to the analytic posterior summaries for all three parameters. The small deviations are well within a range of  $\pm 2 \times$  the corresponding MCMC error. In other words, at least for this model and dataset, we have validated our MCMC-based approximation against the analytic posterior. However, this validation only works for a few model classes, where the likelihood and prior match in just the right way for the posterior to have a closed form. For all other cases, we need entirely different procedures for MCMC validation, something we will discuss in more detail in [?@sec-bayesian-workflow](#).

## 1.7 Summary

In this chapter, we have learned about the basics of Bayesian linear models. In the process, we have seen a lot of general principles and methods of Bayesian modeling and how to realize

them in brms. This includes not only the core model specifications, but also the principles of sampling, posterior predictions, and some variations of likelihoods and priors. All of these principles and methods will continue to be relevant across this book.

# 2 Model Comparison

## 2.1 Setup

```
library(magrittr)
library(posterior)
library(ggplot2)
library(patchwork)
library(brms)
```

## 2.2 Introduction

In practice, we will almost never fit only a single model to our data. Too large is the space of potential assumptions we could make, or choose not to make, and too big are the a priori unknowns such that just knowing “the best model” right away is implausible<sup>1</sup>. Thus, we need to be able to somehow compare multiple models which each other in order to come up with a model (or a set of models) whose inference we choose to trust.

Model comparison is not an easy topic. Despite – or perhaps because of – having worked quite a bit on this topic, I am not feeling super comfortable writing about it. There are just so many things to keep in mind and take into account when comparing, choosing, or averaging over models. Many of you may take my thoughts expressed here seriously and follow my suggestions in their work. Just let me say that often, your understanding of the science, study design, and data will be as valuable in coming up with reasonable models as the statistical comparison methods, which tend to be a bit blind to the specifics of your case.

---

<sup>1</sup>In pre-registered studies, the concept of data-based model selection becomes challenging. In such studies, we often have to fix specific models to fit before even gathering the data. That said, this requirement should not stop us from performing additional analysis with other, potentially better models, once we have gathered the data.

## 2.3 Prediction vs. Explanation

Before starting to compare models, we should first remind us about our actual modeling goals. Often we are so stuck in the customs and traditions in our field that we forget to consciously map our research questions to modeling goals and to subsequently build models aligning with these goals. Looking back, I guess at least I have been guilty of this many times myself.

So what kinds of modeling goals there are? I argue that there are two main classes of modeling goals, which I refer to as *prediction* and *explanation* (Scholz and Bürkner 2023). With prediction I mean that the goal is to predict future data, that is, quantities that are at least in principle observable even if they are not yet available at the time of model building. If prediction is our main goal, we may choose to disregard the model structure (i.e., treat it as a black box) and just focus on how well the model predicts new data (by which I don't want to imply that I recommend disregarding model structure). With explanation I mean that the goal is to understand the latent (i.e., unobservable) state of the world. In this case, we care about the latent parts of the model, that is, its parameters or other latent quantities implied by them. Accordingly, we will seek to build models whose parameters we can argue to represent some latent aspects of the world that we care about.

If our modeling goal is prediction, we will have an easier time with real data. We just train our model on some data and then evaluate its *predictive performance* on new data, unseen during training. In a nutshell, this is the basic logic of building machine learning models. Their parameters are usually not interpretable and generally of secondary matter. All we care about, is how well the model predicts new data. In a way, this is a conceptually easy perspective as our goal is straightforward and we have a good idea how to practically evaluate it.

If our goal is explanation however, things become conceptually and practically more involved. In reality, model parameters may often not truly “exist” in the real world. And even if they did exist in some sense of the word, we will not know their true value in reality. If we did, why bother building any model to estimate them? Accordingly, doing model comparison that directly targets explanation is not something we can do for real data. We can only study it in the context of simulation studies where we know and fully control the ground truth. There, we can investigate a model's *parameter recovery* that is, how close the posterior of its parameters (or some other parameter-implied quantity of interest) is to the true value we aim to estimate, on average across simulated dataset. For me, parameter recovery is the statistical operationalization of explanation, as the latter is often used in a more vague sense in the philosophy of science literature.

In most chapters of this book, our main goal will be explanation, and prediction will only be of secondary concern. For example, in the epilepsy data, we care primarily about understanding the effect of the epilepsy treatment, which is of course a latent concept that we have no direct access to. At the same time, in most chapters of this book, we will analyse real data. So we cannot infer the explanatory capabilities of our models by means of studying their parameter recovery. Instead, we have to look at the models' predictive performance while paying close

addition to their structure. The hope is that, ideally, we can use prediction as a proxy for explanation, such that the better predicting models also provide more valid explanation for the latent concept of interest (e.g., the latent treatment effect). I consider this proxy use valid as long as our models are causally consistent, that is, are sufficiently representing the true causal structure of the involved variables. At least, we have found evidence for this perspective in relatively simple models (Scholz and Bürkner 2023) and I am currently unaware of any evidence contradicting it.

While causality is a highly important topic, and still under-appreciated in many sciences, I do not have the space in this book to talk about causality much. For an introduction to causality from a Bayesian perspective, I recommend reading, for example, Chapter 6 of Statistical Rethinking (McElreath 2019), a book I highly recommend more generally. It even has a brms translation, where all models are fitted with brms instead of with the rthinking package designed to accompany the book natively (Kurz 2019).

The main goal of this chapter is showing how to *practically* performing model comparison. That said, given the complexity of this topic, I need a bit more space for some preparatory thoughts before we can start running code.

## 2.4 Measuring Predictive Performance

There is not a single “best” predictive metric agnostic to our models and modeling goals. Accordingly, we need to choose metrics that fit to our goals. If possible, that is. In practice I often see a rather different approach. It is common to just apply whatever metrics are currently standard in a given field. And I cannot blame anyone for that. As you will see in this chapter and more generally in this book, I am a bit guilty of the same. I have my favorite methods and apply them consistently to most modeling situations I encounter. Perhaps in some of these situations, my choice of predictive metrics for model comparison is actually not particularly good after all.

As explained above, evaluating and comparing the predictive performance of Bayesian models will be the main focus of this chapter out of practical necessity. To make it easier for you to connect all the topics that will follow into a coherent picture, I will give you a quick overview first. Don’t worry if this does not fully make sense yet upon first reading. It will make more sense after you have finished reading this chapter.

First, we distinguish between absolute and relative predictive performance. Absolute predictive metrics *have clearly defined and interpretable optimal value*, indicating “perfect predictions” so to say. That way, a model’s predictive performance can be compared to this optimal value without having to resort to comparing multiple models. Relative predictive metrics do not have such an optimal value, at least not an easily interpretable one, so we have to compare them across multiple models to make any sense of the results. Absolute metrics can always be used in a relative manner, but not the other way round.

Second, we distinguish between prior and posterior predictive performance. Prior predictive metrics measure how well a model performs *before seeing the current data*, that is just based on the information in prior alone – whatever non-current-data information was used derive the prior in the first place. Posterior predictions measure how well a model performs *after seeing the current data*, that is based on both current data and prior information.

Third, we distinguish between in-sample and out-of-sample predictive performance. We measure in-sample predictive performance if the data we evaluate the model on (the *test data*) is the same (or a subset of) the data on which we have trained the model (the *training data*). Contrarily, we measure out-of-sample performance if the test data does not overlap with the training data. This distinction only makes sense when we care about posterior predictions because prior predictions have no training data to begin with. In that sense, prior predictions are always “out-of-sample”.

## 2.5 Absolute predictive performance

Let's start with evaluation of absolute predictive performance that can be performed on the basis of just a single model. Further, let's focus on in-sample predictions for now. We will reuse a bunch of models from the last chapter, thus continuing our case study of modeling the number of epileptic seizures. For more detailed explanation how we came to these models, please see the last chapter. Let's look at a simple Gaussian model first:

```
fit_epilepsy <- brm(count ~ Trt + Base, data = epilepsy)
```

In the last chapter, we have already seen some basic form of evaluating absolute prediction graphically, by means of posterior predictive checks. Two examples, you can find below (see Figure 2.1 and Figure 2.2):

Different PP-checks can be differently informative in different situations, so it is hard to give general recommendations for this. What I like is to start with the default PP-check (`type = "dens_overlay"`) just to see if things roughly make sense marginally. Additionally, the `"error_scatter_avg"` type resembling residual plots provides useful information into the error structure of our model. Both plots clearly point to problems in our simple Gaussian model fit to the epilepsy data. For a detailed overview of available PP-check options, I recommend checking out the help page `?bayesplot::PPC`.

### 2.5.1 Measures of Explained Variance

But what about numerical measures? One measure that may come to mind is the *coefficient of determination*, usually denoted by  $R^2$ , which is also often referred to as percentage of explained

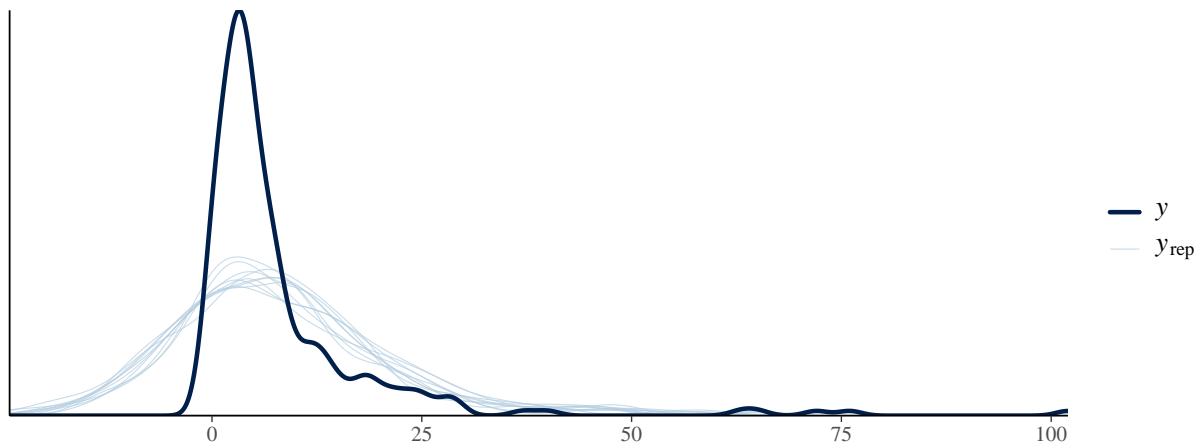


Figure 2.1: Default posterior predictive check (density overlay) for model `fit_epi_gaussian2`.

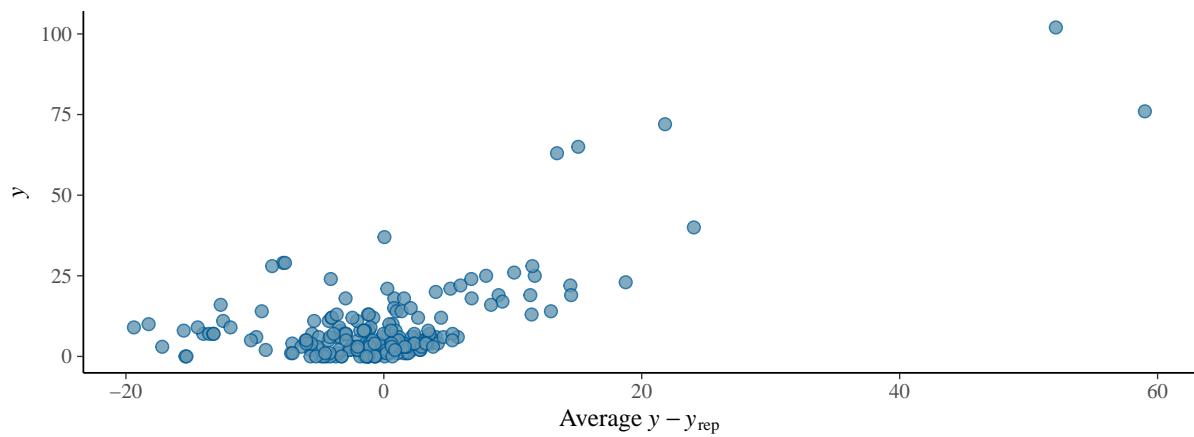


Figure 2.2: Posterior predictive check comparing observed responses (y-axis) with the predictive residuals (x-axis) of model `fit_epi_gaussian2`.

variance or percentage of explained variation. In its basic form,  $R^2$  is defined only for Gaussian linear models as:

$$R_{\text{basic}}^2 = 1 - \frac{\sum_{n=1}^N (y_n - \hat{y}_n)^2}{\sum_{n=1}^N (y_n - \bar{y})^2}$$

The numerator in the ratio describes the sums-of-squares of the errors, as we compute the difference between observed responses  $y_n$  and model-predicted responses  $\hat{y}_i$ . The denominator normalizes these squared errors on the total variation of the responses around their mean  $\bar{y}$ . Accordingly, the ratio describes the percentage of *unexplained variation*; and 1 minus this term is thus a measure of *explained variation*.

The definition of  $R_{\text{basic}}^2$  is not inherently Bayesian, but we can readily turn it into a fully Bayesian measure by considering posterior draws over the mean of the posterior predictive distribution (`posterior_epred`) as our  $\hat{y}_i$ . Then, we can compute draws from the posterior of  $R_{\text{basic}}^2$  as follows:

$$(R_{\text{basic}}^2)^{(s)} = 1 - \frac{\sum_{n=1}^N (y_n - \hat{y}_n^{(s)})^2}{\sum_{n=1}^N (y_n - \bar{y})^2}$$

Since `fit_epi_gaussian2` is just a Bayesian linear regression, we can actually just go ahead and compute (the posterior distribution of) this basic  $R_{\text{basic}}^2$ :

```
# compute draws of the predictive errors based on posterior_epred
errors <- predictive_error(fit_epi_gaussian2, method = "posterior_epred")
str(errors)
```

```
num [1:4000, 1:236] 2.3 2.89 3.63 2.79 4.03 ...
```

```
# sum errors over observations
error_variation <- rowSums(errors^2)
str(error_variation)
```

```
num [1:4000] 15177 15239 15233 15037 15072 ...
```

```
# compute R2_basic
overall_variation <- sum((epilepsy$count - mean(epilepsy$count))^2)
R2_basic_epi_gaussian2 <- 1 - error_variation / overall_variation
posterior_summary(R2_basic_epi_gaussian2)
```

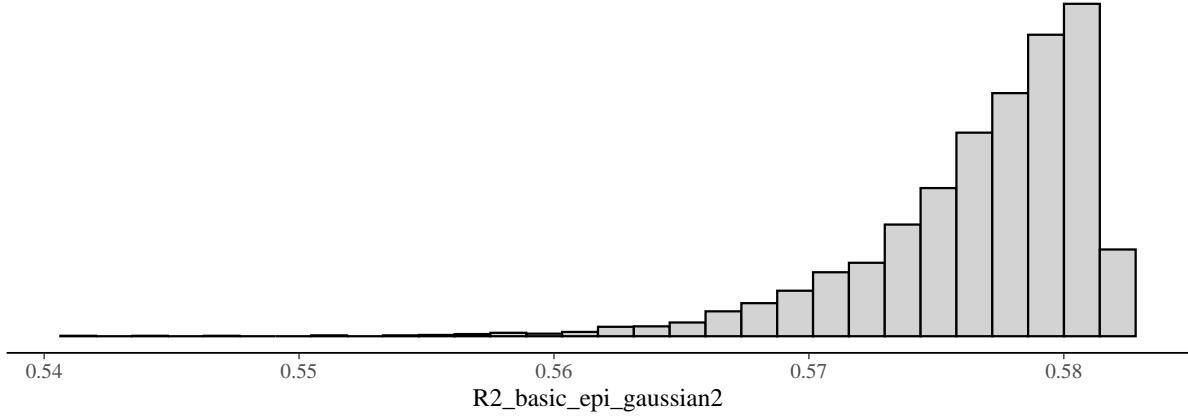


Figure 2.3: Posterior histogram of  $R^2_{\text{basic}}$  for model `fit_epi_gaussian2`.

Estimate	Est.Error	Q2.5	Q97.5
[1,] 0.5766577	0.004486575	0.5651756	0.5816194

The results tell us that roughly between 56% and 58% variation in the responses is explained by the model, least when considering in-sample fit for now. This is quite substantial and likely primarily driven by the `Base` predictor of baseline epileptic seizures unsurprisingly showing a very strong relation to the number of epileptic seizures during treatment.

If we were to compute a basic linear regression with the `lm` function instead of with `brm` and checked out the obtained  $R^2$ , we would see that the point estimate obtained from the former would be in the bulk of the posterior distribution obtained from the latter; at least when using weakly-informative priors. Check it out yourself as an exercise.

The  $R^2_{\text{basic}}$  is a great starting point, but it doesn't readily generalize to models that are much more complicated than Gaussian linear models. In particular it doesn't readily generalize to most other likelihood families. Accordingly, it makes sense to introduce a more general form of  $R^2$  that we can apply to (almost) all brms models, regardless of what likelihood families they have. A quite straightforward general definition was given by Gelman et al. (2019). Following the intuition about  $R^2$ , their measure is based on the ratio of explained variance and the sum of explained and error variance:

$$R^2_{\text{general}} = \frac{\text{explained variance}}{\text{explained variance} + \text{error variance}} = \frac{\text{Var}(\hat{y})}{\text{Var}(\hat{y}) + \text{Var}(\hat{e})}$$

Where  $\text{Var}(\hat{y})$  is the variance of the posterior predicted mean over observations (again `posterior_epred`) and  $\text{Var}(\hat{e})$  is the variance of the model-implied errors over observations, where  $\hat{e}_n = y_n - \hat{y}_n$ . Notice that this is the same error definition as was used in  $R^2_{\text{basic}}$ . Still, even for Gaussian linear models, the two discussed definition of  $R^2$  are not identical, but their interpretation as percentage of explained variance is the same.

We can readily formulate a posterior-draws version of  $R_{\text{general}}^2$  by computing  $\text{Var}(\hat{y})$  and  $\text{Var}(\hat{e})$  for each posterior draw, which then implies posterior draws of  $R_{\text{general}}^2$ :

$$(R_{\text{general}}^2)^{(s)} = \frac{\text{Var}(\hat{y})^{(s)}}{\text{Var}(\hat{y})^{(s)} + \text{Var}(\hat{e})^{(s)}}$$

We could code this up by hand, but brms already supports it natively via the `bayes_R2` method. For our Gaussian linear regression model, we get

```
bayes_R2(fit_ephi_gaussian2)
```

	Estimate	Est.Error	Q2.5	Q97.5
R2	0.5793621	0.02697763	0.5216896	0.6266168

So, in qualitative accordance with the basic  $R^2$ , we see that this model “explains” around 58% of the response’s variance. What is more, we are quite certain about this estimate with the 95% CI of the  $R^2$  posterior ranging from 52% to 63%. By default, the output already provides summaries over posterior draws. If we were interested in the draws themselves, we could add argument `summary = FALSE` as is the case in a lot of other post-processing functions being able to output draws or summaries thereof.

How are things looking with our slightly more complicated but still Gaussian linear model involving an additional interaction term between treatment and baseline epilepsy score?

```
fit_ephi_gaussian3 <- brm(count ~ Trt * Base, data = epilepsy)
```

```
bayes_R2(fit_ephi_gaussian3)
```

	Estimate	Est.Error	Q2.5	Q97.5
R2	0.6065522	0.02529101	0.5519936	0.6499702

Apparently, at least according to  $R^2$ , adding the interaction makes bare any difference and just adds around 2%-3% more explained variance. Additionally noting the strong overlap between  $R^2$  posterior distributions of both model, this difference is unlikely to be substantial in most sensible definitions of that word. That said, I still remember a professor of mine talking about another professor working in neuroscience stating that: “Whenever a model explains for than 1% of the fMRI data’s variance, she gets nervous.” What we can learn is that small changes (in term of model predictions,  $R^2$ , etc.) can mean a lot depending on the context and the signal strength we can reasonably expect to be in our data.

### 2.5.2 Measures of Squared Errors

Another common measure of predictive performance that finds widespread application in statistics and machine learning is that root mean squared error (RMSE). As with  $R^2$ , there is not necessarily one single definitions, but rather several, all following the same intuition of measuring (the root of) an average squared difference between target values and values used to approximate the targets. For example, we can define the RMSE simply as

$$\text{RMSE}_{\text{basic}}^{(s)} = \sqrt{\frac{1}{N} \sum_{n=1}^N (y_n - \hat{y}_n^{(s)})^2},$$

which you will easily identify as the numerator of  $R^2_{\text{basic}}$  defined earlier. As such,  $\text{RMSE}_{\text{basic}}$  is just the “unnormalized” version of  $R^2_{\text{basic}}$ . While the latter is scale free and varies between 0 and 1, the former has the same scale as the response  $y$  and can vary between 0 and infinity in general. If we care about and understand the scale of  $y$ , the RMSE may be more interpretable. For example, in the context of logistics, we are interested, among others, in arrival time predictions of goods. To make sure that our supply chains run smoothly, we would want our arrival time predictions to be as accurate as possible, e.g., not being off by more than a certain number minutes on average, say, 60 minutes. This is something we could easily estimate with an RMSE because we understand the scale of our response (time in minutes) and as such understand the scale of the RMSE. In contrast, an  $R^2$  measure would be a bit useless since we wouldn’t know whether or not, say, 80% of explained variance would mean we would reach our desired average prediction error of less than 60 minutes.

When computing  $\text{RMSE}_{\text{basic}}^{(s)}$ , we can largely reuse code that we already needed for  $R^2_{\text{basic}}$ :

```
# compute draws of the predictive errors based on posterior_epred
errors_epি_gaussian3 <-
  predictive_error(fit_epি_gaussian3, method = "posterior_epred")
str(errors_epি_gaussian3)
```

```
num [1:4000, 1:236] 1.0331 -0.0103 2.949 3.2593 1.4767 ...
```

```
# root mean of squared errors over observations
rmse_basic_epি_gaussian3 <- sqrt(rowMeans(errors_epি_gaussian3^2))
str(rmse_basic_epি_gaussian3)
```

```
num [1:4000] 7.72 7.8 7.77 7.77 7.73 ...
```

```
histogram(rmse_basic_epি_gaussian3)
```

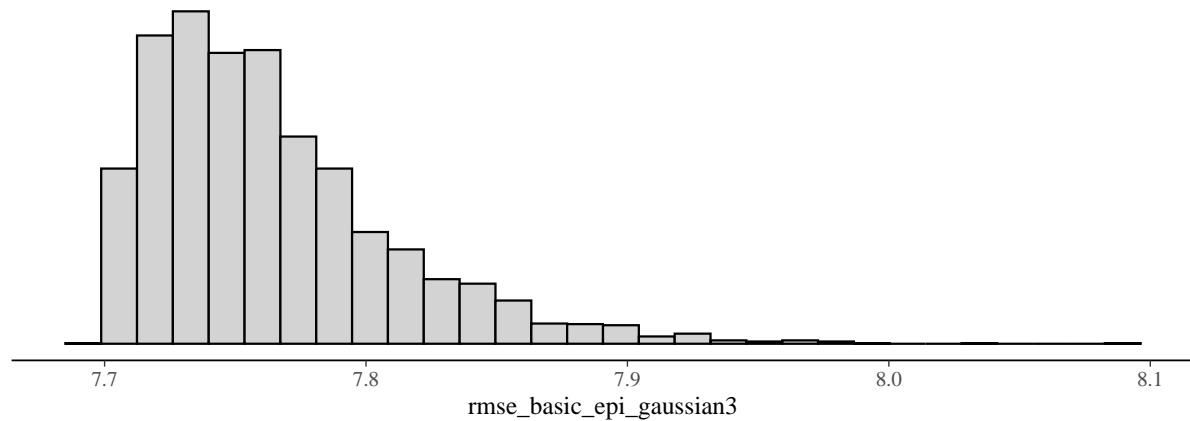


Figure 2.4: Posterior histogram of RMSE<sub>basic</sub> for model `fit_epি_gaussian3`.

Accordingly, the average posterior predictor error of `fit_epি_gaussian3` is very likely between 7.7 and 7.9 seizure counts, which admittedly is quite a lot, but not surprising given how sub-optimal of a model we know all the Gaussian models to be for this data.

We can also define useful RMSEs in other ways. For example, while RMSE<sub>basic</sub> computes a mean square error of observations  $n$  for each posterior draw  $s$ , and thus yields a posterior distribution over RMSE values, we could also exchange the use of  $n$  and  $s$  and compute a mean square error over draws  $s$  for each observation  $n$ :

$$\text{RMSE}_{\text{alt},n} = \sqrt{\frac{1}{S} \sum_{s=1}^S (y_n - \hat{y}_n^{(s)})^2},$$

In this case, we would get a distribution of RMSE over observations, where each individual RMSE value would be computed over the posterior predictive distribution of a single observation. Both of the above RMSE measures are fully Bayesian as they take into account the uncertainty in the posterior distribution, albeit in different ways.

```
# notice that here we have used colMeans instead of rowMeans as for rmse_basics
rmse_alt_epি_gaussian3 <- sqrt(colMeans(errors_epি_gaussian3^2))
str(rmse_alt_epি_gaussian3)
```

```
num [1:236] 1.965 0.979 1.034 1.819 10.955 ...
```

```
histogram(rmse_alt_epi_gaussian3)
```

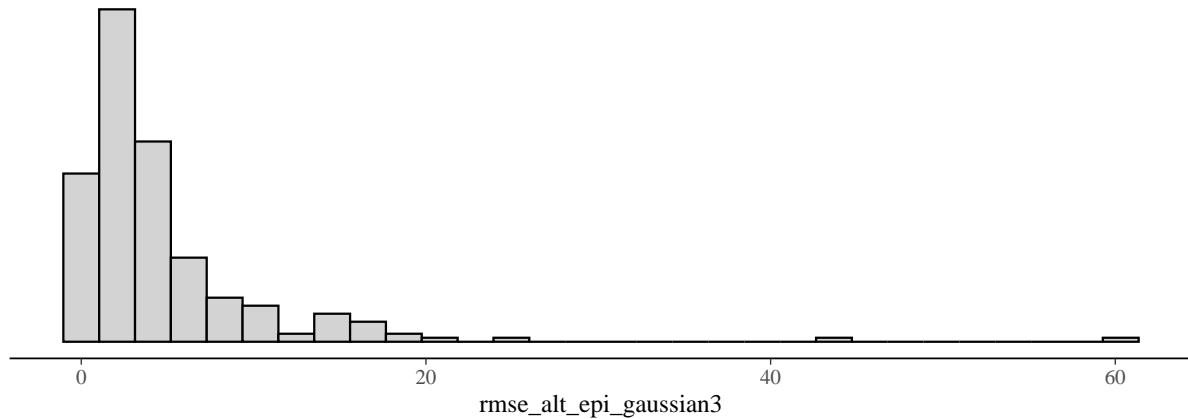


Figure 2.5: Posterior histogram of RMSE<sub>alt</sub> for model `fit_epi_gaussian3`.

Now, that looks completely different. In the above histogram we see one RMSE value per observation, which clearly tell us the predictions for some observations are extremely bad; up to 60 seizure counts even. We didn't see these extreme results in RMSE<sub>basic</sub> because there we averaged the squared error over observations.

Also, both are not the typical RMSE measure you see, for example, in machine learning papers. There, we typically see only a point estimate  $\hat{y}_n$  being used to represent the model-implied predictions, instead of a (posterior) distribution over such predictions for each  $n$ . For example, for a Bayesian model this point estimate could simply be the posterior mean  $\hat{y}_n = \sum_{s=1}^S \hat{y}_n^{(s)} / S$ . When using such a point prediction approach, our RMSE definition becomes

$$\text{RMSE}_{\text{point}} = \sqrt{\frac{1}{N} \sum_{n=1}^N (y_n - \hat{y}_n)^2}.$$

This is pretty similar to RMSE<sub>basic</sub> except that we have gotten rid of the posterior uncertainty already when computing  $\hat{y}_n$ .

```
# extract a point estimate of the predictions per observation
ppmean_epi_gaussian3 <- colMeans(posterior_epred(fit_epi_gaussian3))
str(ppmean_epi_gaussian3)
```

```
num [1:236] 3.27 3.27 1.95 2.48 17.88 ...
```

```
# compute RMSE based on the responses and point predictions
(rmse_point_epi_gaussian3 <- sqrt(mean((epilepsy$count - ppmean_epi_gaussian3)^2)))
```

```
[1] 7.697754
```

This value is smaller than even the smallest draw from the posterior of  $\text{RMSE}_{\text{basic}}$  above, because  $\text{RMSE}_{\text{basic}}$  averages over the posterior before computing the sum of squares. This also tells us that comparing, a model from whom we have obtained posterior draws with a model from whom we have obtained only point estimates can be seriously misleading depending on what model comparison metrics we choose to apply. Accordingly, please make sure that all models contain reasonable estimates of posterior uncertainty, that is, are ideally all fitted with the same posterior approximation algorithm (say, with Stan's main MCMC sampler). Otherwise, we may conclude that one model is better in terms of RMSE simply because its output did not contain posterior uncertainty at all, or because it drastically underestimated said uncertainty.

## 2.6 Relative predictive performance

So far we have only looked at one model at a time, which may, in some cases, be sufficient. But in general, it is more common to compare multiple models against each other and thus investigating their *relative predictive performance*. As a competitor to our Gaussian model, let's consider the corresponding Student-t model from the last chapter:

```
fit_epi_student1 <- brm(
  count ~ Trt * Base, data = epilepsy, family = student()
)
```

RMSE values can be computed as previously, for example, for  $\text{RMSE}_{\text{alt}}$ :

```
errors_epi_student1 <-
  predictive_error(fit_epi_student1, method = "posterior_epred")
rmse_alt_epi_student1 <- sqrt(colMeans(errors_epi_student1^2))
```

We can now even compute the pointwise (per-observation) difference in RMSE values:

```
rmse_alt_diff <- rmse_alt_epi_student1 - rmse_alt_epi_gaussian3
str(rmse_alt_diff)
```

```
num [1:236] 0.22 -0.636 -0.472 0.146 -0.919 ...
```

This RMSE difference has a mean of -0.36 seizure counts, which isn't that large I would say, in particular in light of the substantial variation in pointwise RMSEs as shown in Figure 2.6. We can attempt to understand the “significance” of the RMSE difference by comparing the mean difference with its corresponding standard error, that is simply computed as the standard deviation divided by the square root of  $N$ :

```
se_mean <- function(x) {
  sd(x) / sqrt(length(x))
}

se_rmse_alt_diff <- se_mean(rmse_alt_diff)
```

We get a standard error of 0.37 which is roughly of the same size as the mean difference itself. In another words, the mean difference is only about 1 standard error units away from zero. That is not much. For reference, when assuming a normal distribution and having a not too small sample size, we would call a difference significant (in a frequentist sense) at an  $\alpha$  level of 5% if the mean difference is further away from zero than 2 standard error units. So, in our case, we would probably need to have at least twice the current mean difference to start believing in its reliability.

Accordingly, as seen above, we should consider both the difference between models in “raw” RMSE values and in “scaled” RMSE values relative to the corresponding standard error. As we will see in a bit, both is relevant not only for RMSE but also for other metrics of model performance. In our case here, both the raw and the scaled difference in RMSE values is small so, on this basis, we do not see convincing evidence of one model over the other.

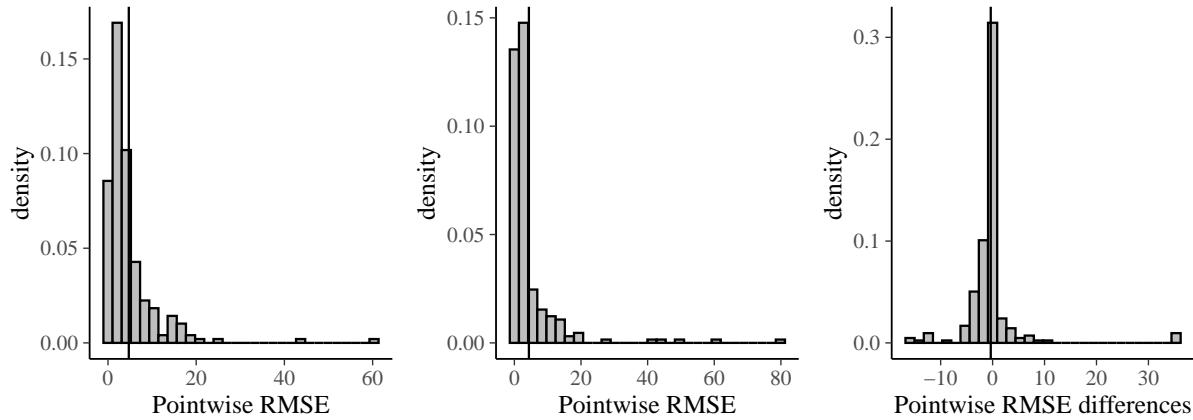


Figure 2.6: Histograms of the pointwise  $\text{RMSE}_{\text{alt}}$  contributions (based on `posterior_epred`) for models `fit_epi_gaussian3` and `fit_epi_student1` as well as for their pointwise  $\text{RMSE}_{\text{alt}}$  differences.

So far, we have been looking at RMSEs that were directly related to common  $R^2$  definitions that we introduced before. However, all of them are ignoring the variation induced by

the likelihood, that is, do not actually incorporate all the model uncertainties. Remember the difference between `posterior_epred` and `posterior_predict` from Chapter 1? Above, we have always used `posterior_epred`, but nothing stops us to compute the RMSEs with `posterior_predict` as well. This would also be closer to more standard posterior-predictive checks which (almost) all use predictions of new responses rather than predictions on the expected response mean. Also, we would expect the Student-t model to better show its strengths there because of its ability to deal with outliers.

Let's compute  $\text{RMSE}_{\text{alt},n}$  again but compute the errors from the `posterior_predict` predictions:

```
# "posterior_predict" is the default method in predictive_error()
errors_pp_epi_gaussian3 <- predictive_error(fit_epi_gaussian3)
rmse_pp_alt_epi_gaussian3 <- sqrt(colMeans(errors_pp_epi_gaussian3^2))

errors_pp_epi_student1 <- predictive_error(fit_epi_student1)
rmse_pp_alt_epi_student1 <- sqrt(colMeans(errors_pp_epi_student1^2))

rmse_pp_alt_diff <- rmse_pp_alt_epi_student1 - rmse_pp_alt_epi_gaussian3
```

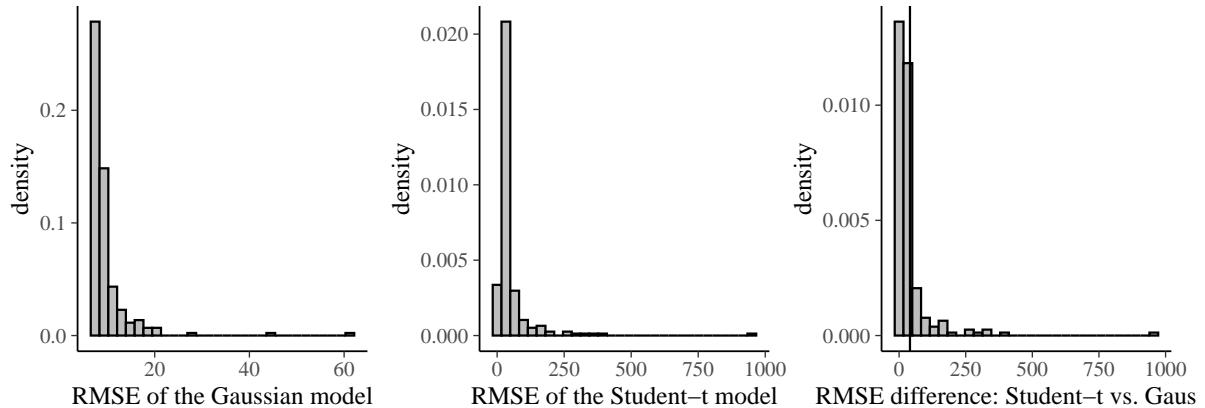


Figure 2.7: Histograms of the pointwise  $\text{RMSE}_{\text{alt}}$  contributions (based on `posterior_predict`) for models `fit_epi_gaussian3` and `fit_epi_student1` as well as for their pointwise  $\text{RMSE}_{\text{alt}}$  differences.

The results are presented in Figure 2.7. We see that, perhaps to our surprise, that the Student-t model has terrible RMSE for some observations; much worse than the normal model. What happened here? Remember the posterior predictive checks of the Student-t model from last chapter whose x-axis we had to restrict in order to see something? The same has happened here: If a Student-t likelihood has small degrees of freedom (as implied here by the data), then the posterior predictive distribution has to produce occasional big outliers. They may

then blow up the scales of plots or strongly influence summary measures that are strongly influencable by outliers such as mean, variance, or standard deviation.

Does that mean the Student-t model is worse than the normal model? As we concluded already in the last chapter, not necessarily. It just depends on what predictive metric we choose to evaluate. As we had seen above, when using expected posterior predicted values (`posterior_epred`), the Student-t model was even slightly better than the Gaussian model even when including the very high RMSEs of several outliers. This is interesting insofar as, conceptually, the RMSE is very closely related to the normal models. If we were to write down the log-likelihood of the normal model, it looks something like this:

$$\log p(y | \theta) = \left( \sum_{n=1}^N -\frac{1}{2} \frac{(y_n - \mu_n)^2}{\sigma^2} \right) - N \log(\sqrt{2\pi}\sigma),$$

where  $\mu_n$  is the predicted mean of observation  $n$  and  $\sigma$  is the residual standard deviation parameter assumed constant over observations. If we just focus on the first term and are ignoring the scaling by  $\sigma$ , we see that the log-likelihood of a Gaussian model (also known as *squared loss function* in machine learning) looks very similar to the RMSE. For the latter, we just have the square-root around the sum but that doesn't really change much here. Put differently, a Gaussian model already tries to get the best mean-squared error possible. Accordingly, we can expect Gaussian model to have some advantage in predictive metrics based on mean-squared errors such as the RMSE. That is fine, if we know that minimizing RMSE is what we want from our models. But if we have no clue what metric we actually care for and just want to have a general purpose metric to compare models "holistically" in some sense, RMSE and close friends may not be ideal. For this reason, we will next learn about more general predictive metrics that are agnostic to the specific model or model structure.

### 2.6.1 Likelihood Density Scores

Above we have talked all kinds of variations of  $R^2$  and RMSE metrics. In particular, we have seen how the RMSE is related to the log-likelihood of Gaussian models. So what if we just go ahead and use the log-likelihood of models as predictive metric more generally? After all the log-likelihood plays a pivotal role not only in to derive the posterior in Bayesian statistics but also to obtain maximum likelihood estimates in a frequentist framework. Intuitively, the higher the likelihood of the data given the model's parameters estimates (represented as either posterior draws or point estimates), the better the fit of the model to the data. Indeed, many important predictive metrics, Bayesian or otherwise, are based on log-likelihood scores.

Remember from [?@sec-bayesian-inference](#) that, if the observations  $y = (y_1, y_2, \dots, y_N)$  are assumed to be independent conditional on the parameters  $\theta$ , the likelihood factorizes into a product of pointwise (per-observation) terms:

$$p(y | \theta) = p(y_1, y_2, \dots, y_N | \theta) = \prod_{n=1}^N p(y_n | \theta)$$

If we then take the log, the product of the pointwise terms becomes the sum of the pointwise log terms:

$$\log p(y | \theta) = \sum_{n=1}^N \log p(y_n | \theta)$$

Since the log is a strictly monotonic transformation, we are not changing anything fundamental by looking at log scores – at least as long as we are careful. However, we are making a lot of math much simpler by now working with sums instead of products. In particular, this concerns computing gradients because the gradient of a sum is just the sum of the individual (pointwise) gradients. In fact, much of the modern statistics and machine learning relies on this property, so we are working with something quite fundamental here. That said, we can also work with log-likelihood metrics if the likelihood doesn't factorize easily or even at all, but the math becomes more cumbersome in this case (Bürkner, Gabry, and Vehtari 2021).

As a regular user of brms, you do not have to worry about these things, since the package comes with a dedicated `log_lik` method that does all the required math for you. Let's us this method to compare the Gaussian with the Student-t model once more. But first, we take a look at the Gaussian model in isolation:

```
ll_epi_gaussian3 <- log_lik(fit_epи_gaussian3)
str(ll_epi_gaussian3)
```

```
num [1:4000, 1:236] -2.97 -2.98 -3.02 -3.03 -3.01 ...
```

The output of `log_lik` has the same structure and `posterior_predict` and friends, that is, it has as many columns as we have observations and as many rows as we posterior draws. Similar to the `RMSEalt` metric earlier, we will average over posterior draws per observation (to be justified more shortly) such that we obtain one log-likelihood value per observation:

```
llm_epi_gaussian3 <- colMeans(ll_epi_gaussian3)
str(llm_epi_gaussian3)
```

```
num [1:236] -3 -2.98 -2.98 -3 -3.97 ...
```

We use `colMeans` here to create mean log-likelihood scores per observation, averaged over draws. You can find the corresponding histogram in Figure 2.8. Again, we clearly see two outlying observations that the Gaussian model predicts very badly – remember higher likelihood values are better – but otherwise the scale of the values is hard to interpret. Although log-likelihood values are usually negative, they might also be positive. There is also no particular value that we could generally identify as the “best-possible” log-likelihood. So we generally need multiple models to compare against in order to make sense of their log-likelihood scores. They are only relative predictive performance metrics.

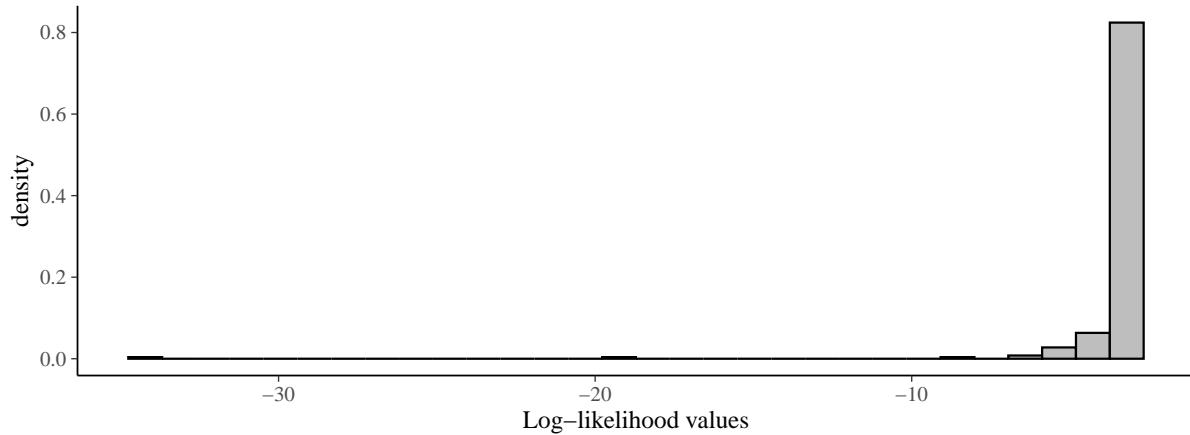


Figure 2.8: Per-observation (pointwise) log-likelihoods of model `fit_epi_gaussian3`.

Let’s compute the log-likelihood scores also for the Student-t model and compare them against those of the Gaussian model:

```
llm_epi_student1 <- colMeans(log_lik(fit_epi_student1))
llm_epi_diff <- llm_epi_student1 - llm_epi_gaussian3
str(llm_epi_diff)
```

```
num [1:236] 0.481 1.128 1.092 0.582 -1.272 ...
```

A visualization of the comparison is provided in Figure 2.7. The mean of the log-likelihood differences (vertical black line on the right plot) is slightly positive which points to a slightly better fit of the Student-t model. For reasons, I will explain later on, it is more typical to work with *sums* instead of *means* of log-likelihood values over observations, a quantity that we call LPD:

$$\text{LPD} = \sum_{n=1}^N p(y_n | y) \approx \sum_{n=1}^N \left( \frac{1}{S} \sum_{s=1}^S \log p(y_n | \theta^{(s)}) \right), \quad \theta \sim p(\theta | y)$$

So we compute our summary of the LPD differences simply as

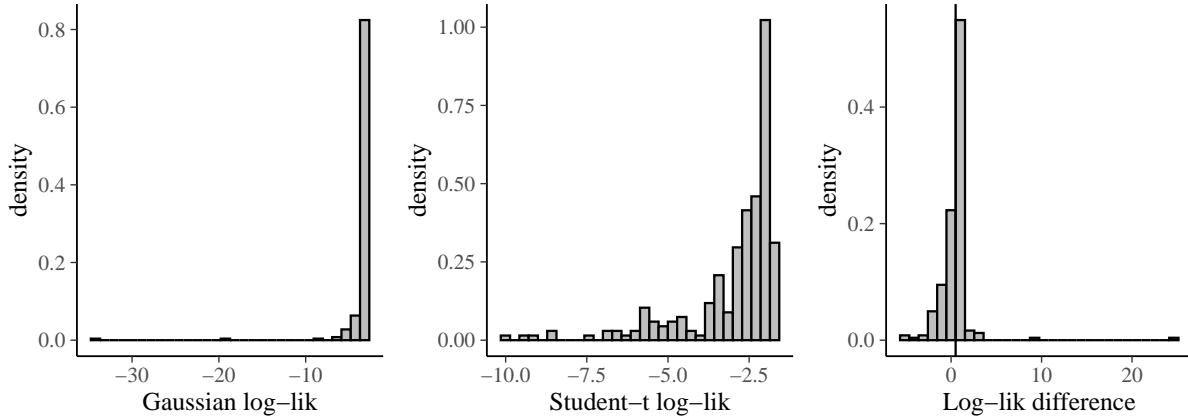


Figure 2.9: Pointwise log-likelihood values of model `fit_epi_gaussian3` and `fit_epi_student1` as well as their pointwise log-likelihood differences.

```
(lpd_epi_diff <- sum(llm_epi_diff))
```

```
[1] 118.1673
```

The corresponding standard error is also not particularly difficult to obtain:

```
se_sum <- function(x) {
  sd(x) * sqrt(length(x))
}

(se_lpd_epi_diff <- se_sum(llm_epi_diff))
```

```
[1] 31.02064
```

Accordingly, the Student-t model is clearly better, both in absolute terms (LPD difference of 118 is quite a bit) and in relative terms (about  $118/31 \approx 3$  SEs away from zero). We had seen this before in the `posterior_epred`-based RMSE comparisons, but saw the opposite pattern for the `posterior_predict`-based RMSEs. Does that mean the `log_lik` is more similar to `posterior_epred`? Actually, quite the opposite. Both `log_lik` and `posterior_predict` include the aleatoric uncertainty in the likelihood in its values whereas `posterior_epred` does not. However, `log_lik` and `posterior_predict` are doing something fundamentally different with this uncertainty. While `posterior_predict` uses it to *sample* model-implied responses from the posterior predictive distribution, `log_lik` merely *evaluates* the posterior predictive distributions at the observed responses. As a result, if a model has long likelihood

tails (e.g., the Student-t model), it will be able to put higher likelihood on outlying observations, thus having higher log-likelihood scores. Since this does not require to sample new responses from the likelihood tails, we do not run into the issue of accidentally creating huge and unrealistic predictions that may strongly affect simulation-based predictive metrics using `posterior_predict`.

When I first encountered this situation, I was quite confused. But it has opened my eyes for all the difficulties and strange things that may happen in model comparison. What you have seen above is my attempt to prepare you mentally for the things you may encounter yourself, or find in papers you read.

Anyway, we cannot not do model comparison regardless of the challenges that may await us. So if I need to make a recommendation, then I would argue for using log-likelihood scores as basis for model comparison. Unless you have a clear reason to prefer another kind of predictive metric on substantive grounds (Vehtari and Ojanen 2012).

That said, please don't use the basic `log_lik`-based metric above, since we have so far ignored one major aspect that concerns on which data to evaluate model fit: In-sample vs. out-of-sample predictions. So far, we have only evaluated model predictions for observations used to fit the model. So the model had the opportunity to adjust to these very observations that we have used to evaluate its model fit. This is likely to bias model comparison in favor of more complicated models as they are more flexible and can adjust to the specific observations better than they see during training, a process known as *overfitting*. There are many simple examples that can illustrate this problem. For example, when have 3 data points and fit a quadratic model of the form  $y_n \sim \text{normal}(b_0 + b_1 x_i + b_2 x_i^2, 1)$ , we have 3 parameters  $(b_0, b_1, b_2)$  trained on 3 observations  $(y_1, y_2, y_3)$ . This results in a model that can perfectly predict these three observations, but is likely to generalize poorly to other observations, unseen during training. I won't spell out this example fully here, but if you want to read more, I recommend Chapter 7 of McElreath (2019).

Accordingly, if we don't want to be fooled by overfitting models, it is essentially to evaluate out-of-sample predictive performance to check the models' generalization abilities Vehtari and Ojanen (2012). This is what we will discuss in detail next.

## 2.7 Out-of-sample predictions

The first question we encounter when trying to evaluate out-of-sample predictions is where to actually get new data to evaluate the model on? In the (usual) absence of any actual new data, we simply go ahead and split our data at hand into two parts, one used for training (the *training data*) and the other we use for model evaluation (the *test data*)<sup>2</sup>. That's easy and

---

<sup>2</sup>In machine learning workflows, we often further differentiate between *validation data* and (actual) test data. The validation data are used as out-of-sample data to evaluate model fit during the model building procedure. The actual test data are then only used upon the final model evaluation. In machine learning challenges such

readily implemented for our case study. Let's say we choose to use roughly 80% as training data roughly 20% as test data, whereas the assignment of each observation to training or test data is done randomly:

```
set.seed(8973)
epilepsy <- epilepsy %>%
  mutate(use = sample(c("training", "test"), n(), replace = TRUE, prob = c(0.8, 0.2)))

table(epilepsy$use)

  test training
    43      193
```

```
epilepsy_train <- epilepsy %>% filter(use == "training")
epilepsy_test <- epilepsy %>% filter(use == "test")
```

We continue to the training data  $y$  as before and introduce the test data  $\tilde{y}$  to indicate that they are separate from the training data. Using a total of  $\tilde{N}$  test observations, we arrive at the out-of-sample version of the LPD, which we will call *expected* LPD, or ELPD in short:

$$\text{ELPD} = \sum_{n=1}^{\tilde{N}} p(\tilde{y}_n | y) \approx \sum_{n=1}^{\tilde{N}} \left( \frac{1}{S} \sum_{s=1}^S \log p(\tilde{y}_n | \theta^{(s)}) \right), \quad \theta \sim p(\theta | y) \quad (2.1)$$

The term “expected” stems from the fact that the above ELPD is an estimate of the log-predictive densities that would be expected if we knew the true data-generating process and could integrate over this process analytically. In some papers, what I call ELPD is simply called LPD or test-data LPD, so it's important to not cling on the specific names but actually check what mathematical definitions are used.

We are now ready to fit the models on the training data and then evaluate and compare model fit via ELPD on the test data.

```
llm_epi_diff_test <- llm_epi_student1_test - llm_epi_gaussian3_test

histogram(llm_epi_diff_test)
```

---

as those organized by Kaggle (<https://www.kaggle.com/>), test data are typically hidden from the model builder such that they cannot overfit on them. In that sense, what we call “test data” here in this chapter is perhaps closer to what we would call “validation data” in machine learning.

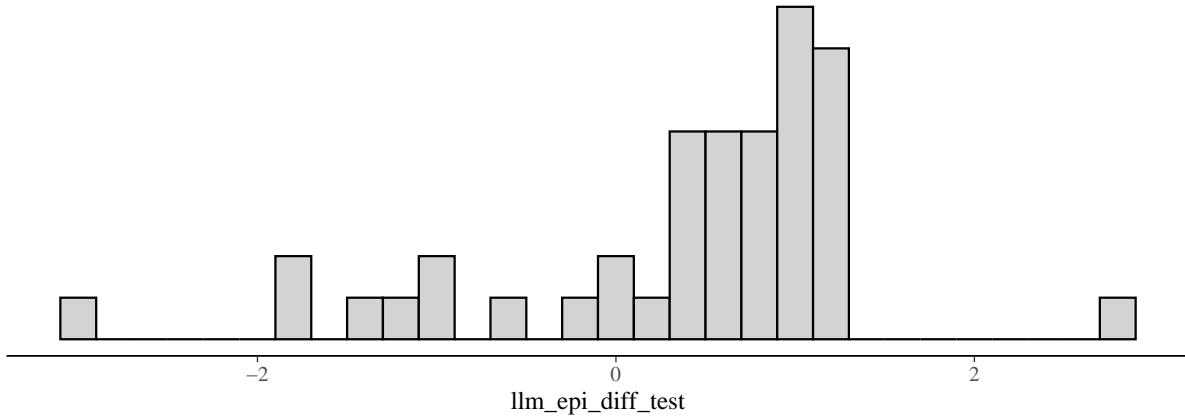


Figure 2.10: Pointwise log-likelihood differences of models `fit_epি_student1` vs. `fit_epি_gaussian3` evaluated on test data.

```
(elpd_epi_diff_test <- sum(llm_epi_diff_test))
```

```
[1] 16.6085
```

```
(se_elpd_epi_diff_test <- se_sum(llm_epi_diff_test))
```

```
[1] 7.012555
```

The (out-of-sample) ELPD difference is roughly 17 with an SE of roughly 7. So clearly in favor of the Student-t model. The ELPD difference is noticeably smaller than the (in-sample) LPD difference from earlier, but that is largely due to the fact that we sum over only 20% of observations' log-likelihoods. In fact, if we multiply 17 by 5 to extrapolate to the full size of the data, we end up at roughly 85, which is roughly in the same ballpark than the LPD difference of 117 we found earlier.

Our approach above has one major problem: We rely on an arbitrary split into training and test data. In fact, had we chosen the split differently, for example, just set another seed to the random split we used, we would have gotten a different result. And in a case where the difference between model was smaller, perhaps a qualitatively different conclusion. We see that it isn't wise to rely on a single training-test split, but rather do many of them and then average (or sum) over the results. This is the basic idea of cross-validation. Multiple times split the data into training and test set, each time fit the models on the training data and then evaluate their predictive performance on the test data.

Cross-validation is procedurally easy but computationally brutal. We don't have to fit the model just once but many times, each time on different training data. If fitting the model

once is already slow, fitting in many times is often prohibitively slow for practical application. Accordingly, we see to use cross-validation schemes that either have few training-tests splits or can be approximated efficiently without actually performing any model refits at all. The first leads us to K-fold cross-validation with small  $K$  (e.g.,  $K = 10$ ). There, we split the data into  $K$  subsets, each time leaving out one of the  $K$  splits as test data and using the rest as training data. It ensures that all observations appear in a test data set exactly once. You can apply this approach to brms models via method `kfold`.

As long as  $K$  is small, we won't need to many model refits. Of course, if our full model based on all data already takes us a few hours, then also  $K = 10$  times additional few hours might hurt quite a bit. Accordingly, we will not use K-fold CV as our default approach but rather try to get away with even fewer, ideally no, model refits. This goal is what will lead us to leave-one-out cross-validation approximated via importance sampling approaches, to be discussed in detail next.

### 2.7.1 Approximate leave-one-out cross-validation

In leave-one-out cross-validation (LOO-CV), we perform  $N$  training-test splits, where each time we are leaving out a single observations, fitting the model on the remaining  $N - 1$  observations before evaluating model fit on that single left-out observation. If we think about that for a second, it doesn't seem like a wise approach computationally. Usually  $N$  is quite large so if we are already unhappy with  $K \ll N$  model refits how the hell shall we now fit  $N$  models in reasonable time? The reason for choosing LOO-CV is certainly not the many training-tests splits. Instead, we make use of the fact that each of the LOO posteriors, trained on a subset of  $N - 1$  observations, is actually pretty close to the full posterior, which is trained on all  $N$  observations. As a result, we are able to approximate the LOO posteriors with the full posterior and a usually minor adjustment.

Said adjustment can be achieved with *importance sampling*. In a nutshell, importance sampling allows us to approximation expectations (means, variances, etc.) of a target distribution by sampling from another distribution (called the *proposal distributions*) that is similar enough to the target distribution. In the LOO-CV case, the targets are the LOO posteriors and the proposal is the full posterior. To the full posterior we have access and can sample from since we have fitted the full model. The LOO posteriors we do not have access to unless we are spending considerable amount of computing power and time, which we seek to avoid.

Let me formulate the mathematical idea of importance sampling a bit more generally. If we had access to the target distribution  $p_t$ , we could approximate its expectations via samples from  $p_t$  itself, as we have done many times in the past chapter, essentially every time we worked with posterior draws. Denoting our parameter-dependent quantity of interest  $f(\theta)$  for which we seek to approximate the expectation over the target, this can be expressed as follows:

$$\mathbb{E}_{p_t}(f(\theta)) \approx \frac{1}{S} \sum_{s=1}^S f(\theta^{(s)}), \quad \theta \sim p_t(\theta)$$

Now, we can actually choose to sample from our proposal  $p_p$  instead of from the target  $p_t$  and still approximate expectations over  $p_t$ . To correct for the difference between the two distributions all we have to do is replace the simple empirical mean above with a *weighted* mean:

$$\mathbb{E}_{p_t}(f(\theta)) \approx \frac{1}{S} \sum_{s=1}^S f(\theta^{(s)}) w(\theta^{(s)}), \quad \theta \sim p_p(\theta)$$

The challenge, is of course to get the weights  $w(\theta^{(s)})$  right. It turns out that the appropriate approach is to first compute values proportional to the density ratio of target and proposal distribution:

$$r(\theta^{(s)}) \propto \frac{p_t(\theta^{(s)})}{p_p(\theta^{(s)})}$$

and then normalizing the ratios such that the obtained weights sum to 1:

$$w(\theta^{(s)}) = \frac{r(\theta^{(s)})}{\sum_{s'=1}^S r(\theta^{(s')})}$$

The fact that the importance ratios only need to be *proportional* to the density ratio is highly important if we apply importance sampling to LOO-CV. There, the target is a LOO posterior  $p(\theta|y_{-n})$  and the proposal is the full posterior  $p(\theta|y)$ . If we plug that into the density ratio, expand the terms following from Bayes Theorem and subsequently simplify, we get

$$\frac{p_t(\theta^{(s)})}{p_p(\theta^{(s)})} = \frac{1}{p(y_n | \theta^{(s)})} \times \frac{p(y_{-n})}{p(y)}.$$

The former term results from the fact that both the prior and the likelihood contributions of all observations except for the  $n$ th appears in both models and thus cancel out. The latter term is the ratio of the marginal likelihood of both models. As we know from **?@sec-bayesian-inference**, marginal likelihoods are notoriously hard to estimate so we try to sneak around them whenever possible. Fortunately, the marginal likelihoods are by definition independent of the parameters, and since we only values proportional to the density ratios, we can simply set the importance ratios to

$$r(\theta^{(s)}) = \frac{1}{p(y_n \mid \theta^{(s)})}$$

Here, the closeness of the proposal and target distribution is reflected in remarkable simple importance ratios.

### 2.7.2 Pareto smoothed importance sampling

We may stop here and just use the thus obtainable importance weights in the ELPD approximation. However, without further adjustments, (a) the thus obtained approximations may be very noisy and (b) we don't know if we can trust the importance sampling approximations in the first place. As a remedy for both, Vehtari et al. (2024) developed Pareto-smoothed importance sampling (PSIS), where the values of the largest percent of importance ratios (usually the largest 15%) are replaced by the respective quantiles of the generalized Pareto distribution, a common extreme value distribution.

By means of this approach, the largest importance ratios are smoothed, thus reducing the noise in the importance sampling approximation. What is more, the generalized Pareto distribution has a shape parameter  $k$ , which indicates how fat the tail of the distribution is. It is very hard to visually grasp differences caused by varying  $k$  (see Figure 2.11). However, to judge the validity of importance sampling, the differences in  $k$  are highly informative as I will explain below. During PSIS, we get an estimate  $\hat{k}$  of this Pareto  $k$  parameter. Jumping to conclusions and simplifying a bit: As long as  $\hat{k} < 0.7$  we can trust our PSIS approximation of the corresponding LOO posterior, as long as our number of draws  $S$  is in the usual range of a couple of thousands.

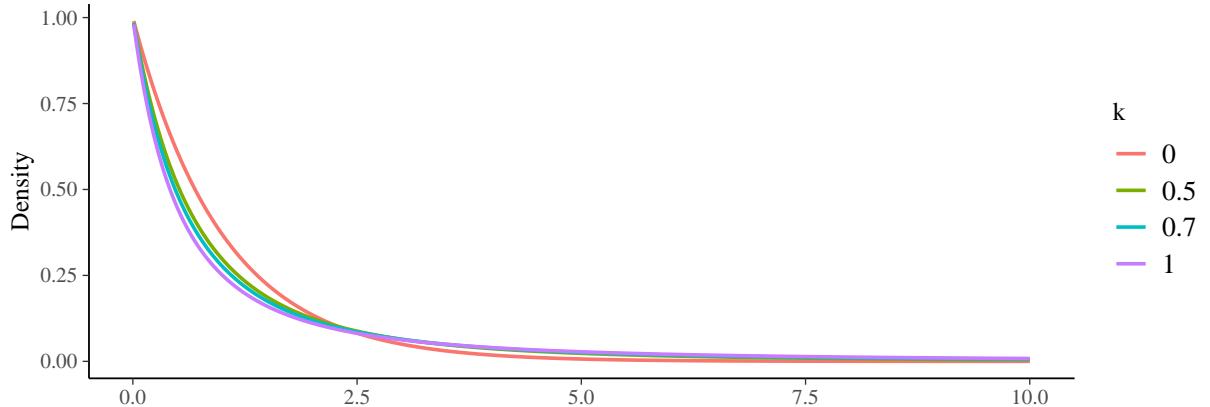


Figure 2.11: Exemplary densities of the generalized Pareto distribution for varying shape parameter  $k$ .

The explanation for the choice of 0.7 as threshold is actually pretty involved. Below, I attempt a high-level summary (see Vehtari et al. 2024 if you want to understand all the details). The

higher the  $k$  parameter, the fatter the tail of the Pareto distribution, which has important consequences for the distribution's moments (mean, variance, etc.): The fatter the tails of a distribution the more moments will be *infinite*. In the language of probability theory, we say a moment *exists* if its absolute value is finite. Existence of moments is quite a strange mathematical concept, but for our purposes it is sufficient to just understanding its practical consequences. Suppose we want to get an estimate of the central tendency of a variable  $\theta$  using the empirical mean  $\bar{\theta} = \sum_{s=1}^S \theta_s / S$  over a set of  $S$  random draws from the distribution  $p(\theta)$ . This  $\bar{\theta}$  estimate can only ever be a good approximation of the distribution's central tendency if the expected value of  $p(\theta)$  (i.e., its true mean) is finite. In other words, if the true mean (or any other moment for that matter) is infinite, their sampling-based approximations will be hopelessly bad in helping us learn something about the underlying distribution  $p(\theta)$ , no matter how many random draws we use.

Back to PSIS and the Pareto distribution, as long as the Pareto  $k$  parameter is smaller than 0.5, both mean and variance exist, which means that any sampling-based approximation of the mean is sensible and will converge quickly to the true value. If  $k$  is greater than 0.5 but smaller than 1, the mean still exist but the variance does not. This means that for large enough  $S$ , we will eventually get a good mean estimate, but the  $S$  required for this purpose may be very very large. Fortunately, as long as  $k < 0.7$ , the sampling-based mean approximation will still converge sufficiently quickly. That is, for the usual number of draws we are working with, having a  $k < 0.7$  will be good enough. For  $k > 0.7$ , convergences becomes too slow, such that we start needing ridiculous number of draws too achieve acceptable mean approximations. Accordingly, it makes sense to use  $\hat{k} < 0.7$  as a diagnostic of the reliability of PSIS and importance sampling more generally. The threshold can be refined even further based on  $S$ , but for our purposes here, using the  $\hat{k} < 0.7$  rule will be good enough.

The complete PSIS LOO-CV procedure is implemented in the `loo` package, which also provides more details on the  $\hat{k}$  threshold choice on the help page `?"pareto-k-diagnostic"`. To perform PSIS LOO-CV on a `brms` model, all we have to do is run the `loo` method, for example:

```
(loo_epi_gaussian3 <- loo(fit_epi_gaussian3))
```

```
Warning: Found 2 observations with a pareto_k > 0.7 in model
'fit_epi_gaussian3'. We recommend to set 'moment_match = TRUE' in order to
perform moment matching for problematic observations.
```

Computed from 4000 by 236 log-likelihood matrix.

	Estimate	SE
elpd_loo	-829.9	41.8
p_loo	19.4	10.8

```

looic      1659.8 83.6
-----
MCSE of elpd_loo is NA.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.6, 1.1]). 

Pareto k diagnostic values:
    Count Pct.    Min. ESS
(-Inf, 0.7] (good)    234   99.2%   1545
(0.7, 1]     (bad)      1   0.4%   <NA>
(1, Inf)    (very bad) 1   0.4%   <NA>
See help('pareto-k-diagnostic') for details.

```

We get a warning about two influential observations and a recommendation (“perform moment matching”) what to do next. We will ignore this for now but come back to the recommendation in Section 2.7.3.

Let’s go through the output in detail. On the very top, we see a short overview of the input data, namely a matrix of log-likelihood values computed for 4000 draws and 236 targets, that is, for our  $N = 236$  LOO posteriors. In the table right below we see summaries for three measures. `elpd_loo` is exactly the quantity we defined in Equation Equation 2.1. The higher the ELPD, the better the model fit. `p_loo` is a measure of the *effective number of parameters*, a quite beautiful concept that I will elaborate on a bit later. Finally, `looic` provides an *information criterion* based on LOO-CV. It is just  $-2 * \text{elpd\_loo}$  so doesn’t really contain any additional information. Yet, it is shown because many people are used to interpreting information criteria such as the Akaike Information Criterion (AIC), so these kind of metrics are familiar for them. Please keep in mind that different from `elpd` smaller `looic` values now indicate better fit. As for the summaries, in the first column, we see that point estimate of the three measures, followed by an approximation of their frequentist standard error, just as we had earlier computed manually for the RMSE. Right under the table, we see information on the Markov Standard Error (MCSE), which tells us how much variation in `elpd_loo` is likely to be caused by the fact the we only used a finite number of posterior draws. For reasons that will become clear in just a bit, we got an `NA` here, so we don’t know about the MCSE in this case. At the bottom of the output, in another table, we see information on the Pareto k diagnostics introduced above. We see that 234 LOO posteriors could well be approximated via our PSIS procedure as indicated by their  $\hat{k}$  values smaller than 0.7. However, two of the LOO posteriors show extremely bad  $\hat{k}$  with values above 1. Apparently, for these two observations, the LOO posterior is too far away from the full posterior for PSIS to bridge the gap. In other words, those two are influential observations whose inclusion vs. exclusion substantially changes the resulting posterior. If you have read the chapter in full until here, you will easily guess which two observations those are. As an immediate implication of these high  $\hat{k}$  values, we cannot trust their ELPD contributions as estimated via PSIS (we also didn’t get an MCSE estimate for this reason). As a result, we should be careful in trusting the whole PSIS LOO-CV results

of the Gaussian model for now, until we have further evaluated the situation. Let's see how LOO-CV does for the Student-t model:

```
(loo_epic_student1 <- loo(fit_epic_student1))
```

Computed from 4000 by 236 log-likelihood matrix.

	Estimate	SE
elpd_loo	-704.6	23.5
p_loo	7.4	0.6
looic	1409.3	46.9
-----		
MCSE of elpd_loo	is 0.1.	
MCSE and ESS	estimates assume MCMC draws (r_eff in [0.5, 1.3]).	
All Pareto k estimates	are good (k < 0.7).	
See help('pareto-k-diagnostic')	for details.	

In terms of  $\hat{k}$  values everything seems good now. Apparently, and unsurprisingly given the nature of the Student-t tails, the two observations in question are no longer strongly influencing the posterior.

Now is a good time to talk about `p_loo`, which I told you was a measure of the “effective number of parameters”. It has its name from the fact that, for well-behaved models with wide priors, `p_loo` roughly equals the total number of parameters. In a way, the number of parameters signals the flexibility of a model. As we increase the informativeness of the priors, the flexibility of the model decreases, so *effectively* we have the flexibility equal to a model with fewer parameters. Accordingly, the effective number of parameter will become smaller than the total number of parameters as we increase the amount of prior information. We will see this in action especially when we start talking about multilevel models in Chapter 4.

What is more, when eyeballing the difference between the LOO-CV results for the Gaussian and Student-t model, the latter seems to be clearly better. We can also see this via the appropriate method to compare LOO-CV results:

```
loo_compare(loo_epic_gaussian3, loo_epic_student1)
```

	elpd_diff	se_diff
fit_epic_student1	0.0	0.0
fit_epic_gaussian3	-125.2	35.9

In `loo_compare` models are ordered from best to worst and all compared to the best model within the set of compared models. It follows that the first row always has zero values since the best model is compared against itself. The first column shows the difference in ELPD values, while the second one shows the SE of this difference. It is kind of magic that this whole procedure allows us not only to obtain error estimates of the individual ELPD values but also of their differences. This SE is not without problems though as shown in Sivula et al. (2022). In particular, if the ELPD differences are close to zero, the SE difference estimates may be biased. Here, our ELPD difference is quite substantial with a value of  $-126$ , so the corresponding `se_diff` of about 36 is probably fine. Through these results, we not only see that the Student-t model is clearly better in absolute but also in relative terms, since the ELPD difference is about a factor  $126/36 = 3.5$  SE units away from zero.

Alternatively to computing the `loo` objects separately and then running `loo_compare`, we can also do the whole process within one step via:

```
loo(fit_epi_gaussian3, fit_epi_student1)
```

Let's add the Skew normal model to our comparisons, which we had introduced in Chapter 1:

```
fit_epi_skew_normal1 <- brm(
  count ~ Trt * Base, data = epilepsy, family = skew_normal()
)
```

The PSIS approximation to LOO-CV works a bit better for the Skew normal model than for the Gaussian model. Only one observation is diagnosed as influential and also not as badly so:

```
(loo_epi_skew_normal1 <- loo(fit_epi_skew_normal1))
```

```
Warning: Found 1 observations with a pareto_k > 0.7 in model
'fit_epi_skew_normal1'. We recommend to set 'moment_match = TRUE' in order to
perform moment matching for problematic observations.
```

Computed from 4000 by 236 log-likelihood matrix.

	Estimate	SE
elpd_loo	-794.3	33.4
p_loo	17.0	6.6
looic	1588.6	66.8
-----		

```

MCSE of elpd_loo is NA.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.5, 1.3]).
```

Pareto k diagnostic values:

		Count	Pct.	Min. ESS
(-Inf, 0.7]	(good)	235	99.6%	70
(0.7, 1]	(bad)	1	0.4%	<NA>
(1, Inf)	(very bad)	0	0.0%	<NA>

See `help('pareto-k-diagnostic')` for details.

If we choose to trust these results for the moment, our model comparison reveals:

```
loo_compare(loo_epi_gaussian3, loo_epi_student1, loo_epi_skew_normal1)
```

	elpd_diff	se_diff
fit_epi_student1	0.0	0.0
fit_epi_skew_normal1	-89.7	24.6
fit_epi_gaussian3	-125.2	35.9

Clearly, Student-t is much better than either Gaussian or Skew normal. But if we care about the difference between the latter two, the above output only gives us indirect evidence since, in `loo_compare` all models are always compared directly only to the best model. So let's remove the Student-t model from the comparison just to see the difference of the other two models:

```
loo_compare(loo_epi_gaussian3, loo_epi_skew_normal1)
```

	elpd_diff	se_diff
fit_epi_skew_normal1	0.0	0.0
fit_epi_gaussian3	-35.6	19.6

Apparently, the Skew normal model is better than Gaussian with a sizable absolute ELPD difference and about 2 SE units of relative difference.

In the above code snippets, we have avoided the recomputation of PSIS LOO-CV by storing the `loo` objects directly. This has the disadvantage that they are now independent objects of the fitted model they belong to, so we need to take care of proper naming etc. to make sure there are clearly marked as related. As a convenient alternative, brms offers to store model fit criteria directly inside the fitted model object via the `add_criterion` function. For example, to store LOO-CV results, we can run:

```
fit_epi_gaussian3 <- add_criterion(fit_epi_gaussian3, "loo")
```

```
Warning: Found 2 observations with a pareto_k > 0.7 in model  
'fit_epi_gaussian3'. We recommend to set 'moment_match = TRUE' in order to  
perform moment matching for problematic observations.
```

Then, when we call things like `loo(fit_epi_gaussian3)`, the previously stored `loo` object is automatically retrieved and recomputation is avoided. That said, if you pass additional arguments to `loo` that may change the result, brms will take it safe and recompute PSIS LOO-CV to provide you with the correct output actually belonging to your input and not to the input you used when running `add_criterion` earlier. For example, if we want to get the importance weights used in the PSIS procedure, we need to save the `psis` object. By default, this is not done such that `loo(fit_epi_gaussian3)$psis_object` returns `NULL`. If we now add `save_psis = TRUE`, brms recognizes the new argument and recomputes LOO-CV instead of retrieving the previously stored object:

```
loo(fit_epi_gaussian3, save_psis = TRUE)$psis_object
```

```
Recomputing 'loo' for model 'fit_epi_gaussian3'
```

```
Warning: Found 2 observations with a pareto_k > 0.7 in model  
'fit_epi_gaussian3'. We recommend to set 'moment_match = TRUE' in order to  
perform moment matching for problematic observations.
```

```
Computed from 4000 by 236 log-weights matrix.  
MCSE and ESS estimates assume MCMC draws (r_eff in [0.6, 1.1]).  
Pareto k diagnostic values:  
          Count Pct.    Min. ESS  
(-Inf, 0.7]   (good)    234  99.2%  1545  
(0.7, 1]     (bad)      1  0.4%  <NA>  
(1, Inf)     (very bad) 1  0.4%  <NA>  
See help('pareto-k-diagnostic') for details.
```

### 2.7.3 Correcting the PSIS approximation

The LOO-CV approximations of both the Gaussian and the skew normal model were flagged as potentially unreliable because of a few observations with high  $\hat{k}$  values. The ELPD differences with the Student-t model are unlikely to be explained by this unreliability, simple because the differences are so large and the influential observations too few. So if you encounter

such a situation in practice, you may not worry further if some of your clearly worse predicting models have a few influential observations. For other cases, especially when ELPD differences are much smaller, or when all models have a non-trivial number of influential observations, you need to be highly careful in interpreting the PSIS LOO-CV results. Below, I will show you how to improve the trustworthiness of these estimates.

The first option is to simply refit the model several times, each time leaving out one of the problematic observations. This is of course computationally expensive, but if we just have to do it a few times, it is often better than performing full on K-fold-CV. In brms, this approach is implemented via the `reloo` method, which requires both the fitted model and a corresponding `loo` object. Alternatively, if have have previously stored the `loo` object inside the model via `add_criterion`, just passing the model is sufficient. Here, we specify both objects explicitly:

```
(reloo_epi_gaussian3 <- reloo(fit_epi_gaussian3, loo = loo_epi_gaussian3))
```

Computed from 4000 by 236 log-likelihood matrix.

	Estimate	SE
elpd_loo	-831.7	43.0
p_loo	21.2	12.1
looic	1663.5	86.0
-----		
MCSE of elpd_loo	is 0.4.	
MCSE and ESS	estimates assume MCMC draws (r_eff in [0.6, 1.1]).	
All Pareto k	estimates are good (k < 0.7).	
See help('pareto-k-diagnostic')	for details.	

If we, for some reason, already know at the time of running `loo` in the first place that we want to perform `reloo` on influential observation, we may alternatively also specify this option directly in the `loo` method:

```
reloo_epi_gaussian3 <- loo(fit_epi_gaussian3, reloo = TRUE)
```

Fitting the two LOO models took some seconds and reliably got rid of the high  $\hat{k}$  warnings as we sampled from the LOO posteriors in question directly via MCMC. The ELPD value as well as all other displayed LOO-CV summaries have changed little from before to after `reloo`:

```
loo_compare(reloo_epi_gaussian3, loo_epi_gaussian3)
```

```

      elpd_diff se_diff
fit_epи_gaussian3  0.0      0.0
fit_epи_gaussian3 -1.9      1.3

```

In particular, the changes in the ELPD estimates are much smaller than and ELPD difference between models we have seen earlier. Of course, this may not always be the case, so take the Pareto  $k$  warnings of PSIS LOO-CV seriously. For completeness, performing `reloo` also for the skew normal models leads to little change there either.

```
reloo_epи_skew_normal1 <- reloo(fit_epи_skew_normal1, loo = loo_epи_skew_normal1)
```

Unsurprisingly, after seeing so little change through `reloo`, comparing the rectified LOO-CV results of the three models leads to no qualitative change in conclusion that the Student-t model is to be preferred:

```
loo_compare(reloo_epи_gaussian3, loo_epи_student1, reloo_epи_skew_normal1)
```

```

      elpd_diff se_diff
fit_epи_student1      0.0      0.0
fit_epи_skew_normal1 -89.1    24.3
fit_epи_gaussian3     -127.1   36.9

```

Using `reloo` is alright if the individual models don't take too long to fit but of course becomes daunting otherwise. As an alternative, we developed a method termed *moment matching*, which extends the range of importance sampling without requiring model refits (Paananen et al. 2021). Without going into the details, moment matching allows to shift and scale the proposal distribution to be closer to the target before performing importance sampling. As a result, the distance between the proposal and the target that can be bridged with importance sampling becomes substantially larger. In brms, we can activate moment matching during `loo` computation via

```
(mmloo_epи_gaussian3 <- loo_moment_match(
  fit_epи_gaussian3, loo = loo_epи_gaussian3
))
```

Computed from 4000 by 236 log-likelihood matrix.

	Estimate	SE
elpd_loo	-831.6	42.9

```

p_loo      21.1 12.0
looic     1663.1 85.8
-----
MCSE of elpd_loo is 0.1.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.6, 1.1]).

All Pareto k estimates are good (k < 0.7).
See help('pareto-k-diagnostic') for details.

```

Also moment matching succeeded in providing good estimates for the two previously problematic LOO posterior and leads to practically identical results as `reloo`:

```
loo_compare(loo_epi_gaussian3, mmlloo_epi_gaussian3, reloo_epi_gaussian3)
```

	elpd_diff	se_diff
<code>fit_epi_gaussian3</code>	0.0	0.0
<code>fit_epi_gaussian3</code>	-1.7	1.2
<code>fit_epi_gaussian3</code>	-1.9	1.3

Due to the implementation details of moment matching, the method requires posterior draws of all variables defined in Stan's parameters block. This is automatically fulfilled for many simple brms models. But for more complex models, brms does not store all such variables by default, because they are otherwise irrelevant for most-processing and just blow up the model sizes. As a result, moment matching may fail, in which case it will suggest you to set `save_pars = save_pars(all = TRUE)` in `brm`. If you initially forgot to set this option and want to apply moment matching, you will have to refit the full model with said option being activated. A bit annoying, I know. Perhaps in the future I will come up with a more convenient and efficient way to handle this case.

Let me finish this section by comparing the `reloo` with the `moment_match` option. If moment matching succeeds in reducing the  $\hat{k}$  below 0.7, I consider it preferable over `reloo`. Not only because it is usually faster, but because – perhaps counter-intuitively – it may sometimes even be more accurate than the brute force `reloo` approach. If you are interested in the details, check out our moment matching paper (Paananen et al. 2021).

## 2.7.4 Leave-one-out $R^2$

We have seen how log scores can be used to form predictive metrics both to evaluate in-sample and out-of-sample fit. The same options exist for  $R^2$  and RMSE metrics too. In order to obtain out-of-sample  $R^2$  and RMSE metrics, we simply have to perform predictions using the posterior obtained from the training data and compare it against test data observations,

unseen by the model during training. Take  $R^2_{\text{basic}}$  as an example. There, we had original drawn predictions from the full model's posterior predictive distribution thus evaluating in-sample fit:

$$(R^2_{\text{basic}})^{(s)} = 1 - \frac{\sum_{n=1}^N (y_n - \hat{y}_n^{(s)})^2}{\sum_{n=1}^N (y_n - \bar{y})^2}, \quad \hat{y}_n^{(s)} \sim p(y_n | y)$$

We can easily create, say, a LOO-CV version of  $R^2_{\text{basic}}$  by drawing  $\hat{y}_n^{(s)}$  from the LOO posterior predictive distribution of the  $i$ th data point:  $\hat{y}_n^{(s)} \sim p(y_n | y_{-n})$  instead. And as with the log scores, we can also use PSIS to approximate the LOO posteriors. A variation of this procedure is implemented in the `loo_R2` method. Is is not exactly the same as this sketched LOO-CV approach for  $R^2_{\text{basic}}$ , but conceptually quite similar. Let's check what `loo_R2` thinks of our Gaussian and Student-t models:

```
loo_R2(fit_epi_gaussian3)
```

	Estimate	Est.Error	Q2.5	Q97.5
R2	0.5664453	0.09304878	0.3614935	0.720993

```
loo_R2(fit_epi_student1)
```

	Estimate	Est.Error	Q2.5	Q97.5
R2	0.3965231	0.04082951	0.3241634	0.4814898

Interestingly, the Student-t model seems to be doing worse here, similar to what we had seen for some of the RMSE metrics. The same pattern can be observed when running `bayes_R2` (try it out as an exercise), so the reason for this behavior does not lie in the fact that we look at out-of-sample predictions. Since both `loo_R2` and `bayes_R2` use `posterior_epred` to create predictions, the Student-t model does not suffer from its abysmal predictions that we may sometimes obtain when running `posterior_predict` with this likelihood. Accordingly, it is not immediately intuitive why the results look like this. I must admit I was a bit puzzled myself when I first saw such results. But it does make sense if you think about it in the following way: Every  $R^2$  contains an estimate of the aleatoric uncertainty in its numerator in some form of error variance. The higher this estimate of aleatoric uncertainty, the lower (worse) the estimate of  $R^2$ . If a model with an insufficiently flexible likelihood, such as our Gaussian model, is faced with non-conforming observations ("outliers" from the perspective of the model), it will shift its mean in their direction rather than (primarily) increasing its estimate of aleatoric uncertainty. As a result, it will yield greater (better)  $R^2$  values than an alternative model with a more flexible likelihood, such as our Student-t model. The latter has an easier time excepting non-conforming observations as aleatoric, therefore increasing its

estimate of aleatoric uncertainty. It is unfortunate that a model being willing to accept greater aleatoric uncertainty is punished for it by  $R^2$  metrics. For this reason, I would advice against comparing models with different understandings of aleatoric uncertainty using  $R^2$  metrics unless one has an explicit rational for doing so. In particular, this concerns models having different likelihood families, but it may also concern models having the same likelihood family, but different modeling of the parameters managing the aleatoric uncertainty. We will discuss the latter kind of model in [?@sec-ldms](#).

## 2.8 Prior predictive performance

So far, we have been focusing on measures of posterior predictive performance. That is, we have first fitted the model and then evaluated it's prediction based on the posterior distribution of parameters after seeing the (training) data. We will now take a fundamentally different approach by evaluating *prior predictive distribution*, that is, by looking at what predictions a model implies before seeing any data. As if our training data were empty and all data were belonging to the test set.

For posterior predictive performance, the prior mattered only insofar it affected the posterior. And if the model is relatively simple compared to the amount of data, the prior will actually not matter so much. In contrast, for prior predictive performance, the prior will always matter strongly, no matter how much data we have. So we have to be much more thoughtful when specifying our priors. In particular we should avoid improper or very vague priors for reasons I will explain soon. With this warning in mind, let's begin.

### 2.8.1 Prior predictive checks

The starting point to investigating prior predictive performance is to perform graphical prior predictive checks. In brms, they can be obtained in an almost identical fashion than posterior predictive checks. All we have to do is to tell brms that it should not consider the data when sampling from the posterior. What remains is, of course, only the prior. For this to work we need to make sure that all parameters have proper priors though, and ideally not any too ridiculous priors either. Let's go ahead and explicitly specify some proper, weakly informative priors for our basic Gaussian interaction model on the epilepsy data:

```
prior_epi_gaussian6 <-  
  prior(normal(6, 3), class = "Intercept") +  
  prior(normal(0, 5), class = "b", coef = "Trt1") +  
  prior(normal(0, 1), class = "b", coef = "Base") +  
  prior(normal(0, 1), class = "b", coef = "Trt1:Base") +  
  prior(normal(0, 15), class = "sigma")
```

In the `normal(5, 3)` prior for `Intercept` we are encoding that a priori we are expecting a mean count of 6 but that it may very well be also 3 or 9; even values of 0 or 12 would be quite possible. In fact around 95% of the prior probability lies between 0 and 12, following the mean  $\pm 2$  SD rule of the 95% central uncertainty interval of normal distributions. Remember that the priors we put on `class = "Intercept"` are not in fact directly targeting the actual intercept when all predictors are taking on the value of zero, but rather an alternative intercept when all predictors are at their mean. This primarily has computational reasons but, as a side effect, also makes prior specification of the intercept easier in brms models.

In the `normal(0, 5)` of the treatment main effect, we are encoding that we assuming a treatment effect in the interval  $[-10, 10]$  with 95% probability, for those patients with a baseline seizure count of 0. Remember that the main effect of a predictor can only be interpreted in the context of corresponding interactions involving said predictor, which doesn't make prior specification any easier. In our case, only very few participants have a baseline seizure count of 0. This renders the main treatment effect an almost hypothetical scenario and highlights the difficulties in prior specification even in such very simple regression models. The specification of the other priors on regression coefficients follows a similar logic, which I invite you think through as an exercise.

The `normal(0, 15)` prior on the residual standard deviation requires a bit more elaboration perhaps. As we remember from earlier, Stan automatically truncates prior at the parameters boundaries. Accordingly, we in fact have specified a *truncated* normal prior with lower truncation bound of zero, since standard deviations cannot become negative. If you want to imagine this prior, just take a normal prior with mean 0 and SD of 15 and then cut off the lower half of it below zero. For this reason a truncated normal prior with a mean of zero before truncation is also referred to as a *half*-normal prior. The SD of 15 encodes that we assume a lot of random noise in our data that we cannot explain by means of our predictors. In fact, we high probability we do expect  $\sigma$  values from very close to zero up to  $2 * \text{SD} = 30$  seizure counts.

You see, in order to even begin to sensibly evaluate prior predictive performance, we need to think really hard about priors. Now that we are done, we are “fitting” the model with the option `sample_prior = "only"`, which ensures that Stan ignores the likelihood contribution to the posterior, such that the posterior directly resembles the prior. To be fair, we could also sample from the prior directly by sampling from the bunch of normal specified above. However, we then had to manually create a brms model around those prior draws, which is not the most convenient thing to do. What is more, some priors are actually hard to sample from directly, so brms comes with the `sample_prior = "only"` to cover all these cases and make the prior predictive checking workflow easier for you to code. We also still need to pass the dataset to the `data` argument such that brms knows about the meaning of all variables in the formula. Otherwise, brms would, for example, not know which variables would be factors and what levels the factors would have.

```

fit_prior_epи_gaussian6 <- brm(
  count ~ 1 + Trt * Base,
  data = epilepsy,
  prior = prior_epи_gaussian6,
  sample_prior = "only"
)

```

When looking at the model summary, we see that the estimated “posteriors” indeed seem to resemble the priors:

```
summary(fit_prior_epи_gaussian6)
```

Warning: There were 1 divergent transitions after warmup. Increasing  
 adapt\_delta above 0.8 may help. See  
<http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup>

Family: gaussian  
 Links: mu = identity  
 Formula: count ~ 1 + Trt \* Base  
 Data: epilepsy (Number of observations: 236)  
 Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
 total post-warmup draws = 4000

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	7.39	36.49	-62.75	77.68	1.00	4391	3049
Trt1	0.04	5.03	-9.61	9.59	1.00	4604	2670
Base	-0.03	1.03	-2.02	1.96	1.00	4369	2665
Trt1:Base	-0.02	1.00	-2.03	1.95	1.00	4259	2822

#### Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	12.06	9.17	0.65	34.50	1.00	2675	1880

Draws were sampled using sampling(NUTS). For each parameter, Bulk\_ESS and Tail\_ESS are effective sample size measures, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).

