NONPARAMETRIC REGRESSION WITH MISSING DATA: THEORY AND APPLICATIONS

Sam Efromovich [1]

Department of Mathematical Sciences

The University of Texas at Dallas, Richardson, Texas, USA

Abstract

Nonparametric regression tries to find a relationship between the predictor and the response without assuming shape of estimated regression function. Its theory and methods are well developed for the case of completely observed vectors of predictors and responses. In many applications, including actuarial, some components of the observed vectors may be missed. Ignoring missing data and using known methods of nonparametric regression may yield inconsistent estimation. Theoretical results on optimal nonparametric regression with missing data are presented, and an applied actuarial example is discussed.

1 INTRODUCTION

Consider a classical heteroscedastic regression model

$$Y = m(X) + \sigma(X)\eta \tag{1.1}$$

where $m(x)$ is the regression function which should be estimated based on a sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of size n from (X, Y), $\sigma(x)$ is an unknown positive scale function, $\eta_1, \eta_2, \ldots, \eta_n$ are independent zero mean and unit variance errors (random variables) which may have different distributions, $X \in [0, 1]$ is the predictor and Y is the response. It is explicitly assumed that the predictor X and the error η are independent. The quality of

[1]Address correspondence to Sam Efromovich, Department of Mathematical Sciences, UTDallas, TX 75080, USA; E-mail: efrom@utdallas.edu

estimation of a regression function is defined by the mean integrated squared error (MISE). For this classical setting sharp-minimax theory of nonparametric regression estimation is well developed, see a discussion in Efromovich (1986,1999,2000) and Efromovich and Pinsker (1996).

This paper considers a more complicated setting when some pairs of observations may not be complete — the so-called situation of missing data. Many examples of missing data and an interesting discussion of the topic can be found in Little and Rubin (2002), Nittner (2003), Chen et al. (2006), Tsiatis (2006), Liang et al. (2007), Mollenberghs and Kenward (2007), Wang and Chen (2009), and Efromovich (2011).

In this paper we restrict our attention to two classical missing mechanisms. (i) Responses are missed at random. Under the missing at random (MAR) mechanism, observations are triplets $\{(X_l, \delta_l Y_l, \delta_l),\ l = 1, 2, \ldots, n\}$ where δ_l is a Bernoulli random variable with

$$\Pr(\delta = 1 | X = x, Y = y) = h(x). \tag{1.2}$$

Function $h(x)$, called the conditional probability of observing the response given the predictor, defines the missing mechanism and in general it is unknown. (ii) Predictors are missed at random. Under the missing at random (MAR) mechanism, observations are triplets $\{(\delta_l X_l, Y_l, \delta_l),\ l = 1, 2, \ldots, n\}$ where δ_l is a Bernoulli random variable with $\Pr(\delta = 1 | X = x, Y = y) = h(y)$.

To understand possible complication that may be caused by missed data, let us consider, as an example, the former case of missed responses. If anyone would like to deal only with complete pairs of observations (as a majority of statistical softwares does by default), then it is important to stress that, if $f^{X,Y}(x, y)$ is the joint density of the predictor and the response, then the conditional joint density of the complete pair predictor-response is biased because

$$f^{X,\delta Y|\delta}(x, y|1) = \frac{h(x)}{\int_0^1 h(u)p(u)du} f^{X,Y}(x, y). \tag{1.3}$$

This is what make the problem complicated.

Let us explain the issue of biased distributions and its connection to missed data via an example and its discussion. This also will be a good technical introduction to the topic of missed data.

Biased Distributions. Suppose that we observe a sample $Y_1, Y_2, \ldots, Y_n$ from a random variable Y with the probability density

$$f^Y(y) := \frac{g(y)}{\int g(z) f^X(z) dz} f^X(y),$$

where $g(y)$ is a given positive function and neither probability density $f^Y(y)$ nor the probability density $f^X(x)$ are known. The problem is to estimate the density f^X which is the probability density of interest. Then the sample $Y_1, \ldots, Y_n$, as well as the corresponding distribution, is called *biased* because the density (1.3) of available observations is different from the density of interest.

Let us present an example which sheds light on the problem of biased data.

