이학석사학위논문

Generalized Partially Linear Models with missing data and its applications

결측 데이터를 포함한 일반화 부분 선형모형 및 적용

2014년 8월

서울대학교 대학원

통계학과

이 준 희

Generalized Partially Linear Models with missing data and its applications
결측 데이터를 포함한 일반화 부분 선형모형 및 적용

지도교수 Byung U. Park

이 논문을 이학석사 학위논문으로 제출함
2014년 6월

서울대학교 대학원
통계학과
이준희

이준희의 이학석사 학위논문을 인준함
2014년 6월

위원장	Myunghee Cho Paik	(인)
부위원장	박병욱	(인)
위원	장원철	(인)

Abstract

Junhee Lee
The Department of Statistics
The Graduate School
Seoul National University

In multivariate data with binary outcomes, it can be difficult to apply generalized linear models when an explanatory variable is missing. Some covariates behave linearly and the others behave nonlinearly when having an effect on the outcome variable, which is the variable to be expected statistically. In this case, one of the ways to deal with a missing explanatory variable is with generalized partially linear models, which consider the aspects of both generalized linear models and generalized additive models. In particular, when data is missing, the WEE (weighted estimating equation) method can be employed in order to estimate the function of interest.

It is interesting that the probability of missing covariates does not affect the biasedness of the estimator. This is because the inverse probability weighting method gives weights to other data near the value of the missing data to complement the missing portion. Also, we apply the kernel regression to estimate the nonparametric term. Since the bandwidth h is necessary, we also check how critical bandwidth selection is by making several simulations and computing the biasedness of estimates.

Contents

Chapter 1

Introduction

For binary outcomes with several covariates, the regression may have both parametric and nonparametric explanatory variables. That is, we assume that some of covariates affect expected outcome in linear terms, and others affect expected outcomes in nonlinear terms, which do which is unknown. In this paper I would like to explain the generalized partially linear model, which is an important result of the efforts that were made to balance the interpretation of GLMs and the flexibility of generalized additive models.

Most approaches to studying generalized partially linear models with missing covariates are used only for observed data. As examples of “complete case analysis,” they exclude the missing data. As a result, the estimator becomes inefficient, and a great deal of information is lost.

There are four approaches in the literature on GLMs with missing covariates: (1) maximum likelihood, (2) multiple imputation, (3) Bayesian, and (4) the weighted estimating equation (WEE). In this paper, the use of the WEE method in GPLMs with missing covariates is examined. Robins et al. developed an efficient WEE method for general parametric models. Wang et al. considered estimation of parametric regression coefficients with unknown missing probability. Liang et al. extended the use of the WEE method to the case of partially linear models (PLMs). Now, our aim is to extend the WEE method to GPLMs with missing covariates. The themes we are concerned with here are different from the common methods for partially linear models in several aspects: (1) We do not have a closed form of the estimator as we do in PLMs. (2) A numerical iteration procedure is needed. (3) Only the WEE can be applied under nonparametric covariates, among four approaches to GLMs.

The article is organized as follows: In chapter 2, we formally introduce the model framework, propose an estimation algorithm, parametrically and nonparametrically consider estimation of the missing probability, correspondingly decide the strategy for the estimation of the parameter of interest, and derive the asymptotic distributions of the estimators. We illustrate the methods with simulation experiments in Section 3, and provide a discussion in Section 4.

Chapter 2

Statistical Methods

2-1. GPLM Model

There are n i.i.d observations $\{(X_i, Z_i, Y_i),\ i = 1, \dots, n\}$ where X_i are the linear covariates, Z_i are the nonlinear scalars, and Y_i are the outcome variables.

The generalized partially linear model (GPLM) is as follow:

$$\mathrm{E}(Y_i|X_i, Z_i) = \mu\{{X_i}^T\beta + \theta(Z_i)\},$$

where μ is a link function, β is a vector, and θ is an unknown smoothing function.

Assume $\mathrm{var}(\mathrm{Y}|\mathrm{X}, \mathrm{Z}) = \sigma^2 V(\mu)$, where $\mu = \mu\{X^T\beta + \theta(Z)\}$.

$\delta = 1$ if X is observed and $\delta = 0$ otherwise.

