Chapter 3

Estimation for Basic Models of Modified Data

This chapter considers basic models where underlying observations are modified and the process of modification is known. The studied topics serve as a bridge between the case of direct data and the case of missing, truncated and censored data considered in the following chapters. Section 3.1 is devoted to density estimation based on biased data. Biased data is a classical statistical example of modified data. The interesting aspect of the presentation is that biased sampling is explained via a missing mechanism. Namely, there are underlying hidden realizations of a random variable of interest X^*, and then a hidden realization may be observed or skipped with the likelihood depending on the value of the realization. This sampling mechanism creates biased data because the distribution of the observed X is different from the distribution of the hidden X^*. As we will see in the following chapters, missing, truncation, censoring and other modifications typically imply biased data. Section 3.2 considers regression with biased responses. Section 3.3 explores regression with biased predictors and responses. Other sections are devoted to special topics. Among them, results of Section 3.7, where Bernoulli regression with unavailable failures is considered, will be often used in the next chapters.

3.1 Density Estimation for Biased Data

We are interested in estimation of the density $f^{X^*}(x)$ of the random variable of interest X^*. If a sample from X^* is available (the case of direct observations), then we know from Section 2.2 how to construct E-estimator of the density. In many applications a direct sampling is impossible and instead the following biased sampling may be possible. There is a hidden sequential sampling from pair (X^*, A) where A is a Bernoulli random variable such that $\mathbb{P}(A = 1|X^* = x) =: B(x)$ and $B(x)$ is called the biasing function. (Recall that a Bernoulli random variable A takes on values 0 or 1). If (X_1^*, A_1) is the first hidden realization of the pair, then we observe $X_1 = X_1^*$ if $A_1 = 1$ and skip the hidden realization otherwise. Suppose that we are interested in a sample of size n. Then the hidden sequential simulation continues until n observations $X_1, \ldots, X_n$ of X are obtained.

Let us present a particular example of biased data. Suppose that a researcher would like to know the distribution of the ratio of alcohol X^* in the blood of liquor-intoxicated drivers traveling along a particular highway. A sample $X_1, \ldots, X_n$ of the ratios of alcohol is available from routine police reports on arrested drivers charged with driving under the influence of alcohol (a routine report means that there are no special police operations to reveal all intoxicated drivers). Because a drunker driver has a larger chance of attracting the attention of the police, it is clear that the available observations are biased (and we may say length-biased) toward higher ratios of alcohol in the blood. Thus, the researcher should make an appropriate adjustment in a method of estimation of the distribution of the ratio of alcohol in the blood of all intoxicated drivers. There are many other similar examples in different sciences where a likelihood for an observation to appear in a sample depends on its value.

To estimate the density $f^{X^*}(x)$, we need to understand how it is related to the density $f^X(x)$ of biased observations. In what follows, we assume that X^* is supported on $[0,1]$ and the biasing function $B(x) := \mathbb{P}(A=1|X^*=x)$ is known and

$$B(x) \geq c_0 > 0. \tag{3.1.1}$$

The joint density of the pair (X^*, A) can be written as

$$f^{X^*,A}(x,1) = f^{X^*}(x)\mathbb{P}(A=1|X^*=x) = f^{X^*}(x)B(x). \tag{3.1.2}$$

Further, we observe $X = X^*$ given $A = 1$. This, together with (3.1.2), allows us to write,

$$f^X(x) = f^{X^*|A}(x|1) = \frac{f^{X^*,A}(x,1)}{\mathbb{P}(A=1)} = \frac{f^{X^*}(x)B(x)}{\mathbb{P}(A=1)}. \tag{3.1.3}$$

Further, (3.1.2) yields

$$\mathbb{P}(A=1) = \int_0^1 f^{X^*,A}(x,1)dx = \int_0^1 f^{X^*}(x)B(x)dx = \mathbb{E}\{B(X^*)\}. \tag{3.1.4}$$

Finally, (3.1.3) allows us to get the following formula

$$\mathbb{E}\Big\{\frac{1}{B(X)}\Big\} = \int_0^1 \frac{f^X(x)}{B(x)}dx = \int_0^1 \frac{f^{X^*}(x)B(x)}{B(x)\mathbb{P}(A=1)}dx = \frac{1}{\mathbb{P}(A=1)}. \tag{3.1.5}$$

This formula points upon a simple sample mean estimator of $\mathbb{P}(A=1)$.

Note that the above-presented formulas are based on the sequential model of creating biased data. In general, instead of exploring the process of collecting biased data, the problem of estimation of $f^{X^*}(x)$ based on a biased sample from X is formulated via the following relation:

$$f^X(x) = \frac{f^{X^*}(x)B(x)}{\int_0^1 f^{X^*}(u)B(u)du}. \tag{3.1.6}$$

In this case $B(x)$ is not necessarily the probability and may take on values larger than 1. On the other hand, according to (3.1.6), the biasing function can be always rescaled to make it not larger than 1, and this rescaling does not change the probability model.

In what follows we will use model (3.1.3)-(3.1.4) rather than (3.1.6) to stress the fact that biased data may be explained via a sequential missing mechanism. As it was already explained, this does not affect the generality of considered model. Also, we will use notation $B^{-1}(x) := 1/B(x)$.

Now, given known $B(x)$, (3.1.1) and (3.1.3)-(3.1.4), we are in a position to explore the problem of estimation of the density f^{X^*} based on a biased sample $X_1, \ldots, X_n$ from X.

First of all, let us stress that if the biasing function $B(x)$ is unknown, then no consistent estimation of the density of interest is possible. This immediately follows from (3.1.6) which shows that only the product $f^{X^*}(x)B(x)$ is estimable. Unless the nuisance function is known (or may be estimated), we cannot factor out $f^{X^*}(x)$ from the product $f^{X^*}(x)B(x)$. This is the pivotal moment in our understanding of the biasing modification of data, and we will observed it in many particular examples of missing and modified data. Only if the biasing function is known (for instance from previous experiments) or may be estimated based on auxiliary data, the formulated estimation problem becomes feasible and consistent estimation of $f^{X^*}(x)$ becomes possible.

The E-estimation methodology of Section 2.2 tells us that we need to propose a sample mean estimator of Fourier coefficients θ_j of the density of interest $f^{X^*}(x)$. To do this, we should first write down θ_j as an expectation and then mimic the expectation by a sample

mean estimator. To make the first step, using (3.1.3) let us write down Fourier coefficients θ_j as follows:

$$\theta_j := \int_0^1 \varphi_j(x) f^{X^*}(x) dx$$

$$= \mathbb{P}(A=1) \int_0^1 \varphi_j(x) f^X(x) B^{-1}(x) dx = \mathbb{P}(A=1)\mathbb{E}\{\varphi_j(X)B^{-1}(X)\}. \tag{3.1.7}$$

Here $\varphi_j(x)$ are elements of the cosine basis on $[0,1]$ (the definition and discussion can be found in Section 2.1). Note that $\theta_0 = \int_0^1 \varphi_0(x) f^{X^*}(x)dx = \int_0^1 f^{X^*}(x)dx = 1$, and hence we need to estimate only Fourier coefficients θ_j, $j \ge 1$. Formula (3.1.7) implies the following plug-in sample mean estimator of Fourier coefficients:

$$\hat{\theta}_j := \hat{P} n^{-1} \sum_{l=1}^n \varphi_j(X_l) B^{-1}(X_l), \tag{3.1.8}$$

where according to (3.1.5)

$$\hat{P} := \frac{1}{n^{-1}\sum_{l=1}^n B^{-1}(X_l)} \tag{3.1.9}$$

is the plug-in sample mean estimator of $\mathbb{P}(A=1)$.

Fourier estimator (3.1.8) yields the density E-estimator $f^{X^*}(x)$ of Section 2.2, and the coefficient of difficulty is

$$d := \lim_{n,j\to\infty} n\mathbb{E}\{(\hat{\theta}_j - \theta_j)^2\}$$

$$= [\mathbb{P}(A=1)]^2 \mathbb{E}\{B^{-2}(X)\} = \mathbb{P}(A=1)\mathbb{E}\{B^{-1}(X^*)\}. \tag{3.1.10}$$

Recall that the coefficient of difficulty of estimation of density f^{X^*}, based on direct observations of X^*, is 1 (see Section 2.2). Is the coefficient (3.1.10) larger? In other words, do biased data make the estimation problem more complicated or some biasing schemes may improve the estimation? To answer the question, let us use the Cauchy-Schwarz inequality (1.3.33) and write,

$$1 = [\int_0^1 f^{X^*}(x)dx]^2 \le [\int_0^1 f^{X^*}(x)B(x)dx][\int_0^1 f^{X^*}(x)B^{-1}(x)dx]$$

$$= \mathbb{P}(A=1)\mathbb{E}\{B^{-1}(X^*)\} = d, \tag{3.1.11}$$

with the equality only for the case of direct observations when $B(x) \equiv 1$. We conclude that biasing makes estimation more complicated and requires larger sample sizes for the same quality of estimation.

Figure 3.1 allows us to both understand the problem of biased data and test performance of the E-estimator of the underlying density f^{X^*}. Its caption explains the simulation and the four diagrams with a corner function being the underlying density. The histograms show simulated biased data. The biasing function is linear $B(x) = a + bx$, it is often referred to as length-biasing and frequently occurs in applications. The particular values of a and b, used in the simulations, are 0.2 and 0.8, respectively. This biasing favors larger values of X^* to be observed and hinders observing smaller values. The latter skews observed data to the right, and this is clearly seen in the left diagram. Here the underlying density f^{X^*} is the Uniform, and the histogram is dramatically skewed to the right. The proposed E-estimator does a perfect recovering of the underlying Uniform density. Let us also stress that if we do not suspect that the data is biased, or suspect that the data may be biased but do not know the biasing function, it is impossible to restore the underlying density. The latter is an important observation that we will repeatedly see in cases of modified and missing data.

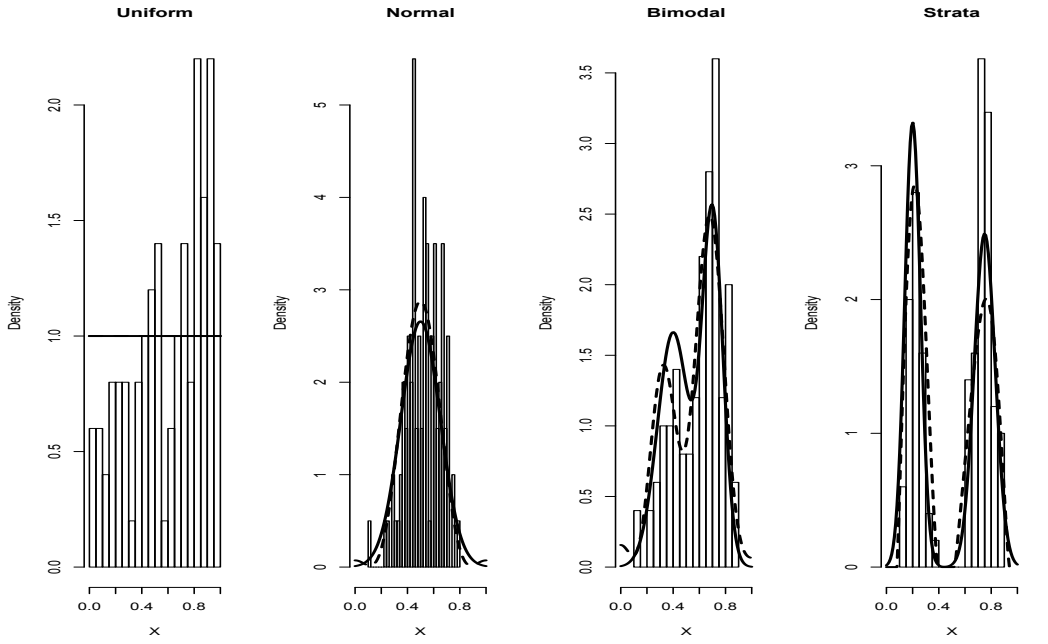

Figure 3.1 *Density estimation for biased data with the biasing function* $B(x) = a + bx$. *Four diagrams correspond to different underlying densities* $f^{X^*}(x)$ *indicated in the title and shown by the solid line. The biased observations are shown by the histogram and the density E-estimate* $\hat{f}^{X^*}(x)$ *by the dashed line.* {*Parameters of the biasing function are controlled by set.B = c(a,b)*} *[n = 100, set.B = c(0.2,0.8), cJ0 = 4, cJ1 = .5, cTH = 4]*

The Normal diagram in Figure 3.1 is another interesting example of the length-biased data. Again we see the skewed to the right histogram of the biased data, and the E-estimator correctly restores the symmetric shape of the Normal density. The funny tails are due to outliers. The Bimodal is another teachable example. This density is difficult for estimation even for the case of direct observations, and here look at how the E-estimator correctly increases (with respect to the histogram) the left mode and decreases the right mode. A similar outcome is observed in the Strata diagram where the E-estimator also corrects the histogram. Let us also shed light on underlying coefficients of difficulty. For instance, for the case of $B(x) = 0.1 + 0.9x$, coefficients of difficulty for our 4 corner densities are 1.4, 1.1, 1.1 and 1.3, respectively.

Figure 3.2 allows us to zoom on biased data and quantify the quality of estimation. The diagrams are similar to those in Figure 3.1, only here confidence bands are added (note that they are cut below the bottom line of the histogram). In the top diagram the underlying density is the Bimodal shown by the solid line. The histogram clearly exhibits the biased data that are skewed to the right. The E-estimate (the dashed line) fairly well exhibits the underlying density, and the 0.95-level confidence bands shed additional light on the quality of estimation. The subtitle shows the integrated squared error (ISE) of the E-estimate, the estimated coefficient of difficulty and the sample size. The bottom diagram exhibits a simulation with the underlying density of interest being the Strata. The left stratum is estimated worse than the right one, and we see that the underlying left strata is beyond the pointwise band but still within the simultaneous one. While the E-estimate is far from being perfect, it does indicate the two pronounced strata and even shows that the left mode is larger than the right one despite the heavily right-skewed biased data.

