이학석사 학위논문

Additive Models for Longitudinal Data with Application to KLIPS

종적자료에서의 가법적 모형과 한국노동패널조사자료에의 적용

2014년 2월

서울대학교 대학원

통계학과

김 민 정

Additive Models for Longitudinal Data with Application to KLIPS

지도교수 박 병 욱

이 논문을 이학석사 학위논문으로 제출함

2013년 10월

서울대학교 대학원

통계학과

김 민 정

김민정의 이학석사 학위논문을 인준함

2013년 12월

위 원 장	김 우 철	(인)
부위원장	박 병 욱	(인)
위 원	조 신 섭	(인)

Additive Models for Longitudinal Data with Application to KLIPS

by

Min Jung Kim

A Dissertation
submitted in fulfillment of the requirement
for the degree of
Master of Science
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
February, 2014

Abstract

Min Jung Kim
The Department of Statistics
The Graduate School
Seoul National University

Many studies have taken to construct models between time-varying variables for longitudinal data. These longitudinal models are widely used in physics, biology and social sciences. In this paper, we introduce a time-varying additive model with smooth backfitting to overcome the limitations of parametric model. This modeling strategy allows us to provide dimension reduction and simultaneously retain flexibility of regression function. Furthermore, an application to Korean Labor and Income Panel Study (KLIPS) data is presented to illustrate the proposed methodologies. This model might be quite useful to fit a data and gives a good explanation on social phenomenon.

Keywords : Additive model, Smooth backfitting, Kernel smoothing, Longitudinal data, KLIPS.

Student Number : 2012-20221

Contents

Chapter 1

Introduction

To analyze longitudinal data, many parametric methods have been proposed and developed, for example, generalized estimating equation (Liang and Zeger (1986)), linear mixed-effects model (Harville (1976) and Laird and Ware (1982)), and nonlinear mixed-effects model (Davidian and Giltinan (1995), Vonesh and Chinchilli (1996)). These models are useful tools in interpreting the relationship between a response variable and covariates in longitudinal studies and predicting a response variable. However, parametric models may be limited for many applications and sometimes unavailable for preliminary data analysis. To overcome this difficulty, several nonparametric and semiparametric regression techniques have been proposed to longitudinal data analysis in recent years, including smoothing spline methods (Brumback and Rice (1998), Guo, W. (2002)), regression spline methods (Shi et al. (1996)), and semiparametric approaches (Zeger and Diggle (1994), Lin and Carroll (2001)). Moreover, there have been many models developed based on kernel-type smoothing (Hoover et al. (1998), Wu and Chiang (2000), Lin and Carroll (2000)). These nonparametric models allow more flexible modeling for the dependence of a response

variable on covariates than parametric models.

In this paper, we especially aim to review the time-varying additive model that overcomes the high-dimensional problem as well as gives flexibility to regression function. It also provides easy, reliable interpretation and fine graphical representations of data. Furthermore, we review a smooth backfitting algorithm to estimate the component functions. The additive model and smooth backfitting for longitudinal data were proposed in Carroll et al. (2009) and Zhang, Park and Wang (2013), which will be investigated further in the next chapter.

Korea is rapidly becoming an aging and academic background-oriented society. Korean government has put more emphasis on social welfare. Due to this trend, the effect of age, working hours and a level of education on a worker's salary is becoming more important issue to establish policies and maintain economic conditions. Therefore, another purpose of this paper is to apply the additive model and local linear smoothing method to KLIPS data, which has researched social and economic phenomena as a longitudinal study since 1998. From there, we explore how educational level, age and work hours have influenced on salary for 11 years and visualize each component functions in graphs. Using this result, we can compare a longitudinal data analysis with a cross-sectional data analysis.

The remainder of this paper is organized as follows. Chapter 2 reviews the additive model and time-varying additive model with smooth backfitting as a nonparametric approach. Chapter 3 describes the KLIPS data. In Chapter 4, we will discuss the KLIPS data application based on the suggested models form Chapter 2. From this result, we analyze the relation between the response variable and the functional covariates. Chapter 5 provides conclusion and discussion.

Chapter 2

Models

This chapter reviews the nonparametric additive model for mean regression problem. In order to deal with longitudinal data, we extend this additive model to the time-varying additive models. We also describe the smooth backfitting to estimate component functions.

2.1 Additive Model

We have n observations on a response variable Y and p linearly independent covariates $\boldsymbol{X} = (X_1, \cdots, X_p)^\top$. Our aim is to construct multiple mean regression model: $m(x) = E(Y|\boldsymbol{X} = \boldsymbol{x})$ where $\boldsymbol{x} = (x_1, \cdots, x_p)^\top$. As standard approach for $m(\boldsymbol{x})$, the linear regression models are defined as

$$Y = \alpha + \sum_{j=1}^{p} X_j \beta_j + \epsilon, \tag{2.1}$$

where $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$. Although this model is a useful method for inference and prediction, the linear assumption often fails to apply real data.