Assume that the X's are missing at random(MAR) in the sense that

$$\pi(Y_i, Z_i) = P(\delta_i = 1|X_i, Z_i, Y_i) = P(\delta_i = 1|\ Z_i, Y_i)$$

2-2. Estimation

Under the generalized partially linear models (GPLM)

$$\mathrm{E}(Y_i|X_i, Z_i) = \mu\{{X_i}^T\beta + \theta(Z_i)\},$$

we will consider several estimators of β : $\widehat{\beta_n}$, $\widehat{\beta_{CC}}$, $\widehat{\beta_{glm}}$, $\widehat{\beta_{n,P}}$, which are obtained under GLM, complete case analysis, parametric model of π, and benchmark estimate(covariates are measured without missingness)

To begin the numerical model, we define several more notations. $\rho_k(t) =$

$\{d\mu(t)/dt\}^k V^{-1}\{\mu(t)\}$ for $\mathrm{k} = 1, 2$, $\Lambda_i = \{1,\ (Z_i - z_0)/h, X_i{}^T\}$

Approximate $\theta(\mathrm{z})$ via Taylor expansion:

$$\theta(\mathrm{z}) = \theta(z_0) + \theta'(z_0)(z - z_0)$$

Then, an iterative algorithm is used to approximate β with several steps following:

(Step 0) Fit a generalized linear model to obtain an initial value $\hat{\beta}$.

(Step 1) For each fixed z_0 and $\hat{\beta}$, $\hat{\theta}(\hat{\beta}, z_0)$ denotes the solution in a_0 of the equation

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(z_0 - Z_i)\Lambda_i\rho_1\left\{\Lambda_i{}^T \times \left(a^T, \hat{\beta}^T\right)^T\right\}[Y_i - \mu\left\{\Lambda_i{}^T \times \left(a^T, \hat{\beta}^T\right)^T\right\}]$$

where $K_h(t) = \frac{K(\frac{t}{h})}{h}$, $\mathrm{K(t)}$ is a kernel function, h is a bandwidth.

(Step 2) Given the estimator $\hat{\theta}\left(\hat{\beta}, z\right)$, and an estimator of β, $\hat{\beta}$ is updated by solving the equation

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}X_i\rho_1\{\hat{\theta}\left(\hat{\beta}, Z_i\right) + X_i{}^T\beta\}[Y_i - \mu\{\hat{\theta}\left(\hat{\beta}, Z_i\right) + X_i{}^T\beta\}]$$

(Step 3) Repeat step1 and 2 until convergence

For convenience, we make some assumption from now on.

<Assumptions>

(a) The function $q_2(t, y) < 0\ for\ t \in (-\infty, \infty)$ and for y in the range of the response variable.

(b) The density function $\mathrm{f(z)}$ of Z is positive and continuous at the point z_0.

(c) The functions $\theta(\cdot)$ and $\theta^2(\cdot)$ are continuous at the point z_0.

(d) $\mathrm{E}\{q_1{}^2(R, Y)|z\}$, $\mathrm{E}\{q_1{}^2(R, Y)X|z\}$, and $\mathrm{E}\{q_1{}^2(R, Y)XX^T|z_0\}$ are twice differentiable in z.

(e) $\mathrm{E}\{{q_2}^2(R,Y)\} < \infty$ and $\mathrm{E}\{{q_1}^{2+\delta}(R,Y)\} < \infty$, for some $\delta > 2$.

Theorem 1. Consider the iterative nonparametric estimator $\hat{\theta}(z)$. Then, as $n \rightarrow \infty$, $h \rightarrow 0$, and $nh \rightarrow \infty$, under the conditions given in <Assumptions>, we have the asymptotic expansion

$$\hat{\theta}(z) - \theta(z) = \frac{\kappa_2}{2}\theta^{(2)}(z)h^2$$

$$+\frac{1}{nf(z)\mathrm{A}(z)}\sum_{i=1}^{n} q_1(R_i,y_i)\mathrm{K}_h(Z_i - z) + o_p\left\{nh^{-1/2} + h^2\right\}$$

and hence

$$nh^{1/2}\left\{\hat{\theta}(z) - \theta(z) - \frac{\kappa_2}{2}\theta^{(2)}(z)h^2\right\} \overset{D}{\rightarrow} N\{0,\mu_0 f^{-1}(z)\mathrm{A}^{-1}(z)\}.$$