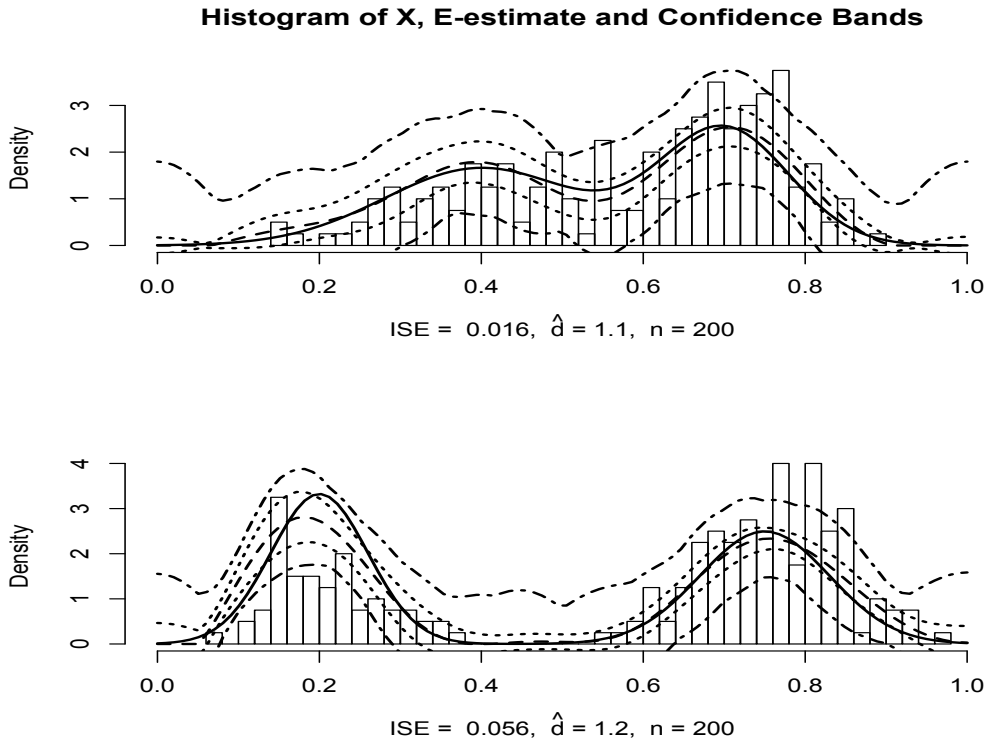

Figure 3.2 *Density estimation for biased data. Results of two simulations are exhibited for the Bimodal and the Strata underlying densities* $f^{X^*}(x)$. *Simulations and the structure of diagrams are similar to Figure 3.1 only here* $1-\alpha$ *confidence bands, explained in Section 2.6, are added. The* $1-\alpha$ *pointwise and simultaneous bands are shown by the dotted and dot-dashed lines, respectively. The exhibited confidence bands are truncated from below by the bottom line of the histogram.* {*Parameters of the biasing function are controlled by set.B =c(a,b), underlying densities are chosen by set.corn, and* α *is controlled by the argument alpha.*} *[n = 200, set.B = c(0.2,0.8), set.corn = c(3,4), alpha = 0.05, cJ0 = 3, cJ1 = 0.8, cTH = 4]*

One theoretical remark is due about the proposed E-estimator. Its natural competitor is the ratio estimator which is based on formula (3.1.3). The ratio estimator is defined as the E-estimator $\hat{f}^X(x)$, based a biased sample, divided by $B(x)/\hat{P}$. It is possible to show that the proposed E-estimator is more efficient than the ratio estimator, and this is why it is recommended. On the other hand, the appealing feature of the ratio estimator is in its simplicity.

We are finishing this section with a remark about the relation between the biased and missing data. Recall that a biased sample may be generated by a sequential sampling from X^* when some of the realizations are missed. Further, the sample size n of biased observations corresponds to a larger random sample size N (stopping time) of the hidden sample from X^*. One may think that $N-n$ observations in the hidden sample from X^* are missed. This thinking bridges the biased data and the missing data. Furthermore, there is an important lesson that may be learned from the duality between the biasing and missing. We know that unless the biasing function is known, consistent estimation of the density of X^* is impossible. Hence, missing data may preclude us from consistent estimation of an underlying density f^{X^*} unless some additional information about the missing mechanism is available.

In other words, missing may completely destroy information about density contained in a hidden sample. In the next chapters we will consider such missing mechanisms and refer to them as destructive missing.

3.2 Regression with Biased Responses

The aim is to estimate the regression function $m(x) := \mathbb{E}\{Y^*|X = x\}$ for the pair of continuous random variables (X, Y^*) where X is the predictor and Y^* is the response. For the case of a direct sample from (X, Y^*) this problem was discussed in Section 2.3 where a regression E-estimator was proposed. Here we are considering a more complicated setting when a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from (X, Y) is available, and

$$f^{X,Y}(x, y) = f^X(x)f^{Y|X}(y|x) = f^X(x)[f^{Y^*|X}(y|x)B(x, y)D(x)]. \tag{3.2.1}$$

Here $B(x, y)$ is a known biasing function and

$$B(x, y) \geq c_0 > 0, \tag{3.2.2}$$

the function $D(x)$ makes the expression in the square brackets a bona fide conditional density and it is defined as

$$D(x) := \frac{1}{\mathbb{E}\{B(X, Y^*)|X = x\}} = \mathbb{E}\{[1/B(X, Y)]|X = x\}, \tag{3.2.3}$$

and $f^X(x)$ is the design density of the predictor X supported on $[0, 1]$, and it is assumed that $f^X(x) \geq c_* > 0$.

Let us explain how a sample with biased responses, satisfying (3.2.1), may be generated. First, a sample $X_1, \dots, X_n$ from X is generated according to the design density $f^X(x)$. Then for each X_l, a single biased observation Y_l is generated according to the algorithm of Section 3.1 with (notation of that section is used) the density $f^{X^*}(y) = f^{Y^*|X}(y|X_l)$ and the biasing function $B(y) := \mathbb{P}(A = 1|X^* = y, X_l) = B(X_l, y)$, where A is the Bernoulli variable. Note that the difference between this sampling and sampling in Section 3.1 is that here biased responses are generated n times with different underlying densities and different biasing functions.

It is a nice exercise to check that the above-described biasing mechanism implies (3.2.1). Write,

$$\begin{aligned} f^{Y|X}(y|x) &= f^{Y^*|A,X}(y|1, x) \\ &= \frac{f^{Y^*,A|X}(y, 1|x)}{\mathbb{P}(A = 1|X = x)} = \frac{f^{Y^*|X}(y|x)\mathbb{P}(A = 1|Y^* = y, X = x)}{\mathbb{P}(A = 1|X = x)}. \end{aligned} \tag{3.2.4}$$

According to the above-described biasing algorithm, $P(A = 1|Y^* = y, X = x) = B(x, y)$, and we also get that

$$\begin{aligned} \mathbb{P}(A = 1|X = x) &= \mathbb{E}\{\mathbb{P}(A = 1|Y^*, X)|X = x\} \\ &= \mathbb{E}\{B(X, Y^*)|X = x\} = 1/D(x). \end{aligned} \tag{3.2.5}$$

Now we plug (3.2.5) in the right side of (3.2.4) and get (3.2.1) with $B(x, y) = P(A = 1|Y^* = y, X = x)$ and $D(x) = 1/\mathbb{P}(A = 1|X = x)$, as we wished to show. Further, the sampling mechanism and formulas (3.2.4)-(3.2.5) shed a new light on the model (3.2.1).

Now we are ready to propose an E-estimator for the regression with biased responses. The good news is that the biasing function $B(x, y)$ is assumed to be known and positive. The bad news is that function $D(x)$, $x \in [0, 1]$ is unknown. Hence our first task is to estimate this function.

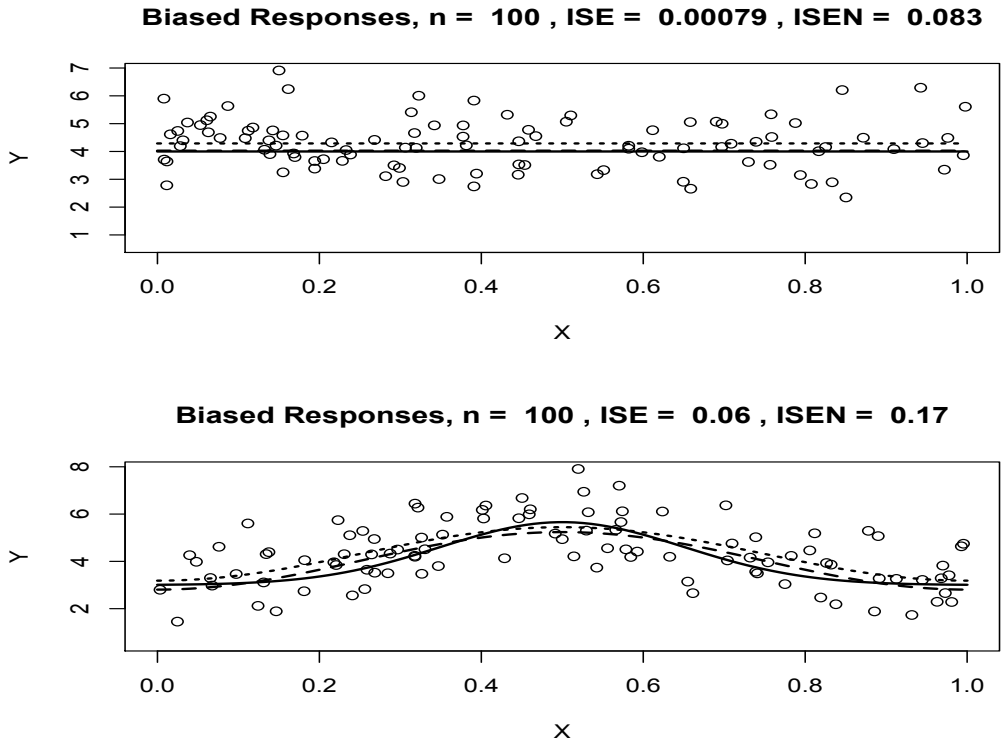

Figure 3.3 *Regression with biased responses. The biasing function is* $B(x,y) = b_1 + b_2x + b_3y$, *the design density is the Uniform, and the hidden regression is* $Y^* = [m(X) + 3\sigma] + \sigma\xi$ *where* $m(x)$ *is a corner function,* ξ *is a standard normal regression error which is independent of the predictor* X, *and* σ *is a parameter. The two diagrams correspond to the Uniform and the Normal functions* $m(x)$, *and the regression functions* $m(x) + 3\sigma$ *are shown by the solid lines. Biased data are shown by circles. Regression E-estimate (the dashed line) and the naïve regression E-estimate (the dotted line), based on biased data, have integrated squared errors indicated as* ISE *and* $ISEN$, *respectively.* {*Parameters of the biasing function are controlled by set.B* $= c(b_1, b_2, b_3)$ *and note that the biasing function must be positive, underlying regressions are chosen by the argument set.corn,* σ *is controlled by the argument sigma.*} *[n = 100, sigma = 1, set.B = c(0.3,0.5,2), set.corn = c(1,2), c = 1, cJ0 = 3, cJ1 = 0.8, cTH = 4]*

As usual, we use E-estimation methodology for estimating $D(x)$. Using (3.2.3) we can write for its Fourier coefficients

$$\kappa_j := \int_0^1 D(x)\varphi_j(x)dx$$

$$= \int_0^1 \mathbb{E}\{[1/B(x,y)]|X = x\}\varphi_j(x)dx = \mathbb{E}\Big\{\frac{\varphi_j(X)}{f^X(X)B(X,Y)}\Big\}. \tag{3.2.6}$$

This implies the plug-in sample mean estimator

$$\hat{\kappa}_j := n^{-1}\sum_{l=1}^{n} \frac{\varphi_j(X_l)}{\max(\hat{f}^X(X_l), c/\ln(n))B(X_l, Y_l)}, \tag{3.2.7}$$

where $\hat{f}^X(x)$, $x \in [0,1]$ is the E-estimator of the density $f^X(x)$ based on $X_1, \ldots, X_n$ (recall that the predictor is not biased).

The Fourier estimator (3.2.7) yields the E-estimator $\hat{D}(x)$, $x \in [0,1]$.

Now we are ready to explain how we can estimate Fourier coefficients of the regression function $m(x) := \mathbb{E}\{Y^*|X = x\}$. Using (3.2.1), Fourier coefficient of $m(x)$, $x \in [0,1]$ can be written as follows,

$$\theta_j := \int_0^1 m(x)\varphi_j(x)dx = \int_0^1 \Big[\int_{-\infty}^{\infty} y f^{Y^*|X}(y|x)dy\Big]\varphi_j(x)dx$$

$$= \int_0^1 \Big[\int_{-\infty}^{\infty} y f^{Y|X}(y|x)[B(x,y)D(x)]^{-1}dy\Big]\varphi_j(x)dx$$

$$= \mathbb{E}\Big\{\frac{Y\varphi_j(X)}{f^X(X)B(X,Y)D(X)}\Big\}. \tag{3.2.8}$$

This yields the following plug-in sample mean estimator of θ_j,

$$\hat{\theta}_j := \int_0^1 \frac{Y_l\varphi_j(X_l)}{\max(\hat{f}^X(X_l), c/\ln(n))B(X_l,Y_l)\hat{D}(X_l)}. \tag{3.2.9}$$

There is one useful remark about the plug-in $\hat{D}(X_l)$. It follows from (3.2.2) and (3.2.3) that if $B(x,y) \le c_B < \infty$, and recall that the biasing function is known, we get $D(x) \ge 1/c_B$. Then $\hat{D}(X_l)$, used in (3.2.9), may be truncated from below by $1/c_B$, that is we may plug in $\max(\hat{D}(X_l), 1/c_B)$.

Fourier estimator (3.2.9) yields the regression E-estimator $\hat{m}(x)$, $x \in [0,1]$ defined in Section 3.3.

As we see, the E-estimator for the regression with biased response is more complicated than the one for the regular regression proposed in Section 2.3 because now we need to estimate the nuisance function $D(x)$.

Figure 3.3 allows us to test the proposed estimator, and its caption explains the simulation and the diagrams. The top diagram shows the scattergram of regression with biased responses when the underlying regression is the Uniform plus 3. Note the high volatility of the biased data. The biasing clearly skews data up, and this is highlighted by the naïve regression E-estimate of Section 2.3 (the dotted line) based solely on the biased data. As we know, without information about biasing, a regression estimator cannot be consistent. The proposed regression estimator (the dashed line) is almost perfect, and this is highlighted by the small ISE. The bottom diagram shows a similar simulation for the Normal plus 3 underlying regression function. Here we have an interesting divergence between the two estimates. The naïve one is better near the mode and worse otherwise. The integrated squared errors quantify the quality of estimation.

It is worthwhile to repeat Figure 3.3 with different parameters and learn to read scattergrams with biased responses.

