What will this chapter tell me?

When I was 8 years old, my parents bought me a guitar for Christmas. Even then, I'd desperately wanted to play the guitar for years. I could not contain my excitement at getting this gift (had it been an *electric* guitar I think I would actually have exploded with excitement). The guitar came with a 'learn to play' book, and after a little while of trying to play what was on page 1 of this book, I readied myself to unleash a riff of universe-crushing power onto the world (well, 'Skip to my Lou' actually). But I couldn't do it. I burst into tears and ran upstairs to hide.[1] My dad sat with me and said 'Don't worry, Andy, everything is hard to begin with, but the more you practise the easier it gets.' In his comforting words, my dad was inadvertently teaching me about the relationship, or correlation, between two variables. These two variables could be related in three ways: (1) *positively related*, meaning that the more I practised my guitar, the better a guitar player I would become (i.e., my dad was telling me the truth); (2) *not related* at all, meaning that as I practised the guitar my playing ability would remain completely constant (i.e., my dad had fathered a cretin); or (3) *negatively related*, which would mean that the more I practised the guitar the worse a guitar player I would become (i.e., my dad had fathered an indescribably strange child). This chapter looks first at how we can express the relationships between variables statistically by looking at two measures: *covariance* and the *correlation coefficient*. We then discover how to carry out and interpret correlations in SPSS. The chapter ends by looking at more complex measures of relationships; in doing so it acts as a precursor to the chapter on multiple regression.

[1] This is not a dissimilar reaction to the one I have when publishers ask me for new editions of statistics textbooks.

Modelling relationships

In Chapter 4 I stressed the importance of looking at your data graphically before running any other analysis on them. I want to begin by reminding you that our starting point with a correlation analysis should be to look at some scatterplots of the variables we have measured. I am not going to repeat how to get SPSS to produce these graphs, but I am going to urge you (if you haven't done so already) to read Section 4.8 before embarking on the rest of this chapter.

Way back in Chapter 2 we started talking about fitting models to your data, and that these models represented the hypothesis you're trying to test. In the previous chapter we started to look at this process using a very specific set of models that are applied to ranked data and are useful when the data contain unusual cases or fail to meet the assumptions we discussed in Chapter 5. However, when these assumptions are met we can start to use a model known as the general linear model, which is an incredibly versatile and simple model. In fact, we've already encountered it. In Section 2.4 we discussed fitting models to the data and I mentioned that everything in statistics boils down to one simple idea (expressed in equation (2.1)):

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

To recap, this equation means that the data we observe can be predicted from the model we choose to fit to the data plus some amount of error. The 'model' in the equation will vary depending on the design of your study, the type of data you have and what it is you're trying to achieve with your model. If we want to model a relationship between variables then we're trying to predict an outcome variable from a predictor variable. Therefore, we need to factor the predictor variable into the model. As we saw in equation (2.3), we usually denote predictor variables with the letter X, so our model will be:

$$\text{outcome}_i = (bX_i) + \text{error}_i$$

This just means 'the outcome for an entity is predicted from the predictor variable plus some error'. As we have seen before, the model is described by a parameter, b, which in this context represents the relationship between the predictor variable (X) and the outcome. We use the sample data to estimate this parameter. Therefore, when we look at linear relationships between variables, this is the model we fit. We're interested in estimating the value of b because this will tell us how strong the relationship between the predictor and outcome is. When there is only one predictor variable in the model, b is known as the Pearson product-moment correlation coefficient (and, just to confuse us, is denoted by the letter r). How might we estimate this parameter? Like a quest for fire, we could search across the land … or, we could use maths.