Theorem 2. Let $\hat{\beta}_n$ be the estimate obtained from step 2. Under the condition given in Appendix A, as $n \longrightarrow \infty$, $nh^4 \rightarrow 0$, and $nh^2/\log(\frac{1}{h}) \rightarrow \infty$, $\hat{\beta}_n$ is a consistent estimate and $n^{\frac{1}{2}}(\hat{\beta}_n - \beta)$ converges to a normal distribution with mean zero and covariance matrix:

$$[E\{\rho_2(R)\tilde{X}\tilde{X}^T\}]^{-1}cov(\frac{1}{\pi}\tilde{X}\epsilon)[E\{\rho_2(R)\tilde{X}\tilde{X}^T\}]^{-1}$$

In this thesis, we are going to generate two sets of $\{(X_i, Z_i, Y_i)\}$ with fixed $\theta(z)$, and control missing probability to see how bias and RMSE depends on missing probability. Also, with some fixed missing probability, the fitted value $\hat{\theta}(z)$ will be plotted, and compared with the real $\theta(z)$.

2-2. Bandwidth selection

The bandwidth of the kernel is a free parameter which exhibits a strong influence on the resulting estimate. The choice of kernel K is not crucial, but the choice of bandwidth h is important. The estimate is very sensitive to the choice of h. Small bandwidths give very rough estimates while larger bandwidths give smoother estimates. The bandwidth should be chosen optimally, in respect of minimizing the estimated MSE.

Theorem. Let $R_x = E(f(x) - \hat{f}(x))^2$ be the risk at a point x and let $\mathrm{R} = \int R_x \, dx$ denote the integrated risk. Assume that f'' is absolutely continuous and that $\int (f'''(x))^2 dx < \infty$. Also, assume that K satisfies the conditions of the kernel function. Then,

$$R_x = \frac{1}{4}\sigma_K{}^4 h_n{}^4 (f''(x))^2 + \frac{f(x)\int K^2(x)dx}{nh_n} + O\left(\frac{1}{n}\right) + O(h_n{}^6)$$

and

$$R = \frac{1}{4}\sigma_K{}^4 h_n{}^4 \int (f''(x))^2 dx + \frac{\int K^2(x)dx}{nh} + O\left(\frac{1}{n}\right) + O(h_n{}^6) \qquad \text{(2-2-1)}$$

where $\sigma_K{}^2 = \int x^2 K(x) dx$.

If we differentiate (2-2-1) with respect to h and set it equal to 0, we see that the asymptotically optimal bandwidth is

$$h_* = \left(\frac{c_2}{c_1{}^2 A(f) n}\right)^{1/5}$$

where $c_1 = \int x^2 K(x) dx$, $\int K(x)^2 dx$ and $A(f) = \int (f''(x))^2 dx$.

In practice, for smooth densities and a normal kernel,

$$h_n = \frac{1.06\hat{\sigma}}{n^{1/5}}. \quad \text{(Normal reference rule)} \qquad \text{(2-2-2)}$$

We will use this property later.

Chapter 3

Simulations

3.1. Controlling Missing Probability

$X \sim N(0, 0.25)$ and $Z \sim U(0,5)$, θ(z) = $\sin(z)$

$$P\{Y = 1 | X = x, Z = z\} := P(x, z) = \frac{exp(0.75x + \sin(z))}{1 + exp(0.75x + \sin(z))}$$

$$P\{X \ is \ missing | Y = y, Z = z\} := \pi(y, z) = \frac{exp(\gamma_1 + \gamma_2 y + \gamma_3 z)}{1 + exp(\gamma_1 + \gamma_2 y + \gamma_3 z)}$$

We tried to plot the relations between the average missing probability and the bias of β or the root mean squared error (RMSE). In detail, we generate n samples with π observing probability (That is, 1 – (missing probability)), and estimate $\hat{\beta}$ and $\hat{\theta}(z_0)$ and repeat this process 30 times. Then, 30 independent estimates are generated and we use them to study our theory.

By manipulating γ_i, we handle the overall rate of missing covariates.

Now we exhibit some results of estimation. In estimation, we choose the normal function, $\mathrm{K(t)} = \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})$, as the kernel function.

Comparing GPLM to GLM by the biasedness of $\mathrm{E(Y|X, Z)}$

To check how effectively GPLM be conducted on our data, which actually contains both parametric and nonparametric covariate part, we estimate β_{glm} under the assumption that X and Z are related to Y as linear terms, and β_{gplm} by assuming that X is a linear term and Z is a nonlinear term.