Example 1. Suppose that a researcher would like to know the distribution of the ratio of alcohol in the blood of liquor-intoxicated drivers traveling along a particular highway. The data are available from routine police reports on arrested drivers charged with driving under the influence of alcohol. Because a drunker driver has a larger chance of attracting the attention of the police, it is clear that the data are biased toward higher ratios of alcohol in the blood. Thus, the researcher should make an appropriate adjustment in a method of estimation of an underlying density of the ratio of alcohol in the blood of all intoxicated drivers.

Now let us consider the following question. How is the biased dataset created? It is created by a missing mechanism. Indeed, there is an underlying set $X_1, \ldots, X_k$ of ratios of alcohol in the blood of all k drivers traveling along the highway at the time when police is present. For the lth driver, let us introduce a Bernoulli random variable δ_l which is equal to 1 if the driver is stopped and arrested, and set $\delta_l = 0$ otherwise. Suppose that $\Pr(\delta = 1 | X = x) = g(x)$. Then the biased dataset is a subsample of Xs corresponding to

$\delta = 1$ from the missing dataset

$$A_M := \{(\delta_1 X_1, \delta_1), \ldots, (\delta_n X_n, \delta_n)\}.$$

The following is an interesting exercise that sheds light on involved mathematical formulae.

Exercise 1. Consider a missing dataset A_M and construct an empirical cdf based on complete cases when $\delta = 1$. Does it converge to the underlying cdf of interest $F^X(x)$? To answer the question, let us write down the empirical cdf based on complete cases,

$$\hat{F}(y) = \frac{\sum_{l=1}^n \delta_l I(X_l \leq y)}{\sum_{l=1}^n \delta_l}.$$

Then we can write,

$$\begin{aligned}
E\{\hat{F}(y)\} &= E\{E\{\hat{F}(x)|(\delta_1, \ldots, \delta_n)\}\} \\
&= E\Big\{\frac{\sum_{l=1}^n E\{\delta_l I(X_l \leq y)|\delta_l\}}{\sum_{l=1}^n \delta_l}\Big\} \\
&= E\Big\{E\{I(X \leq y)|\delta = 1\}\frac{\sum_{l=1}^n \delta_l}{\sum_{l=1}^n \delta_l}\Big\} \\
&= E\{I(X \leq y)|\delta = 1\} = F^{X|\delta}(y|1) \\
&= \int_{-\infty}^{y} f^{X|\delta}(u|1)du.
\end{aligned}$$

We conclude that the empirical cdf is unbiased estimate of $F^{X|\delta}(y|1)$. Further, by the Law of Large Numbers,

$$\begin{aligned}
&n^{-1}\sum_{l=1}^n \delta_l I(X_l \leq y) \xrightarrow{P} E\{\delta I(X \leq y)\} \\
&= E\{\delta F^{X|\delta}(y|1)\} = \mathrm{P}(\delta = 1)F^{X|\delta}(y|1)
\end{aligned}$$

and

$$n^{-1}\sum_{l=1}^n \delta_l \xrightarrow{P} E\{\delta\} = \Pr(\delta = 1).$$

We conclude that $\hat{F}(y) \xrightarrow{P} F^{X|\delta}(y|1)$.

A useful remark is due. Note that $F^X(y) = F^{X|\delta}(y|1)$ iff $g(y)$ is constant. Note that if the latter is correct then the missing mechanism is called Missing Completely at Random (MCAR), and then plainly random variables X and δ are independent.

Let us now return to the problem of density estimation based on biased data. We observe

$$Y^N := (Y_1, \ldots, Y_N), \quad N = \sum_{l=1}^{n} \delta_l \,,$$

or, using our "missing data" notation, we observe $\{(\delta_1 X_1, \delta_1), \ldots, (\delta_n X_n, \delta_n\}$.

Can we estimate the underlying density $f^X(x)$ without knowing $h(y) = \Pr(\delta = 1|X = y)$? Remember that

$$f^Y(y) := \frac{h(y)}{\int h(z) f^X(z) dz} f^X(y),$$

and then the obvious answer is "no."

However, if we can observe an auxiliary/lurking variable Z_l (like speed of a car) and

$$\Pr(\delta = 1|X, Z) = h^*(z),$$

then the answer is "yes" (in this case the missing mechanism is MAR!).