For all practical purposes, our “prior-only” brms model can be post-processed as any other brms model. In particular, we can call pp\_check to obtain prior predictive checks (see Figure 2.12):

```
pp_check(fit_prior_epи_gaussian6, ndraws = 100) + xlim(-150, 150)
```

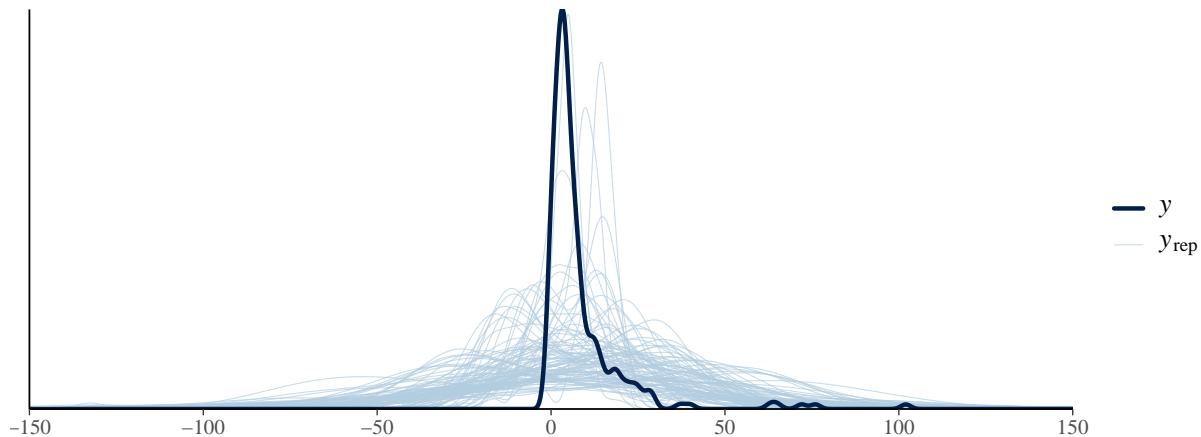


Figure 2.12: Prior predictive check obtained from model `fit_prior_epи_gaussian6`.

We restrict the x-axis a bit to avoid occasional outliers that may cause the plot to be unreadable. When doing so, please make sure that you are not accidentally loosing most of your draws to the axis truncation, which would result in a misleading impression of our prior predictive distribution. The marginal prior predictive check above overall looks reasonable. The prior predictive distribution is wider than the observed data distribution, but roughly on the same order of magnitude. That's what we would hope to see, at least marginally. It doesn't ensure with certainty that the priors are all reasonable but at least non is ridiculously wide. Of course our Gaussian model is still a bit stupid in that it predictive many negative counts. Since its likelihood doesn't have a lower bound at zero, that is not a lot we can do that would not involve making our prior very (perhaps overly) narrow. So we consider this model's prior predictive distribution as "good enough" for our purposes here.

Earlier, we have spend considerable amount of time to compare the Gaussian with the Student-t interaction model. We will continue this theme here. So let's do some prior predictive checks for the Student-t model too. Compared to the Gaussian model, we only have one more parameter, the degrees of freedom  $\nu$  (`nu`). By default,  $\nu$  already has a reasonable, weakly informative `gamma(2, 0.1)` prior, which we specify below manually just to make it explicit:

```
prior_epи_student6 <- prior_epи_gaussian6 +
  prior(gamma(2, 0.1), class = "nu")
```

All other parameters are shared between the Gaussian and the Student-t model. The  $\sigma$  parameter slightly changes its meaning as it is no longer exactly the residual standard deviation in Student-t models. That said, since it still conveys roughly the same thing, we keep its prior untouched for our purposes here, hoping that the prior is still reasonable enough.

```

fit_prior_epic_student6 <- brm(
  count ~ 1 + Trt * Base,
  data = epilepsy, family = student(),
  prior = prior_epic_student6,
  sample_prior = "only"
)

```

After “fitting” the prior-only model, we can again perform prior predictive checks (see Figure 2.13):

```
pp_check(fit_prior_epic_student6, ndraws = 100) + xlim(-150, 150)
```

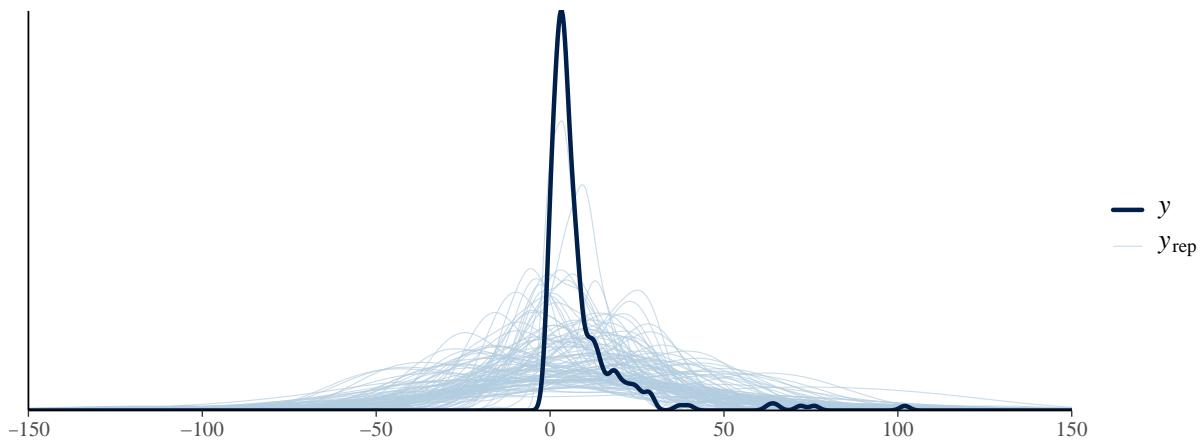


Figure 2.13: Prior predictive check obtained from model `fit_prior_epic_student6`.

That doesn’t look too bad and actually relatively similar to the check of the Gaussian model. If you run this code yourself and pay attention to the warnings, you will see a few more points being excluded from the plot because of the x-axis limits, a result of the fatter tails of the Student-t distribution, which we have discussed extensively above. But nothing too concerning, so it seems we have found some reasonable enough priors for our models to continue our prior predictive endeavors.

### 2.8.2 Marginal likelihood-based metrics

How are we going to mathematically formalize the prior predictive performance of a model? Earlier in Section 2.7 when discussing ELPD, we have evaluated the pointwise likelihood in expectation (averaged) over the posterior. To evaluate prior predictive performance, we could do the same thing but average over the prior. And in fact, that is almost what is done in

practice, except that we do not evaluate the pointwise likelihoods  $p(y_n \mid \theta)$  individually, but rather the whole (joint) likelihood  $p(y \mid \theta)$  directly:

$$p(y) = \int p(y \mid \theta) p(\theta) d\theta$$

You might recognize this quantity as the marginal likelihood that we find in the denominator of Bayes theorem. And indeed it is exactly this quantity. Another name for the marginal likelihood is “evidence”, which is more common in some fields using the marginal likelihood for model comparison. Let’s add the model indicator  $M$  to the equation, to clarify that we are referring the marginal likelihood of model  $M$ , evaluated from the prior and likelihood of said model:

$$p(y \mid M) = \int p(y \mid \theta, M) p(\theta \mid M) d\theta$$

Written this way, we can interpret the marginal likelihood as the likelihood of the data given the model. So instead of doing inference about model parameters based on the (original) likelihood, we can do inference about model themselves based on the marginal likelihood. And just as with the absolute likelihood values  $p(y \mid \theta)$ , the absolute marginal likelihood values  $p(y \mid M)$  are very hard to interpret. We only know that higher is better. Accordingly, we have to compare the marginal likelihoods of two (or more) models to make sense of how good or bad a model fits to the data. The most common such comparative metric is the Bayes factors that is defined as the ratio of two models’ marginal likelihoods:

$$\text{BF}_{12} = \frac{p(y \mid M_1)}{p(y \mid M_2)}$$

We use  $\text{BF}_{12}$  to indicate that model  $M_1$  is in the numerator and model  $M_2$  is the the denominator. If the Bayes factor is greater than 1, the data  $y$  have a higher likelihood given model  $M_1$  compared to  $M_2$ , and vice versa. Admittedly, this interpretation remains a bit awkward. Ideally, we would like to say something like: “Given the data  $y$ , model  $M_1$  is more likely than model  $M_2$ .” And indeed we can make such a statement by considering the posterior odds:

$$\frac{p(M_1 \mid y)}{p(M_2 \mid y)} = \frac{p(y \mid M_1)}{p(y \mid M_2)} \frac{p(M_1)}{p(M_2)} = \text{BF}_{12} \frac{p(M_1)}{p(M_2)}$$

In words, the posterior odds is defined as the ratio of the posterior probabilities of model  $M_1$  vs.  $M_2$  given data  $y$ . It can be computed as the Bayes factor multiplied by the prior odds, which indicates the ratio of prior probabilities we assign to  $M_1$  vs.  $M_2$ . For example, if we a priori (before seeing the data) think that  $M_1$  is more plausible than  $M_2$ , we can set the prior odds to values larger than 1. If we think that a priori both models are equally likely, we simply set the prior odds to 1, in which case the posterior odds are equal to the Bayes factor. In

practice, we often see Bayes factors being interpreted as if they were posterior odds, but keep in mind this only works in the special case of the prior odds being 1. Let me add that often we set the prior odds to 1 not actually because we really believe in models being equally likely a priori, but simply out of convenience; just as we often set wide or even completely flat priors on parameters.

Setting aside the challenges of prior specification, the mathematical theory of Bayes factors and posterior odds is pretty straight forward. However, since the Bayes factor is based on marginal likelihoods, the computational challenges are substantial. Fortunately, there is one class of algorithms that enables reliable computation of (log) marginal likelihood on the basis of posterior draws. This class of algorithms is called bridge sampling (Meng and Schilling 2002). When using bridge sampling, we do not compute the marginal likelihoods directly, e.g., based on prior draws. Rather, we have to obtain draws from the posterior first (e.g., via MCMC), which combined with corresponding log unnormalized posterior density values, enables us to obtain a reliable marginal likelihood estimate. Stan provides us with both, posterior draws and log unnormalized posterior densities, so Stan and bridge sampling play well together. The algorithmic details of bridge sampling are not relevant here, but if you want to learn more, I recommend that you check out the tutorial by Gronau et al. (2017). In R, bridge sampling is implemented in the *bridgesampling* package (Gronau, Singmann, and Wagenmakers 2020).

When wanting to run bridgesampling in brms, we have to keep in mind two special things: (1) Marginal likelihood estimation via bridgesampling usually requires several times more posterior draws than the estimation of posterior moments or quantiles (i.e., what we usually do with posterior draws). Exactly how many posterior draws is sufficient is hard to tell but in my experience something between 3 to 5 times should be good enough, but please don't take my word for it. The bridgesampling package also comes with good diagnostics to inform us if its estimates' reliability is likely insufficient. (2) Bridgesampling requires posterior draws of all variables declared in the parameters block of Stan. This is more of a technical requirement. It only concerns brms users because brms does not necessarily store all these variables, since some of them are irrelevant for the usual post-processing. Except for bridgesampling that is. To ensure all these variables are stored, you have to set `save_pars = save_pars(all = TRUE)` in `brm`. If you fail to do so, and you are fitting anything but a very simple model, chances are some parameter variables were not stored and bridge sampling will fail accordingly. In response, you then need to fit your model again with the above `save_pars` option. A bit annoying, I know, but the alternative would be to always store all kinds of variables in a brms model, thus blowing up model size drastically – only to seamlessly support the (within the brms community) relative niche feature of marginal likelihood estimation.

Anyway, here is the code to fit our Gaussian interaction model with the above discussed priors and all setting enabled to support a reliable marginal likelihood estimation:

```
fit_epi_gaussian6 <- brm(
  count ~ 1 + Trt * Base, data = epilepsy,
  prior = prior_ephi_gaussian6,
```

```
  save_pars = save_pars(all = TRUE),
  iter = 5000, warmup = 1000
)
```

Obtaining an estimate of the log marginal likelihood via bridgesampling is then just a matter of a few seconds.

```
logml_ephi_gaussian6 <- bridge_sampler(fit_ephi_gaussian6)
```

```
summary(logml_ephi_gaussian6)
```

```
Bridge sampling log marginal likelihood estimate
(method = "normal", repetitions = 1):
```

-831.2985

Error Measures:

```
Relative Mean-Squared Error: 7.310076e-07
Coefficient of Variation: 0.0008549898
Percentage Error: 0%
```

Note:

All error measures are approximate.

The log marginal likelihood estimate of about  $-831$  also clarifies why we are estimating the marginal likelihood on the log scale: If we were to exponentiate this value, computers will just straight up round it down to 0 thus rendering the estimate entirely useless. Of course, without a second model to compare against, our log marginal likelihood estimate, while numerically stable, does not tell us much. After all, marginal likelihoods are only estimates of *relative* prior predictive performance. Let's fit out Student-t interaction model then as well with all bridgesampling settings enabled:

```
fit_ephi_student6 <- brm(
  count ~ 1 + Trt * Base, data = epilepsy,
  family = student(),
  prior = prior_ephi_student6,
  save_pars = save_pars(all = TRUE),
  iter = 5000, warmup = 1000
)
```

```
logml_epি_student6 <- bridge_sampler(fit_epি_student6)
```

```
summary(logml_epি_student6)
```

Bridge sampling log marginal likelihood estimate  
(method = "normal", repetitions = 1):

-720.6898

Error Measures:

Relative Mean-Squared Error: 1.449238e-05

Coefficient of Variation: 0.003806886

Percentage Error: 0%

Note:

All error measures are approximate.

Computing the (log) Bayes factor of the two models is then straightforward:

```
bayes_factor(logml_epি_gaussian6, logml_epি_student6, log = TRUE)
```

Estimated log Bayes factor in favor of x1 over x2: -110.60871

All what this method does is subtract the two log marginal likelihood estimates since ratios become differences on the log scale. The log Bayes factors estimate of  $-110$  is very negative, which tells us that first model is much worse than the second model. Here again, it pays off to compute things on the log-scale as we will otherwise get a Bayes factor of 0, which would, if we were to take this value literally imply an “infinitely worse” fit of the Gaussian vs. the Student-t model. To be fair, a log Bayes factor of  $-110$  is also very close to “infinitely worse” already.

We can also change the order of the log marginal likelihood estimates, and get the same result only with a reversed sign for the log Bayes factor. This is because ratios become differences on the log scale, so switching numerator and denominator of the Bayes factor is equivalent to the log Bayes factor switching its sign.

```
bayes_factor(logml_epি_student6, logml_epি_gaussian6, log = TRUE)
```

```
Estimated log Bayes factor in favor of x1 over x2: 110.60871
```

If we just want to compute Bayes factors of two brms models, without having to manually call `bridge_sampler`, we can also just use the `bayes_factor` method directly on the two model objects:

```
bayes_factor(fit_epi_gaussian6, fit_epi_student6, log = TRUE)
```

Be aware though that this will recompute the marginal likelihoods via bridge sampling every time we call the method, unless we have previously stored the estimates in the model itself via `add_criterion`:

```
fit_epi_gaussian6 <- add_criterion(fit_epi_gaussian6, "marglik")
fit_epi_student6 <- add_criterion(fit_epi_student6, "marglik")
```

At the start of this section, I emphasized that measures of prior predictive performance will always remain prior dependent regardless of how much data we throw at the model. As a result, when using such measures, we always have to think carefully about our priors. The two interaction models are not ideally suited to demonstrate the point of prior dependence. So let's fit another model, the Gaussian model without interactions from earlier in the chapter, and compare it to the Gaussian interaction model.

```
prior_epи_gaussian7 <-
  prior(normal(6, 3), class = "Intercept") +
  prior(normal(0, 5), class = "b", coef = "Trt1") +
  prior(normal(0, 1), class = "b", coef = "Base") +
  prior(normal(0, 15), class = "sigma")
```

As you see, we will use the same priors as above just that we remove the prior on the interaction term.

```
fit_epi_gaussian7 <- brm(
  count ~ 1 + Trt + Base, data = epilepsy,
  prior = prior_epи_gaussian7,
  save_pars = save_pars(all = TRUE),
  iter = 5000, warmup = 1000
)
```

```
bayes_factor(fit_epi_gaussian6, fit_epi_gaussian7)
```

We obtained a Bayes factor of around 60 in favor of the Gaussian interaction model, which by common conventions would be classified as strong evidence ( $BF > 10$ ). Not that I want you to follow these conventions. I just want to give you a sense on how people might typically interpret this result.

I promised you to showcase prior sensitivity, so let's replace the `normal(0, 1)` prior on the interaction term by a `normal(0, 10)` prior:

```
prior_epi_gaussian6b <- prior_epi_gaussian6
prior_epi_gaussian6b$prior[4] <- "normal(0, 10)"
```

In the above code, we have made use of the fact that `brmsprior` objects are essentially just data frames and that the prior on the interaction coefficient happened to be in the 4th row as a result of how we defined `prior_epi_gaussian6` initially. Now, let's fit the Gaussian interaction model again with the adjusted prior.

```
fit_epi_gaussian6b <- brm(
  count ~ 1 + Trt * Base, data = epilepsy,
  prior = prior_epi_gaussian6b,
  save_pars = save_pars(all = TRUE),
  iter = 5000, warmup = 1000
)
```

```
bayes_factor(fit_epi_gaussian6b, fit_epi_gaussian7)
```

Apparently, the Bayes factor has reduced substantially and is now around 6. According to common conventions, we would consider this (only) moderate evidence in favor of the interaction model. We are even already nearing the space ( $1/3 < BF < 3$ ), where we would consider the evidence in favor or against any of the models to be inconclusive. Notably, the posterior distribution of the interaction coefficients is virtually unchanged (check the summary output of the models to verify). We see that the Bayes factor may change quite dramatically as a result of changing the prior *even though* the posterior barely changes at all. We can take this even further and change the prior on the interaction coefficient to `normal(0, 100)`, with the result that the Bayes factor shrinks to about 0.6. This would now be in range of inconclusive evidence but with a direction that rather points to the no-interaction model fitting (slightly) better. Again the posterior remains essentially the same. You might argue that both the `normal(0, 10)` and the `normal(0, 100)` prior are ridiculously wide for this interaction coefficient and I agree. My point here is only to showcase the prior sensitivity of Bayes factors (and other marginal likelihood-based metrics), hopefully convincing you to think carefully about your priors if you plan on applying such metrics for model comparison.

After following both the formal and the informal discussions of Bayesians over several years, I came to the conclusions that the dispute of prior vs. posterior predictive performance will not

be resolved any time soon. And that's alright. Both kind of approaches have their sensible use cases and also their shortcomings. We simply have to know when to apply what. Personally, I am more on the posterior predictive side. Not for any strong theoretical or philosophical reasons, but only because prior predictive metrics are just so terribly inconvenient: Even if I have overwhelming amounts of data, where no reasonable prior will have any relevant influence on the posterior, I would still have to think carefully about my priors. And I just don't have the time and energy to do this for most analysis.

Let me finish this section by pointing to the rich, but so far practically underused field of (expert) prior elicitation, which enables the translation of expert knowledge into prior distributions (for a recent review see Mikkola et al. 2023). Even after decades of research we are still struggling to bring prior elicitation into the Bayesian analyses of the masses, because it turns out that both the process of eliciting expert knowledge and translating it into priors is actually really hard. It is one of the topics of my research to make this process feasible and user-friendly for a much larger audience (Bockting, Radev, and Bürkner 2023), but we are still quite far away from that goal.

## 2.9 Model averaging

The results of model comparison can usually be used for two different purposes. The purpose is *model selection* as we have seen above. There, we compare a bunch of models with each other and then select a single one, or a few if we cannot decide for a single one, with whom we will then move forward. The other, non-selected models will then be discarded.

But what if we want to use all models? If we want to use all models "equally" in some sense, for example to evaluate predictive performance, then we can just take posterior predictive draws from all models and pool them together in a single vector of draws. We then process this vector further as if the draws were all coming from a single model. In this case, we wouldn't need to do model comparison at all. But it also sounds a bit wrong to use all models equally, especially if we know that some models are predicting much better than others. So it is perhaps more sensible to give better predicting models higher weight, and more poorly predicting models less weight. Suppose we have two models, and our model comparison procedure indicated that model  $M_1$  was 3 times better on some metric than model  $M_2$ . Translating that into weights  $w$ , that is, non-negative values that sum to 1, this would imply the weight of  $M_1$  to be  $w_1 = 0.75$  and the weight of  $M_2$  to be  $w_2 = 0.25$ . If we wanted to combine the two models' posterior predictive distributions, we could select 75% of the draws from  $M_1$  and only 25% of draws from  $M_2$ , and then only pool the selected draws. The general procedure of combining model's predictions according to some kind of weights is called *model averaging*. Below, I will briefly describe how to obtain weights from marginal likelihood-based and information criterion-based procedures.

### 2.9.1 Weights from marginal likelihoods

In Section 2.8.2, we have discussed comparing the marginal likelihoods of two models using Bayes factors (or posterior odds), but what shall we do if we have more than two options in our set of models being considered? Of course, we would compute the Bayes factors of each of the pairwise comparisons but that would be quite cumbersome to interpret. To improve the situation, we can compute the the posterior model probabilities  $p(M_j | y)$  for all considered models  $M_j$  explicitly, under the assumption that the true model is within the set of considered models. This leads us to apply Bayes theorem but to models instead of to parameters:

$$p(M_j | y) = \frac{p(y | M_j) p(M_j)}{\sum_{j'=1}^J p(y | M_{j'}) p(M_{j'})}$$

In words, we multiple each model's marginal likelihood with its prior model probability and normalize it with the sum of these products over all considered models. In brms, we can compute posterior model probabilities using the `post_prob` method. Here, we assume equal prior model probabilities for the lack of any more sophisticated idea:

```
post_prob(fit_epি_gaussian6, fit_epি_student6, prior_prob = c(0.5, 0.5))
```

```
fit_epি_gaussian6  fit_epি_student6
9.208991e-49      1.000000e+00
```

Unsurprisingly, we obtain a posterior model probability of basically 100% for the Student-t model and one of basically 0% for the Gaussian model. That is, if one of these two models was indeed the true model, than we would be almost 100% sure it is the Student-t model.

Since brms supports several different model weighting methods (see below), it provides the `model_weights` method as a unified interface among all of them. If we want to compute posterior model probabilities using the `model_weights` method, we can use the `weights = "bma"` option, where “bma” stands for Bayesian model averaging. This will internally call `post_prob` so is equivalent to the code above.

Theoretically, posterior model probabilities are a very appealing quantity since they literally encode the probability of each considered model given the data. However, it comes with the key assumption that, as already mentioned above, the true model is among the set of considered models – something we also call the *closed-world assumption* (Bernardo and Smith 1994). In other words, we are assuming that we consider all theoretically possible models in our comparison. If you think about it, that is a very strong assumption and rarely justified unless we are working in a field with very strong theories that we can turn into equivalent statistical models. Perhaps this is justified in some parts of physics, but I struggle to see this being the case much more generally.

If we let go of this assumption, and allow the true model to be outside of the scope of considered models – or if we don’t even assume the existence of a true model in the first place – we talk about an *open-world assumption* (Bernardo and Smith 1994). Typically, measures of prior predictive performance based on marginal likelihoods are considered in the context of a closed world, while measures of posterior predictive performance such as ELPD are considered in the context of an open world. I am not sure this is really a necessary distinction or rather follows purely from the history of individual fields that have concerned themselves with Bayesian model comparison.

Clearly, for the epilepsy data, none of the two models considered above is even close to the truth, which demonstrates the key limitation of posterior model probabilities and their closed-world assumption. Accordingly, in most cases, I would not interpret posterior model probabilities as indicating the probability of a given model being true but rather only as a metric that shows relative prior predictive model performance on an intuitive probability scale.

### 2.9.2 Weights from ELPD scores

In the context of posterior predictions, we have talked a lot about the expected log predictive density (ELPD) as a metric for model comparison. It is surprisingly easy to turn them into model weights as well. The ELPD is essentially the sum of log probabilities  $\log p(\tilde{y}_n | y, M)$  where  $\tilde{y}_n$  are the observations in the test data,  $y$  are the training data (see Section 2.7 for details). Accordingly, we can exponentiate it to obtain the product of probabilities  $p(\tilde{y}_n | y, M)$ . When we assume conditional independence, this product is equal to the joint probability  $p(\tilde{y} | y, M) = p(\tilde{y}_1, \tilde{y}_2, \dots | y, M)$  over all  $\tilde{y}_n$  in the test data given the training data. The higher this joint probability, the better the predictions of the model on test data. This motivates to define ELPD weights simply as the exponentiated ELPD scores normalized against their sum over all considered models:

$$w_j = \frac{\exp(\text{ELPD}_j)}{\sum_{j'=1}^J \exp(\text{ELPD}_{j'})}$$

Sometimes, these weights are also called pseudo-BMA weights due to their structural similarities with posterior model probabilities. If we compute these weights directly, we may run into numerical issues since the ELPD scores may be so small that they underflow to 0 after exponentiation. To mitigate this issue, we first compute the maximum of the ELPD values and subtract it from all ELPD values,  $\Delta\text{ELPD}_j = \text{ELPD}_j - \max(\text{ELPD})$ , before exponentiation:

$$w_j = \frac{\exp(\Delta\text{ELPD}_j)}{\sum_{j'=1}^J \exp(\Delta\text{ELPD}_{j'})} \tag{2.2}$$

Mathematically, this doesn't change the result, but numerically it makes things much stable. I suggest you to try it out yourself as an exercise. And by the way, we also use this trick to make the computation of posterior model probabilities more numerically stable. I just didn't mention it above to focus on other aspects.

In brms, we can compute model weights based on LOO-CV's ELPD values as follows:

```
model_weights(fit_epi_gaussian6, fit_epi_student6, weights = "loo")
```

```
fit_epi_gaussian6  fit_epi_student6
1.076117e-55      1.000000e+00
```

Also here, we see very clear evidence that the Student-t model outperforms the Gaussian model. For this case here, marginal likelihood and ELPD-based model weights are in agreement, but this is not generally the case. Depending on the model, data, and priors, there might as well point in opposite directions.

The model weighting approach presented here does not work only for ELPD values but actually for all methods that can produce values on the “information criterion scale”. Remember that, if we multiply the ELPD with  $-2$ , we get an information criterion, which we simply called LOOIC for ELPDs computed from LOO-CV. If we rewrite Equation 2.2 in terms of information criteria, we get what is known in the literature as *Akaike weights*. If you want to learn more, check out Wagenmakers and Farrell (2004) for a relatively gentle introduction.

### 2.9.3 Weights from stacking of predictive distributions

Lastly, I want to briefly discuss another option for model weighting, which is called *stacking of posterior predictive distributions* (Yao et al. 2018). In a nutshell, we are trying to find the model weights  $w = (w_1, \dots, w_J)$  for our models  $(M_1, \dots, M_J)$  such that the test data have maximum probability under the weighted average of the models' posterior predictive distributions. In math, we can write this as follows:

$$\max_{(w_1, \dots, w_J)} \frac{1}{N} \sum_{n=1}^N \log \sum_{j=1}^J w_j p(\tilde{y}_n | y, M_j)$$

This general approach can be readily combined with all kinds of cross-validation schemes. For example, when combined with LOO-CV, we set  $p(y_n | y_{-n}, M_j)$  as our posterior predictive distribution  $p(\tilde{y}_n | y, M_j)$ . As a result, we can also efficiently approximate LOO-CV based stacking with importance sampling, just as we do for LOO-CV-based ELPD scores.

In order to obtain stacking weights with brms, we can use

```
(ws <- model_weights(fit_epi_gaussian6, fit_epi_student6, weights = "stacking"))
```

Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.

```
fit_epi_gaussian6  fit_epi_student6  
0.1383629        0.8616371
```

For our two example models, stacking leads to somewhat less extreme weights than the other weighting methods we have tried out before, although the Student-t model is still clearly favored. More generally, compared to ELPD (pseudo-BMA) weights, stacking weights seem to perform better in particular when there are many models considered at the same time (Yao et al. 2018). The drawback of stacking weights is mainly the computational overhead created by the optimization step. However, in my experience, this is not too big of a factor and should rarely limit practical applicability. If you want to compute some kind of model weights but have no clear idea or reasoning which specific method to use, I recommend using stacking weights. For this reason, `weights = "stacking"` is also the default in the `model_weights` method.

At the start of this section, I told you that we can use model weights among others to weight posterior predictive distributions. And that's what we are going to now. `brms` comes with the `pp_average` method which takes a bunch of models as well as weights as input and outputs a weighted posterior predictive distribution:

```
ppa <- pp_average(fit_epi_gaussian6, fit_epi_student6, weights = ws,  
                    ndraws = 4000, summary = FALSE)
```

```
str(ppa)
```

```
num [1:4000, 1:236] 9.53 8.63 2.23 5.39 13.22 ...  
- attr(*, "weights")= Named num [1:2] 0.138 0.862  
..- attr(*, "names")= chr [1:2] "fit_epi_gaussian6" "fit_epi_student6"  
- attr(*, "ndraws")= Named num [1:2] 553 3447  
..- attr(*, "names")= chr [1:2] "fit_epi_gaussian6" "fit_epi_student6"
```

Through setting `summary = FALSE`, we got an output that is structurally identical to that of `posterior_epred`, just that parts of the posterior draws (around 14%) have been drawn from the Gaussian model and the remaining draws (around 86%) have been drawn from the Student-t model. In the `weights` attribute, we can even see exactly how many draws were used from which model. In the `weights`, we passed the pre-computed stacking weights directly, but we could have also just named a method like "`stacking`", "`bma`", or "`loo`" to compute

weights on the fly using the `model_weights` method. With `pp_average`, we can not only average posterior predictive distributions, but also, for example, means of posterior predictive distributions (`posterior_epred`) or linear predictors (`posterior_linpred`). Exactly which predictive quantity is to be averaged can be controlled via the `method` argument.

## 2.10 In-distribution vs. out-of-distribution

Above, we have learned about the difference between in-sample and out-of-sample data. The former was the data we fit our model on. The latter was the data we used to evaluate predictive performance of the model in an effort to (ideally) avoid overfitting. Now, we will go one step further and split out-of-sample data into *in-distribution* and *out-of-distribution* data. This is not necessarily a brms topic but understanding it was quite illuminating to me so I want to share it with you. In a nutshell, if new data were generated by the same true distribution as the training data, we call the new data “in-distribution”. In this case, can have justified hopes that our model, if built and trained properly, will be able to predict these new data well. Conversely, if the data were generated from a *different* true distribution than the training data, we call the new data “out-of-distribution”.

Why does this matter? Let me elaborate with a bit more math. Suppose, in reality, the training (in-sample) data  $\tilde{y}$  has been generated by a true data generating distribution  $p^*(y)$ , i.e.,  $\tilde{y} \sim p^*(y)$ . This true distribution  $p^*(y)$ , will almost never be fully known to us, but it suffices to assume  $p^*(y)$  exists. Suppose now we have a new dataset  $y^*$  that we didn’t use for model fitting. Then, we call  $y^*$  “out-of-sample” data (in a strict sense) if it was generated from the same true distribution as the training data, that is, if also  $y^* \sim p^*(y)$ . Now suppose that our model  $p(y | \theta)p(\theta)$  is able to learn a good representation of  $p^*(y)$  from  $\tilde{y}$ , that is, if the posterior predictive distribution of our model,  $p(y | \tilde{y})$  is a good approximation of  $p^*(y)$ . Then, our model should be able to predict  $y^*$  well even though our model hasn’t seen these data during training. This is because the same underlying patterns apply for both  $\tilde{y}$  and  $y^*$ .

However, sometimes the world changes between gathering our training data and the time we use our model for prediction of new  $y^*$ ; or we gathered  $y^*$  at a different spatial location, different population of individuals, etc. In any of these cases,  $y^*$  may no longer be truly generated from the original  $p^*(y)$  but from some other distribution  $p_{\text{new}}^*(y)$ . This kind of new data, originating from a different data-generating process than the training data, we call “out-of-distribution” data. Accordingly, if  $p_{\text{new}}^*(y)$  differs from  $p^*(y)$  in a relevant manner, then, no matter how accurate our model was to approximate  $p^*(y)$ , it may still fail terribly in predicting  $y^* \sim p_{\text{new}}^*(y)$ .

Admittedly, in reality, it is hard to test whether new data is just out-of-sample or actually out-of-distribution, because we don’t have direct access to  $p^*(y)$  and even less so to  $p_{\text{new}}^*(y)$ . There are whole fields of research dealing with questions arising in this context. You can find the related literature, for example, under the terms “out-of-distribution detection” and “anomaly detection”.

The concepts of in- and out-of-distribution data cannot just be applied to true data generating distributions but also more directly to our own Bayesian model  $p(y | \theta)p(\theta)$ . There, we can say that  $y^*$  is “in-distribution” (from the perspective of our model), if it is plausible that  $y^*$  was generated by the posterior predictive distribution of the model, that is, if it is plausible that  $y^* \sim p(y|\tilde{y})$ . Conversely, if the assumption of  $y^* \sim p(y|\tilde{y})$  is *implausible*, than we could call  $y^*$  to be “out-of-distribution” (from the perspective of our model). If used in the context of model comparison via PSIS LOO-CV, we have seen that the Pareto  $k$  diagnostic will tell us something about how influential individual observations are. When our data is not super small, every single data point should only have very little influence. At least if the model is a plausible data generating process, that is, if the data is “in-distribution” for our model. Accordingly, if the Pareto  $k$  diagnostic finds influential observations, those are likely “out-of-distribution” for our model. Hence, our model may not be a good approximation for  $p^*(y)$  after all. In other words, we can use the distribution of the Pareto  $k$  values as an *absolute* fit metric usable to provide evidence against a given model. That said, we should not use the Pareto  $k$  values as evidence in favor of a given model, since even bad models may find all observations as “in-distribution” often enough.

Let's illustrate this idea on a concrete example by investigating the Pareto  $k$  values for the Gaussian and Student-t models fitted above, more specifically `fit_epi_gaussian3` and `fit_epi_student1` both modeling an interaction between treatment and baseline seizure count:

```
khats_epi_gaussian3 <- loo_epি_gaussian3$diagnostics$pareto_k
khats_epi_student1 <- loo_epি_student1$diagnostics$pareto_k
```

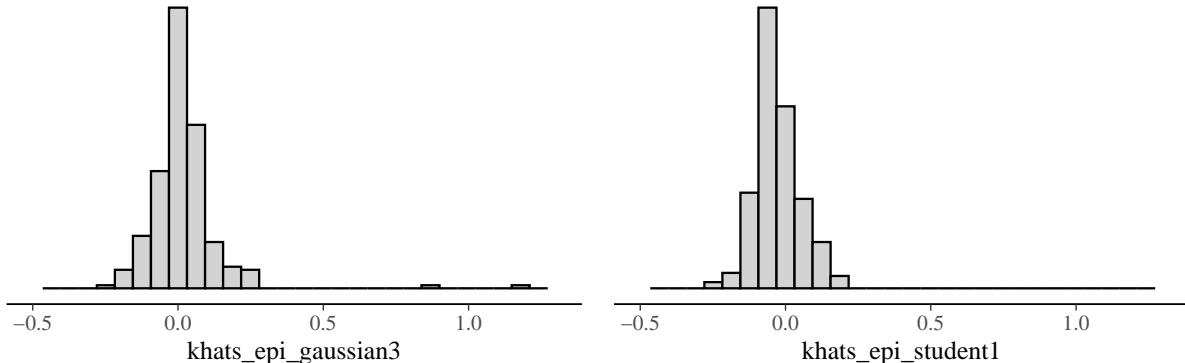


Figure 2.14: Histograms of the LOO Pareto  $k$  estimates obtained for the models `fit_epi_gaussian3` and `fit_epi_student1`.

As we can see from Figure 2.14, the range of estimate Pareto  $k$  estimates is pretty similar for both models, except for two large outliers in the normal model. This indicates that the two corresponding observations are out-of-distribution for the normal model but not for the Student-t model, suggesting that at least the normal model is not such a great fit for the given

data, at least if we trust that the two outlying observations are valid themselves. Of course, the fact that the Pareto  $k$  estimates are all good for the Student-t model doesn't necessarily make it a good model, but at least no observation lies in the out-of-distributions space.

## 2.11 Summary

In this chapter, we have learned about the basics, and some more advanced aspects, of Bayesian model comparison. We have seen that this is quite a challenging topic with many caveats and things to keep in mind. Knowing about this complexity should of course not discourage you from doing model comparison but rather help you understanding the implications and limitations of the decisions you make. In most cases, there is no single "right" way of comparing models. There are simply too many degrees of freedom in the Bayesian workflow (Gelman et al. 2020; Bürkner, Scholz, and Radev 2023a). Attempting to find that single "right" way is thus futile. Rather, I suggest you to aim for what you believe are "good enough" methods, taking into account both the analysis goals and practical constraints you face. For example, LOO-CV is not always the most appropriate cross-validation scheme, but if you have to choose between performing PSIS LOO-CV and no cross-validation at all (e.g., because the computational demands of K-fold CV are too high), then it may very well be sensible to opt for PSIS LOO-CV still – provided that its diagnostics indicate sufficiently trustworthy results. Please just make sure to communicate your model comparison methods' properties, implications, and limitations clearly.

# 3 Generalized Linear Models

## 3.1 Setup

```
library(dplyr)
library(magrittr)
library(ggplot2)
library(patchwork)
library(bayesplot)
library(brms)
library(knitr)
```

## 3.2 Introduction

Generalized linear models (GLMs) are very similar to the linear models we have already seen, except that we are choosing a different likelihood distribution other than normal, ... and then deal with the consequences that follow from it. Suppose our responses  $y$  can only take on positive ( $y > 0$ ) or non-negative ( $y \geq 0$ ) values, as in the `epilepsy` data where our response is the number of epileptic seizures patient have during a certain time interval. Clearly the number of seizures cannot be negative. If we have only one aspect of the likelihood that we can predict, we will probably want to choose a measure of central tendency as our target, usually the mean or the median. Now, in the described case, the mean or the median will of course also be positive (or at least non-negative), but our linear predictor  $\eta$  doesn't know about this restriction if we don't help it a little bit. That is, we either have to transform our response  $y$  to be unbounded and thus on the same scale as  $\eta$  or we have to transform  $\eta$  to be bounded (e.g., positive only) to be on the same scale as  $y$ . Specifically the latter case (transforming  $\eta$ ) is what leads to GLMs in a more strict sense (McCullagh 2019). But for the purpose of this chapter and the book more generally, I use the term "GLM" more generally as a model with a linear predictor where the likelihood is non-normal and/or we have to transform either  $y$  or  $\eta$  for our predictions to stay within the range of possible response values.

### 3.3 GLMs for lower-bounded responses

To illustrate the use of GLMs for lower-bounded responses, we will continue using the `epilepsy` data. Over the course of this book, we will see lots of other GLMs or their extensions modeling lower-bounded responses. Sometimes the responses are counts, as in the `epilepsy` case, sometimes there are (discrete or continuous) times, such as in time-to-event or response time data, or sometimes there are other things that have a lower bound by nature or design. Most of the time, this lower bound will be 0. But even if that bound was some  $c \neq 0$ , we could transform our responses as  $y - c$ , such that the new lower bound would be 0 again. As a result, we will focus on this special case knowing that other bounds can be easily expressed as that special case.

While we discuss count responses here, we will see that most of the discussed challenges and solutions apply more generally to other kinds of lower-bounded responses too.

#### 3.3.1 Modeling log counts

Before we start using likelihoods different than Gaussian, let's first stay in the Gaussian regime a bit longer. As we have seen, the Gaussian likelihood leads to predictions of negative (i.e., impossible) counts, which is definitely something we want to avoid. One way to solving this issue is to transform the response variable such that the assumption of a Gaussian likelihood is more reasonable. One transformation that is commonly applied for positive data is the log transformation. However, for our count response, the log transform is inappropriate since count of zero lead to  $\log(0) = -\infty$ , which clearly is not very useful for further analysis. But we are not giving up so easily. How about just adding 1 to the counts before log transforming, that is, apply  $\log(y + 1)$ ? That works since  $\log(0 + 1) = \log(1) = 0$ . This is kind of a hack but actually quite commonly used in some fields. It's not my preferred approach but it's a good starting point still:

```
fit_epi_gaussian_log1 <- brm(log(count + 1) ~ Trt * scale(Base),  
                               data = epilepsy)
```

```
summary(fit_epi_gaussian_log1)
```

```
Family: gaussian  
Links: mu = identity  
Formula: log(count + 1) ~ Trt * scale(Base)  
Data: epilepsy (Number of observations: 236)  
Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
       total post-warmup draws = 4000
```

Regression Coefficients:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	1.87	0.07	1.74	2.00	1.00	4492	2593
Trt1	-0.23	0.09	-0.42	-0.05	1.00	4633	2854
scaleBase	0.62	0.07	0.49	0.76	1.00	2429	2782
Trt1:scaleBase	0.04	0.09	-0.14	0.21	1.00	2631	2843

Further Distributional Parameters:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	0.69	0.03	0.64	0.76	1.00	4283	2799

Draws were sampled using sampling(NUTS). For each parameter, Bulk\_ESS and Tail\_ESS are effective sample size measures, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).

```
pp_check(fit_ephi_gaussian_log1)
```

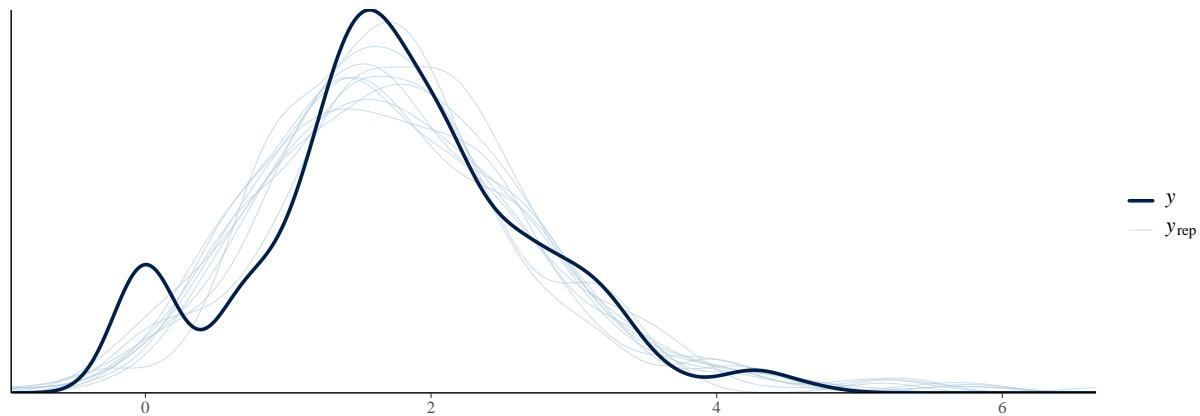


Figure 3.1: Default posterior predictive check for model fit\_ephi\_gaussian\_log1.

Notice that the response scale is now  $\log(y + 1)$ . So we don't see the original scale in Figure 3.1 anymore, because for brms  $\log(y + 1)$  is now the actual response variable. If you want to see predictions on the original scale, we have to apply the inverse transformation ourselves. If  $z = \log(y + 1)$ , then  $y = \exp(z) - 1$ . Accordingly:

```
ypred <- posterior_predict(fit_ephi_gaussian_log1, ndraws = 10)
ypred <- exp(ypred) - 1
pp_check(epilepsy$count, yrep = ypred, fun = "dens_overlay") +
  xlim(c(0, 200))
```

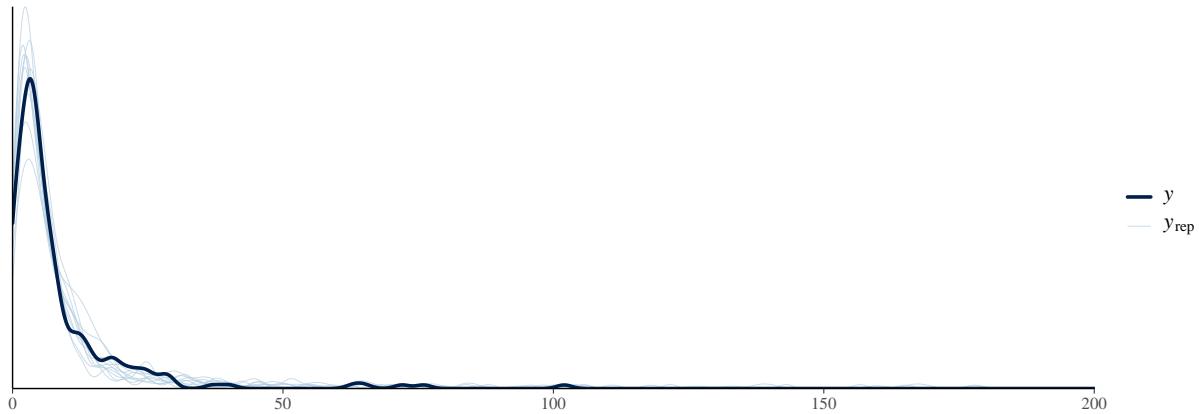


Figure 3.2: Posterior predictive check for model `fit_epi_gaussian_log1` on the back-transformed  $\exp(y) - 1$  scale.

The result shown in Figure 3.2 don't look too bad actually. Notice that, in the above code, we have used the default `pp_check` method that requires us to manually enter the observed responses as the first argument and the predicted responses as the second argument `yrep`. But what about the residuals? Let's plots them at the log-scale:

```
pp_check(fit_epi_gaussian_log1, type = "error_scatter_avg")
```

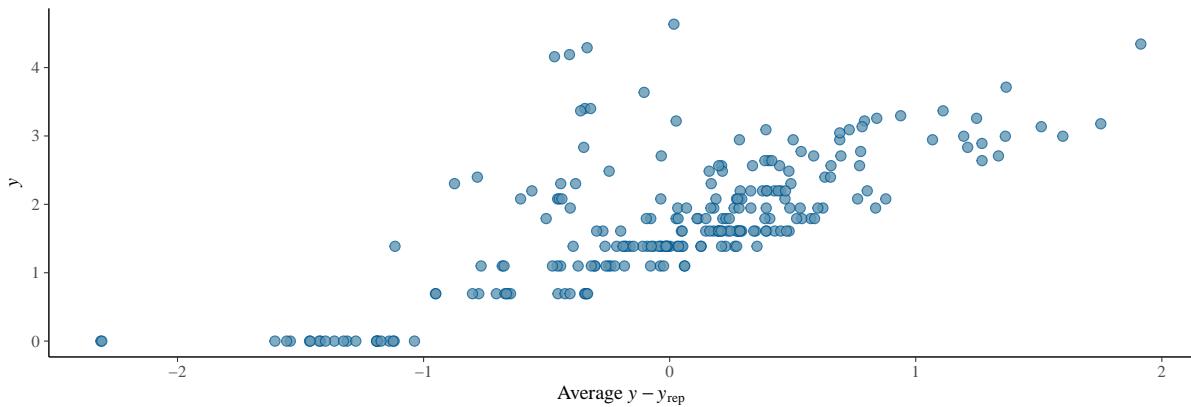


Figure 3.3: Posterior predictive check of residuals for model `fit_epi_gaussian_log1`.

The residual plot in Figure 3.3 shows a clear relationship between residuals and responses, indicating a lot of room for improving our model. We would see the same qualitative picture if we plotted residuals on the original scale, only that outliers would be more clearly visible.

Setting aside our model criticism for now, what does the model think about the effects of seizure baseline and treatment?

```
conditional_effects(fit_epi_gaussian_log1, effects = "Base:Trt")
```

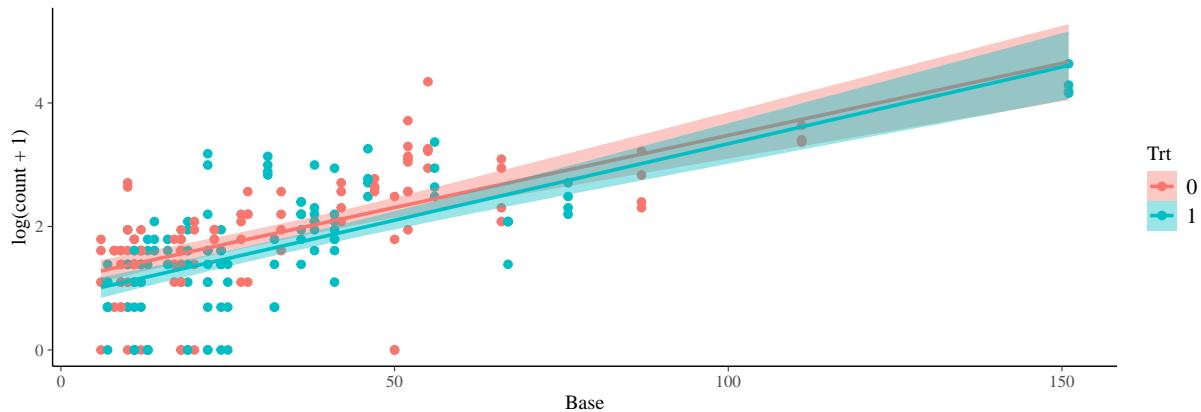


Figure 3.4: Conditional effects of the baseline-treatment interaction for model `fit_epi_gaussian_log1`.

In Figure 3.4, we are seeing conditional predictions on the log-scale where they are linear by choice of using a linear model. Let's build a basic version of `conditional_effects` ourselves to (1) produce conditional predictions on the original scale and (2) illustrate more of the internal workflow that happens inside brms.

First, we set up a dataset containing the predictor values at which we want to evaluate predictions:

```
newdata <- epilepsy %$%
  expand.grid(
    Base = seq(min(Base), max(Base), length.out = 100),
    Trt = unique(Trt)
  )
```

The `%%$%` of the `magrittr` essentially allows us to use the objects stored on the left-hand side (i.e., the `epilepsy` dataset) in the right-hand expressions, as if we had written `epilepsy$<variable>` everywhere. Its base R equivalent is the `with` function.

Second, call `posterior_epred` on the new data to obtain draws from the mean of the posterior predictive distribution:

```
pred <- posterior_epred(fit_epi_gaussian_log1, newdata = newdata)
str(pred)
```

```
num [1:4000, 1:200] 1.41 1.1 1.23 1.13 1.35 ...
```

Third, transform and subsequently summarize posterior predictive mean samples. As a general rule, *always* (if possible) transform first on the level of the posterior draws and summarize only afterwards (see Table 3.1):

```
pred_exp <- pred %>%
  { exp(.) - 1 } %>%
  posterior_summary() %>%
  bind_cols(newdata)
```

```
head(pred_exp, 3)
tail(pred_exp, 3)
```

Table 3.1: Summary of transformed posterior predictions for the first and last three observations in model `fit_epi_gaussian_log1`.

	Estimate	Est.Error	Q2.5	Q97.5	Base	Trt
1	2.60	0.33	2.01	3.30	6.00	0
2	2.73	0.33	2.13	3.43	7.46	0
3	2.86	0.33	2.25	3.57	8.93	0
198	95.64	26.59	53.67	158.47	148.07	1
199	99.30	27.94	55.38	165.38	149.54	1
200	103.10	29.36	57.10	172.41	151.00	1

The reason for me being particular about this is that the transformed posterior summary will not, in general, be equal to the summary of the transformed posterior draws. Since we aim for an accurate representation of a quantity’s posterior, we always want to see the posterior first, which in our case means getting draws from it. And afterwards, we can summarize as we please. Let’s illustrate this point and compute summaries in the “wrong” way (see Table 3.2):

```
pred_exp_wrong <- pred %>%
  posterior_summary() %>%
  { exp(.) - 1 } %>%
  bind_cols(newdata)
```

```
head(pred_exp_wrong, 3)
tail(pred_exp_wrong, 3)
```

Table 3.2: Summary of transformed posterior predictions for the first and last three observations in model `fit_epi_gaussian_log1`. An incorrect order of transformations was used.

	Estimate	Est.Error	Q2.5	Q97.5	Base	Trt
1	2.59	0.09	2.01	3.30	6.00	0
2	2.71	0.09	2.13	3.43	7.46	0
3	2.84	0.09	2.25	3.57	8.93	0
198	92.20	0.31	53.67	158.47	148.07	1
199	95.65	0.31	55.38	165.38	149.54	1
200	99.22	0.32	57.10	172.41	151.00	1

Let's carefully compare each of the summary statistics. For the posterior mean results are different. Sometimes by more sometimes by less. But they overall seem to be roughly in the same ballpark (more on this below). The posterior SD however, is completely different, for our data by a factor of about 3 to 4! Both the mean and SD fall under the category of expectation-based summaries. For them, we have Jensen's inequality to tell us that, for a convex function  $g(x)$  (e.g.,  $g(x) = \exp(x) + 1$ ), it hold that

$$g(\mathbb{E}(x)) \leq \mathbb{E}(g(x)).$$

This inequality we see clearly represented in the differing results for posterior mean and SD. The quantiles on the other hand seem to not care about the order of transformation and summary: We say they are *equivariant* under these operations. There is a mathematical rule also behind this result, which says that quantiles are equivariant under any monotonically increasing function  $g(x)$  (e.g., for  $g(x) = \exp(x) + 1$ ):

$$g(Q(x)) = Q(g(x))$$

where I am using  $Q(x)$  to denote an arbitrary quantile extract from the variable  $x$ . Finally, after this little rant, let's plot the predictions:

```
pred_exp %>%
  ggplot(aes(Base, Estimate, color = Trt, fill = Trt,
             ymin = Q2.5, ymax = Q97.5)) +
  geom_smooth(stat = "identity") +
  geom_point(data = epilepsy, aes(Base, count, color = Trt),
             inherit.aes = FALSE) +
  ylab("count")
```

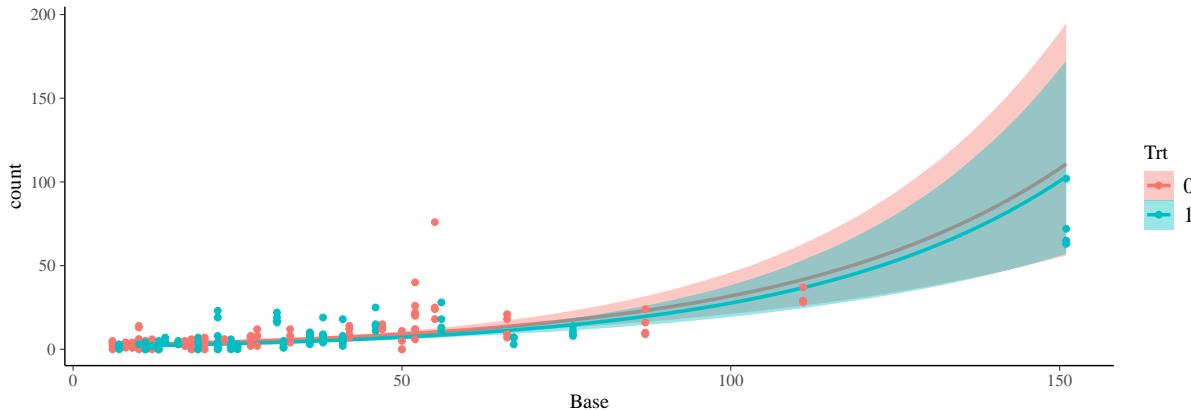


Figure 3.5: Conditional effects of the baseline-treatment interaction for model `fit_epি_gaussian_log1` on the back-transformed  $\exp(y) - 1$  scale.

Despite our efforts, these transformed predictions shown in Figure 3.5 do not represent draws for the mean of the posterior predictive distribution on the original scale. On the log-scale, it predictions represented both the mean, the median and the mode at the same time since the predictive distribution of a Gaussian likelihood is symmetric and unimodal. However, after the log transform our results do only represent the median of the posterior predictive distribution on the original scale. Again, this is because the median, and all quantile-measures for that matter are equivariant under monotonically increasing transformations. In contrast, neither the mean nor the mode are equivariant under (non-linear) transformation, as we have seen from the example of Jensen's inequality.

Accordingly, in this simple model, we have seen (non-)equivariance already coming up twice. First, when we discussed transformation and summary of posterior draws. Second, when we discussed transforming means and medians of the posterior predictive distribution. These are actually two different things: We can have draws from the mean or median of the posterior predictive distribution and can then in turn summarize both of these draws by mean or median. I understand this is confusing. Take a moment to reflect on this to make sure the difference is clear.

### 3.3.2 Log transform both response and baseline counts

As we have seen, the baseline counts (variable `Base`) are a very strong predictor of future counts (variable `count`). But on which scale do we expect them to be related? Above, we have assumed that `Base` is linearly related to  $\log(\text{count})$ . Does that make sense? In Figure 3.6, we see some simple scatter plots that help us getting a better feeling for the appropriate predictor scale. The curves that were put through the points are `loess` curves the default in `geom_smooth` for small datasets. It its not something I would use for inference, but in order to get a quick glance, I think it's nice actually. For more details on this method, see

?stats:::loess. Admittedly, none of the relationships looks particularly linear to me, but the bottom plots predicting log counts at least don't look like a exponential growth.

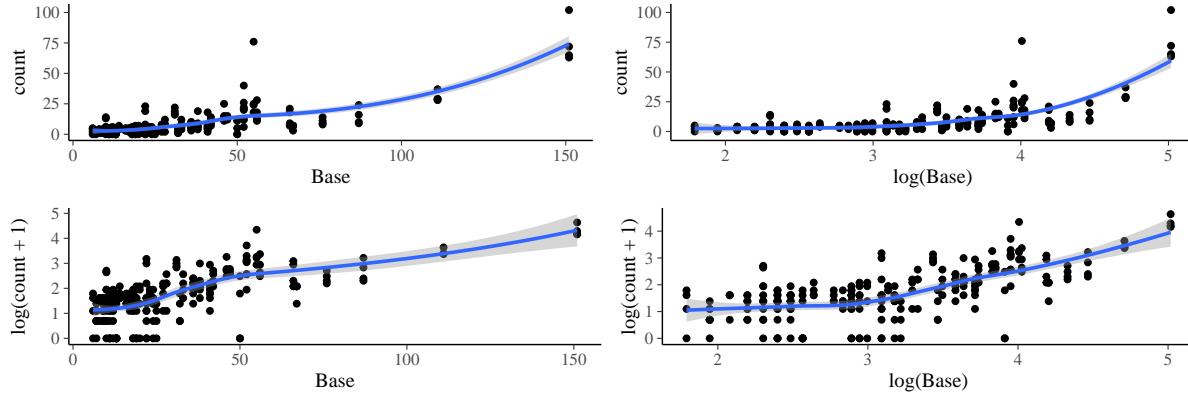


Figure 3.6: Scatter plots of the epilepsy data relating the `Base` predictor to the response `count` on both the original and the log scale.

In the original analysis, of the epilepsy dataset (Thall and Vail 1990), the authors used a log score of `Base`, so let's do the same now:

```
fit_epi_gaussian_log2 <- brm(log(count + 1) ~ Trt * scale(log(Base)),
                               data = epilepsy)
```

```
summary(fit_epi_gaussian_log2)
```

```
Family: gaussian
Links: mu = identity
Formula: log(count + 1) ~ Trt * scale(log(Base))
Data: epilepsy (Number of observations: 236)
Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
      total post-warmup draws = 4000
```

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	1.89	0.07	1.76	2.02	1.00	5022	3040
Trt1	-0.28	0.09	-0.46	-0.11	1.00	5334	2969
scalelogBase	0.56	0.06	0.44	0.68	1.00	3580	2753
Trt1:scalelogBase	0.16	0.10	-0.03	0.35	1.00	3461	2766

#### Further Distributional Parameters:

Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
----------	-----------	----------	----------	------	----------	----------

sigma	0.70	0.03	0.64	0.77	1.00	4244	3065
-------	------	------	------	------	------	------	------

Draws were sampled using sampling(NUTS). For each parameter, Bulk\_ESS and Tail\_ESS are effective sample size measures, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).

Qualitatively, the overall patterns remain the same. The higher (log) Base, the higher the predicted number of seizures, and treatment may have a small effect at least for people with small baseline of seizures. This is confirmed also visually (see Figure 3.7):

```
conditional_effects(fit_epi_gaussian_log2, effects = "Base:Trt")
```

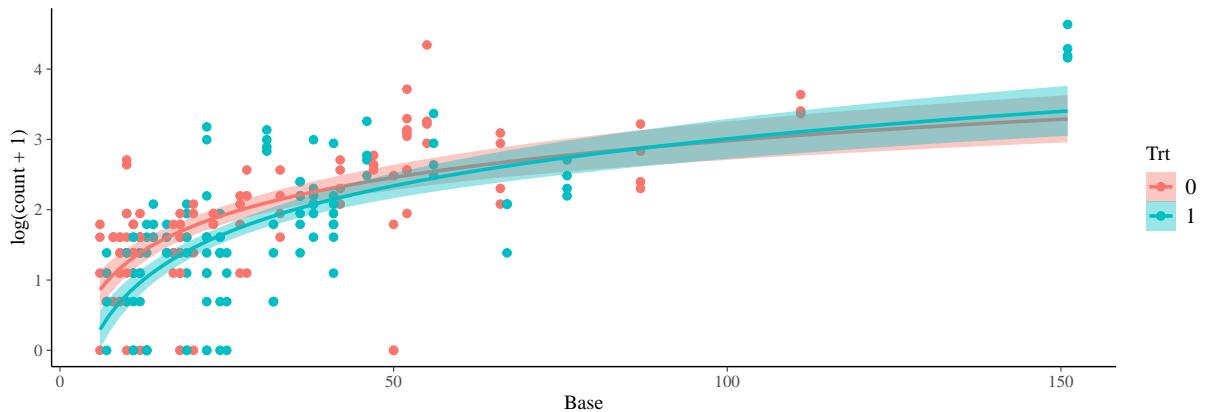


Figure 3.7: Conditional effects of the baseline-treatment interaction for model fit\_epi\_gaussian\_log2.

Notice that we still see the untransformed `Base` variable on the x-axis. Accordingly, the relationship is now logarithmic on the log scale of `count`, as a result of assuming a linear relation between `log count` and `log Base`. In order to see the predictions on the original scale of `count`, we again use the manual approach via `posterior_epred` shown above, which leads to the plot shown in Figure 3.8.

We see that the relationship of `count` and `Base` on the original scales is now almost linear, but this just a coincidence. The regression coefficients were different, the relationship could indeed be quite non-linear. We can see this by consider the following relationship between two variables  $x$  and  $y$  that resembles our situation:

$$\log(y) = b_0 + b_1 \log(x)$$

When we now exponentiate both sides and use the math rules of exponential functions, we get:

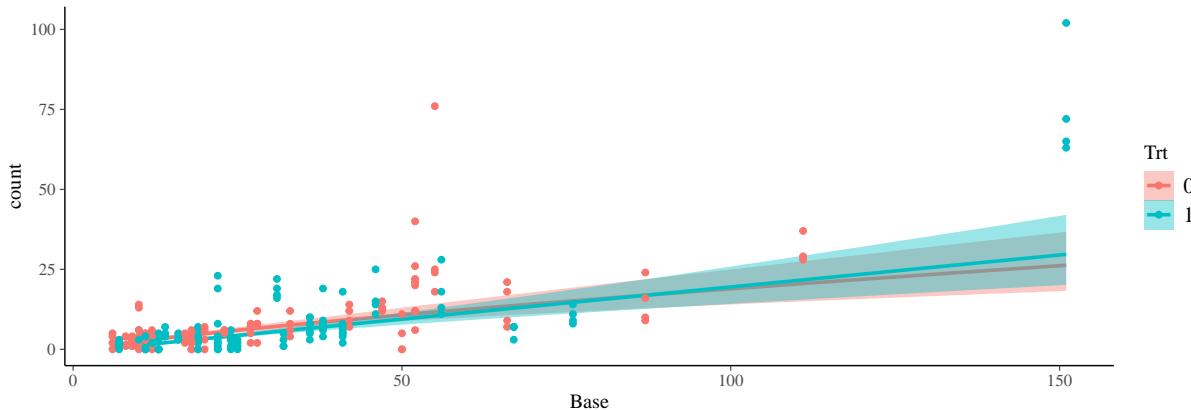


Figure 3.8: Conditional effects of the baseline-treatment interaction for model `fit_epi_gaussian_log2` with predictions back-transformed to the original count scale.

$$y = \exp(b_0 + b_1 \log(x)) = \exp(b_0)x^{b_1}$$

Accordingly, we relationship will be linear only if the slope parameter  $b_{-1}$  is not too far from 1. The functional form we have just created is also known as “power law”.

Back to our actual modeling task, we can ask which of the two models, using `Base` or `log(Base)`, leads to better predictions:

```
loo_compare(loo(fit_epi_gaussian_log1), loo(fit_epi_gaussian_log2))
```

	elpd_diff	se_diff
<code>fit_epi_gaussian_log1</code>	0.0	0.0
<code>fit_epi_gaussian_log2</code>	-2.6	4.6

Very similar indeed, certainly not a difference that we can base a decision on. Below, I will continue to use `log(Base)` as predictor for because assuming a power law relationship of baseline and current seizure counts seems like a theoretically sensible choice.

### 3.3.3 Lognormal models

The combination of Gaussian distribution on a log-transformed variable is so common that it has its own name: the lognormal distribution. To express this family in brms, we can either transform the response variable and then use `family = gaussian()` as we have done above, or use `family = lognormal()` on the “original” (here counts + 1) response:

```
fit_epi_lnorm1 <- brm(count + 1 ~ Trt * scale(log(Base)), data = epilepsy,
family = lognormal())
```

When you investigate the results, you will see that they are identical (up to MCMC error) to the corresponding Gaussian model on the log-scale. The advantage of the lognormal approach is that we are automatically getting predictions on the original scale, which makes visualizations and model comparisons easier. For example, if we were to run `conditional_effects`, we would immediately get the plot for the original scale (`counts + 1` in our case).

### 3.3.4 Poisson models

At the start of this chapter, I promised you GLMs. But all we saw so far were different variations of the normal likelihood. This is about to change. If people ask me what distributions they should consider as likelihoods, I always recommend to use some that are (what I call) *structurally faithful*, by which I mean that it fits naturally to the kind of response being modeled (Bürkner, Scholz, and Radev 2023b). For example, if my responses are lower-bounded counts, I should prefer distributions that are defined exactly for such kind of data. Probability theory and statistics are quite old fields by now, so for almost every imaginable kind of variable, someone has probably thought of corresponding structural faithful distributions. For lower-bounded counts, the Poisson distribution is the canonical choice (as an initial starting point). For all  $y = 0, 1, 2, \dots$ , the Poisson density looks as follows:

$$p(y | \lambda) = \frac{\lambda^y}{y!} \exp(-\lambda)$$

It just has one parameter,  $\lambda$ , the mean of the Poisson distribution. In Figure 3.9, you can see a visualization of the Poisson distribution for different  $\lambda$ .

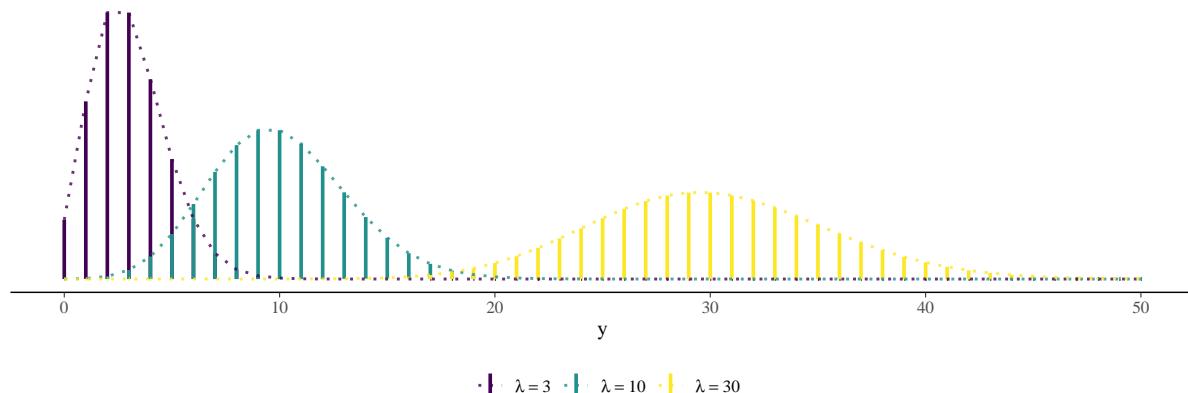


Figure 3.9: Three example densities of the Poisson distribution varying in the mean  $\lambda$ .

Since  $\lambda$  is the mean parameter, it constitutes the ideal prediction target in the context of GLMs. There is only one problem:  $\lambda$  is the mean of counts, so it must necessarily be positive. But a linear predictor  $\eta = b_0 + b_1 x_1 + \dots$  may take on any real values, also negative ones. So how do we make sure this won't cause trouble? The answer is that we won't set  $\lambda = \eta$  but rather  $\lambda = \exp(\eta)$ . We can write this equivalently as  $\log(\lambda) = \eta$ . Accordingly, in comparison to our previous approaches, we do not log transform the response variable directly but rather the mean parameter. We say that we have applied the *log link function*, which links the likelihood's mean parameter to the linear predictor.

The inverse link function, here `exp` is also called *response function*. I personally find it more intuitive to think in terms of response functions rather than link functions, because the former follows the natural logic of computation. I first compute my linear predictor  $\eta$ , than I apply the response function to obtain  $\lambda$ , then I plug  $\lambda$  into the Poisson likelihood. That being said, in the established Statistics literature, we call these transformation always by their link function name and I guess it's too late to change this habit now. Accordingly, also in `brms`, we specify the link rather than the response function. Let us run our first Poisson model in `brms`:

```
fit_epi_poisson1 <- brm(count ~ Trt * scale(log(Base)), data = epilepsy,
                           family = poisson("log"))
```

The main change is the specification of `family = poisson(link = "log")`. We might have also written just `family = poisson()` leading to the same result, because `log` is the default link for Poisson models. Notice also, that we no longer need the awkward `count + 1` approach, since we now log transform the Poisson mean  $\lambda > 0$ , which can never become exactly zero.

```
summary(fit_epi_poisson1)
```

```
Family: poisson
Links: mu = log
Formula: count ~ Trt * scale(log(Base))
Data: epilepsy (Number of observations: 236)
Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
      total post-warmup draws = 4000
```

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	1.90	0.04	1.82	1.98	1.00	1932	2023
Trt1	-0.38	0.06	-0.50	-0.26	1.00	1354	2291
scalelogBase	0.71	0.03	0.65	0.78	1.00	1591	1838
Trt1:scalelogBase	0.31	0.05	0.22	0.40	1.00	1340	1939

Draws were sampled using `sampling(NUTS)`. For each parameter, Bulk\_ESS

and Tail\_ESS are effective sample size measures, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).

As opposed to the lognormal model from earlier on, the Poisson model suggests a substantial interaction effect of baseline seizures and treatment. Not only is the posterior mean about twice as large, but also the posterior standard deviation is about twice as small. Which of these two results should we rather trust? Let's first approach this question by investigating the Poisson model fit graphically (see Figure 3.10):

```
pp_check(fit_ephi_poisson1)
```

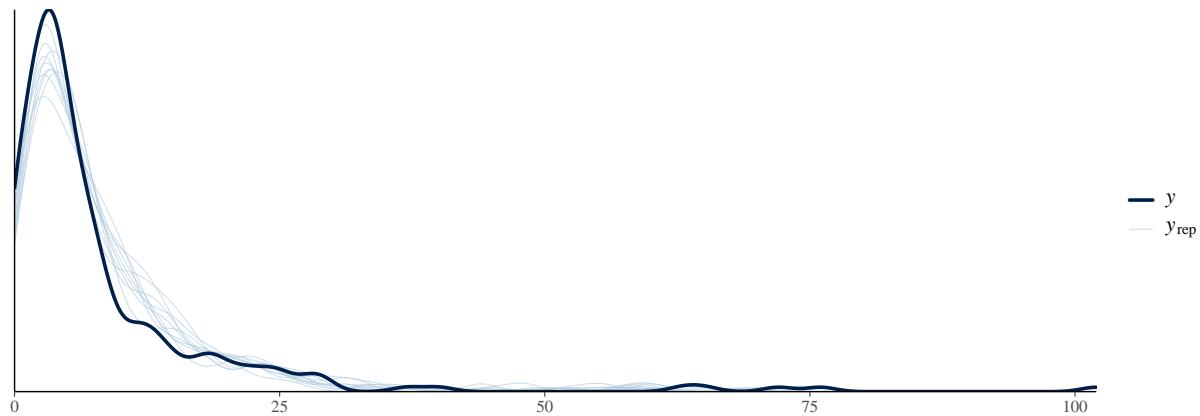


Figure 3.10: Default posterior predictive check for model `fit_ephi_poisson1`.

As you know by now, the default posterior predictive check produces a density, which is of course not an ideal representation for count data. In the present case, it still does the job. We see that the marginal fit is quite good overall. However, already in this plot, we see seemingly small but, to the experienced eye, clearly visible issues: The bulk of the data distribution close to zero is underestimated by the Poisson model and so is the right tail. There is a specific reason for this pattern: The Poisson distribution has just a single parameter  $\lambda$ , which happens to be not only the mean but also the variance of the distribution; a property called *equidispersion*. However, most real-world count data is actually *overdispersed*, that is, has higher variance than mean, something that the Poisson distribution is not able to handle.

In the context of regression models, we should perhaps better speak of *conditional equidispersion* and *conditional overdispersion*, because what matters are the data distribution after conditioning on the predictors. This is actually a very important point that I think can barely be stressed enough regardless of the statistical framework you fit our model in: The likelihood distribution should be appropriate for the noise in the responses that remains after conditioning on the predictors. Using normal linear regression as an example, we do not need to care

if the marginal (i.e. unconditional) response distribution is actually normal, but if the (conditional) distribution of the residuals is normal. The same holds for all kinds of other likelihood distributions, including Poisson models. While the above posterior predictive check shows only the marginal distribution, if the conditional distributions are overdispersed, we will likely see this in the marginal too.

### 3.3.5 Negative binomial models

So how can we deal with conditional overdispersion in count data models? The negative binomial distribution is perhaps the most popular option. It has two parameters and a mean parameterization that is commonly used for the purpose of GLMs: One parameter is still being the mean  $\lambda > 0$  that we also know from Poisson and a second is a shape (precision) parameter  $\phi > 0$  which controls the amount of overdispersion. As  $\phi$  increases towards infinity, the negative binomial distribution approaches equidispersion, and thus the Poisson distribution. As such the negative binomial is a generalization of Poisson. The density of the negative binomial (in the parameterization used for GLMs) is quite involved, and I show it here just for completeness:

$$p(y | \lambda, \phi) = \binom{x + \phi - 1}{x} \left( \frac{\lambda}{\lambda + \phi} \right)^x \left( \frac{\phi}{\lambda + \phi} \right)^\phi$$

Let's get a better intuition by visualizing some examples of the negative binomial density:

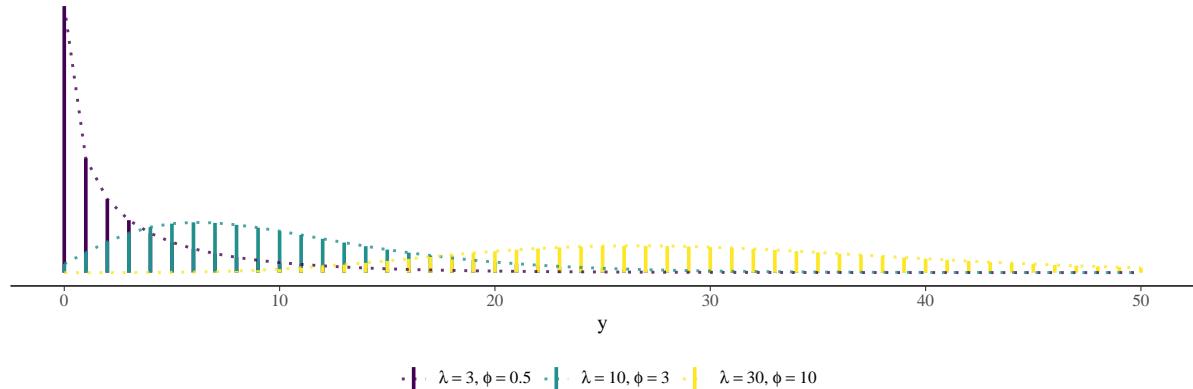


Figure 3.11: Three example densities of the negative binomial distribution varying in the mean  $\lambda$  and shape  $\phi$ .

In brms, we specify a negative binomial likelihood via `family = negbinomial()`. As for the Poisson, the default link function is "log" so if we don't specify a link ourselves, we will get the log link anyway. We are now ready to fit our first negative binomial model in brms:

```
fit_epi_negbin1 <- brm(count ~ Trt * scale(log(Base)), data = epilepsy,
                        family = negbinomial())
```

```
summary(fit_epi_negbin1)
```

```
Family: negbinomial
Links: mu = log
Formula: count ~ Trt * scale(log(Base))
Data: epilepsy (Number of observations: 236)
Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
      total post-warmup draws = 4000
```

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	1.91	0.07	1.77	2.05	1.00	4756	2738
Trt1	-0.31	0.10	-0.52	-0.10	1.00	4243	2671
scalelogBase	0.69	0.06	0.56	0.82	1.00	2910	3117
Trt1:scalelogBase	0.20	0.10	0.01	0.39	1.00	2981	3044

#### Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
shape	2.55	0.35	1.93	3.27	1.00	4180	2636

Draws were sampled using sampling(NUTS). For each parameter, Bulk\_ESS and Tail\_ESS are effective sample size measures, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).

Let us check the results. First of all, we see that the model seems to have converged well, for some parameters even with super efficiency (higher ESS than the total number of draws). The posteriors of the regression coefficients look qualitatively similarly than the ones of the Poisson model. That is, the coefficient of the treatment (for average baseline scores) is negative, the coefficient of baseline (for the control group) is strongly positive, and the interaction coefficient is positive. That said, both the posterior mean and the posterior uncertainty vary clearly between negative binomial and Poisson: The former not only has smaller coefficients on average (posterior means closer to zero), but also larger posterior uncertainty (larger posterior standard deviations and credible intervals). This together implies that the negative binomial model is far less certain about the predictive value of the covariates than the Poisson model. But why is this the case?

Let's investigate the `shape` parameter at the bottom of the summary output. It's mean is around 2.57 with a credible interval between roughly 1.94 and 3.33. That doesn't tell us much

admittedly. We only know that if `shape` would be infinite, the negative binomial and Poisson distribution would be the same. But how far is 2.57 away from infinity in our case? That is, how can we interpret the `shape` parameter? For this purpose, it is useful to investigate the relationship of the `shape` parameter  $\phi$  with the variance of the negative binomial distribution:

$$\text{Var}(y) = \lambda + \lambda^2/\phi$$

From the above formula for the variance, we see that the impact of  $\phi$  on the variance also depends on the mean  $\lambda$ . So let's take a look at the ratio of variance and mean for different combinations of  $\lambda$  and  $\phi$ , as shown in Figure 3.12. On the y-axis, we see the variance-mean ratio, that is:

$$\frac{\text{Var}(y)}{E(y)} = \frac{\lambda + \lambda^2/\phi}{\lambda} = 1 + \lambda/\phi.$$

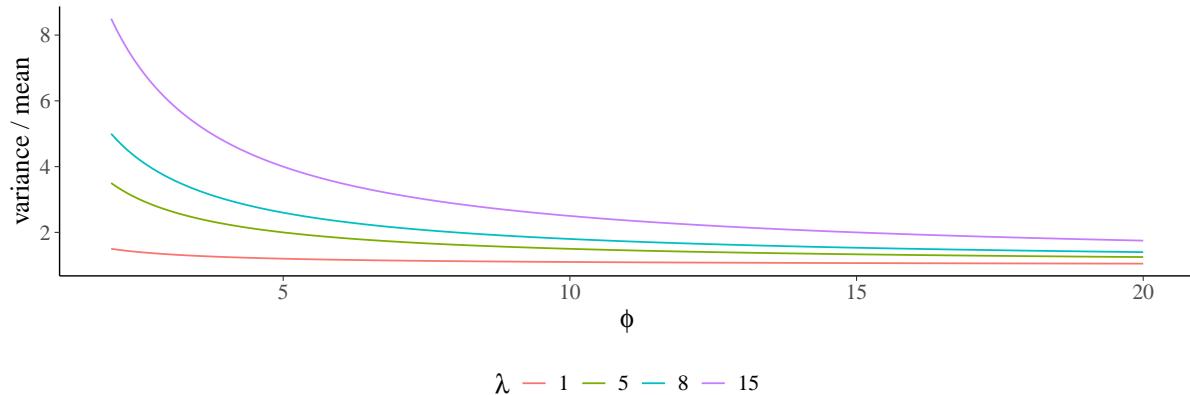


Figure 3.12: Variance-mean ratio of the negative binomial distribution as a function of the mean  $\lambda$  and shape  $\phi$ .

Based on the formula and as also visible in Figure 3.12, we see that the variance-mean ratio converges to 1 (i.e., to equidispersion) for  $\phi \rightarrow \infty$ . We also see that for small  $\phi$  values, the ratio can be substantially bigger than zero, especially if the mean  $\lambda$  is not very small. To make this more concrete, let us take a look what ratio we can expect for our model. First, we use a quick and dirty approach by using the exponential transformed posterior mean of `Intercept`  $\approx \exp(1.91) \approx 6.75$  as  $\lambda$  (i.e., the mean when all predictors are set to zero) and the posterior mean of `shape`  $\approx 2.57$  as  $\phi$ . Then we find the ratio to be  $1 + \lambda/\phi = 1 + 6.75/2.57 = 3.62$ . Accordingly, variance-mean ratios of around 3 to 4 are not unrealistic for our negative binomial model; ratios that are indeed far away from what Poisson would expect (i.e., a ratio of 1).

For the purpose of practicing working with posterior draws, let us check out the overdispersion a bit more systematically for our model. First, we will extract the draws of the posterior predictive means for each observation as well as the draws from `shape`

```

draws_pp_means <- posterior_epred(fit_epi_negbin1)
str(draws_pp_means)

num [1:4000, 1:236] 3.33 3.24 3.54 3.84 3.82 ...

draws_shape <- as.data.frame(fit_epi_negbin1, variable = "shape")$shape
str(draws_shape)

num [1:4000] 2.66 2.3 2.81 2.33 2.04 ...

```

The `draws_pp_means` matrix has one column per observation and one row per posterior draw. The `draws_shape` vector has one element per posterior draw, since we assumed `shape` to be constant across observations. R stores matrices in column major order. That is, when we compute transformations involving both matrices and vectors, the vector is implicitly extended to match the size of a matrix, in a way that the first column is filled first, then the second column, and so on. In other words, if we work with vectors, matrices, or arrays where the first dimension (i.e., the rows) are indicating the draws, we can just combine matrices, array, and vectors in the same vectorized R expression, provided that their number of draws match. Suppose I want to get the posterior means of the variance-mean ratio for each observation. Then I can write:

```

draws_var_mean_ratios <- 1 + draws_pp_means / draws_shape
means_var_mean_ratios <- colMeans(draws_var_mean_ratios)
str(means_var_mean_ratios)

```

```
num [1:236] 2.36 2.36 1.78 2.01 8.04 ...
```

A histogram of these posterior means can be found in Figure 3.13. Indeed, there are a lot of observations with a variance-mean ratio higher than 3. One observation even has a ratio of about 20.

All of the above analyses suggest substantial overdispersion in the data. This in turn suggests that the negative binomial family is a better choice than the Poisson family. To be fair, just looking at the `shape` parameter in the summary output makes this pretty clear, but I wanted to show you these additional analyses to help you understand my rational better. Let's see if LOO-CV agrees with our conclusion.

```

loo_epi_poisson1 <- loo(fit_epi_poisson1)
loo_epi_negbin1 <- loo(fit_epi_negbin1)

```

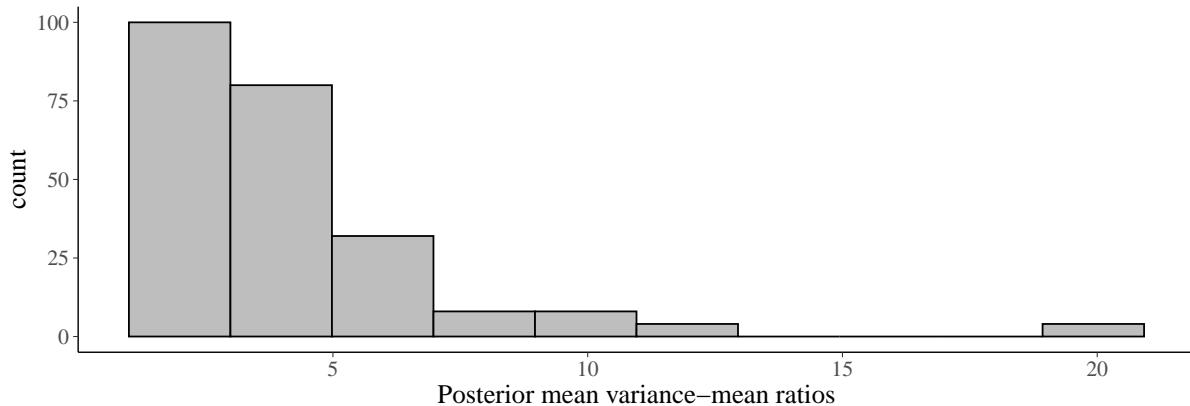


Figure 3.13: Histogram of posterior mean variance-mean ratios per observation of the negative binomial model.

```
loo_epi_poisson1
```

Computed from 4000 by 236 log-likelihood matrix.

	Estimate	SE
elpd_loo	-864.8	80.1
p_loo	22.3	7.2
looic	1729.5	160.2
-----		
MCSE of elpd_loo	is NA.	
MCSE and ESS estimates assume MCMC draws (r_eff in [0.3, 1.1]).		

Pareto k diagnostic values:

		Count	Pct.	Min.	ESS
(-Inf, 0.7]	(good)	235	99.6%	91	
(0.7, 1]	(bad)	1	0.4%	<NA>	
(1, Inf)	(very bad)	0	0.0%	<NA>	

See `help('pareto-k-diagnostic')` for details.

While computing approximate LOO-CV, we encountered the above warning. It suggests that PSIS failed for one observation in the Poisson model (Pareto  $\hat{k} > 0.7$ ). A closer inspection reveals that one of these problematic observation number 49 containing the largest observed count in the data with `count = 102`. This observation we already saw poorly predicted in the posterior predictive check earlier. This alone points to some potential misfit of the Poisson model. And, since the warning didn't occur for the negative binomial model, it provides some

further evidence that the latter is more appropriate. Remember from Section 2.7.2 that the Pareto  $\hat{k}$  has a double interpretation in the context of LOO-CV. First, it tells us how well PSIS worked to approximate the LOO-posterior for the full posterior of a given observation. Second, it tells us how influential individual observations are for the posterior. And our observation number 49 seems to be quite influential indeed. Anyway, if we set the issue with this observation aside for a second, what does the LOO-based model comparison currently tell us?

```
loo_compare(loo_epi_poisson1, loo_epi_negbin1)
```

	elpd_diff	se_diff
fit_epi_negbin1	0.0	0.0
fit_epi_poisson1	-208.4	66.0

Indeed, and unsurprisingly in light of what we saw before, the negative binomial model is overwhelmingly better, both in absolute ELPD points and relatively to the corresponding standard error. This comparison is certainly biased because PSIS failed badly for a single observation, but the difference is so brutal that a single failed PSIS is highly unlikely to explain it. Accordingly, we could in good faith just leave it at that and go with the negative binomial model. But out of curiosity, let's fix PSIS by running moment matching:

```
loo_epi_poisson1_mm <- loo_moment_match(fit_epi_poisson1, loo_epi_poisson1)
loo_epi_poisson1_mm
```

Computed from 4000 by 236 log-likelihood matrix.

