Semiparametric Regression

DAVID RUPPERT
Cornell University

M. P. WAND
Harvard University

R. J. CARROLL
Texas A&M University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United States of America

Typeface Times 10.5/13 pt. *System* AMS-TEX [FH]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication data
Ruppert, David, 1948–
Semiparametric regression / David Ruppert, M.P. Wand, R.J. Carroll.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-78050-0 – ISBN 0-521-78516-2 (pb.)
1. Regression analysis. 2. Nonparametric statistics. I. Wand, M. P. (Matthew P.).
II. Carroll, Raymond J. III. Title.

QA278.2.R87 2003
519.5′36 – dc21 2002041460

ISBN 0 521 78050 0 hardback
ISBN 0 521 78516 2 paperback

Contents

1

Introduction

Semiparametric regression can be of substantial value in the solution of complex scientific problems. The real world is far too complicated for the human mind to comprehend in great detail. Semiparametric regression models reduce complex data sets to summaries that we can understand. Properly applied, they retain essential features of the data while discarding unimportant details, and hence they aid sound decision-making.

Figure 1.1 depicts a complex data set corresponding to a cancer study in the Upper Cape Cod region of Massachusetts. Apart from the geographical location of cancer occurrences, there are data on age and smoking status. These data are for females.

One question of interest is whether there are elevated lung cancer rates, relative to all cancers and after adjustment for confounders, in any particular geographical locations. There is clearly a lot of relevant information represented by the one thousand points in this plot. However, it is very difficult to draw any conclusions from this alone. A semiparametric regression analysis leads to Figure 1.2.

Each of the graphics in Figure 1.2 displays an easy-to-comprehend estimate of the effect of smoking status, age, and geographical location on the occurrence of

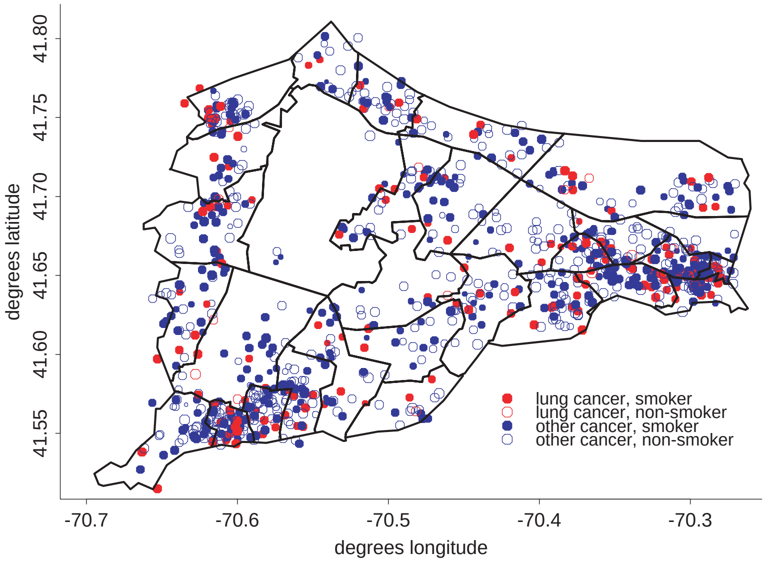


Figure 1.1 One thousand randomly chosen occurrences of female cancer in Upper Cape Cod, Massachusetts, for the period 1986–1994. The data are categorized according to lung cancer (red) or other (blue) and smoker (closed circle) or nonsmoker (open circle). The size of the circle is proportional to age. For confidentiality reasons, the data have been jittered.

