Section 1.2. Regression Models and Their Uses

Note. In this section we give a brief history of regression analysis (focusing on least-squares linear regression), state the postulates of a regression model, discuss choosing a regression model, and give some purposes of the use of regression models.

Note. Kutner et al. comment that "[r]egression analysis was first developed by Sir Francis Galton..." (see page 5). However, the story has a much richer history. Astronomers throughout the 1700s attempted to take observations of solar system objects (which would contain uncertainties in measurements) and use Newton's law of gravitation to predict the future location of the objects or the gravitational influence of previously unknown objects (asteroids or unknown planets of the time). One such success was the explanation of the motion of the moon by statistical methods by Tobias Mayer (February 17, 1723–February 20, 1762) in 1750. Such attempts were not always successful and the prolific Swiss mathematician Leonhard Euler (April 15, 1707–September 18, 1783) attempted but failed to find a statistical approach to explain the non-periodic motion in the orbits of Jupiter and Saturn. In 1755, Roger Joseph Boscovich (May 18, 1711–February 1787) published a work on the precise shape of the Earth (long known not to be a perfect sphere, but instead a prolate sphere). His technique of analyzing data is known as "Boscovish's Method" and it is the most famous predecessor to the method of least squares (Pierre-Simon Laplace [March 23, 1749–March 5, 1827] also considered this problem in the late 1780s). This work of the 1700s culminated in Adrien-Marie Legendre's (September 18, 1752–January 9, 1833) publication of *Nouvelles méthodes pour la*

détermination des orbites des comètes (New Methods for the determination of the Orbits of Comets) in 1805 (with supplements added to the work over the next 15 years). In this work, he published the first use of the method of least squares. In connection with least-squares, Legendre comments in his 1805 work: "We see, therefore, that the method of least squares reveals, in a manner of speaking, the center around which the results of observations arrange themselves, so that the deviations from that center are as small as possible."

The only known image of Legendre is this caricature by J-L Boilly form the MacTutor History of Mathematics Archive biography of Legendre webpage

Carl Friedrich Gauss (April 30, 1777–February 23, 1855) in 1809 claimed that he had been using the method of least squares since 1795, leading to a priority dispute with Legendre (this was not the only such dispute in which Gauss was involved; see my online presentation on Hyperbolic Geometry). The use of least squares was a standard tool in astronomy and geodesy in much of Europe by 1815 and in England by 1825. These historical comments are based on the Introduction and "Chapter 1. Least Squares and the Combination of Observations" of Stephen Stigler's *The*

History of Statistics: The Measurement of Uncertainty before 1900 (Belknap Press of Harvard University Press, 1986). Francis Galton (February 16, 1833–January 17, 1911) was largely involved in the application of statistical ideas to biological data. Based on earlier work of Adolphe Quetelet (February 22, 1796–February 17, 1874), he often applied regression ideas to fit normal distributions to biological data (such as an individuals height). He found that in a population, individuals tend to have quantitative traits that cluster ("revert" or "regress") near the population mean. He published *Hereditary Genius* in 1869 in which he argued that "talent runs in families" (as Stigler puts it; see his page 267). Stigler also comments that "as a statistical investigation it was naive and flawed, and Galton seems to have realized it" (see his page 268). Galton, along with Francis Edgeworth and Karl Pearson, cleared the way for statistical analysis of social science data. The Wikipedia page on Francis Galton declares him "a proponent of social Darwinism, eugenics and scientific racism."

Note/Definition. In a broad sense, we consider data based on two variables: a *predictor variable* X and a *response variable* Y. Variable Y has a tendency to vary with X in a systematic way. More precisely, we impose two postulates is a regression model:

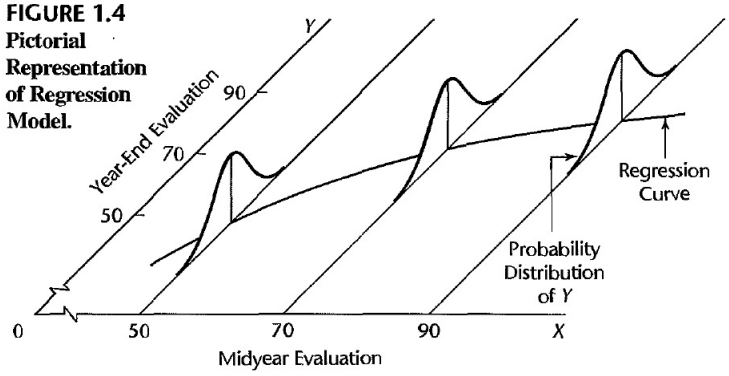
Postulate 1. There is a probability distribution of Y for each value of X.