	Estimate	SE
elpd_loo	-865.2	80.2
p_loo	22.8	7.5
looic	1730.5	160.4
-----		
MCSE of elpd_loo	is 0.1.	
MCSE and ESS estimates assume MCMC draws (r_eff in [0.3, 1.1]).		
All Pareto k estimates are good (k < 0.7).		
See help('pareto-k-diagnostic') for details.		

Indeed, it seemed to have worked as now all Pareto  $\hat{k} < 0.7$ . Let's compare the two models again:

```
loo_compare(loo_epi_poisson1_mm, loo_epi_negbin1)
```

	elpd_diff	se_diff
fit_epi_negbin1	0.0	0.0
fit_epi_poisson1	-208.9	66.1

There seems to be barely any difference to the comparison before moment matching, but there might as well have been a much bigger one. So please, take high Pareto  $\hat{k}$  values seriously, especially if there are several of them and the difference between models is not as huge as it was in our current case. With all this evidence, we are finally able to let the Poisson likelihood go and choose to move forward with negative binomial models.

I have worked with many unbounded count responses in the past, and the negative binomial model was almost always better than the corresponding Poisson model. So why is real data so often overdispersed? If the events come in over time at a certain constant rate, the Poisson distribution arises naturally. That is why some people used to see very clean data, e.g., in some parts of physics, will equate any unbounded count process with the Poisson distribution. But medical, psychology, or biological data from living beings is rarely that clean. Even if measurement itself is relatively precise, living beings especially humans or other higher developed species tend to vary quite strongly from individual to individual, and even within the same individual over time. This variation is almost impossible to fully account for by means of the predicting variables we have access to. Any unaccounted within- or between-individual variation will necessarily become an additional source of noise from the perspective of the model. This noise is what causes the overdispersion.

### 3.3.6 More adventures into model comparison

While reading this chapter, you may have asked yourself already why we didn't compare the lognormal models fitted at the start to the Poisson and negative binomial models. The reason is that the former have a continuous likelihood distribution while the latter have a discrete likelihood distributions. So, for the former, probabilities are integrals under the density, while for the latter, probabilities are the density values themselves. So really quite something different.

But is it valid to compare continuous and discrete likelihood models using density-based metrics such as ELPD? The short answer is *no*. The long answer is that it actually works approximately, at least in certain cases. If you are happy with the short answer and had enough of math today, you can safely skip this section and move right to the next case study in Section 3.4.

For those of you who stayed, let's just check the LOO comparisons for a start:

```
loo_epi_lnorm1 <- loo(fit_epi_lnorm1)
loo_compare(loo_epi_lnorm1, loo_epi_poisson1_mm, loo_epi_negbin1)
```

Warning: Not all models have the same y variable. ('yhash' attributes do not match)

	elpd_diff	se_diff
fit_epi_negbin1	0.0	0.0
fit_epi_lnorm1	-8.9	5.8
fit_epi_poisson1	-208.9	66.1

We get a warning that the response variable is not the same across models. This is not because of the continuous vs. discrete situation but because we have added 1 to the count to make them usable with the lognormal likelihood. If we take these results seriously for a second, we would conclude that the lognormal model is perhaps slightly worse than the negative binomial model but substantially better than Poisson. But can we trust these conclusions? If we remember the discussion about ELPD from Section 2.7, we see that it is built on log-likelihood density values. For LOO-CV via ELPD to be fair, the log-likelihood density values need to be comparable. And for the latter to be the case, they need to be on the same scale. For discrete densities, that is easy since we just have to make sure that the density values sum to 1. Let's verify this for the negative binomial model, by exemplary summing over the density values of count 0 to 500 for a person in the treatment group with a Base seizure count of 5:

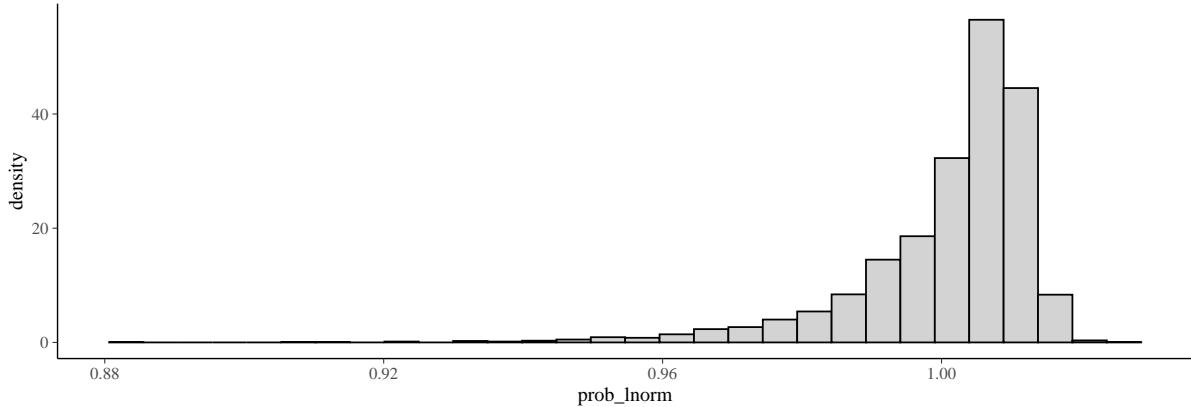
```
dat_all_base <- data.frame(Trt = 0, Base = 5, count = seq(0, 500, 1))
```

```
prob_negbin <- log_lik(fit_epi_negbin1, newdata = dat_all_base) %>%
  exp() %>%
  rowSums()
```

All of these probabilities evaluate to 1, more precisely almost 1 since we didn't add the vanishingly small probability of predicting seizure counts over 500. Now, let's do the same with the lognormal model:

```
prob_lnorm <- log_lik(fit_epi_lnorm1, newdata = dat_all_base) %>%
  exp() %>%
  rowSums()
```

```
data.frame(prob_lnorm = prob_lnorm) %>%
  ggplot(aes(prob_lnorm, after_stat(density))) +
  geom_histogram(fill = "lightgrey", color = "black", bins = 30)
```



Okay, so not exactly 1 but actually quite close for most draws. First of all, why not exactly 1? As mentioned before, for continuous densities, probabilities are defined as integrals under the density not as sums of density values at a bunch of discrete points. So there is no reason to expect such a sum to be 1 exactly. And yet we found sums which are actually quite close to 1. Was that just luck or is there actually something systematic behind it?

What we did was to evaluate the lognormal distribution only at the positive natural numbers  $k = 1, 2, \dots$  and then sum over the resulting density values. Since these values are equidistantly placed with a distance of one, we can turn these discretized density values into a rectangles of width 1 with height equal to the density values  $p(k)$ . For  $k = 1$  the rectangular goes from  $y = 0.5$  to  $y = 1.5$ , for  $k = 2$  it goes from  $y = 1.5$  to  $y = 2.5$ , and so on. Now, we can interpret  $p(k)$  as approximating the average density in the whole surrounding interval

$$p(k) \approx \int_{k-0.5}^{k+0.5} p(y) dy.$$

As a result, we get

$$\begin{aligned} \sum_{k=1}^{\infty} p(k) &\approx \sum_{k=1}^{\infty} \int_{k-0.5}^{k+0.5} p(y) dy \\ &= \int_{0.5}^{\infty} p(y) dy \\ &= 1 - \int_0^{0.5} p(y) dy \\ &\approx 1. \end{aligned}$$

The second approximation we have made is to ignore all the density between in the interval  $[0, 0, 5]$ , which is justified as long as the log-mean parameter  $\mu$  of the lognormal distribution

is not very small (see Figure 3.14). This shows that, with a little bit of squinting, using likelihood-density based measures such as ELPD to compare actual count-data models with their continuous approximations can be a valid approach, if we interpret only large differences. With regard to the current case study, I wouldn't use ELPD to choose among the negative binomial and the lognormal model because their performance is so similar that the small errors induced by the continuous vs. discrete comparison may actually matter. But we can safely conclude that both are substantially better than Poisson.

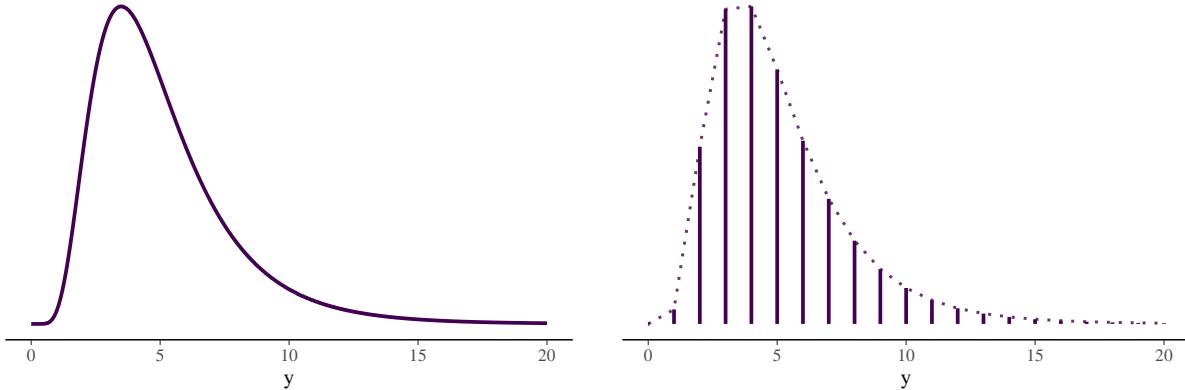


Figure 3.14: Lognormal density with log-mean  $\mu = 1.5$  and log-SD  $\sigma = 0.5$ . Left: Regular density function. Right: Discretized version of the density on the positive natural numbers ( $1, 2, \dots$ ).

With this, let us leave the epilepsy dataset for now. We will come back to it later in the context of multilevel models in Chapter 5.

### 3.4 GLMs for double-bounded responses

In this section, we will take an initial look at Bayesian GLMs for double-bounded responses, that is, responses that have both an upper and a lower bound. In particular, binary (0 or 1) data falls within this class and we will focus on them for now. As a case study, we will predict the fate of passengers on the fatal maiden voyage of the famous ocean liner “Titanic”. Among others due to insufficient number of live boats, 1502 out of 2224 passengers and crew died in the Titanic accident. Our goal is to figure out which groups of passengers had a higher chance of surviving. For this, we use the dataset shipped with the `titanic` R package (Hendricks 2015), which resembles the one provide in the `titanic` Kaggle challenge (<https://www.kaggle.com/competitions/titanic>). For each of the passengers, we have quite a bit of data available, in particular their socio-economic class (variable `Pclass`; coded as first, second, and third class), their `Sex` (a binary indicator; in 1912 we weren't that far in terms of acknowledging sex and gender diversity), their `Age` (in years), as well as some other information, such as the kind of tickets, fare, and the cabin they were staying in (if known). The

response variable `Survived` is 1 if a person survived and 0 if they died. An overview of the data is shown in Table 3.3.

Table 3.3: First six observations of the `titanic` dataset.

Survived	Pclass	Sex	Age	Ticket	Fare	Cabin
0	3	male	22	A/5 21171	7.25	
1	1	female	38	PC 17599	71.28	C85
1	3	female	26	STON/O2. 3101282	7.92	
1	1	female	35	113803	53.10	C123
0	3	male	35	373450	8.05	
0	3	male	NA	330877	8.46	

We will primarily focus on the predictors `Pclass`, `Sex`, and `Age`. The former two we will turn into factors to ensure proper behavior within R formula syntax:

```
titanic <- titanic::titanic_train %>%
  mutate(Pclass = factor(Pclass),
        Sex = factor(Sex))
```

As likelihood, we will choose the Bernoulli distribution. Due to responses being either 0 or 1, it has a remarkably simple form. We say that  $y = 1$  (i.e., a person survived the titanic accident) happened with a to-be-estimated probability  $\theta$  and, accordingly,  $y = 0$  (the person died) happens with probability  $1 - \theta$ . We can write this conveniently into a single equation which constitutes the density of the Bernoulli distribution:

$$p_{\text{Bernoulli}}(y | \theta) = \theta^y (1 - \theta)^{1-y}$$

Now, we of course don't want to just estimate a single overall  $\theta$  for the whole dataset. That would be a bit boring. Instead, we want to understand how  $\theta$  varies between people. For this purpose, we will again set up a linear predictor  $\eta$  that we will connect to  $\theta$  via an appropriate response function. Since  $\theta$  is a probability parameter by definition, it can only take on values in the unit interval  $[0, 1]$ . Accordingly, we need a response function that maps real values to values in the unit interval. Here, we choose the logistic response function, which is the most common choice in most fields:

$$\theta = \text{logistic}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} = \frac{1}{1 + \exp(-\eta)}$$

The corresponding link (inverse response) function is called the *logit* link and is given by:

$$\eta = \text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right) = \log(\theta) - \log(1-\theta)$$

First, we want to predict  $\theta$  by the class and sex of the passengers. We use dummy coding for these predictors with 3rd class and females as reference categories, respectively. This leads us to the following linear predictor for the  $n$ th observation:

$$\eta_n = b_0 + b_1 \text{Class1}_n + b_2 \text{Class2}_n + b_3 \text{SexMale}_n$$

$$b_i \sim \text{normal}(0, 3) \quad i = 0, \dots, 3$$

In brms syntax, this model can be specified as follows:

```
fit_titanic1 <- brm(
  Survived ~ Pclass + Sex,
  data = titanic, family = bernoulli("logit"),
  prior = prior(normal(0, 3), class = "Intercept") +
    prior(normal(0, 3), class = "b")
)
```

```
summary(fit_titanic1)
```

```
Family: bernoulli
Links: mu = logit
Formula: Survived ~ Pclass + Sex
Data: titanic (Number of observations: 891)
Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
      total post-warmup draws = 4000
```

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	2.29	0.22	1.88	2.75	1.00	2974	2711
Pclass2	-0.83	0.25	-1.31	-0.37	1.00	3130	2682
Pclass3	-1.90	0.21	-2.33	-1.49	1.00	3152	2479
Sexmale	-2.64	0.19	-3.01	-2.30	1.00	3371	2916

Draws were sampled using sampling(NUTS). For each parameter, Bulk\_ESS and Tail\_ESS are effective sample size measures, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).

What do we learn from this summary? Looking at the `Pclass` predictor first we see that the regression coefficients of the second and third class – relative to the first class as reference – are clearly negative. This means the probability of survival is clearly lower in the second and third classes than in the first class. But how much lower? This is hard to tell from these numbers alone, since they are all on logit scale. With time you will get a feeling for regression coefficients on logit scales and can tell how big of an effect it roughly corresponds to on the probability scale. We will learn more about this later on. For now, let it be sufficient to say that an absolute group difference of roughly 2 is quite large already, as we also see visualized in Figure 3.15.

```
conditional_effects(fit_titanic1, "Pclass")
```

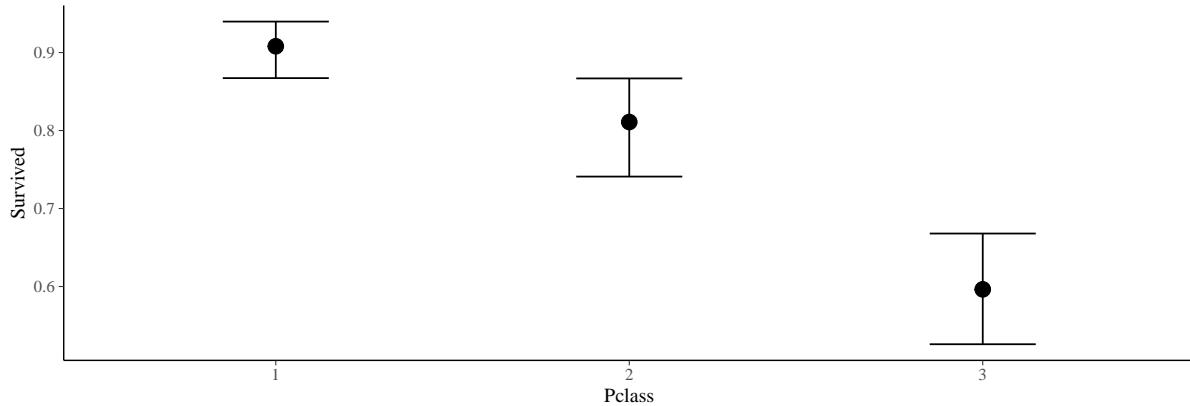


Figure 3.15: Conditional effects of `Pclass` for model `fit_titanic1`.

The same is true for the difference in survival probabilities between men and women, with man being estimated to have much lower survival probability than women overall. In the summary output, we see this from the clearly negative regression coefficient of `Sexmale`.

We modeled only the main effects of class and sex, so on the logit scale their predictions of the survival probability are only additive by definition. However, since the exponential transform (part of the logistic response function) turns additive relationships into multiplicative ones, as we have already discussed in the context of the count data models in Section 3.3. That is, if we look at the predictions on the response scale, say via `conditional_effects`, we can see the non-linearity of the response function to have an influence (see Figure 3.16).

```
conditional_effects(fit_titanic1, "Pclass:Sex")
```

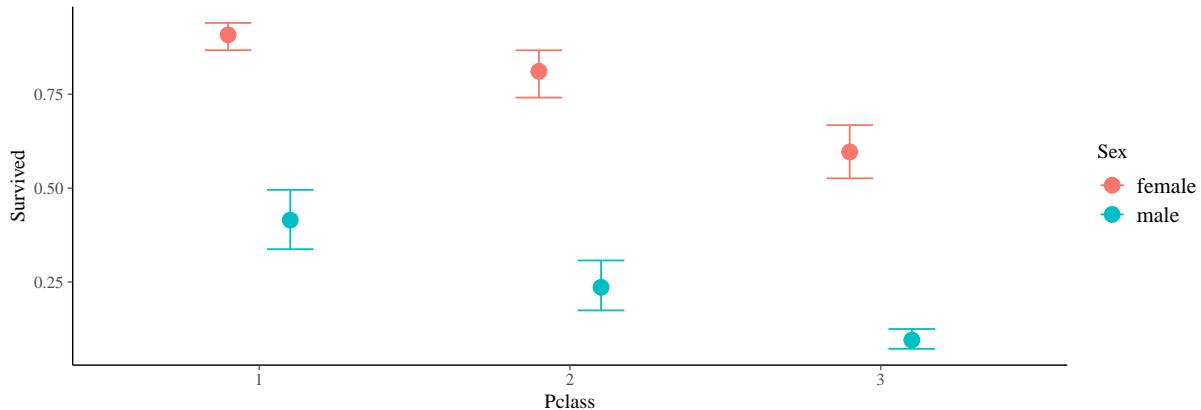


Figure 3.16: Conditional effects of the `Pclass:Sex` interaction for model `fit_titanic1`.

If the effects of the predictors were purely additive on the response scale, then the lines we could put between through the two blue and red points, respectively, would be exactly parallel to each other. But as you can see, they are not quite parallel. Instead they are “bend” towards the boundaries of  $\theta = 0$  or  $\theta = 1$ . This is a consequence of the response function ensuring in a smooth way that predictions will never cross these boundaries. In other words, even if things are linear on the latent scale, a certain amount of non-linearity is unavoidable in GLMs with non-identity link functions.

How good does our model fit the data though? Let’s start with some posterior predictive checks (see Figure 3.17):

```
pp_check(fit_titanic1, ndraws = 10)
```

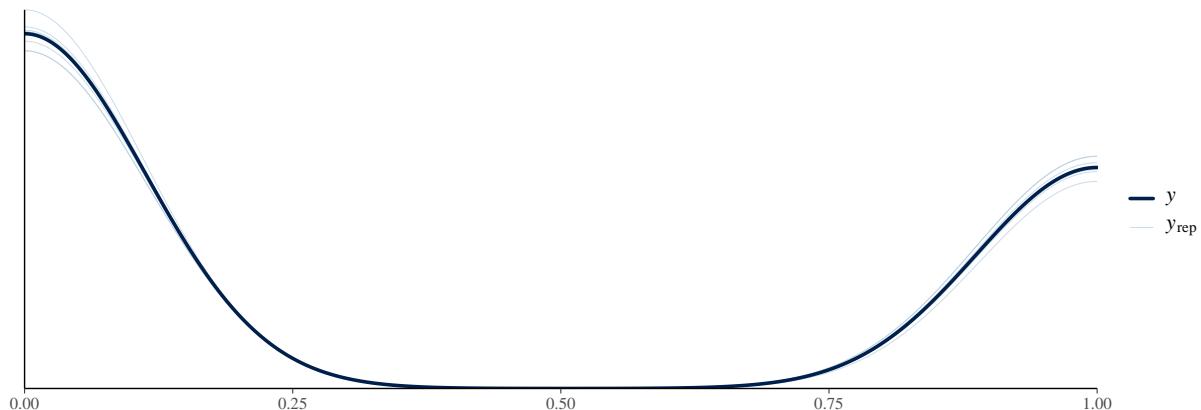


Figure 3.17: Default posterior predictive check for model `fit_titanic1`.

Admittedly, this predictive check is a bit pointless. It merely tells us that the model got

the ratio of 0s and 1s in the dataset right, something that a Bernoulli model with only an intercept can already do perfectly. More generally, a lot of the standard predictive checks are not sensible for binary responses as the space of model predictions (the unit interval) is different from the space of possible responses (only 0 or 1).

One predictive check that kind of works is the "error\_binned" type, because it bins responses into groups and then averages them per group. The resulting binned responses, and the corresponding residuals, are now on the same scale as the predictions (i.e., within the unit interval), so comparison becomes sensible:

```
pp_check(fit_titanic1, type = "error_binned", ndraws = 10)
```

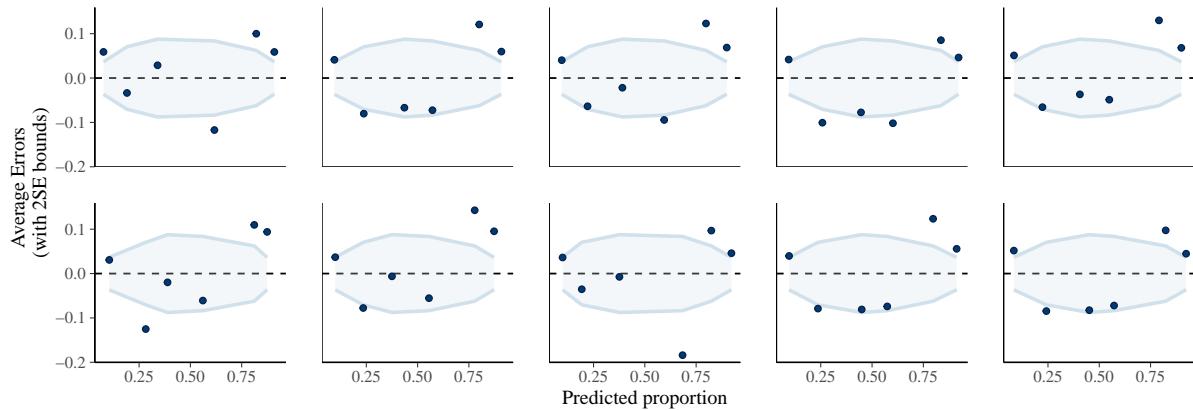


Figure 3.18: Posterior predictive check of binned errors for model `fit_titanic1`.

Every facet in Figure 3.18 corresponds to one posterior draw. Here, we decided to plot 10 draws such that the plot does not get too large. Bins are done according to the predicted proportions. That is, data points that the model assigned similar predictions to will be binned together. We can then check two things: (a) How big are the residuals of the binned responses on the probability scale and (b) how well does the uncertainty in the posterior predictive distribution reflect these residuals? With regard to (a), we see that the absolute residuals often exceed 10%, a quite substantial prediction error I would say. With regard to (b), the blue shadowed areas should contain roughly 95% of the corresponding points if the model's predictions were well calibrated. But here it seems to be the case for only roughly 50% of the points. In other words, the model is *overconfident*, that is, it underestimates the error in its own predictions. Together, this indicates that we should think about how to further improve our model.

With `fit_titanic1` we have seen that women, overall, survived with a substantially higher probability than men, suggesting that the policy of evacuating women first had worked, at least to some degree. However, we have also seen that second and third class passengers, overall, survived with a substantially lower probability than first class passengers. This begs

the question whether the difference in survival probability between woman and men was the same across classes. In other words, whether the interaction between sex and class is predictive of the survival probability.

Before we move on to more complicated models, let's quickly run LOO-CV and store the results so we can reuse it efficiently later.

```
fit_titanic1 <- add_criterion(fit_titanic1, "loo")
loo(fit_titanic1)
```

Computed from 4000 by 891 log-likelihood matrix.

```
    Estimate    SE
elpd_loo   -417.3 17.0
p_loo        3.7  0.2
looic      834.6 34.0
-----
MCSE of elpd_loo is 0.0.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.8, 1.1]).
```

All Pareto k estimates are good (k < 0.7).  
See help('pareto-k-diagnostic') for details.

The loo summary output shows that PSIS has worked out well (all Pareto  $k < 0.7$ ). What is more, we see an effective number of parameters  $p_{\text{loo}} = 3.69$ . This coincides nicely with the number of parameters themselves (4), which come in the form of 4 regression coefficients. This also illustrates that our priors were indeed weakly informative. Had we chosen (much) more informative priors, we would have seen  $p_{\text{loo}}$  to reduce noticeably. Try it out yourself, for example, by assigning highly informative `normal(0, 0.1)` priors to the regression coefficients and then running `loo` again.

### 3.4.1 Adding interactions

Was the policy of evacuating women first applied with equal success across the different passenger classes? This call for a model including the interaction of class and sex. Leaving everything else the same, our new model looks as follows:

```
fit_titanic2 <- brm(
  Survived ~ Pclass * Sex,
  data = titanic, family = bernoulli("logit"),
```

```
prior = prior(normal(0, 3), class = "b")  
)
```

Sampling and convergence are good so we directly check out the posterior of the regression coefficients (see Table 3.4):

```
summary_table(fit_titanic2)
```

Table 3.4: Posterior summary for model `fit_titanic2`.

variable	mean	sd	q5	q95
b_Intercept	3.22	0.50	2.45	4.12
b_Pclass2	-0.69	0.63	-1.74	0.35
b_Pclass3	-3.20	0.52	-4.15	-2.39
b_Sexmale	-3.74	0.53	-4.66	-2.92
b_Pclass2:Sexmale	-0.49	0.70	-1.61	0.67
b_Pclass3:Sexmale	1.85	0.57	0.96	2.87

Remember that first class female passengers are the reference category here. In the first class, women had a clearly higher survival rate than men as indicated by the posterior of the main coefficient of `Sexmale` lying strongly below zero. When looking at the posterior of the interaction coefficient related to third class male passengers (coefficient `Pclass3:Sexmale`), we see that it is strongly positive. This means that, compared to first class, the difference in survival rates between women and men was much smaller in the third class, although still in favor of women.

It is much easier to get an understanding of the regression coefficients' meaning once we visualize the predictions:

```
conditional_effects(fit_titanic2, "Pclass:Sex")
```

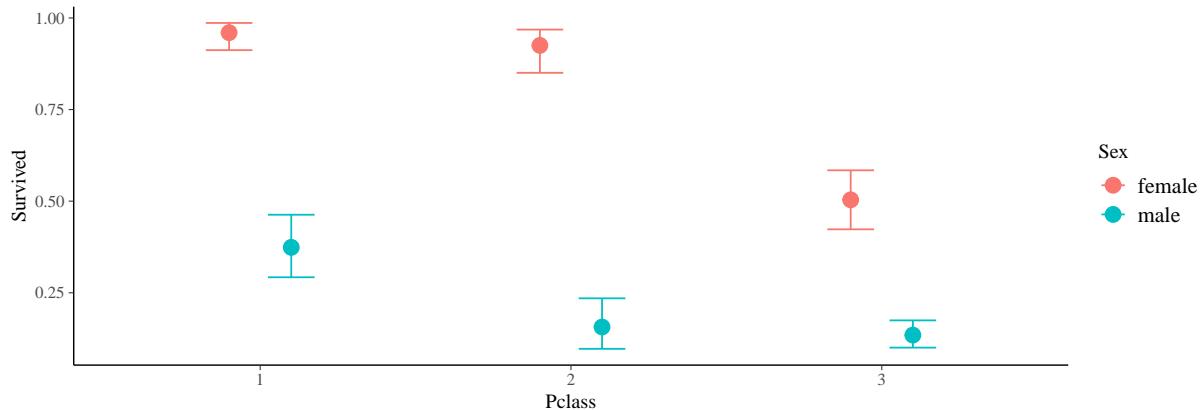


Figure 3.19: Conditional effects of the `Pclass:Sex` interaction for model `fit_titanic2`.

In Figure 3.19, we see clearly that the difference between women's and men's survival is much smaller for the third class passengers than for first and second class passengers. Apparently, the policy of "women first" didn't work out so well in the third class. Does adding the interaction term improve model fit? Let's first look at some posterior predictive checks:

```
pp_check(fit_titanic2, type = "error_binned", ndraws = 10)
```

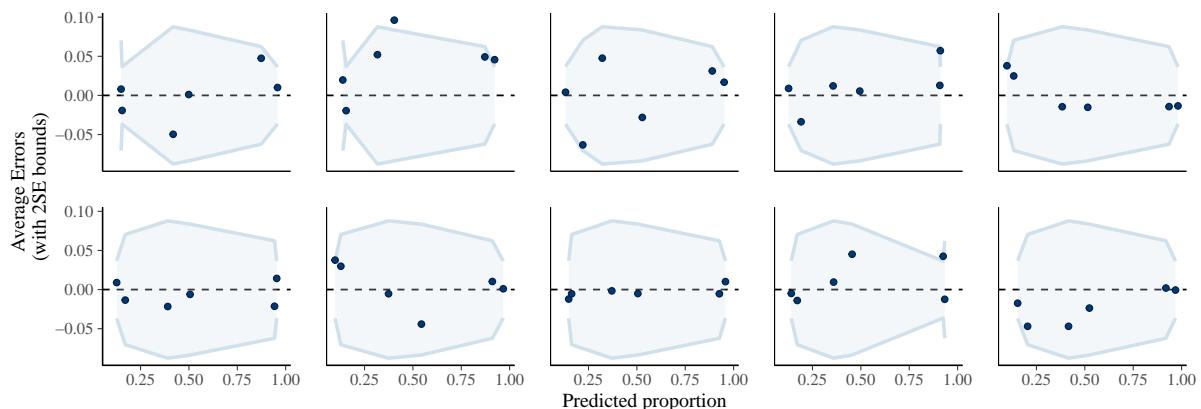


Figure 3.20: Posterior predictive check of binned errors for model `fit_titanic2`.

The result shown in Figure 3.20 clearly looks much better than for `fit_titanic1`: The residuals are smaller (mostly below 5%) and the posterior uncertainty seems to be much better calibrated. But in these plots, we still look at in-sample fit, which is prone to overfitting as we know. For a more formal comparison, let's compare the LOO-CV performance:

```
fit_titanic2 <- add_criterion(fit_titanic2, "loo")
loo(fit_titanic2)
```

Computed from 4000 by 891 log-likelihood matrix.

	Estimate	SE
elpd_loo	-405.0	17.0
p_loo	5.8	0.6
looic	810.0	34.0
-----		
MCSE of elpd_loo	is 0.1.	
MCSE and ESS	estimates assume MCMC draws (r_eff in [0.2, 1.2]).	
All Pareto k estimates	are good (k < 0.7).	
See help('pareto-k-diagnostic')	for details.	

```
loo_compare(fit_titanic1, fit_titanic2)
```

	elpd_diff	se_diff
fit_titanic2	0.0	0.0
fit_titanic1	-12.3	5.1

The `loo_compare` results clearly point to an improved out-of-sample predictions of `fit_titanic2` suggesting that the interaction between class and sex is indeed an important aspect of the data to be modeled.

### 3.4.2 Centering Predictors

Interaction terms are notoriously hard to interpret, as we have to consider them in the context of the main effects terms of the involved variables. In model `fit_titanic2`, we had to consider which reference category was used for both `Pclass` and `Sex` as the interpretation of the interaction coefficients rested on the comparisons with these reference categories. This get awkward pretty quickly as more predictors and their interactions enter the model as we will see in a bit. Let's already come up with a solution now. We know that, for categorical predictors, we will always have to compare against *some* baseline, the question is just which value we choose as that baseline. In the default coding in R, this will be the first category of each predictor, something that we call treatment coding, also known as dummy coding. But what if we would use the mean of all categories instead? The coding that leads to coefficients interpretable this way is called sum coding, also known as effect coding. At least for me, sum

coding creates a more intuitive baseline compared to treatment coding, especially once we enter the realm of interaction effects.

Let us look into the idea of sum coding in detail. Below, we will denote the mean of group  $i$  as  $\mu_i$ , the overall mean (i.e., the mean of group means) as  $\mu$  and the difference between  $\mu_i$  and  $\mu$  as  $\delta_i$ . First, we write the contrast coefficients ( $\mu$  and  $\delta_i$ ) as a function of the group means  $\mu_i$ :

$$\begin{aligned}\mu &= \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 \\ \delta_2 &= \mu_2 - \mu = -\frac{1}{3}\mu_1 + \frac{2}{3}\mu_2 - \frac{1}{3}\mu_3 \\ \delta_3 &= \mu_3 - \mu = -\frac{1}{3}\mu_1 - \frac{1}{3}\mu_2 + \frac{2}{3}\mu_3\end{aligned}$$

Conversely, we can also write the group means as a function of contrast coefficients:

$$\begin{aligned}\mu_1 &= 1\mu - 1\delta_2 - 1\delta_3 \\ \mu_2 &= 1\mu + 1\delta_2 \\ \mu_3 &= 1\mu + 1\delta_3\end{aligned}$$

If the contrast coefficients are set up as in the first set of equations, we can compute the group means from the contrast coefficients as per the second set of equations. This is a one-to-one relation, so we can do this process in the other direction too. That is, if we *want* the contrast coefficients to be those in first set of equations, we have to set up the contrast variables as per the second set of equations. We can set this up in R as follows:

```
contrast_matrix <- rbind(
  c(1/3, 1/3, 1/3),
  c(-1/3, 2/3, -1/3),
  c(-1/3, -1/3, 2/3)
)
rownames(contrast_matrix) <- c("mu", "delta2", "delta3")
colnames(contrast_matrix) <- c("mu1", "mu2", "mu3")
```

Table 3.5: Contrast matrix for sum coding of three groups.

	mu1	mu2	mu3
mu	1/3	1/3	1/3

Table 3.5: Contrast matrix for sum coding of three groups.

	mu1	mu2	mu3
delta2	-1/3	2/3	-1/3
delta3	-1/3	-1/3	2/3

```
coding_matrix <- solve(contrast_matrix)
```

Table 3.6: Coding matrix for sum coding of three groups.

	mu	delta2	delta3
mu1	1	-1	-1
mu2	1	1	0
mu3	1	0	1

The above exemplified approach works regardless of the specifically desired contrasts: (1) Set up the contrast equations, that is, the desired contrasts defined as linear functions of the group means. (2) Invert the implied system of linear functions via inversion of the coefficient matrix. (3) Use the inverted matrix to obtain the dummy variables from the original grouping variable. For a most of the common contrast codings, including treatment and sum coding, R provides built-in implementations already, so for them we do not have to make Step (1) and (2) ourselves. What is more, step (3) will be done automatically by R’s system to create design matrices from formulas (via the `model.matrix` function), a system that brms also builds upon. That is, all we have to do ourselves it to tell R which contrast coding to apply. The built-in approach works via the `contr.*` assignment functions. For example, we could apply sum coding to `Pclass` via:

```
contrasts(titanic$Pclass) <- contr.sum(3)
```

However, this approach is a bit awkward for the following reasons. First, this syntax isn’t nicely integratable into the tidyverse approach to data wrangling to the `contrasts` function appearing on the left-hand side of the assignment. Second, the not shown category will be the last one for sum coding, but the first one for treatment coding, creating confusing inconsistencies when switching the codings. Third, the dummy variables do not have names, so the resulting dummy variable names will be rather uninformative. For these reasons, I always use my own implementation of sum coding, which nicely integrates with the tidyverse, uses the first category as “reference”, and creates sensible dummy variables names:

```

# sum (effect) coding in a less awkward way
# codes the first category with -1
# x: categorical vector to be sum coded
# lvls: optional factor levels of x
sum_coding <- function(x, lvls = levels(as.factor(x))) {
  nlevels <- length(lvls)
  stopifnot(nlevels > 1)
  cont <- diag(nlevels)[, -nlevels, drop = FALSE]
  cont[nlevels, ] <- -1
  cont <- cont[c(nlevels, 1:(nlevels - 1)), , drop = FALSE]
  colnames(cont) <- lvls[-1]
  x <- factor(x, levels = lvls)
  contrasts(x) <- cont
  x
}

```

Since the built-in treatment contrast matrix is set up in a nicer way already, my corresponding function for tidyverse integration looks much simpler than that for sum coding:

```

# treatment (dummy) coding in a less awkward way
# codes the first category with 0
# x: vector to be sum coded
# lvls: optional factor levels of x
treatment_coding <- function(x, lvls = levels(as.factor(x))) {
  x <- factor(x, levels = lvls)
  contrasts(x) <- contr.treatment(levels(x))
  x
}

```

With these functions in hand, it is easy to change apply the desired coding schemes to our categorical predictors within a tidyverse workflow:

```

titanic <- titanic %>%
  mutate(
    Pclass = sum_coding(Pclass),
    Sex = sum_coding(Sex)
  )

```

Let's fit the interaction model again:

```

fit_titanic3 <- brm(
  Survived ~ Pclass * Sex,
  data = titanic, family = bernoulli("logit"),
  prior = prior(normal(0, 3), class = "b")
)

```

As you can see, the `brm` code for `fit_titanic3` is exactly the same as for `fit_titanic2`. The models only differ in terms of the predictor coding, which is implicitly contained only in the `titanic` data frame object.

```
summary(fit_titanic3)
```

```

Family: bernoulli
Links: mu = logit
Formula: Survived ~ Pclass * Sex
Data: titanic (Number of observations: 891)
Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
      total post-warmup draws = 4000

```

#### Regression Coefficients:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	0.33	0.14	0.07	0.62	1.00	1.00	1579	2256	
Pclass2	0.07	0.21	-0.34	0.47	1.00	1.00	2688	2455	
Pclass3	-1.26	0.16	-1.57	-0.96	1.00	1.00	1855	2704	
Sexmale	-1.70	0.14	-1.99	-1.43	1.00	1.00	1502	2164	
Pclass2:Sexmale	-0.41	0.20	-0.80	-0.02	1.00	1.00	2633	2773	
Pclass3:Sexmale	0.77	0.16	0.47	1.09	1.00	1.00	1695	2218	

Draws were sampled using `sampling(NUTS)`. For each parameter, `Bulk_ESS` and `Tail_ESS` are effective sample size measures, and `Rhat` is the potential scale reduction factor on split chains (at convergence, `Rhat = 1`).

What are the coefficients' posteriors telling us? In terms of the main effects of class, we see that 3rd class has a substantially smaller survival probability than the grant mean across the classes, while the 2nd class is pretty much right on the grant mean. We don't see the coefficient of the 1st class directly in the output but we will compute it in a bit. Of course, interpretation of the coefficients' size remains difficult because we are working on the logit scale. Investigation of the main effect of sex, tells use that men have substantially smaller survival probability than the grant mean across men and women. Notice that the coefficient is about half of the one we found in `fit_titanic1`, that is, the model with treatment coding and no interactions. This is no coincidence. Since the grant mean is here just the mean of two groups, the doubled

difference between each of the groups and the grant mean is equal to the difference between the two groups. Lastly, in terms of the interactions, we see a negative interaction effect of 2nd class males and a positive interaction effects of 3rd class males. If we view this in relation to the strongly negative main effect of males, this means that the difference in survival probability between males and females becomes even stronger than average in the 2nd class and less strong (but still substantial) in the 3rd class.

Sum coding has the advantages that the intercept always represents the grant mean over groups (more precisely: the mean over group means on the link scale), and thus the predictor-specific coefficients can be interpreted as deviations of the groups from the grant mean. However, we still had to “sacrifice” one level whose coefficient does not immediately show up in the summary. In `fit_titanic3`, for example, we got the coefficients for `Pclass = 2` and `Pclass = 3` (which we called  $\delta_2$  and  $\delta_3$ ) but not for `Pclass = 1` (which we will now call  $\delta_1$ ). But we can still compute  $\delta_1$  from the model after the fact. By construction of sum coding, the  $\delta$  coefficients need to sum to zero:  $\delta_1 + \delta_2 + \delta_3 = 0$ , otherwise the intercept wouldn’t be the grant mean. Accordingly, we have  $\delta_1 = -\delta_2 - \delta_3$  or, formulated as a per-draw operation  $\delta_1^{(s)} = -\delta_2^{(s)} - \delta_3^{(s)}$ . That is, we could extract the posterior draws of  $\delta_2$  and  $\delta_3$  from the model and do the transformation ourselves. Alternatively, we can just use a `brms` method particularly built for such tasks:

```
hypothesis(fit_titanic3, "- Pclass2 - Pclass3 = 0")
```

Hypothesis Tests for class b:

	Hypothesis	Estimate	Est.Error	CI.Lower	CI.Upper	Evid.Ratio	Post.Prob	Star
1	(-Pclass2-Pclass3) = 0	1.19	0.23	0.76	1.69	NA	NA	*

---

'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.

'\*': For one-sided hypotheses, the posterior probability exceeds 95%;

for two-sided hypotheses, the value tested against lies outside the 95%-CI.

Posterior probabilities of point hypotheses assume equal prior probabilities.

In the first and only row of the returned table, we see the summary of the posterior of  $\delta_1$ . Clearly,  $\delta_1 > 0$  almost surely showing us again that, in comparison to passengers in lower classes, passengers in the first class had a much higher chance of survival.

This also mean that if we change our mind and suddenly prefer our treatment coded coefficients, we could compute it from the sum coded coefficients. For example, if the first class was the desired reference category, and we wanted the contrast it with the second and third classes, we would want to compute

$$\begin{aligned}\mu_2 - \mu_1 &= (\mu + \delta_2) - (\mu + \delta_1) = \delta_2 - \delta_1 \\ &= \delta_2 - (-\delta_2 - \delta_3) = 2\delta_2 + \delta_3\end{aligned}$$

and analogously with  $\mu_3 - \mu_1$ . Thus to obtain the treatment coded coefficients via hypothesis, we can write:

```

hyps <- c(
  "2 * Pclass2 + Pclass3 = 0",
  "2 * Pclass3 + Pclass2 = 0"
)
hypothesis(fit_titanic3, hyps)

```

Hypothesis Tests for class b:

	Hypothesis	Estimate	Est.Error	CI.Lower	CI.Upper	Evid.Ratio	Post.Prob	Star
1	(2*Pclass2+Pclass3) = 0	-1.11	0.41	-1.98	-0.34	NA	NA	*
2	(2*Pclass3+Pclass2) = 0	-2.44	0.34	-3.16	-1.83	NA	NA	*

---

'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.  
'\*': For one-sided hypotheses, the posterior probability exceeds 95%;  
for two-sided hypotheses, the value tested against lies outside the 95%-CI.  
Posterior probabilities of point hypotheses assume equal prior probabilities.

These are still not the coefficients that we got in `fit_titanic2` because there, both predictors (class and sex) were treatment coded while here, we still (implicitly) used the sum coding for sex. We could go ahead and transform the complete set of sum coded coefficients to the complete set of treatment coded coefficients (or vice versa). It's just a bit more math and I leave this for you as an exercise, if you like. What I am trying to say here is that the different coding are *equivalent* in the sense that we can transform their set of coefficients in a one-to-one manner. This means we can in principle freely choose our coding used during model fitting and obtain any other coding afterwards by transforming the posterior draws.

Practically, two things stand in the way though. The first is the math itself and it's easy to get the transformation wrong somewhere. So it is usually just safer to use a coding that already contains the contrasts of interest as parameters or at least allows them to be computed easily during the post-processing. Second, and more Bayesian specific, priors break the equivalence of the different codings. That is, if we choose informative priors in one coding (i.e., in one parameterization), then the resulting transformed coefficients of another coding are in general no longer the same as the coefficients we would have obtained by directly going for that other coding. Accordingly, another way to choose our preferred coding is to think about what parameter we can easily specify priors on. Do we understand specific group difference well, or is it easier to a prior judge the deviation of groups from the overall mean? Usually, in my experience, the parameters that I am interested in are also those I will most likely have prior knowledge on (if I have any). As such, both the math and the prior argument would point us in the same direction.

In our example here, we did specify wide priors, for the logit scale at least, so for practical purposes, we can consider the results arising from our different coding as equivalent. In fact, if we ran some model comparison via, say, LOO-CV we would see that the two models with treatment and sum coding respectively would provide the same predictions up to MCMC error.

### 3.4.3 Even More Interactions

Another variable which we expect to affect survival rates is the passengers' age, as children were supposed to be prioritized in the evacuation. So let's include age (in years) and model it directly in interaction with class and sex. For this purpose, it makes sense that transform the `Age` variable to simplify interpretation. First, we will divide age by 100 such that it is roughly on a 0 to 1 scale. This ensures that age-related coefficients on the logit scale do not become too small. As an exercise, fit the model below with the unscaled `Age` variable and check how small the coefficients will become. Second, we will center age around its mean, such that a centered `Age` value of 0 will indicate the mean age in the data. In a way, mean centering is “the sum coding for continuous predictors”, as it allows us to interpret coefficients as deviations from the predictor's mean.

```
titanic <- titanic %>%
  mutate(
    Age100 = Age / 100,
    Age100c = Age100 - mean(Age100, na.rm = TRUE)
  )
```

The model is then specified as follows:

```
fit_titanic4 <- brm(
  Survived ~ Pclass * Sex * Age100c,
  data = titanic, family = bernoulli("logit"),
  prior = prior(normal(0, 3), class = "b")
)
```

The summary table of coefficients shown in Table 3.7 is now quite crowded, with any up to 3-way interaction coefficients:

Table 3.7: Posterior summary for model `fit_titanic4`.

variable	mean	sd	q5	q95
b_Intercept	0.27	0.16	0.00	0.54
b_Pclass2	-0.05	0.23	-0.43	0.33

Table 3.7: Posterior summary for model `fit_titanic4`.

variable	mean	sd	q5	q95
b_Pclass3	-1.45	0.18	-1.76	-1.15
b_Sexmale	-1.65	0.16	-1.92	-1.40
b_Age100c	-3.27	1.04	-4.99	-1.53
b_Pclass2:Sexmale	-0.74	0.24	-1.14	-0.35
b_Pclass3:Sexmale	0.87	0.19	0.56	1.17
b_Pclass2:Age100c	-3.31	1.46	-5.77	-0.97
b_Pclass3:Age100c	-0.16	1.17	-2.12	1.73
b_Sexmale:Age100c	-2.38	1.08	-4.13	-0.65
b_Pclass2:Sexmale:Age100c	-0.14	1.49	-2.59	2.31
b_Pclass3:Sexmale:Age100c	1.88	1.20	-0.04	3.85

Admittedly, even with all variables centered, interpretation is not so easy. And the logit scale certainty doesn't help. With regard to the coefficients of age, the first thing we observe is the strong negative main coefficient, indicating a strong drop in survival rates for older people. This effect is not constant across passengers groups though. In fact, the it appears to be even stronger (i.e., more negative) for men and second class passengers. For third class passengers the effect of age is not very different from the main effect, likely because the survival rates of third class passengers were already generally low to begin with. Another thing that becomes apparent is that age coefficients are quite uncertain with posterior standard deviations of over 1 and large credible intervals. All of these aspects can also be found in the corresponding `conditional_effects` plot (see Figure 3.21):

```
conditional_effects(
  fit_titanic4, effects = "Age100c:Pclass",
  conditions = make_conditions(fit_titanic4, "Sex")
)
```

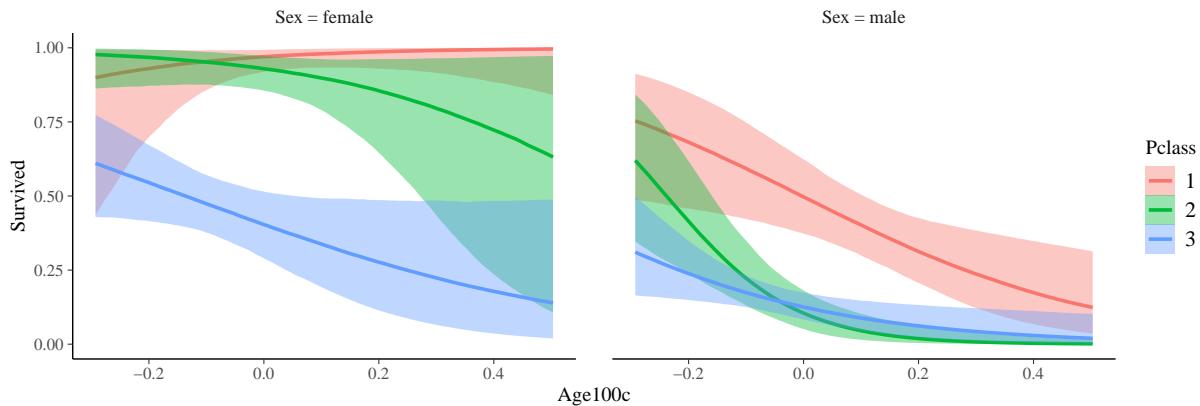


Figure 3.21: Conditional effects of for model `fit_titanic4` using `Sex` for facetting.

The `effects` argument can only handle two predictors at once, the first being shown on the x-axis, the second being shown as separate color coded lines. Accordingly, if we want to illustrate three predictors together, we need a different mechanism, which comes in the form of the `conditions` argument. This is just a data frame and for each of its rows, we will get a separate facet in the plot where we conditioned on the predictor values in the corresponding row. You can set up this data.frame manually or, for additional convenience, use `make_conditions`, which will also give you nice facet labels. The data frame used above looks as follows:

```
make_conditions(fit_titanic4, "Sex")
```

	Sex	cond__
1	female	Sex = female
2	male	Sex = male

Accordingly, it contains all the levels of the `Sex` predictor along with a `cond__` column that will be used as facet labels. If you are specifying your conditions data.frame manually and want to control facet labels, you have to set the `cond__` column yourself. Of course, when we want to use continuous predictors to define the facets, showing all realizations of the predictors is most likely not sensible. Instead, `make_conditions` will create three facets by default, one for the predictor mean and two more of the predictor mean  $\pm$  the predictor standard deviation:

```
make_conditions(fit_titanic4, "Age100c")
```

	Age100c	cond__
1	-1.452650e-01	Age100c = -0.15
2	-2.050084e-17	Age100c = 0
3	1.452650e-01	Age100c = 0.15

This reduces information in the plot but at least ensure that it will not have dozens of facets by default:

```
conditional_effects(
  fit_titanic4, effects = "Sex:Pclass",
  conditions = make_conditions(fit_titanic4, "Age100c")
)
```

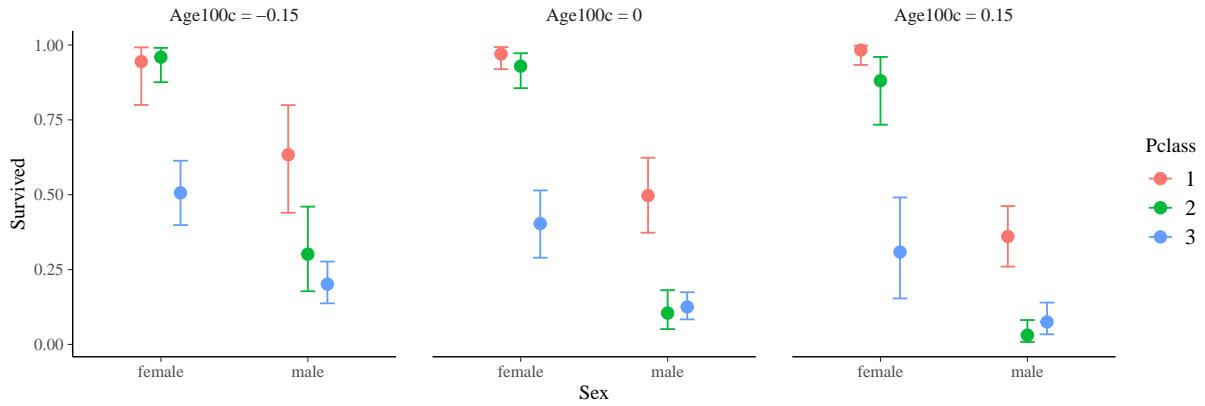


Figure 3.22: Conditional effects of model `fit_titanic4` using `Age` for facetting.

At least to me, Figure 3.22 looks a bit more messy than Figure 3.21 above, but perhaps that's up to taste. My personal recommendation for a mix of continuous and categorical predictors is to show one of the continuous predictors on the x-axis such that we can see the full value range at least for that predictor.

The results above suggest that age is indeed an important variable. But does it also show up in out-of-sample predictive performance? Let's compare the model with and without age via LOO-CV:

```
loo_titanic3 <- loo(fit_titanic3)
loo_titanic4 <- loo(fit_titanic4)
loo_compare(loo_titanic3, loo_titanic4)
```

```
Error : Not all models have the same number of data points.
```

That didn't work apparently. A closer inspection reveals that `Age` has some missing values which lead to a removal of the corresponding rows in the data by default (with a warning). So while the full dataset has 891 observations, `fit_titanic4` was only fit with 714 of them. Comparing models on different test datasets would not be fair, hence `loo` complained. We will

learn about how to properly deal with missing values in `?@sec-missing-values`. For now, we will have to find a workaround: What if we ask the model without age to only predict the observations with non-missing age values? We first create the corresponding data subset

```
titanic_sub <- titanic %>%
  filter(!is.na(Age))
```

and then feed in this dataset into the `newdata` argument of `loo`:

```
loo_titanic3_sub <- loo(fit_titanic3, newdata = titanic_sub)
loo_compare(loo_titanic3_sub, loo_titanic4)
```

	elpd_diff	se_diff
fit_titanic4	0.0	0.0
fit_titanic3	-13.6	7.5

Now the comparison seems to work and suggests that indeed including age along with its two-way and three-way interactions with class and sex improve predictive performance.

But was this a fair comparison now? Well, it depends on the perspective. If we are attempting to win a Kaggle challenge, and want to choose which of the two models we are submitting as our final model, then this comparison is fair. We would just care about the (out-of-sample) predictive performance and ignoring a subset of the data due to missing values is then “the model’s problem”. Put differently, we do not care about the probabilistic model itself, but only about its test data predictions based on whatever training data it could use during learning.

But there is also a second perspective. What if we care about the epistemic virtue of age as predictor in the titanic accident? That is, what if we want to judge whether the policy of rescuing young people first was actually implemented? In other words, we would care about the true relevance of age. Our two models would then be representatives of the two hypothesis “age doesn’t matter” and “age does matter”, respectively. In this case, we really care about the inner workings of the model and the role age plays in it; even if we evaluate the model’s performance based on out-of-sample predictions.

In the second perspective, the above model comparison would be biased against the model including age, as it had less training data to learn from and hence smaller potential to predict well just for this reason. Accordingly, to address this epistemic question fairly, we would want to fit both models on the same training data, that is, would need to refit `fit_titanic3` on the `titanic_sub` dataset before running `loo`. I leave this to you as an exercise.

I have to add that there is an ongoing debate on whether cross-validation is at all a valid procedure for the discussed second perspective, something we already touched upon in Section 2.3.

Summarizing the results so far, we may conclude something along the following: Young and/or female people had a substantially higher probability of survival than older men. Unless you were a third class passenger. There, the only people with with a decent chance of survival were young women and girls.

### 3.4.4 Changing the Link Function

The link function is an essential part also of GLMs for binary data, and more generally for all double bounded responses. brms offers a wide range of options, comprising the most common link functions, including logit (default), probit, cloglog, and cauchit, an illustration of which you can find in Figure 3.23.

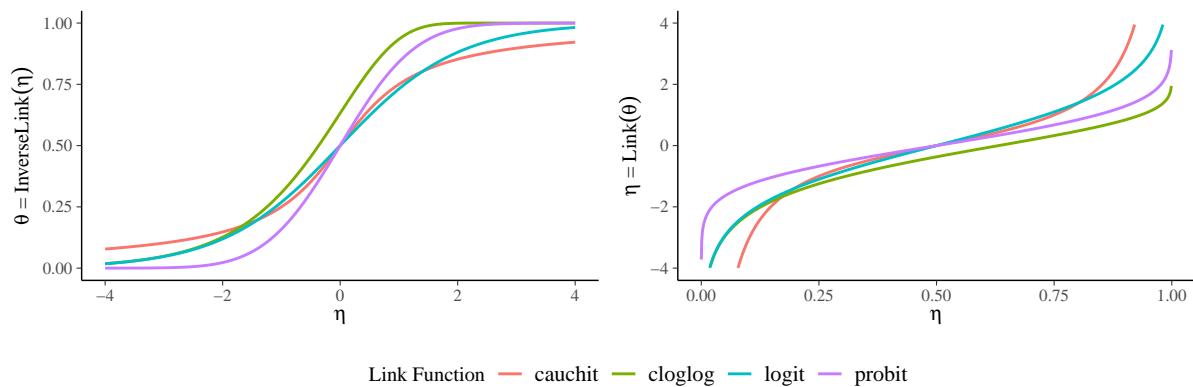


Figure 3.23: Illustration of several common links and corresponding response (inverse link) functions.

At the start of this chapter, I have motivated link functions primarily through pure necessity in order to ensure that predictions remain in the valid (location) parameter range. But for these double-bounded link function applied to binary data – and their generalization – we have access to another interpretation that goes as follows: Assume there is a unbounded latent (i.e., unobservable) variable that gives rise to the binary response through a thresholding process. If the latent variable is smaller than or equal to the threshold, the binary response will be 0 (“no event”). If the latent variable is larger than the threshold, the binary response will be 1 (“event”). You can find an illustration of this idea in Figure 3.24. Our linear predictor can be thought of as the mean of the latent variable and our threshold can be thought of a being fixed to 0 on the latent scale. As the linear predictor increases, we are more likely to exceed the threshold. Of course, this latent variable is not deterministic but has a distributions, which is implicitly given by the link function. In fact, the response function of a double bounded link is nothing else than the cumulative distribution function (CDF) of the latent variable’s distribution.

The assumption of such a latent process variable is not necessary to justify utilizing these links, but it may at least help with intuition. Sometimes, we can actually argue for the existence

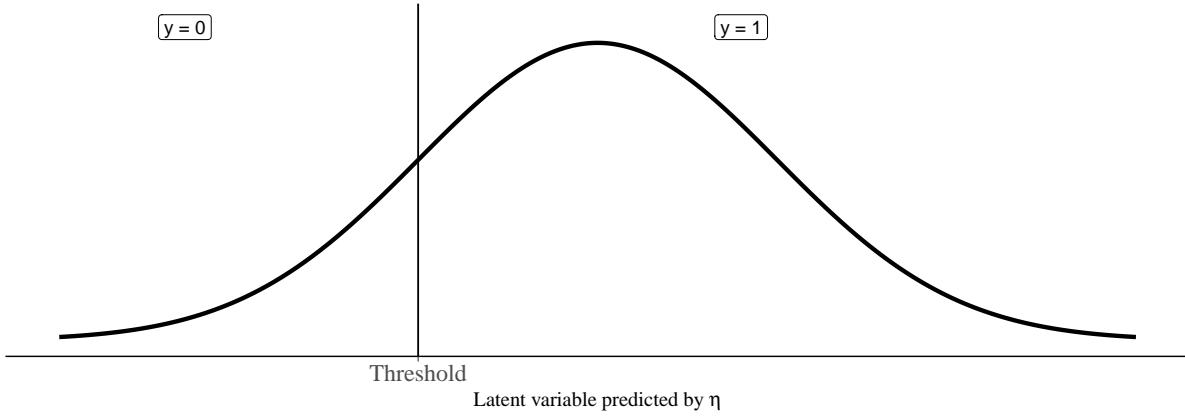


Figure 3.24: Illustration of the thresholding procedure implied by link functions for models of binary data.

of such a latent variable, for example, in some psychological applications: Suppose we ask an individual whether or not they like a particular book, giving them only the option of “I like” or “I don’t like” the book. The individual will certainly be able to give a much more nuanced view of the book, all of which will contribute to how much they like the book overall. This “overall liking” may be well understood as an unbounded continuous variable existing in the individual’s mind. Of course, we cannot observe this variable directly and our access to it is only indirect by asking the “do you like it or not” question. Accordingly, this question induces a thresholding process in the individual’s mind, forcing them to reduce their nuanced opinion to a binary response. We will revisit and extend this perspective in the context of ordinal modeling in [?@sec-ordinal-models](#).

Back to our running example, the survival of passengers on the Titanic may or may not “truly” have an underlying latent variable, but it doesn’t matter. We can freely choose our link function in any case. For example, we can use the `cloglog` link, which implies an asymmetric response function with a fatter right tail (see Figure 3.23). That is, even if the odds of survival are stacked against an individual, say, because they are an old man in the 3rd class, they may still be some non-negligible chance of surviving due to random lucky factors. At least a higher chance than a symmetric link function would assign to this individual under the same circumstances.

```
fit_titanic5 <- brm(
  Survived ~ Pclass * Sex * Age100c,
  data = titanic, family = bernoulli("cloglog"),
  prior = prior(normal(0, 3), class = "b")
)
```

Although sampling worked well, a closer inspection of the chains reveals a bunch of warnings that all look like this:

```

Chain 1: Rejecting initial value:
Chain 1: Log probability evaluates to log(0), i.e. negative infinity.
Chain 1: Stan can't start sampling from this initial value

```

This indicates the the chain had trouble starting at the randomly chosen initial values. Eventually the chains started so we don't need to worry. But sometimes, finding good random initial values is so difficult that Stan terminates after (by default) 100 attempts. In fact, this is something we actually see quite often in cloglog models. For example, if we used the unscaled and uncentered age in the formula, i.e., `Survived ~ Pclass * Sex * Age` an reran `fit_titanic5` with it, we would see exactly this problem. In addition or alternative to centering and scaling predictors, you can also change the range where initial values are drawn. If you are using the `rstan` backend to `brms` (the default), this can be done via the `init_r` argument. Setting it to values smaller than the default of 2, say, to 0.1 or so will enforce initial values to be much closer to zero on the unconstrained parameter scale, which often makes it easier to find good initial values. If you are using the `cmdstanr` backend, you can use the `init` argument for the same purpose. In both backends `init = 0` will initialize all parameters at zero on the unconstrained scale. This is not ideal as it implies the same initial values for all chains, which may hamper thorough posterior exploration, but I have often had practical success with this option if all of the previously described options failed.

Back to our actual cloglog model, let's check out the summary output:

```
summary(fit_titanic5)
```

```

Family: bernoulli
Links: mu = cloglog
Formula: Survived ~ Pclass * Sex * Age100c
Data: titanic (Number of observations: 714)

```

#### Regression Coefficients:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-0.54	0.09	-0.72	-0.37	1.00	2483	1788
Pclass2	-0.12	0.14	-0.41	0.14	1.00	2098	2043
Pclass3	-0.81	0.12	-1.04	-0.58	1.00	2561	2293
Sexmale	-1.03	0.09	-1.21	-0.86	1.00	2508	2692
Age100c	-3.16	0.61	-4.38	-1.97	1.00	2391	2597
Pclass2:Sexmale	-0.59	0.15	-0.90	-0.32	1.00	2029	2217
Pclass3:Sexmale	0.36	0.12	0.13	0.60	1.00	2766	3036
Pclass2:Age100c	-2.57	0.93	-4.44	-0.76	1.00	2033	2837
Pclass3:Age100c	0.27	0.79	-1.26	1.83	1.00	2226	2984
Sexmale:Age100c	-2.18	0.60	-3.38	-1.03	1.00	2628	2798
Pclass2:Sexmale:Age100c	-1.45	0.95	-3.34	0.42	1.00	2436	2554
Pclass3:Sexmale:Age100c	1.45	0.79	-0.08	2.96	1.00	2636	2865

Compared to the logit model, the coefficients' posteriors of the cloglog model are a bit smaller (closer to zero): That does not mean that the cloglog model estimates relationships are actually smaller. It is at least partially an artifact of the latent scale differences induced by the different link function: The SD of the latent variable implied by the logit model is equal to the SD of the standard logistic distribution, which is  $\pi/\sqrt{3} \approx 1.81$ . In contrast, the cloglog link implies a standard Gumbel distribution (aka. Type-1 generalized extreme value distribution) on the latent scale with an SD of  $\pi/\sqrt{6} \approx 1.28$ . The latent scale is not identified by the binary response data, so the SD difference is essentially arbitrary; only the distribution's shape matters.

In an attempt to achieve more comparable estimates, we can take the posterior of the coefficients and divide it by their posterior SD, as exemplified below:

```
post_Sexmale4 <- as.data.frame(fit_titanic4)[["b_Sexmale"]]
post_Sexmale5 <- as.data.frame(fit_titanic5)[["b_Sexmale"]]
post_Sexmale4_scaled <- post_Sexmale4 / sd(post_Sexmale4)
post_Sexmale5_scaled <- post_Sexmale5 / sd(post_Sexmale5)
```

```
posterior_summary(cbind(post_Sexmale4_scaled, post_Sexmale5_scaled))
```

Table 3.8: Posterior summary of regression coefficients scaled by their posterior standard deviations for models `fit_titanic4` and `fit_titanic5`.

	Estimate	Est.Error	Q2.5	Q97.5
post_Sexmale4_scaled	-10.23	1	-12.27	-8.40
post_Sexmale5_scaled	-11.37	1	-13.36	-9.47

As we can see in Table 3.8, there are in fact differences in the coefficients' posterior not attributable to the scale differences. For the above example the difference is quite small, but for other coefficients, it actually is much bigger. As an exercise, repeat the above code for some of the interaction coefficients (e.g., `b_Pclass3:Sexmale`). That said, in terms of predictions, the change of link function barely seems to matter. Indeed, if we run `pp_check`, `conditional_effects` or `loo`, we see that the models perform very similarly. I don't show this here and leave it to you as an exercise.

### 3.4.5 Binomial Models

So far, the individual responses were an indicator (0 or 1) whether a person survived the titanic accident. But often, double bounded count data are not just 0 or 1 but  $0, 1, 2, \dots, M$  for a given, an (usually) a priori known  $M$ . In fact, we can transform our present data to be in this format. Suppose we are just interested in the predictors `Pclass` and `Sex` as well as

their interactions. Then, it is sufficient to just count how many people in a group (say, women in the first class) survived out of all people in that group. To make this more concrete, let's restructure the data this group format:

```
titanic_group <- titanic %>%
  group_by(Pclass, Sex) %>%
  summarise(Survived = sum(Survived), M = n())
```

Table 3.9: The titanic data grouped after passenger class and sex.

Pclass	Sex	Survived	M
1	female	91	94
1	male	45	122
2	female	70	76
2	male	17	108
3	female	72	144
3	male	47	347

As we will see, the data set shown in Table 3.9 and formally containing just 6(!) rows is sufficient to estimate the main effects and interactions of class and sex, which together implies 6 regression coefficients. Since we assumed each individual survival indicator to be Bernoulli distributed with survival probability  $\theta$ , it follows that the sum of  $M$  such indicators is binomial distributed with the same  $\theta$  and number of trials  $M$  (here number of people in a group). The term “trials” comes from situations where we are repeating an experiment independently multiple times, say, we throw the same coin  $M$  times and count how often the coin lands on head. Of course, in our titanic example the indicator we are measuring is not really a “trial” but an individual human. However, the statistical model neither knows nor cares whether the data we feed it with are coin throws or indicators of individual humans’ survival. We as data analysts are the ones who need to care.

Let's fit our binomial model to the grouped data:

```
fit_titanic_binom1 <- brm(Survived | trials(M) ~ Pclass * Sex,
                           data = titanic_group, family = binomial("logit"))
```

Sampling works well and fast, with excellent convergence. Apparently, the model did not mind that the number of “observations” (rows) in the data is the same as the number of regression coefficients. If we tried the same on 6 rows of Bernoulli (binary) data, the model will fail miserably because the amount of information contained there is way smaller than the amount of information contained in a data point comprising the observations of all individuals in a group.

When investigating the summary output of the binomial model, we see that its results are the same as the corresponding Bernoulli model `fit_titanic1`, up to MCMC error. This is no coincidence. We can either specify our binomial model on the level of the individual “trials” or on the level of groups. But sometimes grouping isn’t possible, at least now without loss of information. For example, if we wanted to include `Age` as additional predictor in the binomial model, without loss of information, we would need to create groups per year of age. The resulting reduction in data rows would not be that big. We can see this by computing

```
titanic_group2 <- titanic %>%
  group_by(Pclass, Sex, Age) %>%
  summarise(Survived = sum(Survived), M = n())
```

where we find `titanic_group2` to have `nrow(titanic_group2)` rows, compared to `nrow(titanic)` rows in the original dataset. Of course, we could choose to compress age into groups, say, from 0 to 10, 11 to 20, etc. but this would result in a loss of information in the age predictor that we should seek to avoid. As an exercise, you may want to fit binomial model to the clustered age data (each cluster having an age range of 5, 10, or 20 years) and see how the posterior changes in comparison to the models without such clustering.

## 3.5 Summary

We have seen that GLMs can be easily specified in `brms` by changing the `family` argument, while keeping the model formula the same – or only adding one term as in binomial models. The likelihood families we have used in the examples are only a small subset of all the families implemented in `brms`. We will see many of them in use in the following chapters. If you want to get an overview right now, check of the `?brmsfamily` help page. Regardless of the GLM you choose, the principles that we learned here stay the same: We are changing the family argument, and perhaps the link function if we don’t want to use the default. And then we are ready to go. The price we pay is a more challenging interpretation of the regression coefficients, but we can help with that with some nice and quick predictive visualizations. In many of the upcoming chapters, we will use GLMs as the basis for more complicated models.

# 4 Linear multilevel models

## 4.1 Setup

```
library(dplyr)
library(ggplot2)
library(patchwork)
library(bayesplot)
library(brms)
```

## 4.2 Introduction

Most real-world datasets are not just a collection of independent observations measured cleanly in an experimental setting with only one observation per individual. Instead, we are more often than not confronted with complex dependency structures, for example, because of natural groupings of individuals or repeated measurement of the same individuals. Here, “individuals” can mean many different things, for example, humans, animals, cells, classes, laboratories, studies, etc. In a longitudinal study in psychology, for example, the individuals will be humans measured multiple times over the course of the study. Naturally, the observations belonging to the same individual are highly likely to be dependent on each other, as they share the same source, namely said individual, even if they had been gathered under different experimental conditions. Also, some individuals are often more similar than others, as the result of groups occurring in the data. For example, students in the same class, are highly likely dependent as they share the same class rooms, peers, teachers etc. We can of course choose to ignore these dependencies in our modeling, but then pay the price in the form of overconfident and potentially biased inference, as we will see later on in this chapter. Multilevel models, on the other hand, are built to express such dependencies explicitly, thus helping to obtain trustworthy inference from complex-structured (multilevel) datasets. Due to the frequency with which datasets contain multilevel structure, I consider multilevel models as the *default* model class for statistical modeling, Bayesian or otherwise.

As a running example, we will use the `sleepstudy` dataset originally published by Belenky et al. (2003), and available in R from the `lme4` package (Bates et al. 2015). It contains the average reaction times per day across 10 consecutive days for a selection 18 subjects in a

sleep deprivation study. After the first two days of adaptation and training, sleep deprivation started after day 2, which meant that subjects were allowed to sleep for only a maximum of 3 hours per night. If you have ever slept this little for several nights in a row (I definitely have), you will know that this messes with your concentration so a lot of tasks become slower and require more concentration. It is not hard to hypothesize that several days of consecutive sleep deprivation will lead to slower reaction times overall, the details of which we will analyse below. Let's load the data first and get an initial overview (see Table 4.1):

```
data("sleepstudy", package = "lme4")
```

```
head(sleepstudy)
```

Table 4.1: First few observations of the `sleepstudy` data.

Reaction	Days	Subject
249.56	0	308
258.70	1	308
250.80	2	308
321.44	3	308
356.85	4	308
414.69	5	308

The datasets is very simple in structure, containing only 3 variables: `Reaction` contains the average reaction time (in milliseconds) for a specific `Day` (0 to 9) and a specific `Subject`. In total, we have access to 180 observations, 10 day for each of 18 subjects. Clearly, it contains multilevel structure in the form of repeated measurements of the same subjects. In other words, individual observations are nested within subjects. Let's investigate the data graphically:

```
sleepstudy %>%
  ggplot(aes(Days, Reaction)) +
  geom_point() +
  facet_wrap("Subject", nrow = 3) +
  scale_x_continuous(breaks = c(0, 3, 6, 9))
```

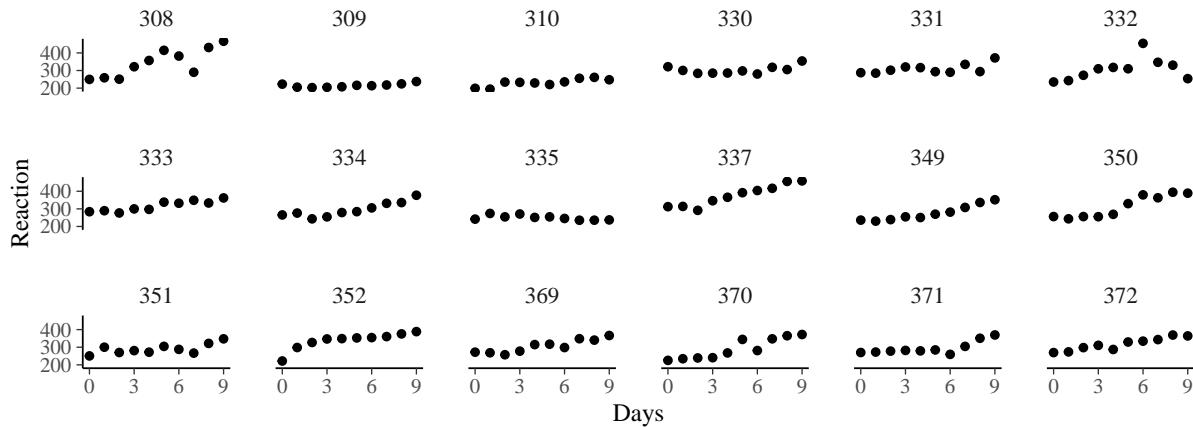


Figure 4.1: Scatter plots of the `sleepstudy` data split by subject.

When looking at Figure 4.1, we see that we may not need any modeling to understand that the subjects' reaction times increase over the days of consecutive sleep deprivation. Moreover, we see that the performance varies quite drastically between subjects, both in the initial baseline reaction time at Day 0 and in how much reaction times increase over the days. What we don't see immediately are the results averaged across subjects: In many real-world analysis scenarios, we are not interested in the specific behavior of specific subjects, but rather in what we learn over the population of individuals more generally. For example, we might ask about the average reaction time at baseline as well as the average change of reaction time over the days. In other words, we want to learn something *general* (average) about how humans react to sleep deprivation. What is more, we would like to understand how much individuals tend to *vary* around those average results. And on top of that, we would also like to obtain trustworthy *uncertainty estimates* of both the average results and their variations across the population of individuals. Bayesian multilevel models can provide us with answers to all of these three questions. How good those answers are depends, as always, on how well our modeling assumptions align with the data being analysed.

Before we start modeling, we will remove the first two days containing data from the adaptation and training phases. Subsequently, we will shift the `Days` variable such that the baseline (Day 2) will be coded as 0:

```
sleepstudy <- sleepstudy |>
  filter(Days >= 2) |>
  mutate(Days = Days - 2)
```

## 4.3 Complete pooling

As our first model, we will – for the moment – throw all of these considerations of multilevel structures out of the window and just fit a basic linear models, with just an average intercept and slope:

$$y_n \sim \text{normal}(b_0 + b_1 x_n, \sigma)$$

We call this a *complete pooling* model as we ignore potential variations across individuals and thus completely pool all their data together to only from a single set of average coefficients; as if all individuals had the same intercept and the same slope.

```
fit_sleep1 <- brm(Reaction ~ 1 + Days, data = sleepstudy)
```

```
summary(fit_sleep1)
```

```
Family: gaussian  
Links: mu = identity  
Formula: Reaction ~ 1 + Days  
Data: sleepstudy (Number of observations: 144)  
  
Regression Coefficients:  
Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
Intercept 267.95     7.53   253.00   282.39 1.00    4062    3283  
Days       11.44     1.79    7.98    14.95 1.00    3770    3118
```

```
Further Distributional Parameters:  
Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
```

```
sigma     51.06     3.00   45.70   57.64 1.00    4193    2446
```

We learn from the results that, on average, the model estimates a baseline of around 268 milliseconds and a slope of about 11.42 additional milliseconds per day, with a 95% credible interval of 7.6 to 15.3 milliseconds. Visually, the predictions look as follows:

```
plot(conditional_effects(fit_sleep1), points = TRUE)
```

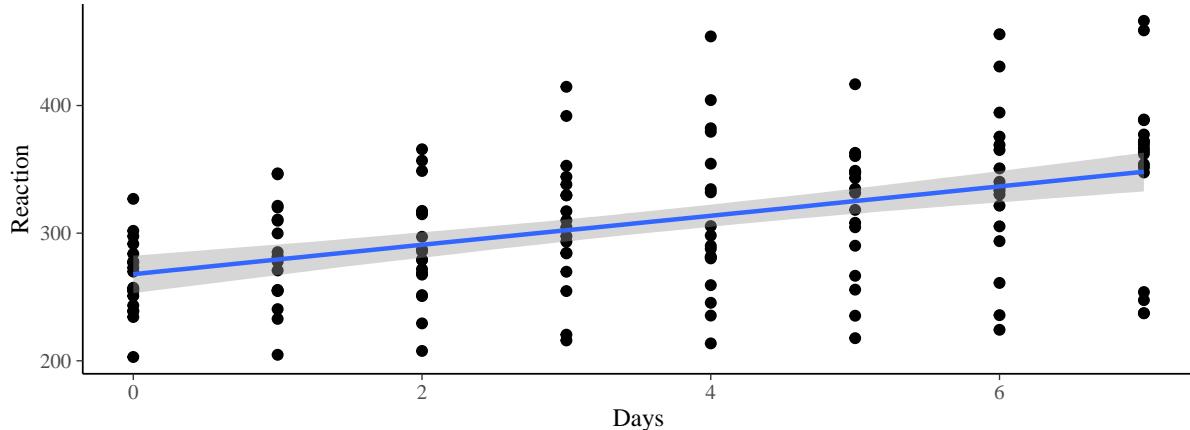


Figure 4.2: Conditional effects for model `fit_sleep1`. Observations are shown as points.

As we see clearly in Figure 4.2, average reaction times increase over the course of the 8 days, and the model is quite confident by how much reaction times are expected to increase. In this simple example, we are using just a single predictor (`Days`), so the comparison of model-based predictions and observations in this plot can be used as sensible indication of model fit. We see that the variances of observations at the early days of the study are much smaller than the variances towards the end of the study. Let's see what more standard posterior predictive plots tell us about the model:

```
pp_check(fit_sleep1, type = "dens_overlay")
```

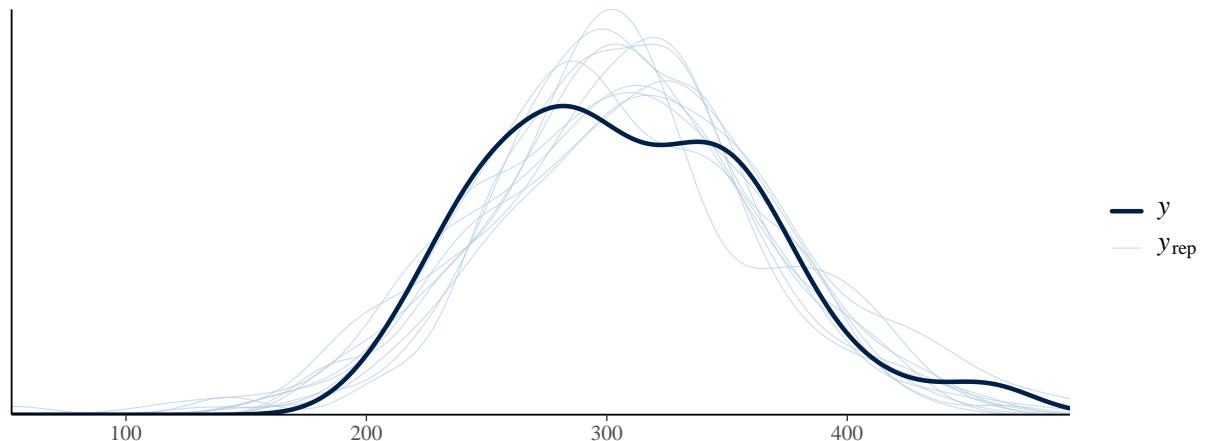


Figure 4.3: Default posterior predictive check for model `fit_sleep1`.

The default predictive check with `type = "dens_overlay"` shown in Figure 4.3 is painting the picture of a supposedly reasonably fitting model. At least there are no huge indications

that the model gets the marginal distribution of the response variable wrong. If you have read the previous chapters carefully, you will already know to not trust this result too much. So what do we learn from a plot of predictive errors?

```
pp_check(fit_sleep1, type = "error_scatter_avg")
```

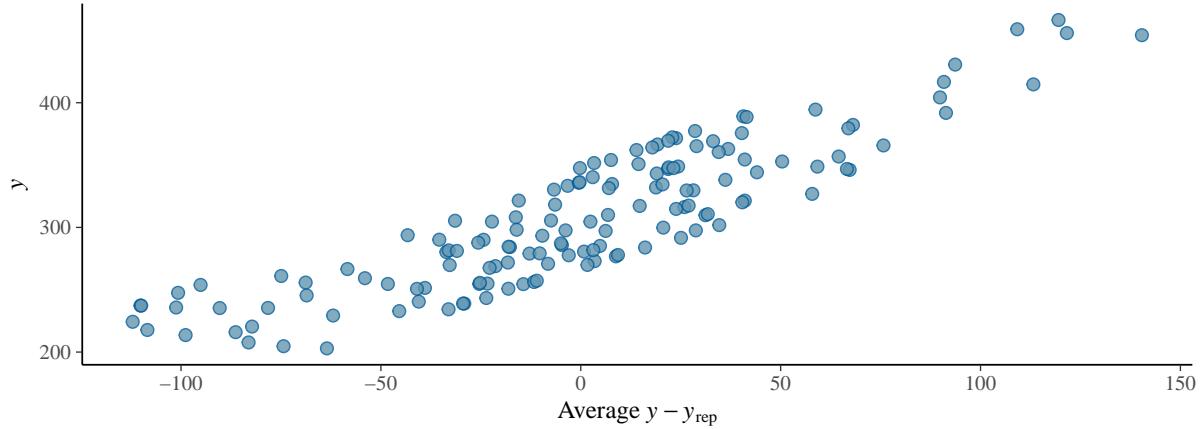


Figure 4.4: Posterior predictive check of residuals for model `fit_sleep1`.

The predictive check shown in Figure 4.4 doesn't look so good anymore. Remember that, for a reasonable fitting model, we would expect to see a big point cloud in such a plot, without any clearly visible patterns. Any deviation from that points to some potential misfit. And indeed, there is a strong linear pattern in this plot. Specifically, large negative errors (lower observed than predicted responses) are associated with small observed responses. Conversely, large positive errors (higher observed than predicted responses) are associated with large observed responses. Accordingly, there is apparently some strong signal in the data that our complete pooling model fails to account for. Based on our discussion at the start of this chapter, you should already have a pretty good hunch on which that is. So let's dive right in and start fitting multilevel models.

## 4.4 Partial pooling: Varying intercepts

As a first step, we will assume the intercept  $b_0$  to vary across subjects, by replacing  $b_0$  with a set of intercepts  $b_{0j}$ , one for each subject  $j$ :

$$y_n \sim \text{normal}(b_{0j[n]} + b_1 x_n, \sigma)$$

I use the notation  $j[n]$  to indicate the subject  $j$  to whom observation  $n$  belongs. This likelihood alone does not yet specify a multilevel model yet. In fact, it is just a linear model with cell

mean coding on the subject factor. The magic begins as soon as we set a special kind of prior on the varying intercepts:

$$b_{0j} \sim \text{normal}(b_0, \sigma_0)$$

If we were to fix the mean  $b_0$  and the standard deviation  $\sigma_0$  to some values, this would just be a regular normal prior that we have seen already in the previous chapters. However, in multilevel models, we will not only assume the  $b_{0j}$  but also  $b_0$  and  $\sigma_0$  to be parameters, estimated from the data. In other words, we have a hierarchy of parameters, where the prior on the lower level parameters  $b_{0j}$  is determined by the high-level (hyper-)parameters  $b_0$  and  $\sigma_0$ . To complete the model, we now also have to set priors on  $b_0$ ,  $\sigma_0$ , and the residual standard deviation  $\sigma$ , but we do not focus on these priors for now. It is really the hierarchical prior above that sets multilevel models apart from standard regression models.

First, let us talk about what assumptions we convey with the hierarchical prior. Since every subject's intercept  $b_{0j}$  has the same prior  $\text{normal}(b_0, \sigma_0)$ , we a priori do not provide any information which of the subjects may have a smaller and which may have a larger intercept. We do of course assume there will be variation between subjects (unless  $\sigma_0 = 0$ ), but this assumption is symmetric as it treats all subjects equally. Does that mean we assume subjects to be *independent*? Not quite, since all subjects share the same hyperparameters. If, for example, the overall intercept  $b_0$  increases, the subjects' individual intercepts will tend to increase too. Thus, speaking more generally, by means of the hierarchical prior with estimated hyperparameters, we induce a positive correlation between the lower-level parameters. So the intercepts are in fact dependent, but in a symmetric manner. We call this assumption *exchangeability*. Mathematically speaking, their joint prior is the same regardless of the subjects' ordering. For example, we could reverse the ordering of the subjects' indices and the hierarchical prior would still be the same:

$$p(b_{01}, b_{02}, \dots, b_{0j-1}, b_{0j}) = p(b_{0j}, b_{0j-1}, \dots, b_{02}, b_{01})$$

This is intuitively clear when we look at the hierarchical prior specification above, but is still crucial to point out. It helps us understand when the assumptions of multilevel models are justified, and when they are not. But what do we actually achieve by means of this prior and its exchangeability assumption?

The key lies in understanding the difference between independence and exchangeability. If we had assumed the intercepts to be mutually independent, as is done by a standard regression model, the intercepts could not share information with one another. In other words, if some intercepts were large that would have no influence on the other intercepts, at least not as conveyed through the prior. Now, with our hierarchical prior, if some intercepts are large, this will increase the overall intercept  $b_0$ , which in turn influences the prior of all intercepts, pushing all intercepts up. Of course, this exchange of information is symmetric across subjects, so each intercept influences all other intercepts via its influence on the hyperparameters. We

can also say it like this: Through the hierarchical prior, all lower-level parameters are shrunk towards their hierarchical mean parameter. This sharing of information is what I call *partial pooling* following Gelman and Hill (2006). The degree of shrinkage induced by partial pooling is influenced (a) by the distance between the individual parameters and the hierarchical mean (higher distance implies higher shrinkage) and (b) the size of the hierarchical variations parameters (larger variation implies smaller shrinkage). In our example, if the intercept of a specific subject is far away from the hierarchical mean intercept  $b_0$ , it will be shrunk more than those subjects' intercepts closer to  $b_0$ . But the degree of that shrinkage will be modulated by the size of the hierarchical standard deviation  $\sigma_0$ .