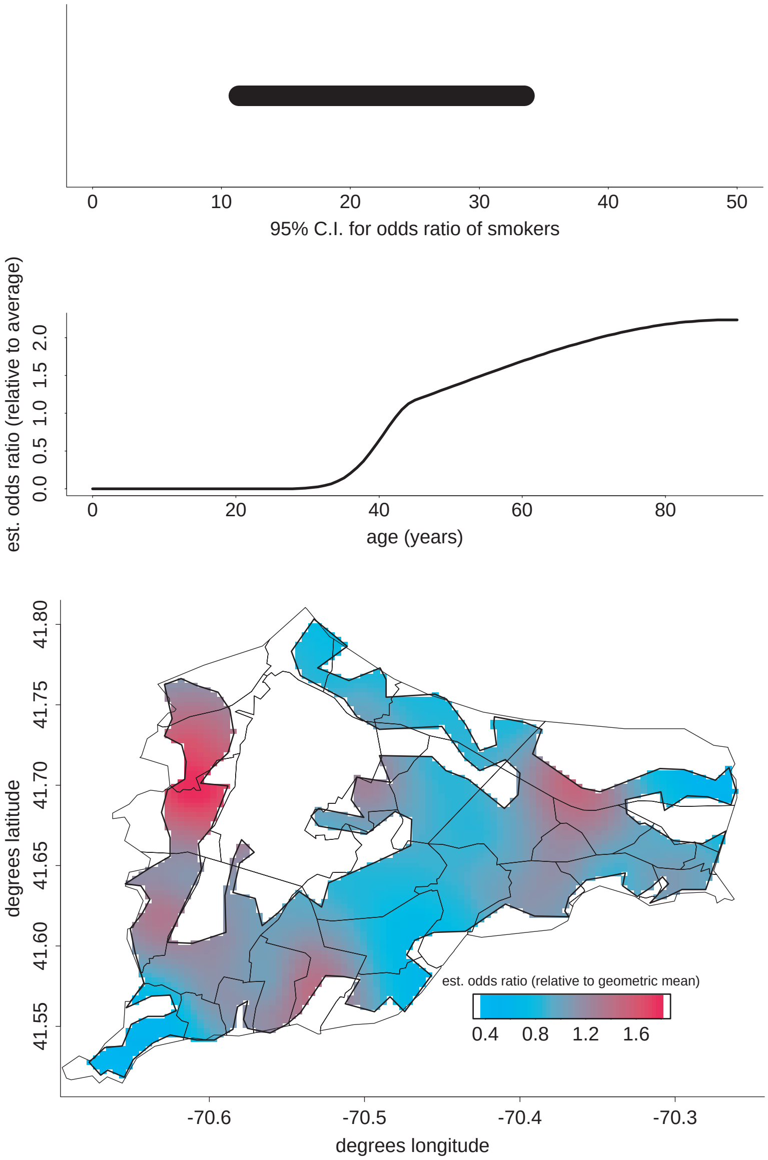


Figure 1.2 Graphical outcomes from a semiparametric regression analysis of Upper Cape Cod lung cancer data: top panel, point estimate and approximate 95% confidence interval for the odds ratio of lung cancer among smokers who have some type of cancer; middle panel, estimated odds ratio as function of age; bottom panel, estimated odds ratio as function of geographic location. Higher values correspond to high estimated probabilities of lung cancer, given cancer, measured through the odds ratio.

lung cancer, relative to cancer, while controlling for each of the other two variables. Smoking status is a binary variable, so its effect can be modeled through a single parameter. This the simplest type of parametric modeling. The graphic shows an odds ratio estimated to be in the range 11 to 33. Age is a continuous variable and, in this instance, its effect can be modeled reasonably well using parametric regression techniques. However, the nonparametric estimate shown in the middle panel suggests an unusual type of nonlinearity and so nonparametric regression techniques may lead to an improved fit. The effect of geography is difficult to model using traditional parametric models, and the map in Figure 1.2 is the result of a bivariate nonparametric regression technique. It clearly shows

The *odds ratio* of an event A, relative to an event B, is defined to be the ratio of the odds of A to the odds of B. The *odds of* A is the probability of A occurring divided by the probability of A not occurring.

Table 1.1 Observed mammary tumor rates with phenolphthalein. For example, 32 of the 50 animals exposed at 25,000 ppm had tumors at the time of death. Of these, 18 died during the experiment and 32 were sacrificed at the end of the experiment, with 15 of the sacrificed animals being among the 17 with tumors.

0 ppm			25,000 ppm		
Tumor rates		Mean body weight[b]	Tumor rates		Mean body weight
Overall	Terminal[a]		Overall	Terminal	
32/50	25/30	287	17/50	15/32	254

[a] Tumors found at terminal sacrifice time.
[b] Average body weight at 12 months.

regions with elevated lung cancer levels, something that is not easy to discern in Figure 1.1. Since the effects of smoking, age, and location have been modeled using a combination of parametric and nonparametric regression techniques, we call this a *semiparametric regression* analysis.

In the next sections we look at other important scientific investigations where semiparametric regression can play a useful role. We give detailed analyses of these studies (or at least references to where careful analyses can be found) in Chapter 18, after we have developed methodology to tackle them; Chapters 2–17 will be spent describing this methodology.

1.1 Assessing the Carcinogenicity of Phenolphthalein

The U.S. National Toxicology Program (NTP) routinely conducts animal experiments to measure the toxicity of certain foods and drugs. One such example is the assessment of the possible carcinogenicity of *phenolphthalein*, an ingredient of over-the-counter laxatives that was recently withdrawn by the U.S. Food and Drug Administration.

A topic of recent interest in the analysis of carcinogenicity data is how to deal with body weight. A recent editorial in *Science* magazine was highly critical of risk assessment agencies for not controlling for the possible confounding effect of weight, since weight loss caused by a toxic substance might protect against cancer and mask a carcinogenic effect (Abelson 1995). It is not uncommon for control animals to weigh substantially more than the treated animals throughout the course of an experiment owing to toxic effects of the chemical. Several sources have reported a lower incidence of tumors corresponding to lower body weights (Hart et al. 1995; Haseman, Bourbina, and Eustis 1994; Seilkop 1995). Thus, dose-related differences in body weights could affect the conclusions drawn from these studies. Indeed, many studies conducted by the NTP have shown protective effects of the chemical being tested on certain tumor incidences. These apparent reductions in tumor incidence across dose may be due to differences in body weight (Hart et al. 1995). This phenomenon is illustrated in Table 1.1, taken from the NTP study in phenolphthalein.

Figure 1.3 shows nonparametric estimates of the probabilities of four carcinogenic outcomes as a function of weight based on a large NTP set of data on

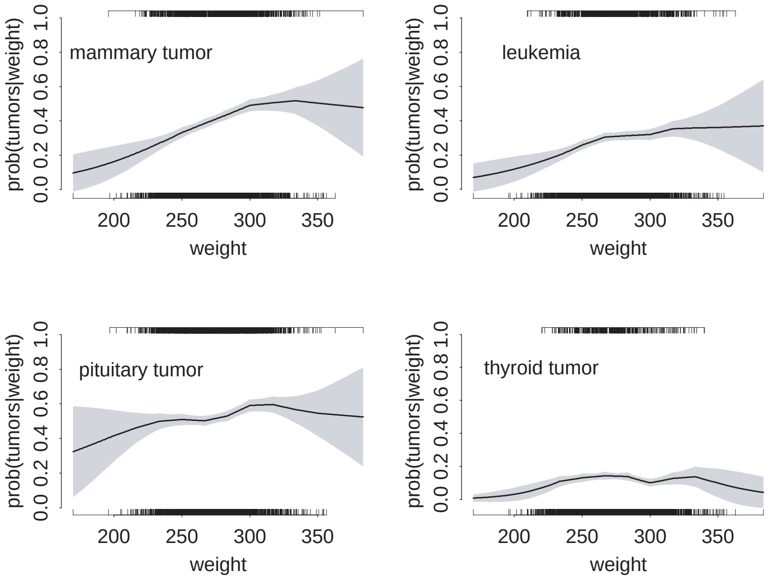


Figure 1.3 Estimated probability of mammary tumor, leukemia, pituitary tumor, and thyroid tumor as a function of weight for a set of NTP historical controls. The shaded region represents plus and minus twice the estimated (pointwise) standard error.

controls. It is apparent from these plots that nonlinear relationships exist and that semiparametric models for incorporation of weight data would be beneficial.