We compare cases from two different assumptions by estimating the bias of P, that is, $\frac{1}{n}\sum_i(\widehat{P}(Y_i|X_i,Z_i) - P(Y_i|X_i,Z_i))$, where

$$\widehat{P}_{glm}(Y_i|X_i,Z_i) = \frac{\exp(X_i\beta_{glm,1} + Z_i\beta_{glm,2})}{1 + \exp(X_i\beta_{glm,1} + Z_i\beta_{glm,2})}$$

in the GLM case and

$$\widehat{P}_{gplm}(Y_i|X_i,Z_i) = \frac{\exp(X_i\beta_{gplm} + \widehat{\theta}(Z_i))}{1 + \exp(X_i\beta_{gplm} + \widehat{\theta}(Z_i))}$$

in the GPLM case.

(a) Estimate parameters assuming X, Z are linear covariate. (GLM case)

Missing probability	$\beta_{glm,1}$	$\beta_{glm,2}$	(Bias of $\widehat{P}_{glm}$)*	(Bias of $\widehat{P}_{gplm}$)**
0.0	1.453865916	0.349263207	0.07820229	0.01825714
0.1	1.460649725	0.348901703	0.07856461	0.0588392
0.2	1.444921107	0.35084615	0.0777923	0.01362938
0.3	1.472080742	0.342613298	0.07853354	0.01681808
0.4	1.472763039	0.356677347	0.07868422	0.06758422
0.5	1.510790153	0.353829741	0.07891318	0.02278331

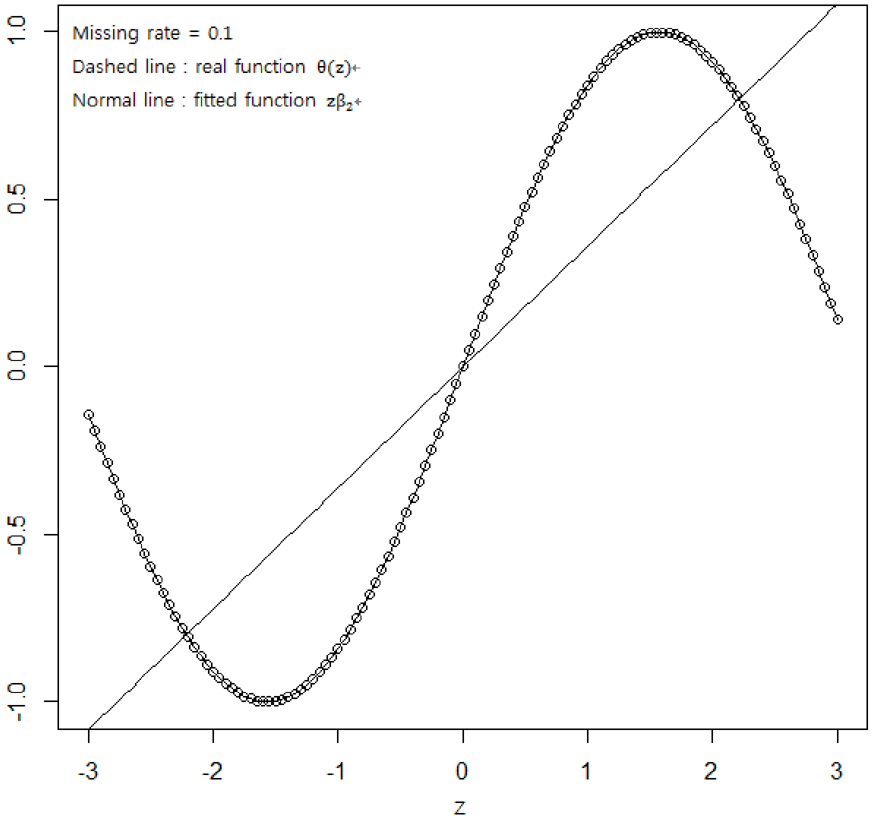
$^*\frac{1}{n}\sum\{\widehat{P}_{glm}(Y_i|X_i,Z_i) - P(Y_i|X_i,Z_i)\}$, $^{**}\frac{1}{\mathrm{n}}\sum\{\widehat{P}_{gplm}(Y_i|X_i,Z_i) - P(Y_i|X_i,Z_i)\}$

(Table 1: Comparing the bias of $\widehat{P}_{glm}$ and $\widehat{P}_{gplm}$)

To get reasonable estimates, we conducted simulations with fixed missing probability, 10 times independently for each value. Then, we get the average of those estimates. With this value, we compute the bias.