Is *partial pooling* something we actually want? I would argue that it is, indeed, desirable at least in many cases. Thinking about the sleepstudy example, all subjects are healthy, adult humans so they inevitably share some similarities with one another. For example, if most subjects have an average response time of around, say, 300 ms for the task at hand, it becomes quite unlikely that some subjects will have an average response time of, say, 10 seconds. In other words, by knowing something about some subjects, we also know something about the other subjects. That is not else than what the partial pooling property induced by the hierarchical prior attempts to express.

Before we start fitting multilevel models in brms, we need to make one further change in the model. Nothing that actually changes its meaning. Just a reformulation that makes model specification more compatible with R formula syntax. Instead of directly modeling the  $b_{0j}$  as centered around  $b_0$ , we define a new set of intercepts  $\tilde{b}_{0j}$ , which we model as centered around 0 instead:

$$\tilde{b}_{0j} \sim \text{normal}(0, \sigma_0)$$

Now that  $b_0$  no longer appears in the hierarchical prior, we add it directly to the likelihood:

$$y_n \sim \text{normal}(b_0 + \tilde{b}_{0j[n]} + b_1 x_n, \sigma)$$

The two formulations are equivalent since  $b_{0j} = b_0 + \tilde{b}_{0j}$  so the model assumptions remain the same. That said, the second formulation we just introduced has the advantage that all its terms, the overall intercept  $b_0$ , the overall slope of  $b_1$  of predictor  $x$ , and the varying intercept  $\tilde{b}_{0j}$  all appear in the same place, namely the mean parameter of the likelihood. It is now easy to express this mean parameter as an R formula.

Inspired by the syntax of the `lme4` package, multilevel terms are specified in brms as `(terms | group)`. Accordingly, if we want to add a varying intercept over subjects, we have to add `(1 | Subject)` to the model formula. Together with the overall intercept and the overall slope of `Days`, our varying intercept model formula becomes `Reaction ~ 1 + Days + (1 | Subject)`. The order of the terms doesn't matter so we can also write, for example, `Reaction ~ 1 + (1 | Subject) + Days`, but it is more common to add multilevel terms at the end.

With this formula in hand, and for now trusting the default priors, we are ready to fit our first multilevel models in brms:

```
fit_sleep2 <- brm(Reaction ~ 1 + Days + (1 | Subject),
                   data = sleepstudy)

summary(fit_sleep2)

Family: gaussian
Links: mu = identity
Formula: Reaction ~ 1 + Days + (1 | Subject)
Data: sleepstudy (Number of observations: 144)

Multilevel Hyperparameters:
~Subject (Number of levels: 18)
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)    44.73     8.89    31.05    64.98 1.00      799     1316

Regression Coefficients:
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     267.83    11.36   244.92   289.87 1.00      588     1006
Days          11.44     1.11     9.31    13.59 1.00     4206     2834

Further Distributional Parameters:
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma       30.51     1.92    26.97    34.52 1.00     3836     3170
```

In the summary output, notice the new part named “Multilevel Hyperparameters”, with the single row termed `sd(Intercept)` summarizing the standard deviation  $\sigma_0$  of the intercepts over subjects. We can, of course, also understand the overall intercept  $b_0$  as a hyperparameter, but since this parameter is also present for non-multilevel models we just leave it under “Regression Coefficients”. Based on model results,  $\sigma_0$  has a posterior mean of around 44 ms with a 95% credible interval ranging between 31 and 62. Judged by these credible interval bounds relative to the mean, we see that the posterior of  $\sigma_0$  is somewhat right skewed since the distance of the mean to the upper bound is larger than its distance to the lower bound. When combining the estimates of  $\sigma_0$  with the posterior mean of  $b_0 \approx 268$  ms, we can infer that most subjects have a baseline reaction time of roughly 150 to 390 ms.

The summary output does not show summaries of the individual intercepts for the subjects because that would quickly clutter the output for more than a handful of grouping levels. Similarly to the `summary` method, also the `plot` method will not show any varying coefficients by default.

```
plot(fit_sleep2)
```

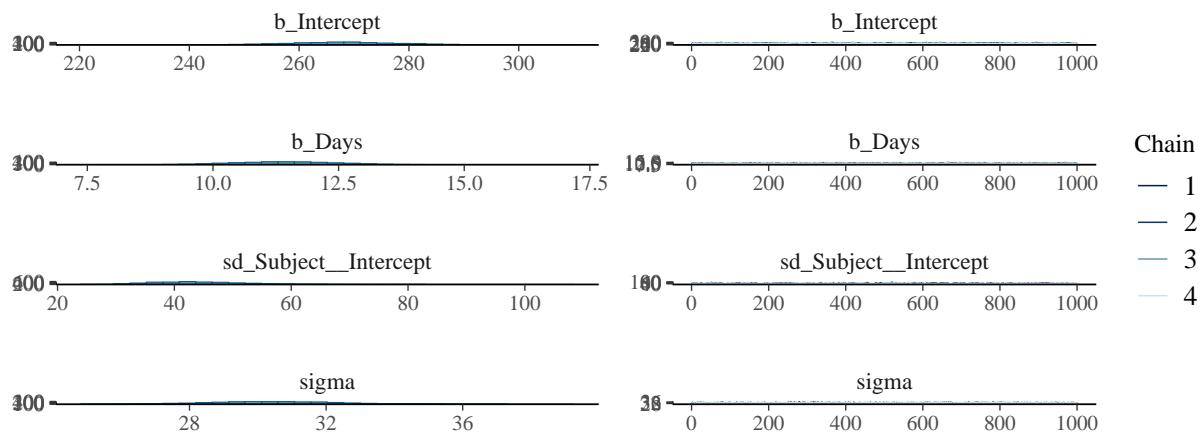


Figure 4.5: Posterior histograms and trace plots for model `fit_sleep2`.

In Figure 4.5, we nicely see the slightly right skewed posterior distribution of  $\sigma_0$  ranging roughly between 30 and 60 ms, confirming graphically what we have seen numerically in the summary output. We can also use `conditional_effects` as usual to visualize the mean predictions associated with the covariates (see Figure 4.6):

```
plot(conditional_effects(fit_sleep2), points = TRUE)
```

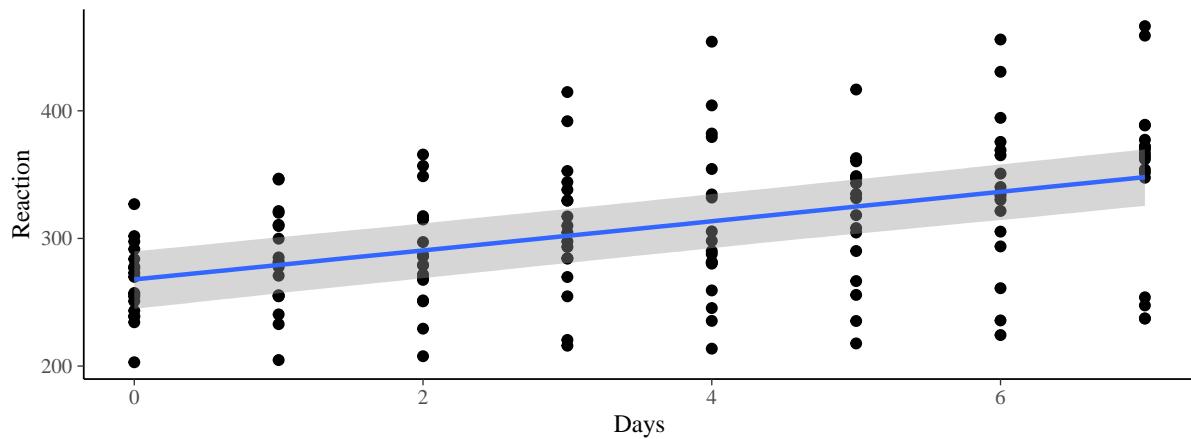


Figure 4.6: Conditional effects for model `fit_sleep2`. Observations are shown as points.

By default, `conditional_effects` will ignore all varying coefficients, that is, assume them to be 0. As a result, the prediction line we are seeing is the posterior of  $b_0 + b_1x$ , just like

it was for the basic linear model without any multilevel terms. Yet, the predictions shown in Figure 4.6 clearly look different than that of the basic linear model, in particular more uncertain. What happened is that, by telling the model about the different groups and their varying intercepts, the overall intercept  $b_0$  became more uncertain since it is now only informed by 18 “data points”, the 18 varying intercepts that are also uncertain themselves, rather than by all the 144 actual observations.

In most real-world applications of multilevel models that I came across, people do not actually care about the varying coefficients. We just need to model them to appropriately to better calibrate the uncertainty of the overall coefficients. That said, we can of course investigate the varying coefficients using several methods dedicated for this purpose. The `ranef` method will return the varying coefficients as estimated by Stan, that is, those  $\tilde{b}_{0j}$  centered around zero:

```
ranef(fit_sleep2)$Subject[, , "Intercept"]
```

Table 4.2: Posterior summary of the first five subject-specific zero-centered intercepts for model `fit_sleep2`.

	Estimate	Est.Error	Q2.5	Q97.5
308	52.96	14.49	25.77	82.28
309	-86.75	14.31	-114.90	-59.06
310	-64.24	14.43	-93.12	-36.41
330	-6.15	14.20	-34.29	21.70
331	7.01	14.29	-20.76	34.92

For example, we see that Subject 308 is around 53 ms slower at baseline than the mean across subjects, while Subject 309 is around 87 ms faster at baseline than said mean. We also see that all these intercept estimates have considerable uncertainty with 95% credible intervals spanning around 60 ms. This is not surprising if we consider that each subject has only 8 associated data points. Even though the partial pooling property allows to borrow information also from the other subjects, the combine information is still relatively sparse leading to the large uncertainty estimates.

While the  $\tilde{b}_{0j}$  coefficients are the ones being estimated by Stan, it is usually the “actual” varying coefficients  $b_{0j} = b_0 + \tilde{b}_{0j}$  that we are care about. We could go ahead and extract the posterior draws of  $b_0$  and  $\tilde{b}_{0j}$  and than add them manually as we have learned in the past chapters, but since this is such a common task in multilevel models, brms has the `coef` method readily made just for this purpose:

```
coef_sleep2 <- coef(fit_sleep2)
coef_sleep2$Subject[, , "Intercept"]
coef_sleep2$Subject[, , "Days"]
```

Table 4.3: Posterior summary of the first five subject-specific intercepts for model `fit_sleep2`.

	Estimate	Est.Error	Q2.5	Q97.5
308	320.78	11.35	298.49	343.00
309	181.07	11.13	159.76	202.70
310	203.59	11.31	181.03	225.64
330	261.68	11.00	239.27	282.71
331	274.84	11.20	253.18	296.66

Table 4.4: Posterior summary of the first five subject-specific slopes for model `fit_sleep2`.

	Estimate	Est.Error	Q2.5	Q97.5
308	11.44	1.11	9.31	13.59
309	11.44	1.11	9.31	13.59
310	11.44	1.11	9.31	13.59
330	11.44	1.11	9.31	13.59
331	11.44	1.11	9.31	13.59

The output of `coef`, shown in Table 4.3 and Table 4.4, follows the same structure as that of `ranef`. However, `coef` shows all coefficients even those without varying coefficients associated with it. As a result, we also see per-subject coefficients of `Days`, which are however constant across subjects and just resemble the overall coefficient that we already saw in the summary output.

When we compare the results of `coef` and `ranef`, we see that not only `coef` shows estimates around  $b_0$  instead of around 0 but also that the uncertainty of `coef` is actually *smaller* than that of `ranef` as can be seen by both the posterior standard deviations (column "Est.Error") and the credible interval range. This is curious since, in `coef`, we actually have added something, namely (the posterior of)  $b_0$ , which should intuitively lead to an increase in uncertainty rather than a decrease. To understand what has happened, let's look at the posterior of  $b_0$  as well as the  $b_{0j}$  for the first two subjects via a posterior pairs plot:

```
vars <- c("b_Intercept", "r_Subject[308,Intercept]", "r_Subject[309,Intercept]")
pairs(fit_sleep2, variable = vars)
```

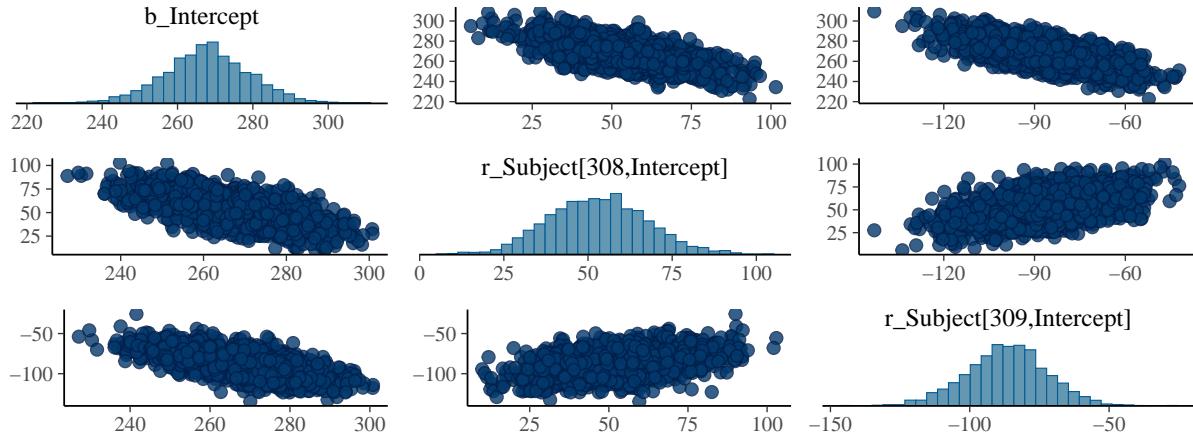


Figure 4.7: Posterior pairs plot of the overall intercept and two subject-specific intercepts.

In Figure 4.7, we see a strong negative correlation between  $b_0$  on the one hand and the  $b_{0j}$  on the other hand. Conversely, the  $b_{0j}$  are positively correlated with one another, as we also see from computing the correlations explicitly (see Table 4.5):

```
draws_sleep2 <- as.matrix(fit_sleep2, variable = vars)
colnames(draws_sleep2) <- c("b0", "b0_308", "b0_309")
cors_sleep2 <- cor(draws_sleep2)
```

Table 4.5: Posterior correlation of the overall intercept and two subject-specific zero-centered intercepts for model `fit_sleep2`.

	b0	b0_308	b0_309
b0	1.00	-0.64	-0.65
b0_308	-0.64	1.00	0.46
b0_309	-0.65	0.46	1.00

This happens because, in the likelihood, we add  $b_0$  and  $b_{0j}$  together such that the both “fight” for the same information in the data, namely the baseline reaction times. As  $b_0$  increases, so must the  $b_{0j}$  decrease (and vice versa) in order to properly fit the empirical baseline reaction times. Since all of the  $b_{0j}$  fight with  $b_0$  but not with each other, since they relate to data from different subjects, they end up being positively correlated with each other. These correlations are largely not due to any partial pooling effect but simply an artifact of how we have parameterized our model. But what happens if we look at the “actual” varying coefficients obtained via `coef`?

```
coef_sleep2 <- coef(fit_sleep2, summary = FALSE)
cor(coef_sleep2$Subject[, 1:5, "Intercept"])
```

Table 4.6: Posterior correlation of five subject-specific intercepts for model `fit_sleep2`.

	308	309	310	330	331
308	1.00	0.12	0.12	0.13	0.15
309	0.12	1.00	0.19	0.12	0.12
310	0.12	0.19	1.00	0.15	0.12
330	0.13	0.12	0.15	1.00	0.14
331	0.15	0.12	0.12	0.14	1.00

As we can see exemplarily for the first five subjects in Table 4.6, their varying intercepts are indeed all mildly correlated with each other with a correlation of roughly  $r \approx 0.1$ . These correlations are now actually due to partial pooling and it shows how the multilevel model enables sharing of information across coefficients. While  $r \approx 0.1$  doesn't look like much, remember that it applies to all pairs of varying intercept, such that their *joint* information exchange adds up. For example, when we take the posterior of an individual subject's intercept and correlate it with the posterior of the mean of all other subjects' intercepts, we see a correlation of about  $r \approx 0.3$ , which is no longer that small. I suggest you to compute this correlation yourself as an exercise. These posterior correlations are not the only way in which hierarchical priors find their way into the posterior and we will come back to this later on.

Let us move on to investigating the per-subject predictions by including the varying intercepts into the prediction plots created via `conditional_effects` (see Figure 4.8):

```
conditions <- make_conditions(sleepstudy, "Subject")
ce <- conditional_effects(
  fit_sleep2, conditions = conditions,
  re_formula = NULL
)
plot(ce, ncol = 6, points = TRUE)
```

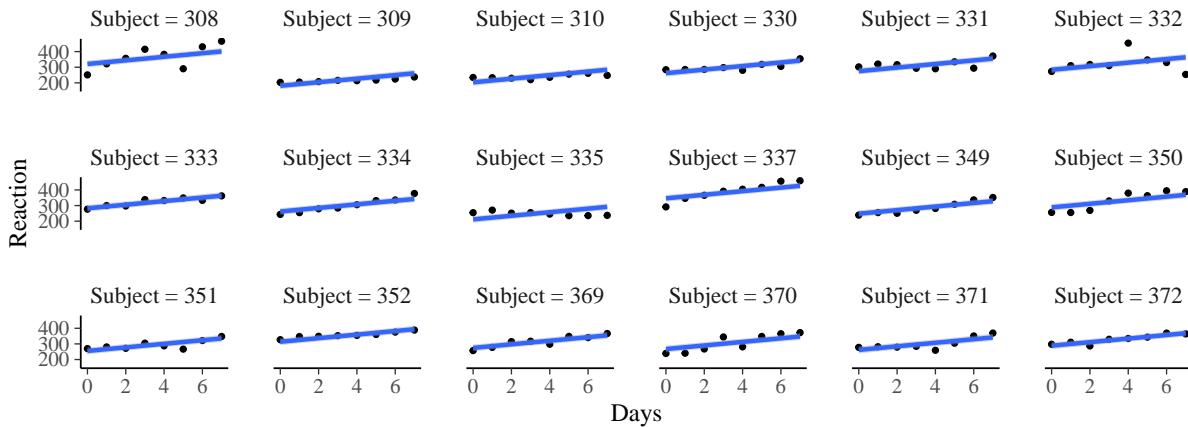


Figure 4.8: Subject-specific conditional effects for model `fit_sleep2`. Observations are shown as points.

With the `conditions` argument, we indicate that we want to generate one plot per row of the supplied data frame, here one row per subject. By setting `re_formula = NULL` we indicate that we want the varying coefficients are included in the predictions. By default, `conditional_effects` would ignore them, as we have seen above. As we see in Figure 4.8, the model provides quite reasonable predictions for all subjects, but obviously fails to account for the differences in the slopes. That is, for some subjects with more extreme (large or small) slopes, there is an easily visible difference in the data patterns and the model's predictions. The fact that the varying intercept model fits better than our complete pooling model is also confirmed by the LOO-CV results, which show an ELPD difference of about 65, with a standard error of about 13, in favor of the varying intercept model:

```
loo_sleep1 <- loo(fit_sleep1)
loo_sleep2 <- loo(fit_sleep2)
loo_compare(loo_sleep1, loo_sleep2)
```

	elpd_diff	se_diff
<code>fit_sleep2</code>	0.0	0.0
<code>fit_sleep1</code>	-65.1	13.3

The posterior predictive checks shown in Figure 4.9 also look better than those of the complete pooling model. But it still shows a clearly visible relationship between residuals and responses, pointing to some signal in the data that we have failed to model still.

```
pp_check(fit_sleep2, "error_scatter_avg")
```

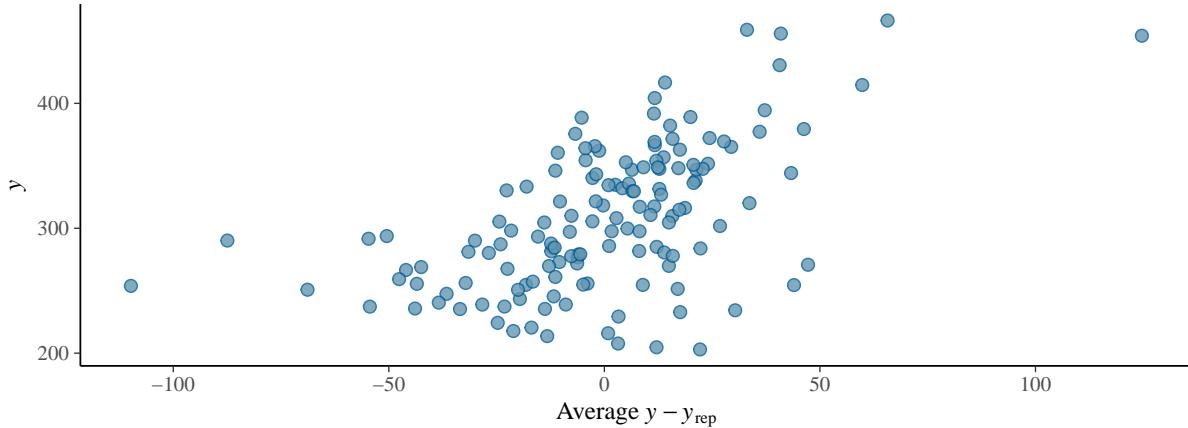


Figure 4.9: Posterior predictive check of residuals for model `fit_sleep2`.

So far, we have not talked about priors for the hyperparameter  $\sigma_0$  yet, and instead just used brms' default priors. So what was the default prior actually?

```
prior_summary(fit_sleep2)
```

	prior	class	coef	group	lb	ub	tag	source
	(flat)	b						default
	(flat)	b		Days				(vectorized)
student_t(3, 303.2, 65.5)	Intercept							default
student_t(3, 0, 65.5)	sd				0			default
student_t(3, 0, 65.5)	sd			Subject	0			(vectorized)
student_t(3, 0, 65.5)	sd	Intercept	Subject	Subject	0			(vectorized)
student_t(3, 0, 65.5)	sigma				0			default

As we can see from the `prior_summary` output, the default prior on standard deviations of the varying coefficients (class "sd") for this model is `student_t(3, 0, 65.5)`; the same as the default prior of the residual standard deviation (class "sigma"). Remember from earlier chapters that, internally, such priors for standard deviations are truncated at zero such that we are looking at a half-Student-t prior with small degrees of freedom (heavy right tails) and a data-dependent scale, that is for linear models equal to the standard deviation of the response variable. As is the case for many brms' default priors, I don't claim it is necessarily the best default prior on standard deviation parameters. However, it fulfills two important criteria: (a) It is highly unlikely to be more than weakly informative independent of the specific dataset or model due to the small degrees of freedom and data-adaptivity. (b) It often improves sampling speed and convergence compared to completely flat priors. Later on, in Section 4.10, we will discuss an example on how to specify informative priors for standard deviations of varying coefficients. For now, we move on directly to more complex multilevel models.

## 4.5 Partial pooling: Varying intercepts and slopes

Above, we have seen that only letting the intercept vary across subjects was insufficient since, apparently, also the effect of the days of sleep deprivation varies across subjects. So let's model this appropriately now. Our likelihood continues to be normal but now the its mean parameter is obtained from both a varying intercept  $b_{0j}$  and a varying slope  $b_{1j}$  of the respective subject to which the  $n$ th observation belongs:

$$y_n \sim \text{normal}(b_{0j[n]} + b_{1j[n]}x_n, \sigma)$$

So far so simple. But now we have to figure out how to write down an appropriate multilevel prior for the pair  $(b_{0j}, b_{1j})$ . The most immediate solution is specify two independent multilevel priors, one for each kind of coefficient. This would then look as follows:

$$b_{0j} \sim \text{normal}(b_0, \sigma_0)$$

$$b_{1j} \sim \text{normal}(b_1, \sigma_1)$$

We now have four hyperparameters, namely, the overall intercept  $b_0$ , overall slope  $b_1$  as well as the standard deviations  $\sigma_0$  and  $\sigma_1$  for the varying intercepts and slopes, respectively. That is a valid and sensible way to set up multilevel priors. However, if we choose the “independent prior” approach, we miss the opportunity for the intercepts and slopes to inform each other. For example, it is conceivable that subjects who are slow to react already before sleep deprivation are more affected by said deprivation, thus also having a larger slope. Or, conversely, it may be that subjects who are faster at the start are more strongly affected by sleep deprivation.

Admittedly, none of these two directions is very plausible for the `sleepstudy` data. So let us dive into a little thought experiment where such dependencies are more plausible. Suppose we are analyzing how much students learn from a math class by evaluating students math skills both before and after the class. It could be that the math class (and the exams) deal with easy topics, which the good math students already know. Accordingly, the good students wouldn't learn much from the class, a pattern that we call a “ceiling effect” in psychology. However, the not-so-good students, who struggled with the exam before the class, have now understood more of the concepts and subsequently improve quite a bit in the exam after class. In this context, smaller intercepts (i.e., lower exam scores before the class) would go hand in hand with higher slopes (i.e., larger differences between the exams after and before the class). Speaking in terms of multilevel models, this would imply a *negative* correlations between varying intercepts and slopes. The same argument can be made for a very difficult math class that only the good students would benefit from, while the math skills of the not-so-good students would remain about the same. This would induce a “floor effect”, which implies a *positive* correlation between varying intercepts and slopes. In both of these scenarios,

our estimates of the varying coefficients would benefit from a multilevel prior that supports the exchange of information Not only among the coefficients of the same kind (e.g., varying intercepts of different students), but also among the different kinds of varying coefficients (i.e., the varying intercept and slope of the same student).

The way to achieve this is by imposing a *multivariate* normal prior on the pairs  $(b_{0j}, b_{1j})$  of varying intercepts and slopes, with a pair of overall mean coefficients  $(b_0, b_1)$  and a covariance matrix  $\Sigma_{01}$  that models the variation of the intercepts and slopes, respectively, as well as their dependency as motivated above:

$$(b_{0j}, b_{1j}) \sim \text{multinormal}((b_0, b_1), \Sigma_{01})$$

Personally, I struggle interpreting variances and I struggle even more interpreting covariances. Fortunately, we can decompose any covariance matrix into a set of standard deviations and a correlation matrix, both being more intuitive quantities for most people I believe. Explaining this for our specific case at hand, the diagonal elements of  $\Sigma_{01}$  are the variances, which we obtain from the standard deviations  $\sigma_0$  and  $\sigma_1$  by simply squaring them. The covariance between the varying intercepts and slopes can be expressed as the product of the two standard deviations  $\sigma_0$  and  $\sigma_1$  and their correlation  $\rho_{01}$ . Putting this together, we obtain

$$\Sigma_{01} = \begin{pmatrix} \sigma_0^2 & \sigma_0\sigma_1\rho_{01} \\ \sigma_0\sigma_1\rho_{01} & \sigma_1^2 \end{pmatrix}.$$

We now have  $\sigma_0$ ,  $\sigma_1$ , and  $\rho_{01}$  as hyperparameters in addition to  $b_0$  and  $b_1$ . Comparing this to the “independent prior” approach above, we have actually only added  $\rho_{01}$  to the set of hyperparameters. Since we only have two varying coefficients per subject in our case study, the correlation matrix in fact consists of only one relevant element, namely  $\rho_{01}$ . If we had more than two coefficients per grouping-level the number of estimated correlation hyperparameters would of course increase correspondingly. We will dive deeper into the specification of priors on correlations and correlation matrices in Section 4.7. For now, it suffices to say that brms applies a uniform prior in  $[-1, 1]$  on  $\rho_{01}$  by default for this model. The “zero-centered” version used in brms works in basically the same way as before. That is, we set

$$(\tilde{b}_{0j}, \tilde{b}_{1j}) \sim \text{multinormal}((0, 0), \Sigma_{01})$$

and then compute the predictor term on the likelihood mean as

$$\mu_n = b_0 + b_1 x_n + \tilde{b}_{0j[n]} + \tilde{b}_{1j[n]} x_n.$$

In brms, we can indicate that we want both varying intercepts and varying slopes of Days across subjects by adding `(1 + Days | Subject)` to the model formula, in addition to the overall intercept and overall slope. If we do not want to model the correlation hyperparameters,

we can replace the `|` symbol with `||`, that is, `(1 + Days || Subject)` for our present example. Not modeling the correlations will sometimes lead to quite substantial speedups during model estimation so can be a useful shortcut if some models run too slowly. In general, I would recommend to include the correlation parameters to improve the exchange of information among the varying coefficients. That being said, the complete brms model with correlation (and default priors) looks as follows:

```
fit_sleep3 <- brm(
  Reaction ~ 1 + Days + (1 + Days | Subject),
  data = sleepstudy
)

summary(fit_sleep3)

Family: gaussian
Links: mu = identity
Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
Data: sleepstudy (Number of observations: 144)

Multilevel Hyperparameters:
~Subject (Number of levels: 18)
  Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)    33.54     7.76   21.03   51.60 1.00    1642    1954
sd(Days)         7.51     1.89    4.44   11.95 1.00    1595    2240
cor(Intercept,Days) 0.17     0.31   -0.41    0.74 1.00    1276    1651

Regression Coefficients:
  Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    268.03     8.99   251.01   285.92 1.00    1508    2496
Days          11.37     2.01    7.44    15.41 1.00    1519    1995

Further Distributional Parameters:
  Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma       25.99     1.81   22.70   29.74 1.00    2856    2915
```

From the summary output, we see that there is substantial variation in the slopes across subjects, with the posterior mean of  $\sigma_1$  being around 7.4 and a 95% credible interval spanning roughly between 4.3 and 11.9 ms. There appears to be no noteworthy correlation between varying intercepts and slopes as we can see from the results of the correlation hyperparameter  $\rho_{01}$ . Its credible interval is quite wide, which makes sense since the correlation is based on only 18 “data points”, the 18 subjects included in the dataset. As you may know from your experience with correlations more generally, we usually need much larger datasets to

achieve precise correlation estimates, so the high uncertainty here is not surprising from this perspective.

Before we start studying the varying coefficients in detail, let us better understand the implications of adding multilevel structure on the overall intercepts and slopes. In Figure 4.10, we see the conditional effect of `Days` for the complete pooling model vs. the varying intercept-slope model. Two relevant patterns become apparent. First, the mean prediction line is very similar for both models. This is not a coincidence but follows from the fact that all subjects have the same number of observations. We can think of the complete pooling model as obtaining its coefficients by averaging over all 144 observations. The partial pooling model, in contrast, obtains its overall coefficients by averaging over the 18 subjects. If each subject has the same number of observations, both averages are the same. Yet, the fact that the posterior mean lines are highly similar does not mean that the choice of model becomes irrelevant. That is because, second, the posterior credible intervals of the predictions are very different. Specifically, the partial pooling model implies clearly more uncertainty than the complete pooling model. When I have shown this comparison to students, I have often gotten the question on why higher uncertainty should be considered better and I imagine some of you might ask yourself the same question. The truth is, neither higher nor lower uncertainty are necessarily preferable. Rather, broadly speaking, it is about finding *just the right amount of uncertainty*, which accurately reflects our knowledge. This can be formalized statistically, where it is commonly known under the term *calibration*. I will provide a practical example of how to study calibration in Section 4.8. I recommend you to also check out Section 3.2 in Bürkner, Scholz, and Radev (2023a), specifically the parts about uncertainty estimation.

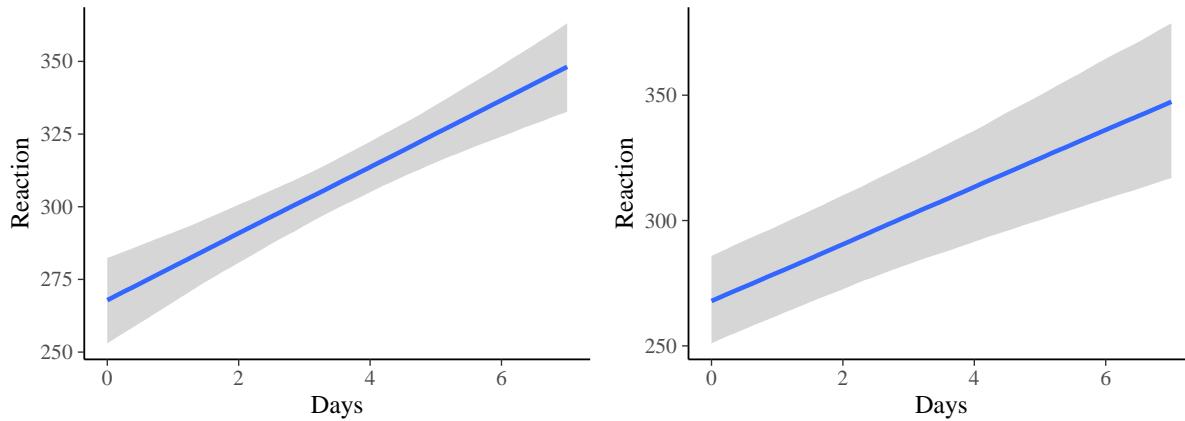


Figure 4.10: Conditional effects of the complete pooling model (left) vs. the partial pooling model (varying intercepts and slopes; right). The blue line indicates the mean predictions while the gray area indicates the 95% credible intervals of the predictions.

We will now turn our attention to the varying coefficients. When we run `coef`, we see that both also the slope posteriors are now varying across subjects:

```
coef(fit_sleep3)$Subject[, , "Days"]
```

Table 4.7: Posterior summary of the first five subject-specific slopes for model `fit_sleep3`.

	Estimate	Est.Error	Q2.5	Q97.5
308	19.97	3.26	13.65	26.27
309	3.15	3.41	-3.48	9.99
310	3.91	3.42	-2.90	10.63
330	9.09	3.19	2.63	15.31
331	7.74	3.32	0.97	13.79

As we can see in Table 4.7, some subjects (e.g., Subject 309) have a posterior mean slope of close to 3 ms. That is, their reaction times seem to be basically unaffected by sleep deprivation. Other subjects, for example, Subject 308, have a slope of about 20 ms. That is their reaction times increase by 20 ms per day in expectation, which, when extrapolated over the 7 days of sleep deprivation, ultimately leads to a quite substantial increase of about 140 ms. We can also see this visualized in the per-subject `conditional_effects` plots (see Figure 4.11):

```
ce <- conditional_effects(
  fit_sleep3, conditions = conditions,
  re_formula = NULL
)
plot(ce, ncol = 6, points = TRUE)
```

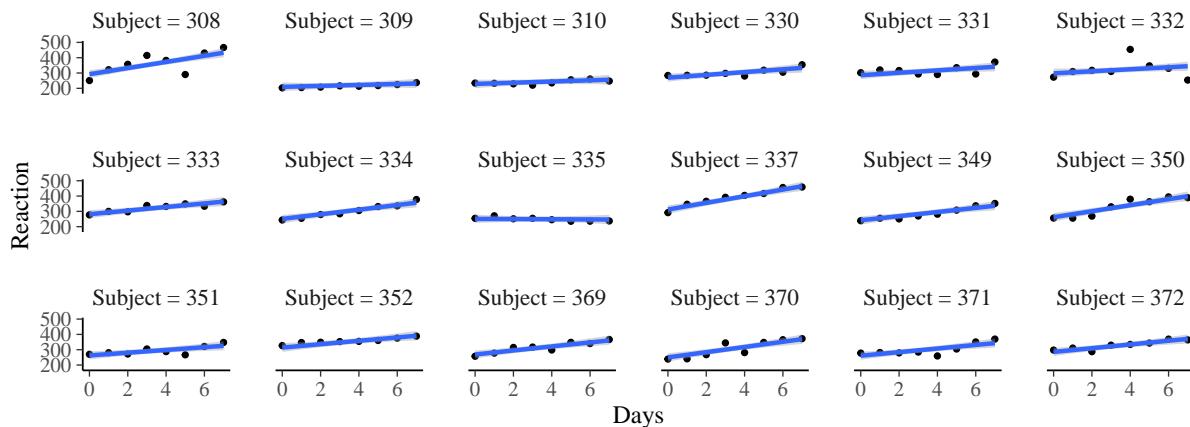


Figure 4.11: Subject-specific conditional effects for model `fit_sleep3`. Observations are shown as points.

Clearly, any sensible model of this data should somehow account for this strong variation across subjects. Indeed, the posterior predictive checks are looking much improved compared to the model with only varying intercepts (see Figure 4.12):

```
pp_check(fit_sleep3, "error_scatter_avg")
```

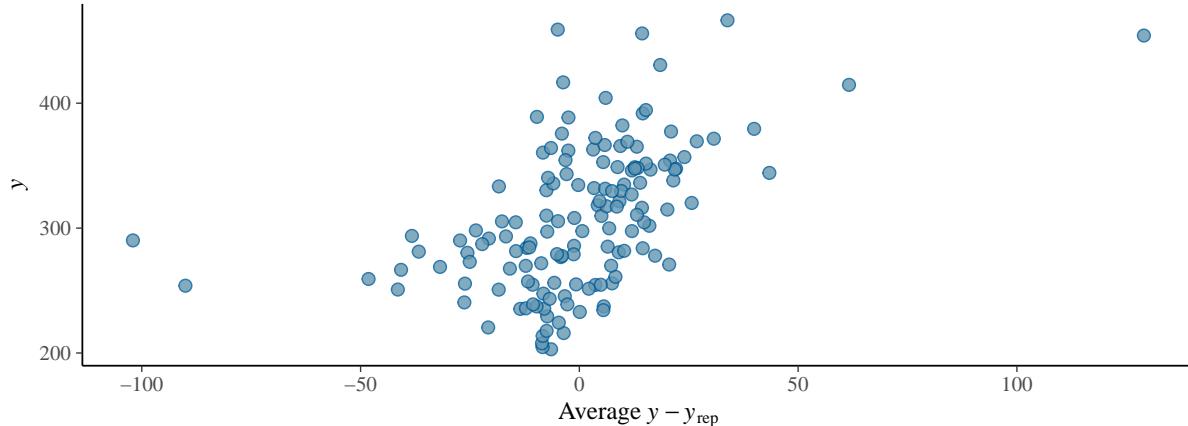


Figure 4.12: Posterior predictive check of residuals for model `fit_sleep3`.

The only remaining issue in model fit we see in this plot are 3 or 4 outlying data points, which we can also identify from the above `conditional_effects` plots. It seems that some subjects sometimes had an uncharacteristically good or bad day, defying the overall trend in their reaction time changes. Based on the available data, there is nothing that provides us with a better understanding why that happens exactly. As a remedy, we could for example change the likelihood from Gaussian to Student-t to make the model less affected by such outliers. I leave this modeling task to you as an exercise.

To compare our models more formally, we again run LOO-CV:

```
loo_sleep3 <- loo(fit_sleep3)
```

```
Warning: Found 3 observations with a pareto_k > 0.7 in model 'fit_sleep3'. We recommend
to set 'moment_match = TRUE' in order to perform moment matching for problematic
observations.
```

We get a warning that our PSIS approximation of LOO-CV has failed for 3 observations, clearly the exact same observations we identified in the above `pp_check` plot. We choose to ignore this warning for now and proceed with interpreting the results. The effective number of parameters evaluates to roughly 34, which is clearly smaller than the nominal 42 parameters we have in

our model (36 varying coefficients plus 5 hyperparameters plus the residual standard deviation  $\sigma$ ). This is something we usually see in multilevel models as the partial pooling property of the multilevel priors reduce the de-facto complexity of the model. Let's compare the LOO-CV performance of our three models to each other:

```
loo_compare(loo_sleep1, loo_sleep2, loo_sleep3)
```

	elpd_diff	se_diff
fit_sleep3	0.0	0.0
fit_sleep2	-14.8	9.5
fit_sleep1	-79.9	20.4

The evidence in favor of the varying intercept-slope model `fit_sleep3` being the best is pretty strong, to a degree that it is unlikely caused by PSIS failing for 3 observations. Together with the large differences in varying slopes seen previously, we decide that we have gathered sufficient evidence for choosing the varying intercept-slope model as the best among our three candidate models. If the evidence was less clear, we would probably need to run moment matching to improve the trustworthiness of our LOO-CV approximation. Since the results remain basically unchanged, I will just show the code below for your reference.

```
# we need to refit the model while saving all parameters
fit_sleep3 <- update(fit_sleep3, save_pars = save_pars(all = TRUE))
mmlloo_sleep3 <- loo_moment_match(fit_sleep3, loo = loo_sleep3)
loo_compare(loo_sleep1, loo_sleep2, mmlloo_sleep3)
```

## 4.6 Predicting coefficients of new levels

One of the really cool things we can do with multilevel models is to predict coefficients of new levels of grouping variables (aka new groups), not previously seen by the model. Let's ask to model to make predictions about days 0 and 7 for Subjects 308 and 309, both present in the original data, and for a new subject that we call "`new`". It doesn't really matter how we call it. As long as it is not in the list of names in the original data, brms will recognize it as new:

```
newdata <- expand.grid(Days = c(0, 7), Subject = c("308", "309", "new"))
posterior_epred(fit_sleep3, newdata = newdata)
# Error: Levels 'new' of grouping factor 'Subject' cannot be found in the
# fitted model. Consider setting argument 'allow_new_levels' to TRUE.
```

brms throws an error on our first attempt to make predictions for a new group to make sure that we really indented it. Since the predictions of new and old subject will be behave quite

differently (see below), I don't want users to accidentally make new level predictions through a typo in the old names. The error message however informs us what to do if we really want to make predictions for new levels, namely setting `allow_new_levels = TRUE`. Additionally, we can choose among several options of *how* to create varying coefficients for new levels, via argument `sample_new_levels`. We will choose the `sample_new_levels = "gaussian"` option for now. It implies that, for each posterior draw, we will sample varying coefficients from a normal distribution with hyperparameters set to the hyperparameter values in this posterior draw:

```
posterior_epred(
  fit_sleep3, newdata = newdata, allow_new_levels = TRUE,
  sample_new_levels = "gaussian"
) %>%
  posterior_summary() %>%
  cbind(newdata, .)
```

Table 4.8: Posterior summary of expected predictions for model `fit_sleep3` for three subjects: two existing subjects and one new subject.

Days	Subject	Estimate	Est.Error	Q2.5	Q97.5
0	308	292.20	13.64	265.01	318.58
7	308	432.86	15.10	403.27	462.40
0	309	209.24	15.09	179.03	238.84
7	309	231.78	15.35	201.47	261.11
0	new	268.30	35.33	197.60	337.27
7	new	349.66	69.00	210.24	486.64

The variation in the new-level predictive distribution shown in Table 4.8 contains both the uncertainty in the hyperparameters' posterior and the variation implied by sampling from a normal distribution with these hyperparameters. In other words, the new-level predictive distribution contains also variation *across* levels and hence will be much wider than the posterior distributed from levels existing in the original data. This is exactly what we see reflected in the output above. It may help with intuition to see all of this visualized:

```
conds <- data.frame(Subject = c("308", "309", "new"))
rownames(conds) <- conds$Subject
conditional_effects(fit_sleep3, conditions = conds, re_formula = NULL,
  sample_new_levels = "gaussian")
```

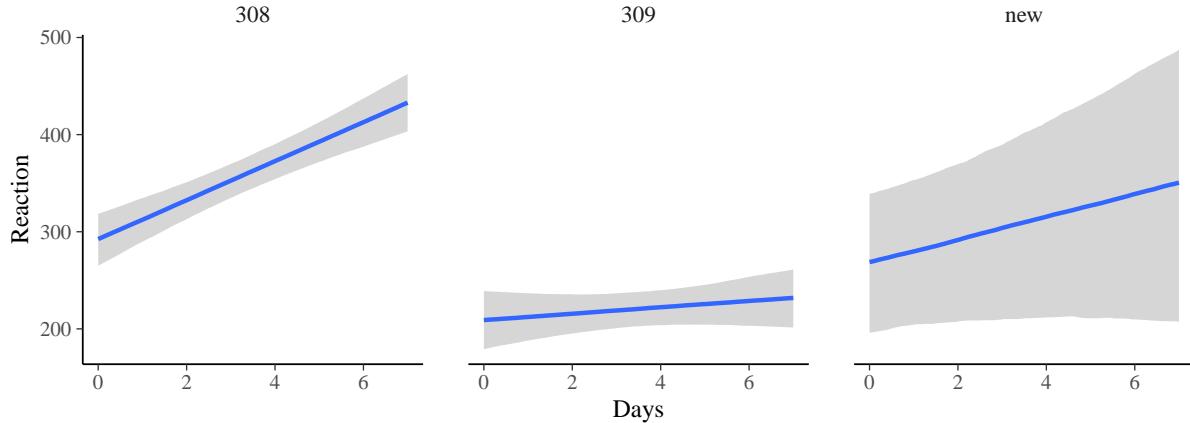


Figure 4.13: Conditional effects for model `fit_sleep3` for three subjects: two existing subjects (left and center) and one new subject (right).

Indeed, as we see on the right-hand side of Figure 4.13, the new-level predictive distribution covers almost the entire range of the original subjects' predictive distributions. Subjects 308 and 309 are on opposite ends of the distribution of varying coefficients and indeed their predictions are roughly at the borders of the new-level predictive distribution.

The above predictions always combined the posterior uncertainty in the hyperparameter with the subsequent predictive uncertainty on the values of the new varying coefficients. What if we want to only get the latter uncertainty, while keeping hyperparameter values constant? We can solve this in brms by obtaining several predictions for the same posterior draw. Say, we want to obtain 20 new-level predictions for posterior draw 7, we can achieve this as follows:

```
posterior_epred(
  fit_sleep3, newdata = newdata, allow_new_levels = TRUE,
  sample_new_levels = "gaussian", draw_ids = rep(7, 20)
) %>%
  posterior_summary() %>%
  cbind(newdata, .)
```

Table 4.9: Posterior summary of expected predictions for model `fit_sleep3` for three subjects: two existing subjects and one new subject. Only a single posterior draw was used to create the predictions.

Days	Subject	Estimate	Est.Error	Q2.5	Q97.5
0	308	287.05	0.00	287.05	287.05
7	308	441.00	0.00	441.00	441.00
0	309	218.28	0.00	218.28	218.28

Table 4.9: Posterior summary of expected predictions for model `fit_sleep3` for three subjects: two existing subjects and one new subject. Only a single posterior draw was used to create the predictions.

Days	Subject	Estimate	Est.Error	Q2.5	Q97.5
7	309	194.85	0.00	194.85	194.85
0	new	268.13	29.85	216.32	318.57
7	new	331.71	63.10	222.91	431.97

In Table 4.9, notice how the old level predictions are constant, since we always used the same posterior draw there. In contrast, the new-level predictions still have variation as a result of sampling from the new-level predictive distribution 20 times, even though hyperparameter values were constant. Again, it helps to see this visualized (see Figure 4.14):

```
set.seed(7332)
ce <- conditional_effects(
  fit_sleep3, conditions =conds, re_formula = NULL,
  sample_new_levels = "gaussian", draw_ids = rep(7, 20),
  spaghetti = TRUE
)
plot(ce, spaghetti_args = list(colour = "blue"),
     line_args = list(colour = alpha("white", 0)))
```

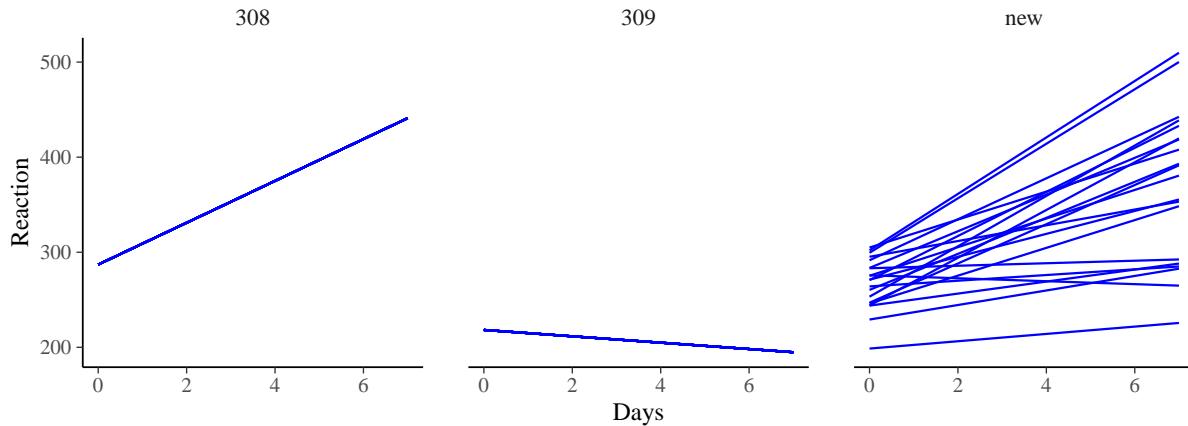


Figure 4.14: Conditional effects for model `fit_sleep3` for three subjects, two existing subjects (left and center) and one new subject (right). Individual posterior predictive draws are shown as lines.

I used `spaghetti = TRUE` such that we see each prediction line separately rather than having them combined to form an uncertainty band around the mean predictions. Indeed, in

Figure 4.14, we see 20 separate lines for the new-level predictions, while there is just a single line for the existing level predictions. The arguments passed to the `plot` method of `conditional_effects` are just some tricks make the individual lines non-transparent, but you will see the same patterns also with the default settings.

Above we have seen the `sample_new_levels = "gaussian"` option throughout, but brms also supports other ways to sample new grouping levels. The default "uncertainty" option samples not from a normal distribution but from the posterior draws of the old levels' coefficients, where for each posterior draw, the old level to draw from is newly chosen. The result are predictions quite similar to that of the "gaussian" option: If the "across-level" distribution of the old levels' coefficients was indeed normal, the two options would basically be the same. However, it might be that the old levels' coefficients are in fact not normally distributed. After all, the multilevel normal distribution we assumed for them was just a prior after all. In a way, the "gaussian" option is a mix of posterior and prior assumptions: We sample the hyperparameters from the posterior but then sample the coefficients from a normal distribution conditional on the posterior hyperparameter values. In contrast, through the "uncertainty" option, we sample purely from the posterior, but with the drawback that the "across-level" posterior is only approximated by the coefficients of the levels in the original data. Accordingly, if a grouping factor only contained few levels in the original data, the approximation of "across-level" posterior may be inaccurate.

Lastly, a third `sample_new_levels` option is "old\_levels", which means that for every new level, we will randomly choose a *single* old level to take all posterior draws of its coefficients from. Accordingly, the new-level predictions obtained this way will, for each new level, be much more certain than with the "gaussian" or "uncertainty" options. Let's sample coefficients of 3 different new levels this way and see what happens:

```
set.seed(3552)
conds <- data.frame(Subject = c("new1", "new2", "new3"))
rownames(conds) <- conds$Subject
conditional_effects(fit_sleep3, conditions = conds, re_formula = NULL,
                    sample_new_levels = "old_levels")
```

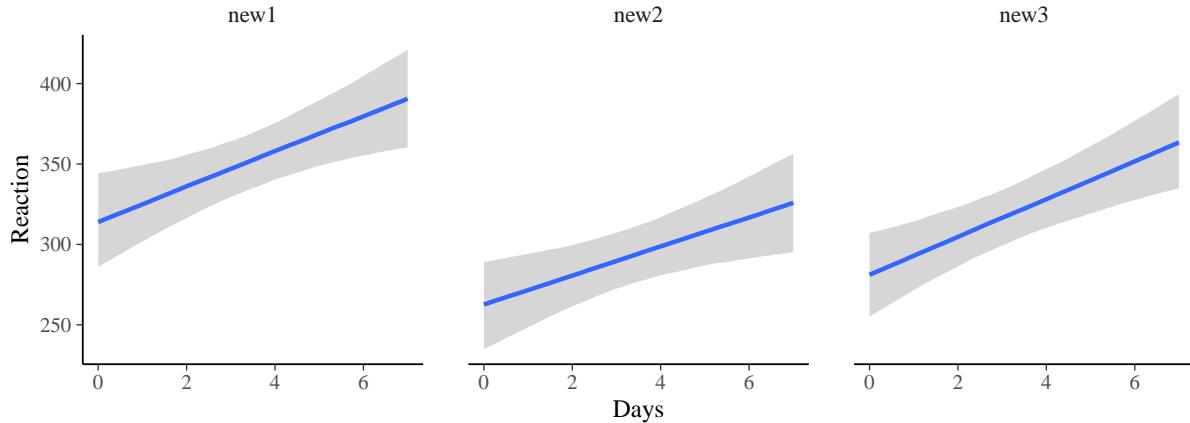


Figure 4.15: Conditional effects for model `fit_sleep3` for three new subjects. For each new subject, coefficients were drawn randomly from the set of existing subjects' coefficients.

Indeed, as we see in Figure 4.15, each of the lines is pretty certain while there might be quite some variation across new levels. It may of course also be that we choose the same old level for two new levels, in which case their predictions become the same. Try it out yourself by varying the random seed above until you get two levels to yield the exact same predictions. On my current setup, I see this happening for example when using the seed 3556.

## 4.7 Priors on correlation matrices

So far, we have not discussed priors on the correlations between varying coefficients, but this is about to change. As we have mentioned in Section 4.5, a covariance matrix can always be decomposed into a vector of standard deviations and a set of correlations organized in the form of a correlation matrix. As you probably know from your basic stats training, a correlation matrix is symmetric with diagonal elements consisting of 1s and off-diagonal elements (the correlations we care about) being in the interval  $[-1, 1]$ . For example, in the 3-dimensional case, a correlation matrix  $\Omega$  looks as follows:

$$\Omega = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{pmatrix},$$

where always  $\rho_{ij} = \rho_{ji}$ . Accordingly, specifying a prior on  $\Omega$  requires only the lower-diagonal elements to vary since the diagonal elements are 1 and the upper-diagonal elements are the same as the lower-diagonal ones. Still, defining a prior on correlation matrices is not a trivial endeavor, since we need to ensure that only *valid* correlation matrices are possible. To put

it intuitively, by valid I mean that a correlation matrix could theoretically be obtained from data, or equivalently, that we could use it to simulate data. Mathematically, this translates to the correlation matrix being *positive definite*. I mention this term just that you have heard of it, in case you come across it again at some other point.

Not all correlation matrices are valid even if all included correlations individually are within the interval  $[-1, 1]$ . Let's make a simple example as an illustration. Think of three variables  $x_1, x_2, x_3$  where  $x_1$  and  $x_2$  are very highly correlated, say,  $\rho_{12} = 0.99$ . This now restricts which correlations of  $x_1$  and  $x_3$  as well as of  $x_2$  and  $x_3$  are possible. Suppose we also have  $\rho_{13} = 0.99$ . Then, say,  $\rho_{23} = -0.99$  would be impossible (i.e., invalid). Intuitively, this can be understood as follows: If one variable ( $x_1$ ) is highly positively correlated with two other variables ( $x_2, x_3$ ), then these two other variables also have to be (relatively) highly positively correlated with each other, certainly not highly negatively correlated. Each of the individual correlations are valid. Only if viewed jointly, they become invalid. As the dimensionality of the correlation matrix grows, the number of such interdependencies between correlations grows, such that the space of valid correlation matrices becomes smaller.

Within the space of valid correlation matrices of given dimension  $D$ , we can now choose if we want to favor certain kind of matrices over others. The prior of choice for correlation matrices in brms and Stan is the LKJ prior (Lewandowski, Kurowicka, and Joe 2009), named after the initials of its inventors. This prior has a single positive shape parameter  $\eta$ . If  $\eta = 1$  all valid correlation matrices of dimension  $D$  are equally likely. That is,  $\eta = 1$  implies a “uniform” prior over the space of valid correlation matrices. If  $\eta > 1$ , correlation matrices with correlations closer to zero are favored and if  $\eta < 1$  correlation matrices with correlations further away from zero are favored. In any case, the LKJ prior is always symmetric around zero and has the same marginal distribution for all individual (off-diagonal) correlations that are part of the matrix. For dimensions  $K = 2, 3, 9$  and varying  $\eta$ , its marginal distributions for the individual correlations are illustrated in Figure 4.16.

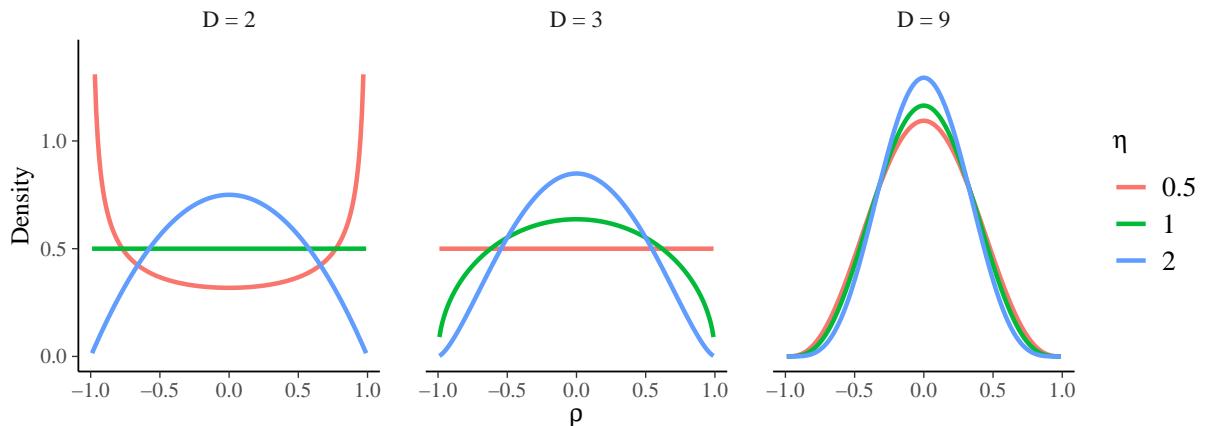


Figure 4.16: Marginal (per-correlation) distribution of the LKJ prior for varying shape parameter  $\eta$  and varying matrix dimension  $D$ .

For  $D = 2$ , the uniform case  $\eta = 1$  indeed means that all correlations in  $[-1, 1]$  are equally likely. This makes sense since, for  $D = 2$ , the correlation matrix only contains a single correlation that is free to vary. Still for  $D = 2$ , if we set  $\eta < 1$  or  $\eta > 1$ , we see the above described patterns of a U-shaped or an inverse U-shaped prior, respectively. This changes as soon as  $D > 2$ . Already at  $D = 3$ , we see how the supposedly “uniform” choice of  $\eta = 1$  implies very much non-uniform marginal prior where highly positive or negative correlations become less likely. As the simple example above illustrates, this makes total sense, but it is certainly not the most intuitive behavior. Incidentally,  $\eta = 0.5$  now implies a uniform marginal prior for  $D = 3$ . At  $D = 9$ , we see an even stronger tendency towards zero correlations, where there is even barely any difference anymore between the hyperparameter values of  $\eta = 0.5, 1, 2$ .

As a user, you will rarely have to worry about the specifics of the LKJ prior or valid correlation matrices more generally. In my experience, setting  $\eta = 1$  is usually a safe choice and also the default in brms. If we want to make more extreme correlations a priori less likely, we may choose something like  $\eta = 2$  or even  $\eta = 4$  but this will only have a noticeable effect on the marginal prior distributions in lower dimensions as the  $D = 9$  case in Figure 4.16 has shown us.

In brms, we can set LKJ priors on correlation matrices via the pattern `prior(lkj(<eta>), class = "cor")`. The dimensions of the correlation matrices are automatically inferred from the model and do not need to be specified in the prior argument. As an exercise, refit the varying intercept-slope model `fit_sleep3` with varying LKJ priors (say,  $\eta = 0.5, 1, 2, 4$ ) and investigate how the posterior on the correlation between the varying intercepts and slopes changes.

## 4.8 Simulation study of Type 1 errors

I hope I have convinced you so far that multilevel models are a reasonable choice for the `sleepstudy` data, at least more reasonable than ignoring the repeated measurements per subject altogether. But what implications does it have if we decide to ignore the multilevel structure still? One thing I know a lot of researchers are caring about are Type 1 error rates in the context of null hypothesis significance testing. That is, given a truly null effect, how likely is it that the statistical model or test will incorrectly lead us to conclude that there is a non-null effect?

For our example this concretely means the following: Suppose there truly was no effect of sleep deprivation on the reaction times (i.e.,  $b_1^* = 0$ ), how likely is that the model would lead us to conclude that there is an effect of sleep deprivation? This is a question at the heart of frequentist statistics, but I see no issues using Bayesian models to answer frequentist questions. There is no guarantee that a Bayesian model has correct frequentist calibration but neither is there a guarantee that a frequentist model has correct frequentist calibration, except in very simple scenarios that are unlikely in reality.

Of course, we should still understand the frequentist properties of our Bayesian models before we use them for this purpose. To make this more concrete, if our models have a correct frequentist calibration for the parameter  $b_1$  we should, for example, see that its 95% posterior credible interval contains the true value of  $b_1^* = 0$  in 95% of the datasets generated from this true value. We cannot use real datasets for the purpose of checking the calibration of any (latent) model parameter, since we will never know its true value exactly; even if we could argue that the parameter itself would somehow “exist” for this dataset. Accordingly, we need simulated data. And not just one simulated dataset, but many of them, since frequentist calibration makes a statement about the data-generating process, not about any specific dataset. In other words, we will need a simulation study.

First, we set up a function for our assumed data-generating process. Here, we will assume that the varying intercept-slope model is indeed the “true” model. To make the code a bit easier to read, we will only allow for equal number of observations per subject and assume the correlation between varying intercepts and slopes to be always zero. Since this roughly resembles the situation in our real data, I will be just fine for our purposes.

```
# simulate datasets that resemble the structure of the sleepstudy data
sim_sleepstudy_data <- function(beta0, beta1, tau0, tau1, sigma,
                                    ngroups = 18, nobs_per_group = 8) {
  u0 <- rnorm(ngroups, 0, tau0)
  u0 <- rep(u0, each = nobs_per_group)
  u1 <- rnorm(ngroups, 0, tau1)
  u1 <- rep(u1, each = nobs_per_group)
  Subject <- factor(rep(1:ngroups, each = nobs_per_group))
  Days <- rep(0:(nobs_per_group - 1), each = ngroups)
  e <- rnorm(ngroups * nobs_per_group, 0, sigma)
  Reaction <- beta0 + u0 + (beta1 + u1) * Days + e
  data.frame(Reaction, Days, Subject)
}
```

Second, we set up a function that fits the models of interest to a simulated dataset. We chose to focus on comparing the complete pooling model `fit_sleep1` and the varying intercept-slope models `fit_sleep3`:

```
# fit the models of interest to a simulated dataset
# the simulation index j is unused here but still provided
# as an argument to prevent it from being passed further
# returns a matrix of posterior draws of b1 with
# one column per fitted model and one row per posterior draw
sim_sleepstudy_models <- function(j, ...) {
  dat <- sim_sleepstudy_data(...)
```

```

fit1 <- update(fit_sleep1, newdata = dat, chains = 1)
draws1 <- as.data.frame(fit1, variable = "b_Days")$b_Days

fit2 <- update(fit_sleep3, newdata = dat, chains = 1)
draws2 <- as.data.frame(fit2, variable = "b_Days")$b_Days

return(cbind(draws1, draws2))
}

```

Third, we will actually run our simulations. Since, for each simulated dataset, we are fitting multiple Bayesian models, this may take a while. It is thus highly beneficial to parallelize the code in order to run multiple simulation trials at the same time. There are many ways to set up parallel code in R but I personally prefer using the `future` package as it abstracts away a lot of the complications that arise in parallel computing.

```

library(future)
library(future.apply)

```

We will run the simulations for 250 trials, that is, 250 simulated datasets using 8 cpu cores (called `workers` here). When running these simulations on your own machine, make sure that you do not request more cpu cores than you have on your machine; ideally one or two cores less if you plan on doing anything else on your machine while the simulations run.

```

nsim <- 250
# set up a parallel environment to run futures in
plan(multisession, workers = 8)
# future_lapply is like lapply just with parallel execution
sim_results <- future_lapply(
  1:nsim, sim_sleepstudy_models,
  beta0 = 250, beta1 = 0, tau0 = 25,
  tau1 = 7, sigma = 25,
  future.seed = TRUE
)
# go back to sequential evaluation after finishing the simulations
plan(sequential)

```

Lastly we will transform and summarize the results. First, we will look at the histogram of the posterior quantile in which the true parameter value  $b_0^* = 0$  lies. That is, for each simulation trial, we compute how many (in percent) of the posterior draws are smaller than the true value:

```

# combine the results into an array
sim_results_array <- abind::abind(sim_results, along = 3)
# compute posterior quantiles of the true parameter value
true_beta1 <- 0
true_beta1_quantiles_1 <- colMeans(sim_results_array[, 1, ] < true_beta1)
true_beta1_quantiles_2 <- colMeans(sim_results_array[, 2, ] < true_beta1)

```

Under perfect frequentist calibration, the distribution of these quantiles across simulation trials should be uniform. Any deviation from uniformity beyond mere chance provides evidence for miscalibration of the posterior. So let's check out the histograms:

```

histogram(true_beta1_quantiles_1, bins = 10) +
  histogram(true_beta1_quantiles_2, bins = 10)

```

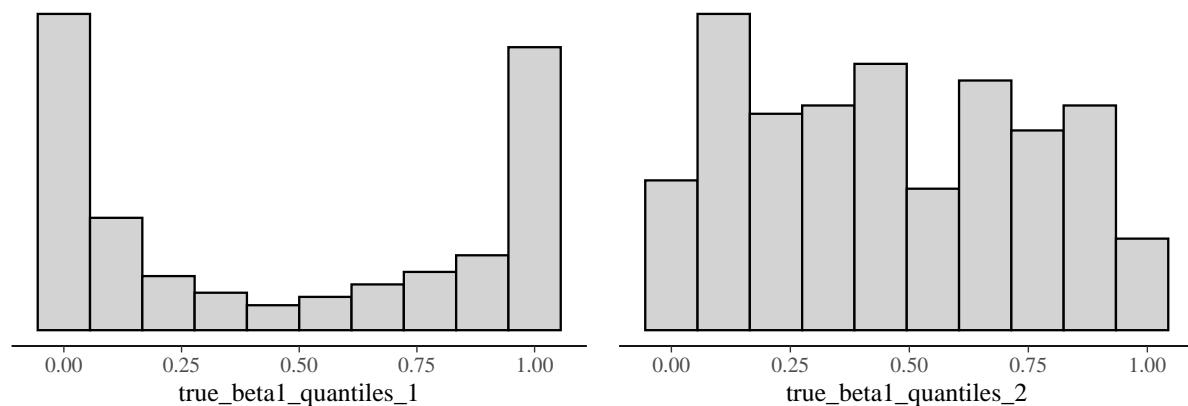


Figure 4.17: Histograms of the true-value posterior quantiles of slope parameter  $b_1$  for the complete pooling model (left) vs. the partial pooling model (varying intercepts and slopes; right). Under good calibration these histogram are expected to be approximately uniform.

On the left-hand side, for the complete pooling model, we see a strong deviation from uniformity in the way that very small quantiles and very high quantiles are more likely than would be expected under good calibration. This means that the posterior of  $b_1$  under this model is too narrow such that the true value of  $b_1^* = 0$  is often in the tails of the posterior. For the varying intercept-slope model, results look more uniform but it is a bit unclear if the variation we see indicate some systematic miscalibration or just random noise due to our limited simulation budget of 250 simulation trials and binning artifacts. We could add some confidence intervals for the histogram bars, but we will opt for a different method right away. Instead of histograms, we will plot the empirical cumulative distribution function (ECDF) of

the above computed quantiles. Under perfect frequentist calibration, the ECDF should resemble the CDF of a uniform distributions, that is, simply a diagonal line. If the deviation from uniformity is so large that is cannot be explained by chance, this points to some kind of miscalibration. For ECDFs we can obtain simultaneous confidence envelopes that take into account the ECDF curve as a whole (Säilynoja, Bürkner, and Vehtari 2022) and we have access to this kind of plot via the `bayesplot` package:

```
true_betas1 <- rep(0, nsim)
ppc_pit_ecdf(true_betas1, sim_results_array[, 1, ]) +
  ppc_pit_ecdf(true_betas1, sim_results_array[, 2, ])
```

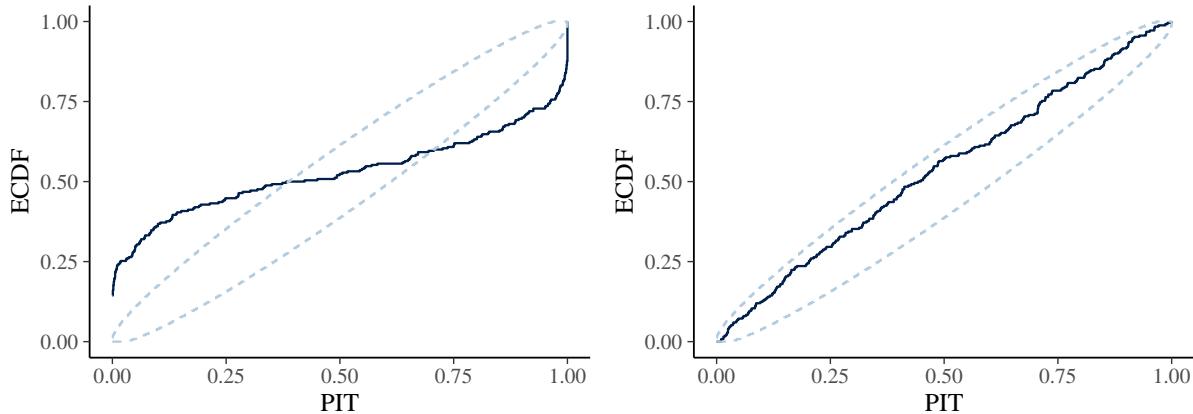


Figure 4.18: ECDF plots of the true-value posterior quantiles of slope parameter  $b_1$  for the complete pooling model (left) vs. the partial pooling model (varying intercepts and slopes; right). Under good calibration the ECDF line (dark blue) is expected to lie within the confidence envelope (light blue lines).

In the above ECDF plots, the dark blue lines indicates the ECDF and the line blue lines indicate the borders of the region within which the ECDF will highly likely fall (with probability 99% by default), if the underlying distribution was indeed uniform. Confirming what we have seen in the histograms, the complete pooling model is clearly miscalibrated for small and large parameter values, while the varying intercept-slope model shows overall good calibration.

These results are very closely related to the expected Type 1 error rates. Namely, if we see too many extreme true-value quantiles than would be expected under uniformity, we will have an inflated Type 1 error rate. To make this more concrete, suppose we would use central 95% credible intervals to indicate “significance”: If the value under the null hypothesis (here  $b_1^* = 0$ ) was outside of the credible interval, we would declare significance. To check how often this was the case in our simulations, all we have to do is to check how many true-value quantiles are smaller than 2.5% or larger than 97.5%. For the first model, we compute:

```
mean(true_beta1_quantiles_1 < 0.025 | true_beta1_quantiles_1 > 0.975)
```

```
[1] 0.484
```

That is, we have a Type 1 error rate of about 50%(!), while it should nominally be around only 5%. Such kind of extreme error rate inflations are not uncommon when ignoring multilevel structure and should serve as a clear warning for anyone ignoring or not sufficiently thinking about such structure in their data. In contrast, the same computation for the varying intercept-slope model yields an error rate of 0.04, which is very close to the nominal 5%. This is merely a sanity check since we build our simulations assuming said model to be true. If there was relevant signal in the data not captured by this model, it would also be miscalibrated of course. The point I want to make here was merely to showcase what kind of bad things can happen if you ignore relevant structure in your data.

## 4.9 No pooling

Clearly, we should not ignore multilevel structure in our data. But what happens if we just model a separate intercept and separate slope for each subject without a dependency inducing prior, almost as if we modeled each subject's data completely separately? This is what we refer to as *no pooling* model. Such a model will just require basic R formula syntax without any multilevel terms and we can specify it as follows:

```
fit_sleep4 <- brm(  
  Reaction ~ 0 + Subject + Subject:Days,  
  data = sleepstudy  
)
```

I avoid printing the summary here because it shows the intercepts and slopes for every subject rendering the output very long. The only thing that distinguishes this model from just a separate model per subject is that we estimate a single `sigma` parameter across subjects. Remember when we found small positive correlations between varying coefficients that I told you were due to partial pooling? Let's check out the posterior correlations again with the no pooling model.

```
vars_intercepts <- paste0("b_Subject", unique(sleepstudy$Subject))  
coef_sleep4 <- as.matrix(fit_sleep4, variable = vars_intercepts)  
cor(coef_sleep4[, 1:5])
```

Table 4.10: Posterior correlation of five subject-specific intercepts for model `fit_sleep4`.)

	b_Subject308	b_Subject309	b_Subject310	b_Subject330	b_Subject331
b_Subject308	1.00	-0.01	0.01	-0.01	0.00
b_Subject309	-0.01	1.00	-0.03	-0.02	-0.01
b_Subject310	0.01	-0.03	1.00	-0.02	0.00
b_Subject330	-0.01	-0.02	-0.02	1.00	0.01
b_Subject331	0.00	-0.01	0.00	0.01	1.00

As we can see from Table 4.10, virtually zero correlations, with the minimal deviations fully attributable to MCMC error. But there is another, arguably much more important effect of partial pooling on the coefficient estimates, namely that these priors induce *shrinkage* towards the overall mean. In Figure 4.19, we can see how shrinkage plays out for our models. Indeed we see noticeably differences between the no pooling and the partial pooling estimates, with differences that tend to be higher for subjects whose coefficients are further away from the overall mean. For some subject this shrinkage is actually quite substantial. For other subjects, close to the overall mean, there is barely any shrinkage at all. If we look at the plot in more detail, we see some curious details. For example, the two subjects' estimates on the lower left corner (small intercepts, small slopes) seem to be moving *further away* from the overall mean, at least in the Y-direction. That is, through partial pooling their slopes are estimated to be even smaller than in the no pooling model. This is admittedly unintuitive, but does not indicate a bug in our multilevel model, rather a misconception in the intuition. Partial pooling will induce *joint* shrinkage, that is shrinkage in the overall coefficient vector, which may mean that for some individual coefficients there is no shrinkage at all, or even a “reversed” shrinkage as is the case for the slopes of the two subjects in question.

What influence do the multilevel priors have on prediction? Usually, they will improve out-of-sample predictive performance compared to the respective no pooling models. The differences in predictive performance are larger if groups have few observations relative to the number of coefficients estimated per group. The lower the number of observations per group or the higher their number of coefficients, the more likely the no pooling model will overfit (i.e., treat noise as signal), thus reducing its out-of-sample predictive performance. In our case, we have 8 observations and only 2 coefficient per subject. In this case, the improvements in predictions induced via partial pooling are likely not that large. Let's check it out:

```
loo_sleep4 <- loo(fit_sleep4)
mmlloo_sleep4 <- loo_moment_match(fit_sleep4, loo = loo_sleep4)
loo_compare(mmlloo_sleep3, mmlloo_sleep4)
```

	elpd_diff	se_diff
fit_sleep3	0.0	0.0
fit_sleep4	-2.2	3.0

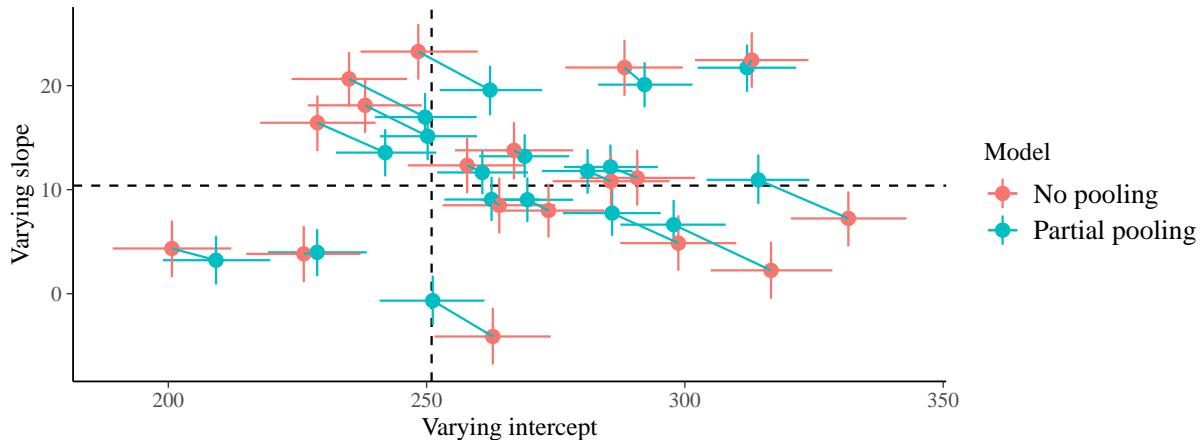


Figure 4.19: Varying coefficients of the no pooling model (red) vs. the partial pooling model (varying intercepts and slopes; blue). Points indicate posterior means and lines indicate central 50% credible intervals.

Indeed, while the partial pooling model provides a little better predictions in terms of ELPD difference, the corresponding standard error is of about the same size. Accordingly, the evidence for better out-of-sample predictions of the partial pooling vs. the no pooling model is small in our case. This would drastically change if we decreased the number of observations per subject. For example, when we fit the two models to a reduced dataset containing only data from Days 0, 2, 4, and 7 (i.e., four observations per subject), the LOO-CV performance of the partial pooling model becomes way better than that of the no pooling model. Try it out yourself as an exercise.

## 4.10 Models with more than two levels

So far, we have seen only how to handle a single grouping factor, but `brms` allows modeling, in principle, arbitrarily many of them. To illustrate this, let's pretend that the sleep study was conducted in three different sleep labs, each of which monitored 6 of the 18 subjects:

```
sleepstudy <- sleepstudy %>%
  mutate(Lab = rep(c("A", "B", "C"), each = 6 * 8))
```

We may not know exactly how the sleep labs differ from each other and how that might affect the subjects' performance during the study. One example for a potential mechanism driving such variation could be that the computer systems on which subjects perform the reaction time tasks differ across labs. If such a difference is a priori plausible, then, without accounting for the labs in the model, the assumption of a priori exchangeable of subjects is no longer justified. We may not know which lab leads to faster or slower reaction times. Yet, we cannot

arbitrarily change the order of the subjects anymore without changing our prior assumptions, since changing the order may imply subjects “switching labs”.

In an effort to restore a priori exchangeable, we decide to model the intercept as also varying across labs. Building on and extending the varying intercept-slope model from Section 4.5, the observation level likelihood mean  $\mu_n$  is then computed as

$$\mu_n = b_0 + b_1 x_n + \tilde{b}_{0j[n]} + \tilde{b}_{1j[n]} x_n + \tilde{\alpha}_{0l[n]}$$

where the lab-specific intercepts  $\tilde{\alpha}_{0l}$  have the usual zero-centered normal prior

$$\tilde{\alpha}_{0l} \sim \text{normal}(0, \gamma_0),$$

with standard deviation parameter  $\gamma_0$  also estimated from the data. In brms, we achieve this by simply adding `(1 | Lab)` to the model formula.

In many intro books and papers about hierarchical/multilevel models, you may see the more-than-two-level case being introduced a bit differently, by defining equations not only at observation level as done above, but also at all the other levels. For our example, this would mean not writing the varying coefficients of labs in the same equation as the coefficients of subjects, but rather writing the coefficients of subjects *as a function of* the coefficients of labs (since subjects are nested in levels). I have always struggled with this “multi-equation” formulation of multilevel models since the notation needed to adequately express it quickly blows up. Understanding that, due to linearity, all of those equations can be just written equivalently as a single equation, has made it much simpler for me to understand and reason about multi-level models. The “single-equation” formulation also generalizes much better to even more complicated setting as we will see in later chapters. It is used consistently in brms.

Back to our `sleepstudy` model: Since `Lab` has only three levels, its standard deviation parameter  $\gamma_0$  will be hard to estimate from data alone (think of estimating a standard deviation from just three data points). Accordingly, we should help the model a bit by setting an informative prior on  $\gamma_0$ . Let’s say, we assume the baseline reaction time differences between labs to be small, say, unlikely to be larger than 60 ms between the labs with the fastest and the lab with the slowest reaction times. If we interpret these 60 ms as the 95% uncertainty interval of the varying intercepts prior  $\text{normal}(0, \gamma_0)$  then this implies  $\gamma_0$  should not be larger than 15 ms, since roughly 95% of the values of a normal distribution lie in the interval [Mean - 2 SD, Mean + 2 SD]. Of course,  $\gamma_0$  might very well be smaller than 15 ms. We can encode such assumptions with various kind of priors but one straightforward choice is to use a half-normal prior with standard deviation equal to half the maximal expected  $\gamma_0$ :

$$\gamma_0 \sim \text{normal}_+(0, 7.5)$$

Putting it all together, results in the following brms model:

```

fit_sleep5 <- brm(
  Reaction ~ 1 + Days + (1 + Days | Subject) + (1 | Lab),
  data = sleepstudy,
  prior = prior(normal(0, 7.5), class = "sd", group = "Lab"),
  sample_prior = "yes"
)

```

```
summary(fit_sleep5)
```

Family: gaussian  
 Links: mu = identity  
 Formula: Reaction ~ 1 + Days + (1 + Days | Subject) + (1 | Lab)  
 Data: sleepstudy (Number of observations: 144)

Multilevel Hyperparameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept) ~Lab	(Number of levels: 3)	5.24	4.04	0.18	14.99	1.00	2758	1966	

~Subject (Number of levels: 18)

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept) ~Subject	34.04	8.12	21.07	52.33	1.00	1792	2579		
sd(Days) ~Subject	7.61	1.91	4.44	11.94	1.00	1409	2114		
cor(Intercept,Days) ~Subject	0.18	0.30	-0.41	0.75	1.00	1226	1877		

Regression Coefficients:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	267.82	9.92	248.23	287.61	1.00	1668	2235		
Days	11.45	2.11	7.37	15.64	1.00	1508	2095		

Further Distributional Parameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sigma	25.96	1.81	22.76	29.69	1.00	3099	3016		

We added the `sample_prior = "yes"` option to also obtain prior draws in addition to the posterior draws, which enables easy comparison of prior and posterior (see Figure 4.20):

```

vars_subject <- c("sd_Subject_Intercept", "prior_sd_Subject")
vars_lab <- c("sd_Lab_Intercept", "prior_sd_Lab")
mcmc_plot(fit_sleep5, type = "hist", variable = c(vars_subject, vars_lab))

```

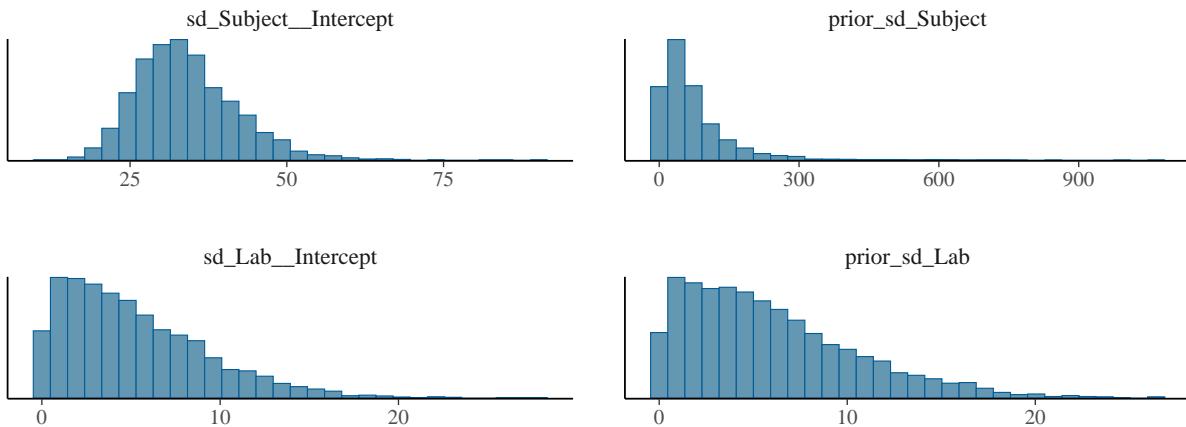


Figure 4.20: Posterior histograms of selected parameters for model `fit_sleep5`.

For the SD parameter  $\sigma_0$  of the subject-specific intercepts, we have implicitly used the default prior of brms, which is a half-Student-t prior with small degrees of freedom and data-adaptive scale. Comparing the prior and posterior of  $\sigma_0$ , we clearly see that we have learned a lot about  $\sigma_0$ . In comparison, there is very little difference in the prior and posterior of the SD parameter  $\gamma_0$  of the lab-specific intercepts. For once, we already started with a much more informative prior and second, we only have the intercepts of three labs (i.e., three “data points”) to inform the posterior. Had we used a more weakly informative prior, we would have seen stronger prior-posterior differences for  $\gamma_0$ , yet not nearly as much as for  $\sigma_0$ . I suggest you to try out different priors on  $\gamma_0$  and  $\sigma_0$  as an exercise.

As we see from the model formula above, it is syntactically not difficult to extend multilevel models to as many levels as you like. The decisions we have to make primarily involve (a) to identify which relevant grouping factors are present in the data and (b) to decide which of the predictors’ coefficients can and should be modeled as varying across which grouping factors. For the present model, we have made the choice to let the intercept vary across both subjects and labs while we allowed the slopes of the `Days` predictor to only vary across subjects. This is choice that may or may not be justified. We could have as well chosen to model the slopes of `Days` to vary across both subjects and labs or even only across labs. This is not to say that all of these choices are sensible, of course. In fact, there has been much debate about the process that leads us to decide which multilevel terms to include or to exclude, a topic I will come back to in Chapter 5.

## 4.11 Different covariance matrices by group

Let us expand our model in yet another direction by pretending that each subject belongs to either one of two groups: Those who like to sleep long and those who like to sleep short. Here is how I chose to assign subjects to these hypothetical two groups:

```

sleepstudy <- sleepstudy %>%
  mutate(
    Sleeper = ifelse(
      Subject %in% c(308, 331, 332, 337, 349, 350, 352, 370, 372),
      "long", "short"
    )
  )

```

It is likely that people who tend to sleep longer compared to shorter will have more trouble with sleep deprivation. We can express this statistically by modeling an interaction effect of `Days` and `Sleeper`, in addition to their individual main effects. The main effects and interaction can be conveniently written as `Days * Sleeper` in the model formula.

With the overall coefficients settled, we turn to the varying coefficients. The first question that may come to mind is whether it makes sense to model subject-specified coefficients of `Sleeper` as well? This would come down to the question of asking whether the habit of sleeping long vs. short affects different people differently. It is a potentially interesting question but cannot be answered by the present data since each subject is classified as either a long or a short sleeper. In other words the predictor variable is constant within each of the grouping levels. Accordingly, it is impossible to statistically infer any variation of the sleeper effects across subjects. If we want to put this into a single sentence to remember the principle behind it, it goes as follows: “If you want to model a predictor’s coefficient as varying *between* the levels of a grouping factor, the predictor’s values need to vary *within* said levels.”

