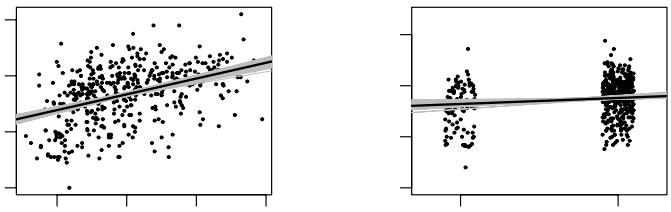

Figure 3.11 *Data and regression of child's test score on maternal IQ and high school completion, shown as a function of each of the two input variables (with light lines indicating uncertainty in the regressions). Values for high school completion have been jittered to make the points more distinct.*

```
  ylab="Child test score")
for (i in 1:10){
  curve (cbind (1, x, mean(mom.iq)) %*% beta.sim[i,], lwd=.5,
     col="gray", add=TRUE)
}
curve (cbind (1, x, mean(mom.iq)) %*% beta.hat, col="black", add=TRUE)
```

3.6 Assumptions and diagnostics

We now turn to the assumptions of the regression model, along with diagnostics that can be used to assess whether some of these assumptions are reasonable. Some of the most important assumptions, however, rely on the researcher's knowledge of the subject area and may not be directly testable from the available data alone.

Assumptions of the regression model

We list the assumptions of the regression model in *decreasing* order of importance.

1. *Validity.* Most importantly, the data you are analyzing should map to the research question you are trying to answer. This sounds obvious but is often overlooked or ignored because it can be inconvenient. Optimally, this means that the outcome measure should accurately reflect the phenomenon of interest, the model should include all relevant predictors, and the model should generalize to the cases to which it will be applied.

   For example, with regard to the outcome variable, a model of earnings will not necessarily tell you about patterns of total assets. A model of test scores will not necessarily tell you about child intelligence or cognitive development.

   Choosing inputs to a regression is often the most challenging step in the analysis. We are generally encouraged to include all "relevant" predictors, but in practice it can be difficult to determine which are necessary and how to interpret coefficients with large standard errors. Chapter 9 discusses the choice of inputs for regressions used in causal inference.

   A sample that is representative of all mothers and children may not be the most appropriate for making inferences about mothers and children who participate in the Temporary Assistance for Needy Families program. However, a carefully