8 Multiple linear regression in R-INLA

In Chapter 7 we explained how to obtain posterior distributions of regression parameters in R-INLA. Before diving into models with dependency structures (e.g. repeated measurements, or spatial and temporal data) we will analyse a relatively simple data set in this chapter using multiple linear regression. It allows us to discuss topics like fitted values, residuals, model validation, model selection, model visualisation, and simulations. Once we have this knowledge we will deal with dependency in all remaining chapters.

Prerequisite for this chapter: Knowledge of R and multiple linear regression is required.

8.1 Introduction

The data that will be analysed in this chapter are taken from Hopkins et al. (2013), who studied tool use by captive chimpanzees (*Pan troglodytes*). An experiment was carried out to simulate termite fishing in wild chimpanzees. Small PVC pipes were filled with food and the animals were given thin sticks, which allowed them to eat the food (provided they could figure out how to put the stick into the PVC pipe). The underlying question that we will address in this chapter is whether tool-use skills differ by sex, age, and rearing experience.

In the Hopkins et al. paper, tool-use skill is quantified as the time required per successful dip. This variable is called ‘latency’. For each chimpanzee a minimum of 50 successful attempts are recorded and average latency score for each of the 243 monkeys is provided in the online material of the paper. The provided latency scores are standardised. The larger the latency score, the longer it took (on average) the chimp to get food.

The covariates are sex (males vs. female), age (in years), rear (MR = ‘mother rear’, HR = ‘human rear’, WC = ‘wild caught’), and colony (chimps came from two research units).

8.2 Data exploration

Figure 8.1 shows Cleveland dotplots of all the variables. We temporarily coded the categorical variables colony, rear, and sex as numerical, otherwise the `dotplot` function in R gives an error message. Each point in the Latency panel represents the average motor performance score for a specific chimpanzee. Note that there are two animals with a relative large score. This means that on average it took these two animals a long time to figure out how to get the food! These chimps were removed from the analysis in the Hopkins et al. (2013) paper, but we will keep them in. The Age panel shows the age (in years) of the chimpanzees; there are no animals that are considerably younger or older. Sample sizes differ per colony, rear, and sex.

There are no spatial or temporal dependency aspects in this data set. Because we have an average score per animal, we do not have repeated measurements from the same animal. Assuming that the animals don't learn from one another or are in any other way related there is no need for regression models with complicated dependency structures. Actually, within a colony the chimpanzees are genetically linked, which in principle causes pseudoreplication. If you have this then you may want to look into models with phylogenetic correlation. Lajeunesse and Fox (2015) provide an easy-to-understand starting point.

Further data exploration steps did not indicate any major problems.

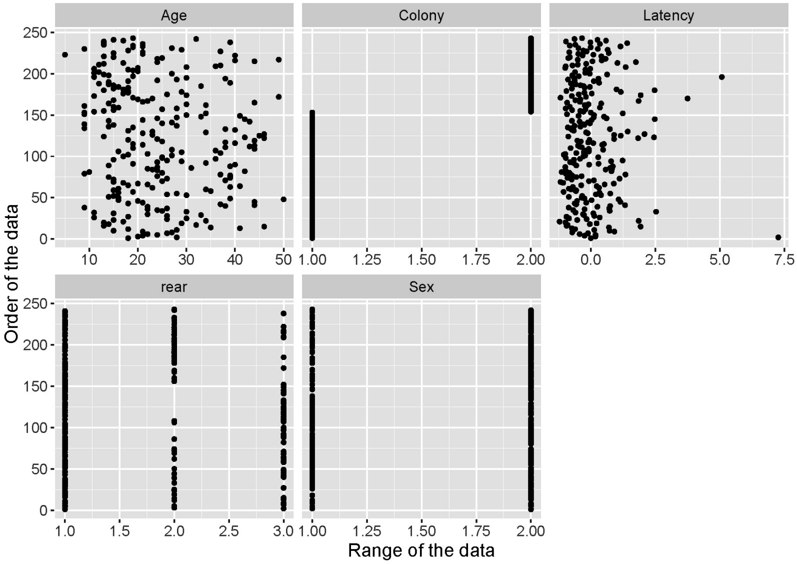

Figure 8.1. Cleveland dotplots of all the variables. The horizontal axes show the values of the variables and the vertical axes are the row numbers as imported from the data file.