So modeling varying coefficients for `Sleeper` is not possible for the given data. But this doesn’t mean `Sleeper` is necessarily irrelevant for the multilevel terms of our model. For example, it is conceivable that short sleepers are more similar to each other than long sleepers when it comes to how they handle sleep deprivation. In other words, we would expect a smaller SD of the varying slopes in the short vs. the long sleeper group. If we have reasons to believe something like this to be the case, we again are in a situation where a priori exchangeability of all subjects is no longer justified, as changing the subjects’ ordering may change their sleeper status in a way not fully accounted for by the overall coefficients. The formula syntax in brms allows to handle such a case, using the `gr` function on the right-hand side of multilevel terms. By replacing `Subject` in the `(Days | Subject)` term with `gr(Subject, by = Sleeper)` we tell the model that we want the multilevel hyperparameters (standard deviations and correlations) to vary by the factors of the `by` variable, here by the `Sleeper` categories:

```

fit_sleep6 <- brm(
  Reaction ~ Days * Sleeper + (Days | gr(Subject, by = Sleeper)),
  data = sleepstudy
)

```

```
summary(fit_sleep6)
```

Family: gaussian  
Links: mu = identity  
Formula: Reaction ~ Days \* Sleeper + (Days | gr(Subject, by = Sleeper))  
Data: sleepstudy (Number of observations: 144)

Multilevel Hyperparameters:

~Subject (Number of levels: 18)

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat
sd(Intercept:Sleeperlong)	37.75	12.88	18.07	69.56	1.00
sd(Days:Sleeperlong)	8.52	3.33	3.86	16.71	1.00
sd(Intercept:Sleepershorts)	26.59	11.49	9.16	53.79	1.00
sd(Days:Sleepershorts)	6.95	2.89	2.62	14.02	1.00
cor(Intercept:Sleeperlong,Days:Sleeperlong)	-0.28	0.35	-0.82	0.50	1.00
cor(Intercept:Sleepershorts,Days:Sleepershorts)	0.42	0.40	-0.47	0.97	1.00
	Bulk_ESS	Tail_ESS			
sd(Intercept:Sleeperlong)	1518	2061			
sd(Days:Sleeperlong)	1565	2189			
sd(Intercept:Sleepershorts)	1936	2140			
sd(Days:Sleepershorts)	1681	2261			
cor(Intercept:Sleeperlong,Days:Sleeperlong)	1716	2382			
cor(Intercept:Sleepershorts,Days:Sleepershorts)	1637	2097			

Regression Coefficients:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	283.08	14.47	254.16	311.81	1.00	1792	1937
Days	14.56	3.30	8.00	21.06	1.00	1624	2091
Sleepershorts	-30.49	18.42	-67.49	6.16	1.00	1762	1989
Days:Sleepershorts	-6.10	4.41	-14.58	2.91	1.00	1792	2131

Further Distributional Parameters:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	25.98	1.79	22.72	29.78	1.00	2771	3055

Looking at the summary output, specifically, under “Multilevel Hyperparameters”, we see only little differences between the standard deviations and correlations of the two **Sleeper** groups. However, the overall interaction effect of **Days** and **Sleeper** as shown under “Regression Coefficients” is quite strongly negative. That is, according to the results, shorter sleepers seems to be less affected by sleep deprivation than longer sleepers. Please keep in mind that I added the **Sleeper** variable artificially to showcase some of the functionality of brms on this data example. So this is not an actual empirical finding.

## 4.12 Some benefits of multilevel models

I want to close this chapter by a short discussion and summary on the benefits of multilevel models, both in general and specifically when applied in a Bayesian framework. I will start with the general points.

For me, perhaps the most important benefit of multilevel models is that they help us better calibrate the uncertainty of the overall coefficients in the face of structured data. This point we have seen demonstrated among others in the small simulation study on Type 1 error rates. Further benefits include the convenient estimation of variation in different levels of the hierarchy, such that we can, for example, see how much variation in the data is attributable to variation between vs. within subjects. In scenarios where we care about the varying coefficients, additional benefits come into play such as their improved estimation accuracy through partially pooling information across groups. This also leads to better out-of-sample predictions especially in cases where the number of observations per group is small relative to the number of coefficients per group. Lastly, multilevel models allow us to easily predict the coefficients of new levels by using the posterior over the estimated varying coefficients as a prior.

In terms of Bayes-specific benefits – or at least benefits more often harvested in Bayesian modeling – I particularly see the greater modeling flexibility. This tends to allow the estimation of more complex multilevel models including more grouping factors and more varying coefficients without running into serious convergence issues. This is driven by two main aspects: Firstly, specifying priors on the multilevel hyperparameters can help stabilizing estimation and obtain reasonable posteriors even in the face of sparse data or over-parameterized models where the data alone are insufficient to inform the parameters. Secondly, advanced sampling algorithms such as the MCMC methods implemented in Stan tend to be more powerful than optimization-based approaches in obtaining a good presentation of the posterior (or any form of frequentist equivalent). The use of sampling algorithms also enables the joint estimation of overall and varying coefficients, something that is much harder to achieve in optimization-based approaches. The price of sampling is usually an increase of estimation time by an order of magnitude or more compared to optimization. That said, there are algorithms that bridge the gap between optimization and sampling, including integrated laplace approximation (INLA) (Rue, Martino, and Chopin 2009), automatic differentiation variational inference (ADVI) (Kucukelbir et al. 2017), and pathfinder (Zhang et al. 2022). Specifically INLA is well suited for estimating Bayesian multilevel models both accurately and efficiently and you should give it a try if your model is too slow to estimate in brms or Stan.

# 5 Generalized linear multilevel models

## 5.1 Setup

```
library(ggplot2)
library(dplyr)
library(tibble)
library(brms)
```

## 5.2 Introductions

Generalized linear multilevel models (GLMMs) are the combination of generalized linear models (Chapter 3) and linear multilevel models (Chapter 4): That is, they describe models that have a non-normal likelihood, usually combined with a non-linear link function, as well as linear predictors that contain multilevel terms. While conceptually a simple generalization, this opens up a lot of possibilities to express our data by means of statistical regression models. I don't have hard evidence for it, but my impression is that GLMMs are one of the primary reasons people use brms, especially because it provides a large range of model likelihoods. In this chapter, I will exemplify the application of GLMMs with brms in three case studies all requiring qualitatively different kinds of likelihoods and modeling considerations.

## 5.3 GLMMs for lower-bounded count responses

In multiple previous chapters, we have taken a look at the `epilepsy` dataset where we try to infer the efficacy of a treatment on the number of epileptic seizures. We had built models along the lines of `count ~ Trt * scale(log(Base))` to predict the seizure count by the treatment indicator `Trt` and the baseline number of seizures `Base` modeled on the log scale and standardized for easier interpretation of the treatment main effect. While we acknowledged the presence of multilevel structure in the form of repeated measurements over patients (four visits per patient), we previously didn't have the means to model said structure appropriately. But now we do: We start by adding a varying intercept over patients in a Poisson model with (default) log link:

```

fit_epi_pois1 <- brm(
  count ~ Trt * scale(log(Base)) + (1 | patient),
  data = epilepsy, family = poisson("log")
)

```

Multilevel terms in GLMMs follow the same syntax as for their linear counterparts, so if you have read the previous chapter, the above syntax should not come as a surprise. The only difference to the linear case is that we need to interpret the varying coefficients and their hyperparameters in light of the non-linear scale imposed by the link function. To make this more concrete, let's check out the model summary:

```
summary(fit_epi_pois1)
```

```

Family: poisson
Links: mu = log
Formula: count ~ Trt * scale(log(Base)) + (1 | patient)
Data: epilepsy (Number of observations: 236)

Multilevel Hyperparameters:
~patient (Number of levels: 59)
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)    0.55      0.06     0.44     0.69 1.01      959     1790

Regression Coefficients:
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept        1.80      0.11     1.58     2.02 1.00      798     1312
Trt1           -0.36      0.16    -0.67    -0.04 1.00      819      950
scalelogBase      0.66      0.11     0.45     0.87 1.00      875     1470
Trt1:scalelogBase  0.21      0.15    -0.09     0.53 1.01      861     1326

```

All terms, both overall and varying coefficients, are on the log scale as a result of using the log link. For example, varying intercept standard deviation `sd(Intercept)` with a posterior mean of 0.56 does *not* imply that patient vary on average by 0.56 epileptic seizures, which would indeed be quite little. Instead, it says that patients vary on average by 0.56 in *log* epileptic seizures, whatever that means. We would also say that, roughly (remember Jensen's inequality!), patients vary on average by *a factor of*  $\exp(0.56) \approx 1.75$  in their number of epileptic seizures. So nearly a factor of two difference between an average patient and a patient being one standard deviation above average, which is indeed quite a substantial variation. What does that imply to the uncertainty of our model predictions? We can plot the conditional effects of the `Base:Trt` interaction both without and with including variation attributable of the variation over patients (see Figure 5.1 and Figure 5.2):

```
conditional_effects(fit_epi_pois1, "Base:Trt")
```

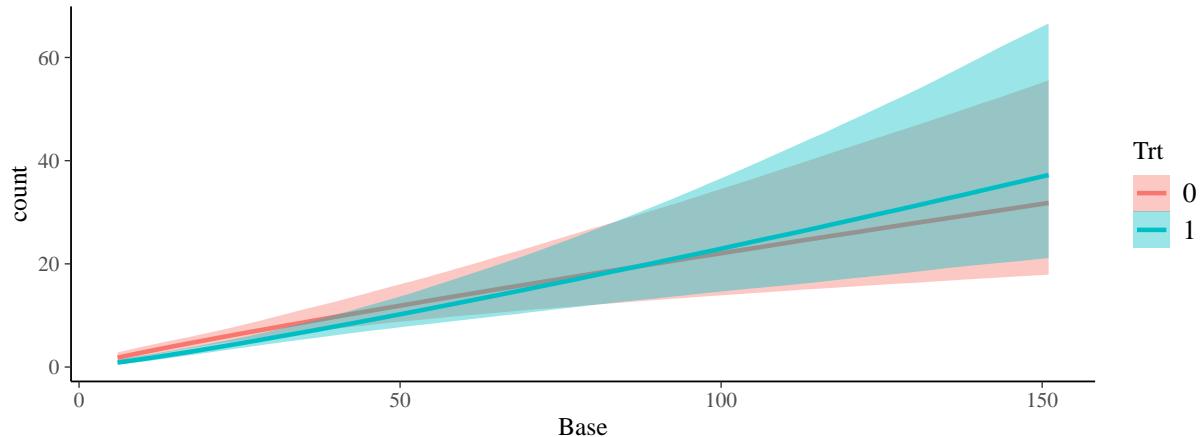


Figure 5.1: Conditional effects of model `fit_epi_pois1` excluding the uncertainty captured by multilevel terms.

```
conditional_effects(fit_epi_pois1, "Base:Trt", re_formula = NULL)
```

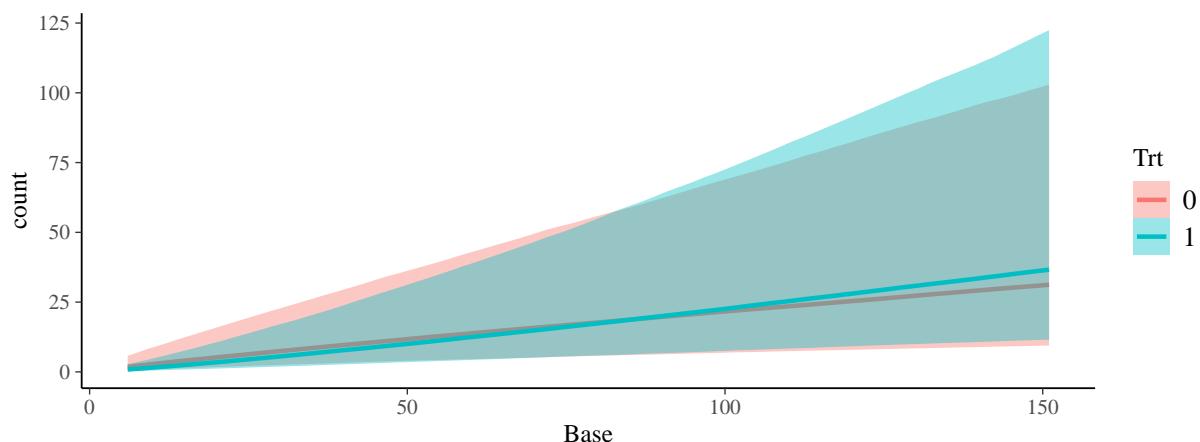


Figure 5.2: Conditional effects of model `fit_epi_pois1` including the uncertainty captured by multilevel terms.

Indeed, the between-person variation dwarfs any minor potential treatment effect the motivated researcher may try to identify in patients below a baseline of about 50 seizures.

### 5.3.1 Modeling overdispersion

In Chapter 3, we discussed how to model overdispersion in count data, for example, with a negative binomial likelihood. Back then, we saw clear evidence for the presence of overdispersion. But perhaps this was just because we hadn't modeled between-patient variation then? In other words, we may ask: Is there still evidence of overdispersion even when including varying intercepts over patients? That's easy to answer. We just take our GLMM above and replace the Poisson with the negative binomial likelihood. We keep using the log link as default.

```
fit_epi_negbin1 <- brm(  
  count ~ Trt * scale(log(Base)) + (1 | patient),  
  data = epilepsy, family = negbinomial()  
)
```

Performing model comparison via LOO-CV tells us that clearly, there is still overdispersion present and that we should favor the negative binomial model over Poisson:

```
loo_epi_pois1 <- loo(fit_epi_pois1)  
loo_epi_negbin1 <- loo(fit_epi_negbin1)  
loo_compare(loo_epi_pois1, loo_epi_negbin1)
```

	elpd_diff	se_diff
fit_epi_negbin1	0.0	0.0
fit_epi_pois1	-57.1	22.9

The substantive conclusions in terms of treatment efficacy remain the same also for the negative-binomial model. Accordingly, I don't show the model summary etc. here. But you can of course fit the model yourself and look at all the results in detail.

Now that we have access to multilevel terms, I can show you another approach to handling overdispersion in models with likelihoods that have no natural overdispersion parameter, such as Poisson or binomial: We define residuals ourselves via observation-level varying intercepts. With observation-level I mean that *every observation gets its own varying intercept*. To achieve this in brms, we first define a new grouping variable that goes from 1 to the number of rows in the data.

```
epilepsy$obs <- 1:nrow(epilepsy)
```

I call this variable `obs` here but you are free to name it however you like. Now, we add a varying intercept over `obs` to our above Poisson GLMM:

```

fit_epi_pois2 <- brm(
  count ~ Trt * scale(log(Base)) + (1 | patient) + (1|obs),
  data = epilepsy, family = poisson()
)

```

For brms, there is nothing special about the `(1|obs)` term. It is just another multilevel term like every other. Only the fact that we have one group per observations makes this behave like a latent residual. Since we have additionally included the `(1 | patient)` term, we can interpret `(1|obs)` as expressing *within-person* variation beyond what is naturally implied by the equidispersion assumption of Poisson. What does the summary say about our multilevel terms?

```
summary(fit_epi_pois2)
```

```

Family: poisson
Links: mu = log
Formula: count ~ Trt * scale(log(Base)) + (1 | patient) + (1 | obs)
Data: epilepsy (Number of observations: 236)

Multilevel Hyperparameters:
~obs (Number of levels: 236)
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept) 0.37 0.04 0.29 0.46 1.00 1455 2349

~patient (Number of levels: 59)
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept) 0.52 0.07 0.39 0.67 1.00 1638 2537

Regression Coefficients:
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept 1.75 0.11 1.53 1.97 1.00 2620 2697
Trt1 -0.36 0.16 -0.67 -0.05 1.00 2961 3093
scalelogBase 0.66 0.11 0.45 0.87 1.00 2171 2345
Trt1:scalelogBase 0.21 0.16 -0.10 0.53 1.00 2217 2735

```

First of all, convergence looks good across the board. This may seem surprising if we acknowledge the fact we have just fitted a model with  $4 + 59 + 236 = 299$  parameters on a dataset with only 236 observations. But it is just another example of how priors (here the hierarchical normal priors on the multilevel terms) enable us to fit models that would be impossible to identify from data alone. It didn't even fit much slower than the negative binomial model despite

having 235 more parameters, thanks to the magic of Hamiltonian Monte Carlo implemented in Stan.

One advantage of the Poisson model with latent residuals over the negative binomial model is that in the former, we can easily compare the residual variation with the variation explained by other terms in the model. For example, comparing the standard deviations of the `patient` vs. `obs` intercepts in the summary output, we see that there seems to be more variation between patients than within patients. We may also ask the `hypothesis` function to provide us with the variance ratio of the two terms:

```
hyp <- "sd_patient__Intercept^2 / sd_obs__Intercept^2 = 0"
(hyp <- hypothesis(fit_ephi_pois2, hyp, class = NULL))
```

```
Hypothesis Tests for class :
Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob Star
1 (sd_patient__Inte... = 0     2.07      0.83     0.89     4.16      NA      NA   *
```

```
plot(hyp, chars = NULL)
```

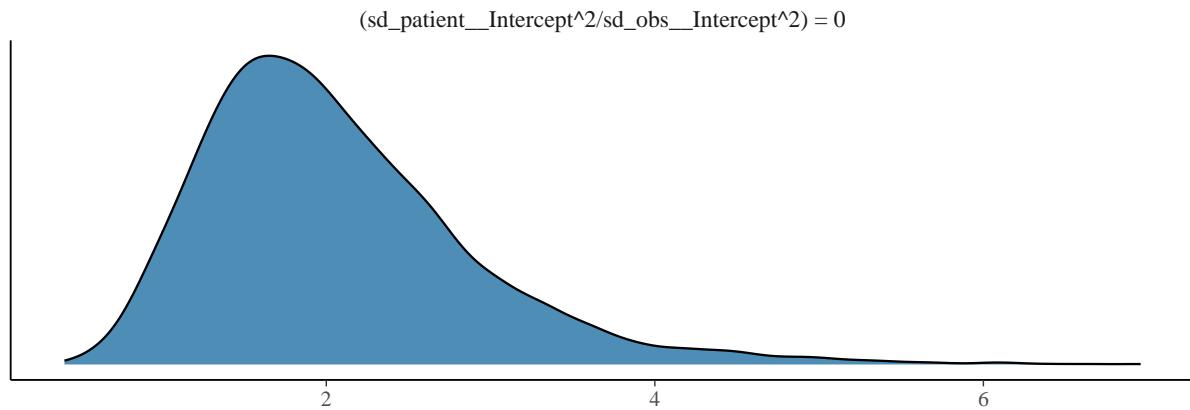


Figure 5.3: Posterior of the ratio of the between vs. within-person variances in model `fit_ephi_pois2`.

Accordingly, with high probability and after accounting for treatment and baseline effects, the overdispersion attributable to variance between patients is roughly 1 to 4 times as large as the corresponding variance within patients. Getting a similar estimate out of the overdispersion `shape` parameter of negative binomial models would be a much more cumbersome procedure.

With the Poisson model leading in interpretability, we may ask which of the two overdispersion models are to be preferred in terms of predictive performance. First, let us run LOO-CV for our new Poisson model:

```
(loo_epis2 <- loo(fit_epis2))
```

```
Warning: Found 55 observations with a pareto_k > 0.7 in model 'fit_epis2'. We recommend to set 'moment_match = TRUE' in order to perform moment matching for problematic observations.
```

```
Computed from 4000 by 236 log-likelihood matrix.
```

	Estimate	SE
elpd_loo	-594.2	13.9
p_loo	106.5	6.9
looic	1188.4	27.8

-----

```
MCSE of elpd_loo is NA.
```

```
MCSE and ESS estimates assume MCMC draws (r_eff in [0.4, 1.9]).
```

```
Pareto k diagnostic values:
```

		Count	Pct.	Min.	ESS
(-Inf, 0.7]	(good)	181	76.7%	122	
(0.7, 1]	(bad)	45	19.1%	<NA>	
(1, Inf)	(very bad)	10	4.2%	<NA>	

```
See help('pareto-k-diagnostic') for details.
```

There are a lot of too high Pareto k estimates. What has happened? Remember that in LOO-CV we each time leave out a single observation to be then predicted by a model trained on all other observations. PSIS LOO-CV attempts to approximate the required LOO posteriors from the full posterior via importance sampling. This may fail if the LOO posteriors are too different from the full posterior. In our Poisson model with latent residuals, every observation has its own parameter in the form of its varying intercept representing its residual. Accordingly, the LOO posterior not only has one observation less than the full posterior, it also has one parameter less. Or, to put it differently, if all parameters were included in the LOO model, one parameter would only be informed by the prior since it has no corresponding data. Without going further into the mathematical details, let me just say that importance sampling generally struggles a lot if asked to bridge the gap between two models where some parameters are not informed by data in the target sub-model. Also adding moment matching will not fix this satisfactorily. As I am writing this chapter, we are still actively working on methods to solve this challenge, but I am confident that we will soon have some viable solutions.

We may still choose to compare the two overdispersed models for educational purposes:

```
loo_compare(loo_epi_pois2, loo_epi_negbin1)
```

	elpd_diff	se_diff
fit_epi_pois2	0.0	0.0
fit_epi_negbin1	-19.8	4.6

If we didn't know about the warnings, we would conclude that, clearly, the overdispersed Poisson model makes better LOO predictions. But this huge difference is just an artifact of the failed PSIS approximation. If we compute exact LOO-CV for all the problematic observations, we see a different picture. Below you can find the code to do it rather "efficiently" using parallel computation, but it will still take a while even for this simple model. For larger models with lots of high Pareto k estimates, exact LOO-CV via `reloo` is unlikely to be computationally feasible.

```
library(future)
plan(multisession, workers = 4)
reloo_epio_pois2 <- reloo(fit_epi_pois2, loo = loo_epi_pois2, chains = 1)
plan(sequential)
```

Checking out the new LOO output confirms that all Pareto k issues have been resolved:

```
reloo_epio_pois2
```

Computed from 4000 by 236 log-likelihood matrix.

	Estimate	SE
elpd_loo	-608.8	16.4
p_loo	121.1	9.7
looic	1217.6	32.9
-----		
MCSE of elpd_loo	is 0.9.	
MCSE and ESS estimates assume MCMC draws (r_eff in [0.4, 1.9]).		
All Pareto k estimates are good (k < 0.7).		
See help('pareto-k-diagnostic') for details.		

Model comparison now indicates only a small advantage of the overdispersed Poisson vs. the negative binomial model:

```
loo_compare(reloo_epi_pois2, loo_epi_negbin1)
```

	elpd_diff	se_diff
fit_epi_pois2	0.0	0.0
fit_epi_negbin1	-5.3	1.4

As for general preferences, I tend to prefer using negative binomial models over overdispersed Poisson not only because the former are easily compatible with PSIS LOO-CV, but also because the fitted model objects are a lot smaller since they don't need to store posterior draws of  $N$  additional parameters. As  $N$  grows, this will lead to very large objects, quickly becoming several gigabytes large.

Most of the models we have been fitting so far in this book have been converging well and easily so. Let me show you a model that fails very badly. Suppose we accidentally kept the  $(1|obs)$  term when we switched to the negative binomial likelihood. Then, we account for the overdispersion twice. As a result, the involved parameters start "fighting" for the same information, which often leads to very bad sampling behavior:

```
fit_epi_negbin2 <- brm(  
  count ~ Trt * scale(log(Base)) + (1 | patient) + (1 | obs),  
  data = epilepsy, family = negbinomial(),  
  seed = 232  
)
```

Indeed, model fits very slowly and the summary output shows completely broken parameter estimates:

```
summary(fit_epi_negbin2)
```

Warning: Parts of the model have not converged (some Rhats are  $> 1.05$ ). Be careful when analysing the results! We recommend running more iterations and/or setting stronger priors.

Warning: There were 286 divergent transitions after warmup. Increasing adapt\_delta above 0.8 may help. See  
<http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup>

```
Family: negbinomial  
Links: mu = log  
Formula: count ~ Trt * scale(log(Base)) + (1 | patient) + (1 | obs)
```

```

Data: epilepsy (Number of observations: 236)

Multilevel Hyperparameters:
~obs (Number of levels: 236)
    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.17      0.08      0.06      0.38 2.17      11      30

~patient (Number of levels: 59)
    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.46      0.07      0.36      0.61 1.95       6       8

Regression Coefficients:
    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept          1.73      0.11      1.58      1.94 2.37       7      22
Trt1              -0.31      0.14     -0.56     -0.12 2.12       7      31
scalelogBase        0.69      0.09      0.52      0.80 1.76       7      35
Trt1:scalelogBase   0.13      0.09      0.01      0.40 2.82      10      32

Further Distributional Parameters:
    Estimate           Est.Error l-95% CI u-95% CI Rhat
shape 50454784694097928.00 86378875901532800.00      6.53 200044384732470752.00 8.63
          Bulk_ESS Tail_ESS
shape        4         NA

```

This is true in particular for the `shape` shape parameter: It is extremely large indicating that the negative binomial likelihood is estimated to be essentially equal to Poisson, which of course does make sense in this case. A look at the marginal posterior and trace plots reveal more interesting patterns:

```
plot(fit_epi_negbin2, variable = c("^sd_ ", "shape"), regex = TRUE)
```

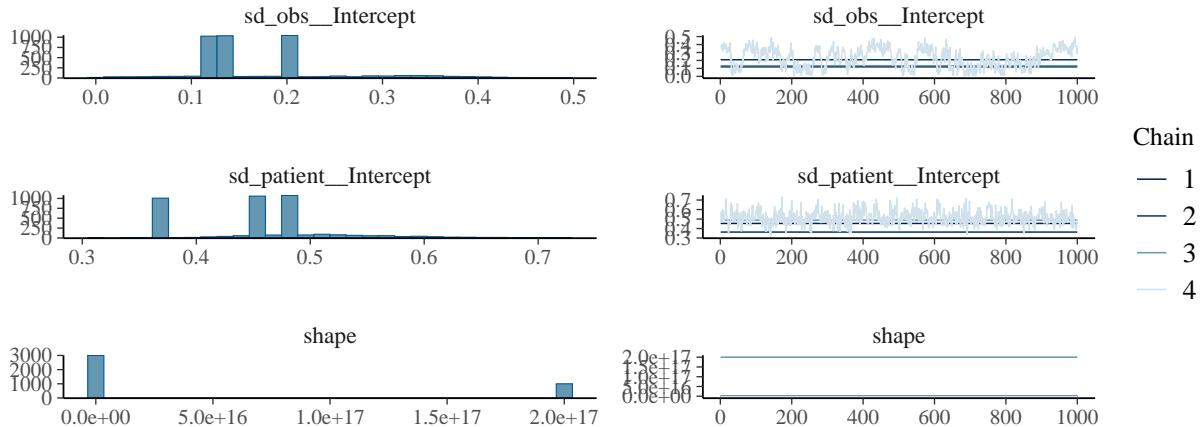


Figure 5.4: Marginal posterior plots of model `fit_epi_negbin2`. Left: marginal histograms. Right: trace plots. The plots indicate very strong convergence issues.

Apparently, for some chains, the  $(1|obs)$  term “wins” leading to extremely large posterior draws for `shape`, while for other chains `shape` “wins” leading to `sd(Intercept)` draws that are essentially zero. It seems that in fact only a single chain has mixed properly at all. To be fair, this is an example of extreme failure of convergence but something not unusual for models that are not properly identified.

### 5.3.2 Adding varying slopes

Before we complete the analysis of the epilepsy data, let us fit one last model, in which we also include varying slopes over patients. Ideally, we would like to know the the treatment effect varies over patients. However, since in the available data a patient either received the new treatment or not, we cannot model if there are varying treatment effects. This is not to say there are no varying treatment effects or even that we truly assume there are none. It simply means that we do not have the necessary information in the data to learn about such potential effects. For the same reason, we cannot model varying slopes for the `Base` predictor either. That said, there is a relevant question to be asked of this data that calls for the modeling of varying slopes, namely whether the severity of epilepsy changes over time – within the time frame of the study at least. Remember that each patient was coming in for 4 sessions (`visit` 1 to 4) so we do have some longitudinal data available to track each patient’s development. For the sake of simplicity, and to not immediately blow up the number of varying slopes, we will model `visit` as a numeric variable from 1 to 4 and call this numeric variable `nvisit`:

```
epilepsy$nvisit <- as.numeric(epilepsy$visit)
```

It is not only plausible that epilepsy severity changes differently over time for different patients (justifying the varying slopes of `nvisit` over patients) but also that the treatment effect in-

teractions with `nvisit`. For example, it is conceivable that the treatment only starts working properly after a few weeks or even months such that we could clearly see its effect only in the later sessions. These thoughts lead us to the model below:

```
fit_epi_negbin3 <- brm(
  count ~ Trt * (scale(log(Base)) + nvisit) + (1 + nvisit | patient),
  data = epilepsy, family = negbinomial()
)
```

Just looking at the `conditional_effects` of the treatment-visit interaction clarifies that neither is there any relevant main effect of `nvisit` nor is there any interaction with the treatment effect (see Figure 5.5).

```
conditional_effects(fit_epi_negbin3, "nvisit:Trt")
```

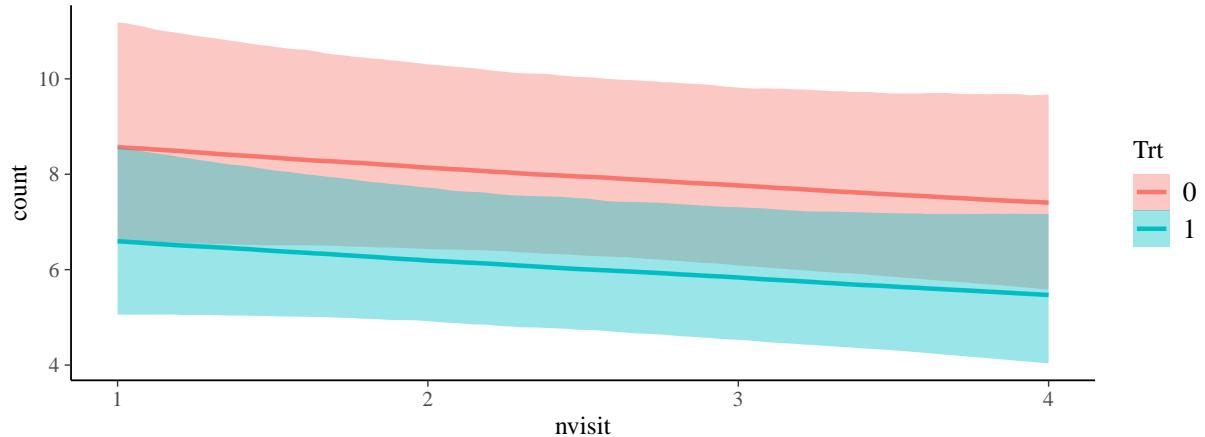


Figure 5.5: Conditional effects of model `fit_epi_negbin3`.

What is more, there seems to be very little variation across patients when it comes to changes over time, as we see from the marginal posterior histogram of the corresponding standard deviation parameter being “pushed” against zero (see Figure 5.6).

```
mcmc_plot(fit_epi_negbin3, variable = "sd_patient__nvisit", type = "hist")
```

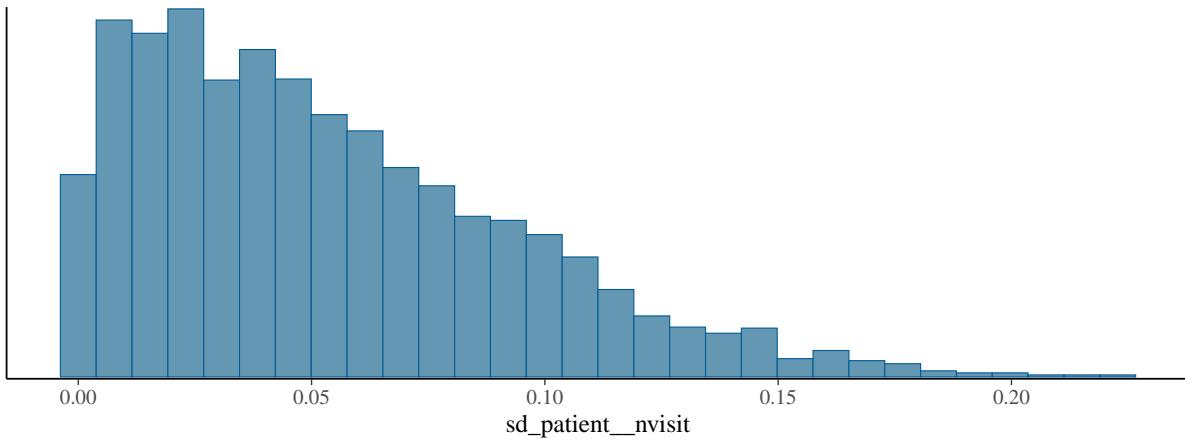


Figure 5.6: Posterior histogram of the between-patient standard deviation of model `fit_epi_negbin3`.

LOO-CV confirms our suspicion that there are also no improvements in predictive performance due to the modeling of the number of visits:

```
loo_epi_negbin3 <- loo(fit_epi_negbin3)
loo_compare(loo_epi_negbin1, loo_epi_negbin3)

      elpd_diff se_diff
fit_epi_negbin1  0.0      0.0
fit_epi_negbin3 -3.0      1.9
```

The predictive performance of the varying slope model `fit_epi_negbin3` even becomes slightly worse than the simpler model `fit_epi_negbin1`, clearly indicating that modeling visits leads to an unnecessarily complex model for the data at hand.

## 5.4 GLMMs for lower-bounded continuous responses

In the second case study, we will learn how to model lower-bounded continuous responses, specifically time-to-event data. In such data, we are measuring the time it takes until a certain event takes place. This could, for example, be the time until the failure of a machine or death due to an illness, a common case that has motivated the alternative name *survival analysis* for the field that deals with time-to-event data.

Here, we are going to analyse the time until the occurrence of a kidney disease. The dataset, originally included in McGilchrist and Aisbett (1991), describes the first and second recurrence

time (in months) of infection in kidney patients using portable dialysis equipment. In addition, information on the risk variables age, sex and disease type is provided. The `kidney` data is shipped directly with `brms` as one of its example datasets. Here is how the first few rows look:

```
head(kidney)
```

	time	censored	patient	recur	age	sex	disease
1	8	0		1	28	male	other
2	23	0		2	1	48	female
3	22	0		3	1	32	male
4	447	0		4	1	31	female
5	30	0		5	1	10	male
6	24	0		6	1	16	female

Our goal will be to predict the recurrence `time` as accurately as possible while taking into account the intricacies of the data structure. More on this in a bit. Let us first start with some theoretical basics of time-to-event modeling that provide necessary vocabulary to confidently work with such data.

#### 5.4.1 Basics of time-to-event modeling

One of the most important functions in time-to-event modeling is the *survival function*. For every time  $t \geq 0$ , it describes the probability for the event to happen later than  $t$ :

$$S(t) = p(t^* > t),$$

where  $t^*$  indicates the time of the event. Conversely, we can define the *lifetime distribution function* as the complement of the survival function:

$$F(t) = p(t^* \leq t) = 1 - S(t)$$

This function describes the probability that the event happened not later than  $t$ , that is, between 0 and  $t$ . As we know from the basics of Bayesian inference ([?@sec-bayesian-inference](#)), the density of a distribution is equal to the derivative of the distribution function. Accordingly, when we have  $F(t)$ , we can obtain the *event density*  $p(t)$  by taking the gradient of  $F(t)$  with respect to  $t$ :

$$p(t) = F'(t)$$

Having access to  $p(t)$  is important because it constitutes the main function we have to supply to Stan in order to express our Bayesian survival model. That said, there is one function that play an even bigger role in survival analysis, namely the *hazard function* defined as the ratio of even density and survival function:

$$h(t) = \frac{p(t)}{S(t)}$$

In words, the  $h(t)$  indicates the event rate at time  $t$  conditional on survival until time  $t$  or later, that is, conditional on the event not happening before time  $t$ . If we pretend for a second that we deal with a discrete measure of time, we may also say it as follows: The hazard function indicates the probability of the event happening at time  $t$  conditional on the probability that the event did not happen before time  $t$ .

I personally find the concept of an event density more intuitive than that of a hazard function, but I guess this is merely a matter of training. In any case, the four mentioned functions are equivalent in the sense that once we have specified one of the four functions, we have all the other functions specified too. For example, if we have the  $p(t)$ , we can first integrate it to obtain  $F(t)$ . Then we compute the complement of  $F(t)$  to obtain  $S(t)$  and finally we combine  $p(t)$  and  $S(t)$  to obtain  $h(t)$ . This works in all possible directions. In other words, we can equivalently specify a time-to-event model by defining any one of these four functions. While it is common in many areas of statistics to name models after the name of their likelihood densities, in survival analysis, models are often named after the hazard function they imply. I will follow this approach below when going through a set of increasingly complex time-to-event models.

#### 5.4.2 Constant hazard models

The arguably most simple time-to-event model assumes a *constant* hazard function. That is, intuitively speaking, it is assumed that the probability of the event happening at  $t$  conditional on the event not happening before is the same across all time points. It turns out that a constant hazards function implies the density of the exponential distribution:

$$p_{\text{exp}}(t) = \lambda \exp(-\lambda t)$$

The exponential distribution has a single *rate* parameter  $\lambda$  that is the inverse of the distributions mean  $\mu = 1/\lambda$ . Some examples of exponential distributions are shown in Figure 5.7:

When computing the hazard function implied by  $p_{\text{exp}}(t)$ , we see that the hazard is independent of time:

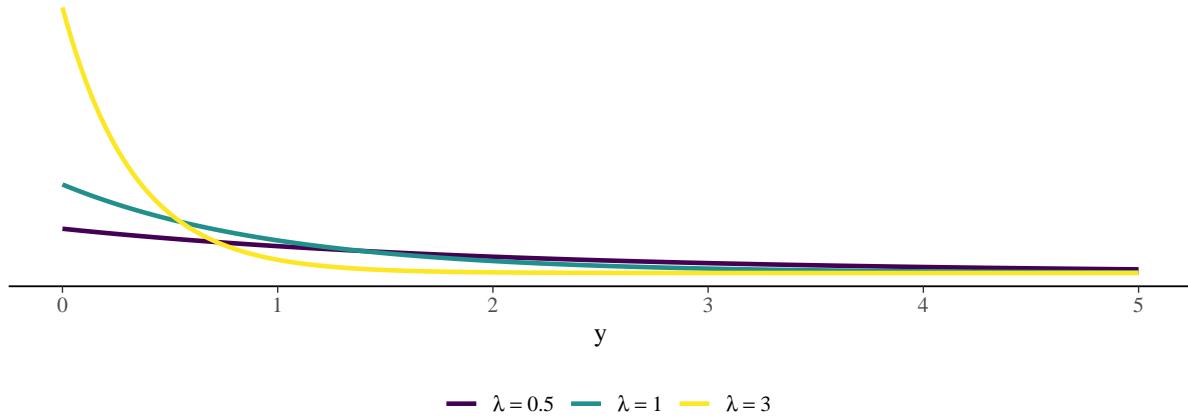


Figure 5.7: Exemplary densities of the exponential distribution with rate parameter  $\lambda$ .

$$h_{\text{exp}}(t) = \frac{\lambda \exp(-\lambda t)}{\exp(-\lambda t)} = \lambda$$

Hence the name “constant hazard model” is indeed justified when using an exponential likelihood for our event times.

In brms, we model the mean  $\mu = 1/\lambda$  to follow the convention that, if possible, the main parameter of a brms family (always called  $\mu$ ) should ideally be the mean or at least some measure of central tendency. As a result, higher predictor term values will imply *larger event times* rather than *larger hazard*. The latter would imply smaller event times instead. Since the event times can only be positive, so will their mean always be positive. To accomodate this in brms, all the time-to-event distributions, including `exponential`, use `log` as their default link. We will apply the `log` link in all the models of this case study.

Back to our `kidney` data, we will predict the time until recurrence of kidney diseases, by the `sex` (treated as binary), `age` (in years), and `disease` type. The latter is a categorical variable with four levels: `GN` (glomerulo nephritis), `PKD` (polycystic kidney disease), `AN` (I assume it refers to adenovirus nephritis), as well as `other`. What is more, since we have observed two recurrance times for all patients, we model the resulting dependency via a varying intercept term. Together, this leads us to the following model:

```
fit_kidney1 <- brm(time ~ age + sex + disease + (1 | patient),
                      data = kidney, family = exponential("log"))
```

```
summary(fit_kidney1)
```

```
Family: exponential
Links: mu = log
```

```

Formula: time ~ age + sex + disease + (1 | patient)
Data: kidney (Number of observations: 76)

```

Multilevel Hyperparameters:

```

~patient (Number of levels: 38)

```

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept)	0.53	0.21	0.08	0.95	1.01	914	670

Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	3.84	0.57	2.74	4.96	1.00	3552	2854
age	-0.01	0.01	-0.03	0.02	1.00	1833	2449
sexfemale	1.29	0.37	0.57	1.99	1.00	3063	2996
diseaseGN	-0.11	0.46	-1.03	0.79	1.00	2245	2619
diseaseAN	-0.45	0.45	-1.36	0.44	1.00	1985	2168
diseasePKD	0.78	0.67	-0.51	2.11	1.00	1731	2480

It seems that the most important predictor of recurrence time is people's sex with women having clearly longer recurrence times. Concretely, based on the credible interval bounds of the coefficients, the recurrence time for women is between  $[\exp(0.5), \exp(2)] \approx [1.65, 7.39]$  times higher than for men. Remember that this is a multiplicative rather than additive relationship on the original response scale since we have assumed a log link. We can also interpret the other coefficients in a similar multiplicative fashion after exponentiation. We do not discuss this model further here since it is missing one crucial aspect that renders its results potentially misleading, as we will see next.

### 5.4.3 Censoring

When it comes to accurately describing the data's generating process, the above exponential model is not only limited due to its constant hazard assumption. Rather, it completely ignores another, very important aspect common in time-to-event data: *censoring*. In the data, each patient has two data points, one for the first and one for the second recurrence of their kidney disease. For many patients, the second recurrence did not occur until the end of the study period. This was certainly good for the patients but makes our life as statistical modelers harder. Now we have to decide how to best model this information. We could, of course, choose to remove the observations in question, but that wouldn't seem quite right. After all, we know at least until which point in time, the disease did not recur. We call this case *right censoring* since the actual value – here the actual second recurrence – must be located somewhere to the right of the value we have logged – here the time until the end of the study. Rather than excluding the censored data, or just ignoring that censoring even took place, as we have done initially, we should attempt to properly express this information.

Let's look at the situation mathematically. First, consider an observation  $i$  where the event actually took place within the study period at time  $t_i$ . We inform the model about such an observation by setting its likelihood contribution to the event density evaluated at  $t_i$ :  $p(t_i | \theta_i)$ . Here, I am using  $\theta_i$  to express all the terms and parameter directly affecting the likelihood of the  $i$ th observation. Speaking in general statistical terms,  $p(t | \theta)$  is the probability density function (PDF) of the assumed likelihood distribution.

If  $i$  is a right censored observation instead, the above expression is inadequate since the corresponding  $t_i$  no longer indicates the actual time-to-event, but rather that the event did not occur until  $t_i$ . In other words, the event must occur sometime *after*  $t_i$ . We can express this mathematically by writing

$$p(t > t_i | \theta_i) = \int_{t_i}^{\infty} p(t | \theta_i) = S(t_i | \theta_i) = 1 - F(t_i | \theta_i),$$

where  $S(t_i | \theta_i)$  and  $F(t | \theta_i)$  are the survival and lifetime distribution functions, respectively. Speaking in general statistical terms,  $F(t | \theta)$  is the cumulative distribution function (CDF) and  $S(t | \theta)$  is the complement of the CDF (CCDF) belonging to the PDF  $p(t | \theta)$ . For most common distributions, their (C)CDFs are easy evaluate, either because they are analytic or because we have efficient algorithms for fast and accurate approximations. And Stan also has all these functions readily implemented, which allows brms to handle censored data efficiently.

In the `kidney` data, there is a variable called `censored`, which indicates if a disease recurrence actually took place during the study (`censored = 0`) or whether the recurrence was right censored

(`censored = 1`). For all censored observations, the logged `time` value was just the time until the end of the study. In brms, we will express censoring information using the `cens` response addition term. As a reminder, we express such addition terms in the formula on the left-hand side of the tilde `~` separated from the response variable by a bar `|`. That is, for our data, the formula has the general form `time | cens(censored) ~ <predictors>`. On the right-hand side of the formula, we can still add any kind of terms we want. Here is how our right-censored exponential model looks for the `kidney` data:

```
fit_kidney2 <- brm(
  time | cens(censored) ~ age + sex + disease + (1 | patient),
  data = kidney, family = exponential()
)
```

```
summary(fit_kidney2)
```

```
Family: exponential
Links: mu = log
```

```
Formula: time | cens(censored) ~ age + sex + disease + (1 | patient)
Data: kidney (Number of observations: 76)
```

Multilevel Hyperparameters:

```
~patient (Number of levels: 38)
```

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept)	0.45	0.24	0.03	0.91	1.00	887	1044

Regression Coefficients:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	3.76	0.57	2.69	4.89	1.00	3930	3005
age	-0.00	0.01	-0.03	0.02	1.00	2144	2626
sexfemale	1.62	0.38	0.87	2.36	1.00	3138	2853
diseaseGN	-0.10	0.48	-1.06	0.85	1.00	2202	2347
diseaseAN	-0.50	0.48	-1.42	0.47	1.00	1982	2413
diseasePKD	1.09	0.72	-0.33	2.55	1.00	2315	2558

Overall, for the present data, the results happen to be similar than for the non-censored model. This is not something we can rely on in general though. So always model censored data appropriately! One small difference we do see is that now the coefficient of `sex` is even larger, predicting between  $[\exp(0.86), \exp(2.38)] \approx [2.36, 10.80]$  times longer recurrence times for women than for men.

If we want to investigate graphically how well censored models fit to the data, things get a little tricky. When using standard posterior-predictive checks, as the one shown in Figure 5.8, censored observations will be excluded. This is because censored responses are not valid to compare against the model's predictions which are always uncensored.

```
pp_check(fit_kidney2)
```

Warning: Censored responses are not included.

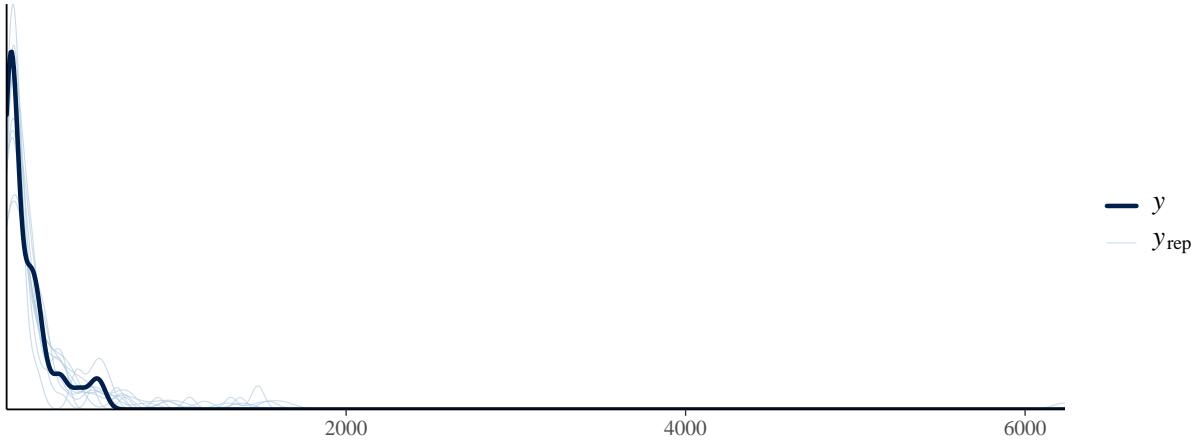


Figure 5.8: Posterior predictive check (marginal densities) of model `fit_kidney2`.

The `kidney` data has about 24% censored observations, too much to safely ignore in our checks. Instead, we should be using special types of posterior-predictive checks that can handle (right) censored data by design. The "km\_overlay" type plots an estimate of the survival function of the observed data vs. several survival functions implied by the model's predictions. The term "km" stands for Kaplan-Meier estimate (Kaplan and Meier 1958), which is a common method to estimate survival functions in the face right censored data. For the "km\_overlay" type, brms ensures internally that the censored data is treated correctly, so we can simply run:

```
pp_check(fit_kidney2, type = "km_overlay")
```

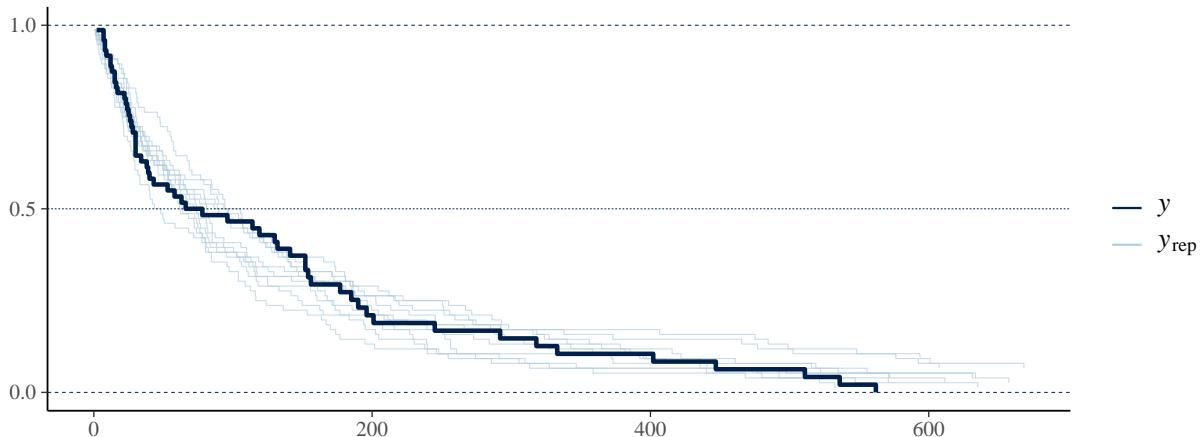


Figure 5.9: Posterior predictive check (marginal survival function) of model `fit_kidney2`. Right censoring of the observed data was taken into account.

The result shown in Figure 5.9 indicates a good fit of our censored exponential model to the

`kidney` data, at least visually. Just as a demonstration, how would the check look if `pp_check` would treat all observations as uncensored? Let's find out by creating the plot manually:

```
y <- kidney$time
yrep <- posterior_predict(fit_kidney2, ndraws = 10)
# status_y = 1 stands for an event and status_y = 0 stands for right censored
status_y <- rep(1, length(y))
bayesplot::ppc_km_overlay(y, yrep, status_y = status_y)
```

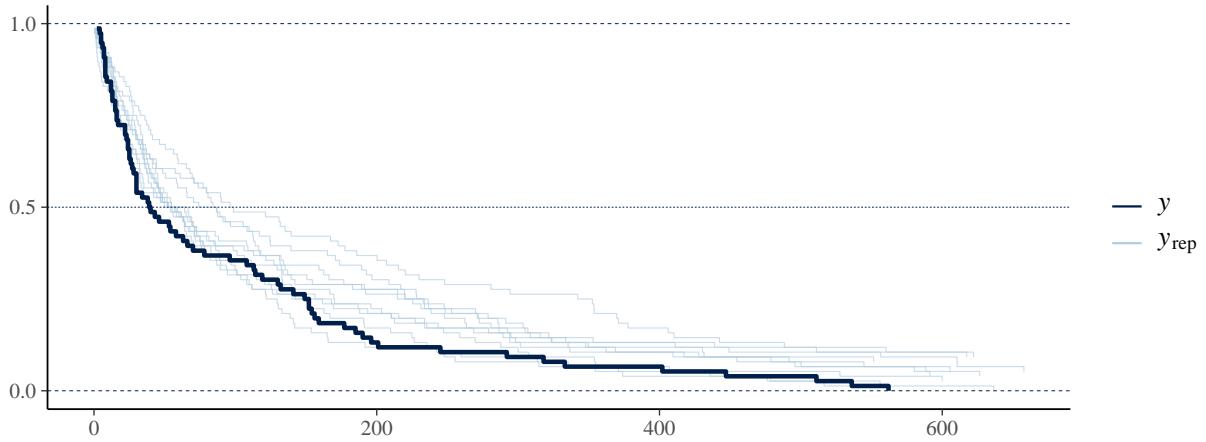


Figure 5.10: Posterior predictive check (marginal survival function) of model `fit_kidney2`. Right censoring of the observed data was ignored.

As we see from Figure 5.10, the posterior predictive check ignoring censoring would indicate that the model tends to overestimate the times until recurrence of the kidney diseases. But we know this is just an artifact of not handling censored data properly.

#### 5.4.4 Varying hazards models

As mentioned earlier, the constant hazard assumption is very restrictive, usually too restrictive to be true for most real-world data. To increase the flexibility, we may assume that the hazard is non-constant in a monotonic way. That is, it may accelerate (or decelerate) over time, such that the hazard may be higher (or lower) for later than for earlier points in time. One distribution that allows for such monotonically changing hazard functions is the Weibull distribution, visualized in Figure 5.11.

The hazard function of the Weibull distribution evaluates to

$$h_{\text{weibull}}(t) = \lambda\alpha (\lambda t)^{\alpha-1},$$

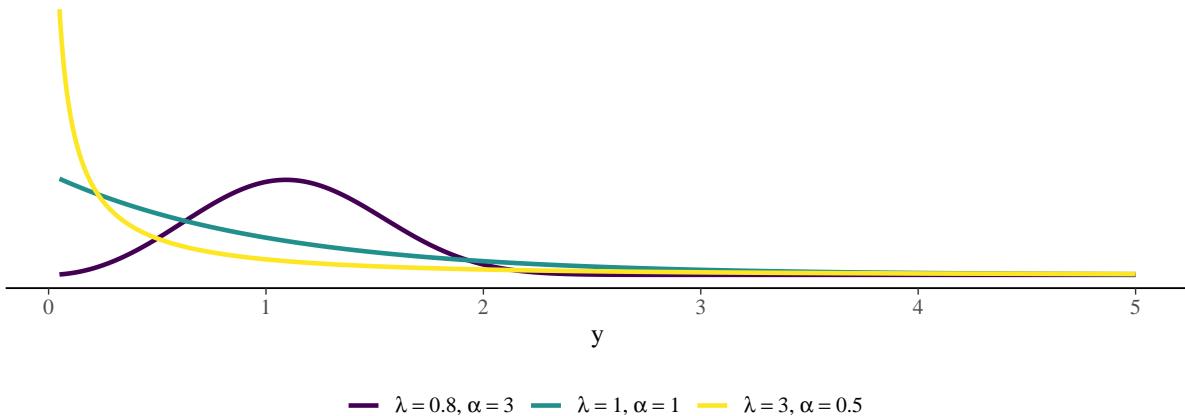


Figure 5.11: Exemplary densities of the Weibull distribution with shape  $\alpha$  and rate  $\lambda$ .

where  $\alpha > 0$  is a shape and  $\lambda > 0$  is a rate parameter. If  $\alpha > 1$ , the hazard increases over time and if  $\alpha < 1$ , the hazard decreases over time. In the special case of  $\alpha = 1$ , the hazard function simplifies to the constant case  $h_{\text{weibull}}(t) = \lambda$ . From this we see that the exponential distribution is in fact a special case of the Weibull distribution. In brms, we don't predict the rate  $\lambda$  but rather the mean  $\mu$ . The parameters relate to each other via the equation  $\lambda = \Gamma(1 + 1/\alpha)/\mu$ . As a regular user, you do not have to worry about such mathematical details. All you need to know is that the main parameter to predicted is the mean and that the shape  $\alpha$  controls the acceleration or deceleration of the hazard rate.

We are now ready to fit a censored Weibull model to the `kidney` data:

```
fit_kidney3 <- brm(
  time | cens(censored) ~ age + sex + disease + (1 | patient),
  data = kidney, family = weibull()
)
```

```
summary(fit_kidney3)
```

```
Family: weibull
Links: mu = log
Formula: time | cens(censored) ~ age + sex + disease + (1 | patient)
Data: kidney (Number of observations: 76)
```

```
Multilevel Hyperparameters:
~patient (Number of levels: 38)
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.55      0.24      0.05     1.00 1.00       718      1113
```

Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	3.74	0.57	2.65	4.87	1.00	3220	2942
age	-0.00	0.01	-0.03	0.02	1.00	2124	2054
sexfemale	1.62	0.39	0.86	2.36	1.00	2916	2839
diseaseGN	-0.13	0.49	-1.10	0.81	1.00	1890	2551
diseaseAN	-0.54	0.48	-1.45	0.46	1.00	1763	1873
diseasePKD	0.95	0.72	-0.50	2.37	1.00	1812	2121

Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
shape	1.12	0.16	0.84	1.46	1.00	1335	2404

As we can see from the summary output, the results are relatively similar to that of the censored exponential model. This makes sense since the `shape` parameter is estimated very close to 1 without much posterior uncertainty. Also the posterior predictive checks look very similar and I hence don't show them here. What does a LOO-CV tell us about the predictive performance of the two models?

```
loo_kidney2 <- loo(fit_kidney2)
loo_kidney3 <- loo(fit_kidney3)
loo_compare(loo_kidney2, loo_kidney3)
```

```
elpd_diff se_diff
fit_kidney3 0.0      0.0
fit_kidney2 -0.5     0.9
```

Indeed, little difference. In fact, if at all, the Weibull model even fits slightly worse, likely due to the fact that it has an additional parameter (`shape`) that does not improve predictive performance. So it seems that, for the given data and the considered linear predictor term, a constant hazard function is actually justified. We are not giving up just yet though and try another, even more complicated model in the next section to verify our conclusions.

#### 5.4.5 Proportional hazards models

Using Weibull models for survival analysis already provides some flexibility to model varying hazards, but we are still limited to either an increasing or decreasing (or constant) hazard rate over time, but not a mix of them. Of course, in reality, the hazard rate may change over time in various ways. In our kidney example, consider the hypothetical case that there is first a time frame right after recovery with high probability of the disease returning, followed by time where this is unlikely, only for the hazard rate to surge again after a couple of years.

Essentially, in this scenario, multiple overlaying mechanisms would drive the hazard up at different points in time. Our Weibull model could likely not account for such cases. We need something more flexible. This is where the Cox proportional hazards model (Cox 1972) comes into play. In this model, we assume that the hazard function is given by the product of a *time-dependent* baseline hazard function  $h_0(t)$  and our exponentiated predictor term  $\eta$  usually equipped with a log link such that  $\lambda = \exp(\eta)$ :

$$h_{\text{cox}}(t) = h_0(t) \lambda \quad (5.1)$$

If we took the ratio of  $h_{\text{cox}}(t)$  for two different  $\lambda$  values, we would see the time-dependent baseline hazard  $h_0(t)$  to cancel out. Hence the name *proportional hazards model*. For the same reason, when we estimate the Cox model with maximum likelihood, we can get around evaluating  $h_0(t)$  at all. However, for fully Bayesian estimation, we need the complete event density of this distribution:

$$p_{\text{cox}}(t) = h_{\text{cox}}(t) S_{\text{cox}}(t)$$

So we cannot just remove the baseline hazard somehow. In other words, we need to explicitly model  $h_0(t)$  and brms uses M-splines for this purpose. Without going into the details here, all you need to know is that the implementation ensures  $h_{\text{cox}}(t)$  being positive and easy to integrate such that obtaining  $S_{\text{cox}}(t)$  is easy too. We will learn more about splines in the context of non-linear modeling in Chapter 8.

```
fit_kidney4 <- brm(
  time | cens(censored) ~ age + sex + disease + (1 | patient),
  data = kidney, family = cox("log")
)
```

As opposed to the `exponential` and `weibull` families, in the `cox` family we are directly predicting the rate  $\lambda$  instead of the response distribution's mean. As a result, since the hazard is modeled to *increase* with increasing  $\lambda$  as per Equation 5.1, the mean time-to-event will *decrease* with increasing  $\lambda$ . For this reason, the signs of the regression coefficients are reversed in the `cox` model in comparison to the other models we have looked at thus far. Something we see clearly representdc in the summary output:

```
summary(fit_kidney4)

Family: cox
Links: mu = log
Formula: time | cens(censored) ~ age + sex + disease + (1 | patient)
Data: kidney (Number of observations: 76)
```

```

Multilevel Hyperparameters:
~patient (Number of levels: 38)
    Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.65      0.30     0.07    1.26 1.01      807     1389

Regression Coefficients:
    Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept       3.02      0.75     1.62    4.60 1.00      2429    2080
age            0.00      0.02    -0.03    0.04 1.00      2358    2508
sexfemale     -1.84      0.48    -2.83   -0.95 1.00      2130    2101
diseaseGN      0.11      0.56    -0.97    1.24 1.00      2049    1824
diseaseAN      0.55      0.56    -0.55    1.69 1.00      1729    1690
diseasePKD    -1.18      0.82    -2.83    0.44 1.00      1704    2376

```

For consistency within brms, I would have preferred to parameterize the `cox` family differently. However, for consistency with virtually every other survival modeling software, I decided it would be better to use the proportional hazards parameterization also in brms.

When it comes to post-processing, the `cox` family is a bit awkward. Due to how we model the baseline hazard, with a time-dependent function, sampling for the implied distribution is quite cumbersome and not implemented in brms at the time of writing this book. As a result, functions like `posterior_predict` and `pp_check` will not work with Cox models unfortunately.

However, `log_lik` works such that we can compute ELPD and perform things like PSIS-LOO-CV:

```

loo_kidney4 <- loo(fit_kidney4)
loo_compare(loo_kidney2, loo_kidney3, loo_kidney4)

```

	elpd_diff	se_diff
<code>fit_kidney4</code>	0.0	0.0
<code>fit_kidney3</code>	-1.1	1.3
<code>fit_kidney2</code>	-1.6	1.5

While the Cox model fits slightly better than its exponential and Weibull competitors, the difference is very small for the data at hand. It seems as if a constant hazard is indeed a reasonable assumption here. At least after accounting for our three predictors `age`, `sex`, and `disease` and modeling varying intercepts over patients.

## 5.5 GLMMs for double-bounded responses

As our third case study, we consider an example from item response theory (IRT), which concerns itself with the modeling of psychological measurements. I provide a detailed introduction to Bayesian IRT with brms in Bürkner (2021). Here, we will analyse the `VerbAgg` data used throughout De Boeck and Wilson (2004) and shipped with the `lme4` package (Bates et al. 2015):

```
data("VerbAgg", package = "lme4")
```

This data set contains responses of 316 participants on 24 items of a questionnaire on verbal aggression. For each `item`, participants (variable `id`) answered whether they would engage in certain verbally aggressive thoughts or behaviors. The ordinal response options (variable `resp`) were "yes", "perhaps" or "no". To make the modeling simpler for now, we will focus on the binary response variable `r2`, which results from `resp` by collapsing the "yes" and "perhaps" answers into one category. See `?@sec-ordinal-models` for details on how to model ordinal response variables with more than two categories. In the data, several item and person covariates are provided as well and we will talk about them in a bit. The first few rows of the data are given in Table 5.1.

Table 5.1: First ten rows of the `VerbAgg` data.

Anger	Gender	item	resp	id	btype	situ	mode	r2
20	M	S1WantCurse	no	1	curse	other	want	N
11	M	S1WantCurse	no	2	curse	other	want	N
17	F	S1WantCurse	perhaps	3	curse	other	want	Y
21	F	S1WantCurse	perhaps	4	curse	other	want	Y
17	F	S1WantCurse	perhaps	5	curse	other	want	Y
21	F	S1WantCurse	yes	6	curse	other	want	Y
39	F	S1WantCurse	yes	7	curse	other	want	Y
21	F	S1WantCurse	no	8	curse	other	want	N
24	F	S1WantCurse	no	9	curse	other	want	N
16	F	S1WantCurse	yes	10	curse	other	want	Y

In IRT we are often directly interested in the difficulty (or easiness) of items and the ability of people. That is, if items and people are modeled as varying coefficients, we are directly interested in these coefficients. This stands in contrast to treating them only as nuisance parameters which need to be modeled but are otherwise not really relevant for us. As a result, formulas without predictors, such as `r2 ~ 1 + (1 | item) + (1 | id)`, make total sense in an IRT context. We are predicting a response variable only by varying item and person coefficients with the intention to then analyse said coefficients. As family we choose `bernoulli`

with the default `logit` link to accommodate our binary response `r2`. Since the dataset is quite large and model fitting takes roughly a minute by chain, we fit the MCMC chains in parallel by setting `cores = 4`.

```
fit_va1 <- brm(
  r2 ~ 1 + (1 | item) + (1 | id),
  data = VerbAgg, family = bernoulli("logit"),
  cores = 4
)
```

In the summary output, we see that both items and persons show substantial variation. That is, some items are easier to say “yes” to than other items and some people tend to say “yes” to more items than other people. For an average item and an average person, the probability of responding “yes” or “perhaps” (vs. “no”) is close to 50% as we can see from the overall intercept being close to 0 on the logit scale.

```
summary(fit_va1)
```

```
Family: bernoulli
Links: mu = logit
Formula: r2 ~ 1 + (1 | item) + (1 | id)
Data: VerbAgg (Number of observations: 7584)

Multilevel Hyperparameters:
~id (Number of levels: 316)
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.39      0.07     1.25     1.54 1.01      934     1993

~item (Number of levels: 24)
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.21      0.19     0.90     1.63 1.00      654     1139

Regression Coefficients:
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept      -0.14      0.26    -0.63     0.36 1.01      291      547
```

While it is not visible in the formula, this model contains a special kind of multilevel structure, namely a *crossed structure*. We say that two grouping variables are crossed if neither one of them is nested inside the other. Clearly, this is the case of the grouping variables `item` and `id` since every person sees every item and vice versa. What I like about the multilevel formula syntax used by `brms` and several other packages is that we don’t have to worry about how

the different grouping variables are related to each other. That is, it doesn't matter for the formula if they are hierarchically nested or crossed. We just specify our multilevel terms and the model will automatically figure out the relation between all the grouping variables from the provided data.

If we want to learn which items are particularly easy or difficult, we can use the `coef` method which adds the overall intercept to the varying item coefficients. By convention in IRT, we will treat the overall intercept as “belonging” to the items rather than the person coefficients. A few lines of tidyverse code allow us to obtain a nice plot of the item coefficients' posteriors:

```
coef(fit_va1)$item[, , "Intercept"] %>%
  as_tibble(rownames = "item") %>%
  mutate(item = as.numeric(item)) %>%
  ggplot(aes(item, Estimate, ymin = Q2.5, ymax = Q97.5)) +
  geom_pointrange() +
  coord_flip() +
  labs(x = "Item Number")
```

```
Warning: There was 1 warning in `mutate()` .
i In argument: `item = as.numeric(item)` .
Caused by warning:
! NAs introduced by coercion
```

```
Warning: Removed 24 rows containing missing values or values outside the scale range
(`geom_pointrange()`).
```

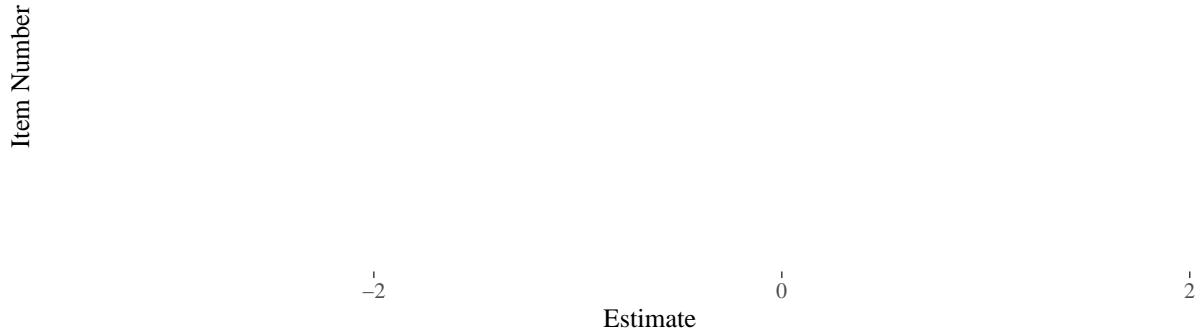


Figure 5.12: Varying item coefficients of model `fit_va1` on the logit scale.

As shown in Figure 5.12, there is indeed substantial variation across items. Some items (e.g., Item 1 or 4) ask whether people would *want to curse* in a given scenario. Those are very easy items in the sense that many people respond with “yes” or “perhaps”. Other items, (e.g.,

Item 21 or 24) ask whether people would *actually shout* in a given scenario. Those are very hard items in the sense that only few people respond with “yes” or “perhaps”. Estimates on the logit scale are admittedly a bit abstract. So let us transform the estimates back to the probability scale under the assumption of a person with average verbal aggression tendencies answering the items.

```
inv_logit <- function(x) 1 / (1 + exp(-x))

coef(fit_va1, summary = FALSE)$item[, , "Intercept"] %>%
  inv_logit() %>%
  posterior_summary() %>%
  as_tibble(rownames = "item") %>%
  mutate(item = as.numeric(item)) %>%
  ggplot(aes(item, Estimate, ymin = Q2.5, ymax = Q97.5)) +
  geom_pointrange() +
  coord_flip() +
  labs(x = "Item Number", y = "Probability")
```

Warning: There was 1 warning in `mutate()`.

i In argument: `item = as.numeric(item)`.

Caused by warning:

! NAs introduced by coercion

Warning: Removed 24 rows containing missing values or values outside the scale range (`geom\_pointrange()`).

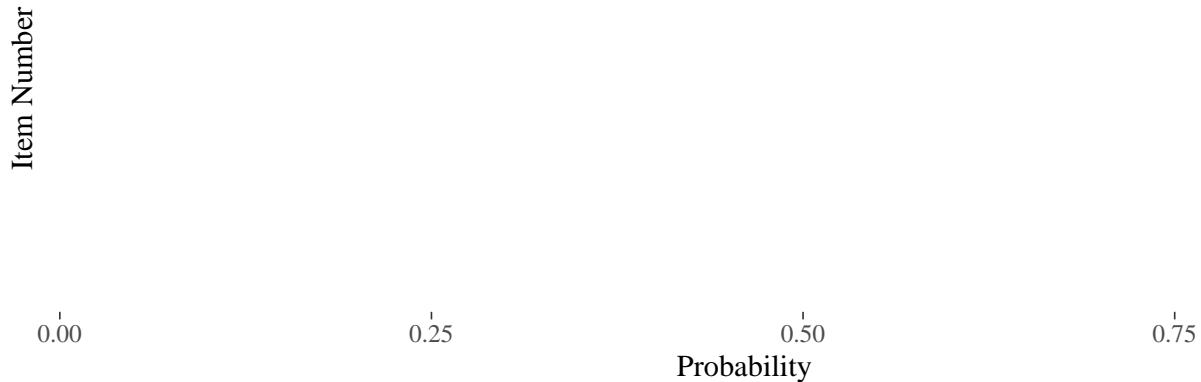


Figure 5.13: Varying item coefficients of model `fit_va1` on the probability scale.

In the above code, we first extracted the raw posterior draws by setting `summary = FALSE`, transformed these draws via `inv_logit`, and only then summarized them via

`posterior_summary`. This is to follow the general rule that any transformation should happen on the level of posterior draws not on the level of summary statistics. See Section 3.3.1 for an in-depth discussion of this approach.

We may not only look at items but also at the person coefficients. This time, we will use `ranef` since we do not want to include the overall intercept in the person coefficients by IRT convention. What is more, we will sort the coefficients, with higher values indicating people with higher verbal aggressive tendencies.

```
ranef(fit_va1)$id[, , "Intercept"] %>%
  as_tibble() %>%
  rownames_to_column() %>%
  arrange(Estimate) %>%
  mutate(id = seq_len(n())) %>%
  ggplot(aes(id, Estimate, ymin = Q2.5, ymax = Q97.5)) +
  geom_pointrange(alpha = 0.7) +
  coord_flip() +
  labs(x = "Person Number (Sorted)")
```

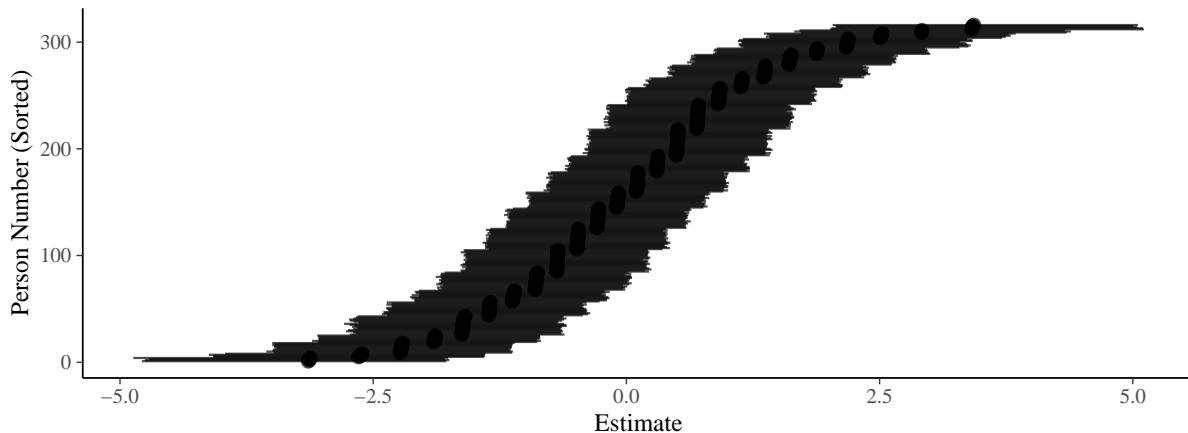


Figure 5.14: Varying person coefficients of model `fit_va1` on the logit scale.

The curve shown in Figure 5.14 looks almost perfectly like the empirical distribution function of a normal distribution, in good alignment with our choice of using an hierarchical normal prior for the varying coefficients. It doesn't have to be this way of course. However, as is the case with most complex human traits, verbal aggressive tendencies are influenced by the sum of many genetic and environmental factors. And as you may remember from basic statistics training, the sum of many (more or less) independent variables will be approximately normally distributed, a law of nature we clearly see at work here.

### 5.5.1 Adding item covariates

For all users who have never heard or cared about IRT, the previous model was perhaps a bit strange since we didn't actually have any predictors to analyse. This is about to change. Already above we noticed that items differ as a function of their specifics, for example, whether they asked for cursing or shouting or whether the referred to just wanting to engage in a certain behavior vs. actually doing it. In fact, the items are created by crossing three item factors, namely `btype` (the kind of verbally aggressive behavior: "curse", "scold", or "shout"), `situ` (does the scenario refer to me or another person?), and `mode` (is it about wanting or actually doing the behavior?). In our next model, we will include these three item factors as predictors.

```
fit_va2 <- brm(  
  r2 ~ btype + situ + mode + (1 | item) + (1 | id),  
  data = VerbAgg, family = bernoulli(),  
  cores = 4  
)  
  
summary(fit_va2)  
  
Family: bernoulli  
Links: mu = logit  
Formula: r2 ~ btype + situ + mode + (1 | item) + (1 | id)  
Data: VerbAgg (Number of observations: 7584)  
  
Multilevel Hyperparameters:  
~id (Number of levels: 316)  
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
sd(Intercept)     1.39      0.07     1.26     1.54 1.00     1259     2177  
  
~item (Number of levels: 24)  
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
sd(Intercept)     0.42      0.09     0.29     0.62 1.00     1376     2189  
  
Regression Coefficients:  
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
Intercept       1.78      0.23     1.32     2.22 1.00     1412     1682  
btypescold     -1.06      0.22    -1.50    -0.61 1.00     1618     2028  
btypeshout     -2.11      0.23    -2.56    -1.68 1.00     1806     2130  
situself       -1.06      0.19    -1.43    -0.69 1.00     1677     2082  
modedo         -0.71      0.19    -1.08    -0.34 1.00     1694     2320
```

We are learning a few interesting facts about this model from the summary output. First, it is less likely that people scold or shout that curse as we see from the clearly negative posteriors of the `btypescold` and `btypeshout` coefficients. Since we don't see a `btypecurse` coefficient and also didn't enforce any specific factor coding, we know dummy coding was used by default with `curse` as the reference category. See also Section 3.4 if you need a refresher on the coding of factors in R and brms. Second, people are less likely to engage in verbally aggressive behaviors when they themselves are involved in the situation, as we see from the clearly negative posterior of the `situself` coefficient. And lastly, people are less likely to actually doing rather than only wanting to engage in verbally aggressive behavior. Exemplary, we look at the latter effect also graphically on the probability scale, confirming what the posterior summary statistics have told us (see Figure 5.15).

```
conditional_effects(fit_va2, "mode")
```

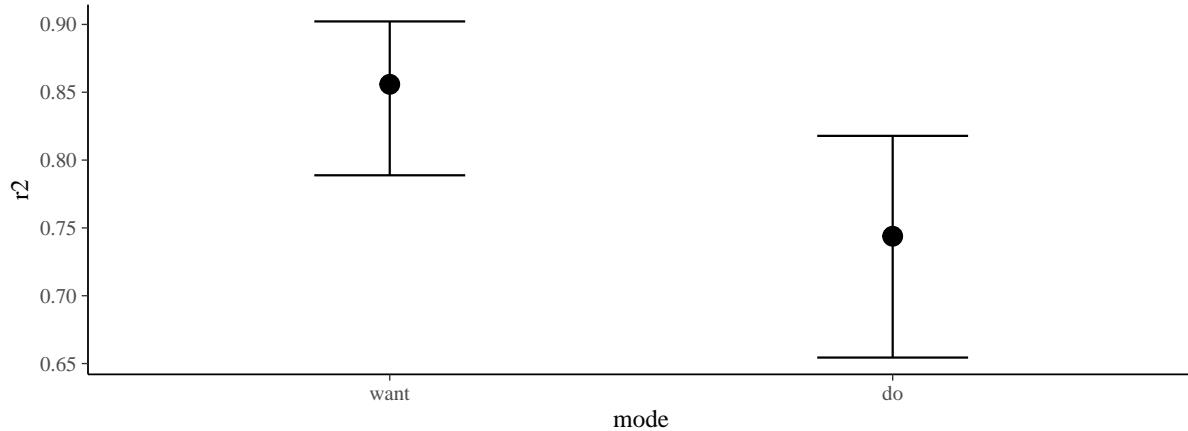


Figure 5.15: Conditional effects of model `fit_va2`.

Given that there are very clear effects of the three item covariates, we would expect predictive performance to improve quite strongly, right? Let's check out the LOO-CV results:

```
loo_va1 <- loo(fit_va1)
loo_va2 <- loo(fit_va2)
loo_compare(loo_va1, loo_va2)
```

	elpd_diff	se_diff
<code>fit_va2</code>	0.0	0.0
<code>fit_va1</code>	-0.3	1.5

Actually, there seems to be no noticeable difference in the LOO-CV performance between the model with and without item covariates. The reason is that all the variation explainable

by the item covariates can also be explained by the varying intercepts over items that were already present in the first model without covariates. After all, item covariates just explain differences between items. The varying intercepts over items may not explain *why* items differ but they certainly model *that* they differ. And the latter is completely sufficient for predictive performance measures. This is an excellent example where measures of posterior predictive performance (such as LOO-CV) fail to follow our intuition on what is a “better” model. Such cases happen particularly often in crossed multilevel structures because predictors tend to naturally relate to one or the other grouping factor, such as item covariates relate to items.

That said, it is not only the explanability perspective that favors our covariate model `fit_va2`. Remember the caveat that groups shall be exchangeable for our hierarchical normal priors to be justified (see Chapter 4)? Our strong item covariate effects tell us that, without accounting for these covariates, items were actually not exchangeable. So we have in fact violated the exchangeability assumption on our no-covariate model `fit_va1`. In this case, it does not seem to have hampered predictive performance, presumably because each item’s intercept is separately informed by the responses of 316 people, such that sharing of information across items barely affected the posterior anyway. If we would have had fewer observations per item, we might have actually seen differences in predictive performance. Try it out as an exercise by subsetting the `VerbAgg` data to model responses of, say, only 10 or 20 people.

In the next step, we will include varying slopes of some of the item covariates. Specifically, we will model the coefficients of `mode` as varying over persons via `(0 + mode | id)`. By means of `0 + mode`, we enforce cell mean coding for `mode`, which means that we will get an estimate for both `mode:want` and `mode:do` rather than an intercept and a different coefficients between the two categories. This has the advantage that we will directly see the correlation between the modes’ coefficients, which presumably is large. After all, someone who tends to want to engage in verbally aggressive behavior more is also more likely to do it (and vice versa).

```
fit_va3 <- brm(
  r2 ~ btype + situ + mode + (1 | item) + (0 + mode | id),
  data = VerbAgg, family = bernoulli(),
  cores = 4
)
```

```
summary(fit_va3)
```

```
Family: bernoulli
Links: mu = logit
Formula: r2 ~ btype + situ + mode + (1 | item) + (0 + mode | id)
Data: VerbAgg (Number of observations: 7584)

Multilevel Hyperparameters:
~id (Number of levels: 316)
```

```

          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(modewant)      1.47      0.09     1.30     1.65 1.00      1806     2853
sd(modedo)        1.67      0.10     1.48     1.86 1.00      1726     3056
cor(modewant,modedo) 0.77      0.04     0.69     0.84 1.00      1571     2461

~item (Number of levels: 24)
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)    0.46      0.09     0.31     0.67 1.00      1672     2917

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept       1.89      0.24     1.42     2.37 1.00      1767     2250
btypescold     -1.12      0.25    -1.62    -0.65 1.00      1921     2252
btypeshout     -2.23      0.25    -2.71    -1.74 1.00      1970     2310
situself        -1.12      0.20    -1.52    -0.71 1.00      2113     2302
modedo         -0.78      0.21    -1.21    -0.38 1.00      2239     2823

```

The within-person correlation between wanting and doing is indeed quite high with a posterior mean of close to  $r = 0.8$ . We also see this visualized in Figure 5.16.

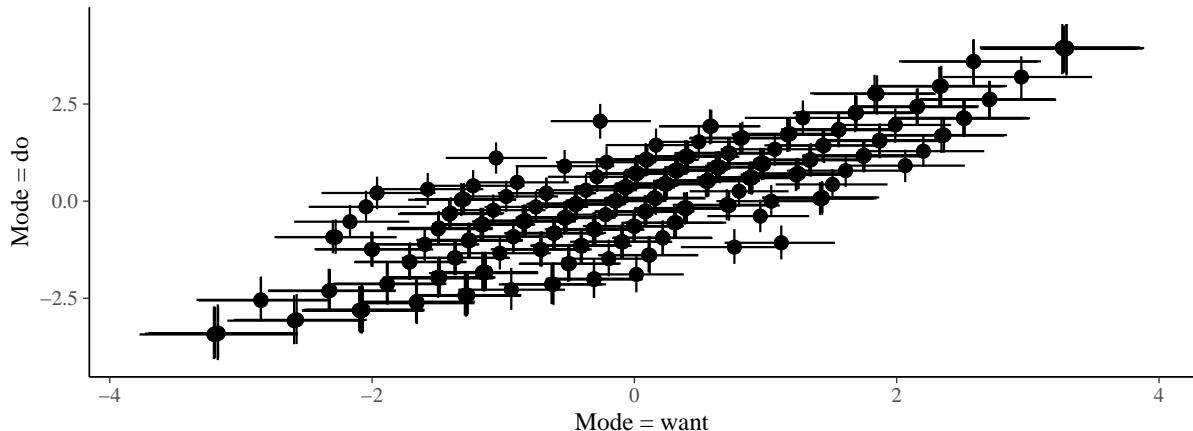


Figure 5.16: Varying person coefficients of modes ‘want’ vs. ‘do’ in model `fit_va3`.

Given the strong correlation, we might ask if it is at all necessary to have different varying person coefficients for wanting and doing. To answer this question, LOO-CV is sensible again since no other previously included term was able to express the same variation.

```

loo_va3 <- loo(fit_va3)
loo_compare(loo_va2, loo_va3)

```

```
elpd_diff se_diff
```

```

fit_va3    0.0      0.0
fit_va2 -80.9     12.8

```

Apparently, modeling different varying person coefficients for wanting and doing strongly improves predictions, so we have good reason to keep these coefficients in the model as we keep adding more terms. In model `fit_va3`, we have chosen model varying coefficients only for `mode` and not for `btype` and `situ`, which was a rather arbitrary decision done for illustration purposes. Of course, it is quite likely that people will vary in their responses also as a function of these other item covariates. You may want to try out modeling additional varying person coefficients as an exercise. Let me just remind you that we *cannot* model varying coefficients of item covariates over the items. The reason is that for each item, item covariates are constant. And no variation of a predictor within a grouping factor (here: items) means that corresponding varying coefficients are not identified by the data. The formula syntax does not care which predictor varies within or between which grouping factor. That is something we, as users, have to keep in mind to ensure all of the modeled terms are reasonable given the data structure.

### 5.5.2 Modeling person covariates

In the data, we also have access to two person covariates, namely people's `Gender` and their general `Anger` tendencies measured with a separate questionnaire. We choose to model an interaction between `Gender` and `mode` to understand if men and women differ from each other in their tendencies to want vs. do engage in verbally aggressive behavior. And we also choose to model varying item coefficients for `Gender` using cell-mean coding via `(0 + Gender | item)`. Here is how our chosen model with both item and person covariates looks like:

```

fit_va4 <- brm(
  r2 ~ Anger + Gender*mode + btype + situ +
    (0 + Gender | item) + (0 + mode | id),
  data = VerbAgg, family = bernoulli(),
  cores = 4, seed = 63242
)

summary(fit_va4)

Family: bernoulli
Links: mu = logit
Formula: r2 ~ Anger + Gender * mode + btype + situ + (0 + Gender | item) + (0 + mode | id)
Data: VerbAgg (Number of observations: 7584)

Multilevel Hyperparameters:
~id (Number of levels: 316)

```

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sd(modewant)	1.49	0.09	1.32	1.68	1.00	1673	2461		
sd(modedo)	1.59	0.10	1.40	1.78	1.00	1907	2997		
cor(modewant,modedo)	0.78	0.04	0.70	0.85	1.00	1709	2358		
 ~item (Number of levels: 24)									
	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sd(GenderF)	0.52	0.11	0.35	0.76	1.00	1420	2015		
sd(GenderM)	0.35	0.11	0.16	0.59	1.00	1974	2487		
cor(GenderF,GenderM)	0.78	0.18	0.33	0.99	1.00	2049	2559		
 Regression Coefficients:									
	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	0.71	0.44	-0.16	1.57	1.00	1597	2052		
Anger	0.06	0.02	0.02	0.10	1.00	1464	2113		
GenderM	-0.09	0.24	-0.55	0.41	1.00	1575	2310		
modedo	-0.99	0.23	-1.44	-0.54	1.00	2090	2798		
btypescold	-1.03	0.22	-1.47	-0.59	1.00	2197	2239		
btypeshout	-2.42	0.24	-2.90	-1.94	1.00	2035	2567		
situself	-1.04	0.18	-1.40	-0.68	1.00	2225	2295		
GenderM:modedo	0.90	0.23	0.45	1.36	1.00	3201	3209		

Among others, we learn from the summary that there indeed seems to be an interaction between `mode` and `Gender`. Specifically, the clearly positive interaction coefficient `GenderM:modedo` tells us that men (report to) actually engage in verbally aggressive behavior more often than would be expected based on the main gender and mode coefficients alone. A bit awkward to understand what this actually means especially since we are working on the logit scale. Let's look at a plot instead (see Figure 5.17):

```
conditional_effects(fit_va4, "mode:Gender")
```

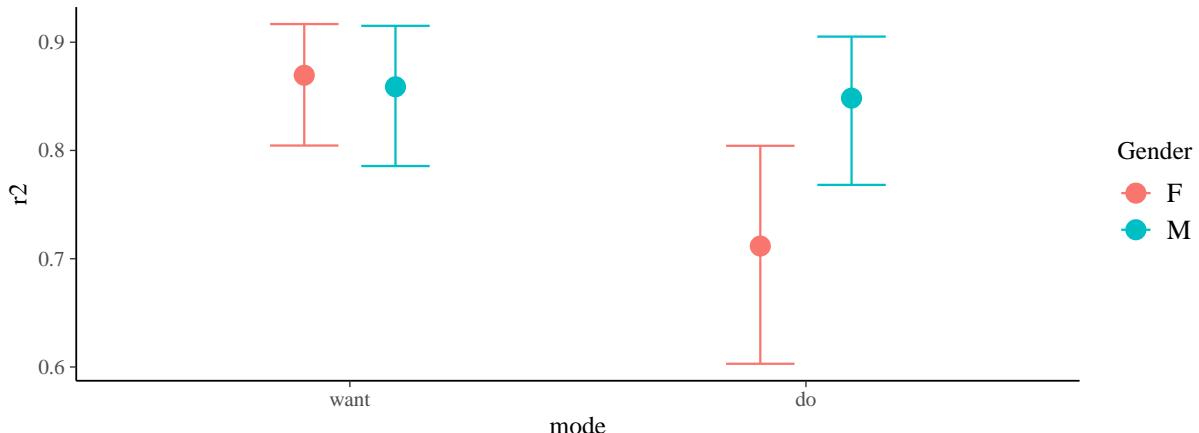


Figure 5.17: Conditional effects of model `fit_va4`.

