碩士學位論文

함수적 데이터 분석의 비모수적 추정법에 관한 연구

韓國外國語大學校 大學院

統 計 學 科

임 슬 기

碩士學位論文

함수적 데이터 분석의 비모수적 추정법에 관한 연구

A Study on Nonparametric Method for Functional Data Analysis

指導 姜 奇 勳 敎授

이 論文을 碩士學位 請求論文으로 提出합니다.

2014年 12月

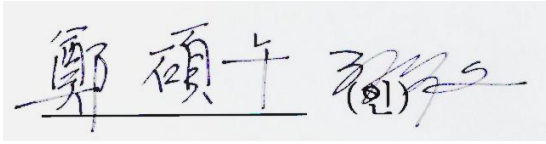
韓國外國語大學校 大學院

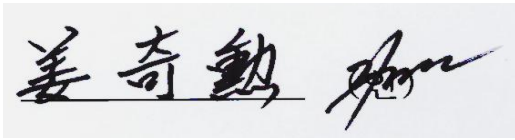
統 計 學 科

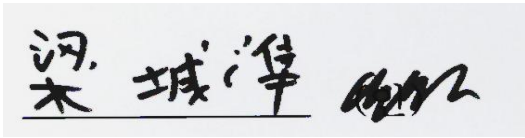
임 슬 기

이　論文을 임슬기의 碩士學位 論文으로 認定함.

2014年　　12月　　01日

審査委員 

審査委員 

審査委員 

韓國外國語大學校 大學院

요 약

시간에 관련된 데이터를 분석하는 대부분의 접근 방법은 시간에 따른 동태(dynamic)에 대한 정보를 무시하고 데이터를 단면적으로 처리하는 경향이 있다. 이러한 접근 방법들은 시간에 대한 정보를 잃게 되고, 기본적인 요약 통계량을 제시하는 것으로 한하여 시간에 관련된 규칙성을 연구한다. 이러한 한계점을 해결하기 위해 제한된 여러 분석 방법 중 하나가 Ramsay와 Silverman (2002)이 제안한 함수적 데이터 분석(functional data analysis)이다. 이는 시간에 대한 정보의 손실을 최소화 하여 직접 모형화하고, 시간에 따른 동태의 변화를 파악하기 위해 분석하는 방법이다.

본 연구에서는 Ramsay와 Silverman (2002)이 제안한 함수적 데이터 분석 방법을 실제 데이터에 적용하여 데이터의 동태를 파악하고 함수적 회귀분석(functional regression analysis)을 수행하는 방법으로 분석하였다.

목 차

1. 서론 ····· 1

2. 함수적 데이터 분석 ····· 3

2.1 데이터 표현 ····· 4

2.2 데이터 조정 ····· 6

2.3 함수적 선형 모형 ····· 7

3. 함수적 데이터 분석 기법의 적용 ····· 9

3.1 옥션 데이터 ····· 9

3.2 함수적 데이터 분석 수행 단계 ····· 12

3.2 함수적 회귀 분석 ····· 18

3.3 함수적 F 검정 ····· 24

4. 결론 및 토의과제 ····· 27

참고문헌 ····· 28

<표> 목 차

1. <표 1> 연속형 변수에 대한 요약 통계량 ····· 11
2. <표 2> 이산형 변수에 대한 요약 통계량 ····· 11

[그 림] 목 차

1. <그림 1> 경매 기록 ····· 10
2. <그림 2> 두 개의 옥션 데이터 자료 산점도 ····· 12
3. <그림 3> 선형보간법 전과 후의 그래프 ····· 13
4. <그림 4> 입찰 횟수에 대한 히스토그램 ····· 15
5. <그림 5> 페널티를 부과한 평활화 스플라인 ····· 16
6. <그림 6> 입찰가격에 대한 플랏과 미분계수에 대한 플롯 ····· 17
7. <그림 7> 상태에 대한 입찰가격의 함수적 선형 모형의 회귀 계수 ····· 19
8. <그림 8> 상태에 대한 가격 예측 값 ····· 21
9. <그림 9> 점프비딩에 대한 입찰가격의 함수적 선형 모형의 회귀 계수 ·· 22
10. <그림 10> 점프비딩에 대한 가격 예측 값 ····· 23
11. <그림 11> 상태 변수에 대한 순열검정 ····· 25
12. <그림 12> 점프비딩 변수에 대한 순열검정 ····· 26

1. 서론

현실에 존재하는 다양한 데이터는 시간에 의존하여 관측되며 함수 형태로 표현되어질 수 있다. 하지만 이러한 데이터에 대한 대부분의 접근 방법은 데이터의 동태에 대한 정보를 무시하고 단면적으로 처리하는 경향이 있다. 이의 경우 데이터의 정보를 크게 손실할 위험이 있다. 또한 기존의 접근 방법으로는 요약 통계량을 계산하는 것에 국한하게 된다. 이러한 문제점에 대한 대안으로 함수적 데이터 분석(functional data analysis) 기법이 있다. 이에 대한 일반적인 개관에 대해서는 Ramsay와 Silverman (2005)를 참고하면 된다.