1.2 Salinity and Fishing in North Carolina

This example comes from a larger project to predict the annual shrimp (or prawn) harvest in Pamlico Sound, North Carolina, where shrimping occurs in the summer and autumn. It was believed that low salinity in the sound was detrimental to the shrimp harvest and that salinity values during certain crucial springtime periods would be useful predictors.

Salinity values were not measured regularly during the years prior to the project. However, discharges from rivers that empty into Pamlico Sound were known. The goal of the project was to develop a prediction model that could be used during the spring, early enough to help the fishing industry decide whether to rig for shrimp or instead to harvest some other species such as bluefish.

The data set has 28 cases taken from the spring periods of years 1972 to 1977. In each case, salinity was measured at the current time period and two weeks earlier, giving the variables `salinity` and `lagged.sal`. Two other variables were measured, `discharge` and `trend`. The variable `trend` indicated which of six biweekly periods during March to May a case came from. It was felt that `trend` might model the effects of increasing evaporation as the weather warmed, but no effect of `trend` was detected and so that variable will be ignored.

Figure 1.4 is a scatterplot matrix of the salinity data. One can see the strong, seemingly linear, relationship between `salinity` and `lagged.sal`. The relationship between `salinity` and `discharge` is somewhat weaker and possibly nonlinear. There is not a strong relationship between `lagged.sal` and

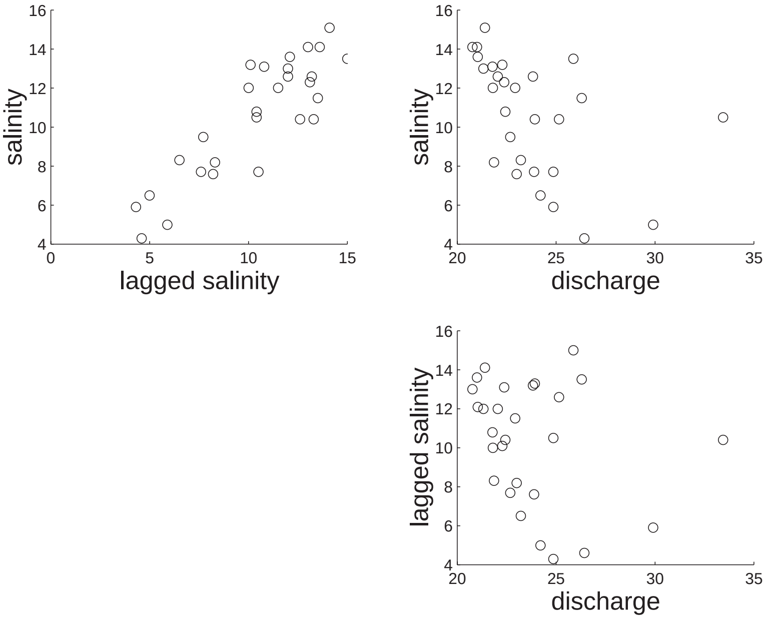


Figure 1.4 Scatterplot matrix of the salinity data.

`discharge`, so their effects upon `salinity` should be individually estimable with good precision.

The relationship between `salinity` and `discharge` is easier to see if we remove the effects of `lagged.sal`. To do this, we regressed `salinity` on `lagged.sal` using a straight line model (see Section 2.2). The residuals (i.e., the differences between `salinity` and the predicted values) are plotted against `discharge` in Figure 1.5. The nonlinearity is now more evident, especially because a *scatterplot smooth* has been added. This suggests that a semiparametric regression approach will be beneficial. The observation with `discharge` equal to nearly 34 is a “high leverage point,” meaning that it has a potentially high influence on the fitted curve. In fact, the fitted curve bends upward in the figure but would not do so if the leverage point were excluded. However, unlike a linear fit, the curved fit is only influenced locally – that is, on the right. We will discuss this point further when we return to this example in Chapter 18.

The notion of smoothing a scatterplot will be described extensively in Chapters 3 and 5.

1.3 Management of a Retirement Fund

Bryant and Smith (1995) describe a managerial problem based on a real data set, but with names changed to protect confidentiality. It concerns a company, Best Retirement Inc. (BRI), that sells retirement plans to corporations around the United States. To capture a market niche, it has decided to target smaller firms: those with 500 or fewer employees. The major portion of their revenue comes from retirement packages.