Accordingly, while both men and women are roughly equally likely to (report to) *want* to engage in verbally aggressive behavior, women are clearly less likely to actually *do* it, while for men there is barely any difference in the probability of reported wanting and doing. We could try to run LOO-CV in an attempt to verify that the newly added terms improve predictive performance but, as we have seen earlier, we would primarily learn about the relevance of the additional varying coefficients. So instead, we have to be content with the evidence the posterior itself provides about the credibility of the interaction effect.

In the next section, we will discuss how to further improve sampling speed for complex and relatively large models such as the generalized linear multilevel models we fitted on the `VerbAgg` dataset.

### 5.5.3 Speeding up sampling using within-chain parallelization

Estimating the previous model, `fit_va4`, was quite slow relative to what we have encountered previously. This is because we have around 7500 observations in the `VerbAgg` dataset and because the multilevel structure we modeled was quite involved. On my laptop each chain of 2000 iterations of which the first 1000 were warmup ran for about 100 seconds. We can extract this information from the model via the `elapsed_time` method:

```
brms:::elapsed_time(fit_va4)
```

	chain_id	warmup	sampling	total
1	1	50.651	19.881	70.532
2	2	54.220	19.856	74.076
3	3	50.028	19.332	69.360
4	4	47.605	19.382	66.987

Around 100 seconds per chain is not much, especially when we can easily parallelize chains as done above. But what if we didn't have only a few thousand observations but hundreds of thousands or even more? And what if our model structure was even more complicated? In such cases, our chains may easily run many hours or even days, which becomes highly impractical. For example, some computing clusters only allow jobs to run for a maximum of 2 days and any job not finishing in time will be canceled. We could of course run our chains for fewer iterations but we still need relevant amount of warmup iterations to make sure the sampler works accurately and efficiently. What is more, the warmup phase often makes up most of the total sampling time, as was the case also for `fit_va4` with warmup making up about 75% of the total time. So running more chains and parallelizing them across cores is not always a viable option to sufficiently reduce a model's sampling time.

Accordingly, what we need is to speed up individual chains. One approach is to find a way to achieve *within-chain parallelization* (aka threading) so that we can use multiple cores within the same MCMC chain. Since, in most models, observations are assumed to be conditionally independent given the parameters, we can split the log-likelihood into multiple parts and then evaluate each part with a different core before pooling all the evaluations together. Stan provides the required functionality in the form of the `reduce_sum` function, which `brms` interfaces. All we need to do is to set argument `threads` in `brm`. For example, we can set `threads = threading(4)` to indicate the we want to use 4 CPU cores *per chain*. We could combine this with the `cores` argument to have both between and within-chain parallelization. For example, we could set `cores = 2` and together with `threads = threading(4)` this would indicate we want to run 2 chains in parallel each of which uses 4 cores. That is, we would use 8 cores in total.

A detailed discussion of within-chain parallelization in `brms`, can be found in the [brms\\_threading vignette](#). Here, I just want to illustrate the use of threading for a single model by showcasing how to speed up `fit_va4`. To simplify the comparison, I only use within-chain and no between-chain parallelization:

```
fit_va5 <- brm(
  r2 ~ Anger + Gender*mode + btype + situ +
    (0 + Gender | item) + (0 + mode | id),
  data = VerbAgg, family = bernoulli(),
  cores = 1, threads = threading(4),
  seed = 63242
)
```

The posterior of `fit_va5` is the same as for `fit_va4` up to MCMC error so we can just focus on the sampling speed:

```
brms:::elapsed_time(fit_va5)
```

	chain_id	warmup	sampling	total
1	1	18.588	7.485	26.073
2	2	19.288	7.492	26.780
3	3	19.808	7.464	27.272
4	4	19.561	7.459	27.020

As we see, each chain now took about 30 rather than 100 seconds, roughly a 3x speed up. This is close but not quite the 4x speed up that we would ideally hope for since we have used 4 cores. While different MCMC chains are embarrassingly parallel, cores used within the same chain still need to talk to each other, which reduces efficiency and causes some diminishing returns. Often, we see that using 4x as many cores may only speed up the sampling by about 2x or less, so seeing a 3x speedup in our example is actually very good. Using more and more cores will eventually lead to little speedup though. This is because the individual chunks of data for which the log-likelihood is evaluated within one core becomes so small that the non-parallelized parts of evaluating the model’s log-posterior start dominating the overall sampling time. Try it out as an exercise by changing the number of cores for within-chain parallelization for the above model.

## 5.6 Summary

GLMMs open up a vast space of modeling options, especially when combined with the conceptual and computational perks of Bayesian inference. Using three case studies, this chapters gave you a glimpse into the capabilities of Bayesian GLMMs. In some of the next chapters, such as in [?@sec-ordinal-models](#) and Chapter 7, we will dive deeper into some specific model classes within the GLMM framework. A lot of brms users will likely never have left the space of GLMMs so far. That said, there are many more modeling options beyond GLMMs both in probabilistic modeling generally and in brms specifically. Among others, this concerns non-linear modeling ([?@sec-nl-theory](#) and Chapter 8) as well as distributional regression ([?@sec-ldms](#) and [?@sec-gldms](#)), to name two prominent directions. I am excited to further explore with you all the advanced modeling opportunities that brms has to offer.

# 6 Multivariate Models

## 6.1 Setup

```
library(ggplot2)
library(patchwork)
library(bayesplot)
library(brms)
```

## 6.2 Introduction

So far in this book, we have only looked at models predicting a single response variable. In this chapter, we will generalize what we have learned to models predicting multiple response variables at once, called *multivariate models* for short. The multivariate models I will talk about here are not just collections of mutually independent univariate models. Rather, in one way or another, they express *dependencies* between the univariate models. In fact, if we didn't include any notion of dependency, nothing would be gained from fitting a multivariate model as it would yield the same results as the individual univariate models. Below, we will take a detailed look into two case studies, each highlighting the importance of different kinds of dependencies we can express with multivariate models in brms.

## 6.3 Modeling residual confounding

For our first case study, we will look into an example of causal inference. In our scenario we will analyse data from a hypothetical study that investigates the effect of a psychological treatment on social anxiety. The general idea of this scenario is inspired from a blog post on instrumental variables (Magnusson 2018), which I have simplified to focus on the aspects that motivate multivariate modeling in brms.

In the hypothetical scenario, patients are assigned to either the treatment group or a (waitlist) control group, encoded in the binary treatment variable `treat`. As the outcome of interest, we will study the social anxiety after treatment. The whole scenario is complicated by two aspect. First, patients in the treatment group may adhere to the treatment differently. For example,

if they miss many of their therapy sessions or don't practice their newly acquired skills in real life, the treatment cannot be very effective. We will capture these aspects in the variable **adherence**. It will be zero for people in control group but vary for people in the treatment group. In the end, we are not interested in the causal effect of the treatment assignment (**treat**) on the **outcome**. After all, people in the treatment group not adhering to the treatment at all are not getting more treatment than people in the control group. Instead, we are interested in the causal effect of **adherence** on the **outcome**, because **adherence** is the variable actually capturing what happens if the treatment is followed properly. This complicates the causal analysis because **adherence** is not fully randomized but rather partially observed, which brings us to the second point.

The social anxiety after treatment (**outcome**) will clearly be influenced by the social anxiety pre-treatment, which we will capture in the variable **base**. What is more, how anxious patients are at pre-treatment will likely also influence how much they adhere to the treatment (if they were assigned to the treatment group). Different directions of this effect are plausible. More socially anxious people may be more motivated to adhere to the treatment because they suffer more from their anxieties. Or they may be too scared to even go to the therapy sessions or trying things out in real life, leading to less adherence. Whatever the direction and specific functional form may be, it is clear that **base** is highly likely to affect both **outcome** and **adherence**. This renders **base** a *confounder*. We need to control for confounders in our analysis to have a chance of obtaining trustworthy estimates of the true causal effect.

The complete causal graph of this scenario is summarized in Figure 6.1. Below, we simulate some data following this causal graph. We assume that 100 people are assigned to the treatment group and 30 people to the control group. We assume all effects to be linear for simplicity. The causal effect of interest (adherence on outcome) is set to  $\beta = 0.7$ . The remaining parameters required to simulate all variables are chosen so that the patterns of interest are clearly visible in the models that follow. That said, you would see these patterns with many other parameter configurations as well, since the patterns are a result of the underlying causal graph not the specific parameter choices.

```
set.seed(9463)
beta <- 0.7
Nt <- 100
Nc <- 30
N <- Nt + Nc
treat <- c(rep(1, Nt), rep(0, Nc))
base <- rnorm(N)
adherence <- 5 - 2 * base + rnorm(N, sd = 0.5)
adherence <- ifelse(adherence < 0, 0, adherence) * treat
outcome <- base + beta * adherence + rnorm(N)
datRC <- data.frame(treat, base, adherence, outcome)
```

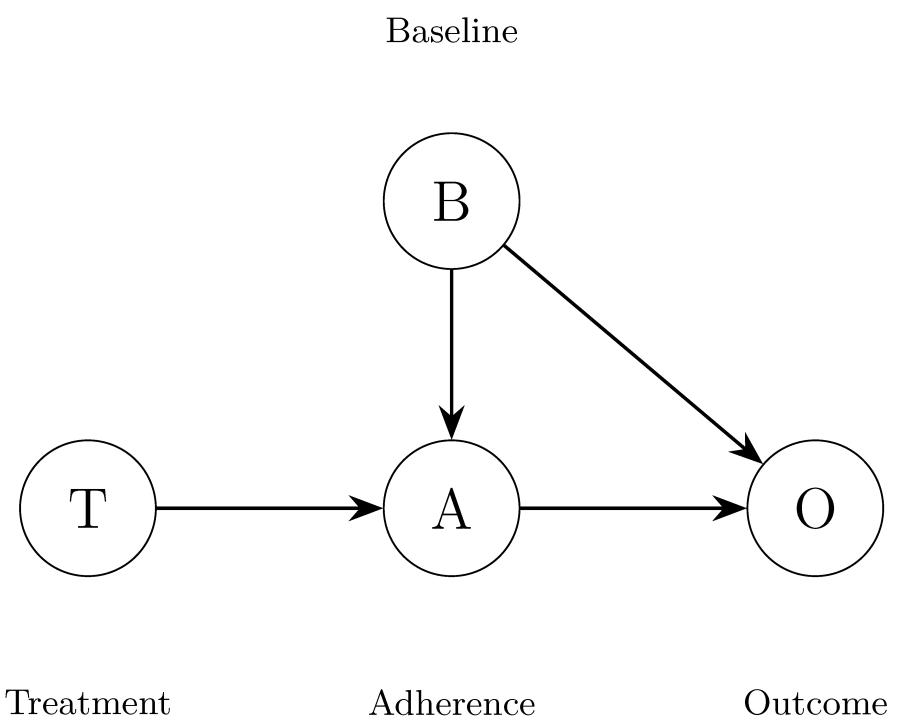


Figure 6.1: Causal graph of the social anxiety treatment scenario.

### 6.3.1 Correcting for confounding

In what follows, we will analyse the data with a series linear models, which eventually become multivariate. Since the models themselves have only few parameters and we have plenty of data, the brms default priors will be perfectly fine. Also convergence is not an issue for any of the models so we skip discussing it. Rather, we focus directly on the posterior summaries and their implications for the causal effect estimation.

First, we fit a naive model just predicting `outcome` by `adherence`, completely ignoring the `base` confounder.

```
fit_rc1 <- brm(outcome ~ adherence, data = datRC)
```

The results shown in Table 6.1 indicate that we underestimate the true causal effect. The posterior mean of the regression coefficient is  $\hat{\beta} \approx 0.49$  while the true causal effect is  $\beta = 0.7$ . The fact that this model yields biased estimates is of course unsurprising given that we have ignored an important confounder.

Table 6.1: Posterior summary for model `fit_rc1`.

variable	mean	sd	q5	q95
b_Intercept	0.99	0.23	0.60	1.37
b_adherence	0.49	0.05	0.41	0.57

Next, we include the confounder in our model:

```
fit_rc2 <- update(
  fit_rc1, outcome ~ adherence + base,
  newdata = datRC
)
```

We have used the `update()` method to avoid recompling the model since the underlying Stan code of `fit_rc1` and `fit_rc2` is the same. As we can see in Table 6.2, the posterior of the adherence coefficient is now very close to the true causal effect. So controlling for the confounder has removed the bias we have seen before.

Table 6.2: Posterior summary for model `fit_rc2`.

variable	mean	sd	q5	q95
b_Intercept	-0.05	0.20	-0.39	0.28
b_adherence	0.73	0.04	0.66	0.81

Table 6.2: Posterior summary for model `fit_rc2`.

variable	mean	sd	q5	q95
b_base	1.18	0.12	0.98	1.37

However, in practice, we rarely have direct access to the confounder but only to a noisy measurement of the confounder. Let's add such a noisy measurement to our simulated data:

```
datRC$base_noise <- datRC$base + rnorm(N)
```

We fit our next model replacing `base` with `base_noise`.

```
fit_rc3 <- update(
  fit_rc1, outcome ~ adherence + base_noise,
  newdata = datRC
)
```

The results displayed in Table 6.3 indicate that we no longer obtain an unbiased estimate of the true causal effect. The posterior mean of the adherence coefficient is  $\hat{\beta} \approx 0.57$ . So not as badly biased as the coefficient obtained by model `fit_rc1` where we didn't control for the confounder at all. But also not as good as when we had access to the actual, noise-free confounder. Given how often we only have access to measured proxies important variables, you can imagine of big of a problem this is in practice (e.g., see Westfall and Yarkoni 2016).

Table 6.3: Posterior summary for model `fit_rc3`.

variable	mean	sd	q5	q95
b_Intercept	0.63	0.22	0.27	0.98
b_adherence	0.57	0.05	0.50	0.65
b_base_noise	0.50	0.10	0.34	0.67

### 6.3.2 Correcting for residual confounding

We can fix the bias in our causal estimate even without access to the actual confounder. And this is where multivariate models come into play. We will model both `outcome` and `adherence` jointly following our understanding of the causal graph and corresponding data-generating process. We continue to predict `outcome` by `adherence` and `base_noise` because this is what the graph dictates. Following the same logic, we predict `adherence` by the treatment indicator `treat` and its interaction with `base_noise`, concretely `adherence ~ treat +`

`treat:base_noise`. We intentionally do not include a main effect of `base_noise` because in the control group (`treat = 0`) the treatment adherence is 0 by design, a value which cannot be affected by the baseline functioning of the patient.

Now that we have specified the individual formulas for our two response variables, we need to add something else to the model such that the two parts can exchange information. To be precise, there is still residual correlation between `outcome` and `adherence` following from the fact that we have only partially controlled for the confounder. In other words, the part of `base` not expressed by `base_noise` remains unmodeled so will show itself in the residuals, causing them to be correlated across the two model parts.

Let's put this in math. We assume that the two responses are multivariate normally distributed:

$$(\text{outcome}_i, \text{adherence}_i) \sim \text{multinormal}((\mu_{O_i}, \mu_{A_i}), \Sigma)$$

The means  $\mu_O$  and  $\mu_A$  of outcome and adherence, respectively, are predicted as follows:

$$\mu_{O_i} = b_{O0} + \text{adherence}_i b_{O1} + \text{base\_noise}_i b_{O2},$$

$$\mu_{A_i} = b_{A0} + \text{treat}_i b_{A1} + \text{treat}_i \times \text{base\_noise}_i b_{A2}.$$

The residual covariance matrix  $\Sigma$  is decomposed into residual standard deviations  $\sigma_O$  and  $\sigma_A$ , as well as the residual correlation  $\rho$ :

$$\Sigma = \begin{pmatrix} \sigma_O^2 & \sigma_O \sigma_A \rho \\ \sigma_O \sigma_A \rho & \sigma_A^2 \end{pmatrix}$$

The latter parameter,  $\rho$ , is unique to multivariate models and indicates how strongly the residuals of two response variables correlate after accounting for the predictor terms. For the above multivariate model,  $\rho$  is the only parameter that “connects” the two univariate models. If it wasn't included, the results of the multivariate model and separately estimated univariate models would be the same. In other words, without  $\rho$  we wouldn't gain anything from multivariate models in this case study.

Expressing this multivariate model in brms is straightforward:

```
bform_rc4 <- bf(outcome ~ adherence + base_noise) +
  bf(adherence ~ treat + treat:base_noise) +
  set_rescor(TRUE)
```

First, we define the individual univariate models via their formulas and wrap them inside `bf`. This is short for `brmsformula`, an extended formula object that we need for many of the advanced models that brms supports. We then add the formula objects together via `+`. Equivalently, we could also use `mvbf(bf(...), bf(...))`. To ensure that residual correlations are estimated as well, we further add `set_rescor(TRUE)`. This combined object is then passed to the `brm` in the same way that a more basic formula would:

```
fit_rc4 <- brm(bform_rc4, data = datRC)
```

The brms default prior for  $\rho$  is uniform in  $[-1, 1]$ . If we had more than two response variables, resulting in more than one residual correlation parameters, this would generalize to a uniform distribution over correlation matrices (see also Section 4.7 for more details).

For `fit_rc4`, we will investigate the whole `summary` output rather than just a reduced table to show you how summaries of multivariate models differ from those of univariate models.

```
summary(fit_rc4)
```

```
Family: MV(gaussian, gaussian)
Links: mu = identity
       mu = identity
Formula: outcome ~ adherence + base_noise
          adherence ~ treat + treat:base_noise
Data: datRC (Number of observations: 130)
```

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
outcome_Intercept	0.14	0.25	-0.38	0.62	1.00	2635	2583
adherence_Intercept	0.05	0.24	-0.44	0.53	1.00	3847	3153
outcome_adherence	0.69	0.06	0.58	0.81	1.00	2451	2282
outcome_base_noise	0.61	0.10	0.41	0.80	1.00	3581	2592
adherence_treat	5.06	0.28	4.52	5.61	1.00	3741	3013
adherence_treat:base_noise	-1.05	0.11	-1.26	-0.85	1.00	4729	2911

#### Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma_outcome	1.40	0.10	1.23	1.61	1.00	3515	2897
sigma_adherence	1.34	0.09	1.19	1.52	1.00	4528	3163

#### Residual Correlations:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
rescor(outcome,adherence)	-0.43	0.09	-0.59	-0.25	1.00	2535	2766

For the most part the summary looks familiar. On the top we see the families, formulas and data information. Also the posterior tables looks similiar except that the parameters now have the corresponding response variable as prefix. Checking the regression coefficient of `adherence` on `outcome`, (labeled as `outcome_adherence`), we find a posterior mean of  $\hat{\beta} \approx 0.71$ , so clearly very close to the true causal effect of  $\beta = 0.7$ . And we managed that without knowing the actual, noise-free confounder. This appears almost like magic but in fact only works because the treatment variable `treat` is an *instrumental variable* for `adherence`. This means that `treat` predicts with `adherence` and has no other causal paths to the `outcome`. In practice, finding instrumental variables is hard since the assumptions to satisfy for a variable to be instrumental are pretty strong. If you want to learn more about instrumental variables and causal inference more generally see, for example, Pearl (2009).

Despite having found an instrumental variable, the fact that we only used a noisy proxy of the confounder is still visible in the model results. First, we see the the residual correlation parameter at the very bottom of the summary output is cleary non-zero. It happen to be negative for our specific data, but could as well have been positive too if we had specified our data-generating process a little differently. For our purposes, the sign doesn't matter. What matters is that the unmeasured part of the confounder is represented in the residual correlation (with the help of an instrumental variable), thus no longer biasing our causal estimate.

Yet, while we managed to make the causal estimate unbiased, we still don't do as well as we could have if we had included the actual confounder: The posterior of the causal estimate is wider for `fit_rc4` than for `fit_rc2` as we illustrate in Figure 6.2.

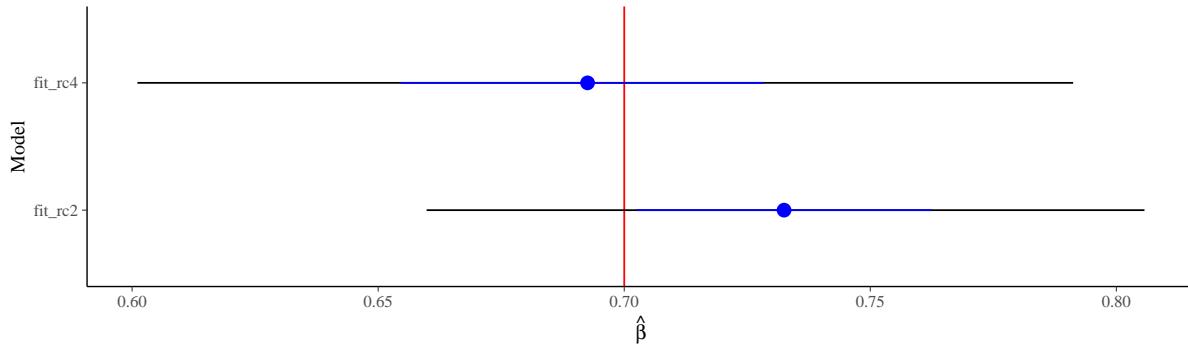


Figure 6.2: Posterior summary of the causal effect estimates obtained from models `fit_rc2` and `fit_rc4`. The central dot indicates the posterior mean, the outer black range the 90% CI, and the inner blue range the 50% CI. The vertical red line indicates the true causal effect.

Lastly, as a sanity check, let's replace the noisy confounder with the actual confounder in the multivariate model. Since the residual correlation was caused by us only partially controlling for the actual confounder, including the latter in the model should eliminate the residual correlation.

```
bform_rc5 <- bf(outcome ~ adherence + base) +
  bf(adherence ~ treat + treat:base) +
  set_rescor(TRUE)
```

```
fit_rc5 <- brm(bform_rc5, data = datRC)
```

Indeed, as shown in Table 6.4, this is exactly what happens. The residual correlation is almost zero with only small uncertainty and The posterior of causal estimate is unbiased and almost identical to the one obtained in corresponding univariate model `fit_rc2`.

Table 6.4: Selected posterior summary for model `fit_rc5`.

variable	mean	sd	q5	q95
b_outcome_Intercept	-0.07	0.20	-0.41	0.27
b_adherence_Intercept	0.00	0.08	-0.13	0.14
b_outcome_adherence	0.74	0.04	0.66	0.81
b_outcome_base	1.18	0.12	0.99	1.38
b_adherence_treat	5.10	0.09	4.94	5.25
b_adherence_treat:base	-2.03	0.05	-2.11	-1.95
rescor__outcome__adherence	-0.06	0.09	-0.21	0.09

### 6.3.3 Checking model fit

Before we finish this case study, let us go back to our model `fit_rc4` to discuss the appropriateness of the normal likelihood assumption. The main problem of this assumption is the fact that `adherence` is constant 0 for patients in the control group, which implies the distribution shown in Figure 6.3. The spike at zero cannot appropriately be modeled by a normal distribution alone. Some of you may already recognize this situation as calling for a *zero-inflated* model, where we add a special component to handle excess zeros. We cover zero-inflated models separately in ?@sec-gldms.

In our case, we already know by design of the study that the normal likelihood assumption of `adherence` is partially inappropriate. Yet, it is useful to investigate how different posterior predictive checks highlight this issue (see Figure 6.4).

```
pp_check(fit_rc4, type = "dens_overlay", resp = "adherence")
pp_check(fit_rc4, type = "ecdf_overlay", resp = "adherence")
pp_check(fit_rc4, type = "error_scatter_avg", resp = "adherence")
```

The default check type "dens\_overlay" correctly shows a bimodal distribution of `adherence` but smooths over the zero mode because it fits a continuous density to the data. This effectively

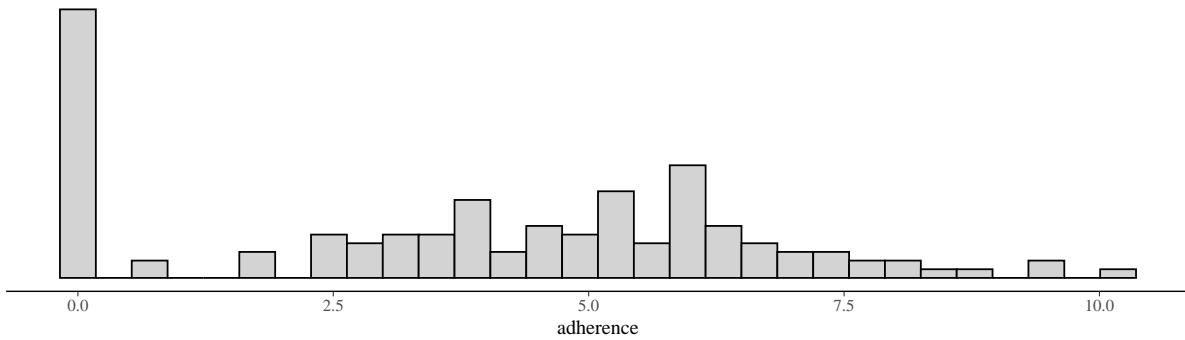


Figure 6.3: Histogram of the `adherence` variable in the `datRC` data.

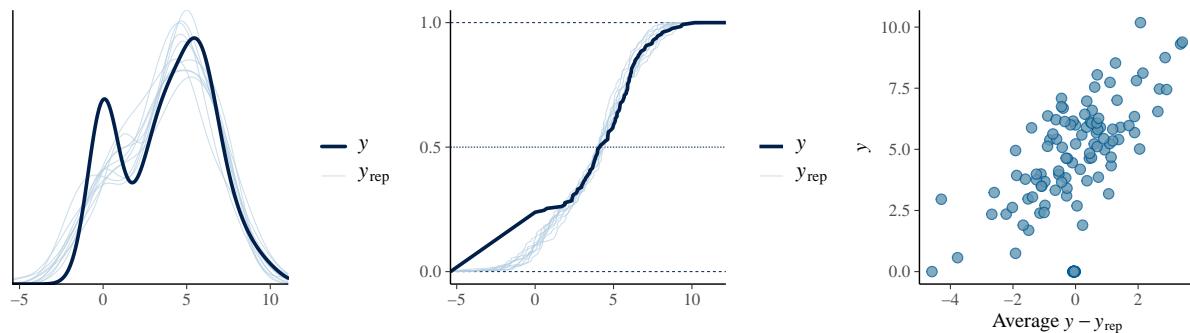


Figure 6.4: Three posterior predictive checks of the `adherence` response in model `fit_rc4`.

hides that this mode consists only of exact zeros. Yet, we still see that our model doesn't cover this zero mode appropriately. The second check type "ecdf\_overlay" is a bit more revealing as it shows the zero mode as a vertical spike in the empirical cumulative distribution function. And it also shows that the normal likelihood smooths over the spike. When checking the residuals, via type "error\_scatter\_avg", we only see the zero mode if we look very closely and notice that the point at  $(0, 0)$  is shining more brightly due to many semi-transparent points overlapping each other there. But if we didn't already know what we are looking for, this detail would easily be missed. What we do see in this plot though is a correlation between residuals and response. This results from the fact that `fit_rc4` only includes the noisy confounder. If we instead looked at the same plot for `fit_rc5`, the correlation between residuals and response would vanish (try it out yourself as an exercise).

In some of the previous chapters, for example in Chapter 3 and Chapter 5, we have spent a lot of time finding the best possible likelihood families for the response variables at hand. So the fact that we just stucked to a normal likelihood here may feel wrong. We could change the family for `adherence` to something else than normal in the hope that this would be more appropriate. However, to achieve the goal of this case study – obtaining an unbiased causal estimate of the adherence effect in the face of residual confounding – it may not help or even be detrimental. The reason is that the key aspect of our multivariate model that helps us achieving our goal (modeling the residual correlation between `outcome` and `adherence`) is not that easily available for many other families. We will come back to this point later on in the context of our second multivariate case study (see Section 6.4.3). Viewed from a slightly different angle, we primarily care about the `outcome` model component, where a normal likelihood is appropriate. And we do not need to perfectly model `adherence` as long as we get the aspects right that are central to our goal, namely predicting `adherence` via an instrumental variable (the treatment assignment) and modeling the residual correlation between `adherence` and `outcome`. I cannot give a general rule when certain model-misfits will or will not matter for your specific inference goal. The space of models and inference goals, even only the space covered by brms is simply too large. If in doubt, I recommend running small simulation studies to investigate how different model assumptions and misfits between model and data affect the estimates of interest. You can find an example of such a simulation study in Section 4.8.

## 6.4 Modeling the phenotype of young birds

As a second case study, we will look at real-world example from ecology. Hadfield et al. (2007) analyzed data of the Eurasian blue tit, a small singing bird. They predicted the (scaled) `tarsus` length as well as the (scaled) `back` color intensity of  $N = 828$  chicks. Half of the brood were put into another `fosternest`, while the other half stayed in the fosternest of their own `dam` (genetic mother). This allows to separate genetic (`dam`) from environmental (`fosternest`) factors. Additionally, we have information about the (scaled) `hatchdate` and `sex` of the chicks (the latter being known for 94% of the animals, otherwise coded as unknown). An overview of the data is provided in Table 6.5.

```
data("BTdata", package = "MCMCglmm")
```

Table 6.5: First few observations of the **BTdata**.

tarsus	back	animal	dam	fosternest	hatchdate	sex
-1.89	1.15	R187142	R187557	F2102	-0.69	Fem
1.14	-0.76	R187154	R187559	F1902	-0.69	Male
0.98	0.14	R187341	R187568	A602	-0.43	Male
0.38	0.26	R046169	R187518	A1302	-1.47	Male
-0.08	-0.30	R046161	R187528	A2602	-1.47	Fem
-1.14	1.56	R187409	R187945	C2302	0.35	Fem

The goal is to understand the degree to which genetic and environment factors predict **tarsus** length and **back** color, as well as the dependencies between these two chick properties. It is these dependencies that require us to use multivariate models.

Due to the multilevel structure of the data, our first multivariate model on the **BTdata** will already be quite involved. For now, we will assume that both tarsus length and back color are normally distributed, but we will change this assumption later on. The joint likelihood of the two response variables is then multivariate normal

$$(\text{tarsus}_i, \text{back}_i) \sim \text{multinormal}((\mu_{Ti}, \mu_{Bi}), \Sigma),$$

with means predicted by

$$\mu_{Ti} = b_{T0} + \text{hatchdate}_i b_{T1} + \text{sex}_i b_{T2} + b_{Tf[i]} + b_{Td[i]},$$

$$\mu_{Bi} = b_{B0} + \text{hatchdate}_i b_{B1} + \text{sex}_i b_{B2} + b_{Bf[i]} + b_{Bd[i]}.$$

The means include both the sex and hatchdate predictors as well as varying coefficients for fosternest and dam. Here,  $T$  stands for **tarsus** and  $B$  stands for **back** color. Further,  $f[i]$  and  $d[i]$  indicate the **fosternest** and **dam**, respectively, to which chick  $i$  belongs. The residual covariance matrix  $\Sigma$  is decomposed into residual standard deviations  $\sigma_T$  and  $\sigma_B$ , as well as the residual correlation  $\rho$ :

$$\Sigma = \begin{pmatrix} \sigma_T^2 & \sigma_T \sigma_B \rho \\ \sigma_T \sigma_B \rho & \sigma_B^2 \end{pmatrix}$$

As we have already discussed in the first case study above, the  $\rho$  parameter is unique to multivariate models. It indicates how strongly the residuals of two response variables correlate after accounting for the predictor terms.

We complete the model by specifying the multilevel structure of both the `fosternest` and `dam` coefficients:

$$(b_{T0f}, b_{B0f}) \sim \text{multinormal}((b_{T0F}, b_{B0F}), \Sigma_F),$$

$$(b_{T0d}, b_{B0d}) \sim \text{multinormal}((b_{T0D}, b_{B0D}), \Sigma_D).$$

Here,  $F$  stands for `fosternest` and  $D$  stands for `dam`. We have two varying intercepts per level of each grouping variable, one for `tarsus` length and one for the back color of the chicks. The two intercepts might be correlated, so we decompose  $\Sigma_F$  and  $\Sigma_D$  into two standard deviations and a correlation, just as we did for the residual covariance matrix  $\Sigma$ :

$$\Sigma_F = \begin{pmatrix} \sigma_{TF}^2 & \sigma_{TF} \sigma_{BF} \rho_F \\ \sigma_{TF} \sigma_{BF} \rho_F & \sigma_{BF}^2 \end{pmatrix}$$

$$\Sigma_D = \begin{pmatrix} \sigma_{TD}^2 & \sigma_{TD} \sigma_{BD} \rho_D \\ \sigma_{TD} \sigma_{BD} \rho_D & \sigma_{BD}^2 \end{pmatrix}$$

The correlation parameters  $\rho_F$  and  $\rho_D$  are of particular scientific interest since they tell us how tarsus length and back color intensity correlate due to environmental factors ( $\rho_F$ ) and genetic factors ( $\rho_D$ ). The mathematical formulation may look quite intimidating, but the corresponding formula syntax of brms is much more concise in comparison:

```
bform_bt1 <-  
  bf(tarsus ~ sex + hatchdate + (1|p|fosternest) + (1|q|dam)) +  
  bf(back ~ sex + hatchdate + (1|p|fosternest) + (1|q|dam)) +  
  set_rescor(TRUE)
```

As can be seen in the model code, we have used two `bf` calls and added them together via a `+` to combine multiple univariate models. This notation tells brms that both `tarsus` and `back` are separate response variables each with their own model formula. Since the right-hand sides of both formulas are the same, we could have also used a shorter notation via

```
bf(mvbind(tarsus, back) ~ sex + hatchdate + (1|p|fosternest) + (1|q|dam))
```

The `mvbind` notation is internally expanded to two `bf` calls. So using `mvbind` is really just for convenience in case you are modeling many response variables with the same predictors. We add `set_rescor(TRUE)` to ensure that residual correlations are being estimated. Inside the model formulas, the term `(1|p|fosternest)` indicates a varying intercept over `fosternest`. By writing `|p|` in between, we indicate that the varying effects of `fosternest` should be modeled as correlated. This makes sense since we actually have two model parts, one for `tarsus` and one for `back`. The indicator `p` is arbitrary and can be replaced by other symbols that comes into your mind. Similarly, the term `(1|q|dam)` indicates correlated varying effects of the genetic mother of the chicks.

With the model formula prepared, we are ready to fit our first multivariate brms model:

```
fit_bt1 <- brm(bform_bt1, data = BTdata, chains = 2, cores = 2)
```

The model results are readily summarized by

```
summary(fit_bt1)
```

```

Family: MV(gaussian, gaussian)
Links: mu = identity
       mu = identity
Formula: tarsus ~ sex + hatchdate + (1 | p | fosternest) + (1 | q | dam)
         back ~ sex + hatchdate + (1 | p | fosternest) + (1 | q | dam)
Data: BTdata (Number of observations: 828)

Multilevel Hyperparameters:
~dam (Number of levels: 106)
                                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(tarsus_Intercept)                  0.49     0.05     0.40     0.59 1.00    953
sd(back_Intercept)                   0.25     0.08     0.10     0.40 1.00    347
cor(tarsus_Intercept,back_Intercept) -0.51     0.22    -0.94    -0.07 1.00    499

~fosternest (Number of levels: 104)
                                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(tarsus_Intercept)                  0.26     0.05     0.16     0.37 1.00    752
sd(back_Intercept)                   0.35     0.06     0.23     0.46 1.00    655
cor(tarsus_Intercept,back_Intercept)  0.70     0.21     0.24     0.99 1.00    280

Regression Coefficients:
                                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
tarsus_Intercept      -0.41     0.07    -0.54    -0.27 1.00    1614    1332
back_Intercept        -0.01     0.07    -0.14     0.12 1.00    2777    1356

```

tarsus_sexMale	0.77	0.06	0.66	0.89	1.00	3770	1545
tarsus_sexUNK	0.22	0.13	-0.03	0.47	1.00	3960	1658
tarsus_hatchdate	-0.04	0.06	-0.16	0.07	1.00	1807	1431
back_sexMale	0.01	0.07	-0.13	0.14	1.00	4401	1505
back_sexUNK	0.15	0.15	-0.15	0.44	1.00	4673	1297
back_hatchdate	-0.09	0.05	-0.20	0.02	1.00	2673	1791

Further Distributional Parameters:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma_tarsus	0.76	0.02	0.72	0.80	1.00	2331	1503
sigma_back	0.90	0.02	0.85	0.95	1.00	2289	1181

Residual Correlations:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
rescor(tarsus,back)	-0.05	0.04	-0.13	0.02	1.01	3099	1496

The summary output indicates that, across dams, tarsus length and back color seem to be negatively correlated, while across foster nests the opposite is true. This indicates differential effects of genetic and environmental factors on these two chick characteristics. Further, the small residual correlation `rescor(tarsus, back)` on the bottom of the output indicates that there is little unmodeled dependency between tarsus length and back color.

Although not necessary at this point, we will already compute and store the LOO-CV results of `fit_bt1`, which we will use for model comparisons.

```
fit_bt1 <- add_criterion(fit_bt1, "loo")
```

Next, let's take a look at some posterior-predictive checks, which give us a first impression of the model fit.

```
pp_check(fit_bt1, resp = "tarsus")
pp_check(fit_bt1, resp = "back")
```

This looks pretty solid, at least marginally, but we notice a slight unmodeled left skewness in the distribution of `tarsus`. We will come back to this later on.

Next, we want to investigate how much variation in the response variables can be explained by our model. We use a Bayesian generalization of the  $R^2$  coefficient for this purpose (see also Section 2.5.1).

```
bayes_R2(fit_bt1)
```

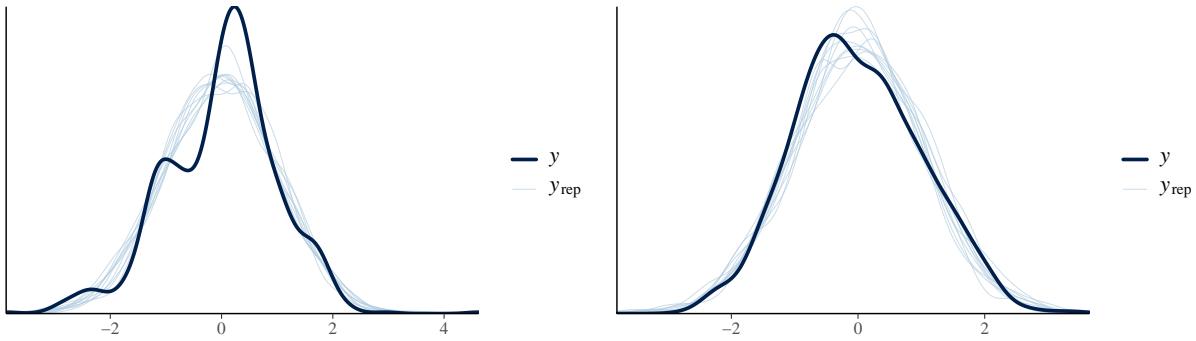


Figure 6.5: Default posterior predictive check for both response variables of model `fit_bt1`.

Table 6.6: Posterior summary of the Bayes  $R^2$  coefficients for model `fit_bt1`.

	Estimate	Est.Error	Q2.5	Q97.5
R2tarsus	0.43	0.02	0.39	0.48
R2back	0.20	0.03	0.14	0.25

Clearly, there is much variation in both animal characteristics that we cannot explain, but apparently we can explain more of the variation in tarsus length than in back color. Not only is `dam` explaining more variance in `tarsus` length than in `back` color, but also do we find a clear `sex` effect only for `tarsus` (male chicks are larger) but not for `back`. For the `hatchdate` predictor it is actually the other way around (larger coefficient for `back` than for `tarsus`) but the difference is less pronounced.

Let's graphically illustrate the difference in `sex` effect across the two response variables (see Figure 6.6):

```
conditional_effects(fit_bt1, "sex", resp = "tarsus")
conditional_effects(fit_bt1, "sex", resp = "back")
```

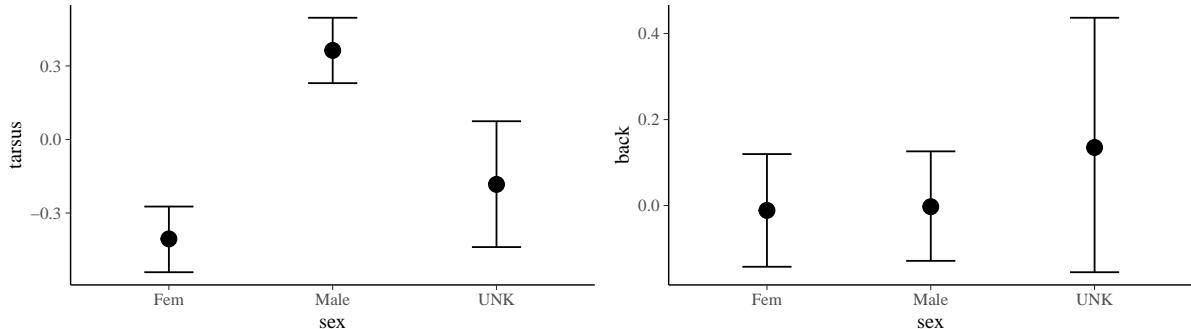


Figure 6.6: Conditional effects of `sex` on both `tarsus` length and `back` color for model `fit_bt1`.

Here again, it is important to specify the `resp` argument in the post-processing method, if we want to access the predictions of specific response variables. As we can see in Figure 6.6, there is a notable sex difference only for tarsus length. Further, the predicted tarsus length of the unknown sex category (UNK)<sup>241</sup> is somewhere in the middle of the two sexes, suggesting that the both some male and female chicks remained unidentified. The relatively larger credible intervals for the unknown category can be explained by the fact that only about 6% of chicks fell into this category.

### 6.4.1 Different formulas across models

Now, suppose we only want to control for `sex` in `tarsus` but not in `back` and vice versa for `hatchdate`. Given the size of the data, estimating one predictor more or less is unlikely to lead to any overfitting, but it allows us to illustrate that we can use different formulas for different response variables.

```
bf_tarsus <- bf(tarsus ~ sex + (1|p|fosternest) + (1|q|dam))
bf_back <- bf(back ~ hatchdate + (1|p|fosternest) + (1|q|dam))

fit_bt2 <- brm(bf_tarsus + bf_back + set_rescor(TRUE),
                 data = BTdata, chains = 2, cores = 2)
fit_bt2 <- add_criterion(fit_bt2, "loo")
```

Let's find out how model fit changed due to excluding certain effects from the initial model:

```
loo_compare(loo(fit_bt1), loo(fit_bt2))

      elpd_diff  se_diff
fit_bt2  0.0      0.0
fit_bt1 -2.5     1.4
```

Apparently, there is no noteworthy difference in the model fit. Accordingly, we do not really need to model `sex` and `hatchdate` for both response variables, but there is also no harm in including them, so we decide to keep both in both model parts.

### 6.4.2 Different families across models

So far, the multivariate models used the same `family` argument across model components. We will now deviate from this pattern and assign different families to different sub-models. Remember the slight left skewness of `tarsus`, which we will model with a `skew_normal` family instead of `gaussian`. Families can simply be added to `brmsformula` objects (as done below) or provided via the `family` argument of `bf`.

```
bf_tarsus_bt3 <- bf(tarsus ~ sex + hatchdate + (1|p|fosternest) + (1|q|dam)) +
  skew_normal()

bf_back_bt3 <- bf(back ~ sex + hatchdate + (1|p|fosternest) + (1|q|dam)) +
  gaussian()
```

Since don't have a multivariate normal (or Student-t) model anymore, estimating residual correlations in brms is no longer directly possible. The reason is that we lack an explicit multivariate distribution for which one marginal is normal and the other one is skew normal. There are ways to work around that, which we will cover in a bit. For now, we will explicitly deactivate the estimation of residual correlations by adding `set_rescor(FALSE)`.

```
fit_bt3 <- brm(
  bf_tarsus_bt3 + bf_back_bt3 + set_rescor(FALSE),
  data = BTdata, chains = 2, cores = 2,
  control = list(adapt_delta = 0.95)
)
fit_bt3 <- add_criterion(fit_bt3, "loo")
```

Again, we summarize the model and look at some posterior-predictive checks.

```
summary(fit_bt3)
```

```
Family: MV(skew_normal, gaussian)
Links: mu = identity
       mu = identity
Formula: tarsus ~ sex + hatchdate + (1 | p | fosternest) + (1 | q | dam)
         back ~ sex + hatchdate + (1 | p | fosternest) + (1 | q | dam)
Data: BTdata (Number of observations: 828)

Multilevel Hyperparameters:
~dam (Number of levels: 106)
                                         Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(tarsus_Intercept)                  0.47      0.05    0.38    0.57 1.00     628
sd(back_Intercept)                   0.25      0.07    0.10    0.39 1.01     287
cor(tarsus_Intercept,back_Intercept) -0.54      0.22   -0.96   -0.12 1.01     369

~fosternest (Number of levels: 104)
                                         Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(tarsus_Intercept)                  0.26      0.06    0.16    0.38 1.00     526
sd(back_Intercept)                   0.35      0.06    0.23    0.47 1.00     437
cor(tarsus_Intercept,back_Intercept)  0.67      0.22    0.15    0.98 1.01     170

Regression Coefficients:
                                         Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
tarsus_Intercept      -0.41      0.07   -0.55   -0.28 1.00     1065    1438
back_Intercept        -0.01      0.06   -0.14    0.11 1.00     1579    1555
tarsus_sexMale        0.78      0.06    0.66    0.89 1.00     4203    1508
```

tarsus_sexUNK	0.22	0.12	-0.01	0.45	1.00	2839	1516
tarsus_hatchdate	-0.04	0.06	-0.16	0.07	1.00	948	1380
back_sexMale	0.01	0.07	-0.12	0.14	1.00	3626	1389
back_sexUNK	0.15	0.15	-0.16	0.44	1.00	3220	1309
back_hatchdate	-0.09	0.05	-0.19	0.01	1.00	1474	1628

#### Further Distributional Parameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sigma_tarsus	0.76	0.02	0.72	0.80	1.00	2164	1211		
sigma_back	0.90	0.02	0.85	0.95	1.00	2684	1310		
alpha_tarsus	-1.19	0.46	-1.88	0.20	1.00	1561	774		

Under **Further Distributional Parameters** in the `summary` output, we see from the negative `alpha` (skewness) parameter of `tarsus` that the residuals are indeed slightly left-skewed, which we also see reflected in a posterior predictive check (Figure 6.7).

```
pp_check(fit_bt3, resp = "tarsus")
```

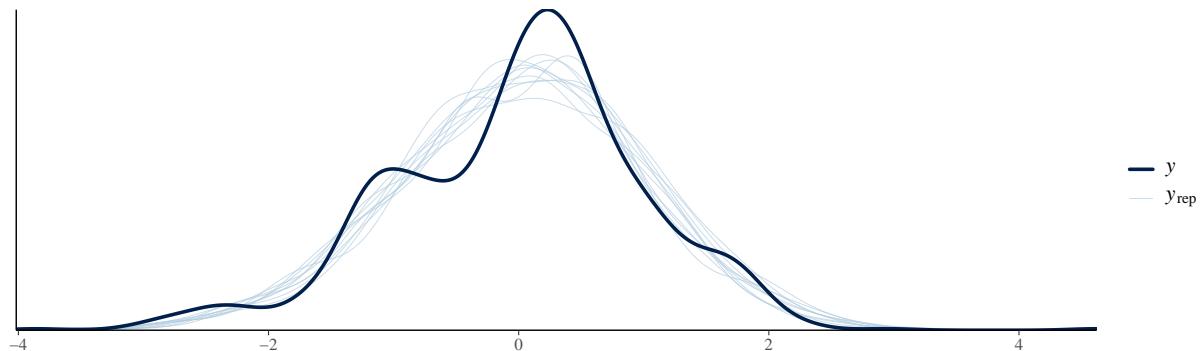


Figure 6.7: Default posterior predictive check for both response variables of model `fit_bt1`.

The model parameters of interest, specifically the hyperparameters of the varying coefficients of `fosternest` and `dam` change little. Apparently, already the multivariate normal model, with or without residual correlations, is doing a good job in capturing the unmodeled uncertainty in the responses.

#### 6.4.3 Modeling residual correlations in the general case

As explained above, unless all responses use a `gaussian` or `student` family, brms does not yet directly support modeling residual correlations. In the current case study, the residual

correlation between tarsus length and back color was very small. So ignoring it for the purpose of changing the families would likely be a good tradeoff – had other families shown any improvements in model fit.

Yet, for other datasets and models, we may be forced to use different families, for example, if one response was count data while the other response was a positive continuous variable. So how are we going to model residual correlations in such cases? The answer has multiple layers. Let's unpack them step by step.

In order to model residual correlations, we need to have residuals in our model in the first place. Let us start with a hypothetical model of the two responses  $y$  and  $z$ , where  $y$  is a count variable and  $z$  is a positive continuous variable. We choose to use a `poisson` family for  $y$  and an `exponential` family for  $z$ . Assuming a single predictor  $x$ , a basic multivariate brnms model without residual correlations would then look as follows:

```
bform_y <- bf(y ~ x) + poisson()
bform_z <- bf(z ~ x) + exponential()
bform <- bform_y + bform_z + set_rescor(FALSE)
```

Since neither the Poisson nor the exponential distribution have an overdispersion parameter, they do not have residuals in the way we would need it here. Ironically, this works in our favor, since we can add residuals to the model via observation-level varying coefficients, that is one coefficient per observation (see also Section 5.3.1). We define a variable `obs` in our data that is simply counting observations. Then, we can add both residual terms and their correlation to the model as follows:

```
bform_y <- bf(y ~ x + (1|r|obs)) + poisson()
bform_z <- bf(z ~ x + (1|r|obs)) + exponential()
bform <- bform_y + bform_z + set_rescor(FALSE)
```

Remember using the same indicator, here `r`, in-between `||` in both varying terms will imply modeling correlations between the coefficients, which in our case qualifies as the residual correlation between  $y$  and  $z$ .

Now suppose that  $z$  would actually be an *unbounded* continuous variable that we would model as normally distributed. Generalizing from the above formulation, we would have:

```
bform_y <- bf(y ~ x + (1|r|obs)) + poisson()
bform_z <- bf(z ~ x + (1|r|obs)) + gaussian()
bform <- bform_y + bform_z + set_rescor(FALSE)
```

If we fitted such a model, we would run into serious convergence issues because we now had two sets of residuals for the  $z$  response: one expressed via `(1|r|obs)` and the other one expressed

via the residual standard deviation `sigma`. The model is unidentifiable because we have multiple parameters to model the same thing. We can try to work around it by fixing `sigma` to a small value, for instance:

```
bform_z <- bf(z ~ x + (1|r|obs), sigma = 0.01) + gaussian()
```

However, if the fixed `sigma` is too small, we again get convergence issues, because the posterior becomes close to non-differentiable and Hamiltonian Monte Carlo fails. And if we fix `sigma` to a too large value, only parts of the residuals will be modeled as correlated while the rest remains uncorrelated. In other words, fixing overdispersion parameters such as `sigma` is not a pretty solution. But it can work as a reasonable approximation, at least for `gaussian` families.

For other families, such as `skew_normal`, it may unfortunately fail, even if we fix `sigma` to a value that is neither too high nor too low. The reason is that the influence of the skewness parameter `alpha` on the modeled skewness depends on `sigma`. Accordingly, if we fix `sigma`, we are effectively changing the meaning of `alpha`. Let me illustrate this behavior on our case study. We first add the `obs` index over observations to our dataset:

```
BTdata$obs <- 1:nrow(BTdata)
```

Then, we specify the extended formula with observation-level varying coefficients and both `sigma` fixed to 0.1. This is small relative to the scale of the responses but not too small to run into serious convergence issues for this example (e.g., `sigma = 0.01` would already fail here).

```
bf_tarsus_bt4 <- bf(
  tarsus ~ sex + hatchdate + (1|p|fosternest) +
  (1|q|dam) + (1|r|obs),
  sigma = 0.1
) + skew_normal()

bf_back_bt4 <- bf(
  back ~ sex + hatchdate + (1|p|fosternest) +
  (1|q|dam) + (1|r|obs),
  sigma = 0.1
) + gaussian()
```

```
fit_bt4 <- brm(
  bf_tarsus_bt4 + bf_back_bt4 + set_rescor(FALSE),
  data = BTdata, chains = 2, cores = 2,
  control = list(adapt_delta = 0.95)
)
```

We can run this model and get reasonable convergence. If we investigate the residual correlation parameter, now to be located under `Multilevel Hyperparameters` in the `summary` output, we see find a posterior mean of  $-0.05$  with a 95% CI of  $[-0.12, 0.03]$ , which is almost the same as what we found for our initial model `fit_bt1`. This indicates that we were indeed able to reasonably estimate the residual correlation with our workaround. However, the skewness parameter `alpha_tarsus` now looks completely different than previously with posterior mean of  $0.01$  and large 95% CI of  $[-6.70, 8.99]$ , essentially indicating no evidence of skewness. Apparently, in our effort to include residual correlations, we accidentally broke the model in another place.

I have seen many people use this approach, sometimes in valid and sometimes in invalid circumstances. I hope these explanations help you in understand when modeling residual correlations via observation-level coefficients is a reasonable approach and when we should better avoid it. In the latter cases, there is currently no way to modeling residual correlations in brms, but there may be in the future. I hope that in the second version of this book, I can cover it.

## 6.5 Summary

In this chapter, we have seen how multivariate models can be specified in brms. They simply consist of a set of univariate models whose formulas are added together. Multivariate models only offer a benefit over separate univariate models, if the univariate models are somehow modeled as dependent. We have discussed two main mechanisms to establish such dependencies. First via residual correlations, if permitted by the assumed likelihoods, and second via correlations of multilevel coefficients. These two are not the only ways to establish dependencies between univariate models. For example, one could imagine equating parameters across model components, such that these parameters would be informed by multiple responses. At the time of writing this chapter, this feature is not yet implemented in brms, but it is likely to be available in the future. More advanced ways of multivariate modeling include so-called *joint modeling* where the predictor terms of one response is also used as part of the predictor term of another response. Such joint models can be highly useful in dose-response and time-to-event modeling where multiple responses are tracked over the course of a longer experiment (Brilleman et al. 2019). These kinds of models are often highly problem-specific. While it may be possible to implement them in brms with a lot of Stan code injection, other software packages focusing on joint models are likely to be more useful for this purpose.

# 7 Meta-Analysis

## 7.1 Setup

```
library(ggplot2)
library(patchwork)
library(tidyr)
library(dplyr)
library(metafor)
library(brms)
library(priorsense)
```

## 7.2 Introduction

The field of meta-analysis is concerned with aggregating results from multiple studies. Despite increasing efforts in open science, raw data of the primary studies is still rarely available. Instead, we have to extract summary statistics about the effects of interest from the results reported in the corresponding papers. Finding all relevant papers and then extracting all the required information are some of the major challenges of meta-analysis. Since this book is about statistics, I will not talk about these topics in detail here. You can read more about the whole process of meta-analysis, for example, in Cooper, Hedges, and Valentine (2019). In my experience, after having contributed to many meta-analyses, the statistics is often one of the easier parts of the whole workflow. That doesn't mean it is without its challenges, as we will see.

### 7.2.1 Computation of effect sizes

In standard meta-analyses, we are not working with the raw data of the studies but with summary statistics, usually directly extracted from the corresponding papers. We then transform the summary statistics into standardized effect size measures that can be compared and aggregated across studies. To illustrate this procedure, let's consider a real-world case study. We will meta-analyse the effectiveness of the Bacillus Calmette-Guerin (BCG) vaccine against tuberculosis. This vaccine uses living attenuated bacteria that are similar to the dangerous

bacteria causing tuberculosis. Being exposed to the harmless bacteria will then hopefully increase the immune response to their dangerous counterparts. Concretely, we will use a dataset from the meta-analysis of Colditz et al. (1994) containing data of 13 studies, which is shipped with the metadat package (Viechtbauer et al. 2025):

```
data("dat.bcg", package = "metadat")
dat_bcg <- dat.bcg
```

Table 7.1: First six rows of the BCG data.

author	year	tpos	tneg	cpos	cneg	ablat	alloc
Aronson	1948	4	119	11	128	44	random
Ferguson & Simes	1949	6	300	29	274	55	random
Rosenthal et al	1960	3	228	11	209	42	random
Hart & Sutherland	1977	62	13536	248	12619	52	random
Frimodt-Møller et al	1973	33	5036	47	5761	13	alternate
Stein & Aronson	1953	180	1361	372	1079	44	alternate

The first few rows of the BCG dataset are shown in Table 7.1. All studies have a treatment group, which received the vaccine, and a control group which did not. It was then recorded how many people in the treatment group got Tuberculosis within the study period (variable `tpos`) and how many did not (variable `tneg`). The same information was recorded for the control group and is available in the variables `cpos` and `cneg`, respectively.

We could choose to model these variables directly, and we will do so later on in Section 7.6. However, for now, we will go the standard route of meta-analysis and aggregate them into a single effect size per study. We choose the log risk ratio to be computed as follows. First, for each study  $i$ , we calculate the positive rates  $p_{Ti} = a_i/(a_i + b_i)$  and  $p_{Ci} = c_i/(c_i + d_i)$ , where  $a_i$  and  $b_i$  is the number of Tuberculosis positives and negatives in the treatment group, respectively, and  $c_i$  and  $d_i$  are the corresponding numbers in the control group. On this basis, we calculate the log risk ratio as

$$y_i := \log \text{RR}_i = \log(p_{Ti}/p_{Ci}).$$

We use the log risk ratio rather than the just risk ratio because the former is unbounded and approximately normally distributed already for smaller sample sizes. We will need this assumption at least for the more basic meta-analytic models presented first. Every effect size measure comes with some uncertainty induced by the fact that it was computed from the finite data of a given study. We usually express this uncertainty via an (approximate) standard error  $s_i := \widehat{SE}(y_i)$  of the effect size  $y_i$ . If  $y_i$  is sufficiently normally distributed,  $s_i$  will indeed be

a good measure of the uncertainty induced by sample variation. For the log risk ratio, an approximate standard error can be computed as

$$s_i := \sqrt{\frac{1}{a_i} - \frac{1}{a_i + b_i} + \frac{1}{c_i} - \frac{1}{c_i + d_i}}.$$

The metafor package (Viechtbauer 2010) provides us with a nice interface to compute all sorts of effect sizes. In order to compute the log risk ratio, we have to specify `measure = "RR"` and provide the required data:

```
dat_bcg <- metafor::escalc(
  measure="RR", ai=tpos, bi=tneg, ci=cpos, di=cneg, data=dat_bcg
)
```

By default, the effect sizes will be stored in the variable `yi`. The corresponding error variance will be stored in `vi`. To obtain the standard error, we add another new variable to our dataset:

```
dat_bcg$sei <- sqrt(dat_bcg$vi)
```

In our example, and in the vast majority of past and present meta-analyses,  $y_i$  and  $s_i$  are frequentist estimates. But this won't stop us from doing Bayesian meta-analysis with them. While there are fully Bayesian procedures that include also the summary statistics and effect sizes (Blomstedt et al. 2024), we will usually do just fine by working with frequentist effect size estimates, even if we end up analyzing them with Bayesian models afterwards.

### 7.3 Fixed effects meta-analysis

The simplest meta-analytic models are called “fixed-effects models” referring to the their key assumption that all included studies share the same underlying true effect size. For our example of the BCG vaccine this means the model assumes the effectiveness of the vaccine to be constant across all studies. The likelihood of the fixed effects model is simply

$$y_i \sim \text{normal}(\mu, s_i),$$

where  $\mu$  is our only model parameter representing the average (true) effect size across studies. Since we are in a Bayesian framework, we can also add a prior  $p(\mu)$  to our model. In meta-analysis, adding prior information will be particularly important if we have data of only few

studies or if the individual studies are small such that their  $s_i$  are large. In most basic meta-analytic models, we usually assume a normal likelihood, which is why it is important to use effect sizes that are at least approximately satisfying that assumption.

In essence, basic meta-analysis just comes down to computing weighted averages. In fixed effects meta-analysis, the weight of each studies' likelihood component is given by  $w_i = 1/s_i^2$ . This implies that studies with smaller standard errors will receive higher weights because we are more certain about their effect sizes. For the BCG data, Figure 7.1 shows the log risk ratios plotted against the weight of the study as implied by the standard error (left) and against the sample size of the study (right). As we can deduct from comparing these two plots, the weight is largely proportional to the sample size. When performing meta-analysis with appropriate software, the weighting will happen implicitly, so we do not need to worry about it. Yet, it's good to understand that most common meta-analytic models are essentially just weighted linear regressions.

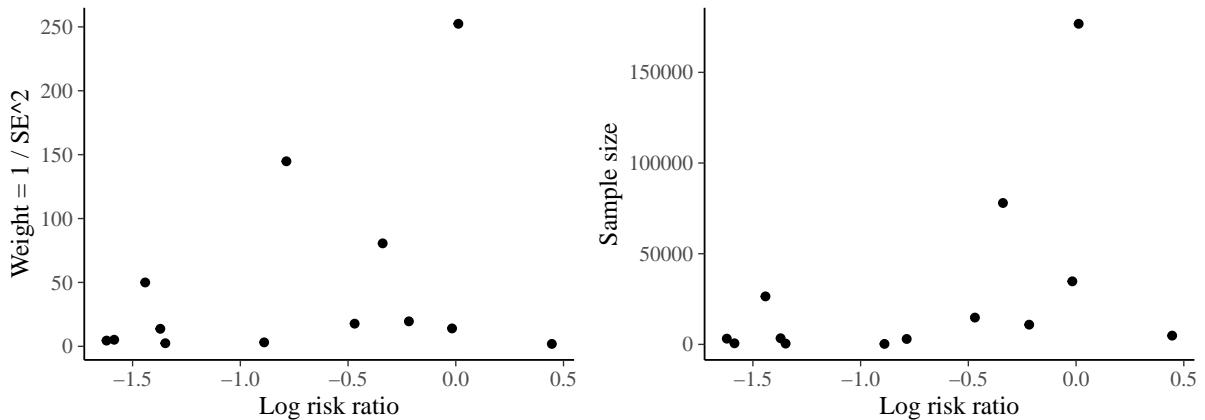


Figure 7.1: BCG data. Scatter plot of the log risk ratio estimates vs. their weight computed as the inverse of the squared standard error (left) and the total sample size of the study (right).

We now have everything together to fit our first meta-analytic model with brms. The main difference to standard linear regression is that we need to tell the model to consider the known standard errors  $s_i$  as replacement of the residual standard deviation parameter  $\sigma$  in the likelihood. This is what we achieve via `se` addition terms, which we can specify on the left-hand side of the formula after the response variable. For the BCG data, the formula for the fixed effects models is simply `yi | se(sei) ~ 1`. When also adding a weakly informative prior `normal(0, 1)` prior on  $\mu$  (represented my the intercept parameter), the whole model specification is given by:

```
fit_bcg_fe <- brm(
  yi | se(sei) ~ 1, data = dat_bcg,
  prior = prior(normal(0, 1), class = "Intercept")
```

```
)
```

```
summary(fit_bcg_fe)
```

```
Family: gaussian  
Links: mu = identity  
Formula: yi | se(sei) ~ 1  
Data: dat_bcg (Number of observations: 13)
```

Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-0.43	0.04	-0.51	-0.35	1.00	1594	2501

Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	0.00	0.00	0.00	0.00	NA	NA	NA

The summary looks pretty much like that of a basic intercept only model. The only difference is that `sigma` was automatically fixed to zero as we replaced it with known standard errors via the `se` addition term. The `Intercept` coefficient tells us that the average log risk ratio  $\mu$  is estimated clearly negatively with a 95% CI of  $[-0.51, -0.35]$ . In other words, the BCG vaccine clearly reduces that risk of Tuberculosis infection on average across studies. Yet, apart from the sign, the log risk ratio values hard to interpret. So let's transform it to back to the risk ratio scale, for example via the hypothesis function:

```
(hyp <- hypothesis(fit_bcg_fe, "exp(Intercept) = 0"))
```

Hypothesis Tests for class b:

	Hypothesis	Estimate	Est.Error	CI.Lower	CI.Upper	Evid.Ratio	Post.Prob	Star
1	$\exp(\text{Intercept}) = 0$	0.65	0.03	0.6	0.7	NA	NA	*

```
plot(hyp)
```

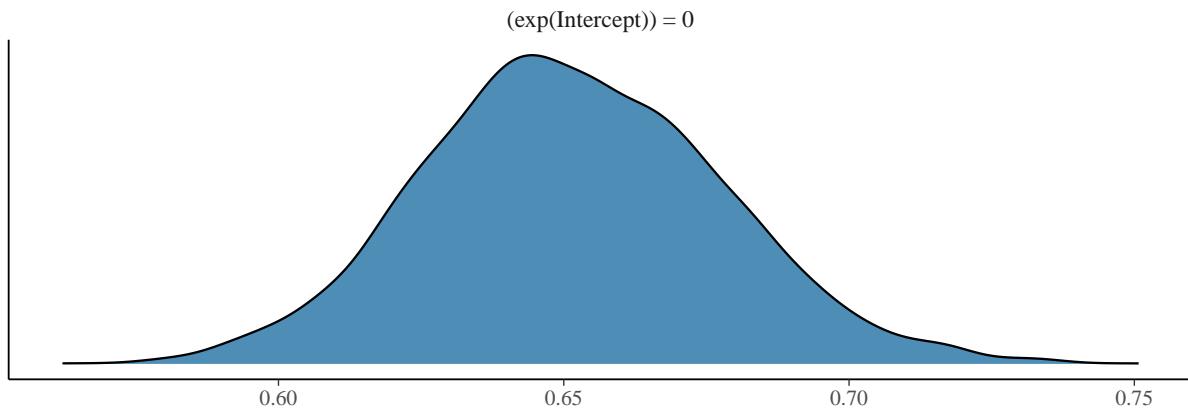


Figure 7.2: Posterior of the risk ratio as estimated by model `fit_bcg_fe`.

Both the printed output and the corresponding plot (Figure 7.2) indicate that the average risk ratio is roughly between 0.6 and 0.7. Accordingly, the BCG vaccine clearly reduces the risk of getting infected by Tuberculosis. Is this an estimate we can trust though? Let's perform some basic posterior predictive checking to get a feel for how good our fixed effects model fits the data:

```
pp_check(fit_bcg_fe)
```

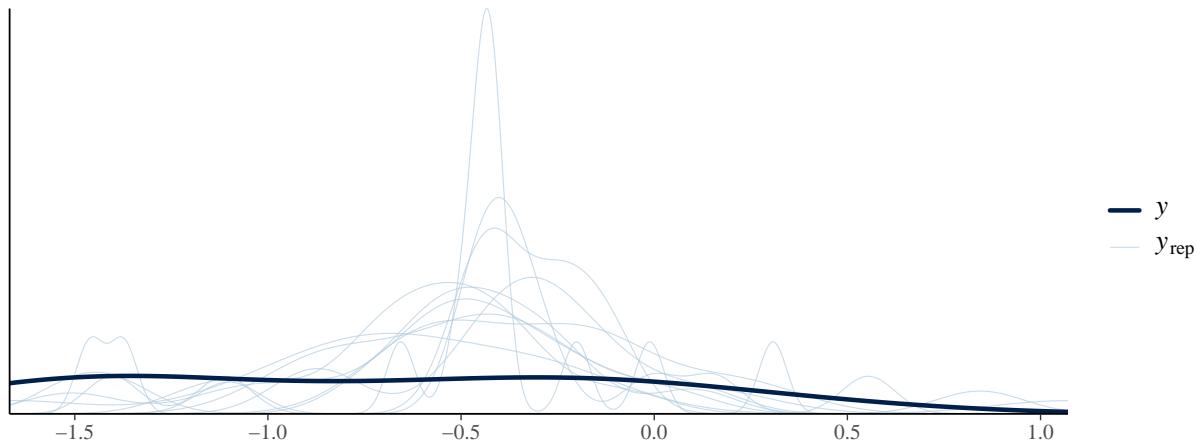


Figure 7.3: Posterior predictive check for model `fit_bcg_fe`.

As we see in Figure Figure 7.3, the model strongly underestimates the variation in the observed effect sizes. Apparently, there is more variation than is covered by the known standard errors  $s_i$ . In other words, there must be variation in the *true* effectiveness of the BCG vaccine across studies. This is what we will investigate next.

## 7.4 Random effects meta-analysis

Meta-analyses assuming the true effect sizes to vary are generally known as “random effects models”. As with the term “fixed effects”, “random effects” has its origins in a frequentist perspective. However, both of these terms are so common in the meta-analysis literature that I am using them here as well even though we are doing Bayesian meta-analysis.

Mathematically, the random effects models generalizes the fixed effects model by assuming a study specific true effect size  $\mu_i$  coming from a distribution of true effect sizes with mean  $\mu$  and standard deviation  $\sigma$ :

$$y_i \sim \text{normal}(\mu_i, s_i)$$

$$\mu_i \sim \text{normal}(\mu, \sigma)$$

In random effects meta-analysis, the study weights are given by  $w_i = 1/(s_i^2 + \sigma^2)$ . Thus, both the within-study variation, quantified by  $s_i$ , and the between-study variation, quantified by  $\sigma$ , play a role in how much each study contributes to the average effect size.

We will estimate both the  $\mu_i$  parameters and the  $\mu$  and  $\sigma$  hyperparameters from the data. So, with the exception of using the known standard errors  $s_i$ , this is just a basic multilevel model with varying intercepts. To obtain an adequate grouping variable, we construct our study indicator by the combination of study authors and year:

```
dat_bcg <- dat_bcg %>%
  mutate(study = paste(author, year))
```

Now, we can easily specify a random effects meta-analysis in brms using its multilevel syntax by adding the term `(1 | study)`. Considering some weakly informative priors, the full random effects model looks as follows:

```
fit_bcg_re1 <- brm(
  yi | se(sei) ~ 1 + (1 | study), data = dat_bcg,
  prior = prior(normal(0, 1), class = "Intercept") +
    prior(normal(0, 0.5), class = "sd")
)
```

```
summary(fit_bcg_re1)
```

```

Family: gaussian
Links: mu = identity
Formula: yi | se(sei) ~ 1 + (1 | study)
Data: dat_bcg (Number of observations: 13)

Multilevel Hyperparameters:
~study (Number of levels: 13)
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept) 0.57 0.14 0.35 0.89 1.00 1251 1721

Regression Coefficients:
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept -0.69 0.19 -1.04 -0.30 1.01 1007 1163

Further Distributional Parameters:
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma 0.00 0.00 0.00 0.00 NA NA NA

```

Under this model, the posterior of the Intercept coefficient (i.e.,  $\mu$ ) has a mean of roughly  $-0.68$  with a 95% CI of  $[-1.06, -0.30]$ . Accordingly, compared to the fixed effects model the average effect size has gotten larger but more uncertain. The larger uncertainty directly results from the assumption of varying true effect sizes across studies. We have seen the same process in other multilevel models also previously in Chapter 4. The reason for the change in posterior mean is a bit more subtle: When we go back to Figure 7.1, we see that studies finding stronger (i.e., more negative) log risk ratios have higher standard errors associated standard errors as a result of having smaller sample sizes. Now, when we assume a random effects model, all the study weights include the constant term  $\sigma$ , which brings the weights closer together compared to the weights in the fixed effects models. As a result, the higher standard error studies get a larger weight, which, in this case, pushes the average effect size estimate to more negative values. In our case, the posterior mean of  $\sigma$  was estimated as  $0.57$  and thus larger than the average of the standard errors  $s_i$ , which is roughly  $0.33$ . If we wanted to, we could even compute the full posterior of the weights under the random effects model by evaluating  $w_i = 1/(s_i^2 + \sigma^2)$  for each posterior draw of  $\sigma$ . I leave this to you as an exercise.

When exponentiating the average log risk ratio (Figure 7.4), we see that the estimated risk ratio is now roughly between  $[-0.35, -0.74]$  according to the 95% CI. Accordingly, we still find clear evidence that, on average across studies, the BCG vaccine is effective against tuberculosis.

```
(hyp <- hypothesis(fit_bcg_re1, "exp(Intercept) = 0"))
```

Hypothesis Tests for class b:

	Hypothesis	Estimate	Est.Error	CI.Lower	CI.Upper	Evid.Ratio	Post.Prob	Star
1	$\exp(\text{Intercept}) = 0$	0.51	0.1	0.35	0.74	NA	NA	*

```
plot(hyp)
```

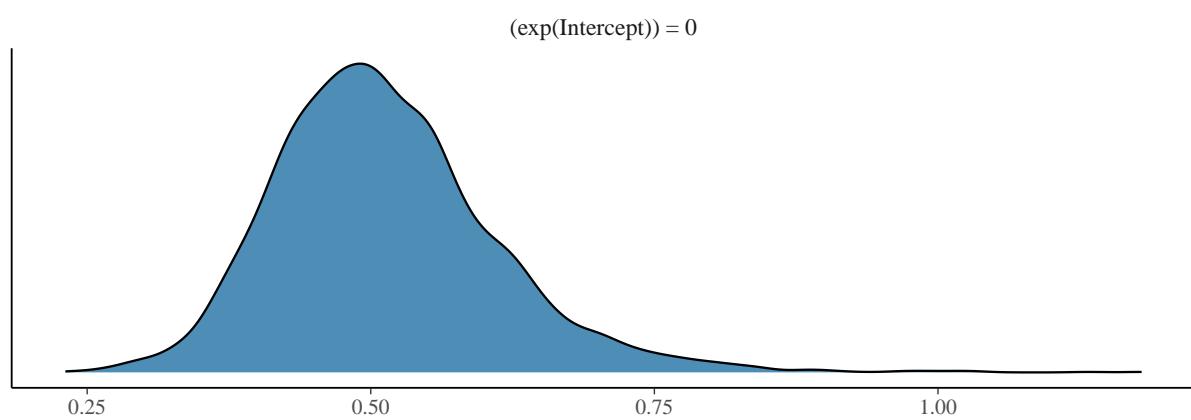


Figure 7.4: Posterior of the risk ratio as estimated by model `fit_bcg_re1`.

Since we are assuming the true effect sizes to vary across studies, we can also look at the individual study estimates, which come in the form of varying intercepts for the `trial` grouping factor. While we could use the `hypothesis` method for this purpose as well, the resulting plot would become a bit convoluted, so we choose to plot posteriors in the form of point ranges (see Figure 7.5):

```
coef(fit_bcg_re1, summary = FALSE)$study[, , "Intercept"] %>%
  exp() %>%
  posterior_summary() %>%
  as_tibble(rownames = "study") %>%
  ggplot(aes(study, Estimate, ymin = Q2.5, ymax = Q97.5)) +
  geom_pointrange() +
  coord_flip() +
  labs(x = "Study", y = "Risk Ratio")
```

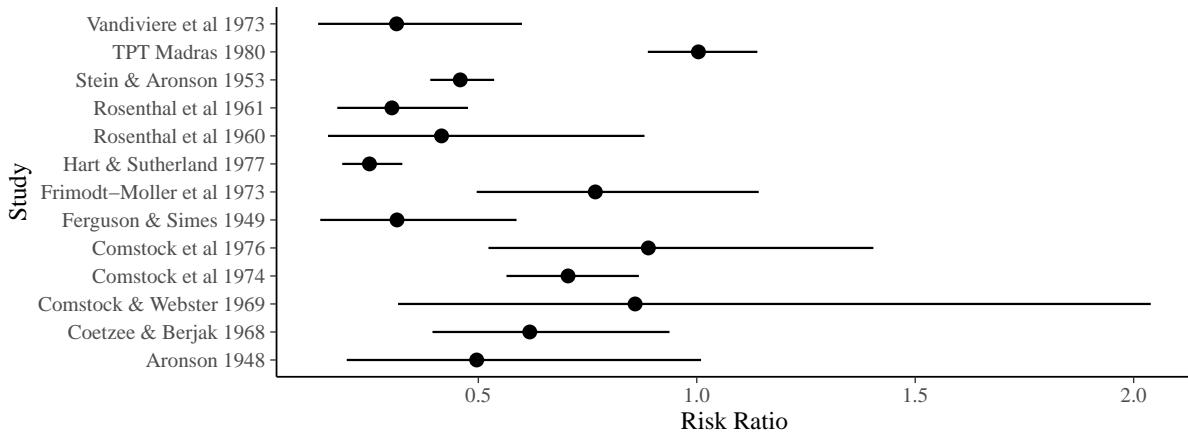


Figure 7.5: Study-specific risk ratio estimates for the `fit_bcg_re1` model.

Clearly, the variation across studies in both posterior means and CIs is quite substantial. But overall, most studies point in the direction of the BCG vaccine being effective (i.e., having a risk ratio smaller than 1).

The default posterior predictive check of the random effects model indicates an overall good marginal fit to the data (Figure 7.6):

```
pp_check(fit_bcg_re1)
```

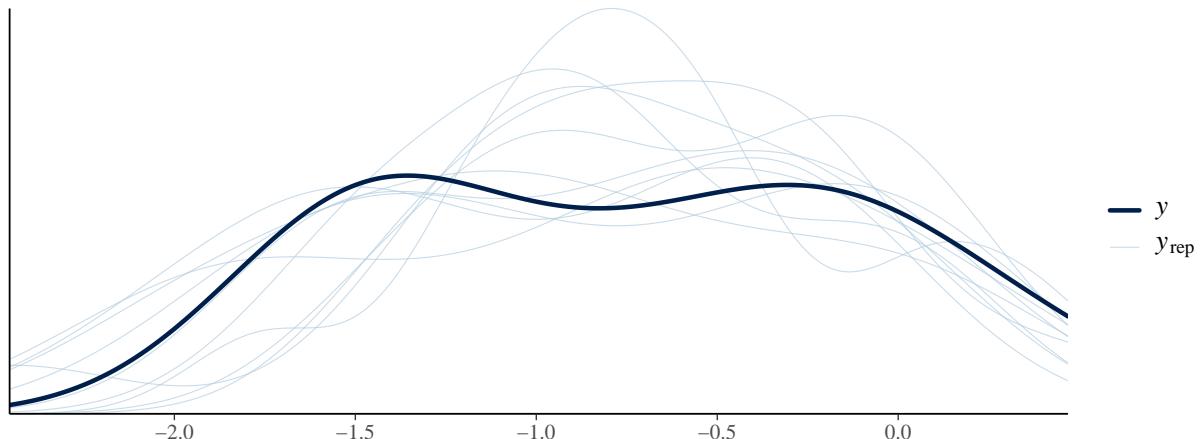


Figure 7.6: Posterior predictive check for model `fit_bcg_re1`.

Knowing the limitations of these checks for the purpose of model comparison, we attempt to additionally perform LOO-CV. However, when trying to run `loo(fit_bcg_re1)`, we get around 10 high Pareto-k warnings. Since, we only have a total of 13 observations (i.e., effect

sizes from 13 studies), this is quite substantial. As you may remember from Section 5.3.1, this issue arises because we are estimating one varying coefficient per observation (i.e.,  $\mu_i$  for each study  $i$ ); a model structure that importance sampling is unable to handle properly in the context of approximate LOO-CV. Since the model is so small, we could easily run exact LOO-CV for the problematic observations. Instead, we will go a different route by marginalizing over study-specific parameters instead.

### 7.4.1 Marginalization of study-specific coefficients

When marginalization over the study-specific parameters  $\mu_i$ , the random effects model can equivalently be written as

$$y_i \sim \text{normal}(\mu, \sqrt{s_i^2 + \sigma^2}). \quad (7.1)$$

That is, the hyperparameters  $\mu$  and  $\sigma$  now appear directly in the likelihood. We also see explicitly that the total residual standard deviation is a combination of the within-study standard error  $s_i$  and the between-study standard deviation  $\sigma$ . In brms, we can express this marginal random effects model by setting argument `sigma = TRUE`, in the `se` addition term. This will not only lead to the family-specific `sigma` being estimated again (instead of fixing it to 0), but also combining  $s_i$  and  $\sigma$  in just the right away, according to Equation 7.1. The complete model is specified as follows:

```
fit_bcg_re2 <- brm(
  yi | se(sei, sigma = TRUE) ~ 1, data = dat_bcg,
  prior = prior(normal(0, 1), class = "Intercept") +
    prior(normal(0, 0.5), class = "sigma")
)
```

```
summary(fit_bcg_re2)
```

```
Family: gaussian
Links: mu = identity
Formula: yi | se(sei, sigma = TRUE) ~ 1
Data: dat_bcg (Number of observations: 13)
```

```
Regression Coefficients:
  Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept -0.69      0.18    -1.05   -0.33 1.00     2872     2279
```

```
Further Distributional Parameters:
```

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	0.57	0.14	0.34	0.89	1.00	3285	2806

As we can see in the summary output, the posterior estimates of  $\mu$  and  $\sigma$  are virtually unchanged. The only difference is that  $\sigma$  appears in a different spot in the summary, since it is now implemented as a regular distributional parameter rather than as a hyperparameter of a multilevel term. All minor changes in the estimates can be fully attributed to Monte Carlo standard error – as it should be for two mathematically equivalent models. Upon closer inspection of the convergence metrics, we see that the marginalized model has somewhat higher effective sample sizes, which reflects the fact that the marginalized posterior is easier to sample from by Stan.

More importantly, we can now run LOO-CV without running into an excessive number of high Pareto-k estimates. The few high Pareto-k estimates in the fixed effects model can easily be fixed via moment matching:

```
loo_bcg_fe <- loo(fit_bcg_fe)
loo_bcg_fe <- loo_moment_match(fit_bcg_fe, loo_bcg_fe)
loo_bcg_re2 <- loo(fit_bcg_re2)
loo_compare(loo_bcg_fe, loo_bcg_re2)
```

	elpd_diff	se_diff
fit_bcg_re2	0.0	0.0
fit_bcg_fe	-79.2	44.2

LOO-CV results indicate strong absolute evidence for the predictive superiority of the random effects model. The standard error of the ELPD difference is quite large due to the fact that individual observations (i.e., studies in our models) contribute a lot to the ELPD difference. Still, together with the far-away-from-zero posterior estimates of  $\sigma$ , the LOO-CV results provide us with clear evidence that assuming a random effects model is the better way forward.

## 7.5 Moderator analysis

We are now convinced that there is indeed variation in the true effect sizes across studies. But why does the efficacy of the BCG vaccine vary? This is a question that calls for what is known as “moderator analysis”. Essentially, it just means adding predictors to the meta-analysis. One readily available predictor for the BCG data is the year in which the study was published. Perhaps the quality the BCG vaccine has improved over time? If that was the case, we would see an increase in effect sizes over the years. Also other aspects of the studies, such as the scientific rigor of the study-protocol may have changed over time, which was cause the effect

sizes to systematically change. If such a cause was at work, the expected direction of the `year` effect is a priori less clear. It might be that higher quality studies will show increases in effect sizes, for example, because the vaccine is better administered. Or it might decrease effect sizes if earlier studies happened to, say, manually rather than randomly assigning people to treatment and control group, in an (perhaps unconscious) obtain results favorable for the vaccine. These are just few examples of how true effect sizes can vary over time. For our moderator model, we choose to scale the year predictor to simplify prior specification. The whole brms model looks as follows:

```
fit_bcg_mod1 <- brm(
  yi | se(sei, sigma = TRUE) ~ scale(year),
  data = dat_bcg,
  prior = prior(normal(0, 1), class = "Intercept") +
    prior(normal(0, 1), class = "b") +
    prior(normal(0, 0.5), class = "sigma")
)
```

```
summary(fit_bcg_mod1)
```

```
Family: gaussian
Links: mu = identity
Formula: yi | se(sei, sigma = TRUE) ~ scale(year)
Data: dat_bcg (Number of observations: 13)
```

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-0.72	0.17	-1.07	-0.37	1.00	3059	2545
scaleyear	0.28	0.19	-0.09	0.64	1.00	3413	2680

#### Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	0.54	0.14	0.32	0.86	1.00	2651	2587

According to the model, the year coefficient is not clearly different from zero, but still points in a positive direction. If we take the sign of the coefficient's mean serious for now, this would imply that effect sizes have become more positive over time. Since it is more negative log risk ratios that imply stronger vaccine efficacy, this means the the vaccine has been less effective in later studies. For a scaled year of 0 (representing year 1966), we find an average effect of roughly  $-0.72$  according to the posterior mean of `Intercept`. Now, for the later studies included in the meta-analysis, say, a scaled year of 2 (representing year 1988), the posterior mean of the effect would roughly be  $-0.72 + 2 \times 0.28 = -0.16$ . This corresponds to a risk

ratio of only about  $\exp(-0.16) = 0.85$ , so very low vaccine efficacy. We can also visualize this generally positive coefficient of year (Figure 7.7):

```
ce_bcg_mod1 <- conditional_effects(fit_bcg_mod1)
plot(ce_bcg_mod1, points = TRUE)
```

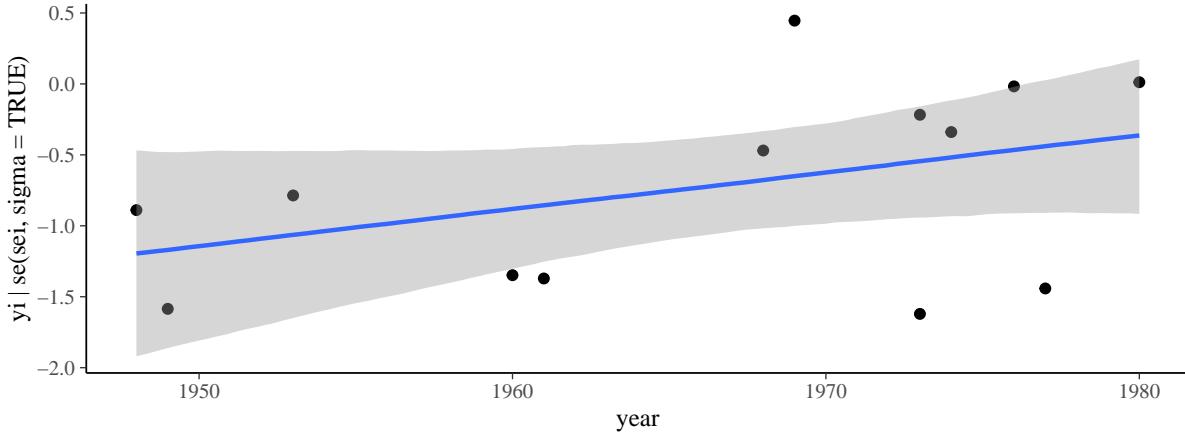


Figure 7.7: Conditional effects of model `fit_bcg_mod1` showing log risk ratio predictions.