3.3 Regression with Biased Predictors and Responses

We are interested in the regression $m(x) := \mathbb{E}\{Y^*|X^* = x\}$ of the response Y^* on the predictor X^*. Realizations of the pair of continuous random variables (X^*, Y^*) are hidden and instead a sample of size n from a biased pair (X, Y) is available. It is known that joint density of the biased pair (X, Y) is

$$f^{X,Y}(x,y) = f^{X^*,Y^*}(x,y)B(x,y)D, \tag{3.3.1}$$

where

$$D = \frac{1}{\mathbb{E}\{B(X^*,Y^*)\}} = \mathbb{E}\{1/B(X,Y)\} \tag{3.3.2}$$

is a constant that makes the joint density bona fide. Further, the biasing function $B(x,y)$ is known and is bounded below from zero,

$$B(x,y) \geq c_0 > 0, \tag{3.3.3}$$

and the design density $f^{X^*}(x)$ of the hidden predictor X^* is supported on $[0,1]$ and $f^{X^*}(x) \geq c_* > 0$, $x \in [0,1]$.

Model (3.3.1) looks similar to the model (3.2.1) for the biased response. The difference is that now the predictor X may be also biased. Indeed, from (3.3.1) we get that

$$f^X(x) = f^{X^*}(x)[D\mathbb{E}\{B(x,Y^*)|X^* = x\}]. \tag{3.3.4}$$

As a result, unless $\mathbb{E}\{B(x,Y^*)|X^* = x\}$ is a constant (an example of the latter is $B(x,y) = B(y)$), the observed predictor is also biased. This is what differentiates models (3.3.1) and (3.2.1).

Let us present a particular example and then explain how biased data may be generated via a sequential missing algorithm. Recall the example of Section 3.1 about the distribution of the ratio of alcohol in the blood of liquor-intoxicated drivers based on routine police reports on arrested drivers. It was explained that the data in reports was biased given that a drunker driver was more likely to be stopped by the police. Suppose that now we are interested in the relationship between the level of alcohol and the age (or income level) of the driver. If it is reasonable to assume that both the level of alcohol and the age (income level) are the factors defining the likelihood of the driver to be stopped (as the thinking goes, your wheels give clues to your age, gender, income level and marital status) then both the level of alcohol and age (income) in the reports are biased.

A possible method of simulation of the biased data is based on a sequential missing. There is an underlying hidden sequential sampling from triplet (X^*, Y^*, A) where A is a Bernoulli random variable such that $\mathbb{P}(A = 1|X^* = x, Y^* = y) = B(x,y)$ satisfying (3.3.3). If (X_1^*, Y_1^*, A_1) is the first hidden realization of the pair, then we observe $(X_1, Y_1) := (X_1^*, Y_1^*)$ if $A_1 = 1$ and skip the hidden realization otherwise. Then the hidden simulation continues until n observations of (X,Y) are available.

Let us check that the simulated sample satisfies (3.1.3). For the joint density of the observed pair (X,Y) we can write,

$$f^{X,Y}(x,y) = f^{X^*,Y^*|A}(x,y|1) = \frac{f^{X^*,Y^*,A}(x,y,1)}{\mathbb{P}(A=1)}$$

$$= \frac{f^{X^*,Y^*}(x,y)\mathbb{P}(A=1|X^*=x,Y^*=y)}{\mathbb{P}(A=1)}. \tag{3.3.5}$$

If we compare (3.3.5) with (3.3.1), then we can conclude that the formulas are identical because $B(x,y) = \mathbb{P}(A = 1|X^* = x, Y^* = y)$ and $D = 1/\mathbb{P}(A = 1)$.

Now let us explain how an underlying regression function $m(x) := \mathbb{E}\{Y^*|X^* = x\}$ can be estimated by the regression E-estimator of Section 2.3. Following the E-estimation methodology, we need to understand how to estimate Fourier coefficients θ_j of the regression function. The approach is to write down Fourier coefficients as an expectation and then mimic the expectation by a sample mean estimator. Write,

$$\theta_j := \int_0^1 m(x)\varphi_j(x)dx = \int_0^1 \Big[\int_{-\infty}^{\infty} y[f^{X^*,Y^*}(x,y)/f^{X^*}(x)]dy\Big]\varphi_j(x)dx. \tag{3.3.6}$$

Using (3.3.1) we get the following expression for the marginal density $f^{X^*}(x)$ (compare with (3.3.4))

$$f^{X^*}(x) = f^X(x)\mathbb{E}\{[1/B(X,Y)]|X = x\}/D. \tag{3.3.7}$$

Using this formula in (3.3.6), together with (3.3.1), we continue,

$$\theta_j = \int_0^1 \Big[\int_{-\infty}^{\infty} \frac{y f^{X,Y}(x,y)}{DB(X,Y) f^{X^*}(x)} dy\Big]\varphi_j(x)dx$$

$$= \mathbb{E}\Big\{\frac{Y\varphi_j(X)}{B(X,Y)f^X(X)\mathbb{E}\{[1/B(X,Y)]|X\}}\Big\}. \tag{3.3.8}$$

This is a pivotal formula that sheds light on the possibility to estimate θ_j. First, we need to estimate two nuisance functions $f^X(x)$, $x \in [0,1]$ and

$$D(x) := \mathbb{E}\{[1/B(X,Y)]|X = x\}, \quad x \in [0,1], \tag{3.3.9}$$

that are used in the denominator of (3.3.8).

The density $f^X(x)$ is estimated by the E-estimator $\hat{f}^X(x)$ of Section 2.2 using the available sample $X_1, \ldots, X_n$. Estimation of $D(x)$ requires developing its own E-estimator. Fourier coefficients of $D(x)$ can be written as

$$\kappa_j := \int_0^1 D(x)\varphi_j(x)dx = \int_0^1 \mathbb{E}\{[1/B(X,Y)]|X = x\}\varphi_j(x)dx$$

$$= \mathbb{E}\{\varphi_j(X)/[f^X(X)B(X,Y)]\}. \tag{3.3.10}$$

This implies the plug-in sample mean estimator of Fourier coefficients,

$$\hat{\kappa}_j := n^{-1}\sum_{l=1}^{n} \frac{\varphi_j(X_l)}{\max(\hat{f}^X(X_l), c/\ln(n))B(X_l,Y_l)}. \tag{3.3.11}$$

In its turn, Fourier estimator (3.3.11) yields the E-estimator $\hat{D}(x)$, $x \in [0,1]$.

The two nuisance functions are estimated, and then (3.3.8) yields the following plug-in sample mean estimator of Fourier coefficients θ_j of the regression function $m(x)$,

$$\hat{\theta}_j := n^{-1}\sum_{l=1}^{n} \frac{Y_l\varphi_j(X_l)}{B(X_l,Y_l)\max(\hat{f}^X(X_l), c/\ln(n))\hat{D}(X_l)}. \tag{3.3.12}$$

One remark about (3.3.12) is due. If the biased data is created by a missing mechanism, then $D(x) \geq 1$, and then its E-estimator may be truncated from below by 1.

Apart of estimation of the regression, in applied examples it may be of interest to estimate the marginal densities f^{X^*} and f^{Y^*} of the hidden predictor X^* and the hidden response Y^*. We estimate these densities in turn.

Estimation of the hidden design density $f^{X^*}(x)$ is based on the following useful probability formula. We divide both sides of (3.3.1) by $DB(X,Y)$, then integrate both sides with respect to y, use (3.3.2), (3.3.9) and get

$$f^X(x) = f^{X^*}(x)\frac{D}{\mathbb{E}\{[1/B(X,Y)]|X = x\}} = f^{X^*}(x)\frac{D}{D(x)}. \tag{3.3.13}$$

Formula (3.3.13) tells us that the density $f^X(x)$ of the observable variable X is biased with respect to the density of interest $f^{X^*}(x)$ with the biasing function $1/D(x)$. Because we already constructed the E-estimator $\hat{D}(x)$, it can be used in place of unknown $D(x)$, and then $f^{X^*}(x)$, $x \in [0,1]$ be estimated by the plug-in density E-estimator of Section 3.1.

For the density of the hidden response Y^*, again using (3.3.1) we obtain the following useful formula (compare with (3.3.13))

$$f^Y(y) = f^{Y^*}(y)\frac{D}{\mathbb{E}\{[1/B(X,Y)]|Y = y\}}. \tag{3.3.14}$$

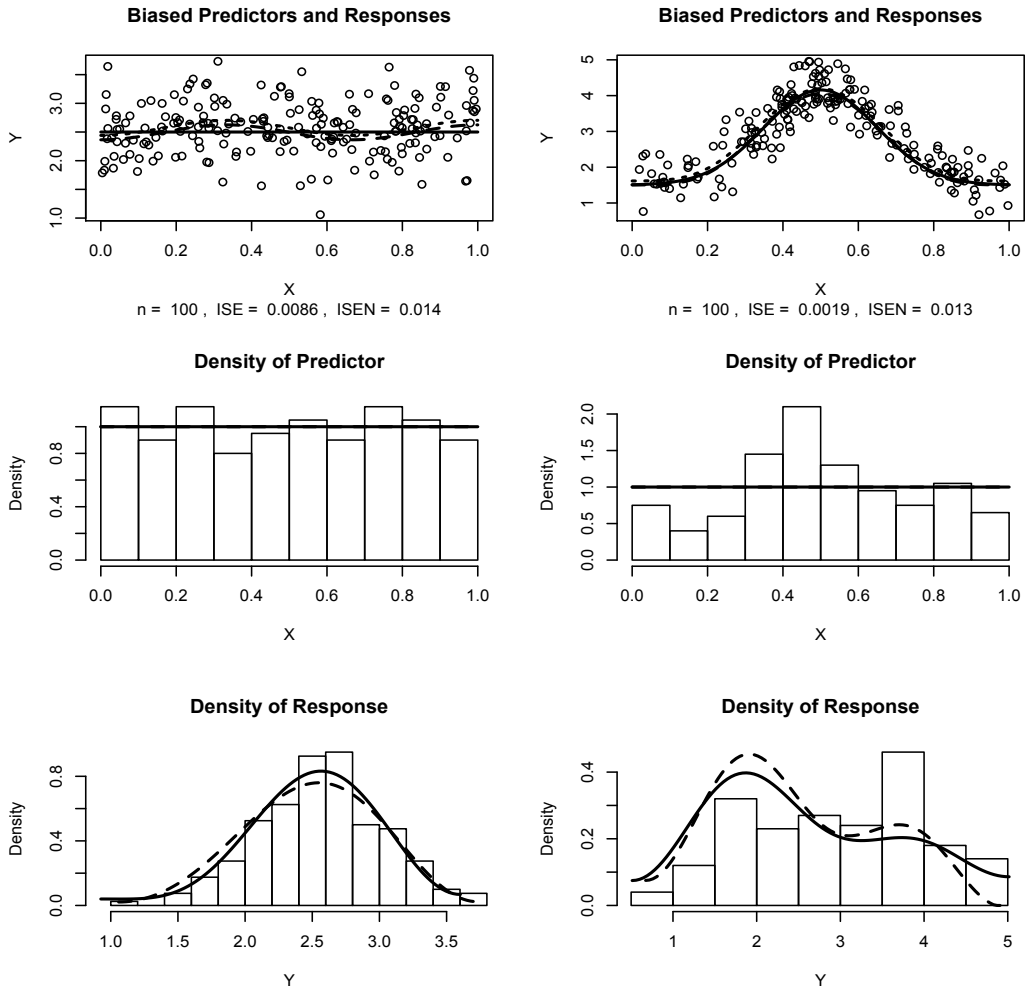

Figure 3.4 *Regression with biased predictors and responses. The underlying regression and the biasing function $B(x, y)$ are the same as in Figure 3.3 whose caption explains the parameters. Two columns of diagrams correspond to simulations with different underlying regression functions. A top diagram shows the scattergram of biased data by circles overlaid by the underlying regression (the solid line), the proposed regression E-estimate (the dashed line) and the naïve regression E-estimate of Section 2.3 based on biased data and not taking into account the biasing. The corresponding integrated squared errors of the two estimates are shown as ISE and ISEN. The middle and bottom diagrams show histograms of biased predictors and responses overlaid by the underlying marginal densities (the solid line) and their E-estimates (the dashed line). These densities are shown over the empirical range of biased observations. [n = 100, sigma = 0.5, set.B = c(0.2,0.5,1), set.corn = c(1,2), c = 1, cJ0 = 3, cJ1 = 0.8, cTH = 4]*

Sure enough, density $f^Y(y)$ is biased with respect to the density of interest $f^{Y^*}(y)$ with the biasing function $1/\mathbb{E}\{[1/B(X, y)]|Y = y\}$. The conditional expectation may be estimated similarly to how $D(x)$ was estimated, and then the density E-estimator of Section 3.1 may be used.

Let us stress one more time that we need to know the biasing function $B(x, y)$ to solve the regression and density estimation problems.

Figure 3.4 allows us to test performance of the proposed regression estimator and the marginal density estimators. Its diagrams and curves are explained in the caption, and note that while the underlying regression and even the biasing function are the same as in Figure 3.3, the biased data are generated according to (3.3.1) and hence the data are different from those generated according to formula (3.2.1) in Section 3.2.

Now let us look at the scattergrams. We begin with the left column where the underlying regression is created by the Uniform and it is shown by the solid line in the left-top diagram. First of all, note that we are dealing with a large volatility in the biased data. It looks like there are several modes in an underlying regression, but we do know that this is not the case. The dotted line shows the naïve regression estimate of Section 2.3 which ignores the known information that the data is biased. It does indicate modes, and recall that the E-estimator follows data while we know the underlying simulation and the hidden regression function. The proposed regression E-estimator, which takes into account the biasing nature of predictors and responses, also shows the same modes but it is much closer to the underlying regression. The latter is also supported by the indicated in the subtitle ISE = 0.0086 and ISEN = 0.014. Overall, despite a not perfect shape, the performance of the E-estimator is impressive keeping in mind the large volatility in the biased data. The middle and bottom diagrams are devoted to estimation of the hidden marginal densities. The density E-estimate of predictor is perfect despite the histogram of biased predictors exhibiting several modes. The bottom diagram is even more interesting because it allows us to look at the elusive marginal density of the response. Note that the histogram is asymmetric, and the density E-estimate is also skewed to the left, but overall it is a good estimate.