For a particular type of retirement plan known as 401(k), data are available on several attributes of the firms from the previous year. It is advantageous that

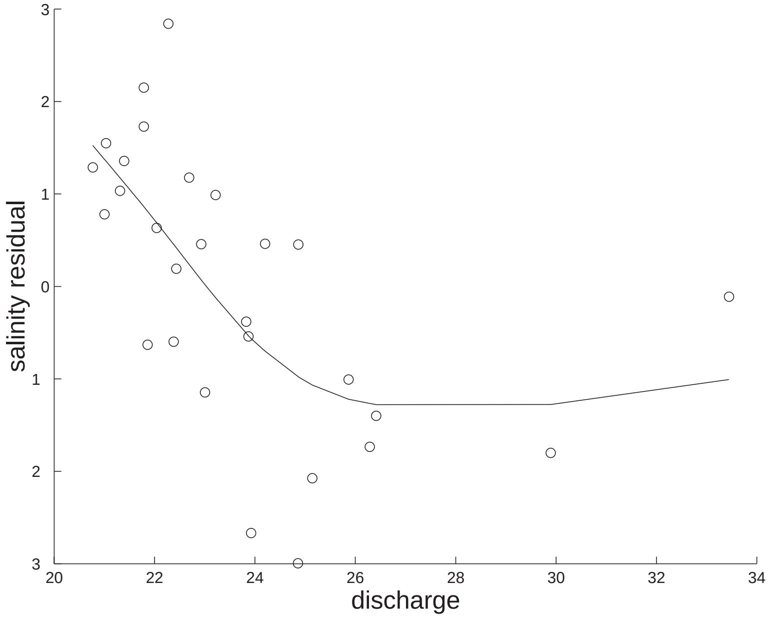


Figure 1.5 Scatterplot of residuals from the regression of `salinity` on `lagged.sal`. A scatterplot smooth has been added. Note the effect of the high leverage point on the extreme right.

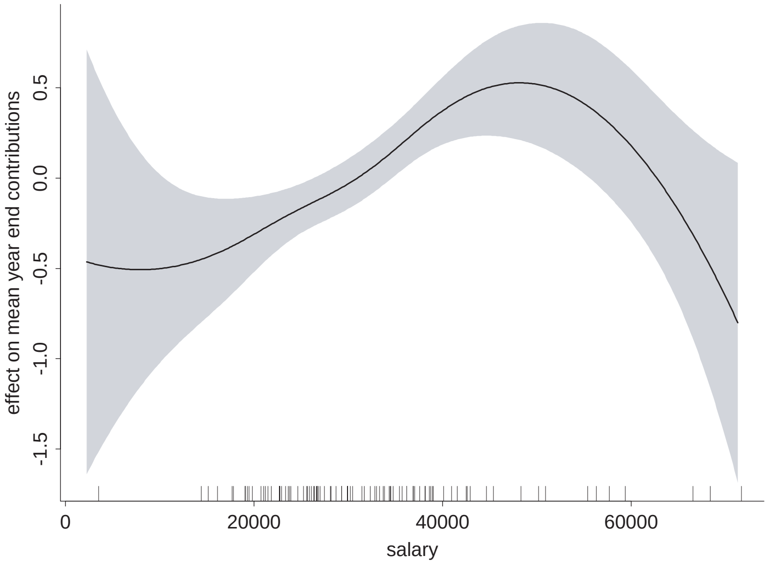


Figure 1.6 Estimated effect of `salary` on contribution to the logarithm of year-end contributions in a semiparametric regression analysis. The shaded region represents plus and minus twice the estimated (pointwise) standard error.

BRI be able to estimate the year-end dollar amount contributed to each plan in advance so that it can make internal revenue and cost projections.

Apart from building a prediction model for year-end contributions, there are some other managerial questions that can be addressed using these data. For example, BRI has a sales representative who has been specifically trained to deal exclusively with 401(k) retirement plans. The company would like to know if her expertise is a factor that influences contributions to such retirement plans.

Figure 1.6 shows the effect of `salary` (average salary of each firm) on the logarithm of year-end contributions as estimated by a semiparametric regression

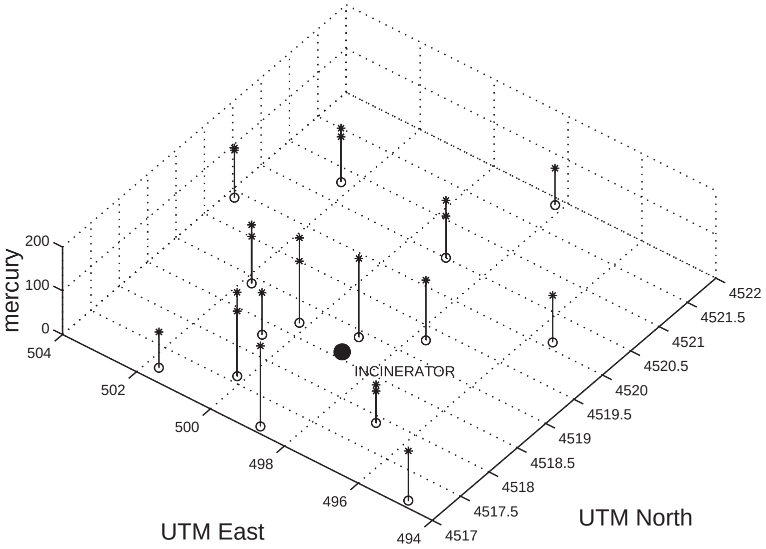


Figure 1.7 Plot of biomonitoring data. Open circles show sampling locations, and asterisks mark the single or replicate values of mercury measured at each sampling location. The large solid circle marks the location of the incinerator.

analysis. There is a pronounced nonlinearity here, which suggests that better predictions and managerial decisions can be realized through the use of semiparametric regression.

1.4 Biomonitoring of Airborne Mercury

Waste incineration is a major source of environmental mercury. As part of an environmental monitoring program in Warren County, New Jersey, pots of sphaghum moss were placed at 15 sampling locations about a solid waste incinerator and exposed to ambient conditions between July 9 and July 23, 1991. The moss was then collected, dried, and assayed for mercury. The resultant data are shown in Figure 1.7.