Therefore, we can consider the general nonparametric model:

$$Y = f(X_1, \cdots, X_p) + \epsilon. \tag{2.2}$$

The above model avoids parametric assumption about the form of the function f, hence making it more flexible. However, as we increase the number of covariates, the general nonparametric model becomes impractical because this model allows all possible interaction terms between covariates and requires a larger sample size. Also, it is difficult to represent graphical result for p bigger than two or three. Thus, to solve such a high dimensional problem, Friedman et al. (1981) and Stone (1985) proposed the additive model:

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{2.3}$$

where α is a constant and $f_j(\cdot)$ is a smooth component function for all j. Thus, the additive model is a powerful model not only to have flexibility in modeling between a response and covariates but also to free from the curse of dimensionality. This model is also more interpretable than general nonparametric model. There are many methods to estimate component function. For example, Buja et al. (1989) introduced ordinary backfitting and Mammen et al. (1999) proposed smooth backfitting. We will focus more specifically on smooth backfitting in subchapter 2.3.

2.2 Time-varying additive model

In this subchapter, we carry out how the additive model (2.3) can be extended for longitudinal analysis. Now, let $Y(t)$ be a longitudinal response variable and $\boldsymbol{X}(t) = (X_1(t), \cdots, X_p(t))^\top$ be longitudinal covariates. An extension of

model (2.3) is described as

$$m(t, \boldsymbol{x}) = E(Y(t)|\boldsymbol{X}(t) = \boldsymbol{x}) = f_0(t) + \sum_{j=1}^{p} f_j(t, x_j), \tag{2.4}$$

where $\boldsymbol{x} = (x_1, \cdots, x_p)^\top$. For identification, the component functions f_k satisfy constraints such as $\int f_k(t, x_k) p_k(x_k|t) dx_k = 0$ for each t and $k = 1, \cdots, p$. This model was introduced by Zhang et al. (2013).

Now, let $(T_{ij}, \boldsymbol{X}_{ij}, Y_{ij})$ be the time points, the covariates and the response variable for subject i $(i = 1, \cdots, n)$ measured at the jth time point $(j = 1, \cdots, N_i)$ where $\boldsymbol{X}_{ij} = \boldsymbol{X}_i(T_{ij}) = (X_{ij1}, X_{ij2}, \cdots, X_{ijp})^\top$. Also, for simplicity, we assume that $\boldsymbol{X}_{ij}, Y_{ij}$ have common support on [0,1].

2.3 Smooth Backfitting

A backfitting algorithm is an iterative procedure used to fit the additive model, which was introduced in Breiman and Friedman (1985). It can estimate each component function by holding other variables constant and iterate this process. In addition, there is no high-dimensional problem in this algorithm because backfitting only depends on one dimensional function. Furthermore, Mammen et al. (1999) suggested the smooth backfitting estimator for additive nonparametric regression. This approach interprets local constant and local polynomial fitting as projections on a subspace of additive functions. Thus, smooth backfitting estimators are obtained by minimizing a smoothed least square.

We adopt the extended smooth backfitting with local linear fitting in Zhang et al. (2013). Let $\boldsymbol{h} = (h_1, \cdots, h_p)^\top$ be the bandwidth vector. We use a boundary corrected kernels $(K_h(\cdot;\cdot))$ in Yu et al. (2009). Let K be a base

kernel function and for a bandwidth h,

$$K_h(w,z) = \frac{1}{h}K\Big(\frac{w-z}{h}\Big)\Big[\int_S \frac{1}{h}K\Big(\frac{u-z}{h}\Big)du\Big]^{-1} I(w,z\in S),$$

where S is the support of the random variables. Also, we suppose that kernel K is symmetric and Lipschitz continuous. Moreover, the multivariate kernel function $\boldsymbol{K_h}$ is defined by a product kernel.

Now, in order to estimate component functions in model (2.4), we consider kernel-weighted least squares and local linear fitting. We first adopt linear approximation to employ local linear smoothing. We approximate $f_0(T_{ij})$ by $f_0(t) + f_{0,1}(t)(T_{ij}-t)/h_0$ where $f_{0,1}(t)/h_0$ is derivative of f_0. Also, for each k $(k=1,\cdots,p)$, we approximate $f_k(T_{ij},X_{ijk})$ by $f_k(t)+f_{k,1}(t,x_k)(T_{ij}-t)/h_0+ f_{k,2}(t,x_k)(X_{ijk}-x)/h_k$ where $f_{k,1}/h_0$ by partial derivative of f_k with respect to first variable, and $f_{k,2}/h_k$ by partial derivative of f_k with respect to second variable. Therefore, the estimator of component functions, $\hat{f}_k(k=1,\cdots,p)$, can be obtained by minimizing, for each t,

$$\hat{p}(t)^{-1}\int \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\Big[Y_{ij}-\mu_{0,0}(t)-\mu_{0,1}(t)\Big(\frac{T_{ij}-t}{h_0}\Big)-\sum_{k=1}^{p}\mu_{k,0}(t,x_k) \tag{2.5}$$

$$-\sum_{k=1}^{p}\mu_{k,1}(t,x_k)\Big(\frac{T_{ij}-t}{h_0}\Big)-\sum_{k=1}^{p}\mu_{k,2}(t,x_k)\Big(\frac{X_{ijk}-x}{h_k}\Big)\Big]^2\boldsymbol{K_h}(t,\boldsymbol{x};T_{ij},\boldsymbol{X}_{ij})d\boldsymbol{x},$$