Postulate 2. The means of these probability distributions vary in some systematic fashion with X (that is, the means are some *regression function* of X).

The graph of the regression function is the *regression curve*.

Note. Consider again the graph in Figure 1.2. We treat the year-end evaluation as response variable Y and the mid-year evaluation as predictor variable X. We then illustrate the idea behind a regression model in Figure 1.4 by thinking of a family of probability distributions (as drawn for $X = 50$, $X = 80$, and $X = 90$), one for each value of X, that when placed along the corresponding line of constant X-value has its mean on the regression curve. Each distribution looks symmetric in this case, but this need not hold in general.

FIGURE 1.4
Pictorial Representation of Regression Model.



Note. A regression model can be based on more that one predictor variable. Kutner et al. list three examples, the third of which involves a medical study of short children. The response variable was the peak plasma growth hormone level and it was based on 14 predictor variables, namely age, gender, height, weight, and 10 skinfold measurements. In the simpler case of only two predictor variables X_1 and X_2, the regression model will give a surface instead of a curve in the case of one predictor variable (as shown in Figure 1.4 above). More than two predictor variables require more than three dimensions to illustrate and so are hard to visualize (but easy to algebraically interpret).

Note. In choosing a particular regression model, one must choose (1) the predictor variables and (2) the type of function that will form the regression relation. Predictor variables should be chosen which are expected to play a causal role in the process. Ideally, these are chosen in a way as to reduce the variation in Y (so that more confidence can be put in the predictions of the model). Other concerns might include the expense of gathering particular date related to predictor variables, or the time it would take to collect such data. Such choices will be further explored in "Chapter 9. Building the Regression Model I: Model Selection and Validation" in Part Two. Multiple Linear Regression. The type of function (i.e., the "functional form") to be used will depend on the predictor variables chosen and, ideally, on some theoretical framework. However, in practice, there may not be any particular functional form indicated by previous research or models. In this event, existing data may empirically suggest the preference of some functional forms over others. One approach is to use linear and quadratic models piecewise to fit the data. One should avoid very complicated functional forms since all predictive value could be lost; for example, any collection of $n+1$ data points of the form (x_i, y_i) (where the x_i are distinct) are contained precisely on the graph of an $n-1$ degree polynomial (it is called the *Lagrange polynomial* containing all of the points); see my online notes for Numerical Analysis (MATH 4257/5257) on Section 3.1. Interpolation and the Lagrange Polynomial. Though this polynomial passes through each of the data points, it will likely oscillate wildly between the points and it is useless for predictions. Also of relevance is the *scope of the model* which is the restriction of the range of the predictor variables; heights and weights cannot be negative, for example.

Note. The three main (but not mutually exclusive) purposes of regression analysis are (1) description, (2) control, and (3) prediction. The idea in description is that the value of the response variable is given in terms of several predictor variables and the regression model "describes" the relationship. In control, the regression model is used to to make decisions concerning the response variable; for example, a central office may determine the optimal budget for its branch offices based on the values of a number of predictor variables (giving the central office a reasonable way to control budgets). In prediction, the goal is to use the measurement of several predictor variables and then use these to estimate the response variable; for example, the dosage of a medication might be predicted by age, gender, body mass index, level of activity, etc.

Note. As you are aware, "correlation does not mean causation." Kutner et al. state this on page 8 as:

> "The existence of a statistical relation between the response variable Y and the explanatory or predictor variable X does not imply in any way that Y depends on X. No matter how strong the statistical relation between S and Y, no cause-and-effect pattern is necessarily implied by the regression model."

In particular, we might have a strong linear relationship between X and Y, but not have the value of Y caused by X but instead have the opposite causal relationship. For example, we might consider some quantitative measure of a child's health X and that child's family income Y. To an extent, it is Y that determines X, not the other way around (that is, having a healthy child does not cause a high family income).