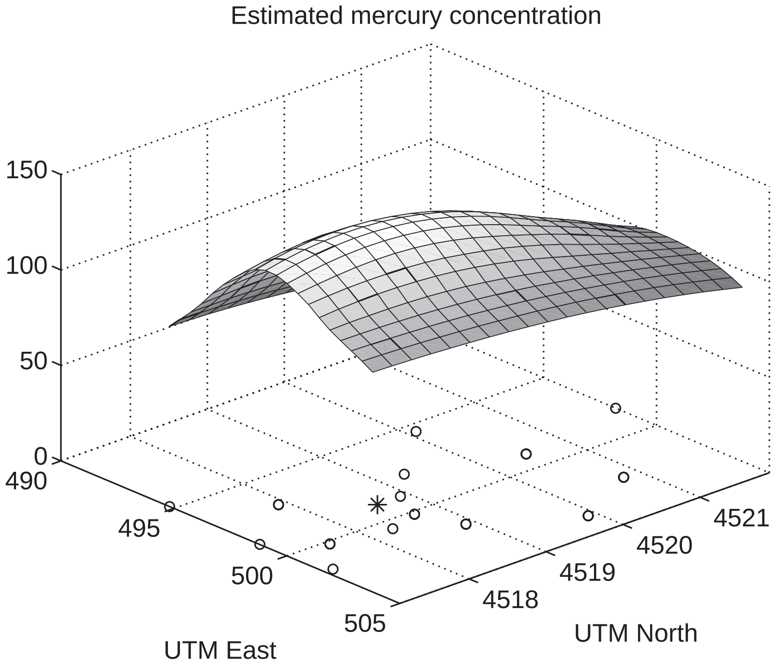
The goals of the study include estimating the distribution of mercury about the incinerator and testing the null hypothesis that the mean mercury concentration is constant.

Figure 1.8 shows estimated levels of mercury concentration that were obtained using nonparametric methods described in this book. The plot indicates that mercury concentration peaks north of the incinerator. There are only 15 sampling locations, with replicate moss pots at 7 of these sites, for a total of 22 observations. With so few data, only gross features of mercury deposition can be resolved, but the nonparametric fit provides a pleasing image of these features.

1.5 Term Structure of Interest Rates

Corporations, municipalities, the U.S. Treasury, and other entities raise money by issuing bonds. The purchase price of the bond is a loan to the issuing entity and the bond is a contract requiring that entity to pay to the bond holder both principal

Figure 1.8 Plot of biomonitoring data with coloring of estimated mercury concentration. There were 15 sampling locations and 7 had replicate samples. Open circles indicate sampling locations; the asterisk marks the incinerator location.



and interest according to a schedule. At the time of expiration of the bond, which is called the *maturity*, the bond holder receives a payment call the *par value*. There are two general classes of bonds, coupon bonds and zero-coupon bonds. At fixed periods, often every six months, the holder of a coupon bond receives a *coupon payment*. Generally, coupon bonds sell at a price near their par value. The par payment at maturity is a repayment of principal while the coupon payments are interest. Zero-coupon bonds have no coupon payments and sell below par. The par payment at maturity represents principal and interest.

Frequently, the initial owner of the bond will sell the bond to another investor. The current price at which bonds trade depends upon the current interest rates. For example, suppose a corporate coupon bond with a 5% coupon rate is issued with the initial price equal to par, so that the coupon payments are 5% of the initial price. If the prevailing interest rate increases to 6% then the price of the bond will drop, so that a new purchaser of the bond will in effect receive a 6% rate.

The interest rates on bonds depend upon their maturities, with long-term bonds frequently (though not always) paying higher rates than short-term bonds. For example, on January 26, 2001, the rate on a 1-year Treasury bill was 4.83% whereas the rate on a 30-year Treasury bond was 6.11%. The term structure of interest rates is a quantitative description of the dependency of rate upon maturity. The estimation of term structure is essential for financial analysts working, for example, with credit derivatives.

Interest rates not only depend upon the maturity, but for any fixed maturity, the interest rate on bonds with that maturity will change over time. In this case study, we are not concerned with such changes. Rather, we will only be concerned with

A financial derivative is a security whose value depends on the value of other *underlying* securities. As an example of a derivative, consider a call option on a stock. A call option gives the owner the right, but not the obligation, to purchase a share of stock at a fixed price on a given date, called the expiration date. The value of the call option depends on the price of the underlying stock and on such how interest rates on a given day depend on maturity. Specifically, in our example, we will model bond interest rates on December 31, 1995.

We will work with continuously compounded interest rates. As an illustration, we will start with an unrealistic assumption that the interest rate is constant, that is, not dependent on maturity. If a bond is worth $P(t)$ dollars at time t and is continuously compounded at a constant rate r, then $P(t)$ satisfies the simple differential equation

$$P'(t) = rP(t) \tag{1.1}$$

and so, at maturity T,

$$P(T) = P(0)\exp(rT). \tag{1.2}$$

The rate r is called the *forward rate.* It is the rate agreed upon at present for interest in the future, that is, *forward* in time.

Interest rates must be inferred from bond prices. Recall that the bond's value at maturity, $P(T)$, is called the par value. Hence, from (1.2) we have

$$P(0) = \texttt{par}\exp(-rT), \tag{1.3}$$

where `par` is the par value. Suppose a 1-year, par \$100 zero-coupon bond is selling now for \$92. This means we can buy the bond now for \$92 and receive \$100 exactly one year from now. Recall that zero-coupon means the bond holder receives no interest payments until maturity. The \$8 difference between the present price and the par value is the only interest payment. Here we have $T = 1$, $P(1) = \texttt{par} = 100$, $P(0) = 92$, and, from (1.2),

$$92 = 100\exp(-r)$$

or

$$r = \log(100/92) = 0.0834.$$

Thus, the annual continuously compounded interest rate over the next year is 8.34%.