where $N=\sum_{i=1}^{n}N_i$ and $\hat{p}$ is the kernel density estimator of T. We cannot obtain the explicit form of the solution for (2.5). Numerically to get a solution, the minimization problem can be solved by an iterative method as in Zhang et al. (2013). Let v_{ijk}, $\hat{\Omega}_{kk}(t,x_k)$ and $\hat{\Omega}_{km}(t,x_k,x_m)$ be defined as follows:
For $k=1,\cdots,p$,

$$v_{ijk}(t,x_k) = (1,(T_{ij}-t)/h_0,(X_{ijk}-x_k)/h_k)^\top,$$

$$\hat{\Omega}_{kk}(t, x_k) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{N_i} v_{ijk}(t, x_k) v_{ijk}(t, x_k)^\top \boldsymbol{K}_{h_0,h_k}(t, x_k; T_{ij}, X_{ijk}),$$

and for $k \neq m$,

$$\hat{\Omega}_{km}(t, x_k, x_m)$$
$$= \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{N_i} v_{ijk}(t, x_k) v_{ijm}(t, x_m)^\top \boldsymbol{K}_{h_0,h_k,h_m}(t, x_k, x_m; T_{ij}, X_{ijk}, X_{ijm}).$$

Also, we obtain local linear kernel estimator $\hat{\mathbf{m}}_k(t, x_k)$ explaining a relation between Y_{ij} and covariates (T_{ij}, X_{ijk}). Thus, we observe the smooth backfitting algorithm in Table 2.1.

Zhang et al. (2013) showed that under regular condition, the system of equations (2.5) has a unique solution with probability tending to one and the smooth backfitting algorithm converges to the solution of (2.5) in the L^2-norm very fast. They also established the oracle property of the smooth backfitting estimator. These properties, conditions and the proofs can be found in Zhang et al. (2013).

Table 2.1: The Smooth Backfitting Algorithm

Step 1: Set initial values $\hat{f}_k^{[0]}$ for $k = 1, \cdots, p$ that satisfy the identification condition.

Step 2: Let $\hat{f}_k^{[r]}$ be the estimates of f_k in the rth iteration. The updating equation is defined as

$$\begin{aligned}\hat{f}_k^{[r]}(t, x_k) =& \hat{\mathbf{m}}_k(t, x_k) - \hat{f}_0(t) \\ & - \sum_{m<k}^{p} \int \hat{\Omega}_{kk}(t, x_k)^{-1} \hat{\Omega}_{km}(t, x_k, x_m) \hat{f}_k^{[r]}(t, x_m) dx_m \\ & - \sum_{m>k}^{p} \int \hat{\Omega}_{kk}(t, x_k)^{-1} \hat{\Omega}_{km}(t, x_k, x_m) \hat{f}_k^{[r-1]}(t, x_m) dx_m\end{aligned}$$

Step 3: Iterate Step 2 until the component functions $\hat{f}_k$ change less than a pre-specified small positive real number.

Chapter 3

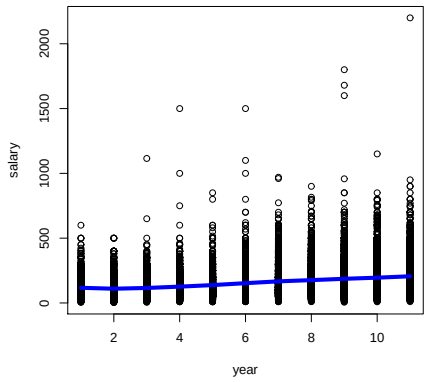
Data Description

Korea Labor Institute has collected the Korean Labor and Income Panel Study (KLIPS) data to track the characteristics of individuals such as the social activities, economic activities, income, and education since 1998. The KLIPS data was selected from all households living in urban areas across the country excluding Jeju Island by using a two-stage stratified clustering sampling method. The number of the KLIPS in 1998 was 5,000 households. The variables included were monthly salary, work hours per week, educational level, and age. We, however, only investigated people aged over 30 years to analyze economically active population and used complete data from 1998 to 2008. Additionally, we only considered salary under 30 million won because some data were distinct from other observation. The original data and variable descriptions are available on the website http://www.kli.re.kr/klips/ko/main/main.jsp and we retrieved the data in May 2013.

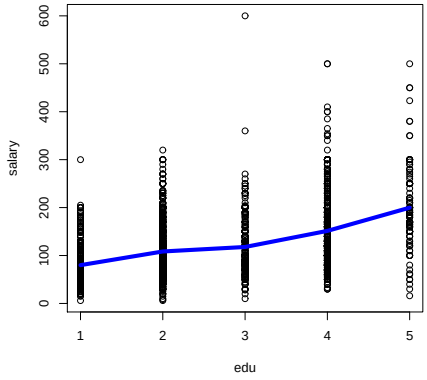
We define $year_{ij} = 1, \cdots, 11$ by investigation years from 1998 to 2008, respectively. Also, $salary_{ij}$, edu_{ij}, age_{ij}, $whour_{ij}$ denote the monthly salary (unit: ten thousand won), educational level, age, and work hours per week for

the ith worker in the jth year ($i = 1, \cdots, 6461$, $j = 1, \cdots, 11$), all of which are time-variant variables. More specifically, edu_{ij} has one of values 1, 2, 3, 4 and 5 if educational level is each of under middle school level, high school level, 2-year colleges, 4-year colleges, and over masters degree.