This ends our discussion of a biased distribution and its relation to missed data.

Now let us present an example of a regression setting that motivated the research.

Example 2. Fairness of using the credit score as an insurance rating variable for college students is a hot topic in actuarial science. One of the specific tasks is to understand how the *Grade Point Average* (GPA) can predict the *Credit Score* via using a nonparametric regression. Credit scores can be found on Internet. If asked by a lecturer to report these two variables, only a proportion of students will do this extra work and return surveys with GPA and credit score, while others will not even return surveys with their GPA. Question: Is it desirable/necessary to get GPA's of the others? In other words, if an underlying data contains missed responses, is it prudent to develop a nonparametric regression based on its complete-case subsample? As we shall see shortly, the answer depends on an underlying setting.

Finally, let us note that the main remedy for missed data, recommended in the literature, is imputation of missed values. Free software packages in R, like MICE (Multiple Imputation via Chained Equations), are available. Let us mention several popular methods of imputation: (a) Mean imputation (average of observed values); (b) Last observed in longitudinal studies; (c) Nearest neighbor; (d) Multiple imputation via generating missing observations based on finding estimates via imputation and then averaging the estimates.

The context of the paper is as follows. Section 2 is devoted to the case of missed responses. An interesting new case of auxiliary covariates is outlined in Section 3. The case of missed predictors is discussed in Section 4. Here an applied example is also presented.

2 SHARP MINIMAX REGRESSION WITH MISSED RESPONSES

The considered nonparametric regression model is (1.1), and the design density $p(x)$ of X is supported.

A missing at random (MAR) sample is generated by the triplet $(\delta Y, X, \delta)$ where δ is Bernoulli with $\Pr(\delta = 1|X = x, Y) =: h(x)$.

A classical Sobolev class of k-fold differentiable regression functions on the unit interval $[0, 1]$ will be considered,

$$\mathcal{H}(k, Q) := \{m(x) : \ m(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x), \sum_{j=0}^{\infty} [1 + (\pi j)^{2k}]\theta_j^2 \le Q\},$$

where $\{\varphi_0(x) = 1, \varphi_j(x) = 2^{1/2} \cos(\pi j x), j = 1, 2, \ldots\}$ is the cosine basis on $[0, 1]$, k is a positive integer and $0 < Q < \infty$.

Assumption 2.1. The nuisance functions $p(x)$, $\sigma(x)$ and $h(x)$ have bounded first derivatives on $[0, 1]$. Further, $p(x)$ and $h(x)$ are bounded below from zero on $[0, 1]$.

Assumption 2.2. The regression model is $Y = m(X) + \sigma(X)\eta$ where η is standard normal.

Set $s := \lfloor \ln\ln(n+20) \rfloor$, $q := \lfloor \ln(n+20) \rfloor$, $J := \lfloor n^{1/(2k+1)}/s \rfloor$.

Define a pivotal Sobolev class

$$\mathcal{F}(m_0, k, Q) := \Big\{ m(x) : \ m(x) = m_0(x) + \mu(x),\ \mu(x) := \sum_{j \geq J} \kappa_j \varphi_j(x),$$

$$\mu(x) \in \mathcal{H}(k, Q) \cap \{\psi : \ \sup_{x \in [0,1]} |\mu(x)| < q n^{-k/(2k+1)} \Big\}.$$

Note that if a regression function $m(x)$ belongs to the pivotal class then its "low-frequency" part is known exactly and the "high-frequency" part is known within the margin $qn^{-k/(2k+1)}$.