The right column shows us results of a simulation with the regression function created by the Normal. Again, the top diagram highlights the large volatility of data. Further, the scattergram and the underlying regression (the solid line) highlight the biased nature of the data (just notice that the data are skewed up). The latter is highlighted by the dotted line of the naïve regression estimate which goes above the underlying regression (the large volatility attenuates the difference). The proposed estimator (the dashed line) practically coincides with the underlying regression. Further, look at the corresponding integrated squared errors and note that taking into account the biased nature of data yields almost a seven-fold decrease in the integrated squared error. The middle diagram is of a special interest on its own. Here we can visualize the histogram of biased predictors, note the influence of the regression function on the distribution of observed biased predictors. It may be a good exercise to write down the density $f^X(x)$ for this simulation and then analyze it. The proposed marginal density E-estimator does a perfect job for this heavily biased data, and recall that this estimator is rather involved and based on estimation of the nuisance function $D(x)$. The bottom diagram is even more interesting because here we are dealing with the elusive marginal density of responses. Note that the hidden marginal density $f^{Y^*}(y)$ has a peculiar asymmetric shape with two modes. Further, look at how "disturbed" the histogram is, and how it magnifies the tiny right mode of the underlying density and, at the same time, diminishes the main left mode. Keeping in mind complexity of the marginal density estimation, which involves estimation of a nuisance conditional expectation, the E-estimator does an impressive job in exhibiting the shape of the underlying marginal density of the response Y^*.

The studied nonparametric regression problem with biased predictors and responses is a complicated one, both in terms of the model and the solution. It is highly advisable to repeat Figure 3.4, with both default and new arguments, and learn more about the biased regression and its consequences.

A remark about a regression with biased predictors is due. It is worthwhile to explain the problem via a particular example of the corresponding data modification. There is a hidden simulation from pair (X^*, A) where X^* is a continuous variable (predictor) supported on

$[0,1]$ and $f^{X^*}(x) \geq c_* > 0$, and A is a Bernoulli random variable generated according to the conditional density $\mathbb{P}(A = 1|X^* = x) =: B'(x) \geq c_0 > 0$. Denote the first realization of the pair as (X_1^*, A_1). If $A_1 = 1$, then $X_1 := X_1^*$ is observed, next the response Y_1 is generated according to the conditional density $f^{Y|X^*}(y|X_1)$, and the first realization (X_1, Y_1) of a regression sample with biased predictors is obtained. If $A_1 = 0$, then the realization (X_1^*, A_1) is skipped. Then the next realization of the pair (X^*, A) occurs. The sequential sampling stops whenever n realizations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are collected. The problem is to estimate the regression of the response Y on the hidden predictor X^*, that is, we want to estimate $m(x) := \mathbb{E}\{Y|X^* = x\} = \int_{-\infty}^{\infty} y f^{Y|X^*}(y|x)dy$.

Let us explore the regression function for the considered regression model with biased predictors. To do this, it suffices to find a convenient expression for the conditional density $f^{Y|X}(y|x)$. We begin with the corresponding joint density,

$$f^{Y,X}(y,x) = f^{Y,X^*|A}(y,x|1) = \frac{f^{Y,X^*,A}(y,x,1)}{\mathbb{P}(A=1)}$$

$$= \frac{f^{Y,X^*}(y,x)\mathbb{P}(A=1|Y=y,X^*=x)}{\mathbb{P}(A=1)}. \tag{3.3.15}$$

According to the considered biased sampling, the equality $\mathbb{P}(A = 1|Y = y, X^* = x) = \mathbb{P}(A = 1|X^* = x)$ holds. This equality, together with the relation $f^{Y,X}(y,x) = f^{Y|X}(y|x)f^X(x)$, the inequality $f^X(x) > 0$, $x \in [0,1]$ and (3.3.15), yield

$$f^{Y|X}(y|x) = \frac{f^{Y|X^*}(y|x)f^{X^*}(x)\mathbb{P}(A=1|X^*=x)}{f^X(x)\mathbb{P}(A=1)}. \tag{3.3.16}$$

Next, we note that

$$f^X(x) = \frac{f^{X^*}(x)\mathbb{P}(A=1|X^*=x)}{\mathbb{P}(A=1)}, \tag{3.3.17}$$

and this formula quantifies the biased modification of the predictors.

Using (3.3.17) in (3.3.16) we conclude that

$$f^{Y|X}(y|x) = f^{Y|X^*}(y|x). \tag{3.3.18}$$

Equality (3.3.18) sheds a light on the case of biased predictors in the regression setting. Here, despite the fact that X is biased, we have the equality (3.3.17) which implies that the regressions of Y on X and Y on X^* are the same. If you think about this outcome, it may seem either plain or confusing. If the latter is the feeling, then think about the fact that X is equal to X^* whenever X^* is observed, and then Y is generated according to $f^{Y|X^*}$. Of course, the same conclusion can be made from our general formula (3.3.1) when the biasing function $B(x,y) = B(x)$.

Finally, let us note that there is a special (and quite different) notion of unbiased predictors in finance theory, namely that forward exchange rates are unbiased predictors of future spot rates. In general, forward exchange rates are widely expected as a good predictor of future spot rates. For instance, any international transaction involving foreign exchange is risky due to unexpected change in currency exchange rates. Forward contract can be used to lower such risk, and as a result, the relation between the forward exchange rate and the corresponding future spot rate is of great concern for investors, portfolio managers, and policy makers. Forward rates are often expected to be unbiased estimator of corresponding future spot rates. It is possible to explore this problem, using our nonparametric technique, via statistical analysis of the joint distribution of the forward and spot rates.

3.4 Ordered Grouped Responses

So far we have explored problems where an underlying data was modified by a biasing mechanism caused, for instance, by observing a realization of a hidden sampling only if a specific event occurs. In this section we are considering another type of modification when it is only known that an underlying observation belongs to a specific group of possible observations. A group, depending on a situation and tradition, can be referred to by many names, for instance the stratum, category, cluster, etc.

Let us present several motivating examples. Strata, categories or clusters may define the socioeconomic statuses of a population: (i) Lowest 25 percent of wage earners; (ii) Middle 50 percent of wage earners; and (iii) Highest 25 percent of wage earners. A car may be driven with speed below 25, between 25 and 45, or above 45 miles per hour. A patient may have no pain, mild pain, moderate pain, severe pain, or acute pain. A patient in a study drinks no beer a day, 1 beer a day, more than 1 but fewer than 2 beers a day, and at least 2 beers a day. The overall rating of a proposal can be poor, fair, good, very good, or excellent.

In the above-presented examples, there is a logical ordering of the groups and hence they may be referred to as *ordinal* responses. To finish with the terminology, *nominal* responses have no natural logical ordering; examples are the color of eyes or the place of birth of a respondent to a survey.

Classical examples of nonparametric regression with grouped regression are the prediction of how a dosage of this or that medicine affects pain, or how the length of a rehabilitation program affects drug addiction, or how the quality of published papers affects the rating of a proposal.

To shed light on grouped (categorical, strata, cluster) nonparametric regression, let us consider the numerically simulated data shown in Figure 3.5. The left diagram shows an example of simulated classical additive regression $Y^* = m(X)+\sigma\eta$ which is explained in the caption. The small sample size $n = 30$ is chosen to improve visualization of each observation. The scatter plot is overlaid by boundaries for 4 ordered groups: $Y^* < -1$, $-1 \le Y^* < 1$, $1 \le Y^* < 3$, and $3 \le Y^*$. Then the data are modified by combining the responses into the above-highlighted groups (categories) shown in the right diagram. Thus, instead of the hidden underlying pairs (X_l, Y_l^*), where $Y_l^* = m(X_l) + \sigma\eta_l$, we observe modified pairs (X_l, Y_l) where Y_l is the number of a group (cell, category, stratum, etc.) for an unobserved Y_l^*. Figure 3.5 visually stresses the loss of information about the underlying regression function, because grouped data give no information on how underlying unobserved responses are spread out over cells. Please look at the right diagram and imagine that you need to visualize an underlying regression function. Further, the fact that heights of cells may be different, make the setting even more complicated.

The interesting (and probably unexpected) feature of the grouped regression is that the regression noise may help to recover an underlying regression. Indeed, consider a case where a regression function is $m(x) = 0$, $\sigma = 0$ and cells are as shown in Figure 3.5. Then the available observations are $(X_l, 2)$, $l = 1, 2, \dots, n$ and there is no way to estimate the underlying regression function. Further, even if there are additive errors but their range is not large enough, for instance $\sigma\eta_l$ are uniform $U(-0.99, 0.99)$, then the modified observations are again $(X_l, 2)$, $l = 1, \dots, n$.

It is a good exercise to repeat Figure 3.5 with different arguments and get used to this special type of data modification.

Now we are ready to explain how an underlying regression function may be estimated based on observed grouped responses.

In what follows it is assumed that the underlying regression model is

$$Y = m(X) + \varepsilon, \tag{3.4.1}$$

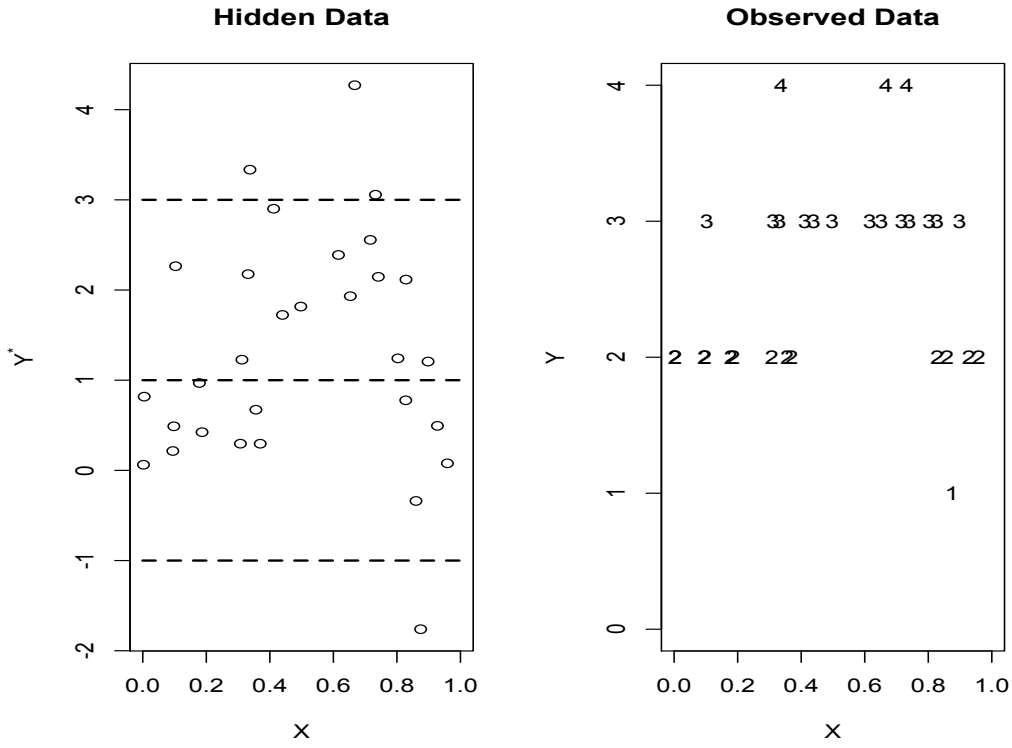

Figure 3.5 *Example illustrating grouped responses in nonparametric regression. A scattergram of underlying observations is shown in the left diagram, and it is generated by the model* $Y^* = m(X) + \sigma\eta$ *where* X *is the Uniform,* η *is standard normal and independent of* X*, and* σ *is a parameter controlling the standard deviation of the regression error. The sample size is* $n = 30$*. Responses are grouped according to 4 groups separated by dashed horizontal lines. The right diagram shows the corresponding grouped data. {Horizontal lines are controlled by the argument bound.set, regression errors are independent additive Normal with zero mean and standard deviation* σ *controlled by argument sigma.} [n = 30, set.corn = 3, sigma = 1, bound.set = c(-50,-1,1,3,50), cJ0 = 3, cJ1 = 0.8, cTH = 4]*

where X is supported on $[0, 1]$ and $f^X(x) \geq c_* > 0$, and the regression error ε is a continuous random variable with zero mean, finite variance and independent of the predictor X.

We begin with the parametric case $m(x) = \theta$ and the model of grouped data shown in Figure 3.5. Let $\bar{p}$ be the proportion of observations that have categories 3 or 4. Then the probability $\mathbb{P}(\theta + \varepsilon \geq 1) =: p$, which is the theoretical proportion of observations in the third and fourth categories, is

$$p = \mathbb{P}(\varepsilon \geq 1 - \theta) = 1 - F^{\varepsilon}(1 - \theta). \tag{3.4.2}$$

By solving this equation we get a natural estimate of θ,

$$\bar{\theta} = 1 - Q^{\varepsilon}(1 - \bar{p}), \tag{3.4.3}$$

where $Q^{\varepsilon}(\alpha)$ is the quantile function, that is, $\mathbb{P}(\varepsilon \leq Q^{\varepsilon}(\alpha)) = \alpha$.

Note that we converted the problem of grouped regression into Bernoulli regression discussed in Section 2.4. The latter is the underlying idea of the proposed solution.

There are three steps in the proposed regression estimator for regression with grouped responses.

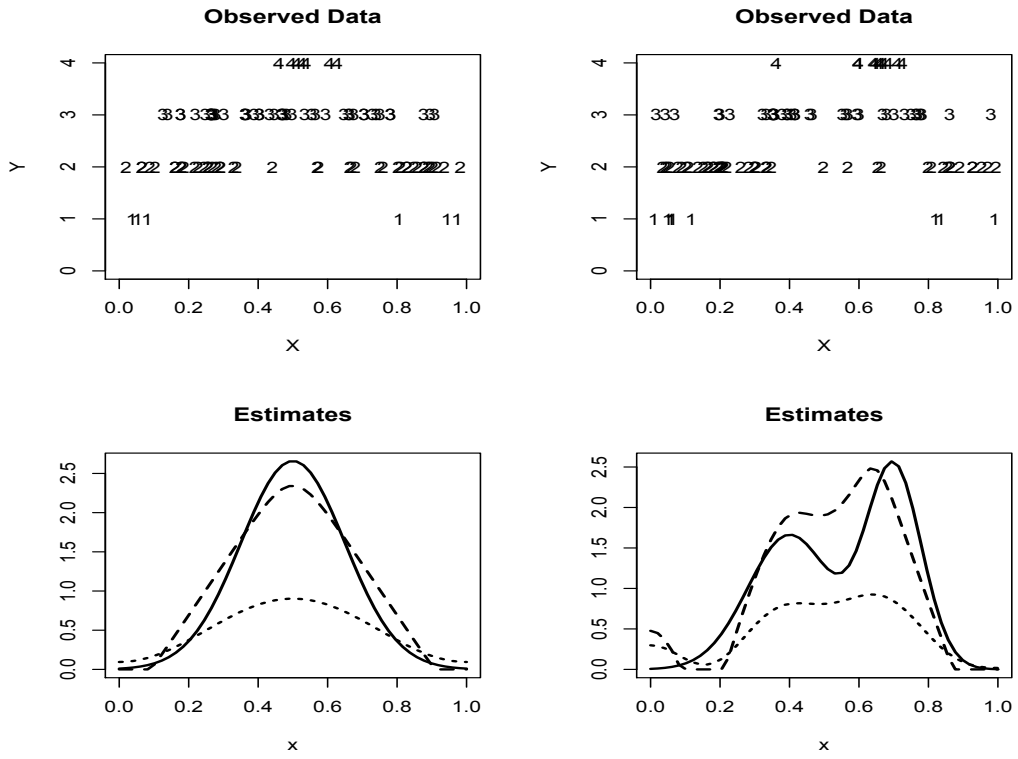

Figure 3.6 *E–estimates for nonparametric regression with grouped responses. The underlying simulation is the same as in Figure 3.5. Dotted and dashed lines show estimates $\hat{p}$ and $\hat{m}$ of the binary probabilities and the regression function, respectively. The solid line is the underlying regression function. [n = 100, set.corn = c(2,3), sigma = 1, bound.set = c(-50,-1,1,3,50),a = 0.005,b = 0.995, cJ0 = 3, cJ1 = 0.8, cTH = 4]*

Step 1. Combine the ordered groups into two groups of "successes" and "failures." Ideally, the boundary in responses that separates these two groups should be such that both successes and failures spread over the domain of predictors. For instance, for the example shown in Figure 3.5, the only reasonable grouping is $\{(1,2),\ (3,4)\}$.