Seen from the table above, for any missing probability value, the average biasedness shows a smaller value in the case of GPLM, as compared to that of GLM. Therefore, we may conclude that as long as some covariates are related to outcome nonlinearly GPLM is more reliable than GLM as the statistical model.

(b) Plotting $\hat{\theta}(z)$.



(Figure 1 : Z - θ(z) plot: Dashed line represents the real $\theta(z)$, which is sin(z). and normal line represents the fitted function $\hat{\theta}(z)$), and since we used GLM model, it is zβ.)

(c) Computing the bias and the RMSE (Root Mean Squared Error)

$$(the\ bias\ of\ \beta) = \frac{1}{m}\sum_{1}^{m}(\hat{\beta} - \beta)$$

$$(the\ RMSE\ of\ \beta) = \frac{1}{m}\sum_{1}^{n}(\hat{\beta} - \beta_0)^2$$

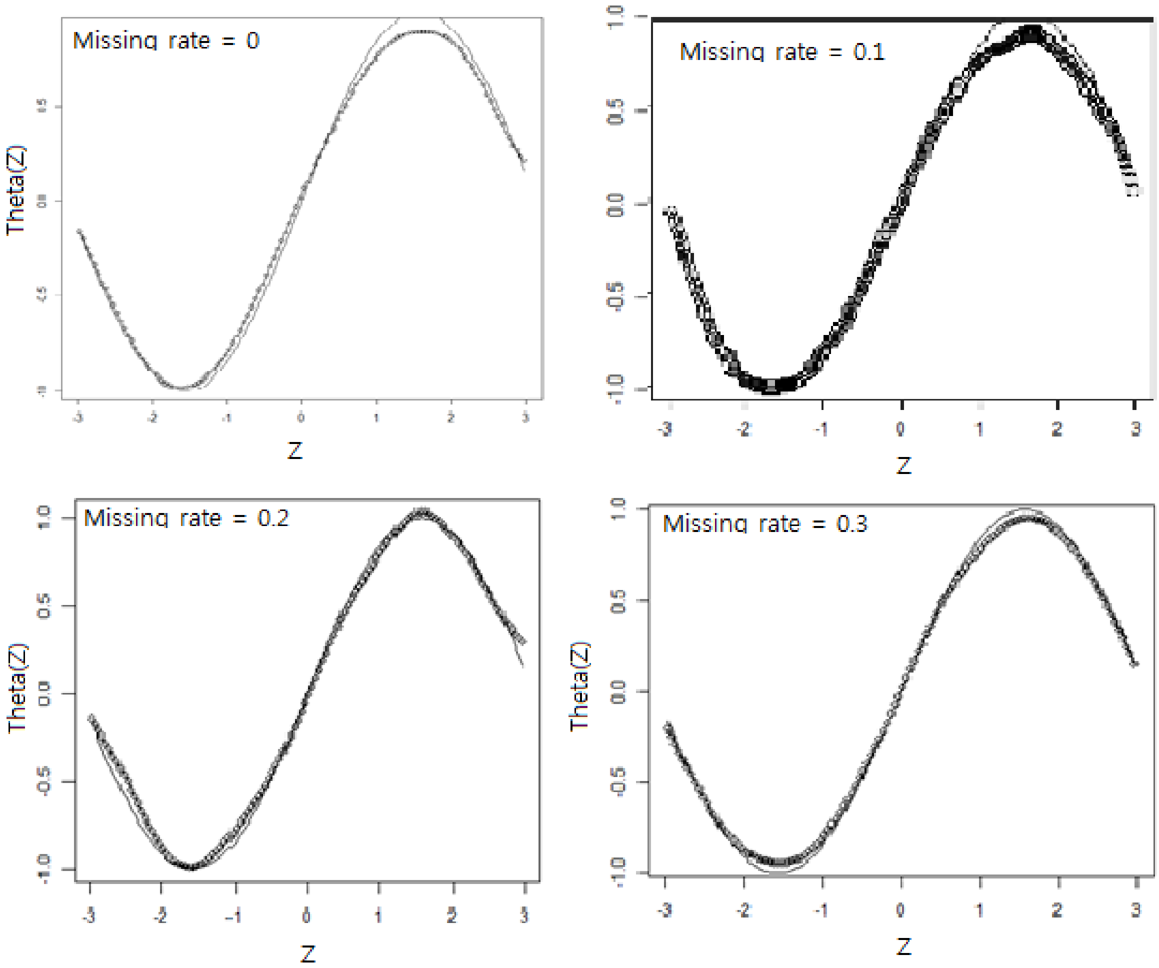
	$missing = 0$	$missing = 0.1$	$missing = 0.2$	$missing = 0.3$
1	1.541455	1.436193	1.532994	1.37934
2	1.402096	1.386127	1.752104	1.315729
3	1.60017	1.357302	1.503846	1.362417
4	1.421242	1.446253	1.553939	1.659089
5	1.577057	1.577733	1.391999	1.2069
6	1.48695	1.429482	1.260213	1.39394
7	1.524898	1.475953	1.467394	1.393641
8	1.443249	1.412329	1.518384	1.636065
9	1.555563	1.215347	1.312279	1.515621
10	1.618768	1.564556	1.429559	1.518577
$(the\ bias\ of\ \beta)$	0.171448	-0.69873	-0.27729	-0.61868
$(the\ RMSE\ of\ \beta)$	0.005468	0.014476	0.017833	0.022032