Theorem 2.1 (Lower Bound for MAR Data). *Under Assumptions 2.1 and 2.2*

$$\inf_{\check{m}} \sup_{m \in \mathcal{F}(m_0,k,Q)} E\{ \int_0^1 (\check{m}(x) - m(x))^2 dx \}$$

$$\geq n^{-2k/(2k+1)} Q^{1/(2k+1)} P(k, p, h, \sigma)(1 + o_n(1)),$$

where the infimum is taken over all oracle-estimators $\check{m}$ depending on: the SIMAR sample, the pivot $m_0(x)$, the underlying parameters (k, Q) of the Sobolev class $\mathcal{F}$, the design density $p(x)$, function $h(x) = \Pr(\delta = 1 | X = x)$ describing the missing mechanism, and the scale function $\sigma(x)$. Further,

$$P(k, p, h, \sigma)$$

$$:= \Big[\frac{k}{\pi(k+1)} \int_0^1 \frac{\sigma^2(x)}{p(x)h(x)} dx \Big]^{2k/(2k+1)} (2k+1)^{1/(2k+1)}.$$

Theorem 2.2 (Upper Bound for MAR Data). *Suppose that Assumption 1 holds as well as Assumption 2 which here can be relaxed to include any zero mean and unit variance regression error ϵ with the finite eighth moment. Then the estimator $\hat{m}(x; S_C(N))$, based on the complete-case subsample is sharp minimax and*

$$\sup_{m \in \mathcal{H}(k,Q)} E\{ \int_0^1 (\hat{m}(x; S_C(N)) - m(x))^2 dx \}$$

$$\leq n^{-2k/(2k+1)} Q^{1/(2k+1)} P(k, p, h, \sigma)(1 + o_n(1)).$$

There are two important conclusions from the asymptotic theory. The first one is that for the considered model complete case pairs allow to construct a sharp minimax nonparametric estimator. In other words, the strategy of standard statistical packages (on how to deal with missing data by ignoring incomplete pairs) is feasible.

Another important conclusion is that the optimal design density of predictors, minimizing the asymptotic MISE, is $p^*(x) = \sigma(x)h^{-1/2}(x)I(x \in [0,1])/\int_0^1 \sigma(u)h^{-1/2}(u)du$. Here $I(\cdot)$ is the indicator function. This result is intuitively clear. One should choose more predictors in areas with larger volatility (scale function) and smaller conditional probability of observing the response. Formula for p^* gives us a specific recipe of how to place predictors, and it may be an underlying idea of an optimal sequential design in a controlled nonparametric regression. Of course, the optimal design density depends on two nuisance functions, the scale $\sigma(x)$ and the conditional probability $h(x)$ of observing the response given the predictor, which are in general unknown. This is the place where a sequential design may shine. Namely, one can use a sequential design where the scale function and the conditional probability of observing the response are sequentially estimated, plugged in the formula for p^*, and then the obtained design density is used to generate next predictor. This procedure may compensate for the loss of information due to missed data. More details can be found in Efromovich (2011a).

3 REGRESSION WITH MISSING RESPONSE AND AUXILIARY COVARIATES

This is a very interesting setting where an underlying sample is from $(Y, X, \mathbf{Z})$, and we are interested in estimation of $m(x) = E(Y|X = x)$. Then the vector-covariate $\mathbf{Z}$ is called auxiliary.

The observed sample is from $(\delta Y, X, \mathbf{Z}, \delta)$ where $P(\delta = 1|Y, X, \mathbf{Z}) = h(X, \mathbf{Z})$. As a result,

the problem would be similar to the above-considered if we had been interested in $E(Y|X, \mathbf{Z})$. But for our setting

$$P(\delta = 1|Y = y, X = x) = \int h(x, \mathbf{z}) f(\mathbf{z}|x, y) d\mathbf{z},$$

which may depend on y.

This is a setting where I will present some outlines of the asymptotic theory.

1. The underlying model can be written as

$$Y = m(X) + [m(X, \mathbf{Z}) - m(X)] + \sigma(X, \mathbf{Z})\eta$$

2. It suffices to know the conditional density $p^{\mathbf{Z}|X}(\mathbf{z}|x)$ and complete cases for sharp minimax estimation.

3. Sharp minimax constant is proportional to

$$\int \frac{1}{p(x) \int \mathcal{I}(x, \mathbf{z}) h(x, \mathbf{z}) p(\mathbf{z}|x) d\mathbf{z}} dx$$

4. A naive rate-optimal estimator has a constant of its MISE convergence proportional to

$$\int \frac{p(\mathbf{z}|x)}{p(x) \mathcal{I}(x, \mathbf{z}) h(x, \mathbf{z})} dx d\mathbf{z}.$$

This is a problem whose solution is in progress. Note that the complex structure of the constant indicates a rather complicated estimator required for an optimal solution.

