Section 1.4. Data for Regression Analysis

Note. We need data to suggest which predictor variables are relevant and which functional form is best in a regression model. The data may come from non-experimental or experimental studies.

Note. Observational data comes from non-experimental studies. As a result, there is no control over the explanatory or predictor variable(s). For example, we might have the explanatory variable of 'age' and the predictor variable of something like 'income.' No experiment can be performed which *assigns* an age to someone! The limitation of observational data is that it may not indicate a causal relationship. It may be that other, more important, explanatory variables should be involved in regression model than those for which there is observational data.

Note. If it is possible to control a predictor variable, then experiments can be performed by varying this variable and measure the response variable. Information that results from such manipulations is called experimental data. When the explanatory variable is manipulated through random assignments (such as randomly assigning patients to a level of drug dosage in a test of medications), the data that results is considered better since the randomization tends to balance out the effects of any other variables that might affect the response variable (such variables are called *lurking variables*; see my online notes for Introduction to Probability and Statistics [MATH 1530] on Chapter 5. Regression). In the terminology of experimental design, for the example of assigning patients to certain drug dosages, the

dosage assigned is called a *treatment*, and the patients themselves are called the *experimental units*.

Note. A statistical experiment in which treatments are assigned to experimental units (or vice versa) at random is called a *completely randomized design*. Experimental units are assigned different treatments (or vice versa) with equal probabilities. A possible disadvantage of a completely randomized design occurs when the experimental units are heterogeneous and fall into a variety of categories; it could be desirable in this case to make sure that representatives of all categories are represented in each treatment category.

Revised: 7/16/2021