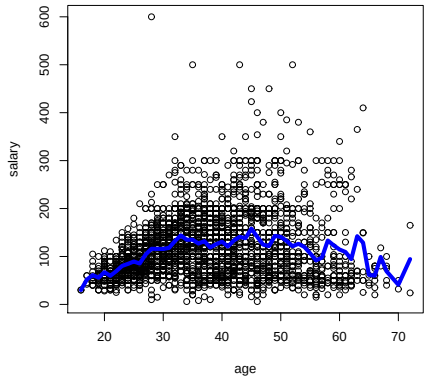
The following Figure 3.1(a) explains that the relation between year and salary is not exactly linear. Although the salary monotonically had increased after 1999 (year 2), it decreased slightly from 1998 (year 1) to 1999 (year 2). Figure 3.1(b) shows the tendency of salary became increased as educational level upgraded in 1998. Furthermore, Figure 3.1(c) presents the plot of salary over age in 1998 and the solid line indicates the mean salary over age. The mean line seems like a quadratic curve. Finally, the effect of work hours on salary at 1998 is described in Figure 3.1(d). As can be seen in this plot, the curve of solid line is not simple but very unusual.



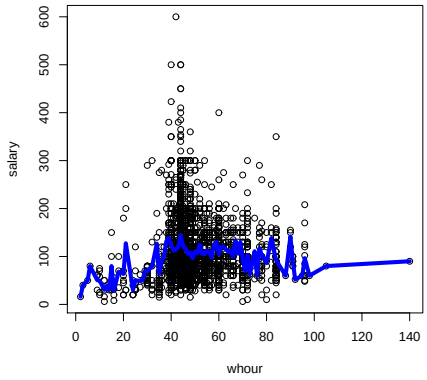
(a) Plot salary over year



(b) Plot salary over education at 1998



(c) Plot salary over age at 1998



(d) Plot salary over work hours at 1998

Figure 3.1: Plots describing KLIPS data

Chapter 4

Application to KLIPS data

Chapter 4 shows that the time-varying additive model and smooth backfitting method in Chapter 2 are applied to the KLIPS data to research the effect of employees' educational level, age and work hours on their salaries.

We implemented the model (2.4) to KLIPS data, as follows:

$$\begin{aligned} &E(salary|year,age,edu,whour) \\ &\qquad = f_0(year) + f_1(year,edu) + f_2(year,age) + f_3(year,whour). \end{aligned} \tag{4.1}$$

Also, the Gaussian kernel was used and the threshold for the convergence of smooth backfitting algorithm was 10^{-4}. We took the bandwidth arbitrary such as (h_0, h_1, h_2, h_3)=(1, 1.5, 5, 8), but if we choose the bandwidth by using some methods such as rule-of-thumb method (Lee et al. (2012)), or plug-in approach (Zhang et al. (2013)), the result would be better.

Figure 4.1 shows fitted overall mean function of the salary, $\hat{f}_0(year)$. As shown in the plot, the salary had slightly declined until 1999 (year 2), and, afterwards the salary had almost linearly increased since 1999. It is similar to the tendency we obtained from our data. The result explained that the IMF crisis shook the very foundation of the Korean economy in 1997, so eco-

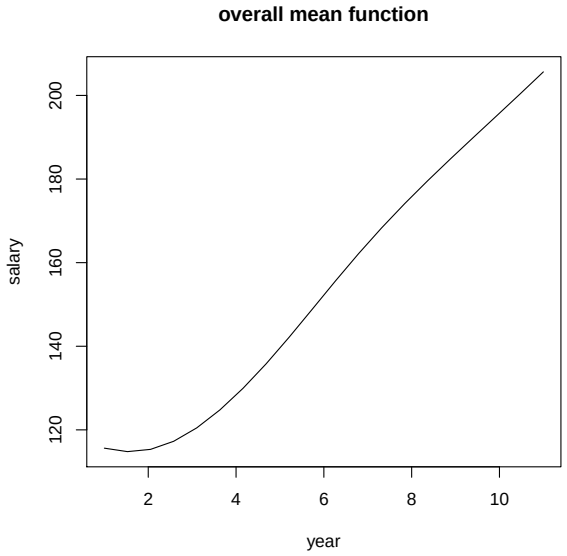


Figure 4.1: Estimated overall mean function, $\hat{f}_0(year)$

nomically active population had decreased and a job market had been even worse from 1997 to 1999. After 1999, the Korean economy has recovered and the urgency of the international situation has being alleviated, so the worker's salary has risen. Here, we draw this figure from multiple-viewing direction for better interpretation. Moreover, Figure 4.2, Figure 4.3, and Figure 4.4 present the component functions, $\hat{f}_1(year, edu)$, $\hat{f}_2(year, age)$ and $\hat{f}_3(year, whour)$, respectively.

In Figure 4.2, the plots describe the estimated component function $\hat{f}_1$ and show the relationship between workers' salaries and education levels. The above plot in Figure 4.2 indicates that the over masters degree group received significantly more salaries than did the group of lower education levels every year. More specifically, as time passed, the salary of works with lower high school education had tended to decrease, while the salary of workers with mas-

ters and doctoral degree had increased. The results, however, suggested that two-year college graduates' wage had remained constant over year. Moreover, the wage gap between workers with higher educational levels and workers with lower educational levels had been widening since 1998.