4 Minimax Regression with Missed Predictors

This section is devoted to the case where predictors are missing at random. Namely, here in place of an underlying sample $S_U(n) = \{(X_1, Y_1, \delta_1), \ldots, (X_n, Y_n, \delta_n)\}$ the observed sample is $S_{MAR}(n) = \{(\delta_1 X_1, Y_1, \delta_1), \ldots, (\delta_n X_n, Y_n, \delta_n)\}$ from $(\delta X, Y, \delta)$ where δ is Bernoulli with $\Pr(\delta = 1|Y, X) = \Pr(\delta = 1|Y) =: h_Y(y)$.

The aim is to estimate the regression function $m(x) = E\{Y|X = x\}$.

Assumption 4.1. The underlying joint density of the pair (X, Y) of the predictor and response is $p(x)f(y|x)$ which is supported on $[0,1] \times \mathcal{Y}$. Further, $f(y|x) = f_{m(x)}(y)$ where $m(x) := E(Y|X = x) = \int_{\mathcal{Y}} yf(y|x)dy$ is a bounded regression function on $[0,1]$.

Assumption 4.1 allows us to formulate all necessary restrictions on the conditional density $f(y|x) = f_{m(x)}(y)$ via its parametric counterpart $f_\theta(y)$. The following regularity conditions are similar ones in classical uniform LAN.

Assumption 4.2. Consider a parametric family of probability densities $\{f_\theta(y),\ \theta \in \Theta,\ y \in \mathcal{Y}\}$ where Θ is an open interval on the real line. Density f_θ is twice-differentiable in θ for all $\theta \in \Theta$, and

$$\mathcal{I}_j(\theta) := \int_{\mathcal{Y}} (\partial^j f_\theta(y)/\partial\theta^j)^2 f_\theta^{-1}(y)dy,\ j = 1, 2,$$

$\mathcal{I}_3(\theta) := \int_{\mathcal{Y}} [\partial f_\theta(y)/\partial\theta]^4 [f_\theta(y)]^{-3}dy$ are uniformly bounded for all $\theta \in \Theta$, and $\mathcal{I}_1(\theta)$ is positive on Θ.

Theorem 4.1 (Lower Bound for MAR Data). *Let Assumptions 2.1, 4.1 and 4.2 hold. Suppose that $m_0(x)$ has a bounded derivative on $[0,1]$ and its range is a subset of Θ. Then*

$$\inf_{\check{m}} \sup_{m \in \mathcal{F}(m_0,k,Q)} E\{\int_0^1 (\check{m}(x) - m(x))^2 dx$$
$$\geq n^{-2k/(2k+1)} Q^{1/(2k+1)} P_{MAR}(k,d)(1 + o_n(1)),$$

where

$$P_{MAR}(k,d) := \Big[\frac{k}{\pi(k+1)}d\Big]^{2k/(2k+1)} (2k+1)^{1/(2k+1)},$$

and

$$d := \int_0^1 [p(x) \int_{\mathcal{Y}} [f'_{m_0(x)}(y)]^2 [f_{m_0(x)}(y)]^{-1} h_Y(y)dy]^{-1} dx.$$

Now let us present an oracle-estimator whose MISE attains the lower bound. Consider a complete-case subsample $S_C(N) := \{(\tilde{X}_l, \tilde{Y}_l),\ \ l = 1, \dots, N\}$. Set $\hat{J} := \lfloor (N+s)^{1/(2k+1)} \rfloor$,

and define the oracle-estimator:

$$\hat{m}(x; S_C(N), n, p, h_Y) := \sum_{j=0}^{\hat{J}} \hat{\theta}_j(S_C(N), n, p, h_Y)\varphi_j(x),$$

where

$$\hat{\theta}_j(S_C(N), n, p, h_Y) := n^{-1} \sum_{l=1}^{N} \frac{\tilde{Y}_l \varphi_j(\tilde{X}_l)}{p(\tilde{X}_l) h_Y(\tilde{Y}_l)}.$$