(Table 2: Comparison among various missing rates)

As shown in the table above, the biasness of beta is not greatly affected by the missing rate. Meanwhile, the RMSE(Root Mean Squared Error) shows some correlations with the missing rate. As the missing rate gets larger, the RMSE also grows.

While solving the equation derived from the score function, which is presented in the section 2-2, we used a number of the Newton-Raphson method. Each step required less than 6 iterations, and each solution required less than 5 repetition for the whole iteration, until convergence to the final solution.

For fixed missing probabilities 0, 0.1, 0.2, 0.3, plot the relations between the value of Z and $\hat{\theta}(Z)$. That figure represents the estimate of $(z,\ \theta(z))$ and a comparison with the real $\theta(z)$ function.



(Figure 2: Z - $\hat{\theta}(z)$. For the varying average missing rate, the dashed line represents the fitted value $\hat{\theta}(Z)$, and the normal line represents the real function $\theta(z)$, which is sin(z)

As we can see from the figure above, variation of missing probability does not affect the accuracy of estimating $\theta(z)$ or expected bias of estimator $\hat{\beta}$. Surprisingly, under the regular conditions, an accurate figure of $\theta(z)$ is estimated regardless of the missing rate.

3.2. Controlling Bandwidth

$\mathrm{X} \sim \mathrm{N}(0,0.25),\ \mathrm{Z} \sim \mathrm{U}(-3,3),\ \theta(\mathrm{z}) = z^3$

$$\mathrm{P}\{\mathrm{Y}=1|\mathrm{X}=\mathrm{x},\mathrm{Z}=\mathrm{z}\} \coloneqq \mathrm{P}(\mathrm{x},\mathrm{z}) = \frac{\exp(\beta x + \theta(z))}{1+\exp(\beta x + \theta(z))}$$

$$\mathrm{P}\{\mathrm{X\ is\ observed}|\mathrm{Y}=\mathrm{y},\mathrm{Z}=\mathrm{z}\} \coloneqq \pi(\mathrm{y},\mathrm{z}) = \frac{\exp(\gamma_1+\gamma_2 y+\gamma_3 z)}{1+\exp(\gamma_1+\gamma_2 y+\gamma_3 z)}$$

◎ Computing bias and RMSE (Root Mean Squared Error)