The y-axis shows the modeled effect sizes, that is log risk ratios, which are hard to interpret. What is more, plotting the observations as points in `conditional_effects` is a bit misleading in meta-analysis, since equally sized points visually indicate equal weighting. I recommend representing the study weights in the form of point sizes. One way to achieve this is to build on the data frames returned by `conditional_effects` and then write our own ggplot code on top of it:

```
ce_year <- ce_bcg_mod1$year
points_year <- attr(ce_year, "points")
# use the posterior median of sigma to compute points estimates of the study weights
median_sigma <- median(as.matrix(fit_bcg_mod1, variable = "sigma"))
points_year$Weights <- 1 / (dat_bcg$sei^2 + median_sigma^2)
# exponentiate the estimates to plot risk ratios
ggplot(ce_year, aes(x = effect1__, y = exp(estimate__))) +
  geom_smooth(aes(ymin = exp(lower__), ymax = exp(upper__)), stat = "identity") +
  geom_point(aes(y = exp(resp__), size = Weights), data = points_year) +
  labs(x = "Year", y = "Risk ratio") +
  # the range is largely arbitrary and chosen purely for aesthetical reasons
  scale_size_continuous(range = range(points_year$Weights) * 1.5)
```

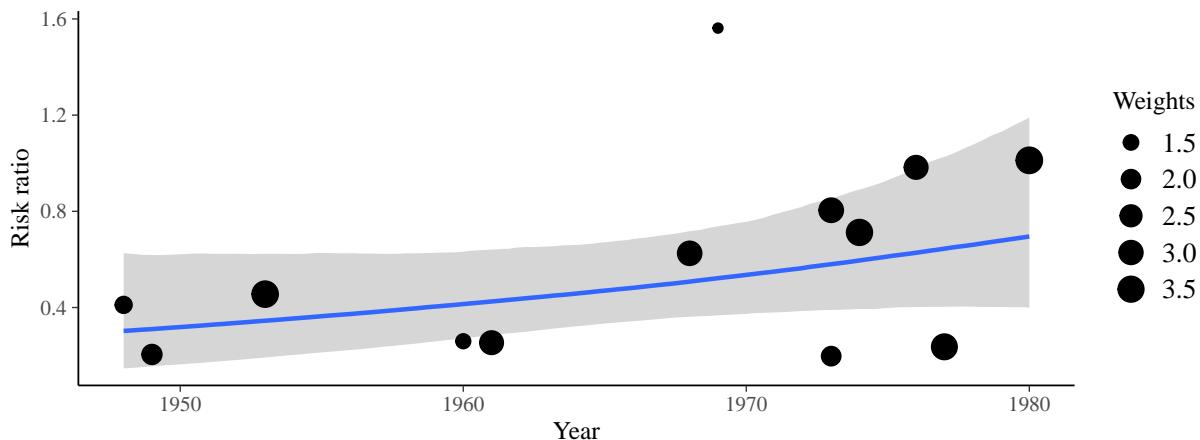


Figure 7.8: Conditional effects of model `fit_bcg_mod1` showing risk ratio predictions and point sizes proportional to the study weights.

The created plot is shown in Figure 7.8. With the help of `brms` and just a few lines of extra code, we have generated a publication-ready figure for our moderator analysis.

But we can trust our finding that the BCG vaccine efficacy may have reduced over time? This essentially becomes a question of causal analysis. Are there perhaps other factors, correlated with time (or whatever else drives the effect of time) that may explain the variation in the true effect sizes? It turns out there is, in the form of the latitude (distance from the equator; variable `ablat`) where the studies were performed. We can add both moderators in the same meta-analytic model just as we do in any other regression model:

```
fit_bcg_mod2 <- brm(
  yi | se(sei, sigma = TRUE) ~ scale(ablat) + scale(year),
  data = dat_bcg,
  prior = prior(normal(0, 1), class = "Intercept") +
    prior(normal(0, 1), class = "b") +
    prior(normal(0, 0.5), class = "sigma")
)
```

```
summary(fit_bcg_mod2)
```

```
Family: gaussian
Links: mu = identity
Formula: yi | se(sei, sigma = TRUE) ~ scale(ablat) + scale(year)
Data: dat_bcg (Number of observations: 13)
```

Regression Coefficients:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-0.71	0.13	-0.98		-0.45	1.00	2822	2315	
scaleablat	-0.40	0.17	-0.70		-0.03	1.00	1514	1313	
scaleyear	0.01	0.18	-0.33		0.39	1.00	1298	1456	

Further Distributional Parameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sigma	0.35	0.16	0.09		0.72	1.00	856	589	

Indeed, the coefficient of latitude is quite clearly different from zero, while the coefficient of year is now indistinguishable from zero. Let's investigate the conditional effects of latitude directly with our prettified version of `conditional_effects`:

```
ce_bcg_mod2 <- conditional_effects(fit_bcg_mod2, effects = "ablat")
ce_ablat <- ce_bcg_mod2$ablat
points_ablat <- attr(ce_ablat, "points")
median_sigma <- median(as.matrix(fit_bcg_mod2, variable = "sigma"))
points_ablat$Weights <- 1 / (dat_bcg$sei^2 + median_sigma^2)
ggplot(ce_ablat, aes(effect1__, exp(estimate__))) +
  geom_smooth(aes(ymin = exp(lower__), ymax = exp(upper__)), stat = "identity") +
  geom_point(aes(y = exp(resp__), size = Weights), data = points_ablat) +
  labs(x = "Latitude", y = "Risk ratio") +
  scale_size_continuous(range = range(points_ablat$Weights) * 1)
```

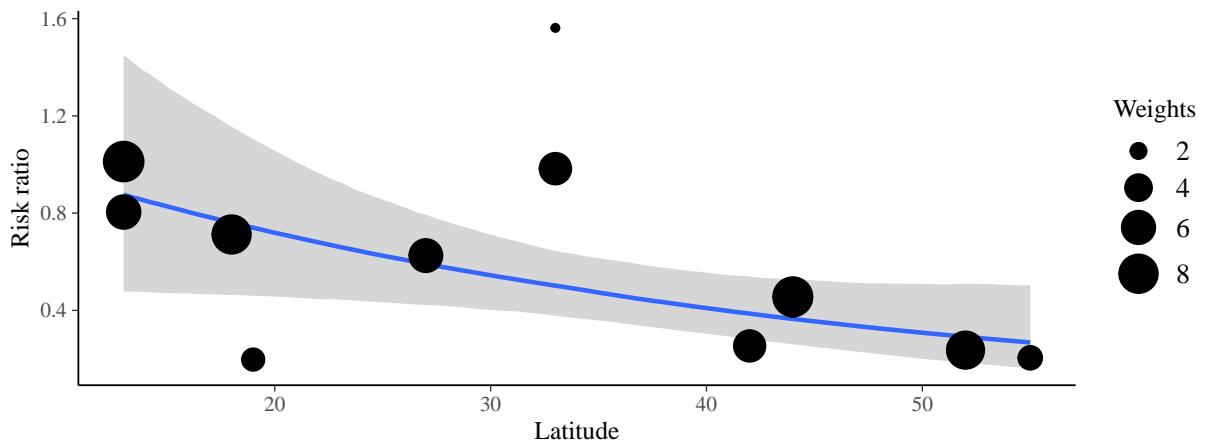


Figure 7.9: Conditional effects of model `fit_bcg_mod2` showing risk ratio predictions and point sizes proportional to the study weights.

Figure 7.9 illustrate how BCG vaccine efficacy increases (more negative risk ratios) as we move further away from the equator. We also see that the most highly weighted (i.e., largest) studies align well with the average prediction line.

Does including our moderators also improve predictive performance? We can easily check this via LOO-CV:

```
loo_bcg_mod1 <- loo(fit_bcg_mod1, moment_match = TRUE)
loo_bcg_mod2 <- loo(fit_bcg_mod2, moment_match = TRUE)
loo_compare(loo_bcg_re2, loo_bcg_mod1, loo_bcg_mod2)
```

	elpd_diff	se_diff
fit_bcg_mod2	0.0	0.0
fit_bcg_re2	-1.2	2.8
fit_bcg_mod1	-2.1	3.1

Indeed, the moderator model `fit_bcg_mod2` containing both latitude and year seems to perform slightly better than the other models. Interestingly, model `fit_bcg_mod1` containing only year as moderator performs worse than the random effect model without any moderators. This points to the fact that modeling year likely leads to overfitting. In light of that finding, it could make sense to fit yet another model including only latitude as moderator. I leave this to you as an exercise.

## 7.6 Meta-analysis on count data

One general principle I try to follow in my data analysis workflows is to model the data “as raw as possible”. That is, to aggregate the data as little as possible prior to the modeling stage. Sometimes, a bit of pre-modeling aggregation is required to ensure comparability of observations or computational feasibility. When looking back at our meta-analytic models above, we have clearly violated the “raw-data principle”. The rawest data available to us are the Tuberculosis positive and negative counts in both treatment and control group. Yet, we chose to aggregate these four counts into two statistics, the log risk ratio and its approximate standard error. From the perspective of standard meta-analysis, this step was necessary to fit into the framework. However, having the full power of brms at our hands, we can just choose to directly model the counts with an appropriate count data model. Since the counts are already comparable and the BCG dataset is small, neither the comparability nor the computational feasibility argument applies here.

For our count data model, we need to restructure the data. First, we compute the total number of people in both treatment and control group. Then, we bring the data into long format such that counts from both groups appear in the same variable:

```

dat_bcg_long <- dat_bcg %>%
  select(-c(yi, vi, sei)) %>%
  mutate(ttotal = tpos + tneg, ctotal = cpos + cneg) %>%
  pivot_longer(
    cols = c(tpos, cpos),
    names_to = c("group", "event"),
    names_pattern = "(.)(.+)",
    values_to = "count"
  ) %>%
  mutate(total = ifelse(group == "t", ttotal, ctotal)) %>%
  select(-c(ttotal, ctotal, tneg, cneg))

```

In the reformatted data, `count` contains the number of people infected with Tuberculosis in a given `group` (treatment or control), while `total` contains the total number of people in that group. Since we know that total number, a binomial model seems appropriate, which we can specify as follows:

```

fit_bcg_bin <- brm(
  count | trials(total) ~ group + (1 | study),
  data = dat_bcg_long,
  family = binomial("logit")
)

```

It is important that we include the `group` variable already here, because it allows the model to differentiate between treatment and control group, enabling us to learn about the efficacy of the BCG vaccine.

```
summary(fit_bcg_bin)
```

```

Family: binomial
Links: mu = logit
Formula: count | trials(total) ~ group + (1 | study)
Data: dat_bcg_long (Number of observations: 26)

Multilevel Hyperparameters:
~study (Number of levels: 13)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.61      0.37     1.06     2.55 1.00      714     1186

Regression Coefficients:
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS

```

Intercept	-4.10	0.44	-4.93	-3.17	1.01	604	724
group	-0.48	0.04	-0.56	-0.39	1.00	2342	2127

```
conditional_effects(fit_bcg_bin)
```

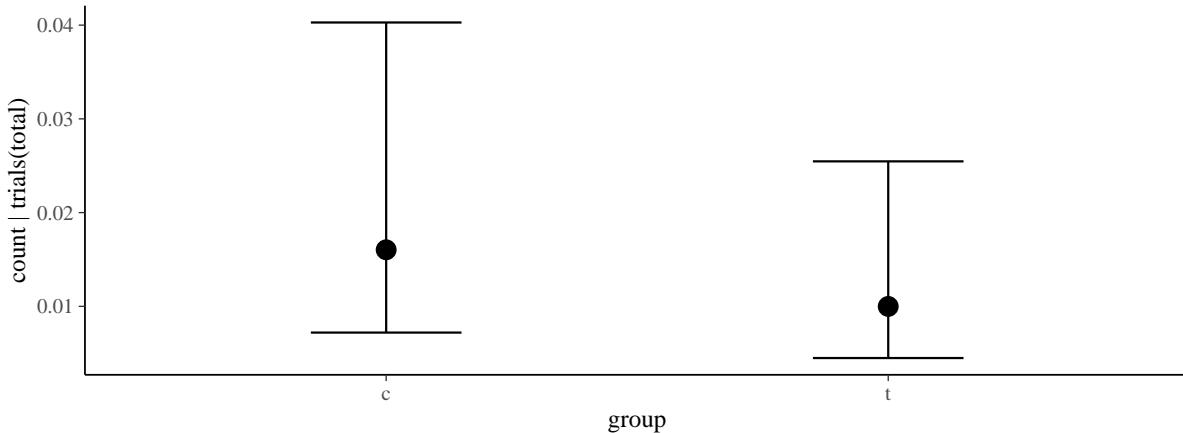


Figure 7.10: Conditional effects of model `fit_bcg_bin` predicting the probability of Tuberculosis infection for treatment and control group separately.

From the summary output we see that the there is a clear difference between groups. However, the conditional effects in Figure 7.10 do not show this as clearly. It appears as if the two groups' probabilities of Tuberculosis infection are not actually that different, at least when considering the posterior uncertainty. The reason is that the shown CIs are independent for the two groups. And both include the posterior uncertainty of the model intercept, which is quite substantial compared to the uncertainty of the group difference coefficient. If we just plotted the latter (e.g., via `hypothesis`) we would indeed clearly see the effect found numerically in the summary.

Using binomial models with a logit link for the meta-analysis of count data is closely related to the log odds ratio effect size measure. Accordingly, for large enough studies with not too extreme counts, the results of the binomial model and a normal random effects model of log odds ratios will yield similar results. Try it out as an exercise. You can obtain the log odds ratio in `metafor::escalc` via `measure = "OR"`. That said, I would always argue going to the binomial approach, because it operates much closer to the true data-generating process of the count data than any meta-analytic model using normal likelihoods and approximate standard errors.

### 7.6.1 Moderator Analysis

We can perform moderator analysis in the binomial model via interaction terms with `group`. For example, to understand the effect of latitude, we can run the following model:

```
fit_bcg_bin_mod1 <- brm(  
  count | trials(total) ~ group * scale(ablat) + (1 | study),  
  data = dat_bcg_long,  
  family = binomial("logit")  
)  
  
summary(fit_bcg_bin_mod1)  
  
Family: binomial  
Links: mu = logit  
Formula: count | trials(total) ~ group * scale(ablat) + (1 | study)  
Data: dat_bcg_long (Number of observations: 26)  
  
Multilevel Hyperparameters:  
~study (Number of levels: 13)  
Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
sd(Intercept) 1.38 0.34 0.88 2.20 1.00 849 1419  
  
Regression Coefficients:  
Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
Intercept -4.08 0.40 -4.84 -3.21 1.00 1219 1545  
group 0.72 0.05 -0.81 -0.62 1.00 2982 2401  
scaleablat 1.05 0.41 0.26 1.90 1.00 1163 1390  
group:scaleablat -0.47 0.04 -0.55 -0.39 1.00 2892 2475
```

Indeed, the interaction term `group:scaleablat` is clearly different from zero so there appears to be bigger group differences for higher latitudes. To make it easier to see differences between groups, we plot predictions with a 66% CI instead of the default 95% CI:

```
conditional_effects(fit_bcg_bin_mod1, effects = "ablat:group", prob = 0.95)  
conditional_effects(fit_bcg_bin_mod1, effects = "ablat:group", prob = 0.66)
```

Figure 7.11 reveals two distinct patterns: First, the BCG vaccine is indeed more effective for higher altitudes. Second, the overall rate of Tuberculosis is much smaller for lower altitudes in both treatment and control group. The latter pattern we couldn't see from the aggregated models fitted earlier, because it was lost in the effect size computation: Since the typical effect

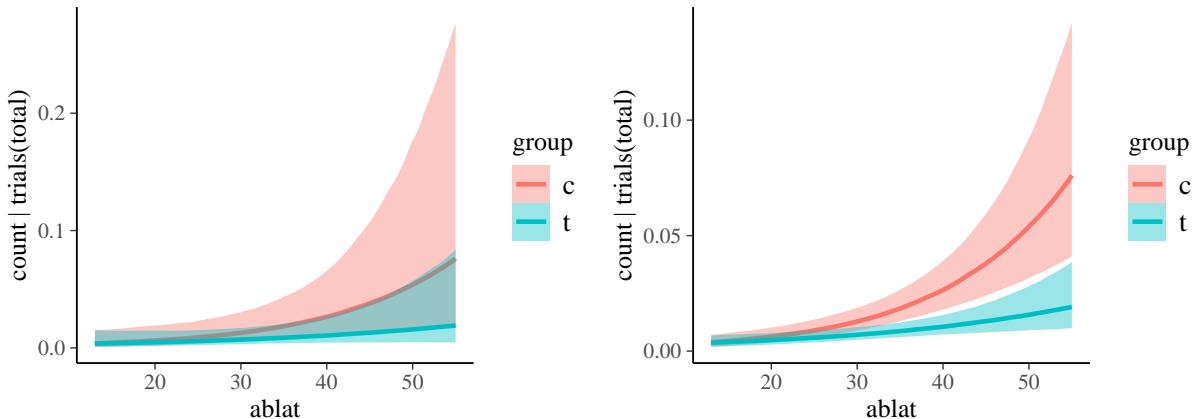


Figure 7.11: Conditional effects of model `fit_bcg_bin_mod` showing the interaction of group with latitude. Left: predictions plotted with a 95% CI. Right: Predictions plotted with a 66% CI.

sizes are defined as the contrast between two groups, patterns common to both groups are eliminated. This provides yet another argument to use lower-level meta-analytic models (here, for the raw counts) if the data allow.

### 7.6.2 What to add to nothing?

In meta-analysis of count data, zeros can cause a lot of trouble for many approaches. The reason is that common aggregated effect size estimates such as the log risk ratio or the log odds ratio are not defined in this case. For example, for zero events in the treatment group, the risk ratio becomes zero such that the log risk ratio becomes minus infinity. As a result, a surprisingly large methods literature has evolved to deal with zeros in the meta-analysis of count data (Beisemann, Doeblner, and Holling 2020). I want to take this opportunity to showcase the issue and illustrate how the binomial model is a natural solution.

First, let's add the results of an artificial study to our dataset, where the BCG vaccine worked so well that none of the 100 people in the treatment group got Tuberculosis:

```
new_study <- data.frame(
  trial = 14, author = "new", tpos = 0, tneg = 100, cpos = 10, cneg = 90,
  ablat = 50, alloc = "random", year = 1970, study = "new 1970"
)

dat_bcg_new <- dat_bcg %>% bind_rows(new_study)
```

When computing the log risk ratio with `metafor`, it seems to run just fine, giving us a small but finite effect size estimate for the new study:

```

dat_bcg_new <- metafor::escalc(
  measure="RR", ai=tpos, bi=tneg, ci=cpos, di=cneg,
  data=dat_bcg_new
)
dat_bcg_new$yi[14]

```

[1] -3.044522

Upon closer inspection of the documentation, we see that metafor actually adds a count of 0.5 to all four cells of a study for which at least one cell is zero. This may be a sensible approach if only few studies have zeros, but may otherwise lead to relevant bias. When we explicitly add nothing to the zero cells, we see that the resulting log risk ratio is indeed not defined and returned as NA:

```

dat_bcg_new <- escalc(measure="RR", ai=tpos, bi=tneg, ci=cpos, di=cneg,
                      data=dat_bcg_new, add = 0)

```

Warning: Some 'yi' and/or 'vi' values equal to +-Inf. Recoded to NAs.

```
dat_bcg_new$yi[14]
```

[1] NA

In contrast, when adding our new study to the data for the binomial model above, no changes to the analysis pipeline or additional assumptions have to be made: Zeros are naturally allowed by the binomial distribution, so pose no difficulty to the model specification. If too many zeros are present, the likelihood may start being only weakly informative, but that is nothing that couldn't be fixed with some more informative priors. As an exercise, run the binomial model again with the updated data including the new zero study. You may also add many such zero studies (in treatment or control group) and reduce the number of non-zero studies. Inspect how the model estimates change as you alter the dataset.

## 7.7 Meta analysis of dependent outcomes

It is not uncommon that the same study reports multiple effects sizes, for example, because it performed multiple sub-studies or estimated the effect of interest via several different outcome measures. Ideally, we would like to include all the available information into our meta-analysis,

but then we have to take into account the dependencies between effect sizes originating from the same study.

As a case study, we will use the meta-analysis of Williams and Bürkner (2017), where we analysed the effect of nasal oxytocin sprays on the symptoms of Schizophrenia. The underlying idea is simple: oxytocin has been shown to mediate social behavior in both animals and humans. So we hope to especially improve negative symptoms of Schizophrenia, such as social deficits, by giving oxytocin. And nasal sprays are an easy way to administer it. Whether the approach is also effective, is very different story as we will see soon. In the process of working on this study, we noticed severe issues in two previous meta-analysis on the same topic, which later on were retracted following our inquiries. In the process, we managed to make one senior professor very unhappy. It was the first time I saw first hand how senior academics try to use their position of power to intimidate younger researchers.

The statistical challenges of this dataset are related to the fact that we measure different kinds Schizophrenic symptoms: positive symptoms (gained through the disorder; e.g., hallucinations), negative symptoms (lost through the disorder; e.g., social deficits), and general psychopathology. Most of the 8 primary studies we found report all three or at least multiple symptom types.

The raw meta-analysis data including the original data preparation and analysis are available on OSF (<https://osf.io/6pyfu/>). For easier access, I have uploaded the required parts of the data also on my website:

```
oxytocin_meta <- readRDS(url("https://paulbuerkner.com/data/oxytocin_meta.rds"))
```

The loaded RDS file contains several variables. For now, we only need the main dataset, an overview of which can be found in Table 7.2. There, we clearly see that each study has multiple rows in the dataset, one row per symptom type.

```
dat_oxy <- oxytocin_meta$data

dat_oxy %>%
  select(-SMD_pre, -SMCR, -se_SMD_pre, -se_SMCR) %>%
  head() %>%
  knitr::kable(digits = 2)
```

Table 7.2: First six rows of the oxytocin data.

	study	type	obs	SMD_post	se_SMD_post
1	Cacotti-Saija et al., 2015	pos	1	0.63	0.28
2	Cacotti-Saija et al., 2015	neg	2	-0.22	0.28

Table 7.2: First six rows of the oxytocin data.

	study	type	obs	SMD_post	se_SMD_post
3	Davis et al., 2014	neg	3	0.33	0.43
5	Feifel et al., 2010	pos	5	-0.39	0.37
6	Feifel et al., 2010	neg	6	-0.49	0.37
7	Feifel et al., 2010	gen	7	-0.22	0.37

We could go ahead and just model the three different symptom types in separate meta-analytic models. And this would probably work well for this dataset too. However, for the purpose of learning how to deal with dependent outcomes in meta-analysis, we use other approaches that include all symptoms data in one model.

As effect size, we will use the standardized mean difference (SMD) between treatment and control groups post treatment. For group means  $\bar{y}_T$  and  $\bar{y}_C$  (say, of negative symptoms), standard deviations  $sd_T$  and  $sd_C$ , and sample sizes  $N_T$  and  $N_C$ , of treatment and control group, respectively, we can compute the SMD as

$$SMD = \frac{\bar{y}_T - \bar{y}_C}{\sqrt{\frac{(N_T-1)sd_T^2 + (N_C-1)sd_C^2}{N_T + N_C - 2}}}.$$

Admittedly, the formula in the denominator is a bit convoluted, but in essence we can just think of it as a weighted mean of the two groups' standard deviations. In our SMD post treatment, we do not account for potential pre-treatment differences between groups. Since the group assignments were fully randomized, this is unlikely to be a major issue. Yet, we could also define more advanced effect sizes that take into account both pre- and post-treatment data. In our case, the results were qualitatively the same independently of whether the pre-treatment data were considered (see Williams and Bürkner 2017 for the details).

### 7.7.1 Multilevel meta-analysis

As a first approach to handling dependent outcomes, we will use so-called “multilevel meta-analysis”, where we model varying intercepts on both the study level (variable `study`) and individual effect size level (variable `obs`). As a result, the corresponding standard deviation parameters will quantify the variation in true effect sizes between and within studies, respectively. Using the SMD as effect size, a simple multilevel meta-analysis on the oxytocin data looks as follows:

```

fit_oxy_smd_post1 <- brm(
  SMD_post | se(se_SMD_post) ~ 1 + (1 | study) + (1 | obs),
  data = dat_oxy,
  data2 = oxytocin_meta,
  prior = prior(normal(0, 0.5), class = Intercept) +
    prior(normal(0, 0.3), class = sd),
  sample_prior = TRUE
)

```

We use `sample_prior = TRUE` to conveniently visualize posterior against prior distributions in a bit.

```
summary(fit_oxy_smd_post1)
```

```

Family: gaussian
Links: mu = identity
Formula: SMD_post | se(se_SMD_post) ~ 1 + (1 | study) + (1 | obs)
Data: dat_oxy (Number of observations: 19)

Multilevel Hyperparameters:
~obs (Number of levels: 19)
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.22      0.13     0.01     0.48 1.00     1156     1098

~study (Number of levels: 8)
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.23      0.13     0.01     0.52 1.00      895      735

Regression Coefficients:
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     -0.14      0.14    -0.42     0.12 1.00     2608     2320

Further Distributional Parameters:
  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma       0.00      0.00     0.00     0.00   NA       NA       NA

```

When averaging across all three symptom types, oxytocin seems to have no noticeable effect. In terms of variation of effect sizes, it seems that both the varying intercepts' standard deviations within and between studies are of roughly similar size with posterior means of about 0.2 to 0.25. However, their uncertainty is substantial, with 95% CIs ranging from almost 0 to about 0.5. Such large uncertainty is not totally unsurprising, since the model has quite a few varying

coefficients: 8 for the `study` grouping factor and 19 for the `obs` grouping factor. Accordingly, it is plausible to assume strong prior sensitivity of the standard deviations' posteriors. Let's investigate this with two different approaches. First, via the `hypothesis` method, which will also show the prior since we set `sample_prior = TRUE`.

```
sd_parameters <- c("sd_obs_Intercept", "sd_study_Intercept")
hyp <- paste(sd_parameters, "= 0")
(hyp <- hypothesis(fit_oxy_smd_post1, hyp, class = NULL))
```

Hypothesis Tests for class :

	Hypothesis	Estimate	Est.Error	CI.Lower	CI.Upper	Evid.Ratio	Post.Prob	Star
1	(sd_obs_Intercept) = 0	0.22	0.13	0.01	0.48	0.80	0.45	*
2	(sd_study_Intercept) = 0	0.23	0.13	0.01	0.52	0.61	0.38	*

```
plot(hyp)
```

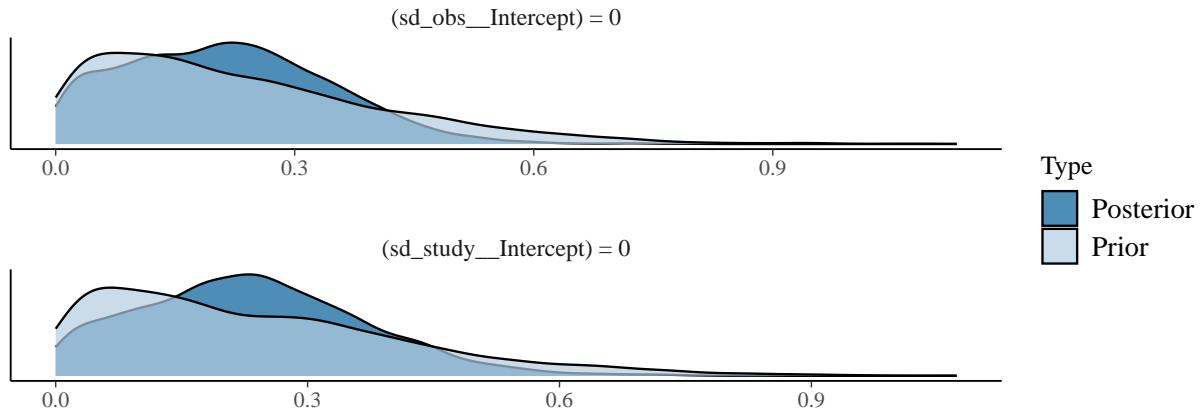


Figure 7.12: Prior and posterior densities of the varying intercepts' standard deviations as estimated by model `fit_oxy_smd_post1`.

As shown in Figure 7.12, the posteriors of the standard deviations are indeed very similar to the prior, indicating that there is only small amounts of information in the data to move the posterior away from the prior.

We can also investigate prior and likelihood (data) sensitivity via methods of the `priorsense` package (Kallioinen et al. 2023). Concretely, `priorsense` power-scales either the prior or the likelihood by a factor  $\alpha \neq 1$ , thus increasing ( $\alpha > 1$ ) or decreasing ( $\alpha < 1$ ) this model component's influence. To avoid refitting the model, importance sampling is applied, similar to what we did for approximation cross-validation in Chapter 2. More details on this approach can be found in Kallioinen et al. (2023). Here, we use the `powerscale_sensitivity` method

as it does not require us to explicitly specify  $\alpha$ , but rather automatically investigate the local change of the posterior around  $\alpha = 1$ .

```
priorsense::powerscale_sensitivity(  
  fit_oxy_smd_post1, variable = sd_parameters  
)  
  
Sensitivity based on cjs_dist  
Prior selection: all priors  
Likelihood selection: all data  
  
      variable prior likelihood           diagnosis  
sd_obs__Intercept 0.084      0.217 potential prior-data conflict  
sd_study__Intercept 0.102      0.081 potential prior-data conflict
```

The output say that there might be prior-data conflict, which in essence means that both the prior and the likelihood seem to have a relevant influence on the posterior for both parameters. This makes sense for our model: The data are in principle informative, so adding more data will change the posterior. However, the data are small enough so that the chosen priors still have relevant influence. I am not sure whether there is necessarily a “conflict” here, simply a low data regime such that the priors’ influence is noticeable.

For multilevel models, such as our meta-analytic model, scaling the likelihood implies increasing the number of observations per group, not the number of groups, which will likely lead to an underestimation of the likelihood informativeness. This also explains one interesting detail in the output: The much higher likelihood informativeness for parameter `sd_obs__Intercept` compared to `sd_study__Intercept`. Since each level in `obs` is only informed by a single observation, the corresponding varying intercepts are highly uncertain too, which propagates to the uncertainty of `sd_obs__Intercept`. If we now added more observations to each level of `obs`, this would imply a relevant decrease in the uncertainty of `sd_obs__Intercept` too. Of course, for our model, we will always have only a single observation per level of `obs` by definition, but priorsense cannot know that.

In terms of prior sensitivity, we see a little stronger sensitivity of `sd_study__Intercept` compared to `sd_obs__Intercept`. This makes intuitive sense given that `sd_obs__Intercept` is informed by 19 varying intercepts while `sd_study__Intercept` is only informed by 8 varying intercepts. It illustrates again that the smaller the number of groups in a multilevel term, the more we need to take care about the priors on the corresponding standard deviation parameters.

The priorsense package also comes with various plotting methods to showcase changes in the posterior as a result of changes in prior or likelihood. For example, we plot the posterior densities as a function of power-scaling prior or likelihood (Figure 7.13):

```
priorsense::powerscale_plot_dens(
  fit_oxy_smd_post1, variable = sd_parameters
)
```

### Power-scaling sensitivity

Posterior density estimates depending on amount of power-scaling ( $\alpha$ ).

Overlapping lines indicate low sensitivity.

Wider gaps between lines indicate greater sensitivity.

Estimates with high Pareto k (dashed lines) may be inaccurate.

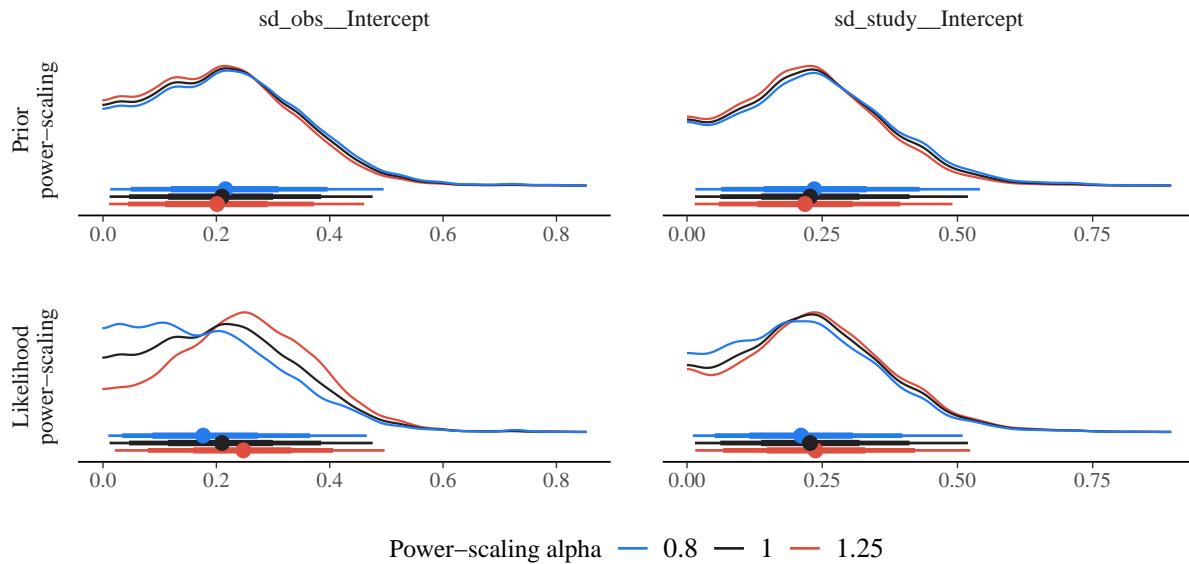


Figure 7.13: Output of the `powerscale_plot_dens` method applied for model `fit_oxy_smd_post1`.

If you want to play around with priorsense a bit more, here are some exercises: (1) Change the priors on the standard deviations (e.g., make them wider or narrower) and see how the prior sensitivity changes. (2) Fit models with either only `(1 | study)` or `(1 | obs)` and see how both prior and likelihood sensitivity changes. Also using `hypothesis` to directly plot posterior vs. prior can yield relevant insights.

### 7.7.2 Adding known correlations

The multilevel meta-analysis accounts for the within-study correlations of effect sizes via the `(1 | study)` term. However, it cannot account for *varying* within-study correlations. In other words, all pairs of effect sizes belonging to the same study will be given the same correlation, as implied by the between-study standard deviation. Since, in the oxytocin data, different effect sizes of the same study measure different symptom types, the equal-correlation assumption is

not ideal. Fortunately, we have correlational data available from Peralta and Cuesta (1994), who investigated the psychometric properties of their “Positive and Negative Syndrome Scale” (PANSS) questionnaire employed in most primary studies included in this meta-analysis. The sample consisted of 100 diagnosed schizophrenic patients, a relatively small sample but still sufficient to give us some rough estimates of the correlations. Peralta and Cuesta (1994) found a correlation of  $r = 0.09$  between positive and negative symptoms,  $r = 0.2$  between positive symptoms and general psychopathology, and  $r = 0.55$  between negative symptoms and general psychopathology. With the exception of the last correlation, these are relatively small, but it won’t hurt taking into account all three of them.

The joint likelihood incorporating known effect size correlations can be written as

$$y \sim \text{multinormal}(\eta, V), \quad (7.2)$$

where  $y$  is the vector of all effect sizes,  $\eta$  is the linear predictor terms (including varying intercepts and moderators), and  $V$  is a fixed covariance matrix.  $V$  is computed from the standard errors  $s$  and the fixed correlations  $r$ . Any diagonal element of  $V$  is just the squared standard error  $V_{ii} = s_i^2$ . For an off-diagonal element quantifying the covariance between two effect sizes of the same study, we have  $V_{ij} = s_i s_j r_{ij}$ , where  $r_{ij}$  depends on which two symptom types  $i$  and  $j$  represent (see above). For an off-diagonal element quantifying the covariance between two effect sizes of different studies,  $V_{ij} = 0$ . The oxytocin meta-analysis data already contains the precomputed  $V$ , so we don’t have to calculate it from scratch. For example,  $V$  belonging to the SMD effect sizes post-treatment, is stored in the `V_SMD_post` variable:

```
oxytocin_meta$V_SMD_post
```

Table 7.3: The first six rows and columns of the covariance matrix of SMD effect sizes post-treatment.

0.08	0.01	0.00	0.00	0.00	0.00
0.01	0.08	0.00	0.00	0.00	0.00
0.00	0.00	0.18	0.00	0.00	0.00
0.00	0.00	0.00	0.14	0.01	0.03
0.00	0.00	0.00	0.01	0.14	0.07
0.00	0.00	0.00	0.03	0.07	0.13

The first six rows and columns of the  $V$  matrix are displayed in Table 7.3. As we can also verify in the main dataset `dat_oxy`, the first two rows belong to the first study, followed by a single row for the second study, and three rows for the third study. Indeed, not all studies reported all three symptom types. The inclusion of  $V$  a brms model is done via the `fcor` term, which accepts covariance or correlation matrices of the outcome vector. For a normal likelihood, adding `fcor` implies the a joint likelihood of the form

$$y \sim \text{multinormal}(\eta, \sigma V). \quad (7.3)$$

So almost exactly what we have specified in Equation 7.2, with the exception of  $V$  being multiplied by  $\sigma$ . When using an `fcor` term, a multilevel meta-analysis with known within-study correlations is specified as follows:

```
bform_oxy_smd_post2 <- bf(
  SMD_post ~ 1 + (1 | study) + (1 | obs) + fcor(V_SMD_post),
  sigma = 1
)
```

As compared to the previous models, notice two important changes implied by adding `fcor`: First, we did not include an `se` addition term anymore, and in fact brms would complain if we tried. The reason is that `fcor` already contains information on the standard errors on the diagonal of the covariance matrix, so we shouldn't include it twice. Second, we have fixed  $\sigma$  to 1 rather than to 0. This is because, as shown in Equation 7.3,  $V$  is multiplied by  $\sigma$  rather than added. So, to eliminate  $\sigma$  while keeping  $V$  as is,  $\sigma = 1$  is the correct choice.

In order to actually fit the model, we need to somehow also pass `V_SMD_post`. This cannot easily be done via the `data` argument because `V_SMD_post` is a matrix rather than a vector. Instead, we have to use the `data2` argument, which accepts a named list of objects of arbitrary shape and size. If a variable cannot be found in `data`, brms will automatically search for it in `data2`. Together with the same weakly informative priors as before, the full model becomes:

```
fit_oxy_smd_post2 <- brm(
  formula = bform_oxy_smd_post2,
  data = dat_oxy,
  data2 = oxytocin_meta,
  prior = prior(normal(0, 0.5), class = Intercept) +
    prior(normal(0, 0.3), class = sd)
)
```

```
summary(fit_oxy_smd_post2)
```

```
Family: gaussian
Links: mu = identity
Formula: SMD_post ~ 1 + (1 | study) + (1 | obs) + fcor(V_SMD_post)
         sigma = 1
Data: dat_oxy (Number of observations: 19)
```

Multilevel Hyperparameters:

```

~obs (Number of levels: 19)
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.21      0.12     0.01     0.46 1.00     1162     1116

~study (Number of levels: 8)
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.22      0.14     0.01     0.51 1.00     1217     1141

Regression Coefficients:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept      -0.15      0.15    -0.44     0.14 1.00     2424     2066

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma        1.00      0.00     1.00     1.00   NA       NA       NA

```

For our data, the results have barely changed compared to the previous model without known within-study correlations. But of course, for other data and contexts, it may very well make a bigger difference.

In some sense, both `(1 | study)` and `fcor(V_SMD_post)` serve the same purpose, which is to express correlations between effect sizes of the same study. In case of `(1 | study)`, this correlation is estimated from the data but assumed constant across all effect sizes within the same study. In contrast, via `fcor(V_SMD_post)`, we can specify all correlations independently. This increases flexibility at the expense of requiring a lot more prior knowledge, since the correlation have to be user-specified rather than estimated. If we would really trust our correlations specified via `fcor`, we could probably leave out the `(1 | study)` term altogether. However, for our analysis, we have merely used some general correlation estimates from the literature, rather than correlations actually specific to each included primary study. So `fcor` alone may not be enough. And in any case, including both terms will likely make our model more conservative, that is, less likely to find an non-zero effect. Usually, when it comes to claiming scientific discovery, we would choose to rather be conservative, so just including both terms seems like a sensible approach. As an exercise, refit the model with only `fcor(V_SMD_post)` and compare the results to the models above.

### 7.7.3 Multivariate meta-analysis

In the previous models of the oxytocin data, we have not differentiated between the different symptom types (negative affect, positive affect, and general symptoms) apart from using different within-study correlations. It is unlikely that intranasal oxytocin will affect all of these symptom types equally, so we should account for that in our model. Since we are interested in the effect on each symptom type, rather than in contrasts between symptoms, we will use

`0 + type` in the formula. It removes the intercept and enforces cell mean coding of `type` (see also (**glm-centering-predictors?**) for details on factor coding). We use this syntax for both overall and varying coefficients, not only estimating an overall effect for each type but also corresponding study-specific effects. Moreover, we will not estimate the correlations among the varying coefficients of `type`, which we indicate via the `||` syntax in the multilevel term `(0 + type || study)` (see also Chapter 4). Our model is already highly parameterized and 8 studies are very much insufficient to get reasonable correlation estimates. Together, the model specification looks as follows:

```
bform_oxy_smd_post3 <- bf(
  SMD_post ~ 0 + type + (0 + type || study) + fcor(V_SMD_post),
  sigma = 1
)

fit_oxy_smd_post3 <- brm(
  formula = bform_oxy_smd_post3,
  data = dat_oxy,
  data2 = oxytocin_meta,
  prior = prior(normal(0, 0.5), class = b) +
    prior(normal(0, 0.3), class = sd)
)
```

```
summary(fit_oxy_smd_post3)
```

```
Family: gaussian
Links: mu = identity
Formula: SMD_post ~ 0 + type + (0 + type || study) + fcor(V_SMD_post)
         sigma = 1
Data: dat_oxy (Number of observations: 19)
```

Multilevel Hyperparameters:

```
~study (Number of levels: 8)
```

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(typeneg)	0.19	0.14	0.01	0.50	1.00	2122	2164
sd(typepos)	0.41	0.15	0.11	0.73	1.00	2372	1676
sd(typegen)	0.15	0.12	0.01	0.44	1.00	2887	2329

Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
typeneg	-0.03	0.15	-0.32	0.27	1.00	4402	3011
typepos	-0.24	0.20	-0.65	0.17	1.00	3183	2866
typegen	-0.20	0.17	-0.54	0.14	1.00	5054	2904

Further Distributional Parameters:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	1.00	0.00	1.00	1.00	NA	NA	NA

```
conditional_effects(fit_oxy_smd_post3)
```

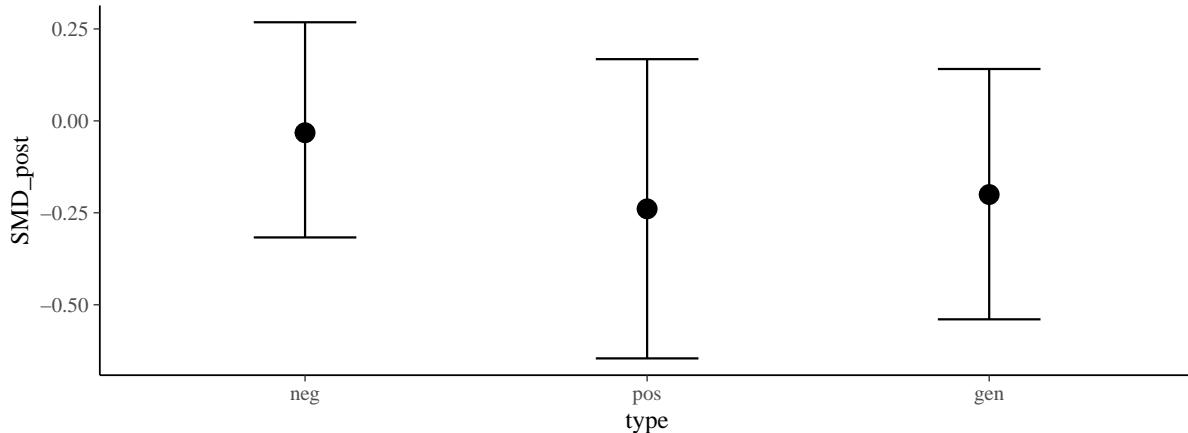


Figure 7.14: Conditional effects of model `fit_oxy_smd_post3`.

As we can clearly see from both the summary and predictive plot (Figure 7.14), there is absolutely no evidence for an effect of intranasal oxytocin on any of the three symptom types, which is also what we concluded in our original meta-analysis (Williams and Bürkner 2017). Apparently, even if oxytocin would in principle help with some schizophrenic symptoms, applying it via a nasal spray is very likely insufficient to induce these effects.

## 7.8 Raw data meta-analysis

In the past few years, scientists have become more open to sharing their raw data, following a general push towards open science coupled with the development of easy-to-use storage platforms such as the Open Science framework (OSF; <https://osf.io/>). As a result, the meta-analysis of raw data has become an important branch of research synthesis. Despite the improvements driven by open science, raw data meta-analysis still requires a huge amount of work, communication, and collaboration among many scientists, especially compared to standard meta-analysis. Accordingly, before engaging in such an endeavor, we have to know if it will be worth the effort.

To illustrate when and why raw data meta-analysis is sensible, I will use a simple example that was presented in the basic statistics course in my very first week of going to university. I have

traced back this example to Engel and Sedlmeier (2010) who cite a German article named “Metusalems machen Kasse: Ein langes Studium zahlt sich in barer Münze aus” (english: “Methuselah make money: a long course of study pays off in hard cash”) published by a German news paper. In the article, it was claimed that the longer people study, the more money they make. They effectively suggest that taking things slow and studying a few extra semesters will lead to higher salary down the line. I couldn’t find the original article, so I have to trust Engel and Sedlmeier (2010) on the correctness of the story.

To illustrate why the headline and its implications are likely highly misleading, let’s reproduce the example of Engel and Sedlmeier (2010), which we then turn into a meta-analytic problem. Suppose we have access to the data of three studies, each gathering data about ex-students’ starting salary and number of semesters studied. Crucially, the three studies differ in the kinds of students they have surveyed. Study 1 asked business students from an applied university, while Studies 2 and 3 investigated physics and chemistry students, respectively.

As a result, not only do the different kinds of students systematically vary in the number of studied semesters, but also in their starting salary: Chemists study the longest and have the highest starting salary on average, followed by physicists and lastly applied business people. To better illustrate the key points of this case study, we simulate hypothetical data as follows:

```
set.seed(43459)

n1 <- 200
dat_business <- data.frame(
  study = "study2", base_salary = 30,
  semester = rnorm(n1, mean = 7, sd = 1)
)

n2 <- 100
dat_physics <- data.frame(
  study = "study1", base_salary = 50,
  semester = rnorm(n2, mean = 10, sd = 1.5)
)

n3 <- 40
dat_chemistry <- data.frame(
  study = "study3", base_salary = 80,
  semester = rnorm(n3, mean = 16, sd = 2)
)

dat_meta_raw <- bind_rows(dat_business, dat_physics, dat_chemistry) %>%
  mutate(salary = base_salary - 1 * (semester - 10) + rnorm(n(), sd = 1))
```

We have 3 key variables in the data: The grouping factor `study`, the predictor `semester` and the outcome `salary`. Let's get a quick overview of the data via a scatter plot:

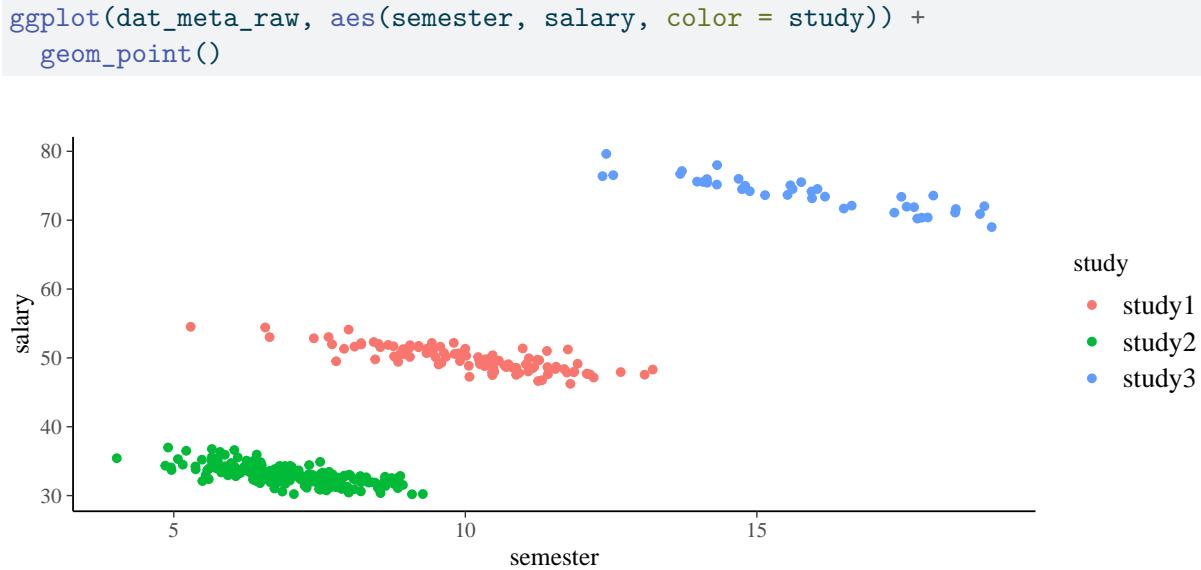


Figure 7.15: Scatter plot of the raw salary data.

Simply plotting the raw data already tells us a lot (see Figure 7.15): When focusing on the variation between studies, there is a strong positive relationship between number of semesters and starting salary. However, within studies, that is within the same group of students, the relationship is clearly negative. Accordingly, the causal claim that one should just “chill a bit more” during ones studies years to earn a higher salary is misleading. Instead, if one wants to make a lot of money right after studies, one should consider studying chemistry. This naturally takes longer because chemists are highly likely to also do a PhD on top of their Bachelor and Master studies, all of which together adds up to their very long study times. Yet, once in the group of chemistry students, those are faster with their studies will have again higher salary, essentially reversing the causal claim. The same within-group pattern can also be seen for the physics and applied business students.

Now suppose we don't have access to the raw data of the three studies but just know their means and standard deviations of the starting salary, together with the average number of semester and sample size  $N$ . Then, we would have to perform a standard meta-analysis with the effect size  $y_i$  simply being the mean salary and corresponding standard error  $s_i = \text{sd}_{\text{salary},i}/\sqrt{N}$ . Below, we create the corresponding aggregated data:

```
dat_meta <- dat_meta_raw %>%
  group_by(study) %>%
  summarise(
```

```

    se_salary = sd(salary) / sqrt(n())),
    salary = mean(salary),
    semester = mean(semester),
    weights = 1 / se_salary^2
)

```

When plotting the aggregated data, we see that only the positive between-study relationship remains visible (see Figure 7.16):

```

ggplot(dat_meta, aes(semester, salary)) +
  geom_point(aes(size = weights))

```

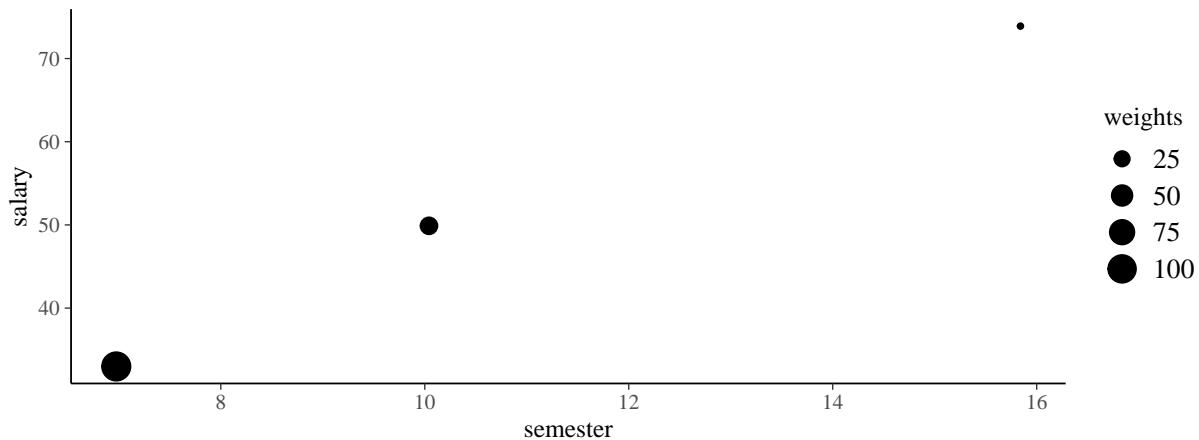


Figure 7.16: Scatter plot of the aggregated salary data on the study level.

Accordingly, when viewing the relationship of semester and salary on an aggregated level, the sign of the causal conclusion is reversed compared to the raw data. This aggregation bias is a special case of Simpson's paradox that probably most of you have heard about at some point during your statistics courses. And yet, many published moderator analysis on (aggregated) meta-analytic data are not properly acknowledging this issue.

That said, let us concretely illustrate when do we benefit from raw meta-analysis data analysis, and when we don't. We first consider the task of just estimating the average starting salary without including any moderators. We fit two meta-analytic models, one on the aggregated data and another on the raw data:

```

fit_meta1 <- brm(
  salary | se(se_salary, sigma = TRUE) ~ 1,
  data = dat_meta,

```

```

prior = prior(normal(50, 20), class = Intercept) +
  prior(normal(10, 10), class = sigma),
control = list(adapt_delta = 0.99)
)

fit_meta_raw1 <- brm(
  salary ~ 1 + (1 | study),
  data = dat_meta_raw,
  prior = prior(normal(50, 20), class = Intercept) +
    prior(normal(10, 10), class = sd),
  cores = 4
)

```

The posteriors of the overall intercepts (average salaries) of the two models are summarized in Table 7.4. Clearly the differences are minimal and can be fully attributed to Monte Carlo variation. Accordingly, if we are just interested in averages of the response, both aggregated and raw data meta-analysis will yield the same or highly similar results, if we compute the study-specific standard errors and corresponding weights correctly. In other words, for this case, we do not need to go through the pain of gathering all the data necessary for a raw data analysis.

Table 7.4: Posterior summaries of the overall intercept of models `fit_meta1` and `fit_meta_raw1`.

	Estimate	Est.Error	Q2.5	Q97.5
Average salary (aggregated)	52.05	9.47	32.32	70.90
Averagey salary (raw)	51.64	9.90	30.27	71.34

This drastically changes once we consider moderators. In our scenario, we are not really interested in the average salary, but rather if the number of semesters studied is predictive of the salary. Below, we specify two moderator models, one using the aggregated data and another using the raw data:

```

fit_meta2 <- brm(
  salary | se(se_salary, sigma = TRUE) ~ semester,
  data = dat_meta,
  prior = prior(normal(50, 20), class = Intercept) +
    prior(normal(10, 10), class = sigma),
  control = list(adapt_delta = 0.99)
)

```

```

fit_meta_raw2 <- brm(
  salary ~ semester + (1 | study),
  data = dat_meta_raw,
  prior = prior(normal(50, 20), class = Intercept) +
    prior(normal(10, 10), class = sd),
  cores = 4
)

```

Table 7.5 shows posterior summaries of the semester regression coefficients for both models. The sign of both the posterior mean and the posterior CIs is now fully reversed. While the aggregated model finds a positive relationship between semester and salary, the raw data model finds a negative relationship. Of course, the data was simulated to illustrate exactly this behavior. In real data, a complete sign reversal may be less likely, but smaller changes in magnitude are probably highly common. Accordingly, we should be careful to trust many published moderator results obtained via standard (aggregated) meta-analysis.

Table 7.5: Posterior summaries of the semester coefficient of models `fit_meta1` and `fit_meta_raw1`.

	Estimate	Est.Error	Q2.5	Q97.5
Average salary (aggregated)	4.61	1.59	1.02	7.98
Average salary (raw)	-1.02	0.04	-1.10	-0.94

Does this also cast doubt on the moderator analyses we performed earlier in this chapter? Fortunately not, since all moderators we analysed – publication year and latitude for the BCG data as well as symptom type for the oxytocin data – do not vary between participants of the same study. Accordingly, they are not at risk of being affected by aggregation bias.

Let's finish the discussion of raw data meta-analysis with one more model. When using raw data, we have access to many observations per study. As a result, modeling varying coefficients of moderators across studies becomes possible. For example, we can ask if semester has different coefficients for different included studies:

```

fit_meta_raw3 <- brm(
  salary ~ semester + (semester | study),
  data = dat_meta_raw,
  prior = prior(normal(50, 20), class = Intercept) +
    prior(normal(10, 10), class = sd, group = study, coef = Intercept) +
    prior(normal(0, 3), class = sd, group = study, coef = semester),
  control = list(adapt_delta = 0.95),
  cores = 4
)

```

Let's directly plot the salary predictions for the three studies (Figure 7.17):

```
conds <- make_conditions(fit_meta_raw3, "study")
ce <- conditional_effects(fit_meta_raw3, conditions = conds, re_formula = NULL)
plot(ce, points = TRUE)
```

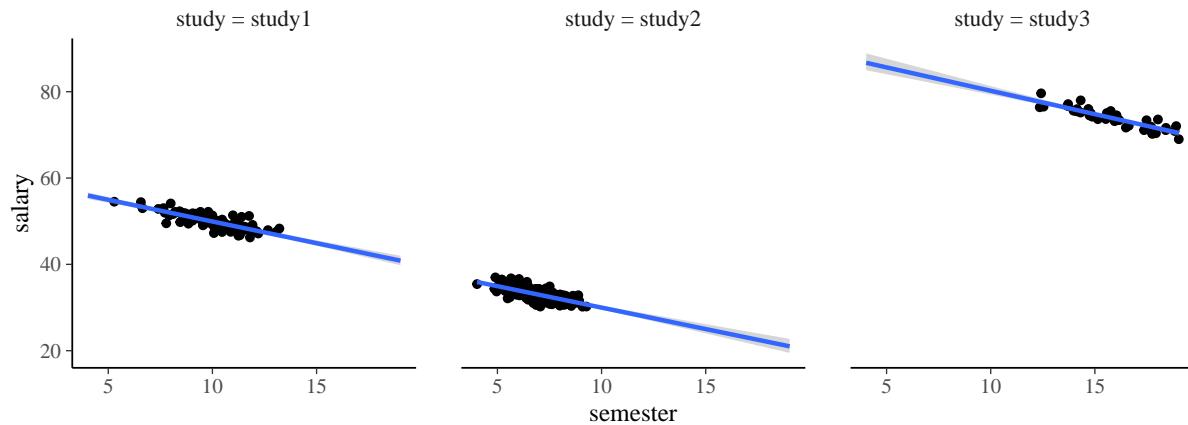


Figure 7.17: Conditional effects of model `fit_meta_raw3` shown separately for each study.

If this was a real dataset, it would be highly likely that the number of semester affected starting salary differently for business, physics, and chemistry students. However, in our data simulation, we assumed the same coefficient for all studies and this is also what we see reflected in Figure 7.17. Model comparison via LOO-CV confirms that there is no benefit of adding varying slopes:

```
loo_compare(loo(fit_meta_raw2), loo(fit_meta_raw3))
```

	elpd_diff	se_diff
<code>fit_meta_raw2</code>	0.0	0.0
<code>fit_meta_raw3</code>	-0.8	1.1

Yet, since we have many raw observations per study, there is also little harm in just keeping the varying slopes in the model regardless. If someone else uses your models as basis for their analysis, they will be more likely to consider varying slopes themselves, thereby potential uncovering relevant signal in the data.

## 7.9 Summary

As we have seen, meta-analytic models can vary strongly depending on the available data and desired types of analysis. Due to its flexibility, all of the common meta-analytic model classes fit very well into the brms framework. From basic fixed and random effects models to advanced models of dependent outcomes and raw data meta-analysis. Once you have understood generalized linear models, brms helps you to access the whole world of meta-analysis with just a few small additions to your model.

# 8 Data-driven non-linear models

## 8.1 Setup

```
library(dplyr)
library(ggplot2)
library(patchwork)
library(bayesplot)
library(brms)
```

## 8.2 Introduction

So far in this book, we have only worked with models that assumed a linear relationship between the predictors and the response one some scale; either directly on the scale of the response (linear models) or one some transformation on the response's mean (generalized linear models). In this chapter, we will now leave the safety of linearity and dive into the deep waters of non-linear modeling. On a higher level, the models discussed here all have the following form:

$$y \sim \text{normal}(b_0 + f(x), \sigma) \quad (8.1)$$

That is, we aim to predict the response  $y$  by a non-linear function  $f$  of the predictors  $x$  without having any idea about the function's form before seeing the data. In other words, the function  $f$  itself is unknown and needs to be estimated from the data. This is the main challenge we will tackle in the following.

Even when “estimating  $f$  from the data”, we still have to make at least some assumptions about  $f$ . That is, we still have to parameterize it somehow, such that we can estimate it within Stan, or any other probabilistic programming language for that matter. To ensure sufficient flexibility, the number of parameters controlling  $f$  will necessarily be high, which presents a consistent challenge for data-driven non-linear models.

In my own mental taxonomy of non-linear function approximators, there exist three main categories, each with their own benefits and drawbacks: Basis function approximations, Gaussian

processes, and neural networks. These categories are not exhaustive and even partially overlap. Yet, in my experience, most approaches applied in practice fall within one of these categories. As of writing this book, brms supports basis function approximations and Gaussian processes, but no neural networks, so it is the former two categories that I will focus on in the following.

To this end, we will keep all other model parts as simple as possible. In particular, we will only fit normal likelihood models. Of course, brms supports non-linear function estimation in combination with other likelihood families or model terms as well, so keeping things simple in these regards does not reflect any limitation in brms' actual flexibility.

### 8.3 Case Study: Housing Rents in Munich

Fahrmeir et al. (2013) use an example about the housing rents in Munich from 1999. The data contain information about roughly 3000 apartments including, among others, the absolute rent (`rent`), rent per square meter (`rentsqm`), size of the apartment (`area`), construction year (`yearc`), and the district in Munich where the apartment is located (`district`). The data can be found in the `gamlss.data` package (Stasinopoulos and Rigby 2016), an overview of which is provided in Table 8.1.

Table 8.1: First few rows of the `rent99` dataset.

rent	rentsqm	area	yearc	district
109.95	4.23	26	1918	916
243.28	8.69	28	1918	813
261.64	8.72	30	1918	611
106.41	3.55	30	1918	2025
133.38	4.45	30	1918	561
339.03	11.30	30	1918	541

For now, we will focus on predicting the rent per square meter by the construction year only. Figure 8.1 shows a scatter plot of the data. Admittedly, it is hard to see much of a pattern there because of so many overlapping points. To simplify matters a bit, we take a random subset of only 200 observations and scale the construction year to the unit interval of 0 to 1. The latter will be helpful when setting up our first non-linear models in a bit. A scatter plot of the sampled subset is shown in Figure 8.2.

```
# scale values to the unit interval
scale_min_max <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}
```

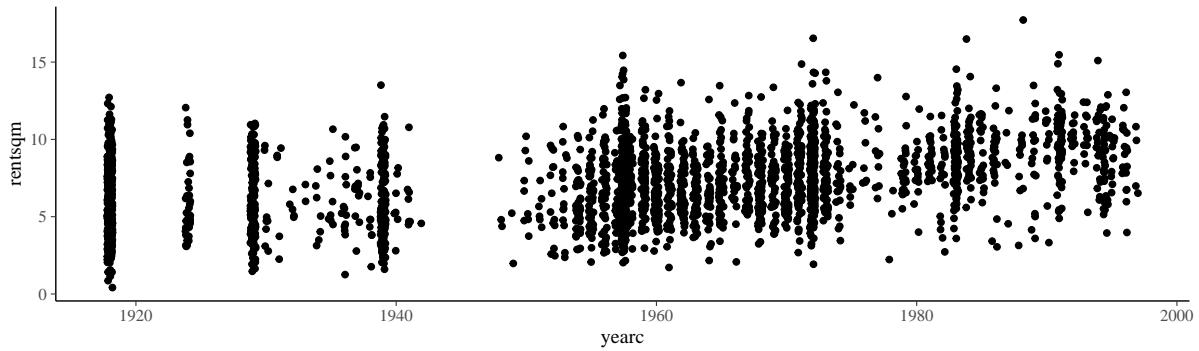


Figure 8.1: Scatter plot of construction year (`yearc`) vs. rent per square meter (`rentsqm`) from the `rent99` dataset.

```
set.seed(1234)
subrent99 <- rent99[sample(1:nrow(rent99), 200), ]
subrent99$ys <- scale_min_max(subrent99$yearc)
```

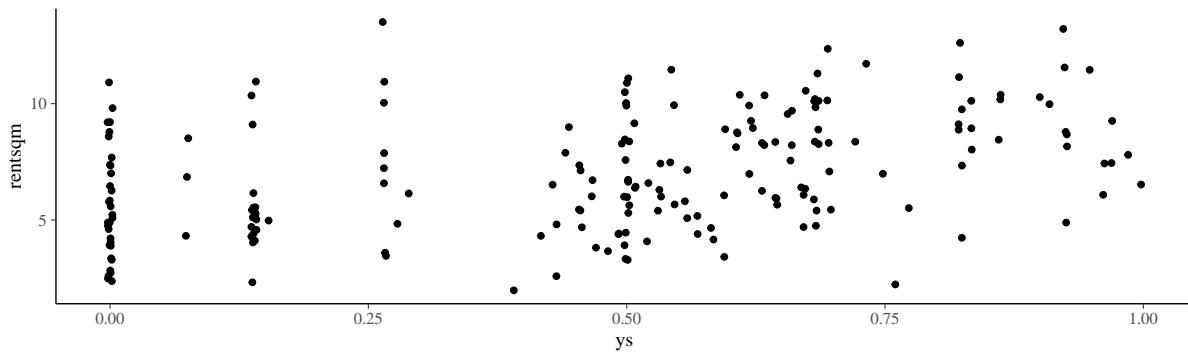


Figure 8.2: Scatter plot of scaled construction year (`ys`) vs. rent per square meter (`rentsqm`) from the a subset of the `rent99` dataset containing 200 observations.

As fundamental baseline for the upcoming non-linear models, we first fit a simple linear regression model to the subset of the renting data:

```
fit_rent1 <- brm(rentsqm ~ ys, data = subrent99)

conditional_effects(fit_rent1, "ys")
```

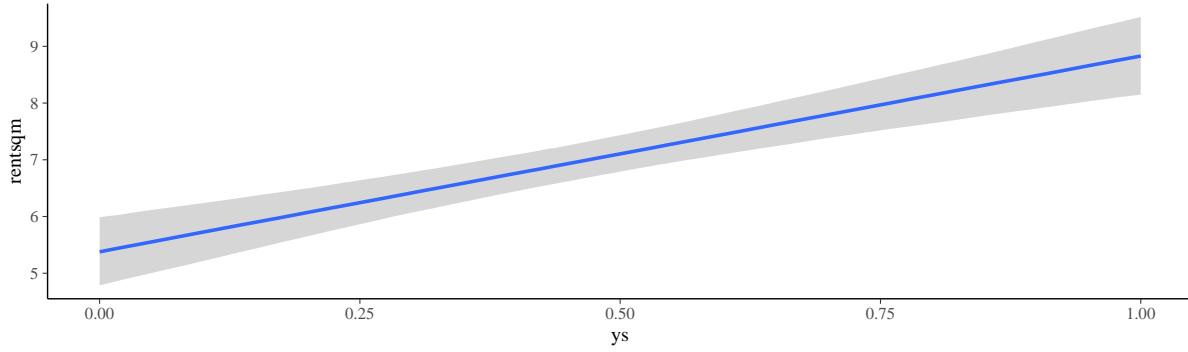


Figure 8.3: Conditional effects for model `fit_rent1`.

As visible in Figure 8.3, the rent per square meter appears to increase over time such that the rent for an apartment of comparable size is way higher for newer buildings. This relationship seems unlikely to be truly linear though and we will have to move to more complex models to uncover that.

## 8.4 Splines

Splines provide one way to solve the challenge of learning the unknown function  $f$  that relates the predictor(s)  $x$  to the response  $y$  as per Equation 8.1. To achieve this, splines are constructed from a set of simple basis functions, each defined only on a subspace of the predictor space, but smoothly connected at their intersections. Instead of explicating this largely informal definition mathematically, we will dive right into our first set of practical spline models, which will then clarify the main ideas in the process.

### 8.4.1 Linear Splines

In the linear model above we saw that buildings with later construction years have, on average, higher rent per square meter. That said, it remains highly questionable whether the relationship of construction year and rent is indeed linear. Unfortunately, we have no a priori idea how the supposed non-linearity is going to look like exactly. Instead, we will have to figure it out in a data-driven manner.

Our first spline model will be a linear spline with two components. Essentially, it's a change point model where we explicitly choose one point at which the linear slope is allowed to change. The change point itself is fixed but the two slopes will be estimated. When squinting at Figure 8.2 long enough, it looks as if the slope may change around the scaled year of 0.4, which roughly corresponds to the construction year of 1950. So this 0.4 will be our change point. Using the terminology of the spline literature, we are setting an (*internal*) *knot* at 0.4.

In order to conveniently express linear splines in a linear regression framework, we will parameterize our model not by one coefficient before and one coefficient after the knot, but by one coefficient before the knot and one coefficient for the *difference* between the slope after and before. This is not strictly out of necessity but more out of convenience. To achieve the difference formulation, we define a function that is 0 for all negative values and identity for all positive values. Readers familiar with deep learning will immediately recognize this as the *relu* activation function and I will just use this name here.

$$\text{relu}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

With *relu* defined, our linear spline model with one internal knot at 0.4 can be formulated as

$$\text{rentsqm} = b_0 + b_1 \text{ys} + b_2 \text{relu}(\text{ys} - 0.4) + \varepsilon$$

where  $b_1$  is the linear slope before the knot and  $b_2$  is the difference in the linear slopes between after and before the knot. As a result, when focusing only on the functional component of our predictor *ys*, we have defined

$$f(\text{ys}) = b_1 \text{ys} + b_2 \text{relu}(\text{ys} - 0.4).$$

In the spline literature, functions like  $\text{relu}(\text{ys} - 0.4)$  that are zero up until a given point and non-zero afterwards are called “hinge” functions, a common way of creating basis functions for splines. Writing this model in R formula syntax is easy. We first define the *relu* function

```
relu <- function(x) ifelse(x >= 0, x, 0)
```

and then set up a linear model with the two predictor terms *ys* and *relu(ys - 0.4)*:

```
fit_rent2 <- brm(rentsqm ~ ys + relu(ys - 0.4), data = subrent99)
```

```
summary(fit_rent2)
```

```
Family: gaussian
Links: mu = identity
Formula: rentsqm ~ ys + relu(ys - 0.4)
Data: subrent99 (Number of observations: 200)
```

Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
--	----------	-----------	----------	----------	------	----------	----------

Intercept	5.73	0.36	5.05	6.43	1.00	2954	2925
ys	1.08	1.34	-1.51	3.69	1.00	2348	2216
reluysM0.4	4.54	2.33	0.04	9.04	1.00	2319	2123

Further Distributional Parameters:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	2.31	0.12	2.09	2.55	1.00	2444	1982

According to the summary, the slope before the knot (coefficient `ys`) is not clearly different from zero. In contrast, the second coefficient (`reluysM0.4`), indicating the difference between the two linear slopes, is clearly positive with a 95% CI that only barely includes zero. The awkward coefficient name originates from taking the predictor term `relu(ys - 0.4)` and removing or replacing all syntactical symbols such as brackets or the minus sign. This admittedly makes the summary output harder to read, but ensures the coefficient name to always be a valid R object name. In turn, this makes a lot of post-processing easier. For example, we are interested in the slope *after* the internal knot, something we can easily compute with the `hypothesis` method when parameters have valid R object names:

```
hypothesis(fit_rent2, "ys + reluysM0.4 = 0")
```

Hypothesis Tests for class b:

Hypothesis	Estimate	Est.Error	CI.Lower	CI.Upper	Evid.Ratio	Post.Prob	Star
<code>(ys+reluysM0.4) = 0</code>	5.62	1.24	3.17	7.95	NA	NA	*

Indeed, the slope after the internal knot is strongly positive. In Figure 8.4 we see a visualization of the resulting predictions:

```
conditional_effects(fit_rent2, "ys")
```

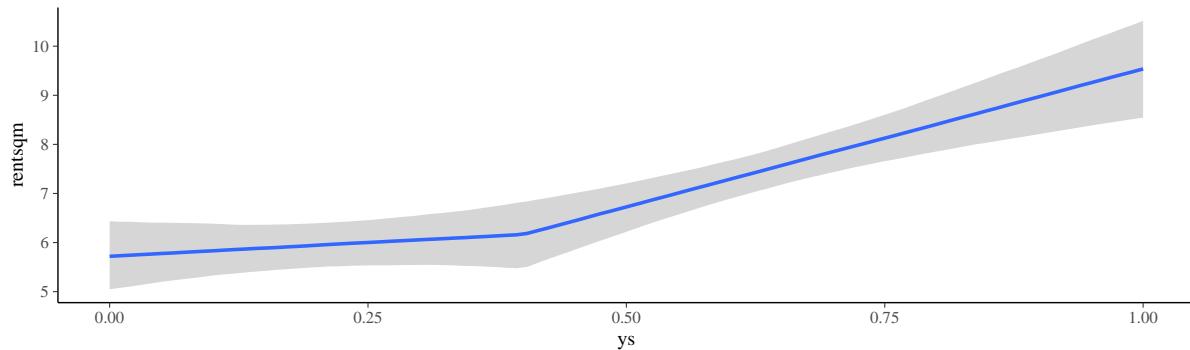


Figure 8.4: Conditional effects for model `fit_rent2`.

Clearly, the slope changes at the internal knot, but the functional form remains “connected” in the sense that we have a single continuous function across the whole predictor space. However, exactly at the knot, the function isn’t particularly smooth, because the linear slope “jumps” from one value to another. Not only is the function not looking very pretty, it is also unrealistic to assume that the relationship of construction year and rent per square meter changes exactly at (scaled) year 0.4. It is of course possible that a policy change right around 1950 qualitatively changed the development of housing rents, which would then motivate this exact change point. However, in our scenario, it is more likely that many factors together contributed to the change in housing rents over time. So a continuous change in slope appears as the more sensible assumption. We will come back to this issue in a bit.

Let’s first find out if using a linear spline with two components improves predictions compared to just a single linear slope:

```
loo_compare(loo(fit_rent1), loo(fit_rent2))
```

	elpd_diff	se_diff
fit_rent2	0.0	0.0
fit_rent1	-0.7	2.0

According to the LOO-CV results, there is a slight predictive advantage of the spline model, but this difference is not substantial both in absolute terms and in comparison with its standard error (for more details on LOO-CV comparison, see Section 2.7). Perhaps there is no non-linear relationship after all, or we placed the internal knot in a suboptimal way; or we had too few spline components? The latter is what we will focus on next.

To create a linear spline with more than two components, we add multiple terms of the form  $\text{relu}(\text{year} - k_i)$  to the model equation:

$$\text{rentsqm} = b_0 + b_1 \text{year} + b_2 \text{relu}(\text{year} - k_2) + \dots \quad (8.2)$$

$$+ b_3 \text{relu}(\text{year} - k_3) + b_4 \text{relu}(\text{year} - k_4) + \varepsilon \quad (8.3)$$

Here, we define a linear spline with 3 internal knots  $k_2, k_3, k_4$ , leading to 4 separate linear components. More generally, a linear spline with  $K$  components can be written as

$$f(x) = \sum_{m=1}^M b_m \text{relu}(x - k_m)$$

where  $k_1$  is chosen small enough such that the first term applies to all  $x$  values. For our current example, we choose  $k_2 = 0.2$ ,  $k_3 = 0.5$ , and  $k_4 = 0.8$  and run the model again:

```

fit_rent3 <- brm(
  rentsqm ~ ys + relu(ys - 0.2) + relu(ys - 0.5) + relu(ys - 0.8),
  data = subrent99
)

```

By adding two more spline components, now having four in total, the functional form is clearly more flexible as we see in Figure 8.5. Yet, the function still doesn't look particularly smooth at the internal knots, which is not something we can fix with linear splines no matter how many components we add. Accordingly, we will look into more advanced types of splines next.

```
conditional_effects(fit_rent3, "ys")
```

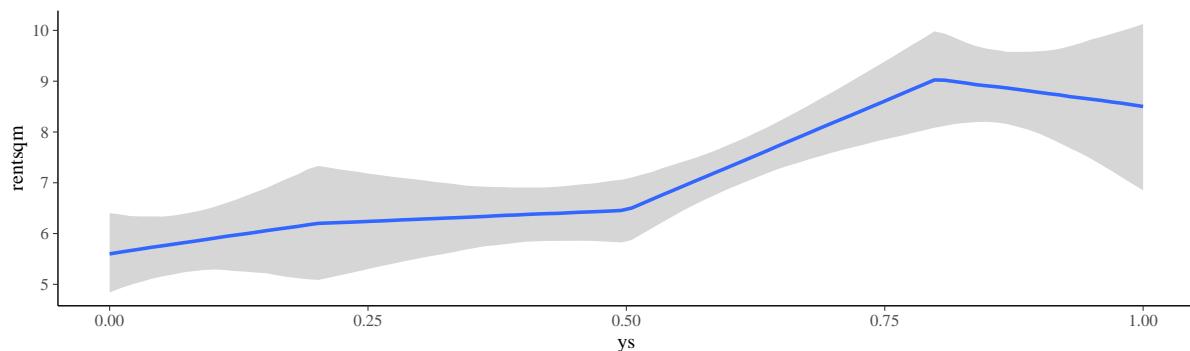


Figure 8.5: Conditional effects for model `fit_rent3`.

#### 8.4.2 Cubic Splines

To improve smoothness, we have to use higher order terms, for instance, cubic terms which leads to what we call cubic splines.

$$f(x) = \sum_{m=1}^M b_m \text{relu}(x - k_m)^3$$

Cubic splines ensure that the first and second derivatives of the function match at the knots, which makes the resulting function look much smoother. The cubic splines defined above via hinge functions are only a special case, since they use only polynomial terms of order 3 but no lower order terms. A complete polynomial of order 3 would have the form  $g(x) = a_0 + a_1x + a_2x^2 + a_3x^3$  for some fixed coefficients  $a_0, \dots, a_3$ , not just  $g(x) = x^3$ . That said, for building a general intuition about splines, focusing on the simple hinge function formulation is sufficient.

```

fit_rent4 <- brm(
  rentsqm ~ I(ys^3) + I(relu(ys - 0.2)^3) +
    I(relu(ys - 0.5)^3) + I(relu(ys - 0.8)^3),
  data = subrent99
)

```

To ensure that R formula syntax executes the  $\wedge 3$  operation correctly, we wrap them inside a call to `I()`. The `I` function does nothing but changing the object class to `AsIs` (read “as is”). Otherwise, the  $\wedge$  operator would be interpreted as a formula operator rather than as a numerical operation.

A direct interpretation of the regression coefficients becomes complicated for cubic splines since we have to interpret them on a scale of cubic (scaled) construction years. Instead, we will resort to only interpreting the graphical representation, rather than the parameter values, an approach that we will continue to use for all the spline models that follow.

```
conditional_effects(fit_rent4, "ys")
```

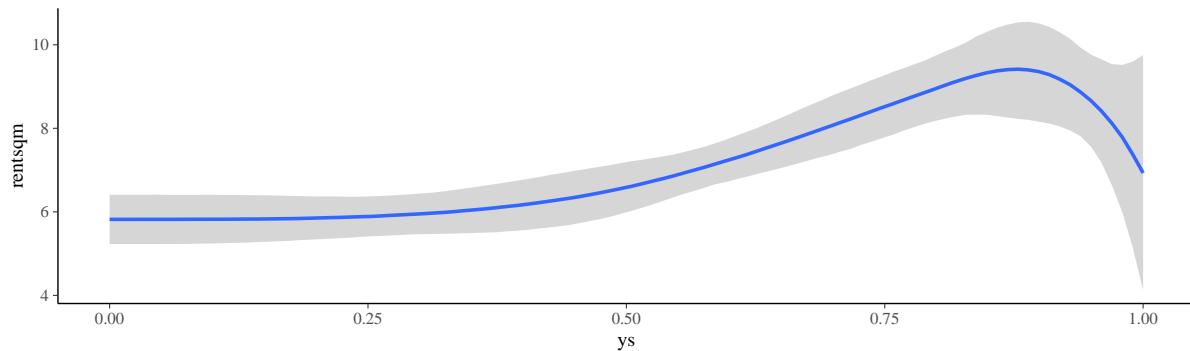


Figure 8.6: Conditional effects for model `fit_rent4`.

From the predictions in Figure 8.6 we can no longer visually identify where our knots have been placed as the function doesn’t look particularly rough at 0.2, 0.5, or 0.8. And there is a reason for it: By using cubic splines, we have ensured the first and second derivative of the implied function to match at the knots. And our eye cannot reliably detect non-smoothness in the 3rd or even higher derivatives. Thus, with cubic splines, we are creating smooth-looking functions that, mathematically, are not in fact perfectly smooth.

In term of substantive interpretation, our simple cubic spline predicts the relationship of construction year and rent per square meter to remain constant until roughly year 0.35, followed by a clear increase in slope up until year 0.85, before becoming sharply negative for the youngest buildings. Especially that latter observations is curious since it is not immediately obvious

why the trend towards increasing rents should have suddenly been reversed after the 1980s. Perhaps this is just a symptom of the model overfitting?

```
loo_compare(loo(fit_rent3), loo(fit_rent4))
```

	elpd_diff	se_diff
fit_rent4	0.0	0.0
fit_rent3	-1.3	1.3

Compared to the linear spline model, predictions seem to improve slightly but that doesn't rule out the possibility of overfitting. Generally, when working with data-driven non-linear models, we have to use many parameters to capture the true, unknown functional form. And in the attempt of achieving that, we may easily fit too complicated functions to the data, which subsequently overfit. In order to reduce this problem, we have to regularize the model. Or, to put it in Bayesian terms, we have to specify priors.

### 8.4.3 Smoothing Splines

The regularized splines implemented in brms are so called *smoothing splines* (S. N. Wood 2017). Viewed from a frequentist perspective, we can derive smoothing splines by finding a function  $f^*$  of the predictor variables  $x$  that minimizes the following objective:

$$f^* = \arg \min_f \|y - f(x)\|^2 + \lambda J(f),$$

where  $J(f)$  is a penalty term and  $\lambda$  is a positive hyperparameter to control the strength of the penalty. From a Bayesian perspective, we can see  $\|y - f(x)\|^2$  as the log-likelihood of a normal distribution and  $\lambda J(f)$  as a log-prior that regularizes the model. Regardless on the statistical viewpoint, the optimal solution  $f^*$  will of course strongly depend on the choice of the penalty  $J(f)$ . For example, we can set  $J(f) = \int f''(x)^2 dx$  as the squared second derivative of  $f$  integrated over the predictor space. This penalty induces a preference for smooth functions whose second derivative (curvature) is overall small, i.e., functions which are not too wiggly. When choosing this penalty, the optimal  $f^*$  is known as a thin plate spline (S. N. Wood 2003). Interestingly, a uni-dimensional thin plate spline is just a cubic spline. So cubic splines are not only convenient to create flexible, smooth-looking functions, they are also theoretically motivated.

For implementing smoothing splines in brms, the formulation of the "log-prior" term  $\lambda J(f)$  is inconvenient since it requires us to deal with  $J(f)$  directly. Fortunately, many smoothing splines (thin-plate or otherwise) and their priors, can be equivalently written in a different form that is easier to express in software that knows how to specify multilevel models, including brms and Stan:

$$f(x) = \sum_{p=1}^P \alpha_p \phi_p(x) + \sum_{l=1}^L \beta_l \psi_l(x) \quad (8.4)$$

In Equation 8.4,  $\phi_p$  and  $\psi_l$  represent simple analytic basis functions (our spline components from earlier), and  $\alpha_p$  and  $\beta_l$  are the corresponding spline coefficients. We write them down as two separate sets of functions because only the  $P$  coefficients  $\alpha_p$  are receiving an explicit multilevel prior of the form

$$\alpha_p \sim \text{normal}(0, s_\lambda).$$

The standard deviation hyperparameter  $s_\lambda$  takes over the role of  $\lambda$  in that it controls the complexity of the resulting spline  $f$ : When  $s_\lambda$  approaches zero, all  $\alpha_p$  will be close to zero too, effectively eliminating the  $\alpha_p \phi_p(x)$  terms from Equation 8.4. As such, the basis functions  $\alpha_p$  are said to be in the *penalized space* of the spline, also called *range space*. As in regular multilevel models, we estimate  $s_\lambda$  from the data, so we learn the appropriate flexibility of the spline from data, rather than fixing it a priori.

The remaining  $L$  basis functions  $\psi_l$  are in the non-penalized space of the spline, the so-called *null space*. Usually, the  $\psi_l$  are just linear functions of  $x$ . Of course, we could still choose to specify an explicit prior for these parameters as well, but they would not share any hyperparameters with the basis functions  $\phi_p$  in the penalized space.

In contrast to the complicated math required behind the scenes, specifying smoothing splines in brms is very simple. All we have to do is wrap our predictors in the `s()` function. Behind the scenes, the data are prepared by mgcv (S. N. Wood 2011), one of the most powerful packages for splines in R. For our example of predicting rent per square meter via (scaled) construction year, we can write the model as follows:

```
fit_rent5 <- brm(
  rentsqm ~ s(ys), data = subrent99,
  control = list(adapt_delta = 0.95)
)

summary(fit_rent5)

Family: gaussian
Links: mu = identity
Formula: rentsqm ~ s(ys)
Data: subrent99 (Number of observations: 200)

Smoothing Spline Hyperparameters:
```

```

    Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sds(sys_1)      1.26       1.05     0.06     4.04 1.00      1060      1519

```

**Regression Coefficients:**

```

    Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept      6.96       0.16     6.65     7.29 1.00      4204      2686
sys_1         2.20       2.93    -5.28     6.84 1.00      1594      1338

```

**Further Distributional Parameters:**

```

    Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma        2.32       0.12     2.10     2.57 1.00      4137      2858

```

While, for specifying the model, the mathematical details of splines are not needed, they help us making sense of the summary output. Under **Regression Coefficients**, we see a single null space coefficient named `sys_1`, which models the linear part of the spline. Indeed, when we correlate the corresponding predictor variable with the original predictor `ys` we get a value of 1. While this is a linear coefficient of `ys`, we should not interpret it in isolation due to the presence of the non-linear (penalized) spline components.

Under **Smoothing Spline Hyperparameters** we find the posterior summary of the spline hyperparameter  $s_\lambda$  for the penalized coefficients, here called `sds(sys_1)`. Again, we should be careful interpreting this parameter, but there is one thing we can still learn from its posterior: The lower bound of its 95% CI is close to zero, which gives us an initial indication that there may be little non-linearity in the spline. We also get the same impression when looking at the `plot` output of this parameter (see Figure 8.7):

```
plot(fit_rent5, variable = "sds_sys_1")
```

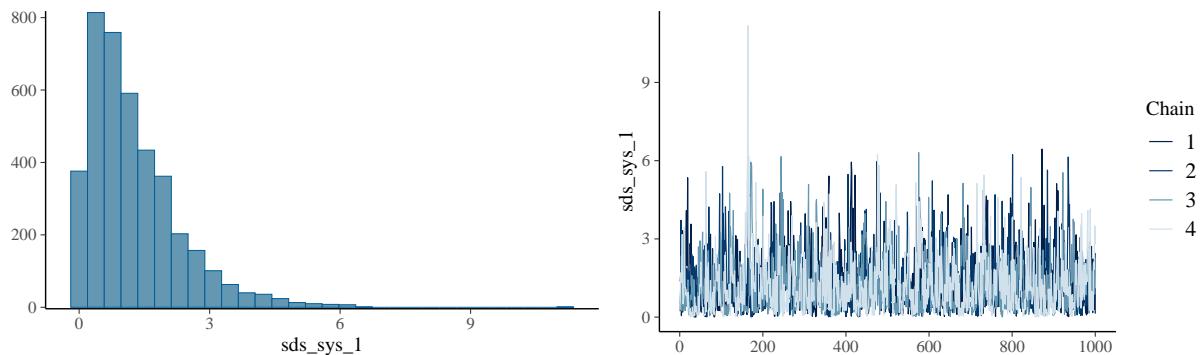


Figure 8.7: Marginal posterior histogram and trace plot of the spline standard deviation parameter `sds_sys_1` (labeled `sds(sys_1)` in the summary) for model `fit_rent5`.

This interpretation of the lower posterior should not be considered as a significance test because, as a standard deviation,  $s_\lambda$  is strictly positive in any case. So its CI can never overlap zero regardless of how little non-linearity there is in the fitted spline.