Suppose, in addition, that a 2-year, par \$100 zero-coupon bond sells for \$85. We assume that this bond pays the just-determined rate of 8.34% the first year but a different interest rate the next year. The rate for the second year, call it r_2, solves

$$83 = 100\exp\{-(0.0834 + r_2)\}$$

or

$$r_2 = \log(100/83) - 0.0834 = 0.1029.$$

Table 1.2 gives the prices on December 31, 1995, of five bonds previously issued by the U.S. communications company AT&T and maturing at some time after that date. These are the prices at which the bonds were traded – that is, purchased by one investor from another. Each bond price is expressed as a percentage of par, the amount AT&T will pay the bond owner at maturity. The maturity is given in years from December 31, 1995. The bonds make semiannual interest payments called coupons. The time in years of the next coupon and the coupon payments are given in the table. The aim is to determine the forward rate of AT&T bonds from these data.

other variables as the time left until expiration. An example of a interest rate derivative is a *cap.* If an interest rate exceeds the cap, then the owner of the cap is paid the difference between the interest rate and the cap. Clearly, the value of the cap depends on the underlying interest rate. A company paying interest at a floating rate might purchase a cap as insurance against rate increases.

Table 1.2 AT&T bond prices on December 31, 1995. Issue, maturity, and next coupon dates are in years from December 31, 1995.

Issue	Maturity	Next coupon	Coupon	Price
−3.9644	5.9781	0.0356	7.1250	109.4580
−1.7726	8.1890	0.2274	6.7500	106.2840
−1.5836	10.3562	0.4164	7.5000	111.4360
−0.8384	11.1041	0.1616	7.7500	115.5090
−0.6384	9.3096	0.3616	7.0000	107.6590

We have been assuming that the forward interest rate is constant over each year. Clearly, this is an oversimplification. Financial analysts model the forward interest rate as a continuous function of time, $r(t)$. If $P(T)$ is the par value of a zero-coupon bond maturing at time T and if $P(0)$ is the current price of the bond, then (1.1) is replaced by

$$P'(t) = r(t)P(t),$$

with solution

$$P(0) = P(T)\exp\left(-\int_0^T r(x)\,dx\right). \tag{1.4}$$

The problem is to estimate $r(t)$ from bond prices, such as those shown in Table 1.2. A further complication is that many bonds, including those in the table, have coupons. A coupon bond can be modeled as a bundle of zero-coupon bonds, one for each coupon payment and one for the final payment at maturity of the par value. The bond price is the aggregate price of all of these coupon bonds. Bond prices such as in Table 1.2 have some random "error" since, for example, they are really prices at last transaction, not exactly at the current time. Therefore, the estimation of the forward rate curve is a statistical problem. Fisher, Nychka, and Zervos (1994) have developed a very elegant spline method for estimating the forward rate curve. Their method works well for Treasury bond data because there are enough Treasury bonds to estimate a continuous forward rate.

A forward price is a price negotiated at the present for the future delivery of some commodity. A forward interest rate means an interest rate that is agreed upon now for a loan in the future.

For corporate bonds, there is often a paucity of data and so the method of Fisher and colleagues cannot be applied directly. Jarrow, Ruppert, and Yu (2001) extend the model of Fisher et al. by assuming that the forward rate for a corporation such as AT&T differs from the Treasury forward rate by a constant or, perhaps, by a low-degree polynomial function of time. The corporate forward rate is greater than the Treasury rate, since Treasury bonds have no risk of default; the U.S. Treasury can always raise money by taxation. The difference between the two rates is called the *risk premium* or *spread* and reflects the extra interest that investors demand when buying corporate bonds (which may default) rather than risk-free Treasury bonds. The model of Jarrow and colleagues is semiparametric in that the Treasury forward rate is modeled as a spline, but the risk premium is modeled parametrically. This case study is typical of semiparametric models in that parts of the model for which there is much data are modeled nonparametrically while parts that are not well supported by data are modeled parametrically.

Figure 1.9 shows the prices of U.S. STRIPS (Separate Trading of Registered Interest and Principal of Securities), a type of zero-coupon Treasury bond. The

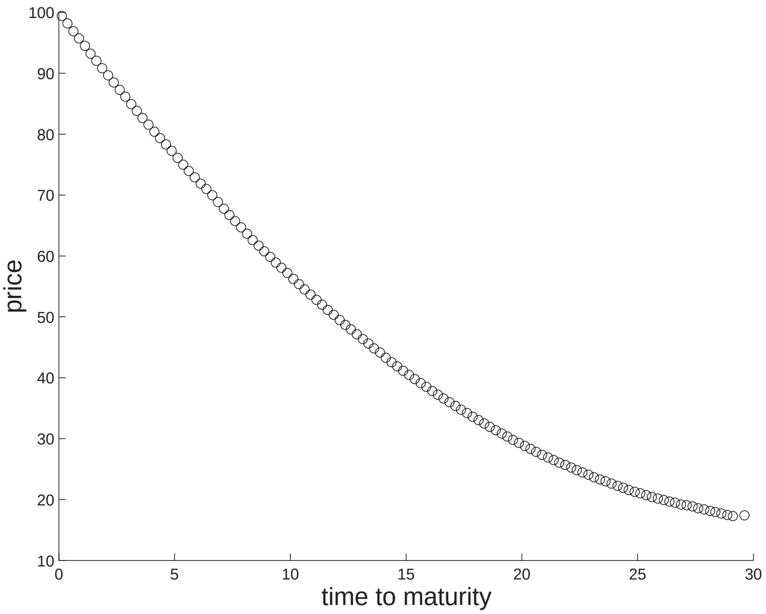


Figure 1.9 U.S. STRIPS prices as a percentage of the par value.