Theorem 4.2 (Upper Bound for Oracle-Estimator). *Let* $\{h_Y : h_Y(y) = Hh^*(y), \sup h(y) = 1, H \in (0,1]\}$. *Let* $h^*(y)$ *and* $p(x)$ *be bounded below from zero on* $\mathcal{Y}$ *and* $[0,1]$, *respectively. Let* $f(y|x)$ *and* $p(x)$ *be bounded on* $\mathcal{Y} \times [0,1]$ *and* $[0,1]$, *respectively, and suppose that* $E(Y^2) < \infty$. *Then for* $m \in \mathcal{H}(k, Q)$

$$E\{\int_0^1 (\hat{m}(x; S_C(N), n, p, h_Y) - m(x))^2 dx$$
$$\le C_*[Hn]^{-2k/(2k+1)}(1 + o_n(1)),$$

where a finite C_* *does not depend on the parameter* H *and* $o_n(1)$ *may depend on* H.

Let us define an estimator that mimics the oracle. Set $J_2 := \lfloor sn^{1/3} \rfloor$ and introduce estimates $\hat{h}$ and $\hat{p}$ based on the complete-case subsample $S_C(N) = \{(\tilde{X}_l, \tilde{Y}_l), l = 1, \dots, N\}$ and $f_Y(y)$:

$$\hat{h}_Y(y) := \min(1/s, \tilde{h}_Y(y)),$$
$$\tilde{h}_Y(y) := n^{-1} \sum_{j=0}^{J_2} \sum_{l=1}^{N} \varphi_j(\tilde{Y}_l)\varphi_j(y)/f_Y(\tilde{Y}_l),$$

and

$$\hat{p}(x) := \min(1/s, \tilde{p}(x)),$$
$$\tilde{p}(x) := n^{-1} \sum_{j=0}^{J_2} \sum_{l=1}^{N} \varphi_j(\tilde{X}_l)\varphi_j(x)/\hat{h}_Y(\tilde{Y}_l).$$

If the marginal density f_Y of the response is unknown then two natural approaches can be recommended for its estimation. The first one is to use additional independent observations $Y'_1, \dots, Y'_{n'}$ of the response with n' being proportional to n, and then set

$$\hat{f}_Y(y) := \min(1/s, \tilde{f}_Y),$$

$$\tilde{f}_Y(y) := (1/n') \sum_{j=0}^{(sn')^{1/3}} \sum_{l=1}^{n'} \varphi_j(Y'_l)\varphi_j(y).$$

For the actuarial example this approach implies a survey of GPAs in similar classes.

The second approach is to use the underlying MAR sample. This approach is definitely more appealing if the MAR sample is available.

Theorem 4.3 (Upper Bound for the Plug-In Oracle-Estimator). *Suppose that $f_Y(y)$, $p(x)$ and $h^*_Y(y)$ are bounded below from zero on $[0,1]$, $p(x)$ and $h^*_Y(y)$ have bounded first derivatives on $[0,1]$, and first-order partial derivatives of $f(y|x)$ in x and y are bounded on $[0,1]^2$. Then for $m \in \mathcal{H}(k,Q)$*

$$E\{\int_0^1 (\hat{m}(x; S_C(N), \hat{p}, \hat{h}_Y) - m(x))^2 dx$$

$$\le C_*[Hn]^{-2k/(2k+1)}(1 + o_n(1)),$$

where a finite C_* does not depend on the parameter H and $o_n(1)$ may depend on H.

Now let us consider application of the proposed methodology for the analysis of a real data with missing predictors. Fairness of using the credit score as an insurance rating variable is a hot topic in actuarial science; see an interesting discussion and further references in Brockett and Golden (2007). Under a grant from the actuarial foundation, the author explored a specific task of understanding how the credit score, here the explanatory variable X, can predict the grade point average (GPA), here the response Y. To obtain data, students taking the same class were asked to get a credit score on Internet and then report it unanimously together with GPA. The top diagram in Figure 1 exhibits the collected data $\{(\delta_l X_l, Y_l),\ l = 1, 2, \ldots, 147\}$ rescaled onto $[0,1]^2$; here only 92 students provided credit scores. The dashed line shows the linear regression based on complete pairs; as we know it may be biased if the missing mechanism is not MCAR. The linear regression supports the well-accepted believe in the insurance industry that the higher the credit score, the better the grade.