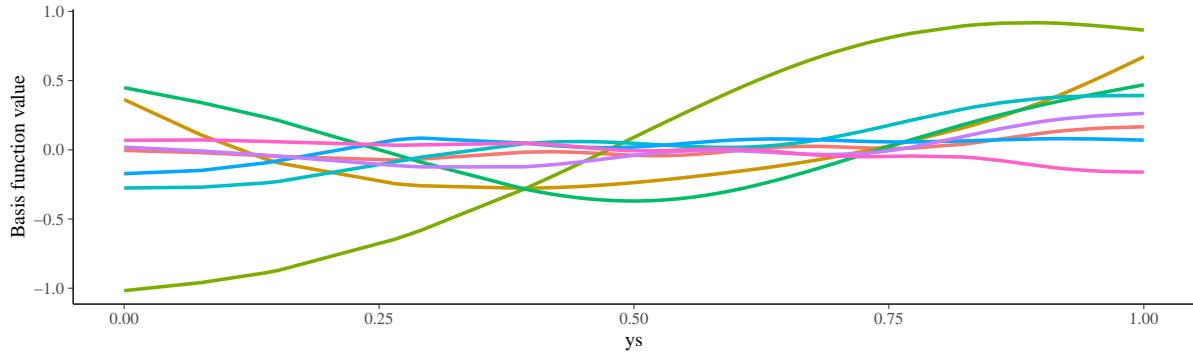


Figure 8.8: Spline basis functions created for the `s(ys)` term of `fit_rent5`.

The non-linear coefficients themselves are not shown in the summary because they are not interpretable in isolation. Only in combination with their corresponding basis functions (see Figure 8.8) do they create the overall spline curve, which we can visualize via the `conditional_smooths` method:

```
conditional_smooths(fit_rent5, "s(ys)")
```

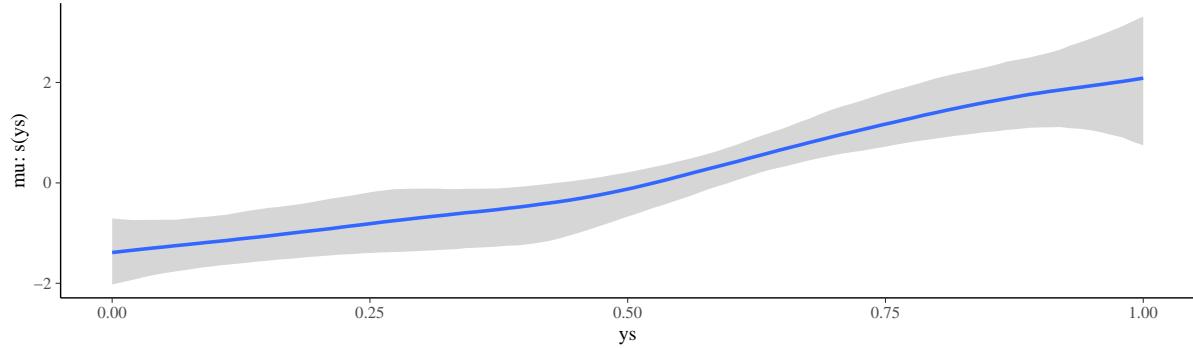


Figure 8.9: Conditional smooths for model `fit_rent5`.

Figure 8.9 shows only the estimated spline and should not be interpreted as a prediction of the (mean) response. For the latter, we of course have `conditional_effects`:

```
conditional_effects(fit_rent5, "ys")
```

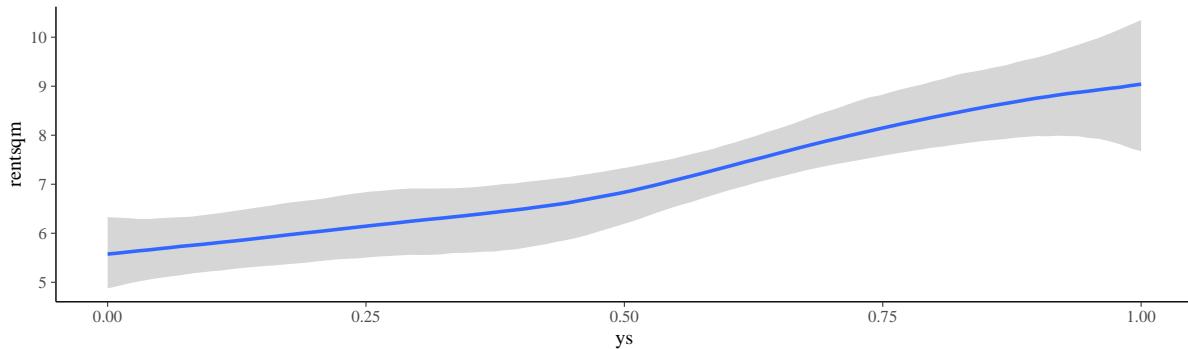


Figure 8.10: Conditional effects for model `fit_rent5`.

We see that Figure 8.10 with Figure 8.9 look highly similar, which makes sense since the predictions of the (mean) response are just the spline plus the intercept for this simple model. For more complex models (e.g., those in `?@sec-ldms`), both plots may differ strongly.

Interpreted with respect to our research question, we again see only a small increase in rent in the first few decades of the observation period followed by a linear increase during later periods. In contrast to the unpenalized cubic spline we fit previously, we don't see a decrease in rent at the end of the observation period anymore. Likely this drop was driven by only a few observations, which are now less influential for the thin-plate spline thanks to its regularization.

So far, we have only fitted our models to a small subset of the data (`subrent99`) consisting of 200 observations. From now on, we will use the full data (`rent99`) with 3082 observations. The same model as `fit_rent5` but with the full data is specified as:

```
fit_rent6 <- brm(
  rentsqm ~ s(yearc), data = rent99,
  cores = 4, control = list(adapt_delta = 0.95)
)
```

We won't analyse this model in detail here but rather directly focus on the next model that predicts rent per square meter by *both* construction year and apartment area. In this initial model considering both predictors, we will use two independent splines:

```
fit_rent7 <- brm(
  rentsqm ~ s(yearc) + s(area), data = rent99,
  cores = 4, control = list(adapt_delta = 0.95)
)
```

Plotting `conditional_effects` shows strong non-linear effects of both predictors:

```
conditional_effects(fit_rent7)
```

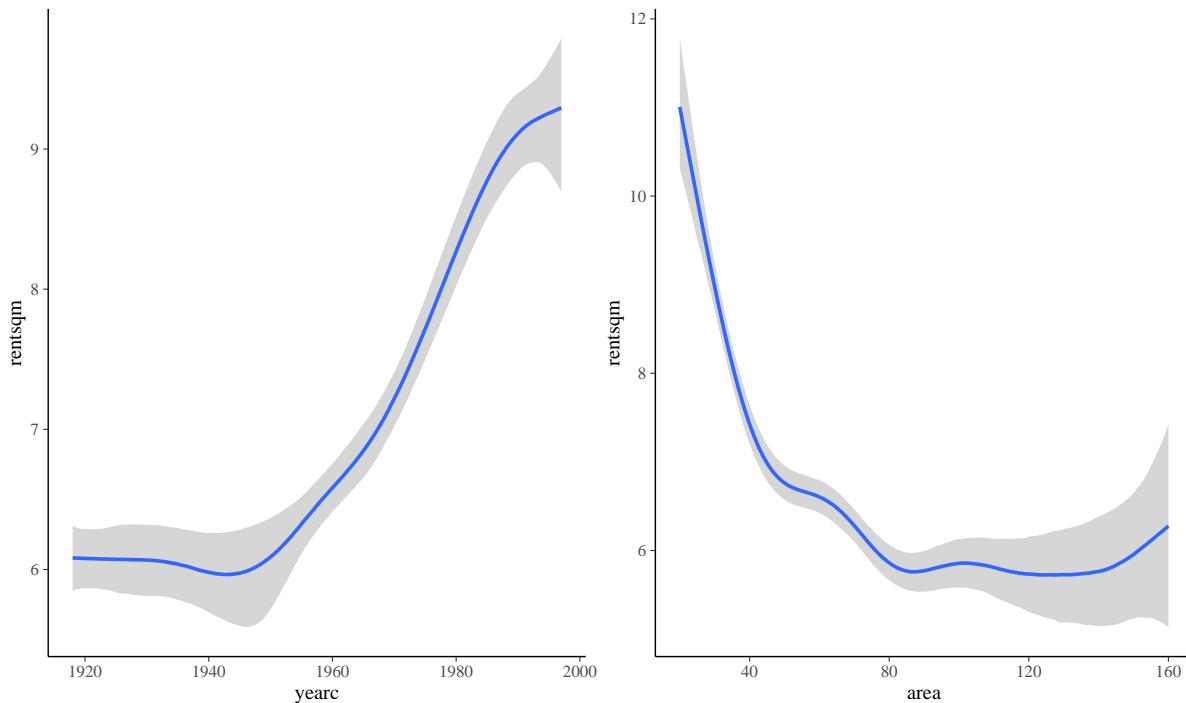


Figure 8.11: Conditional effects for model `fit_rent7`.

As we see from Figure 8.11, construction years later than 1950 indeed imply much higher rent per square meter, while before 1950, rents remained largely constant. We suspected this already based on the small data subset but in the full data, the pattern becomes much more clearly visible. In terms of the apartment area effect, we see the highest predicted rent per square meter for very small apartments, followed by a sharp decline for larger apartments. From roughly 80 square meter onwards, the predicted rent per square meter remains at the same level. The notable increase in predictive uncertainty for very large apartments is the result of only few apartments being this large in the given dataset.

Consistent with the finding that both construction year and apartment area have strong predictive power for the rent per square meter, model comparison via LOO-CV clearly indicates that adding a spline of `area` (vs. just one for `yearc`) improves predictive performance:

```
loo_rent6 <- loo(fit_rent6)
loo_rent7 <- loo(fit_rent7)
loo_compare(loo_rent6, loo_rent7)
```

```
elpd_diff se_diff
```

```

fit_rent7    0.0      0.0
fit_rent6 -262.7     24.9

```

Despite the flexibility of the two individual splines, they only ever model the effects of the two predictors *independently* of one another. Next, we will extend the model to include non-linear *interactions*.

#### 8.4.4 Tensor Product Splines

In linear models, interaction terms are formed by computing the product of both predictor variables (see also Section 1.3.4). This principle of forming interactions can be retained in splines by computing products of the predictors' basis functions. If we set up the spline by computing the products of all basis functions of one predictors with all basis functions of the other predictor, we arrive at *Tensor Product Splines* (Simon N. Wood, Scheipl, and Faraway 2013). In brms, we implement tensor product splines via the `t2` function of mgcv; for our case, `t2(yearc, area)`. The splines created via `t2` have a null space and multiple range spaces each with the individual hierarchical normal priors, essentially generalizing Equation 8.4. For brms users, it suffices to understand that we need to use `t2` when computing interaction splines of predictors that are defined on different scales, such as construction year and apartment area:

```

fit_rent8 <- brm(
  rentsqm ~ t2(yearc, area), data = rent99,
  cores = 4, control = list(adapt_delta = 0.95)
)

summary(fit_rent8)

```

```

Family: gaussian
Links: mu = identity
Formula: rentsqm ~ t2(yearc, area)
Data: rent99 (Number of observations: 3082)

```

```

Smoothing Spline Hyperparameters:
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sds(t2yeararea_1)    4.05      1.52      1.72      7.67 1.00      2712      2830
sds(t2yeararea_2)    6.59      2.32      3.31     12.15 1.00      2097      2678
sds(t2yeararea_3)    3.88      1.87      1.03      8.29 1.00      1682      1552

```

```

Regression Coefficients:
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept       7.80      0.10      7.60      8.00 1.00      2844      3241

```

t2yeararea_1	0.71	0.16	0.40	1.02	1.00	2548	2940
t2yeararea_2	-0.98	0.08	-1.14	-0.82	1.00	3670	3276
t2yeararea_3	-0.13	0.15	-0.42	0.17	1.00	2310	2695

#### Further Distributional Parameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sigma	2.03	0.03	1.98	2.09	1.00	8494	3082		

From the summary, under **Regression Coefficients**, we see three coefficients in the null space, belonging to the *linear components* of the two main effects and the interaction. Under **Smoothing Spline Hyperparameters**, we see the corresponding three standard deviations of the range spaces, belonging to the *non-linear* components of the main effects and interaction. Similar to the previous models, we won't focus too much on interpreting these parameters. Rather, we will directly look at the predictions:

```
effects <- c("yearc", "area", "yearc:area", "area:yearc")
conditional_effects(fit_rent8, effects)
```

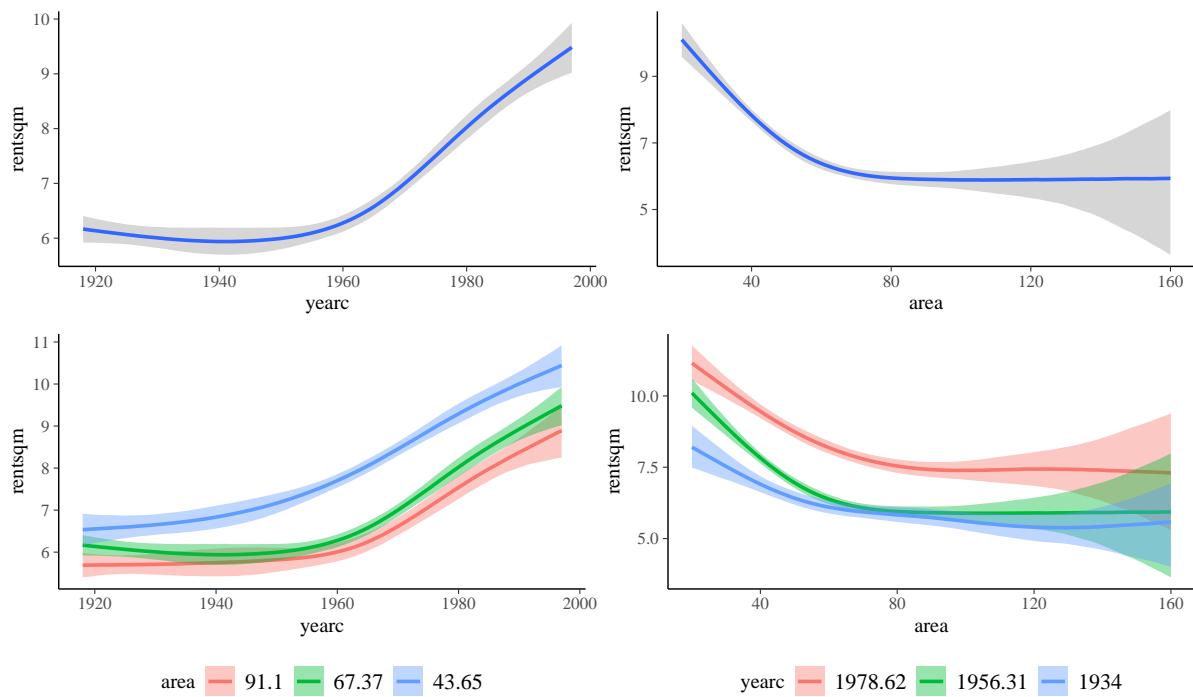


Figure 8.12: Standard conditional effects of model `fit_rent8`.

Using both "`yearc:area`" and "`area:yearc`" in the `effects` arguments implies that we see both predictors once on the x-axis (the first predictor) and once as color-coded groups (the

second predictor). As shown in Figure 8.12, the main effects plots (first row) look very similar to those of the previous model without an interaction. In the second row we see the interaction from the perspective of both predictor orders. If there was a substantial interaction, this would show up as *non-parallel* lines, but the lines in Figure 8.12 are largely parallel. However, we only see three lines at specific values of the second predictor, so there might still be an interaction visible for other predictor values. For this purpose, we can create predictive surface plots as follows:

```
ce8_surface <- conditional_effects(fit_rent8, "yearc:area", surface = TRUE)
```

We can plot the surface either by contour lines or as raster tiles:

```
plot(ce8_surface, stype = "contour")
plot(ce8_surface, stype = "raster")
```

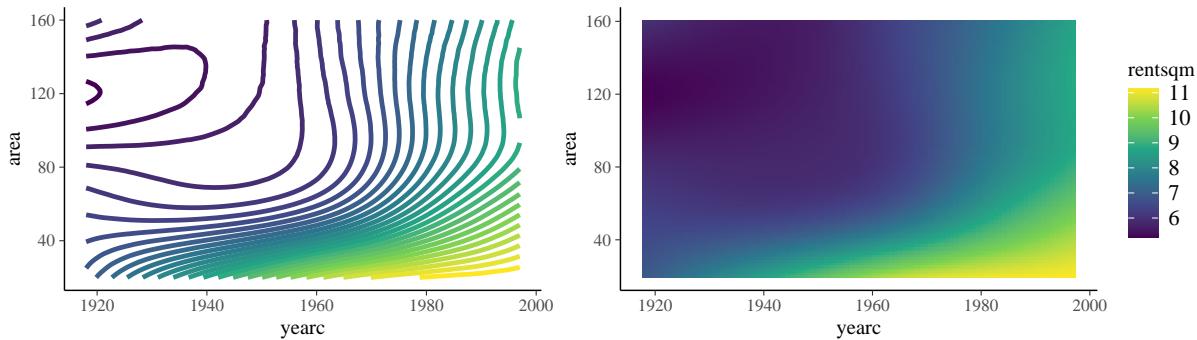


Figure 8.13: Conditional effects surface plots in both contour and raster format of model `fit_rent8`.

The created surface plots are shown in Figure 8.13. For the purpose of detecting interactions, I prefer contour plots as there interactions show up as non-parallel contour lines. Only for old, large buildings do we see non-parallel lines. The disadvantage of surface plots, contour or raster type, is that they don't show the uncertainty but only the medians of the posterior predictions. So is it hard to tell from Figure 8.13 how substantial the non-parallelism is in the top left corner of the predictive space.

Another aspect we need to be careful about when interpreting the predictions of splines is that they are only sensible close to observed data points. In other words, splines may generalize poorly to areas without any data close by, a property shared with other non-linear modeling approaches. To ensure that our predictive surface is not showing artifacts of splines predicting widely in areas without observed data, we can use the `too_far` argument of `conditional_effects`. For example, when setting `too_far = 0.1`, we only show the predictive surface for predictor combinations that are maximally 0.1 away from any observed data point in the Euclidian space after normalizing predictors to the unit interval.

```
ce8_surface_too_far <- conditional_effects(
  fit_rent8, "yearc:area", surface = TRUE, too_far = 0.1
)
```

```
plot(ce8_surface_too_far, stype = "contour")
plot(ce8_surface_too_far, stype = "raster")
```

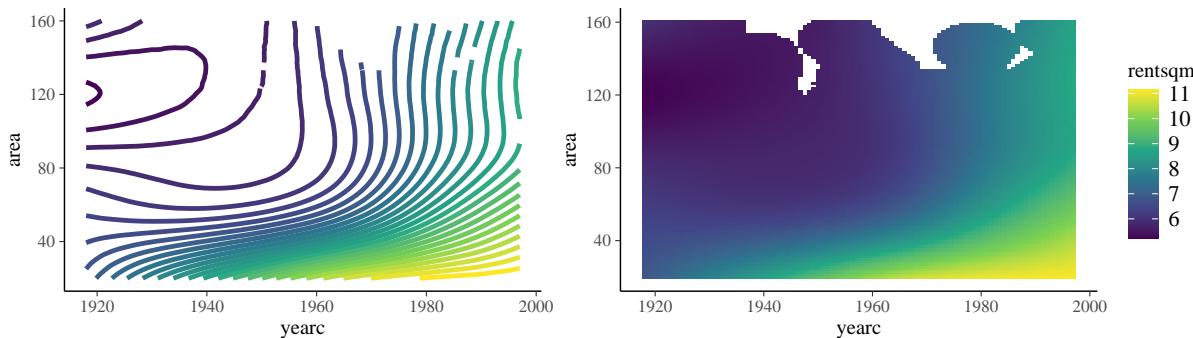


Figure 8.14: Conditional effects surface plots in both contour and raster format of model `fit_rent8`. Predictions are only shown for predictor combinations close to observed data.

As visible in Figure 8.14, we have observed data available for almost the entire two-dimensional surface of the predictors, so our spline predictions are likely sensible across the whole plotted space.

Next, we ask whether modeling the non-linear interaction between construction year and apartment area improves predictive performance:

```
loo_rent8 <- loo(fit_rent8)
loo_compare(loo_rent7, loo_rent8)
```

	elpd_diff	se_diff
<code>fit_rent7</code>	0.0	0.0
<code>fit_rent8</code>	-8.2	7.9

According to the LOO-CV results, model `fit_rent7` containing only the spline main effects has slightly better out-of-sample predictive performance than the interaction model `fit_rent8`. This can be interpreted as evidence against an interaction effect of the two predictors. However, the two models differ in more aspects than just the addition of an interaction. They also use different spline basis types and number of basis functions: Per the defaults of mgcv, `s` uses the thin-plate spline with 10 basis functions (for our data), and `t2` uses (another) cubic spline with

5 basis functions per predictor. The reduction to a default of 5 basis function per predictor is to reduce the total number of basis functions that is created by the tensor product. So the differences in LOO-CV might also be due to different default choices.

We adjust the interaction model to use the same basis as `s` does by default:

```
fit_rent9 <- brm(
  rentsqm ~ t2(yearc, area, k=10, bs="tp"), data = rent99,
  cores = 4, control = list(adapt_delta = 0.95)
)
```

```
loo_rent9 <- loo(fit_rent9)
loo_compare(loo_rent7, loo_rent9)
```

	elpd_diff	se_diff
fit_rent7	0.0	0.0
fit_rent9	-0.4	6.1

According to the LOO-CV results, the main effects and interaction models with otherwise the same spline bases have almost identical predictive performance. Apparently, even with this fairer comparison, there is still no evidence for an interaction effect of construction year and apartment area on the rent per square meter.

#### 8.4.5 Further spline features

Below, I want to briefly highlight some further important features of the splines supported in `brms` via the `mgcv` functions `s()` and `t2()`. The two most important arguments, namely the number of basis functions `k` and the spline type `bs` we have already seen in action above. Other relevant options include the ability to model *linear* interactions with other predictors via the `by` argument. For example, `s(yearc, by = area)` would model a non-linear main effect of `yearc` but a linear interaction with `area`. If the `by` variable was a factor we would fit a separate spline function for each factor level. We can also relate these individual spline functions in a hierarchical manner, in some sense generalizing multilevel models to splines. This can be achieved via the factor smooth basis, `bs = "fs"`, where the last included predictor is the grouping factor. For example `s(yearc, district, bs = "fs")` would fit a spline of `yearc` for each level of the `district` factor, with all individual splines sharing the same hyperparameters, just as in a regular multilevel model. More modeling options and mathematical details related to the splines available in `brms` through `mgcv` can be found in the tutorial paper of Pedersen et al. (2019).

## 8.5 Gaussian processes

An alternative to splines for data-driven non-linear modeling are Gaussian processes (GPs). A GP assumes the set of function values  $f(x_n)$ , for  $n = 1, \dots, N$ , to be multivariate normally distributed with a mean vector  $\mu \in \mathbb{R}^N$  and a covariance matrix  $\Sigma \in \mathbb{R}^{N \times N}$ :

$$(f(x_1), \dots, f(x_N)) \sim \text{multinormal}(\mu, \Sigma) \quad (8.5)$$

Within an additive modeling framework such as brms, we can specify everything that we would usually want to specify in the mean vector  $\mu$  outside of the GP in the form of other additive components. So, for our purposes, we can just focus on zero-mean GPs by setting  $\mu = 0$  without loosing modeling flexibility.

What makes a GP special is its definition of the covariance matrix  $\Sigma$ . It is constructed from a function  $K$ , which takes two input values  $x_i$  and  $x_j$  and outputs their covariance:

$$\Sigma_{ij} = K(x_i, x_j) \quad (8.6)$$

This way, for any set of inputs  $(x_1, \dots, x_N)$ , the behavior of a zero-mean GP is fully determined by the choice of  $K$ . We call  $K$  a *covariance kernel* or simply *kernel*. Since  $K$  controls the dependency for all possible sets of inputs, we can increase the number of input points indefinitely until we have covered *all* possible input values, e.g., all real values. As such, via  $K$ , we do not just define a distribution of the finite number of function values  $f(x_1), \dots, f(x_N)$  but over *all possible function values*, that is, over the function  $f$  itself. We write

$$f \sim \text{GP}(0, K)$$

to make explicit that GPs are, in a way, a distribution over functions. In this sense, we can sample entire functions from a GP in conceptually the same way as we sample a single value from a univariate normal distribution or a finite set of values from a multivariate distribution. If this is the first time you are coming into contact with GPs, don't worry if the idea of a "distribution over functions" does not make sense to you. It took me some time and explanations from different perspectives to grasp the main concepts behind GPs. For our purposes, it is sufficient to understand that GPs can be used as an alternative to splines to model data-driven non-linear functions. Whereas the properties of a spline are controlled by the choice of basis functions, the properties of a GP are controlled by the choice of covariance kernel and its hyperparameters.

A common choice, and also the default in brms, is the squared exponential (SE) kernel

$$K_{\text{SE}}(x, x') = \alpha^2 \exp\left(-\frac{1}{2} \frac{(x - x')^2}{\rho^2}\right)$$

with hyperparameters  $\rho$  and  $\alpha$ . The parameter  $\rho$  is known as *lengthscale* and determines how far the dependency induced through proximity in the input space reaches. For small  $\rho$  (relative to the scale of  $x$ ), two function values  $f(x)$  and  $f(x')$  will only be strongly correlated if their inputs  $x$  and  $x'$  are close together, leading to highly flexible, wiggly functions. For larger  $\rho$ , also function values  $x$  and  $x'$  that are further apart will be more strongly correlated, leading to less wiggly functions. Kernel-implied covariances are exemplified in Figure 8.15 with corresponding illustrations of sampled functions shown in Figure 8.16. The parameter  $\alpha$  is called the *marginal standard deviation* and controls the overall variation of the GP depending on the response's scale. The squared exponential kernel implies functions that are infinitely smooth, i.e., that have infinitely many derivatives. This stands in contrast to, for example, cubic splines for which only the first two derivatives exist (see Section 8.4.2).

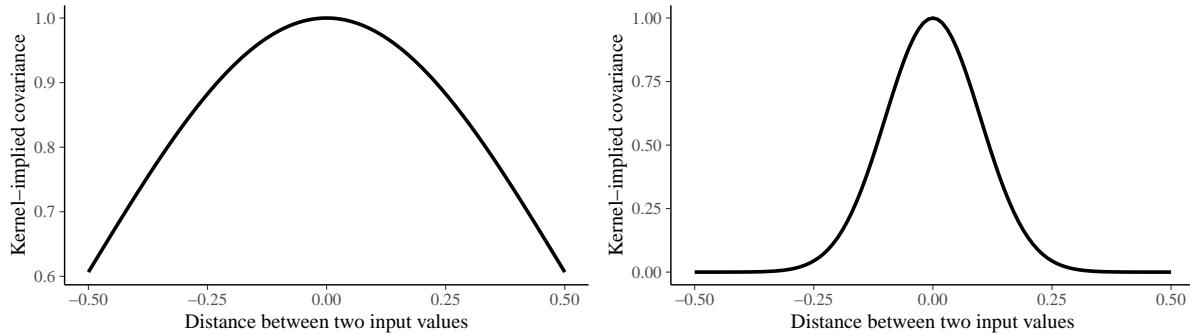


Figure 8.15: Covariances implied by the squared exponential kernel as a function of the distance between two input values. Left: The larger lengthscale  $\rho = 0.5$  implies high covariance across longer distances. Right: The smaller lengthscale  $\rho = 0.1$  implies high covariances only for short distances. The marginal standard deviation was set to  $\alpha = 1$  such that the kernel covariance is in fact a correlation.

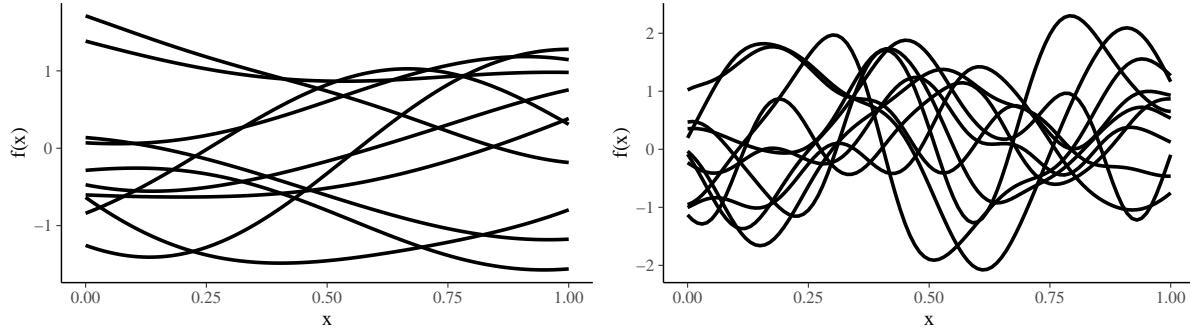


Figure 8.16: Functions sampled from a Gaussian process with squared exponential kernel for  $\alpha = 1$ . Left: The larger lengthscale  $\rho = 0.5$  implies unflexible, non-wiggly functions. Right: The smaller lengthscale  $\rho = 0.1$  implies more flexible, wiggly functions.

In order to sample from the function values implied by GPs we could directly use Equation 8.5. However, in the context of the MCMC algorithms used in Stan, this will cause many convergence issues, especially a high number of divergent transitions. For this reason, we apply a mathematically equivalent but sampling-wise more stable approach by computing:

$$(f(x_1), \dots, f(x_N)) = A z, \quad z = (z_1, \dots, z_N) \sim \text{normal}(0, 1) \quad (8.7)$$

Here,  $A$  is the *Cholesky factor* (matrix square root) of  $\Sigma$  such that  $\Sigma = A^T A$  and  $z$  is a vector of standard normal variables. So, instead of sampling the function values  $f(x_n)$  directly from a multivariate normal distribution (as per Equation 8.5), we first sample standard normal variables  $z_n$  and subsequently transform them into  $f(x_n)$  with the help of (the Cholesky factor of)  $\Sigma$ . We call Equation 8.7 the *non-centered* parameterization, while we call Equation 8.5 the *centered* parameterization of GPs. We use non-centered parameterizations also in other places in brms, for example, in multilevel models (see `?@sec-sec-mlms` and `?@sec-gmlms`).

In brms, we can fit GPs by wrapping the predictor variables inside `gp()`. We will first fit a GP model to the data subset `subrent99` of just  $N = 200$  observations that we started this chapter with. For now, we will just use the default priors of brms, to be discussed in more detail in Section 8.5.1.

```
fit_rent_gp1 <- brm(
  rentsqm ~ gp(yearc), data = subrent99,
  control = list(adapt_delta = 0.95), cores = 4
)
```

```
summary(fit_rent_gp1)
```

```
Family: gaussian
Links: mu = identity
Formula: rentsqm ~ gp(yearc)
Data: subrent99 (Number of observations: 200)
```

#### Gaussian Process Hyperparameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sdgp(gpyearc)	1.97	1.21	0.77	5.90	1.00	446	188
lscale(gpyearc)	0.33	0.37	0.03	1.51	1.01	276	200

#### Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	7.09	1.39	4.54	11.19	1.00	699	211

#### Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	2.30	0.12	2.08	2.54	1.00	3801	2896

Under Gaussian Process Hyperparameters we see the posterior summaries of both the lengthscale  $\rho$ , here named `lscale(gpyearc)`, and the marginal standard deviation  $\alpha$ , named `sdgp(gpyearc)`. By default, the inputs to the GP function are scaled to the unit interval for numerical stability. Accordingly, we should interpret the lengthscale with respect to the unit interval rather than the original predictor scale. As we can see from the summary, the posterior mean of `lscale(gpyearc)` is roughly 0.25, which means that for two input points being roughly 0.5 units away from each other (twice the lengthscale parameter) no relevant dependency between their function values exists anymore. Since 0.5 is half the unit interval, the lengthscale estimate we have obtained for our data can be considered quite large, indicating pretty unflexible, non-wiggly functions. At the same time, the 95% CI of the lengthscale spans almost the entire unit interval suggesting high epistemic uncertainty in the parameter (see Figure 8.17):

```
plot(fit_rent_gp1, variable = "lscale_gpyearc")
```

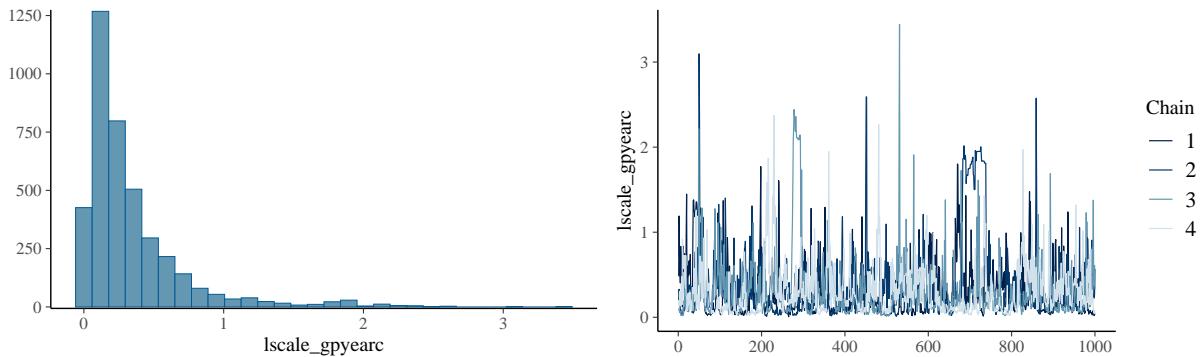


Figure 8.17: Posterior histogram and trace plot of the lengthscale parameter for model `fit_rent_gp1`.

Accordingly, based on observed data subset, the true underlying function may as well be very wiggly or not wiggly at all. Let's check what predictions the GP produces (see Figure 8.18):

```
plot(conditional_effects(fit_rent_gp1), rug = TRUE)
```

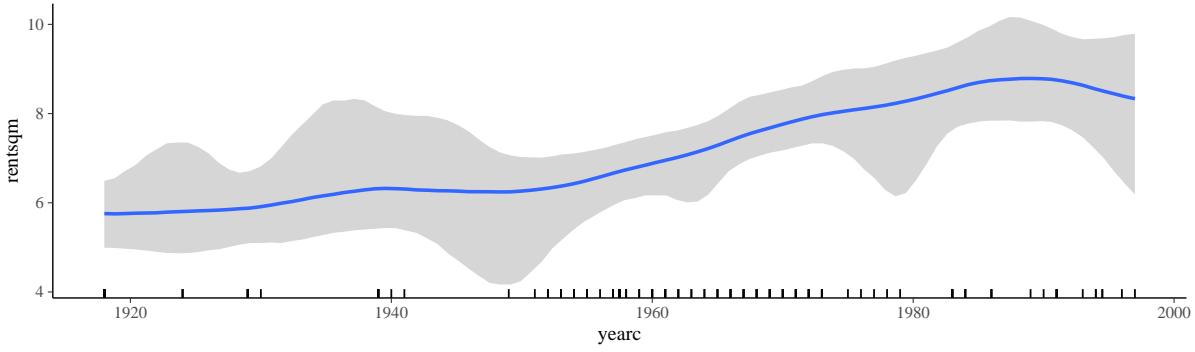


Figure 8.18: Conditional effects for model `fit_rent_gp1`. Ticks on the x-axis show the locations of observed predictor values.

When setting `rug = TRUE`, we see which predictor values on the x-axis were present in the data. And we see in Figure 8.18 that the uncertainty of the GP predictions spikes specifically in those places where no predictor values were observed. Especially when the lengthscale parameter is small, the uncertainty of GPs may increase strongly in areas without observed data. The default `conditional_effects` plot hides the origin of these uncertainty spikes though, because we only see the uncertainty intervals, computed as a summary over the individually predicted functions, not the individual functions themselves. To obtain the latter as one function per posterior draw, we can set `sphagetti = TRUE`:

```
ce_gp1 <- conditional_effects(
  fit_rent_gp1, spaghetti = TRUE, ndraws = 200
)
plot(ce_gp1, rug = TRUE)
```

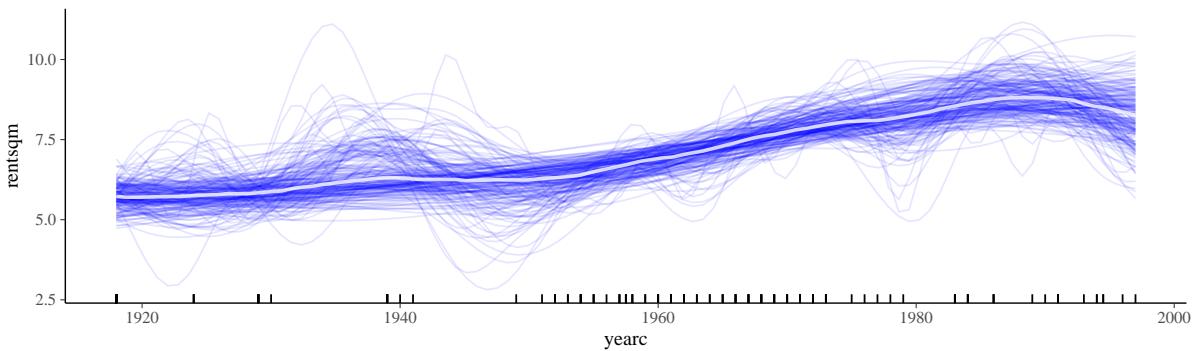


Figure 8.19: Conditional effects for 200 individual posterior draws of model `fit_rent_gp1`. Ticks on the x-axis show the locations of observed predictor values.

In Figure 8.19 only show predictions of 200 draws (i.e., 200 lines) to not clutter the plot too

much. What we gather from this visualization is that the increased uncertainty originates from few very wiggly functions. While many functions are not very wiggly, the data are also compatible with highly wiggly functions, so we see some of them also in the posterior predictions. If we checked exactly which posterior draws would produce these wiggly functions, we would find them to have small lengthscale values. Conversely, particularly non-wiggly functions correspond to posterior draws with high lengthscale values.

Next, let's talk about the speed of fitting GP models. On my machine, each chain took about 8 seconds, which is quite slow for such a small dataset. In comparison, the thin-plate spline model `fit_rent5` took only 1 second per chain for the same dataset. This is not a coincidence. If you have come across GPs before, you will have heard that they “scale badly” to large datasets in the sense that their computation time becomes very high for larger data. Let's unpack what this means concretely.

The computationally expensive part of evaluating GPs is computing the Cholesky factor  $A$  (or equivalently inverting  $\Sigma$ ) needed for Equation 8.7. For  $N$  observations, this requires a number of operations that is roughly proportional to  $N^3$ . We say the computational complexity of GPs scales *cubically*. When  $N$  is large,  $N^3$  becomes huge, which explains why GPs become slow for larger data. When fitting GPs within Stan, we need to compute the Cholesky factor for each sub-step inside each MCMC iteration, making the computational issues particularly drastic. In comparison, splines scale *linearly* with  $N$ , that is, they only require a number of operation that is proportional to  $N$  itself. As such, we expect them to scale much better to larger data.

This is only part of the story though. As we can see from Equation 8.6, the covariance of the function values  $f(x)$  is fully determined by the inputs  $x$ . So if we have two observations with equal  $x$ , their covariances with all other observations will be the same. As such, we actually don't need to compute the Cholesky factor over all  $N$  observations but only over all the  $N_u$  *unique predictor values*. If  $N_u$  remains small, GPs will scale well even as  $N$  increases. For our small data subset of  $N = 200$  observations, the number of unique values of the predictor `yearc` is  $N_u = 48$ , and only this number is what we really need to worry about. In brms, we use this computational trick by default, but if you want to make your life harder for some reason, you could still turn it off via `gp(..., gr = FALSE)`.

When moving on to the full dataset of  $N = 3082$  observations, the number of unique values in `yearc` increases only to  $N_u = 68$ , which is still within a computationally manageable range.

```
fit_rent_gp2 <- brm(
  rentsqm ~ gp(yearc), data = rent99,
  control = list(adapt_delta = 0.95),
  cores = 4, init = 0
)
```

I added `init = 0` as it may happen that one stray chain initializes in an area too far away from

the posterior, such that it never converges properly. This may happen with all GP models, and setting `init = 0` often fixes the issue.

Fitting this model took about 100 seconds per chain on my machine. This is again slower than the comparable thin-plate spline model `fit_rent6` requiring 26 seconds per chain, but not so slow that it would stop us from practically applying the GP model. In fact, if we further increased  $N$  for this dataset while keeping  $N_u$  constant, that is, adding data for more apartments constructed during the same years, GP models would eventually become faster than comparable spline models.

```
summary(fit_rent_gp2)
```

```
Family: gaussian
Links: mu = identity
Formula: rentsqm ~ gp(yearc)
Data: rent99 (Number of observations: 3082)
```

#### Gaussian Process Hyperparameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sdgp(gpyearc)	2.22	1.13	0.90	5.27	1.00	786	1.00	786	1645
lscale(gpyearc)	0.45	0.19	0.16	0.91	1.01	327	1.01	327	272

#### Regression Coefficients:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	7.34	1.45	4.54	10.40	1.00	1101	1.00	1101	1193

#### Further Distributional Parameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sigma	2.21	0.03	2.16	2.27	1.00	5756	1.00	5756	3246

From the summary output, we see that, when fitted to the full data, the lengthscale parameter now has a higher posterior mean of around 0.45 and a much tighter 95%-CI. In particular, the lower CI bound is now around 0.2 suggesting that very small lengthscales are practically ruled out under the posterior. This is also what we see represented in the predictions (Figure 8.20):

```
plot(conditional_effects(fit_rent_gp2, "yearc"), rug = TRUE)
```

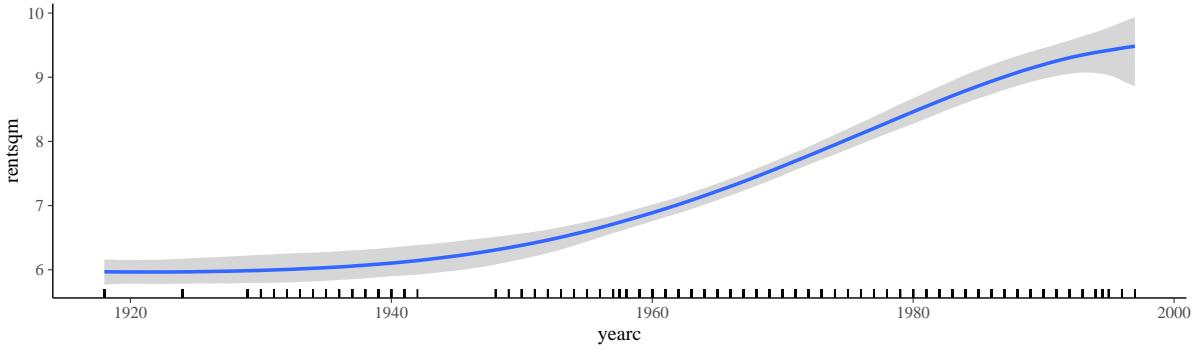


Figure 8.20: Conditional effects for model `fit_rent_gp2`. Ticks on the x-axis show the locations of observed predictor values.

While there are few predictor ranges where there is no data available (specifically around the first and second world war), the GP uncertainty doesn't spike there anymore because the posterior now strongly favors non-wiggly functions. This is also what we see when plotting individual function predictions (Figure 8.21):

```
ce_gp2 <- conditional_effects(
  fit_rent_gp2, spaghetti = TRUE, ndraws = 200
)
plot(ce_gp2, rug = TRUE)
```

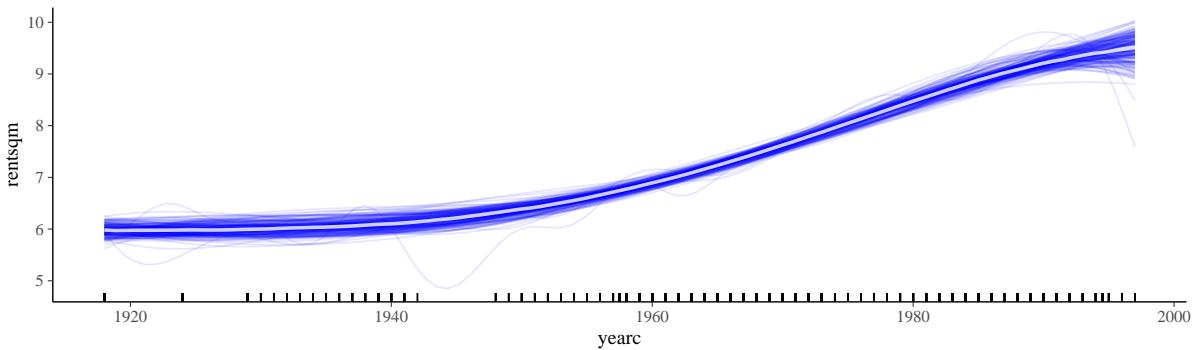


Figure 8.21: Conditional effects for 200 individual posterior draws of model `fit_rent_gp1`. Ticks on the x-axis show the locations of observed predictor values.

Overall, the predictive plots closely resemble those obtained by our spline models (compare with the left-hand side of Figure 8.11). An indeed, also the out-of-sample predictive performance as evaluated by LOO-CV is highly similar:

```
loo_rent_gp2 <- loo(fit_rent_gp2)
loo_compare(loo_rent_gp2, loo_rent6)
```

	elpd_diff	se_diff
fit_rent6	0.0	0.0
fit_rent_gp2	0.0	0.5

Generally, if the true underlying functions are not super wiggly and if there are no larger predictor spaces without observed data, spline and GP models are likely to perform highly similar. However, for true functions that are more wiggly, GPs may be favorable as they can adjust their flexibility in a data-driven manner through their lengthscale parameter (e.g., see Kołczyńska and Bürkner 2024 for empirical evidence). In contrast, most splines have an implicitly fixed lengthscale preventing them from handling very wiggly functions unless we use a very high number of basis functions.

### 8.5.1 Priors on Gaussian process hyperparameters

Specifying priors on GP hyperparameters, specifically for the lengthscale, is a delicate matter. Not only does the lengthscale strongly affect the behavior of the estimated functions (see Figure 8.16), it is often also insufficiently identified by data alone. In general, we can only estimate lengthscale values that are between the minimum and maximum distance between any two input points in the data. The reason is that the data can only provide information on dependencies within that range. Accordingly, dependencies that range longer than the maximum distance or shorter than the minimum distance cannot be identified. This also has practical implication for our MCMC chains. If a chain wanders off into a non-identified space of a lengthscale parameter, we immediately get divergent transitions unless an informative prior holds the posterior in place.

We can use the knowledge about the identified range of lengthscale parameters to define an informative prior that has most of its probability within that range. For example, Betancourt (n.d.) suggests to use an inverse gamma prior with shape and rate parameters  $a$  and  $b$ ,

$$\rho \sim \text{invGamma}(a, b),$$

such that the prior has 1% and 99% quantiles equal to  $Q_{0.01}(\rho) = \min |x - x'|$  and  $Q_{0.99}(\rho) = \max |x - x'|$  for values  $x, x'$  in the observed data. This creates a set of two non-linear equations that we can solve for  $a$  and  $b$ . For our data, when scaling the construction year to the unit interval, the maximal distance between two inputs is 1 by construction and the minimum distance is roughly 0.013, equivalent to one year on the original scale. This results in

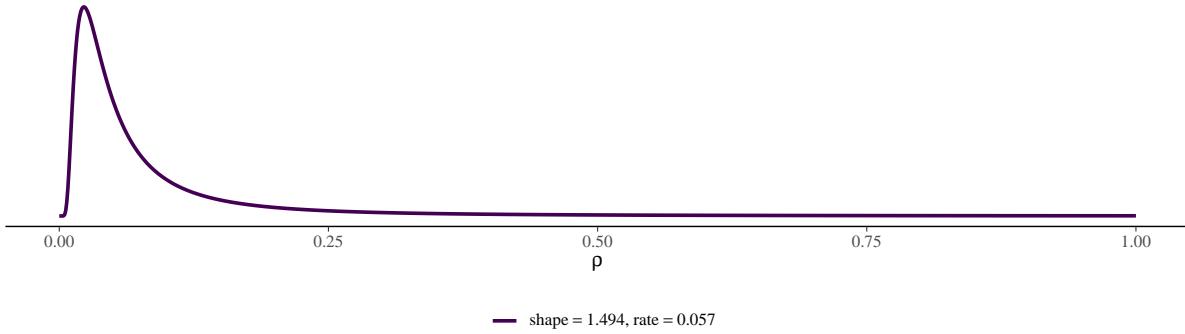


Figure 8.22: Default inverse gamma prior for the lengthscale parameter of model `fit_rent_gp1`.

shape and rate values of  $a \approx 1.494$  and  $b \approx 0.057$  leading to the prior illustrated in Figure 8.22. We can confirm that this was indeed the applied prior by looking at the output of `prior_summary(fit_rent_gp1)`.

If we remove this default prior from the model and instead assume an improper flat prior for the lengthscale, we get very bad convergence issues, which cannot be fixed by the usual means of running longer chains or increasing `adapt_delta`. You can try out the flat prior model yourself by running the following code:

```
prior_flat <- prior("", class = lscale, coef = gpyearc)
fit_rent_gp1_flat <- update(fit_rent_gp1, prior = prior_flat)
```

Apparently some kind of informative prior is required to make this model work. As an alternative informative prior that favors larger lengthscales (i.e., less wiggly functions), we can use a  $\text{normal}(0.5, 0.25)$  prior. It roughly has 95% of its probability in the unit interval, so it is largely located in the identified lengthscale range for the given data.

```
prior_normal <- prior("normal(0.5, 0.25)", class = lscale, coef = gpyearc)
fit_rent_gp1_normal <- update(fit_rent_gp1, prior = prior_normal, cores = 4)
```

```
summary(fit_rent_gp1_normal)
```

```
Family: gaussian
Links: mu = identity
Formula: rentsqm ~ gp(yearc)
Data: subrent99 (Number of observations: 200)
```

```
Gaussian Process Hyperparameters:
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
```

<code>sdgp(gpyearc)</code>	2.44	1.23	0.89	5.57	1.00	2087	2502
<code>lscale(gpyearc)</code>	0.56	0.20	0.21	0.97	1.00	2150	2203

Regression Coefficients:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	7.08	1.59	3.85	10.41	1.00	1916	1873

Further Distributional Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	2.32	0.12	2.09	2.57	1.00	6523	2774

We see that, under the normal prior on the lengthscale, the posterior favors larger lengthscale values than the posterior based on the default inverse gamma prior. This clearly indicates prior sensitivity for the lengthscale at least for the small data subset. Plotting the predictions (Figure 8.23) further confirms this impression as the predicted functions are much less wiggly than what was previously predicted by model with the default inverse gamma prior (Figure 8.18).

```
plot(conditional_effects(fit_rent_gp1_normal), rug = TRUE)
```

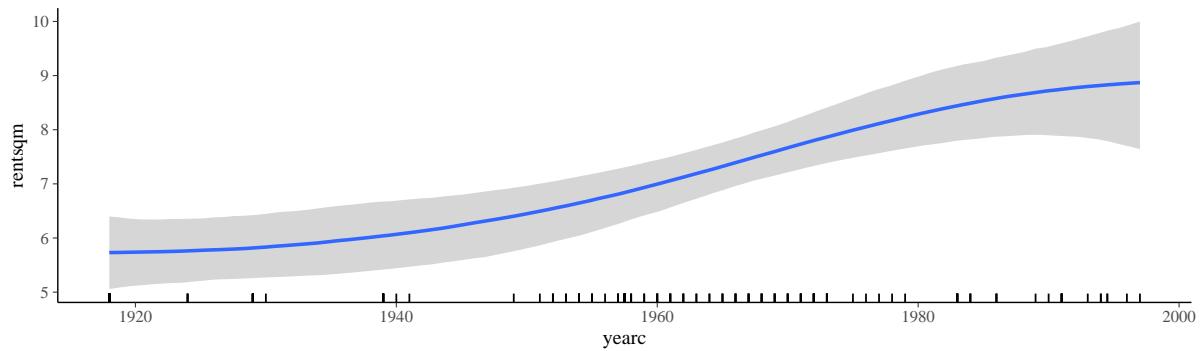


Figure 8.23: Conditional effects for model `fit_rent_gp1_normal`.

For the present data, the less wiggly version obtained from the normal prior model is probably more sensible, especially in light of what we have seen from the models fitted on the full data. In terms of out-of-sample predictive performance the two models only minimally differ from each other:

```
loo_reng_gp1 <- loo(fit_rent_gp1)
loo_rent_gp1_normal <- loo(fit_rent_gp1_normal)
loo_compare(loo_reng_gp1, loo_rent_gp1_normal)
```

	elpd_diff	se_diff
fit_rent_gp1_normal	0.0	0.0
fit_rent_gp1	-0.9	1.1

This makes sense given that the model predictions largely only differ for construction years where there is no observed data at all. Still, what we learn from this prior sensitivity check is that the brms default prior on lengthscale parameters may favor too wiggly functions. So if we a priori prefer less wiggly functions, lengthscale priors with a higher mean – such as  $\text{normal}(0.5, 0.25)$  if we work with auto-scaled predictors – can be a good alternative.

Compared to the prior on the lengthscale, the prior on the marginal standard deviation parameter is less important. Here, brms uses the default scale prior that is also used for other standard deviation parameters such as the residual standard deviation  $\sigma$  or the standard deviation hyperparameters of multilevel models or splines. Namely a half Student-t prior with 3 degrees of freedom and a scale parameter that is proportional to the standard deviation of the response (see also Section 4.4).

### 8.5.2 Other Gaussian process kernels

So far we have only worked with the squared exponential kernel, which assumes very smooth functions with infinitely many existing derivatives. The assumption of infinite smoothness may be unrealistic for real data. For our data, the properties of the time series of the rents in Munich may have changed suddenly due to external events, say, due to policy changes in tenancy law. This would in turn imply a non-smooth functional form not well represented by a GP with squared exponential kernel. If we prefer function that are a bit “rougher”, we can use other kernels from the Matern family. For example, the so called Matern 5/2 kernel

$$K_{M5/2}(x, x') = \alpha^2 \left( 1 + \sqrt{5} \frac{|x - x'|}{\rho} + \frac{5(x - x')^2}{3\rho^2} \right) \exp \left( -\sqrt{5} \frac{|x - x'|}{\rho} \right)$$

implies only twice differentiable functions but otherwise behaves similarly than the squared exponential kernel. As an analogy, the squared exponential kernels is to the family of Matern kernels as the normal distribution is to the family of Student-t distributions. As we increase the degrees of freedom to infinity in the Matern family of kernels, we obtain the infinitely smooth squared exponential kernel.

In brms, we can change the assumed covariance kernel via the `cov` argument of `gp()`, for example, `gp(yearc, cov = "matern52")` to use a Matern 5/2 kernel for the construction year predictor. For our data, the choice of kernel within the Matern class makes little difference, so we don't go further in this direction. As an exercise, try running the above models with different kernels and see if you spot relevant differences that I have overlooked.

### 8.5.3 Multi-dimensional Gaussian processes

So far, we have only fit Gaussian processes of the construction year, but we would also like to include the apartment area as additional predictor in our model. As a first step in this direction, we fit a model with two univariate GPs, one for `yearc` and one for `area`:

```
fit_rent_gp3 <- brm(
  rentsqm ~ gp(yearc) + gp(area), data = rent99,
  control = list(adapt_delta = 0.95), cores = 4,
  init = 0
)
```

If you run this model yourself, you will notice it is pretty slow. A closer inspection reveals that the number of unique values of `area` in the dataset is  $N_u = 132$ . And while this is just about twice the number of unique values of `yearc`, the cubic scaling implies an increase of computation time by roughly a factor of  $2^3 = 8$ . And this does not yet take into account the additional complexity of including two rather than one GP in the same model. We will not actually fit this model here but come back to it later on when discussing approximate GP models that run much faster.

The next, even more complex model is a two-dimensional GP of both `yearc` and `area`. With this GP, we not only model the main effects of the two predictors but also their non-linear interactions. Much like we did with the tensor product splines in Section 8.4.4. Since construction year and apartment area are on very different scales, the wiggliness of the two-dimensional GP may be different with respect to year and area. In the language of GPs this means that we should introduce two lengthscales parameters, one for each predictor. If we give each predictor its own lengthscales, we call it a *non-isotropic* GP. The non-isotropic squared exponential kernel over  $D$  input dimensions is

$$K_{\text{SE}}(x, x') = \alpha^2 \exp \left( -\frac{1}{2} \sum_{d=1}^D \frac{(x_d - x'_d)^2}{\rho_d^2} \right)$$

where  $\rho_d$  is the lengthscale parameter of input dimension  $d$ . In brms, we can specify a two-dimensional non-isotropic GP over construction year and apartment area as follows:

```
fit_rent_gp4 <- brm(
  rentsqm ~ gp(yearc, area, iso = FALSE), data = rent99,
  control = list(adapt_delta = 0.95), cores = 4,
  init = 0
)
```

However, the computational complexity of this model is brutal and would take at least multiple hours to fit. This is because, for multi-dimensional GPs,  $N_u$  becomes the number of unique predictor value *combinations*. For our data, we have  $N_u = 1695$  unique combinations of `yearc` and `area`, which is simply too much to reasonably handle within a full MCMC framework. Fortunately, we have access to approximate GP models that behave almost like exact GPs, but with substantially improved computational properties. One general class of such approximations is implemented in brms and we will focus on it next.

#### 8.5.4 Hilbert space approximate Gaussian processes

The key challenge when developing methods that approximate GPs is to retain their favorable properties as much as possible while alleviating their undesirable properties, specifically their terrible scaling with the number of unique input points. One class of GP approximations that I think strikes a good balance in this regard are Hilbert space approximate Gaussian processes [HSGPs; Solin and Särkkä (2020); Riutort-Mayol et al. (2023)]. They essentially construct a basis function approximation to GPs, thus scaling much like a spline and no longer like an exact GP.

In HSGPs, we approximate covariance kernels by the infinite sum over simple functions, concretely via

$$K(x, x') = \sum_{m=1}^{\infty} S(\lambda_j) \phi_m(x) \phi_m(x'),$$

where  $S$  is the so-called *spectral density* of  $K$ . For the squared exponential kernel, the spectral density is

$$S_{\text{SE}}(\omega) = \sqrt{2\pi} \alpha^2 \rho \exp\left(-\frac{1}{2}\rho^2 \omega^2\right).$$

The values  $\lambda_j$  and basis functions  $\phi$  are independent of the kernel and are usually given by

$$\lambda_j = \frac{j\pi}{2L}, \quad \phi_j(x) = \sqrt{\frac{1}{L}} \sin(\lambda_j(x + L)),$$

where  $L$  is the input value range within which we want the HSGP approximation to be accurate. By default in brms, we choose  $L$  to be 5/4 times the range of the observed  $x$  values, thus ensuring a good approximation slightly beyond the space of the observed input data.

So, in order to turn an exact GP into an HSGP, the main hurdle is to know the spectral density of its kernel; and only some kernels have a spectral density in the first place. The second hurdle is that we cannot practically compute an infinite sum, so we have to choose the number of

initial terms  $M$  of the sum (i.e., the number of basis functions) that we evaluate. Only because we are truncating the infinite sum do HSGPs become an approximation. Otherwise, they would yield exactly the same result as their exact GP counterparts.

The advantage of approximating a kernel with the help of its spectral density is that the Cholesky factor becomes very simple to evaluate. It no longer scales cubically with the number of unique input points but linearly with the number of observations, just like a spline. The HSGP-implied prior over functions  $f$  evaluated at value  $x$  is simply computed as

$$f(x) \approx \sum_{m=1}^M \sqrt{S(\lambda_m)} \phi_m(x) z_m, \quad z_m \sim \text{normal}(0, 1).$$

All involved terms are fast to evaluate, rendering HSGPs much more efficient than exact GPs when fitted to large data.

In brms, we activate HSGPs within the `gp()` function by choosing the number of evaluated basis functions  $M$ , an argument called `k` for consistency with the spline functions `s()` and `t2()`. For our example, we can fit an HSGP of construction year with  $M = 10$  basis functions as follows:

```
fit_rent_hsgp2 <- brm(
  rentsqm ~ gp(yearc, k = 10), data = rent99,
  control = list(adapt_delta = 0.95), cores = 4
)
```

```
summary(fit_rent_hsgp2)
```

```
Family: gaussian
Links: mu = identity
Formula: rentsqm ~ gp(yearc, k = 10)
Data: rent99 (Number of observations: 3082)
```

#### Gaussian Process Hyperparameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sdgp(gpyearc)	2.39	1.30	0.94	6.04	1.00	909	476		
lscale(gpyearc)	0.40	0.22	0.02	0.84	1.01	519	823		

#### Regression Coefficients:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	7.17	1.34	4.16	9.86	1.01	864	394		

#### Further Distributional Parameters:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	2.21	0.03	2.16	2.27	1.00	4228	2696

The structure of the summary is the same as for the exact GP model, since the differences between GPs and HSGPs lie in the latent  $z$  variables not shown in the summary. What we do see though is that the HSGP posterior of the lengthscale parameter favors slightly smaller values than those of the corresponding exact GP posterior (model `fit_rent_gp2`). If we set  $k$  to higher values, say to  $k=20$ , these small differences would vanish too. In terms of predictive performance, the current HSGP with  $k=10$  and its corresponding exact GP show almost the same results already:

```
loo_rent_hsgp2 <- loo(fit_rent_hsgp2)
loo_compare(loo_rent_hsgp2, loo_rent_gp2)
```

	elpd_diff	se_diff
fit_rent_hsgp2	0.0	0.0
fit_rent_gp2	-0.1	0.2

Also the `conditional_effects` curve looks highly similar (see Figure 8.24):

```
plot(conditional_effects(fit_rent_hsgp2), rug = TRUE)
```

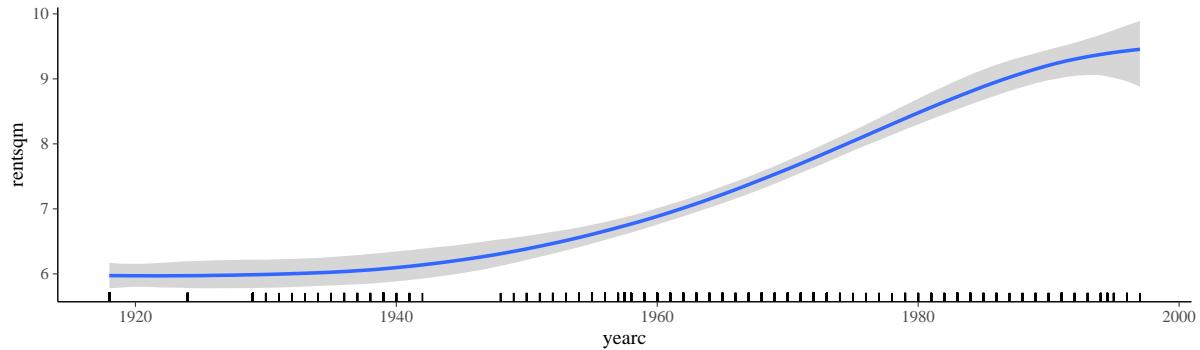


Figure 8.24: Conditional effects for model `fit_rent_hsgp2`.

Of course, in practice, we apply HSGPs only when their exact GP counterparts are computationally too expensive. So we won't usually be able to compare GPs and HSGPs to check if the lengthscale estimates of the latter are already "good enough". Instead, we have to fit the HSGP with a few  $k$  values to check if the posterior and posterior predictions are sensitive to that choice – in the same way as we would do with splines. Even if we fit a few HSGPs this way, their combined runtime is likely to still be lower than that of a single exact GP. In

Riutort-Mayol et al. (2023), we have provided formulas for the minimal number of basis functions required to estimate a given lengthscale. However, since we don't know the lengthscale a priori, I often just start with `k=10` and then increase it further if needed.

Fitting a GP of only construction year was also possible without an Hilbert space approximation. However, once we added a GP of apartment area too, the runtime of the exact GP (model `fit_rent_gp3`) become too long. With HSGPs, that model is now fast to fit:

```
fit_rent_hsgp3 <- brm(
  rentsqm ~ gp(yearc, k = 10) + gp(area, k = 10),
  data = rent99, control = list(adapt_delta = 0.95),
  cores = 4, init = 0
)
```

On my machine, it took just about 131 seconds per chain on average.

```
summary(fit_rent_hsgp3)
```

```
Family: gaussian
Links: mu = identity
Formula: rentsqm ~ gp(yearc, k = 10) + gp(area, k = 10)
Data: rent99 (Number of observations: 3082)
```

#### Gaussian Process Hyperparameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sdgp(gpyearc)	2.23	1.12	0.90	5.05	1.00	2224	2620		
sdgp(gparea)	5.93	2.58	2.57	12.54	1.00	1627	1277		
lscale(gpyearc)	0.16	0.14	0.01	0.50	1.01	744	912		
lscale(gparea)	0.10	0.07	0.02	0.27	1.00	1517	900		

#### Regression Coefficients:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	8.47	1.96	4.56	12.46	1.00	3069	1.00	3069	1877

#### Further Distributional Parameters:

	Estimate	Est.Error	l-95%	CI	u-95%	CI	Rhat	Bulk_ESS	Tail_ESS
sigma	2.03	0.03	1.98	2.08	1.00	5693	1.00	5693	2959

When comparing the GP hyperparameters of `yearc` and `area`, we see that the marginal standard deviation is higher for `area`. Additionally, we see that `area` has a little smaller lengthscale on average, indicating somewhat higher wigginess in its predictive function. Together, this indicates higher variation in the response explained by `area` compared to the variation explained by `yearc`. We can also see this represented visually:

```
conditional_effects(fit_rent_hsgp3)
```

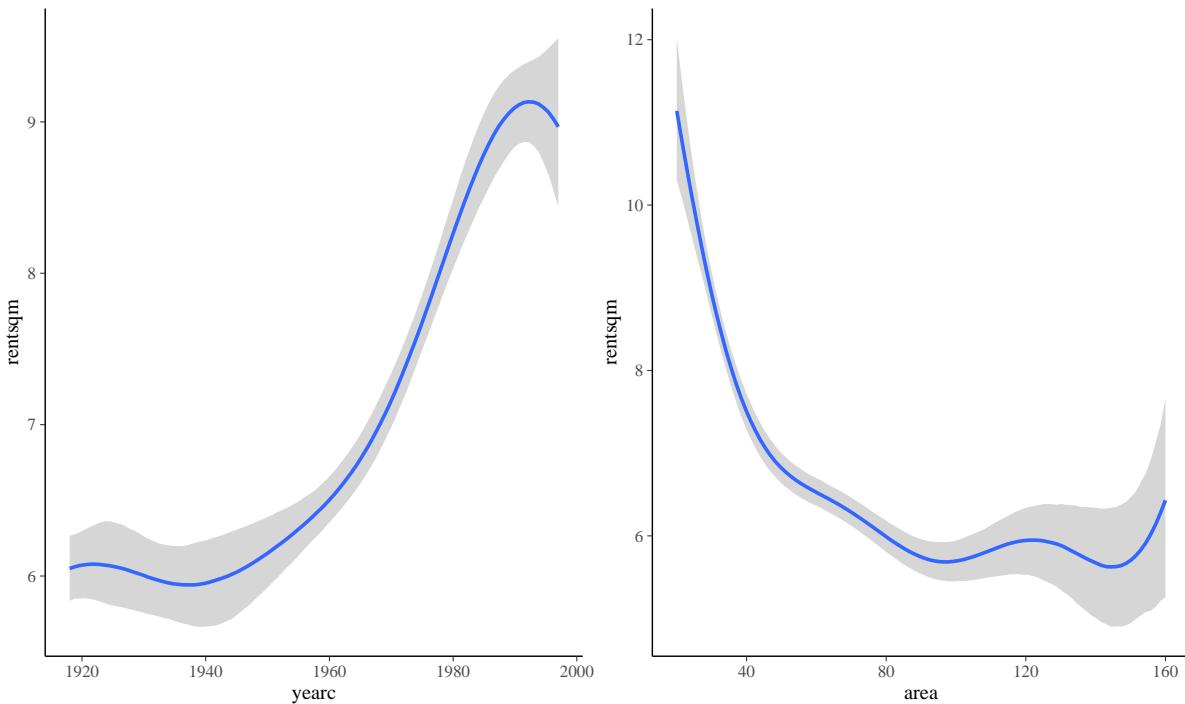


Figure 8.25: Conditional effects for model `fit_rent_hsgp3`.

In Figure 8.25 we see both a higher range of predicted values in the function of `area` and its somewhat higher wigginess compared to the function of `yearc`.

Next, we will fit a two-dimensional HSGP. They work similarly to tensor splines in that we take all the basis functions of one predictor and multiply them with all the basis functions of the other predictor. So, if we set  $k = 10$  this is to be interpreted as 10 basis functions per predictor, such that we have  $10^2 = 100$  basis functions in total for a two-dimensional HSGP. We see that, as we increase the number of predictors, the number of basis functions scales exponentially. As such, HSGPs quickly become inefficient as we increase the number of jointly modeled predictors. Using 3 or less predictors should be fine, but anything beyond that will quickly become computationally infeasible. Also, the more basis functions we model, the more basis coefficients we have to estimate, which can cause convergence issues if the number of coefficients is too high relative to the size of the data.

In order to fit a two-dimensional HSGP of construction year and apartment area with separate lengthscale parameters, we specify:

```

fit_rent_hsgp4 <- brm(
  rentsqm ~ gp(yearc, area, k = 10, iso = FALSE),
  data = rent99, control = list(adapt_delta = 0.95),
  cores = 4, init = 0
)

```

On my laptop, this model took just 131 seconds on average per chain, which is pretty fast considering the size of the data and the complexity of the model. We skip discussion of the summary output as it doesn't yield relevant new insights.

Similar to spline models, we can plot two-dimensional surface plots of the predictions as follows (see Figure 8.26):

```

ce_hsgp4_surface <- conditional_effects(fit_rent_hsgp4, "yearc:area", surface = TRUE)

ce_hsgp4_surface <- conditional_effects(fit_rent_hsgp4, "yearc:area", surface = TRUE)
plot(ce_hsgp4_surface, stype = "contour")
plot(ce_hsgp4_surface, stype = "raster")

```

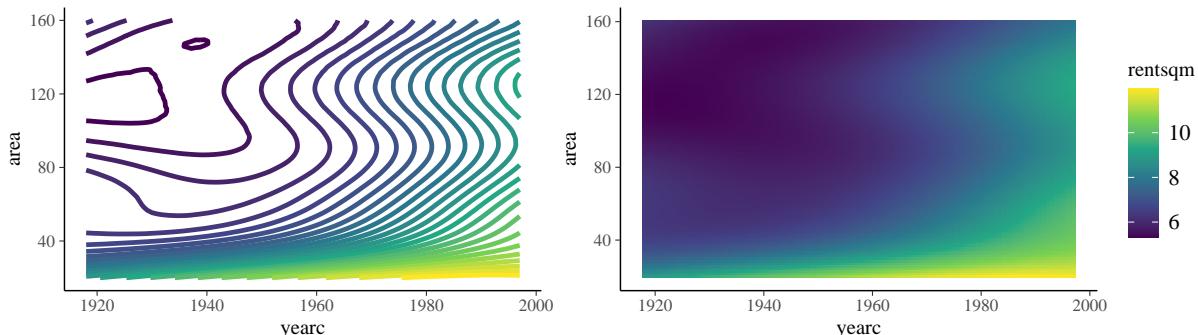


Figure 8.26: Conditional effects surface plots in both contour and raster format of model `fit_rent_hsgp4`.

Again similar to the spline models, the predictive plots provide little evidence of an interaction between construction year and apartment area. This observation is confirmed by the LOO-CV results, which show strong improvements when adding apartment area, but no benefit of modeling a two-dimensional HSGP vs. two one-dimensional HSGPs:

```

loo_rent_hsgp3 <- loo(fit_rent_hsgp3)
loo_rent_hsgp4 <- loo(fit_rent_hsgp4)
loo_compare(loo_rent_hsgp2, loo_rent_hsgp3, loo_rent_hsgp4)

```

	elpd_diff	se_diff
fit_rent_hsgp4	0.0	0.0
fit_rent_hsgp3	-0.8	6.6
fit_rent_hsgp2	-260.5	24.4

## 8.6 Summary

In this chapter, we have learned how to fit data-driven non-linear models, which estimate the form of the predicting functions from the data in a highly adaptive manner. Both splines and GPs are powerful tools for this purpose and brms offers a wide range of options within these two model classes. To achieve the desired flexibility, data-driven non-linear models require a large number of parameters, rendering these models more data hungry than comparable linear models and increase their estimation time. On top of that, the increased flexibility comes at the cost of decreased interpretability, such that often only visualizations of the predicted functions are easily interpretable. In terms of scalability to large data, splines usually do better than GPs and also tend to have less convergence issues. My personal default are thus splines. When requiring the additional flexibility of GPs induced through their estimable lengthscale parameters, I recommend HSGPs as they provide a good tradeoff between GP flexibility and spline scalability. As always in statistical modeling, there is no free lunch. So we have to decide carefully which predictors we model non-linearly and what properties of the function approximators we prioritize.

# References

- Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using `\Pkg{lme4}`.” *Journal of Statistical Software* 67 (1): 1–48.
- Beisemann, Marie, Philipp Doeblner, and Heinz Holling. 2020. “Comparison of Random-Effects Meta-Analysis Models for the Relative Risk in the Case of Rare Events: A Simulation Study.” *Biometrical Journal* 62 (7): 1597–1630. <https://doi.org/10.1002/bimj.201900379>.
- Belenky, Gregory, Nancy J. Wesensten, David R. Thorne, Maria L. Thomas, Helen C. Sing, Daniel P. Redmond, Michael B. Russo, and Thomas J. Balkin. 2003. “Patterns of Performance Degradation and Restoration During Sleep Restriction and Subsequent Recovery: A Sleep Dose-Response Study.” *Journal of Sleep Research* 12 (1). <https://doi.org/10.1046/j.1365-2869.2003.00337.x>.
- Bernardo, José M., and Adrian F. M. Smith. 1994. *Bayesian Theory*. Hoboken: Wiley.
- Betancourt, Michael. n.d. “Robust Gaussian Process Modeling.” [https://betanalpha.github.io/assets/case\\_studies/gaussian\\_processes.html](https://betanalpha.github.io/assets/case_studies/gaussian_processes.html).
- Blomstedt, Paul, Diego Mesquita, Omar Rivasplata, Jarno Lintusaari, Tuomas Sivula, Jukka Corander, and Samuel Kaski. 2024. “Meta-Analysis of Bayesian Analyses.” *arXiv Preprint*. <https://doi.org/10.1214/24-BA1465>.
- Bockting, Florence, Stefan T. Radev, and Paul-Christian Bürkner. 2023. “Simulation-Based Prior Knowledge Elicitation for Parametric Bayesian Models.” *arXiv Preprint*. <https://doi.org/10.48550/arXiv.2308.11672>.
- Breslow, N. E., and D. G. Clayton. 1993. “Approximate Inference in Generalized Linear Mixed Models.” *Journal of the American Statistical Association* 88 (421): 9–25. <https://doi.org/10.1080/01621459.1993.10594284>.
- Brilleman, Samuel L, Michael J Crowther, Margarita Moreno-Betancur, Jacqueline Buros Novik, James Dunyak, Nidal Al-Huniti, Robert Fox, Jeff Hammerbacher, and Rory Wolfe. 2019. “Joint Longitudinal and Time-to-Event Models for Multilevel Hierarchical Data.” *Statistical Methods in Medical Research* 28 (12): 3502–15. <https://doi.org/10.1177/0962280218808821>.
- Bürkner, Paul-Christian. 2021. “Bayesian Item Response Modelling in R with Brms and Stan.” *Journal of Statistical Software*, 1–54.
- Bürkner, Paul-Christian, Jonah Gabry, Matthew Kay, and Aki Vehtari. 2024. “Posterior: Tools for Working with Posterior Distributions.” <https://mc-stan.org/posterior/>.
- Bürkner, Paul-Christian, Jonah Gabry, and Aki Vehtari. 2021. “Efficient Leave-One-Out Cross-Validation for Bayesian Non-Factorized Normal and Student-t Models.” *Computational Statistics* 36 (2): 1243–61.

- Bürkner, Paul-Christian, Maximilian Scholz, and Stefan T. Radev. 2023b. “Some Models Are Useful, but How Do We Know Which Ones? Towards a Unified Bayesian Model Taxonomy.” *Statistics Surveys* 17 (none): 216–310. <https://doi.org/10.1214/23-SS145>.
- . 2023a. “Some Models Are Useful, but How Do We Know Which Ones? Towards a Unified Bayesian Model Taxonomy.” *Statistics Surveys* 17: 216–310. <https://doi.org/10.1214/23-SS145>.
- Colditz, Graham A., Timothy F. Brewer, Catherine S. Berkey, Mary E. Wilson, Elisabeth Burdick, Harvey V. Fineberg, and Frederick Mosteller. 1994. “Efficacy of BCG Vaccine in the Prevention of Tuberculosis: Meta-Analysis of the Published Literature.” *JAMA* 271 (9): 698–702. <https://doi.org/10.1001/jama.1994.03510330076038>.
- Cooper, Harris, Larry V. Hedges, and Jeffrey C. Valentine. 2019. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation.
- Cox, D. R. 1972. “Regression Models and Life-Tables.” *Journal of the Royal Statistical Society: Series B (Methodological)* 34 (2): 187–202. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
- De Boeck, Paul, and Mark Wilson, eds. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York, NY: Springer. <https://doi.org/10.1007/978-1-4757-3990-9>.
- Engel, Joachim, and Peter Sedlmeier. 2010. “Regression Und Korrelation: Alles Klar, Oder Voller Tücken?” [https://www.stochastik-in-der-schule.de/sisonline/struktur/Jahrgang30-2010/Heft%202/2010-2\\_EngelSedlmeier.pdf](https://www.stochastik-in-der-schule.de/sisonline/struktur/Jahrgang30-2010/Heft%202/2010-2_EngelSedlmeier.pdf).
- Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang, and Brian Marx. 2013. *Regression: Models, Methods and Applications*. Springer Science & Business Media.
- Gabry, Jonah, and Tristan Mahr. 2024. “Bayesplot: Plotting for Bayesian Models.” <https://mc-stan.org/bayesplot/>.
- Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis (3rd Edition)*. London: Chapman; Hall/CRC.
- Gelman, Andrew, Ben Goodrich, Jonah Gabry, and Aki Vehtari. 2019. “R-Squared for Bayesian Regression Models.” *The American Statistician* 73 (3): 307–9.
- Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press.
- Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. “Bayesian Workflow.” *arXiv Preprint*.
- Gronau, Quentin F., Alexandra Sarafoglou, Dora Matzke, Alexander Ly, Udo Boehm, Maarten Marsman, David S. Leslie, Jonathan J. Forster, Eric-Jan Wagenmakers, and Helen Stein-groever. 2017. “A Tutorial on Bridge Sampling.” *Journal of Mathematical Psychology* 81: 80–97. <https://doi.org/10.1016/j.jmp.2017.09.005>.
- Gronau, Quentin F., Henrik Singmann, and Eric-Jan Wagenmakers. 2020. “Bridgesampling: An R Package for Estimating Normalizing Constants.” *Journal of Statistical Software* 92 (10): 1–29. <https://doi.org/10.18637/jss.v092.i10>.
- Hadfield, JD, A Nutall, D Osorio, and IPF Owens. 2007. “Testing the Phenotypic Gambit: Phenotypic, Genetic and Environmental Correlations of Colour.” *Journal of Evolutionary*

- Biology* 20 (2): 549–57. <https://doi.org/10.1111/j.1420-9101.2006.01262.x>.
- Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. Springer.
- Hendricks, Paul. 2015. *Titanic: Titanic Passenger Survival Data Set*. <https://CRAN.R-project.org/package=titanic>.
- Juárez, Miguel A., and Mark F. J. Steel. 2010. “Model-Based Clustering of Non-Gaussian Panel Data Based on Skew- *t* Distributions.” *Journal of Business & Economic Statistics* 28 (1): 52–66. <https://doi.org/10.1198/jbes.2009.07145>.
- Kallioinen, Noa, Topi Paananen, Paul-Christian Bürkner, and Aki Vehtari. 2023. “Detecting and Diagnosing Prior and Likelihood Sensitivity with Power-Scaling.” *Statistics and Computing* 34 (1): 57. <https://doi.org/10.1007/s11222-023-10366-5>.
- Kaplan, E. L., and Paul Meier. 1958. “Nonparametric Estimation from Incomplete Observations.” *Journal of the American Statistical Association* 53 (282): 457–81. <https://doi.org/10.1080/01621459.1958.10501452>.
- Kołczyńska, Marta, and Paul-Christian Bürkner. 2024. “Modeling Public Opinion over Time: A Simulation Study of Latent Trend Models.” *Journal of Survey Statistics and Methodology* 12 (1): 130–54.
- Kucukelbir, Alp, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. 2017. “Automatic Differentiation Variational Inference.” *Journal of Machine Learning Research* 18 (14): 1–45.
- Kurz, Solomon. 2019. “Statistical Rethinking with Brms, Ggplot2, and the Tidyverse.” [https://bookdown.org/ajkurz/Statistical\\_Rethinking\\_recoded/](https://bookdown.org/ajkurz/Statistical_Rethinking_recoded/).
- Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2009. “Generating Random Correlation Matrices Based on Vines and Extended Onion Method.” *Journal of Multivariate Analysis* 100 (9): 1989–2001.
- Magnusson, Kristoffer. 2018. “Confounded Dose-Response Effects of Treatment Adherence: Fitting Bayesian Instrumental Variable Models Using Brms.” <https://rpsychologist.com/adherence-analysis-IV-brms>.
- McCullagh, P. 2019. *Generalized Linear Models*. 2nd ed. New York: Routledge. <https://doi.org/10.1201/9780203753736>.
- McElreath, Richard. 2019. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan (2nd Edition)*. London: Chapman; Hall/CRC.
- McGilchrist, C. A., and C. W. Aisbett. 1991. “Regression with Frailty in Survival Analysis.” *Biometrics* 47 (2): 461–66. <https://doi.org/10.2307/2532138>.
- Meng, Xiao-Li, and Stephen Schilling. 2002. “Warp Bridge Sampling.” *Journal of Computational and Graphical Statistics* 11 (3): 552–86. <https://doi.org/10.1198/106186002457>.
- Mikkola, Petrus, Osvaldo A. Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, et al. 2023. “Prior Knowledge Elicitation: The Past, Present, and Future.” *Bayesian Analysis* -1: 1–33. <https://doi.org/10.1214/23-BA1381>.
- Paananen, Topi, Juho Piironen, Paul-Christian Bürkner, and Aki Vehtari. 2021. “Implicitly Adaptive Importance Sampling.” *Statistics and Computing* 31 (2). <https://doi.org/10.1007/s11222-020-09982-2>.
- Pearl, Judea. 2009. *Causality*. Cambridge University Press.

- Pedersen, Eric J, David L Miller, Gavin L Simpson, and Noam Ross. 2019. “Hierarchical Generalized Additive Models in Ecology: An Introduction with Mgcv.” *PeerJ* 7: e6876.
- Peralta, Victor, and Manuel J. Cuesta. 1994. “Psychometric Properties of the Positive and Negative Syndrome Scale (PANSS) in Schizophrenia.” *Psychiatry Research* 53 (1): 31–40. [https://doi.org/10.1016/0165-1781\(94\)90093-0](https://doi.org/10.1016/0165-1781(94)90093-0).
- Riutort-Mayol, Gabriel, Paul-Christian Bürkner, Michael R Andersen, Arno Solin, and Aki Vehtari. 2023. “Practical Hilbert Space Approximate Bayesian Gaussian Processes for Probabilistic Programming.” *Statistics and Computing* 33 (1): 17.
- Rue, Håvard, Sara Martino, and Nicolas Chopin. 2009. “Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (2): 319–92.
- Säilynoja, Teemu, Paul-Christian Bürkner, and Aki Vehtari. 2022. “Graphical Test for Discrete Uniformity and Its Applications in Goodness-of-Fit Evaluation and Multiple Sample Comparison.” *Statistics and Computing* 32 (2). <https://doi.org/10.1007/s11222-022-10090-6>.
- Scholz, Maximilian, and Paul-Christian Bürkner. 2023. “Prediction Can Be Safely Used as a Proxy for Explanation in Causally Consistent Bayesian Generalized Linear Models.” *arXiv Preprint*. <https://doi.org/10.48550/arXiv.2210.06927>.
- Sivula, Tuomas, Måns Magnusson, Asael Alonzo Matamoros, and Aki Vehtari. 2022. “Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison.” *arXiv Preprint*.
- Solin, Arno, and Simo Särkkä. 2020. “Hilbert Space Methods for Reduced-Rank Gaussian Process Regression.” *Statistics and Computing* 30 (2): 419–46.
- Stasinopoulos, Mikis, and Bob Rigby. 2016. *Gamlss.data: GAMLSS Data*. <https://CRAN.R-project.org/package=gamlss.data>.
- Thall, Peter F, and Stephen C Vail. 1990. “Some Covariance Models for Longitudinal Count Data with Overdispersion.” *Biometrics*, 657–71.
- Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. “Rank-Normalization, Folding, and Localization: An Improved \$\widehat{R}\$ for Assessing Convergence of MCMC (with Discussion).” *Bayesian Analysis* 16 (2): 667–718.
- Vehtari, Aki, and Janne Ojanen. 2012. “A Survey of Bayesian Predictive Methods for Model Assessment, Selection and Comparison.” *Statistics Surveys* 6: 142–228.
- Vehtari, Aki, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. 2024. “Pareto Smoothed Importance Sampling.” *Journal of Machine Learning Research* 25 (72): 1–58. <http://jmlr.org/papers/v25/19-556.html>.
- Viechtbauer, Wolfgang. 2010. “Conducting Meta-Analyses in R with the metafor Package.” *Journal of Statistical Software* 36 (3): 1–48. <https://doi.org/10.18637/jss.v036.i03>.
- Viechtbauer, Wolfgang, Thomas White, Daniel Noble, Alistair Senior, and W. Kyle Hamilton. 2025. *Metadat: Meta-Analysis Datasets*. <https://CRAN.R-project.org/package=metadat>.
- Wagenmakers, Eric-Jan, and Simon Farrell. 2004. “AIC Model Selection Using Akaike Weights.” *Psychonomic Bulletin & Review* 11 (1): 192–96. <https://doi.org/10.3758/BF03206482>.
- Westfall, Jacob, and Tal Yarkoni. 2016. “Statistically Controlling for Confounding Constructs

- Is Harder Than You Think.” *PLOS ONE* 11 (3): e0152719. <https://doi.org/10.1371/journal.pone.0152719>.
- Williams, Donald R., and Paul-Christian Bürkner. 2017. “Effects of Intranasal Oxytocin on Symptoms of Schizophrenia: A Multivariate Bayesian Meta-Analysis.” *Psychoneuroendocrinology* 75: 141–51. <https://doi.org/10.1016/j.psyneuen.2016.10.013>.
- Wood, S. N. 2003. “Thin-Plate Regression Splines.” *Journal of the Royal Statistical Society (B)* 65 (1): 95–114. <https://doi.org/10.1111/1467-9868.00374>.
- . 2011. “Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models.” *Journal of the Royal Statistical Society (B)* 73 (1): 3–36. <https://doi.org/10.1111/j.1467-9868.2010.00749.x>.
- . 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman; Hall/CRC.
- Wood, Simon N, Fabian Scheipl, and Julian J Faraway. 2013. “Straightforward Intermediate Rank Tensor Product Smoothing in Mixed Models.” *Statistics and Computing* 23 (3): 341–60.
- Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2018. “Using Stacking to Average Bayesian Predictive Distributions (with Discussion).” *Bayesian Analysis* 13 (3): 917–1007. <https://doi.org/10.1214/17-BA1091>.
- Zhang, Lu, Bob Carpenter, Andrew Gelman, and Aki Vehtari. 2022. “Pathfinder: Parallel Quasi-Newton Variational Inference.” *Journal of Machine Learning Research* 23 (306): 1–49.