21.1 Introduction

As discussed in Chapter 12, regression analysis estimates the conditional expectation of a response given predictor variables. The conditional expectation is called the regression function and is the best predictor of the response based upon the predictor variables, because it minimizes the expected squared prediction error.

There are three types of regression, linear, nonlinear parametric, and nonparametric. Linear regression assumes that the regression function is a linear function of the parameters and estimates the intercept and slopes (regression coefficients). Nonlinear parametric regression, which was discussed in Section 14.3, does not assume linearity but does assume that the regression function is of a known parametric form, for example, an exponential function. In this chapter, we study nonparametric regression, where the form of the regression function is also nonlinear but, unlike nonlinear parametric regression, not specified by a model but rather determined from the data. Nonparametric regression is used when we know, or suspect, that the regression function is curved, but we do not have a model for the curve.

There are many techniques for nonparametric regression, but local polynomial regression and splines are the most widely used, and only these will be discussed here. Local polynomial regression and splines generally work well and, since they usually give similar estimates, it is difficult to recommend one over the other. Local polynomial estimation might be somewhat simpler to understand. Splines are used in many areas of mathematics, such as, for interpolation, and so it is worthwhile to be familiar with them. Also, splines are useful as components in complex models. The R lab at the end of this chapter gives an example.

Models for the evolution of short-term interest rates are important in finance, for example, because they are needed for the pricing of interest rate derivatives. Figure 21.1 contains plots of the monthly risk-free returns in the Capm data set in R’s Ecdat package. This data set has been used for various
Fig. 21.1. Risk-free monthly returns. The returns are 1/12th the yearly rate. (a) Time series plot of the returns. (b) Time series plot of the changes in the returns. (c) Plot of changes in returns against lagged returns and a local linear estimate of the drift. (d) Plot of squared residuals against lagged returns and a local linear estimate of the squared diffusion coefficient.

purposes in several previous chapters. Here we will use it to illustrate nonparametric regression. Panels (a) and (b) are time series plots of the returns and the changes in the returns.

A common model for changes in short-term interest rates is

\[ \Delta r_t = \mu(r_{t-1}) + \sigma(r_{t-1}) \epsilon_t, \]

(21.1)

where \( \Delta r_t = r_t - r_{t-1} \), \( \mu(\cdot) \) is the drift function, \( \sigma(\cdot) \) is the volatility function, also called the diffusion function, and \( \epsilon_t \) is \( N(0,1) \) noise. Many different parametric models have been proposed for \( \mu(\cdot) \) and \( \sigma(\cdot) \), for example, by Merton (1973), Vasicek (1977), Cox, Ingersoll, and Ross (1985), Yau and Kohn (2003), and Chan et al. (1992). The simplest model, due to Merton (1973), is that \( \mu(\cdot) \) and \( \sigma(\cdot) \) are constant. Chan et al. (1992) assume that \( \mu(r) = \beta(r - \alpha) \) and \( \sigma(r) = \theta r^\gamma \), where \( \alpha > 0, \beta < 0, \theta > 0, \) and \( \gamma \) are unknown parameters—this process reverts to a mean equal to \( \alpha \). Chan et al.’s model was used as an example of nonlinear regression in Section 14.14.2. The approach of Yau and Kohn (2001) that is used here is to model both \( \mu(\cdot) \) and \( \sigma(\cdot) \) nonparametrically. Doing this allows one to check which parametric models, if any, fit