Now, just for a moment, let us look at the bottom diagram which exhibits estimates of the nuisance functions. The main one is the estimate of the conditional probability $h(y)$ of

not-missing the credit score shown by the solid line. It indicates the MAR (and not MCAR) nature of the missing data. The dashed line shows the estimate of the design density – it is uniform. The dotted line shows the estimate of the scale function multiplied by factor 3. We see a modest heteroscedasticity which indicates that there is a larger volatility of grades among students with lower credit scores.

Let us return to the top diagram. The solid line exhibits the proposed data-driven nonparametric estimator. Similarly to the linear regression, it exhibits a monotonic relationship between the credit score and the GPA. At the same time, the nonparametric estimate indicates that average GPAs are always smaller than the ones predicted by the linear regression. This is the reflection of the missing mechanism highlighted in the bottom diagram. Furthermore, both the nonparametric confidence band and the band with Bonferroni correction indicate that the linear regression, shown by the dashed line, is questionable. Furthermore, it is easy to see that the nonparametric estimate (the solid line) is not the center of the bands; instead, as it was explained in Section 3.1, here an undersmoothed estimate was used as the central line.

More details about the setting can be found in Efromovich (2011b).

REFERENCES

Brockett, P.L., and Golden, L.L. (2007). Biological and Psychobehavioral Correlates of Credit Scores and Automobile Insurance Losses: Toward an Explication of Why Credit Scoring Works *Journal of Risk and Insurance*, 74: 23-63.

Chen, J., Fan, J., Li, K. and Zhou, H. (2006). Local Quasi-Likelihood Estimation with Data Missing at Random, *Statistical Sinica* 16: 1071-1100.

Efromovich, S. (1986). Adaptive Algorithm of Nonparametric Regression, *Procedures of Second IFAC Symposium on Stochastic Control, Science* 1: 112-114.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications*, New York: Springer.

Efromovich, S. (2000). On Sharp Adaptive Estimation of Multivariate Curves, *Mathematical Methods of Statistics* 9: 117-139.

Efromovich, S. (2007). Sequential Design and Estimation in Heteroscedastic Regression, *Sequential Analysis* 26: 3-25.

Efromovich, S. (2008). Optimal Sequential Design in a Controlled Nonparametric Regression, *Scandinavian Journal of Statistics* 35: 266-285.

Efromovich, S. (2011a). Nonparametric Regression with Responses Missing at Random, *Journal of Statistical Planning and Inference* 141: 3744-3752.

Efromovich, S. (2011b). Nonparametric Regression with Predictors Missing at Random, *Journal of American Statistical Association* 106: 306-319

Efromovich S. and Pinsker, M.S. (1996). Sharp–Optimal and Adaptive Estimation for Heteroscedastic Regression. *Statist. Sinica* 6: 925–945.

Liang, H., Wang, S. and Carroll, R. (2007). Partially Linear Models with Missing Response Variables and Error-Prone Covariates *Biometrika*, 94: 185-198.

Liski, E.P., Mandal N.K., Shah, K.R. and Sinha, B.K. (2002). *Topics in Optimal Design*, New York: Springer.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data* 2nd ed., New York: Wiley

Marron, J.S. and Wand, M.P. (1992). Exact Mean Integrated Squared Error, *Annals of Statistics* 20: 712–736.

Molenberghs, G. and Kenward, M. (2007). *Missing Data in Clinical Studies*, New York: Wiley.

Nittner, T. (2003). Missing at Random (MAR) in Nonparametric Regression - a Simulation Experiment, *Statistical Methodology and Applications* 12: 195-210.

Pukelsheim, F. (1993). *Optimal Design of Experiments*, New York: Willey.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*, New York: Springer.

Wang, D. and Chen, S.X. (2009). Empirical Likelihood for Estimating Equations with

Missing Values, *Annlas of Statistics* 37: 490-517.