We show how workers' age can influence on their salary in Figure 4.3. For example, the wage of people in 30s had consistently declined since 1998 (year 1). This finding suggested that young people's entry into society had been delayed. Moreover, Figure 4.4 draws third component function indicates that people who worked for about 40 hours received the most money every year.

Now, we focus on each component function with $\hat{f}_0$ in the following figures. Figure 4.5 gives the similar results as Figure 4.2. In Figure 4.6, the plots show the estimates of the component functions, $\hat{f}_0(year) + \hat{f}_2(year, age)$ in model (4.1), providing the effect of employees' ages and year on their salaries with all other variables being held fixed. The estimated curve explains that the younger and middle-aged people's salary had risen faster than older people's after 1999 (year 2) based on the above plot. In addition, we can see that the salary reached a peak at the age of 45 and then started to decrease in 1999, whereas it did at age of 48 in 2008 (year 11). That is, the age at which salary reached peak increased. Furthermore, 50-year-old workers earned less money than 41-year-old workers did in 1998 (year 1). However, 50-year-old workers' wage in 2008 was higher than 41-year-old workers' in 2008. The below panel of Figure 4.6 also indicates that the sharp decline of salaries generally occurred as workers' ages increased from 55 to 65 over year.

Figure 4.7 displays the estimated surface for $f_0(year) + f_3(year, whour)$. As shown in this plot, the employees who worked 42 hours per week had earned the maximum wage from 1998 to 2004 (year 7). On the other hand, the employees who worked 37 hours per week had done after 2004. Furthermore,

workers who worked almost 100 hours had tended to receive a high salary after 2005 (year 8). On the whole, when working hours were 30–50 hours per week, the workers got a stable income. However, the salary was fluctuating and very unstable as working hours were more than 100 hours. Based on the limited data available, a rise in salary took place in 2002 (year 5) and 2007 (year 10). This phenomenon might occur because there were two workers; one got 5.8 million won although s/he worked only 3 hours per week in 2002 and the other who worked 5 hours received 4.4 million won in 2007.