Step 2. Use the Bernoulli regression E-estimator $\hat{p}(x)$ of Section 2.4 to estimate the probability of a success as a function in x. If no information about the regression error ε is given, this is the last step. If the distribution of ε is given, then go to step 3.

Step 3. This step is based on the assumption that the distribution of ε is known. Assume that an observed Y_l belongs to the success group iff $Y_l \ge c^*$ where c^* is a constant. Then

$$\hat{m}(x) = c^* - Q^{\varepsilon}\big(1 - [\hat{p}(x)]_a^b\big). \tag{3.4.4}$$

Here $[z]_a^b = \max(a, \min(z, b))$ is the truncation (or we can say projection) of z onto interval $[a, b]$. The truncation allows us to avoid infinite values for $\hat{m}$. The "default" values of a and b are 0.005 and 0.995.

Let us check how the proposed estimator performs. Figure 3.6 exhibits results of two simulations in two columns of diagrams. Underlying regression functions are the Normal and the Bimodal shown by the solid lines in the bottom diagrams. The regression errors are standard normal. The estimates $\hat{p}(x)$ and $\hat{m}(x)$ are shown by dotted and dashed lines, respectively. The datasets are simulated according to Figure 3.5, only here the sample size

$n = 100$. The estimates $\hat{p}(x)$ (the dotted lines) look not too impressive but not too bad either keeping in mind the complexity of the grouped data. After all, we could observe similar shapes in estimates based on $n = 100$ direct observations. The estimate for the Bimodal (see the right-bottom diagram) has a wrong and confusing left tail, but it corresponds to the left tail of the grouped data exhibited in the right-top diagram.

Knowing the distribution of regression error ε dramatically improves the visual appeal of estimates $\hat{m}(x)$ shown in the bottom diagrams of Figure 3.6 by the dashed lines. The estimate for the Normal is truly impressive keeping in mind both complexity of the setting and the small sample size. The estimate for the Bimodal is also a significant improvement both in terms of the two pronounced modes and their magnitudes (just compare with the dotted line which shows the estimate $\hat{p}(x)$).

The reader is advised to repeat this figure with different arguments and get used to this particular data modification and the proposed estimates.

The proposed estimator is not optimal because it is based on creating just two groups from existing groups. Nevertheless, asymptotically the suggested estimator is rate optimal, it is a good choice for the case of small sample sizes where typically only several groups contain a majority of responses, and its simplicity is appealing.

3.5 Mixture

This section presents a new type of data modification that occurs in a number of practical applications, and it will be explained via a regression example.

There is an underlying sample of size n from pair (X, Y) where X is the predictor and Y is the response. It is known that X is a continuous random variable supported on $[0, 1]$, Y is Bernoulli and $\mathbb{P}(Y = 1|X = x) = m(x)$. The problem is to estimate the conditional probability $m(x)$. As we know from Section 2.4, the problem may be treated as a Bernoulli regression because

$$m(x) := \mathbb{E}\{Y|X = x\}. \tag{3.5.1}$$

If the sample from (X, Y) is available, then the E-estimator of Section 2.4 can be used. In the considered mixture model, the responses are hidden and instead we observe realizations from (X, Z) where

$$Z = Y\zeta + (1 - Y)\xi. \tag{3.5.2}$$

Here ζ and ξ are random variables with known and different mean values μ_ζ and μ_ξ.

As we can see, the mixture (3.5.2) is a special modification of an underlying variable of interest Y.

One of the classical practical examples of the mixture is a *change-point* problem in observed time series where $X_l = l/n$ is time and $Y = 1$ if an object functions normally and $Y = 0$ if the object functions abnormally. Then Equation (3.5.2) tells us that while we do not observe Y directly, observations of ζ correspond to the case where the object functions normally and observations of ξ correspond to the case where it functions abnormally. Then changing the regression $m(X)$ from 0 to 1 implies that the object recovers from abnormal functioning.

Now let us propose an E-estimator for the underlying regression function (3.5.1). In what follows it is assumed that in model (3.5.2) $\mu_\zeta \neq \mu_\xi$ and that X is independent of ζ and ξ. Introduce a scaled version of the observed Z defined as

$$Z' := (Z - \mu_\xi)/(\mu_\zeta - \mu_\xi). \tag{3.5.3}$$

The underlying idea of the new random variable Z' is based on the following relation:

$$\mathbb{E}\{Z'|X = x\} = \mathbb{E}\{\frac{Z - \mu_\xi}{\mu_\zeta - \mu_\xi}|X = x\} = \mathbb{E}\{\frac{Y\zeta + (1 - Y)\xi - \mu_\xi}{\mu_\zeta - \mu_\xi}|X = x\}$$

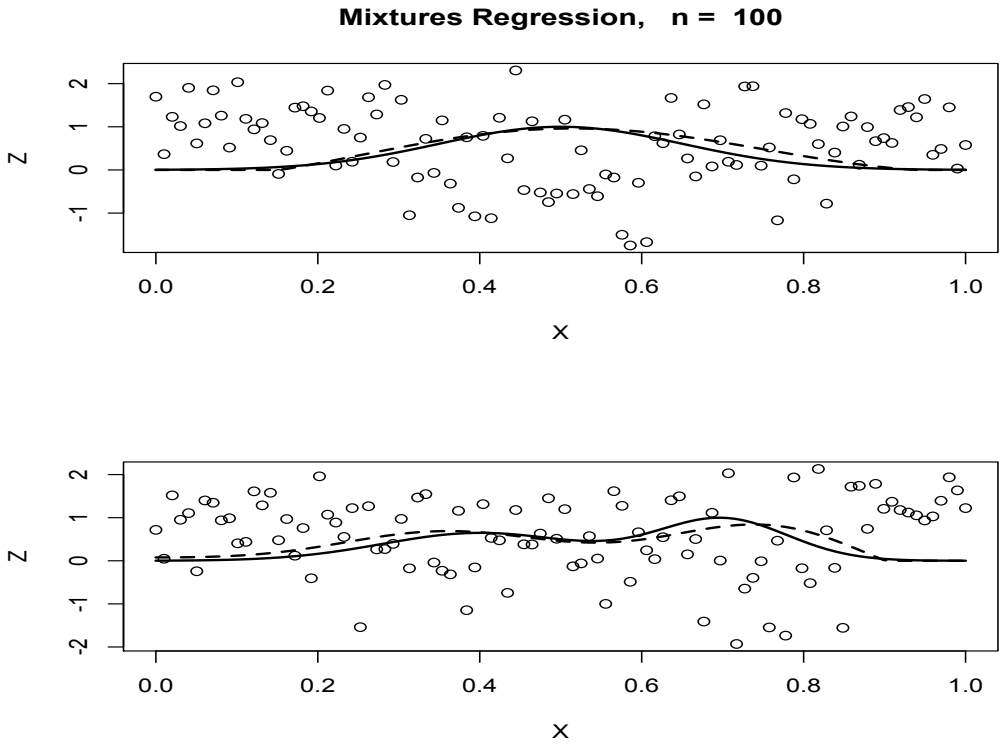

Figure 3.7 *E-estimation for mixtures regression. Underlying regression (the solid line) and its E-estimate (the dashed line) overlaid the scattergram shown by circles. In the top and bottom diagrams the underlying regressions are the Normal and the Bimodal divided by their maximum value. If the Uniform is chosen, then it is equal to 3/4. Realizations of X are equidistant and hence observations imitate a time series.* {*Random variable ξ is Normal(muxi,(sdxi)^2) and ζ is Normal(muzeta,(sdzeta)^2).*} *[n = 100, set.corn = c(2,3), muxi = 1, muzeta = 0, sdxi = 0.6, sdzeta = 0.9, cJ0 = 4, cJ1 = 0.8, cTH = 4]*

$$= \frac{\mathbb{E}\{Y|X=x\}\mathbb{E}\{\zeta\} + (1-\mathbb{E}\{Y|X=x\})\mathbb{E}\{\xi\} - \mu_\xi}{\mu_\zeta - \mu_\xi} = \mathbb{E}\{Y|X=x\} = m(x). \qquad (3.5.4)$$

We conclude that

$$m(x) = \mathbb{E}\{Z'|X=x\}, \qquad (3.5.5)$$

and the problem of the mixtures regression is converted into the traditional regression problem for a sample from (X, Z') for which we can use the E-estimator of Section 2.3. Let us also note that (3.5.1) implies the relation $0 \le m(x) \le 1$ which can be used by the E-estimator for bona fide estimation.

Figure 3.7 allows us to look at a mixtures regression and how the proposed E-estimator performs, and the caption explains the diagrams. We begin our discussion with the top diagram where the underlying regression function is the Normal divided by its maximum value (because the regression should be between zero and one). The regression function is shown by the solid line. Let us look at the scattergram of observations, shown by the circles, and the solid curve which is the estimand. It is clear that in no way the scattergram resembles the regression. This is a very interesting observation because it tells us that a mixture scattergram should be visualized differently than a scattergram for classical regression discussed in Section 2.3. We can realize that larger values of the regression function

correspond to smaller values of Z because the mean of ζ is smaller than the mean of ξ. Further, even with this fact taken into consideration, we can realize the shape of the regression but not its values. This is why we should appreciate performance of the E-estimator with the E-estimate shown by the dashed line. Of course, the estimate is too large around $x = 0.75$. But is this the fault of the E-estimator or do the mixture data indicate the larger tail? This is a teachable issue to explore because it may hint on how to read mixture data. Let us note that there are three relatively small observations of Z around $x = 0.8$. This is what causes the regression estimate to increase its value around this area. Keeping in mind the relatively small sample size $n = 100$, the outcome is impressive.

Now let us look at the bottom diagram where the underlying hidden regression is the Bimodal divided by its maximum value. We already know that estimation of the Bimodal regression is a challenging task even for the case of direct observations. (The reader is advised to return to Section 2.3 and check estimation of the Bimodal regression using Figure 2.7.) After the discussion of the top diagram, it is more clear why the scattergram corresponds to the underlying Bimodal regression. Further, the mixtures shed light on the fact why the main mode shifted to the right while the smaller one shifted to the left. To see this, just look at the smallest realizations of Z. Overall, for this particular simulation the E-estimator performed well.

Of course, the regression setting with mixtures is complicated and another simulation may produce a worse outcome. This is why it is instructive and useful to make more simulations, find a poor outcome, and then try to understand why the E-estimator performed in such a way.

3.6 Nuisance Functions

In many practical situations direct observations of a nonparametric function of interest are not available. A classical example of such a setting, considered in this section, is the case of a heteroscedastic regression with independent observations $(X_1, Y_1), \dots, (X_n, Y_n)$ of the pair of continuous variables (X, Y) where

$$Y = m(X) + \sigma(X)\varepsilon. \tag{3.6.1}$$

Here ε is zero mean, unit variance and independent of X random variable, the nonnegative function $\sigma(x)$ is called the scale (spread or volatility) function, and the predictor X is supported on $[0, 1]$ and $f^X(x) \geq c_* > 0$.

Traditional regression problem is to estimate the function $m(x)$, and then the design density $f^X(x)$ and the scale function $\sigma(x)$ become nuisance ones. These two functions may be of interest on their own. We know from Section 2.2 how to estimate $f^X(x)$, $x \in [0, 1]$ based on the observed predictors. In this section our task is to estimate the scale $\sigma(x)$, $x \in [0, 1]$.

In the statistical literature the same problem of estimating the scale function may be referred to as either estimation of a nuisance function in a regression problem or as estimation based on data modified by a nuisance regression function.

