Chapter 1 Introduction and Motivation

A Brief History

The word regression generally conjures up thoughts of finding a mean prediction function for a "dependent" variable in a set of bivariate (or multivariate) paired data. The technique has come under much scrutiny in the last 114 years. In its origin, the term was used by Francis Galton in 1885 to describe the rather obvious tendency of male offspring to achieve heights closer to the mean than that of their fathers. The regression (as we know it) was used to illustrate that fact. The discovery of the allied least squares method for regression is generally credited to Carl Friedrich Gauss (although perhaps Adrien Legendre published the first work on least squares in 1805). This method and regression were somewhat inseparable until the late 1960's when it became apparent that least squares is not always appropriate. Two very important (and not unrelated) ideas came about in this era.

The idea of weighting (placing greater priority on certain paired observations) became popular through the work of Huber (1964), and Bement and Williams (1969). This technique was used primarily to combat violations of the constant variance assumption. Not surprisingly, the 1960's also ushered in the study of nonparametric regression. Although some would argue that the technique had been used by scientists for years prior to 1964, it was then that Nadaraya (1964) and Watson (1964) are generally credited with formalizing the idea of isolating a function having no particular a priori form. Up to this time it was apropos to find a usual (or parametric) solution to the mean function, meaning the goal was, more specifically, to seek the values of parameters that, although unknown, were part of a known functional form. This new type of regression was predicated upon estimating values of the mean function locally, that is, putting greater weight on values of the dependent variable in closer proximity in the corresponding independent variable(s). This was, in fact, akin to performing a locally weighted constant regression.

If this method was somewhat successful, statisticians queried, how much more success would a locally weighted polynomial produce? Nonparametric regression and its organic weighting scheme would, in turn, produce two hotly debated topics over the next three decades, the “bandwidth” (the area of local weighting) (how large should it be? … how should it be chosen?), and the degree of the local polynomial.

While these issues were being contested, the older form, parametric regression, remained a popular choice. In 1987 and 1993, Einsporn and Birch proposed and studied a new form of regression that combined parametric and nonparametric techniques via a convex mixing parameter $l \in (0,1)$. Originally called the “Hatlink” estimate, the name was later revised to model robust regression (MRR). At about the same time Olkin and Spiegelman (1987) proposed a similar combination for density estimation and gave asymptotic results. Speckman (1988) developed a hybrid technique entitled partial linear regression (PLR). Burman and Chaudhuri (1992) presented a estimate similar to Einsporn and Birch’s MRR, and gave asymptotic results based in part on the Olkin and Spiegelman (1987) paper. Mays (1995) developed a new form of model robust regression (similar in form but not in approach to the Wooldridge (1992) estimate) which always involved the parametrically estimated mean function and utilized the nonparametric estimate to smooth patterns arising from the residuals. Mays called this estimate model robust regression 2 (MRR2) and relegated the name model robust regression 1 (MRR1) to the Einsporn and Birch estimate of several years earlier. This new hybrid (or “semiparametric”) estimate often yielded better results than those of the PLR and MRR1 estimates in small samples. Fan and Ullah (1994) and Rahman, Ghokale and Ullah (1997) published related work and gave asymptotic distribution results for various mixing parameter estimators for the MRR1 estimate, and the Wooldridge (1992) estimate under the null hypothesis that the parametric model is correct.

Advantages and Disadvantages

The nonparametric approach allows the user freedom from functional form, and the ability to capture trends in the data that would otherwise be lost in the parametric approach. It is generally associated with low bias and high variance (in terms of the mean response estimates).
As such, it may be used when the user has little or no knowledge of the true mean function, and wishes to find a possible functional form or forms to be used in subsequent analyses. Its drawbacks are twofold. First, varying bandwidths, specified by the user, may produce very different results. Consequently, a user is subjected to a plethora of bandwidth selection methods (which are discussed later). Second, the nonparametric estimate may easily “capture” a functional form that is not really there, that is, it may attribute to the function that which is actually the stochastic process of the error term.

The parametric approach is based to a large extent on the user’s prior knowledge of the functional form. Assuming that knowledge is correct, the parametric approach provides a simple and quick solution to the regression problem. In addition, most data sets can be modeled well by some parametric function. This approach is generally associated with low variance and higher bias (in terms of the mean response estimates). Its drawbacks are that it may not capture the entire functional form, and that the wrong functional form may be chosen a priori.

The purpose of combining parametric and nonparametric methods is to allow the user to benefit from the positive aspects of each, while hoping to minimize the negative aspects of each.

Preview

In the remaining chapters we wish to investigate the asymptotic properties of various semiparametric estimates. In chapter 2, we give a brief technical overview of the two types of regression as well as a review of model robust regression. In chapter 3, we investigate various forms of the MRR1 estimate asymptotically. In chapter 4, we investigate various forms of the MRR2 estimate asymptotically. Parts 3a and 4a are particularly important, since these sections establish the basic properties of the respective estimates at their asymptotic optima. In chapter 5, we investigate the MRR1 estimate asymptotically as it is employed in quantal regression. There we develop asymptotic results for a specific form of nonparametric logistic regression, as well as a general form of parametric regression (of which forms of parametric quantal regression are special cases), and then provide asymptotic results for the MRR1 mixture of these estimates. In chapter 6, we discuss future work in asymptotic model robust regression.