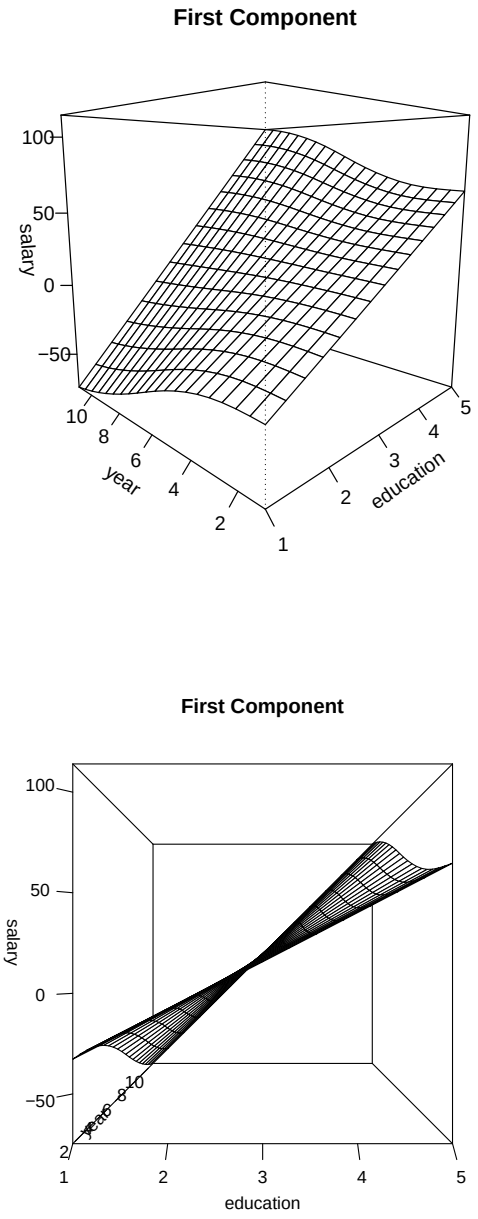


Figure 4.2: Fitted first component function, $\hat{f}_1(year, edu)$

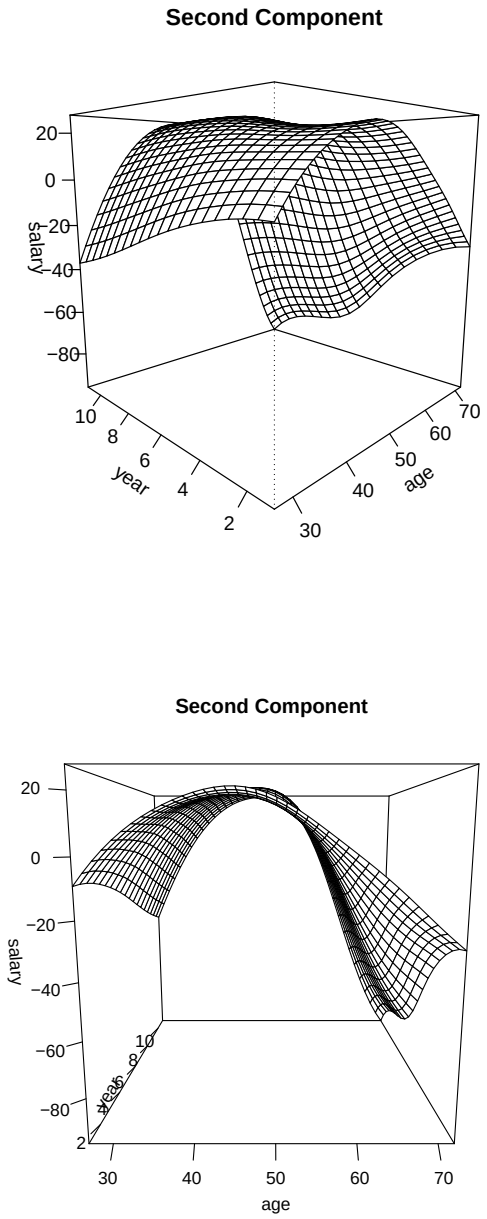


Figure 4.3: Fitted second component function, $\hat{f}_2(year, age)$

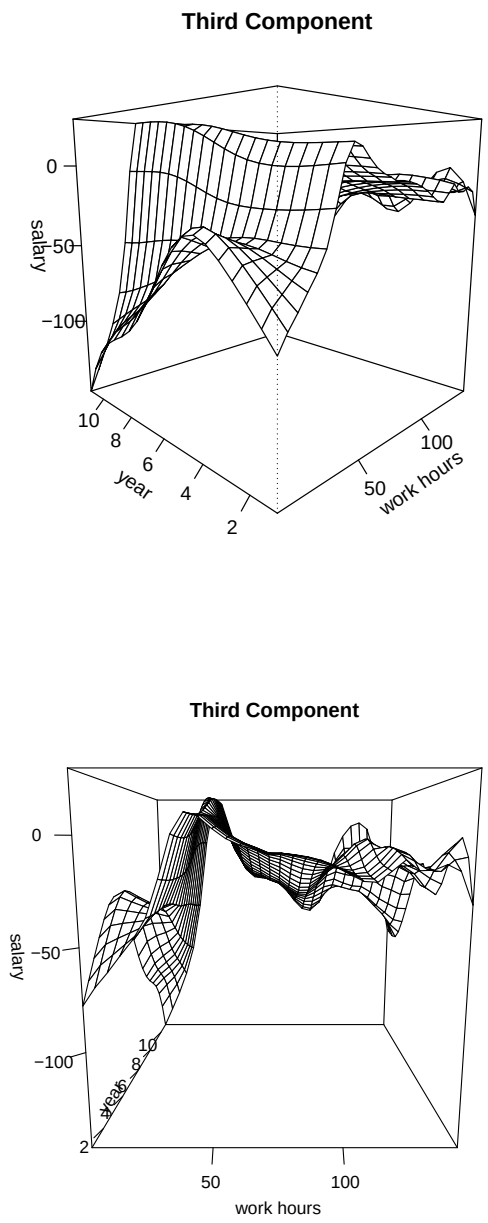


Figure 4.4: Fitted third component function, $\hat{f}_3(year, whour)$

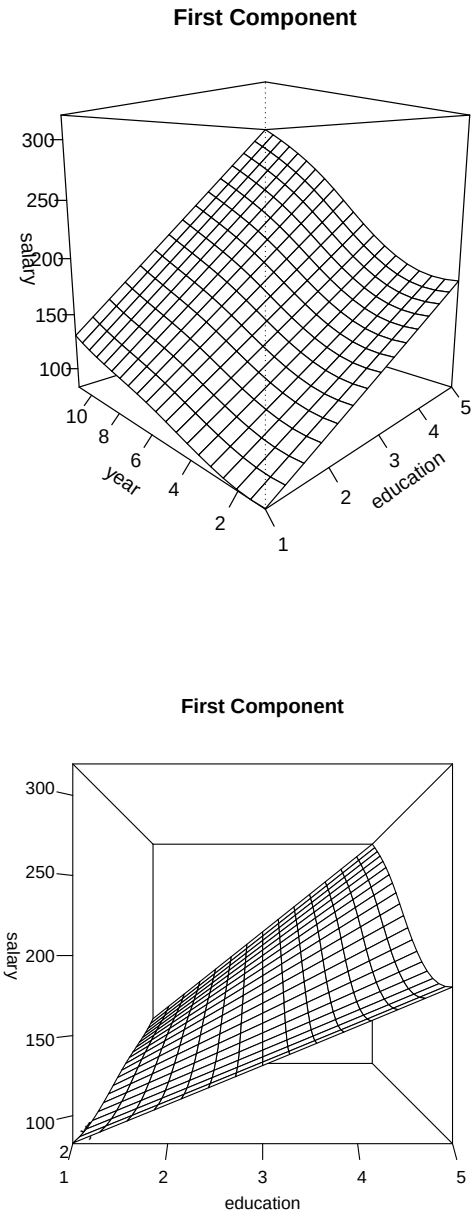


Figure 4.5: Surfaces for $\hat{f}_0(year) + \hat{f}_1(year, edu)$

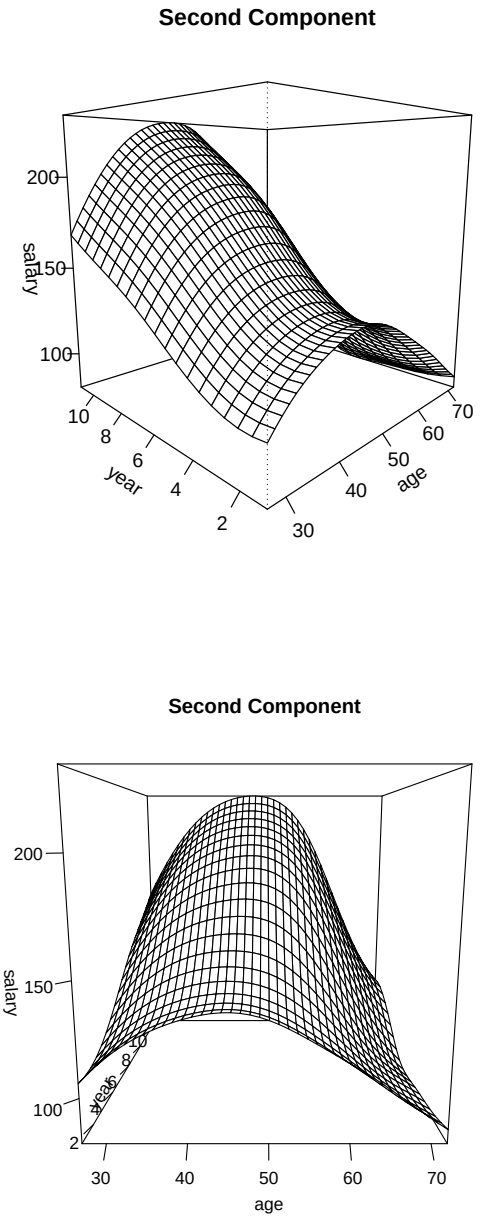


Figure 4.6: Surfaces for $\hat{f}_0(year) + \hat{f}_2(year, age)$

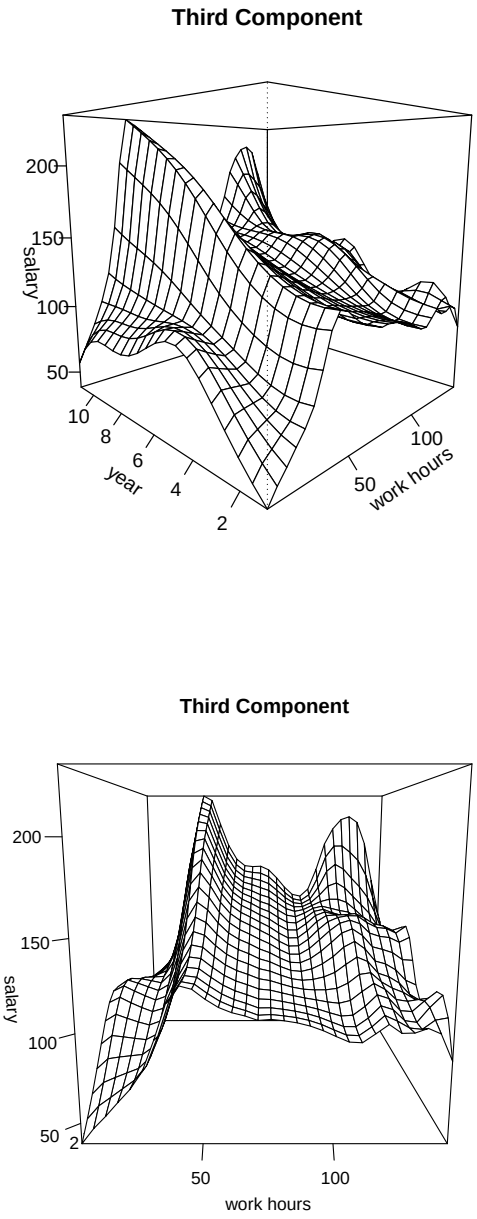


Figure 4.7: Surfaces for $\hat{f}_0(year) + \hat{f}_3(year, whour)$

Chapter 5

Conclusion and Discussion

Our objectives in this paper were that we implemented the nonparametric additive model to KLIPS data and analyzed how workers' age, education levels and work hours could influence on their earnings. Another goal was to draw a comparison between a longitudinal analysis and a cross-sectional analysis.

As a result, this study has indicated that the additive model was more suitable in modeling between an individuals salary, level of education, age and working hours in comparison to the parametric model because the effect of age and work hours on salary was not linearly related. Also, our result on the relation between year and salary is supported by Korean Development Institute (KDI) research. According to KDI, Korea had been in trouble due to IMF crisis and unstable international situation until 1999, so employees' wage had declined. Overall, we observed that employees with a doctorate degree earned the highest salaries among all levels of education. A possible explanation for this findings was that Koreans may place a bigger emphasis on a person's educational background to get high-paying jobs. In addition, the older people's salaries rapidly dropped. This drop might be explained by the fact that older

people retired at the normal retirement age. Moreover, workers who worked 42 hours had received the highest wage until 2004, while workers who worked almost 37 hours did after 2004. This result may be related to the following Labor Standards Act, modified in July 2004. The law proposed statutory working hours to be shortened to 8 hours per day, 40 hours per week. Our finding also noted that the difference between a longitudinal and cross sectional studies may be due to structural problems of Korea such as inflation and economic depression for 11 years.