$$(the\ bias\ of\ \beta)\ = \frac{1}{m}\sum_{1}^{m}(\hat{\beta}-\beta_0)$$

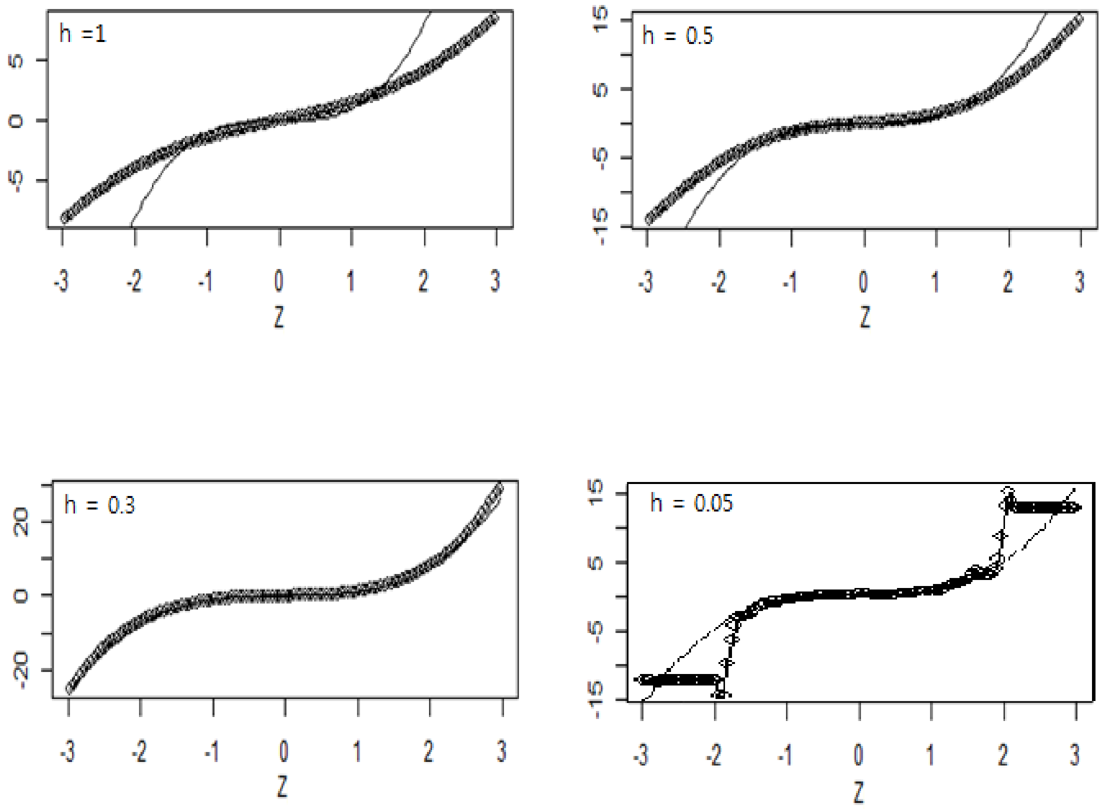
$$(the\ RMSE\ of\ \beta) = \frac{1}{n}\sum_{1}^{n}(\hat{\beta}-\beta_0)^2$$

$\hat{\beta}$	h = 1.	h = .8	h =.5	h = .3	h = .15
1	2.899109	1.936737	2.038326	1.731065	1.951009
2	2.163514	2.342149	2.233557	2.094598	1.406791
3	2.117206	2.493331	1.80058	2.049289	1.872129
4	2.843501	2.242444	2.461985	2.233142	1.827295
5	2.536399	2.047211	1.975731	1.749505	1.703718

$(the\ bias\ of\ \beta)$	1.011946	0.712374	0.602036	0.47152	0.252188
$(the\ RMSE\ of\ \beta)$	1.131564	0.547465	0.413997	0.261681	0.09984

(Table 3: Estimating β under various h)

For several bandwidths h= 1, 0.5, 0.3, 0.05, plot the relations between the value of Z and $\hat{\theta}(Z)$. That figure represents the estimate of $(z,\ \theta(z))$ and compare the real $\theta(z)$ function



(Figure 3: Estimating $\theta(z)$ under various h. The dashed line represents the fitted value $\hat{\theta}(Z)$, and the normal line represents the real function $\theta(z)$, which is sin(z))

As shown in the table and figure above, the biasedness of β decreases as h gets closer to 0.15. Meanwhile, the figure shows that the estimation of $\theta(z)$ is most accurate when h is close to 0.3.

Now we go back to the theoretical approach to optimal bandwidth.

According to the equation (2-2-2), Normal reference rule,

$$h_n = \frac{1.06\hat{\sigma}}{n^{\frac{1}{5}}}$$

$\hat{\sigma} \approx \sigma = \int z^2 f(z) = \int z^2 \left(\frac{1}{6} I_{(-3,3)}\right) dz = 3$, since $z \sim U(-3,3)$. We used n = 8000. Accordingly the theoretical optimal bandwidth is:

$$\mathrm{h}_{theo} = \frac{1.06 * 3}{\sqrt[5]{8000}} = 0.3042625$$

That is approximately 0.3, so this is compatible with our experimental result.

Chapter 4

Discussion

To determine how to make an efficient estimation for β $\theta(z)$, we tried see the effect of missing probability; the length of the interval, which is used for estimating local linear estimation of $\theta(z)$; and the number of interval; and the number *n* of sample; and the bandwidth. At first, we expected that larger missing rate of covariates would cause the inaccuracy of estimation, that is, a larger biasness of the estimator. To prove this suggestion, we conducted several simulations with various missing probabilities. However, as shown in chapter 3, missing probability does not have a significant effect on the accuracy of estimation of β and $\theta(z)$.