prices are expressed as a percentage of the par value and are plotted against time to maturity. If $r(x)$ is constant, say $r(x) = r_0$ for all x, then by (1.4) we have

$$y_i = 100 \exp(-r_0 T_i) \tag{1.5}$$

and

$$\log(y_i) = \log(100) - r_0 T_i. \tag{1.6}$$

Here $P(T_i)$ is the `par`, $P(0)$ is the present price, $y_i = 100P(0)/P(T_i)$ is the "response," and T_i is the maturity for the the ith U.S. STRIPS.

The rough exponential shape in Figure 1.9 suggests that model (1.5) is at least approximately correct. However, in Figure 1.10 we see $\log(y_i)$ plotted against T_i, and the plot is not quite the straight line that (1.6) suggests. In fact, we fit a straight line to $\{T_i, \log(y_i)\}_{i=1}^n$ and plotted the "residuals," which are the differences between the $\log(y_i)$ and the fitted line. This plot, shown as Figure 1.11, shows an obvious deviation from the random cloud that we would expect if the model (1.5) fit the data, thus indicating the need for a nonparametric model. The fitting of straight line models and residual analysis will be discussed in Chapter 2.

1.6 Air Pollution and Mortality in Milan: The Harvesting Effect

In the last decade, a good deal of literature has been published concerning the short-term effect of air pollution on health. Daily mortality counts and hospital admissions have been associated with daily air pollution levels, correcting for several time-dependent confounders. From the public health point of view, the significance of air pollution's short-term effects corresponds to an increase in mortality or morbidity among individuals who would otherwise die much later, not among those who could have died within a few days.

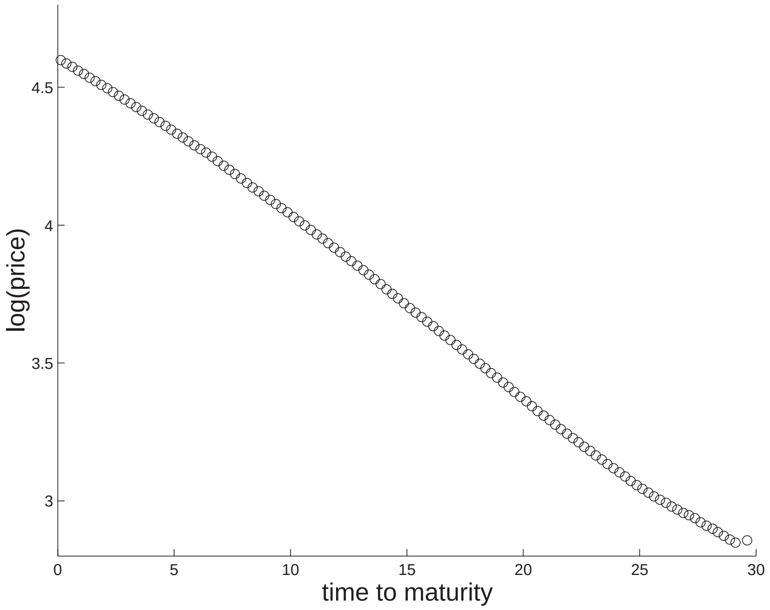


Figure 1.10 Logarithms of U.S. STRIPS prices as a percentage of the par value.

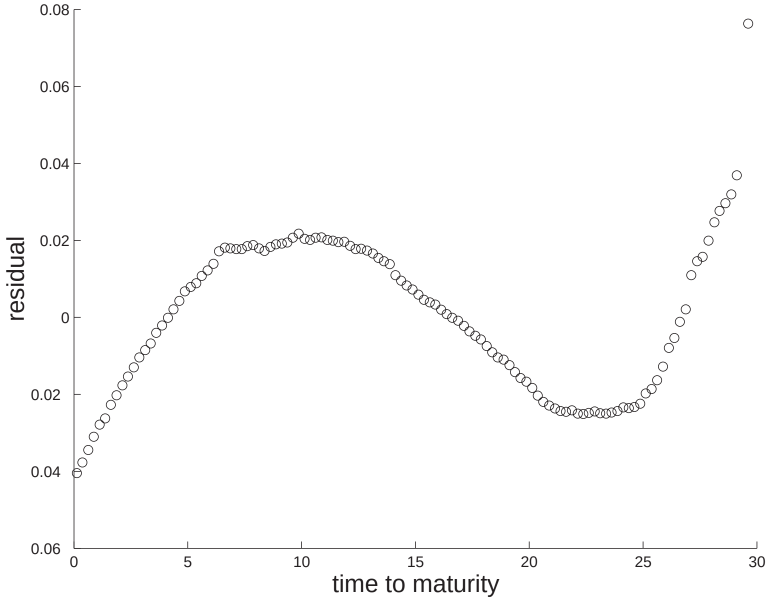


Figure 1.11 Residuals from a straight line fit to the logarithms of the U.S. STRIPS prices as a percentage of the par value.

Figure 1.12 is a schematic representation of the dynamics that arise when air pollution has an impact on mortality. The *risk pool* consists of sick and elderly people. Transitions between this state and the general population are affected by air pollution levels.