Let us explain the latter formulation of the problem of scale estimation. There are hidden observations $Z_l = \sigma(X_l)\varepsilon_l$ of the scale function. If they would be available, we could convert estimation of the scale into a regression problem. To do this, we write

$$Z^2 = \sigma^2(X) + \sigma^2(X)(\varepsilon^2 - 1), \tag{3.6.2}$$

and note that because ε is independent of X and $\mathbb{E}\{\varepsilon^2\} = 1$ we get

$$\mathbb{E}\{\sigma^2(X)(\varepsilon^2 - 1)|X\} = 0. \tag{3.6.3}$$

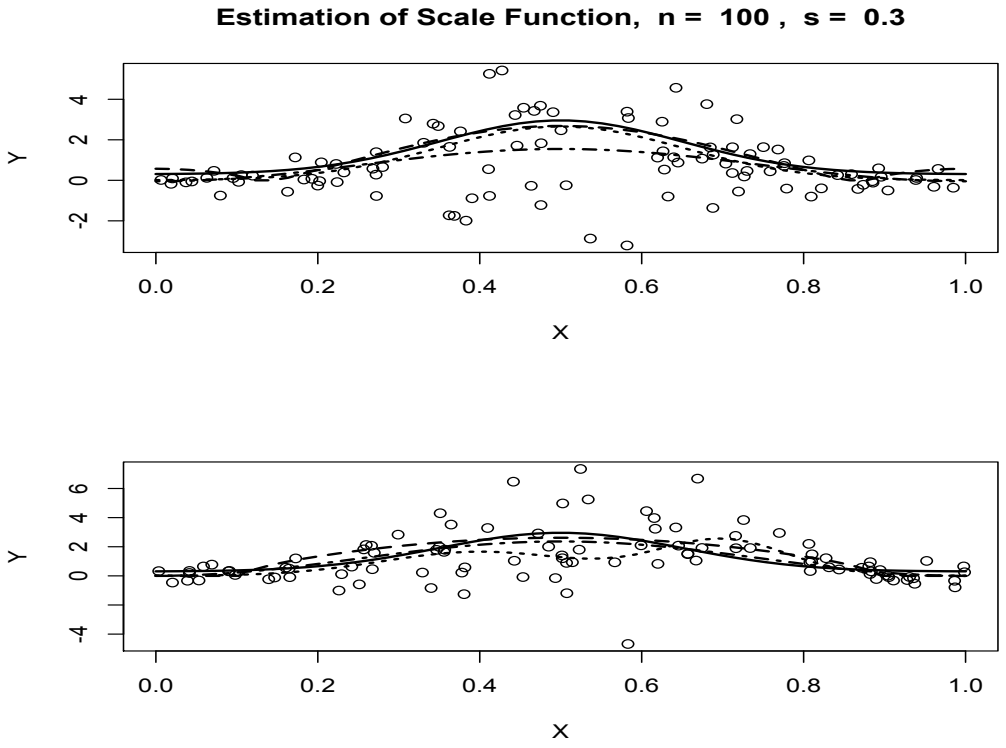

Figure 3.8 *Estimation of the scale function. Each diagram exhibits observations (scattergram) of a heteroscedastic regression generated by the Uniform design density, a regression function (the Normal and the Bimodal in the top and bottom diagrams), and the scale function* $\sigma(x) = s + \sigma f(x)$ *where* $f(x)$ *is the Normal, and* s *and* σ *are positive constants. In each diagram an underlying regression function is shown by the dotted line and its E-estimate by the dot-dashed line, while an underlying scale function and its E-estimate are shown by the solid and dashed lines, respectively.* {*Underlying regression functions are controlled by the argument set.corn, parameter* σ *by sigma, parameter* s *by s, and the choice of* f *is controlled by argument scalefun.*} *[n = 100, set.corn = c(2,3), sigma = 1, s = 0.3, scalefun = 2, cJ0 = 3, cJ1 = 0.8, cTH = 4]*

As a result, (3.6.2) is the regression model discussed in Section 2.3 and we can use the regression E-estimator for estimation of the regression $\sigma^2(x)$. Further, if the E-estimator takes on nonnegative values, then they are replaced by zero. Taking the square root of the above-defined estimator yields the estimator $\tilde{\sigma}(x, Z_1^n)$ of the scale function, and here $Z_1^n := (Z_1, Z_2, \ldots, Z_n)$. Of course, in the regression model (3.6.1) realizations of Z are hidden, instead we observe pairs (X_l, Y_l) such that

$$Y_l = m(X_l) + Z_l, \quad l = 1, 2, \ldots, n. \tag{3.6.4}$$

Equation (3.6.4) explains how the hidden observations Z_l are modified by the nuisance and unknown function $m(X_l)$.

A natural possible solution of a problem with data modified by nuisance functions is to first estimate them and then plug them in. In our case we can estimate the regression function $m(x)$ by the regression E-estimator $\hat{m}(x)$, and then replace unknown Z_l by

$$\hat{Z}_l := Y_l - \hat{m}(X_l). \tag{3.6.5}$$

Finally, we plug obtained $\hat{Z}_l$ in the above-defined estimator $\tilde{\sigma}(x, Z_1^n)$ and get the wished plug-in E-estimator $\hat{\sigma}(x) := \tilde{\sigma}(x, \hat{Z}_1^n)$.

Figure 3.8 allows us to appreciate complications that may be created by nuisance functions and how the proposed estimator performs, and its caption explains the simulation and the diagrams. We begin with the top diagram where the underlying regression function is the Normal, the scale function is the Normal plus $s = 0.3$, and the design density (the density of predictor) is the Uniform. Let us look at the scattergram (available pairs of observations) shown by circles. Can you visualize an underlying regression function that goes through the middle of the cloud of circles? It is not a simple task due to the large volatility of data in the middle of the unit interval. The regression E-estimate (the dot-dashed line) is poor, but it does correctly indicate the unimodal and symmetric nature of the Normal. Further, the regression E-estimate correctly describes the scattergram where we see a large number of negative responses in the middle of the unit interval. Nonetheless, the scale estimate (the dashed line) is impressively good.

The bottom diagram shows a similar simulation only with the underlying regression function being the Bimodal (the dotted line). Again, due to the strong volatility caused by the Normal scale function, it is practically impossible to visualize the underlying regression. The regression estimate (the dot-dashed line) shows a unimodal regression which barely catches characteristics of the Bimodal. This yields the overall poor scale estimate (the dashed line) which, nonetheless, correctly shows the unimodal and symmetric about 0.5 character of the Normal scale.

The presented simulations show the complexity of the studied modification, and having larger samples is the remedy. It is highly advisable to repeat Figure 3.8 and to get better understanding of this complicated problem.

More statistical examples of estimation of nuisance functions and estimation of data modified by nuisance functions will be considered in Chapters 4 and 9.

3.7 Bernoulli Regression with Unavailable Failures

In this section we are considering an important modification of data in Bernoulli regression when only cases with successes are observed while all cases with failures are unavailable. As we will see in the following chapters, this modification occurs in many applied problems and it is the pivot for solving many problems with missing data.

We begin with reviewing classical Bernoulli regression discussed in Section 2.4. Let us briefly recall the problem and the Bernoulli regression E-estimator. We are interested in a relationship between a continuous random variable X^* (the predictor) and a Bernoulli random variable A^*. A Bernoulli random variable takes on only two values 0 and 1, and traditionally the outcome 0 is classified as a "failure" and the outcome 1 as a "success." For instance, every day the level of pollution in a city can be below or above some threshold level, and the outcome is a Bernoulli random variable. Bernoulli random variable is completely defined by the probability of success $w := \mathbb{P}(A^* = 1)$, and then we have the formulae $\mathbb{E}\{A^*\} = w$ and $\mathbb{V}(A^*) = w(1-w)$. These are the basic facts that we need to know about a Bernoulli random variable.

Now let us recall the model of Bernoulli regression considered in Section 2.4. Consider a situation when the probability of the success w is the function of a predictor X^*, which is a continuous random variable with the density $f^{X^*}(x)$ supported on the unit interval $[0, 1]$ and $f^{X^*}(x) \geq c_* > 0$, $x \in [0, 1]$. Introduce the regression function

$$w(x) := \mathbb{P}(A^* = 1|X^* = x) = \mathbb{E}\{A^*|X^* = x\}, \tag{3.7.1}$$

and note that the joint mixed density of the pair (X^*, A^*) is

$$f^{X^*,A^*}(x, 1) = f^{X^*}(x)\mathbb{P}(A^* = 1|X^* = x) = f^{X^*}(x)w(x),$$

$$f^{X^*,A^*}(x,0) = f^{X^*}(x)(1-w(x)). \tag{3.7.2}$$

Using a directly observed sample $(X_1^*, A_1^*), \ldots, (X_n^*, A_n^*)$ from (X^*, A^*), the aim is to estimate the regression function $w(x)$. The proposed E-estimator is based on the E-estimation methodology of constructing a sample mean estimator of Fourier coefficients of an underlying regression function. Following the methodology, a Fourier coefficient $\theta_j := \int_0^1 w(x)\varphi_j(x)dx$ of the regression function $w(x)$, $x \in [0,1]$ can be written as

$$\theta_j = \int_0^1 \mathbb{E}\{A^*|X^* = x\}\varphi_j(x)dx = \mathbb{E}\Big\{\frac{A^*\varphi_j(X^*)}{f^{X^*}(X^*)}\Big\}. \tag{3.7.3}$$

Hence the corresponding sample mean estimator of θ_j is

$$\tilde{\theta}_j := n^{-1}\sum_{l=1}^n A_l^* \frac{\varphi_j(X_l^*)}{f^{X^*}(X_l^*)}. \tag{3.7.4}$$

If the design density $f^{X^*}(x)$ is unknown, then it is replaced by its E-estimator $\hat{f}^{X^*}(x)$ of Section 2.2 truncated below from zero by $c/\ln(n)$. In its turn, the plug-in Fourier estimator (3.7.4) yields the regression E-estimator $\tilde{w}(x)$ of Section 2.4.

Now we are ready to consider the Bernoulli regression problem with unavailable failures. The aim is still to estimate the regression function $w(x)$, $x \in [0,1]$ defined in (3.7.1), but now the sample $(X_1^*, A_1^*), \ldots, (X_n^*, A_n^*)$ is hidden. Instead, a subsample $X_1, \ldots, X_N$ of the predictors $X_1^*, \ldots, X_n^*$, corresponding to successes, is available and also the sample size n of the hidden sample is known. The subsampling is done as follows. If $A_1^* = 1$ then $X_1 := X_1^*$, and otherwise X_1^* is skipped. Then this subsampling continues, and finally if $A_n^* = 1$ then $X_N := X_n^*$ and otherwise X_n^* is skipped. Note that the number N of available predictors in the subsample is

$$N := \sum_{l=1}^n A_l^*. \tag{3.7.5}$$

Further, as usual we do not consider settings with $N = 0$ because there are no data, and in general we also exclude cases with relatively small N that are not feasible for nonparametric estimation. Let us also note that the available data may be equivalently written as $A_1^*X_1^*, \ldots, A_n^*X_n^*$ or as $(A_1^*X_1^*, A_1^*), \ldots, (A_n^*X_n^*, A_n^*)$.

It is convenient to use a different notation X for the observed predictor in a success case because the distribution of X is different from the distribution of the underlying predictor X^*. Indeed,

$$f^X(x) := f^{X^*|A^*}(x|1) = \frac{f^{X^*,A^*=1}(x)}{\mathbb{P}(A^*=1)} = f^{X^*}(x)\frac{w(x)}{\mathbb{P}(A^*=1)}. \tag{3.7.6}$$

This result implies that the observed predictor X has a biased distribution with respect to the hidden predictor X^*, and the biasing function is equal to the regression function $w(x)$.

Recall that biased distributions and biased data were discussed in Section 3.1. As we know from that section (and this also follows from (3.7.6)), based on the biased data we can consistently estimate only the product $f^{X^*}(x)w(x)$. The pivotal conclusion is that we need to know the design density $f^{X^*}(x)$ or its estimate for consistent estimation of $w(x)$.

As a result, we are exploring the following path for solving the problem. Formulas (3.7.3) and (3.7.4) tell us that to estimate Fourier coefficients of the regression function $w(x)$ (and hence to construct a regression E-estimator), it is sufficient to know only predictors X_l^* corresponding to $A_l^* = 1$. As a result, it is sufficient to know only the observed predictors $X_1, \ldots, X_N$. This is good news. The bad news is that we need to know the underlying

density $f^{X^*}(x)$ which, as we already know, cannot be estimated based on the available data.

Suppose that we know values $f^{X^*}(X_l)$, $l = 1, \dots, N$. Then the regression function may be estimated solely on available predictors corresponding to successes in the hidden Bernoulli sample. Indeed, we may rewrite (3.7.4) as

$$\tilde{\theta}_j = n^{-1} \sum_{l=1}^{N} \frac{\varphi_j(X_l)}{f^{X^*}(X_l)}. \tag{3.7.7}$$

This Fourier estimator yields the regression E-estimator $\tilde{w}(x)$, $x \in [0, 1]$. In some practical applications, when design of predictors is controlled, this conclusion allows us to use this regression E-estimator. Further, theoretically this E-methodology implies asymptotically (in n) optimal regression estimation.

If the design density $f^{X^*}(x)$ is unknown, then in some situations it may be possible to get an extra sample $X^*_{E1}, \dots, X^*_{Ek}$ of size $k \ll n$ from X^*; here $\ll$ means "significantly smaller." Then we may use the extra sample to calculate the density E-estimator $\hat{f}^{X^*}(x)$ and plug it in (3.7.7). Because the density estimator is used in the denominator, it is prudent to truncate it from below by $c/\ln(n)$ where c is the new parameter of the E-estimator. Then the (plug-in) sample mean estimator of Fourier coefficients of $w(x)$, $x \in [0, 1]$ is

$$\hat{\theta}_j := n^{-1} \sum_{l=1}^{N} \frac{\varphi_j(X_l)}{\max(\hat{f}^{X^*}(X_l), c/\ln(n))}. \tag{3.7.8}$$

This Fourier estimator yields the regression estimator $\hat{w}(x)$, $x \in [0, 1]$. The asymptotic theory shows that, under a mild assumption, this approach is consistent and implies optimal MISE (mean integrated squared error) convergence.

One more remark is due. In all future applications of the Bernoulli regression with unavailable failures, we need to know $w(x)$ only for $x \in \{X_1, \dots, X_n\}$. This is important information to know because it means that the range of observations in the E-sample should be close to the range of available observations $X_1, \dots, X_n$.

Let us test the proposed E-estimator on several simulated examples. Figure 3.9 presents the first set of four simulations, its caption explains the diagrams and the simulation. Here a left diagram shows the histogram of an extra sample of size k from X^*; the extra sample is referred to as an E-sample. An E-sample is used to estimate $f^{X^*}(X_l)$, $l = 1, \dots, N$. Values of an underlying design density $f^{X^*}(X_l)$ and its E-estimate $\hat{f}^{X^*}(X_l)$ are shown by circles and crosses, respectively. A corresponding right diagram shows via circles observed pairs $(X_l, 1)$. The size of a hidden sample n and the number of available predictors $N = \sum_{l=1}^{n} A^*_l$ are shown in the title. Further, the solid and dashed lines show the underlying regression $w(x)$ and its oracle-estimate based on all n hidden realizations of (X^*, A^*). Crosses show values of $\hat{w}(X_l)$.