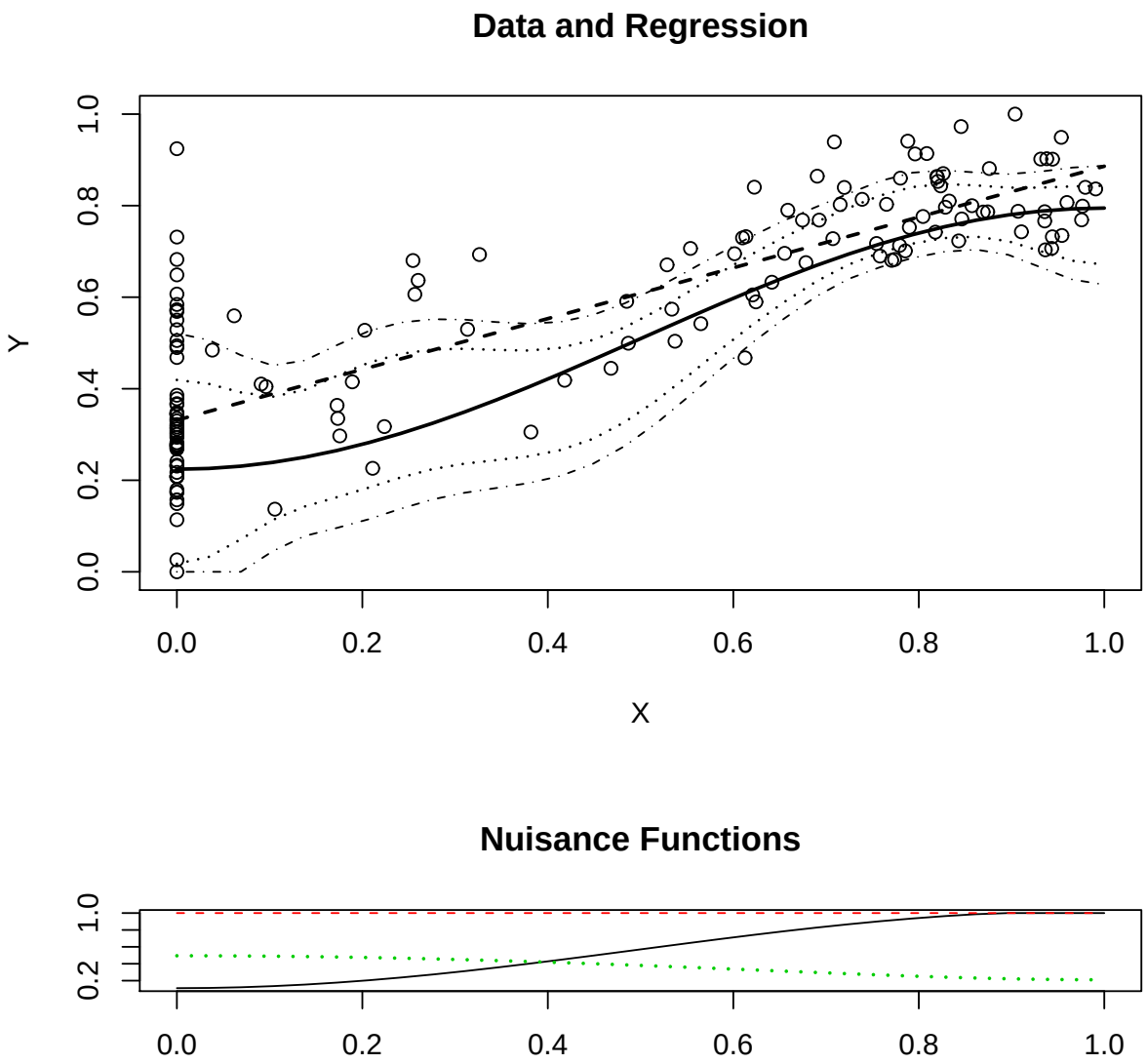

Figure 1: Analysis of a real data with missing predictor. X is the credit score, Y is the GPA, data are shown by circles, $n = 147$ and the number of complete pairs is 92. In the top diagram the solid, dashed, dotted and dashed-dotted lines are the nonparametric estimate, the linear regression based on complete pairs, 95% confidence band and 95% confidence band with Bonferroni correction, respectively. In the bottom diagram the solid, dashed and dotted lines are estimates of the conditional probability $h(y)$ of not-missing the predictor (the credit score), the design density $p(x)$, and the scale $\sigma(x)$ multiplied by factor 3, respectively.