Consider the following lagged regression model of air pollution and generic mortality:

$$\log\{\mathsf{E}(\texttt{mortality}_t)\} = \alpha + \beta_0\texttt{pollution}_t + \cdots + \beta_q\texttt{pollution}_{t-q} + \varepsilon_t,$$

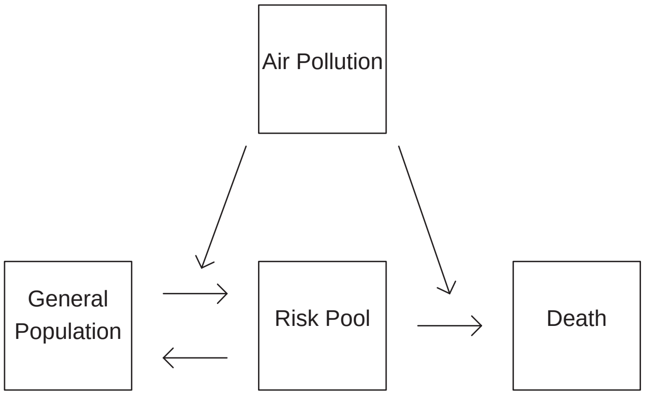


Figure 1.12 Schematic representation of the dynamics that arise when air pollution has an impact on mortality.

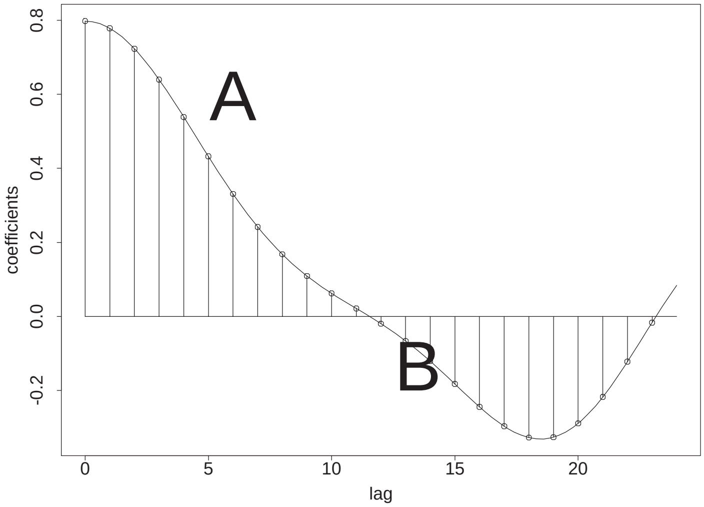


Figure 1.13 Lag structure corresponding to the harvesting effect.

where $\texttt{mortality}_t$ and $\texttt{pollution}_t$ are (respectively) the mortality count and pollution level for day t. The lag structure in Figure 1.13 describes the so-called *harvesting effect*. The horizontal axis is the lag number and the vertical axis shows the coefficients β_ℓ. Each β_ℓ has this interpretation: net effect of pollution level ℓ days ago on mortality.

In the figure, A is the sum of the positive coefficients for low lags and represents the fact that pollution levels in the past few days or weeks have a positive effect on mortality. However, the negative coefficients in B mean that pollution levels a longer period ago have a negative effect. This is due to “depletion of the risk pool,” normally made up of elderly and sick people whose deaths have been hastened a few days or weeks by episodes of high pollution; this is known as “harvesting.” Here A overestimates the public health significance of pollution, since it is really $\mathsf{A} + \mathsf{B}$ (where B is negative) that represents deaths induced by a noticeable amount of time.

Daily data over 10 years are available on mortality, air pollution, and several meteorological variables for the city of Milan, Italy. It is of interest to use these to

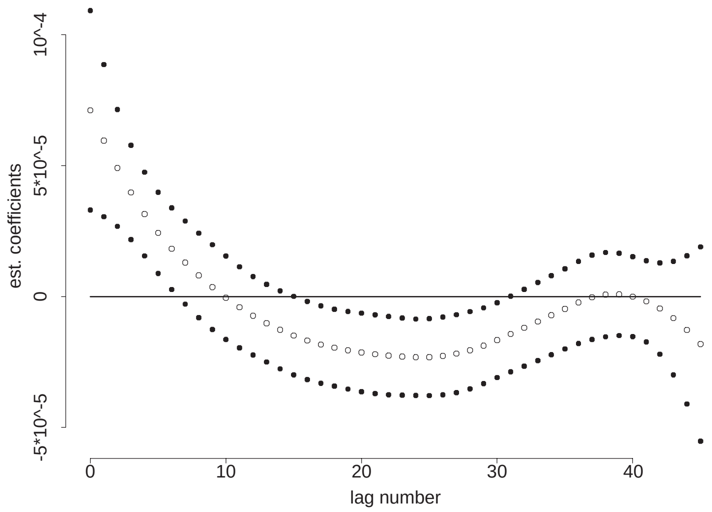


Figure 1.14 Estimates of the coefficients of the lags of sulphur dioxide on mortality in Milan, Italy. The shaded points are plus and minus 2 times the estimated standard error of each coefficient estimate.

quantify the public health significance of air pollution, incorporating the harvesting effect. By constraining the lag coefficients to be on a smooth (but otherwise flexible) curve, we obtained Figure 1.14. This suggests some evidence of harvesting. The construction of this result required some nonstandard semiparametric regression techniques that allowed for the lag coefficients to lie on a smooth curve and also be influenced by data on daily weather conditions.

Chapter 18 provides much fuller analyses and solutions for a selection of the problems presented in this chapter. Between now and then we will need to describe techniques for performing semiparametric regression analysis. The next chapter signals the start of this journey.