This study has taken a step in the direction of defining the relationship only between year, age, educational level, working hour and salary. However, KLIPS included other variables such as workplace type and size, gender, and welfare. Therefore, the further study will be welcomed to include more covariates in time-varying additive model. It would improve the interpretation and bring more accurate prediction. Also, the additive model can get a better result regardless of additional variables because this model is able to handle high-dimensional problem effectively.

In our data, the relation between education and salary seems to be a linear form, while the effect of other covariates on salary has not specified form but different trends in each direction. Therefore, the future research will focus on the partially linear model (Robinson (1988)), where education level is a covariate in parametric part; age and work hour are covariates in nonparametric part. Also, we consider the nonparametric part as the time-varying additive model in the partially linear model. Thus, we can conclude that the result would be useful.

References

Breiman, L. and Friedman, J. H. (1985), "Estimating optimal transformations for multiple regression and correlations (with discussion)," *Journal of American Statistical Association*, **80**, 580–619.

Brumback, B. and Rice, J. A. (1998), "Smoothing spline models for the anlaysis of nested and crossed samples of curves," *Journal of American Statistical Association*, **93**, 961–994.

Buja, A. Hastie, T., and Tibshirani, R. (1989), "Linear smoothers and additive models," *The Annals of Statistics*, **17**, 453–510.

Carroll, R. J., Maity, A., Mammen, E., and Yu, K. (2009), "Nonparametric additive regression for repeatedly measured data," *Biometrika* **96**, 383–398.

Davidian, M. and Giltinan, D. M. (1995), *Nonlinear Models for Repeated Measurement Data,* Chapman and Hall, London.

Friedman, J. H. and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of American Statistical Association* **76**, 817–823.

Guo, W. (2002), "Inference in smoothing spline analysis of variance," *Journal of Royal Statistical Society, Series B*, **64**, 887–898.

Harville, D. A. (1976), "Extension of the Gauss-Markov theorem to include the estimation of random effects," *Annals of Statistics*, **4**, 384–395.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998), "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data," *Biometicka*, **85**, 809–822.

Laird, N. M. and Ware, J. H. (1982), "Random effects models for longitudinal data," *Biometrics*, **56**, 89–97.

Lee, Y., Mammen, E., and Park, B. (2012), "Projection-type estimation for varying coefficient regression models," *Bernoulli*, **18**, 177–205.

Liang, K. Y. and Zeger, S. L. (1986), "Longitudinal data analysis using generalized liner models," *Biometrika.* **73**, 13–22.

Lin, X. and Carroll, R. J. (2000), "Nonparametric function estimation for clustered data when the predictor is measured without/with error", *Journal of American Statistical Association,* **95**, 520–534.

Lin, X. and Carroll, R. J. (2001), "Semiparametric regression for clustered data," *Biometrika*, **88**, 1179–1185.

Mammen, E., Linton, O. B., and Nielsen, J. P.(1999), "The existence and asymptotic properties of a backfitting projection algorithm under week conditions," *The Annals of Statistics*, **27**, 1443–1490.

Robinson, P. M. (1988), "Root-n Consistent Semiparametric Regression," *Econometrica*, **56**, 931–954.

Scott L. Zeger and Peter J. Diggle (1994), "Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrika*, **50**, 689-699.

Shi, M., Weiss, R. E. and Taylor, J. M. (1996), "An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves," *Applied Statistics*, **45**, 151–163.

Stone, C. (1985), "Additive regression and other nonparametric models," *Annals of Statistics*, **13**, 689–705.

Vonesh, E. F. and Chinchilli, V. M. (1996), *Linear and Nonlinear Models for the Analysis of Repeated Measurements,* Marcel Dekker, New York.

Wu, C. O. and Chiang, C. T. (2000), "Kernel smoothing on varying coefficient models with longitudinal dependent variable," *Statistica Sinica*, **10**, 433–456.

Xiaoke Zhang, Byeong U. Park and Jane-Ling Wang (2013), "Time-varying Additive Models for Longitudinal Data," *Journal of the American Statistical Association*, **108**, 983–998.

Yu, K., Park, B. U., and Mammen, E. (2008), "Smooth backfitting in generalized additive models," *The Annals of Statistics* **36**, 228–260.

국문초록

종적 자료(longitudinal data)에서 시간에 따라 변하는 변수들의 모형을 구성하는 방법은 많은 연구에서 행해져 왔다. 이런 종적 모형들은 물리, 생물, 사회과학 등 다양한 분야에서 활용되어 오고 있다. 이 논문에서는 모수적 모형(parametric model)의 한계점을 극복하고 시간에 따라 변하는 가법적 모형(time-varying additive model)과 평활 역적합(smooth backfitting method)을 소개한다. 이 모델의 장점은 차원축소를 하고 회귀함수의 유연함을 제공하는 것이다. 더 나아가서, 앞서 제시된 모델을 설명하기 위해 한국노동패널자료(Korean Labor and Income Panel Study)에 적용이 제시된다. 이 모델은 자료를 적합 하는 데 유용하고 사회적인 현상을 잘 설명하리라 생각된다.

주요어 : 가법적 모형, 평활 역적합, 커널 평활법, 종단자료, KLIPS

학　번 : 2012-20221