Also, I tried to compare the effectiveness of estimation between GPLM and GLM models when some of explanatory variables behave nonlinearly. In my simulation, one variable behaves linearly with a fixed proportional constant, and the other behaves nonlinearly, actually as a smooth function. Experiments were conducted with varying missing probability. Interestingly, every case of missing probability value shows that the GPLM model is more proper for use as a statistical model.

As I explained in chapter 2, the equation derived from the score function does not have a solution in a closed form in general. That is, we cannot expect that the equation will be solved algebraically, and we need to find solutions in numerical approximation. Assuming the nonlinear term is of a smooth function (twice differentiable), the most universal method is the Newton-Raphson method. When I first made an algorithm code and ran it, an overly great biasedness occurred. I tried to fix it by controlling the sample size, the number of intervals, the length of interval, the missing probability, and the range of covariates. However, the estimates did not get close to the real values. Finally when I tried to use an algorithm with several bandwidth values, the estimates began to behave well. And then I started to studied detailed theories of bandwidth, and found the appropriate reason of controlling bandwidth, and found the optimal value of bandwidth at last. Moreover, I firstly chose triangular kernel function, but a normal function seemed to be more efficient and altered it for final results. Selecting a kernel function was not very critical.

The most important finding from this simulation is that, bandwidth selection is very important for increasing the accuracy and efficiency of estimation.

Computing bandwidth based on (2-2-2), the optimal h is 0.3, and this is approximately equal to our experimental result, 0.3. This result shows that our simulation was reasonable and fit well.

Finally, to add some comments, the simulation took a very long time to run, almost 1.5 days for each case. I set the large sample size to minimize the biasedness. The result, however, tells us that sample size is not very critical in our simulation results. So, to save running time and obtain further high-quality results, it might be sufficient to set smaller sample sizes, for example 3000-4000 items.

References

Hua Liang (2008), “Generally partially linear models with missing covariates”, *Journal of Multivariate Analysis* 99, 880-895

T.J. Hastie, R.J. Tibshirani (1990), Generalized Additive Model, Chapman & Hall, New York.

Kaplan, E. L. and P. Meier (1958), “Nonparametric estimation from incomplete observations”, *Journal of the American statistical association*, 53 (282), 457-481.

Peter Hall, Byeong U. Park (2004), “Bandwidth choice for local polynomial estimation of smooth boundaries”, *Journal of Multivariate Analysis* Vol 91 Issue 2, 240-261.

Larry Wasserman (2006), All of Nonparametric Statistics, Springer.

Hua Liang and Haobo Ren (2005), “Generalized partially linear measurement error models”, *Journal of Computational and Graphical Statistics*, vol.14, no.1, 237-250

국 문 초 록

Generalized Partially Linear Models with missing data and its applications

결측 데이터를 포함한 일반화 부분 선형모형 및 적용

이원형 결과값을 가지는 다변량 데이터에서, 설명변수 중의 일부가 결측되었을때에 일반화 선형모형을 적용하기에는 다소 어려움이 있다. 몇몇 설명변수는 선형적으로, 다른 설명변수는 비선형적으로 작용하여 통계적으로 예측하고자 하는 결과변수에 영향을 미친다. 이러한 경우에, 통계모델을 구축하는 방법으로서 일반화 부분선형모형이 있는데, 이는 일반화선형모형(GLM) 과 일반화 가법모형(GAM) 두 모형 모두의 특징을 반영하는 방법이다. 특별히, 데이터가 결측되었을 경우에는 WEE method 를 적용하여 알아내고자 하는 함수를 추정해낼 수 있다.

설명변수의 결측 비율이 추정값의 편향성에 영향을 미치지 않는다는 사실은 흥미롭다. 이는 결측된 부분을 보완하기 위하여 역 확률값 가중치 방법으로 결측치 주변의 값에 가중치를 주기 때문이다. 그리고, 우리는 비모수 요소를 추정하기 위하여 커널회귀를 적용하는데, 여기에서 대역폭 h 가 필요하다. 본 논문에서는 h의 값을 변화시키는 여러 시뮬레이션을 수행하고, 이에 따른 편향성의 변화를 관찰하면서 대역폭 선택의 중요성을 확인하려고 한다.