Now we can look at specific simulations and outcomes shown in Figure 3.9. The top row shows the case of the constant regression $w(x) = 3/4$. The extra E-sample is tiny ($k = 30$) for a nonparametric estimation of the density. The default histogram stresses complexity of the density estimation which is a linear function shown by the circles. The E-estimate is surprisingly good here (look at the crosses). It is fair to say that visualization of data (the histogram) does not help us to recognize the density, and this is why the density E-estimate is impressive. Then the estimated values of the underlying design density are plugged in the regression E-estimator (3.7.8), and results are shown in the right-top diagram. First of all, let us compare the solid line (the underlying regression) and crosses showing values $\hat{w}(X_l)$. The regression estimate is perfect, and this is despite the fact that only $N = 75$ observations are available. Interestingly, the oracle's E-estimate, based on hidden

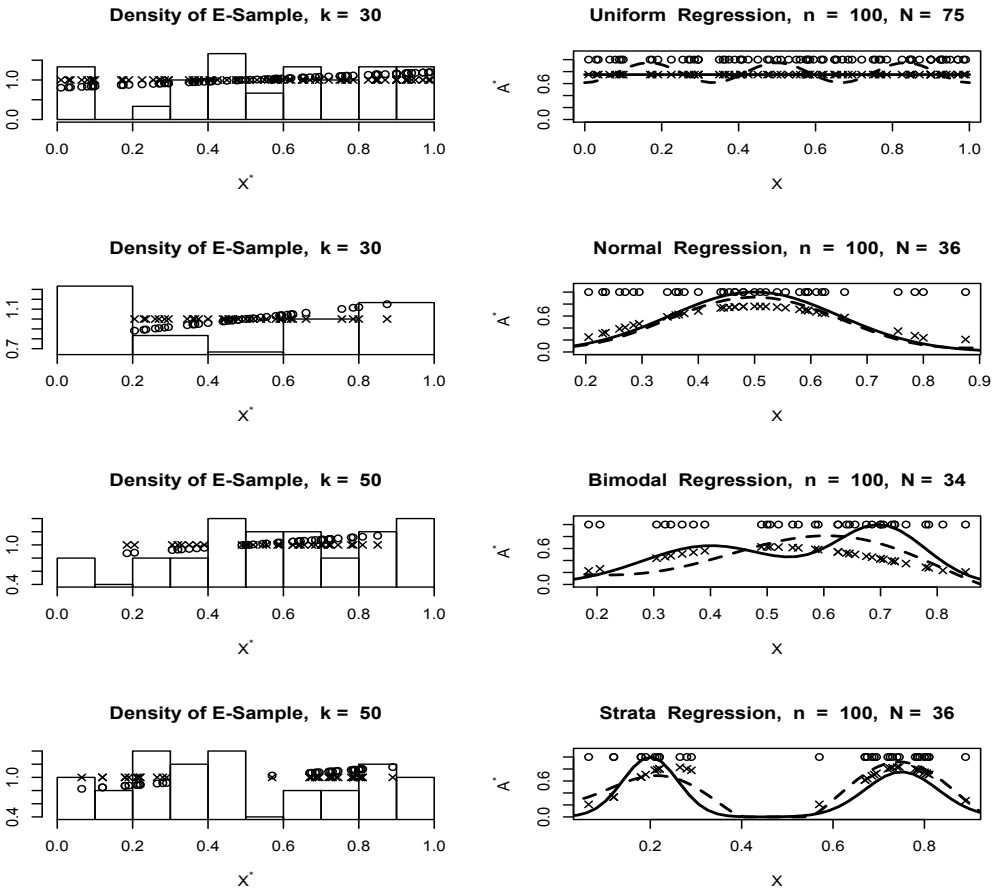

Figure 3.9 *Bernoulli regression with unavailable failures and an extra sample (E-sample) of predictors. Four rows of diagrams exhibit results of simulations with different underlying regression functions $w(x)$ shown by the solid line and named in the title of a right diagram. The sizes of a hidden Bernoulli regression and E-sample are n and k, respectively. A left diagram shows the histogram of E-sample and values of the design density $f^{X^*}(X_l)$and its E-estimate $\hat{f}^{X^*}(X_l)$, $l = 1, \dots, N$ by circles and crosses, respectively. A right diagram shows by circles available observations in the Bernoulli regression, the underlying regression function $w(x)$ (the solid line), oracle's regression E-estimate (the dashed line) based on n hidden observations, and by crosses values of the proposed E-estimator $\hat{w}(X_l)$, $l = 1, \dots, N$. {The figure allows to choose different parameters of E-estimator for the design density (they are controlled by traditional arguments) and the regression E-estimator. Further, for the regression E-estimator, parameters c_{J0} and c_{J1} can be specified for each of the 4 experiments. The latter is done by arguments setw.cJ0 and setw.cJ1. The argument desden controls the shape of the design density which is then truncated from below by the value dden and rescaled into a bona fide density. The argument st.k controls sample sizes of E-samples for each row.} [n = 100, set.k = c(30,30,50,50), desden = "$1 + 0.5 * x$", dden = 0.2, c=1, cJ0 = 3, cJ1 = 0.8, cTH = 4, setw.cJ0 = c(3,3,3,3), setw.cJ1 = c(0.3,0.3,0.3,0.3)]*

pairs (X_l^*, A_l^*), $l = 1, 2, \dots, n$, is much worse (look at the oscillating dashed line). This is an interesting outcome but it is rare, in general the oracle-estimate is much better.

The second (from the top) row of diagrams in Figure 3.9 considers the same setting only with the Normal regression function. Here again only $k = 30$ extra observations of X^* are available for estimating the design density f^{X^*}. Note that the histogram clearly deviates

from the underlying linear design density. However, fortunately for us, we need values of the density E-estimator only for $X_l \in [0.2, 0.9]$ interval, and within this interval the E-estimate is satisfactory. As a result, the right diagram shows us a fair regression estimate despite the fact that only $N = 36$ observations from hidden $n = 100$ are available. The oracle estimate of the regression (the dashed line) is good. In the third row of diagrams the case of the Bimodal regression is considered. Here the larger sample size $k = 50$ of E-sample is used and the design density estimate is fair. Unfortunately, this cannot help the regression E-estimator because the size $N = 34$ of available observations is too small and the regression function is too complicated (recall our simulations in Section 2.3). The poor oracle estimate, based on $n = 100$ hidden observations, sheds additional light on the difficult task. In the bottom diagram we explore the case of the Strata regression function. The design density estimate is fair, and the regression E-estimate is truly impressive given that only $N = 36$ observations are available. Further, this estimate is on par with the oracle estimate.

It is advisable to repeat Figure 3.9 with different parameters and get used to this challenging problem. Further, it is of interest to explore the relation between k and n that implies a reliable estimation comparable with the oracle's estimation. Further, Figure 3.10 allows us to use different parameters for the density estimator and regression E-estimators used in each row. The latter is a nice feature if we want to take into account different sample sizes and shapes of the underlying curves.

In many applications the support of the predictor may not be known. We have discussed this situation in Chapter 2, and let us continue it here because this case may imply some additional complications for our regression E-estimator. Namely, so far it has been explicitly assumed that the design density is bounded below from zero (recall that in Figure 3.9 the design density is not smaller than the argument *dden*). Let us relax these two assumptions and explain how this setting may be converted into the above-considered one.

Suppose that the hidden predictor X^* is a continuous random variable supported on a real line. Then our methodology of E-estimation is as follows. First, we combine N available predictors X_l and k extra observations X_{El} and find among these $N + k$ observations the smallest and largest values X_S and X_L, respectively. Then, using the transformation $(X - X_S)/(X_L - X_S)$ we rescale onto $[0, 1]$ the two available samples, and repeat all steps of the above-proposed regression E-estimation. The only new element here is that the obtained $\hat{f}^{X^*}(x)$ should be divided by $(X_L - X_S)$ to restore its values to the original interval.

Figure 3.10 illustrates this setting and the proposed solution. Its structure is similar to Figure 3.9, only here the regression function is the same in all 4 experiments, it is a custom-made function, and other differences are explained in the caption. Let us look at the top row of diagrams. Here the density E-estimate is fair, keeping in mind the small sample size $k = 30$, and its deviation from the underlying one is explained by the histogram. Please pay attention to the fact that 30 observations from a normal density may not be representative of an underlying density (as we see from the heavily skewed histogram). The deficiency in the density estimate is inherited by the regression estimate. Namely, note that the regression E-estimate (shown by crosses) is significantly smaller for positive values of X, and this is due to larger values of the density E-estimate. In the second row of diagrams results of an identical simulation are shown. Here the density estimate, at the required values X_l, is almost perfect. Of course, recall that the smallest values are truncated from below to avoid almost zero values in the denominator. The corresponding regression E-estimate is better. Overall, keeping in mind the small sample sizes $N = 61$ and $N = 68$ of available observations, the two regression estimates are fairly good and correctly indicate the sigmoid shape of the regression.

Simulations in the two bottom rows in Figure 3.10 use larger size $k = 50$ of E-samples. The second from the bottom row of diagrams exhibits a teachable outcome which stresses the fact that outcomes of small random samples may present surprises. Here we observe the worst density and regression estimates despite the largest $N = 82$. Note how the shortcom-

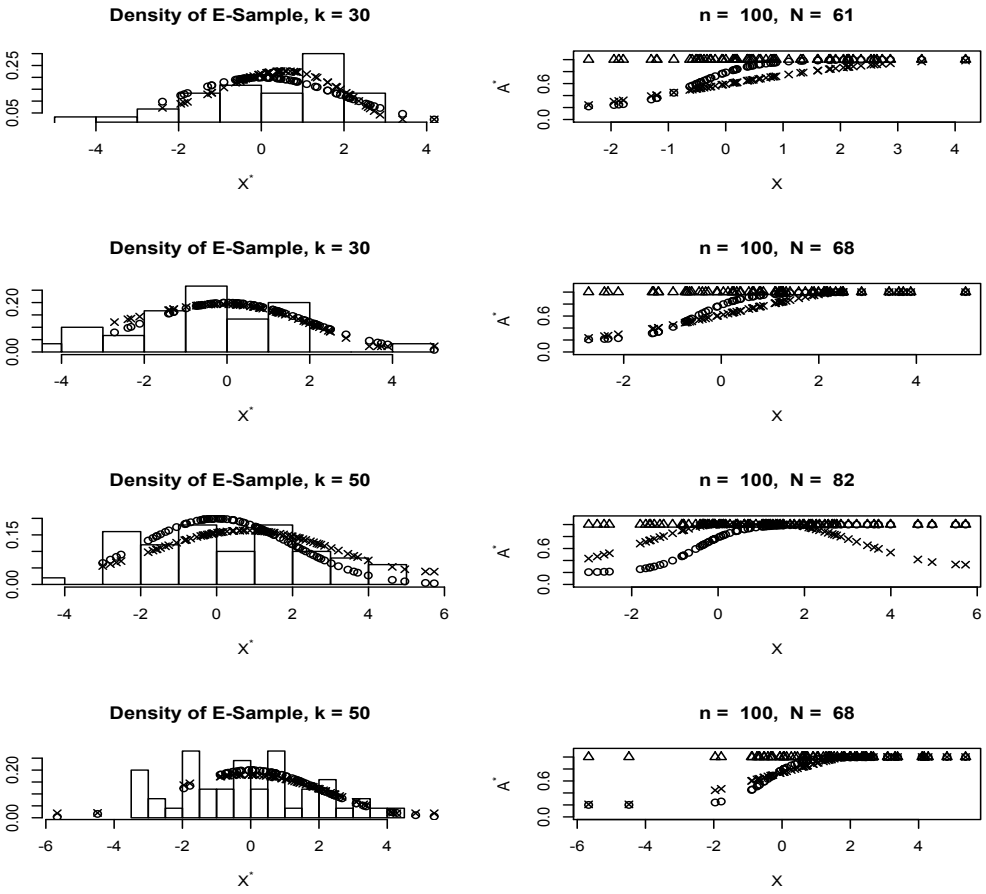

Figure 3.10 *Bernoulli regression with unavailable failures. The structure of the diagrams is the same as in Figure 3.9. The difference is that the design density is Normal*$(0, \sigma^2)$*, in all rows the same underlying regression function* $w(x)$ *is used and it is controlled by the string w, and in a right diagram triangles show available observations while circles and crosses show values of the underlying regression function* $w(X_l)$ *and its E-estimate* $\hat{w}(X_l)$*,* $l = 1, \dots, N$*.* {*The argument sigma controls the standard deviation* σ *of the normal design density. The string w defines the regression function* $w(x)$*, and note that* $w(x) \in [0, 1]$*. All other arguments are the same as in Figure 3.9.*} *[n = 100, set.k = c(30,30,50,50), sigma = 2, w =* $''0.2 + 0.8 * exp(1 + 2 * x)/(1 + exp(1 + 2 * x))''$*, c = 1, cJ0 = 3, cJ1 = 0.8, cTH = 4, setw.cJ0 = c(3,3,3,3), setw.cJ1 = c(0.3,0.3,0.3,0.3)]*

ings in the density estimate are inherited by the regression E-estimate. The bottom row of diagrams exhibits another outcome of the same simulation, and here both the density and the regression E-estimates are very good.

The simulations indicate that the issue that we should be aware of is that the range of the E-sample should be close to the range of available predictors. This remark may be useful if sequential E-sampling is possible.

Overall, we may conclude that if using a relatively small extra sample of hidden predictors is possible, then the proposed regression E-estimator is a feasible solution of the otherwise unsolvable problem of Bernoulli regression with unavailable failures.

It is highly advisable to repeat Figure 3.10 with different parameters and learn more about this important problem that will play a key role in statistical analysis of missing data.

3.8 Exercises

3.1.1 What is the definition of biased data?
3.1.2 Present an example of biased data.
3.1.3 Suppose that an underlying random variable X^* is observed only if it is larger than another independent random variable T. Are the observed realizations of X^* biased?
3.1.4 Verify (3.1.2) and (3.1.3).
3.1.5 Explain all components of formula (3.1.5).
3.1.6* For the setting of Exercise 3.1.3, write down a formula that relates the density of X^* with the density of X.
3.1.7 Is (3.1.8) a reasonable estimator of the Fourier coefficient (3.1.7)?
3.1.8 Find the mean of $\hat{\theta}_j$ defined in (3.1.8). Hint: Begin with the case when $P := \mathbb{P}(A = 1)$ is given, and then look at how using the plug-in estimate $\hat{P}$ affects the mean.
3.1.9* Evaluate the variance of $\hat{\theta}_j$ defined in (3.1.8). Hint: Prove formula (3.1.10).
3.1.10 Verify inequality (3.1.11).
3.1.11 Repeat Figure 3.1 for different biasing functions and explain outcomes.
3.1.12 Repeat Figure 3.2 with different underlying corner densities and biasing functions. What combinations imply worse and better estimates?
3.1.13 How parameters of the biasing function, used in Figure 3.2, affect the coefficient of difficulty for the four corner densities?
3.1.14 A naive estimate for biased data first estimates the density of the observed random variable X and then corrects its using a known biasing function $B(x)$. Write down a formula for this estimator. Hint: Density E-estimator can be used for estimating f^X, then use formula

$$f^X(x) = f^{X^*}(x)B(x)/B \tag{3.8.1}$$

where B is a constant which makes the density $f^X(x)$ bona fide (integrated to 1).

3.2.1 Explain a regression setting with biased responses.
3.2.2 Is the predictor, the response, or both biased under the model (3.2.1)?
3.2.3 Explain all functions in (3.2.1).
3.2.4 How is formula (3.2.3) obtained?
3.2.5 Explain how formula (3.2.4) is obtained. What is its relation to (3.2.1)?
3.2.6 How can a simulation of regression with biased responses be designed?
3.2.7 Consider a model where an underlying response Y^* is observed only if $Y^* > T$ where T is an independent random variable. Is this a sampling with biased responses?
3.2.8 For the setting of the previous exercise, what is the formula for the joint density of the observed pair of random variables (predictor and response)?
3.2.9 Suppose that in the setting of Exercise 3.2.7 the random variable T depends on predictor X. Does this information make a difference in your conclusions about the biased data? Is this a response-biased sampling?
3.2.10 Verify formula (3.2.8).
3.2.11 What is the underlying idea of the estimator (3.2.9)?
3.2.12 Evaluate the bias of estimator (3.2.9).
3.2.13 What is the underlying idea of the estimator $\hat{D}(x)$?
3.2.14 What is the bias of the estimator $\hat{D}(x)$?
3.2.15* The corresponding coefficient of difficulty of the proposed regression E-estimator is

$$d := \mathbb{E}\{[1/(f^X(X)B(X,Y)D(X))]^2\} = \int_0^1 \int_{-\infty}^{\infty} \frac{f^{Y^*|X}(y|x)}{f^X(x)B(x,y)D(x)} dydx. \tag{3.8.2}$$

Prove this assertion, or show that it is wrong and then suggest a correct formula. Hint: Begin with the case when all nuisance functions (like $f^X(x)$ or $D(x)$) are known.

3.2.16 Explain all arguments used by Figure 3.3.
3.2.17 Repeat Figure 3.3 a number of times using different regression functions. Which one is more difficult for estimation? Hint: Use both visual analysis and ISEs to make a conclusion.
3.2.18* Use Figure 3.3 to answer the following question. For each underlying regression function, what are the parameters of the biasing function that make estimation less and more challenging? Confirm your observations using theoretical analysis based on the coefficient of difficulty.
3.2.19* Consider the case $B(x, y) = B^*(y)$ when the biasing is defined solely by the value of the underlying response. Present all related probability formulas for this case, and propose an E-estimator.
3.2.20* In the literature, statisticians often consider a model where

$$f^{X|Y}(x|y) = f^{X|Y^*}(x|y). \tag{3.8.3}$$

Explore this case.

3.3.1 Explain the model of regression with both predictors and responses being biased.
3.3.2 Present an example of regression where both predictors and responses are biased.
3.3.3 What is the relationship between models discussed in Sections 3.2 and 3.3?
3.3.4 Why do we use in formula (3.3.1) a constant factor B and not a function?
3.3.5 Explain a sampling procedure corresponding to (3.3.4).
3.3.6 In formula (3.3.4), how can constant D be estimated?
3.3.7 Explain relations in (3.3.5).
3.3.8 Verify formula (3.3.6).
3.3.9 How can the right-hand side of (3.3.6) be written as expectation?
3.3.10 Explain how formula (3.3.8) is obtained.
3.3.11 Propose an estimator of the function $D(x)$.
3.3.12 Explain the motivation behind the Fourier estimator (3.3.11).
3.3.13 Explain how the marginal density of the underlying predictor X^* may be estimated.
3.3.14* Evaluate the mean, the variance and the coefficient of difficulty of the Fourier estimator (3.3.12).
3.3.15* Consider the model of regression with biased predictors and explain how the regression function can be estimated. Prove your assertion.
3.3.16 Explain the notion of unbiased predictor used in financial literature. Suggest a statistical analysis of this setting.
3.3.17 Repeat Figure 3.4 several times and explain simulated data and the E-estimates.
3.3.18 Using Figure 3.4 explain how parameters of the biasing function affect observations and the estimates.
3.3.19 Which of the underlying regression functions that may be used by Figure 3.4 is more challenging for estimation?
3.3.20 How does parameter σ affect estimation in Figure 3.4?
3.3.21 Using Figure 3.4, propose optimal arguments for the E-estimator for each corner function.
3.3.22 Prove equality (3.3.18), and then explain how it may be used for estimation of the regression function.

3.4.1 Present several examples of grouped observations. Explain the used terminology.
3.4.2 Use Google and find definitions and applications of grouped observations referred to as strata, categories and clusters.
3.4.3 What is the difference between ordinal and nominal categories?
3.4.4 Consider a regression with grouped responses. Can smaller regression errors improve regression estimation?

3.4.5 Suppose that an underlying regression function is a constant θ. Explain why the regression noise can help in estimation of θ.
3.4.6 For the setting of Exercise 3.4.5, explain how one can estimate parameter θ.
3.4.7 Explain the three-step procedure of regression estimation for the case of grouped responses.
3.4.8 How are the grouped observations shown in the left column of the diagrams in Figure 3.6 obtained?
3.4.9 Explain the two types of estimates shown in Figure 3.6.
3.4.10 Using Figure 3.6, create several different groups that make estimation less and more complicated. Explain your results.
3.4.11 Explain how the noise affects the estimation. Use the argument *sigma* of Figure 3.6 to support your conclusion.
3.4.12 Explain the estimator (3.4.3).
3.4.13* Find the expectation and the variance of the estimator (3.4.3).
3.4.14* Find the expectation and the variance of the estimator (3.4.4).
3.4.15 What is the role of parameters a and b in the estimator (3.4.4)?
3.4.16 Using Figure 3.6, explore the effect of parameters a and b on the proposed estimator.
3.4.17* Propose better parameters of the E-estimator used in Figure 3.6. Then explore how the underlying regression function affects the choice. Present empirical and theoretical justifications.

3.5.1 Explain the notion of data modification via mixture.
3.5.2 Present an example which corresponds to model (3.5.2).
3.5.3 How is the model (3.5.2) related to the change-point problem?
3.5.4 Consider a setting where the underlying regression is constant. Write down the corresponding model and propose a procedure of estimation of that constant.
3.5.5* Verify formula (3.5.5). Formulate all necessary assumptions.
3.5.6* Explain the proposed regression E-estimator. Then calculate its coefficient of difficulty.
3.5.7 Knowing the underlying sampling mechanism, explain the scattergrams shown in the diagrams of Figure 3.7. Then repeat the simulation and compare outcomes.
3.5.8* Typically one can visualize (guess about) an underlying regression function as a curve going through the middle of a scattergram. Why is this no longer the case for the mixture regression? Hint: Diagrams in Figure 3.7 may be helpful.
3.5.9 Repeat Figure 3.7 with different corner functions and different sample sizes. What is your conclusion about quality of estimation?
3.5.10* How do distributions of ζ and ξ affect the estimation? Suggest a theoretical explanation and then use Figure 3.7 to check the answer.
3.5.11* It is assumed that means of the random variables ζ and ξ are different. Suppose that they are the same but the corresponding variances are different. Propose an E-estimator for this case.

3.6.1 Explain the heteroscedastic regression model (3.6.1). Presenting several practical examples when using this model may be appropriate.
3.6.2 Assume that in model (3.6.1) the function of interest is the scale $\sigma(x)$. What are the nuisance functions?
3.6.3 Verify equality (3.6.3). Formulate used assumptions.
3.6.4 Does (3.6.3) hold if X and ε are dependent?
3.6.5* Explain how the E-estimator of $\sigma^2(x)$ can be constructed. Then find its coefficient of difficulty.
3.6.6 Explain model (3.6.4).
3.6.7 Explain all steps in construction of the E-estimator $\hat{\sigma}(x)$. Then calculate its coefficient of difficulty. Does it depend on the underlying regression function? Explain your answer.

3.6.8* Consider a sampling from hidden X^* where a realization of X^* is observed only if $X^* \le C$ and C is a continuous random variable. We are interested in estimation of the density of X^*. Do we have nuisance functions here?
3.6.9* Consider a hidden sampling (X_l^*, Y_l^*, T^*), $l = 1, 2, \dots$ where its lth realization is observed only if $Y_l^* > T_l^*$. We are interested in estimation of the regression of Y^* on X^*. Develop the probability model for observed triplets and explore nuisance functions needed for construction a regression E-estimator.
3.6.10 Explain plots shown in Figure 3.8.
3.6.11 In the top diagram of Figure 3.8 the regression estimate (the dot-dashed line) is clearly bad. Nonetheless, the scale estimate is reasonable. Can you explain this outcome?
3.6.12* In a regression estimation problem, larger regression errors typically imply worse estimation. Is this also the case for the scale estimation?
3.6.13 In the bottom diagram of Figure 3.8 the regression estimate (the dot-dashed line) is far from the underlying Bimodal regression function shown by the dotted line. Nonetheless the scale estimate is relatively good. This is due to relatively large scale function. Repeat Figure 3.8 with smaller values of parameter σ and write a report on how this parameter affects estimation of the scale function. Hint: Pay attention to the fact that smaller scale functions improve regression estimation and, at the same time, may make estimation of the scale function more complicated.
3.6.14* Suppose that X and ε in model (3.6.1) are dependent. Propose an E-estimator of the scale function. Hint: Make necessary assumptions.

3.7.1 Explain the model of Bernoulli regression with unavailable failures.
3.7.2 Explain the equality in (3.7.1).
3.7.3 Verify relations (3.7.2).
3.7.4 Explain formula (3.7.3). How can it be used for estimation of the Fourier coefficient θ_j?
3.7.5* Consider a density $f^Z(z)$ of a random variable Z. Is it possible that $f^Z(Z) = 0$? Then use your answer to explain when the design density $f^{X^*}(X_l)$ can be used in the denominator of (3.7.4).
3.7.6* Evaluate the mean and the variance of the Fourier estimator (3.7.4). Explain the used assumptions.
3.7.7 What is an assumption needed for consistency of the plug-in estimator (3.7.4)?
3.7.8 The number N of available predictors is introduced in (3.7.5). What is the distribution of N?
3.7.9 Calculate the mean and standard deviation of the number N of available predictors.
3.7.10* Write down an exponential inequality for the probability $\mathbb{P}(|N/n - \mathbb{P}(A^* = 1)| > t)$ where t is a positive constant. Further, what is the probability of the event $N = 0$?
3.7.11 Explain each equality in (3.7.6).
3.7.12 Are the observed predictors biased with respect to hidden predictors? Explain your answer and, if it is positive, point upon a biasing function.
3.7.13 Is it always possible to propose consistent estimation of the regression function in the Bernoulli regression with missing failures? Explain your answer.
3.7.14 Explain the idea of using an additional (extra) sample from the hidden predictor to estimate the regression function.
3.7.15 Why may it be important to bound the design density E-estimator from zero?
3.7.16 Repeat Figure 3.9 several times and explain shapes of the estimates via analysis of available data.
3.7.17 Why is, in Figure 3.9, the available number N of predictors small with respect to n?
3.7.18 What changes in the experiment of Figure 3.9 will increase the number N of available predictors?

3.7.19 What type of design densities make the regression estimation simpler or more complicated? Check your answers using Figure 3.9.
3.7.20 Figure 3.9 allows us to use different parameters of the regression E-estimator in each row. Use this feature to propose optimal parameters of the estimator.
3.7.21 How does the argument *dden* in Figure 3.9 affect the estimation? Present a heuristic answer and then test it via simulations.
3.7.22* Explain the E-estimator used in Figure 3.10 for the case of a design density with unknown support. Then calculate its coefficient of difficulty.
3.7.23 Repeat Figure 3.10 and then present analysis of obtained estimates.
3.7.24 Test the effect of the argument c of Figure 3.10 on the E-estimator.
3.7.25 How does parameter σ of the normal design density, used in Figure 3.10, affect estimation of the regression function?
3.7.26 Using Figure 3.10, what minimal size k of the extra E-sample would you recommend for a practitioner? Consider several n.
3.7.27 Write down a formula for the rescaled onto $[0, 1]$ design density with unknown support. Hint: Recall the scale-location transformation and how it affects the density. It also may be helpful to recall that the classical z-scoring is used to transfer a normal random variable into a standard normal variable.
3.7.28 In the proposed methodology of rescaling predictors onto the unit interval, the density estimate is divided by $(X_L - X_S)$. On the other hand, no transformation of the regression estimate is mentioned. Is this a mistake and should it be also rescaled? Explain your answer.
3.7.29* Let us complement the proposed approach of regression estimation by another idea of estimation of the Bernoulli regression function. According to (3.7.6) we can write down the regression function as

$$w(x) = \mathbb{P}(A^* = 1)\frac{f^X(x)}{f^{X^*}(x)}. \tag{3.8.4}$$

We know how to estimate the two densities on the right side of (3.8.4). Then suggest a sample mean estimator of $\mathbb{P}(A^* = 1)$, and propose a new estimator of the regression function.

3.9 Notes

3.1 Biased data is a familiar topic in statistical literature, see a discussion in Efromovich (1999a), Comte and Rebafka (2016) and Borrajo et al. (2017). A review of possible estimators may be found in the book by Wand and Jones (1995). Efromovich (2004a) has proved that E-estimation methodology is asymptotically efficient, and then a plug-in estimator of the cumulative distribution function is even second-order efficient, see a discussion and the proof in Efromovich (2004c). A combination of biased and other modifications is also popular in the literature, see Luo and Tsai (2009), Brunel et al. (2009), Ning et al. (2010) and Chan (2013).
3.2-3.3 Biased responses commonly occur in technological, actuarial, biomedical, epidemiological, financial and social studies. In a response-biased sampling, observations are taken according to the values of the responses. For instance, in a study of possible dependence of levels of hypertension (response) on intake of a new medicine (covariate), sampling from patients in a hospital is response-biased with respect to a general population of people with hypertension. Another familiar example of sampling with selection bias in economic and social studies is that the wage is only observed for the employed people. An interesting discussion may be found in Gill, Vardi and Wellner (1988), Bickel and Ritov (1991), Wang (1995), Lawless et al. (1999), Luo and Tsai (2009), Tsai (2009), Ning, Qin and Shen (2010), Chaubey et al. (2017), Kou and Liu (2017), Qin (2017) and Shen et al. (2017).

3.4 Nonparametric estimation for ordered categorical data is considered in the books Simonoff (1996) and Efromovich (1999a). Efromovich (1996a) presents asymptotic justification of E-estimation.

3.5 A discussion of parametric mixture models can be found in the book by Lehmann and Casella (1998). Nonparametric models are discussed in books by Prakasa Rao (1983) and Efromovich (1999a). The asymptotic justification of the E-estimation is given in Efromovich (1996a). For possible further developments see Chen et al. (2016).

3.6 Asymptotic justification of using E-estimation for nuisance functions may be found in Efromovich (1996a; 2004b; 2007a,f).