함수적 데이터 분석에서 자료의 변동성을 알아보기 위해서는 정교한 방법들이 필요하며 대표적으로 함수적 선형 모형(functional linear modelling)이 있고, 함수적 주성분 분석(functional principal components analysis), 함수적 정준상관 분석(functional canonical correlation) 등이 있다. 또한 함수적 데이터 분석에서는 미분계수(derivatives)와 선형 미분 연산자(linear differential operators)들이 유용하게 사용된다. 특히 미분계수를 사용함으로써 데이터의 동태 변화를 파악하여 간단한 그래프적인 의미를 설명하는 데도 이용되고 있다. 함수적 데이터 분석에서는 추후 분석을 위해 데이터를 변환 및 재표현하며 데이터의 패턴과 변동성에 대해 연구한다.

본 연구에서는 함수적 데이터 분석에 대한 개괄적인 방법론들에 대해 고찰하고 실제 자료에 적용시킴으로써 함수적 데이터 분석 기법의 유용성에 대해 살펴볼 것이다. 이를 위하여 옥션 중고장터에서 CANON DSLR 카메라 중 모델명 EOS600D의 입찰 기록 데이터를 이용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 함수적 데이터란 무엇이며 함수

적 데이터 분석을 수행하는 단계에 대해서 알아보겠다. 함수적 데이터 분석의 수행단계 중 첫 번째로는 데이터를 표현하는 방법, 두 번째로 데이터를 조정하는 것, 그리고 세 번째로는 함수적 선형 모형에 대해서 설명하겠다. 3장에서는 함수적 데이터 분석의 기법이 어떻게 적용되는지 살펴보기 위해 사용한 옥션 데이터에 대해 설명하겠다. 실제 데이터를 사용하여 함수적 데이터 분석을 단계별로 수행하고, 이를 함수적 선형 회귀 모형을 적합 시켜 분석을 하겠다. 그리고 적합시킨 모형이 유의한지 순열검정(permutation test)을 통해 확인하겠다. 결론에서는 실제 함수적 데이터 분석의 결과를 통해 기존의 분석 방법의 단점을 개선한 내용과 함수적 데이터 분석의 추후 과제에 대하여 논의하겠다.

2. 함수적 데이터 분석

Ramsay와 Silverman(2005)에 의하면 함수적 데이터는 시간, 빈도, 확률과 같은 연속체에 따라 변화하는 곡선(curves)이나 표면(surfaces)등을 의미한다. 이는 n개의 쌍인 (t_j, y_j)로 이산적으로 관측되고 측정되며 y_j는 $t=t_j$에서 함수 값 $x(t)$의 관측이나 기록이다. 이렇게 시간에 대한 함수(function) 자체에 초점을 맞추어 분석 하는 것이 함수적 데이터 분석(functional data analysis)이다.

함수적 데이터 분석을 수행하기 위해서 가장 먼저 해야 할 것은 이산형으로 관측된 데이터를 함수 형태로 나타내는 데이터 표현(data representation) 작업이다. 만약 이산형의 관측치가 잡음(noise)이 없다고 여겨지면 보간법(interpolation)을 사용하고, 제거해야할 잡음이 있다면 평활화(smoothing)기법이 사용하다. 두 번째로 해야 할 것은 데이터 배열(data registration) 또는 조정(alignment)이다. 시간 축을 적절히 옮겨서 관측치를 조정하는 과정이 필요하다. 세 번째로는 데이터 시각화(data display)로 함수형태로 변환한 결과를 다양한 방식의 그림으로 표현하여 결과를 확인한다. 마지막으로 함수값 뿐만 아니라 함수를 m차 미분한 미분계수를 그림으로 확인한다. 특별히 관심을 가지게 되는 미분계수의 추정은 많은 함수적 데이터 분석에서 중요하게 이용된다. 이것은 단변량 함수 x의 m차 미분계수를 $D^m x$라 나타내며, t에서의 m차 미분계수의 값은 $D^m x(t)$라고 표현한다.

2.1 데이터 표현(data representation)

함수적 데이터 분석을 수행하는 첫 번째 단계는 데이터를 표현하는 것이다. 이산형 데이터를 함수 형태로 평활화(smoothing)하는 방법과 가장 유사한 방법 중에 하나는 충분히 큰 K개의 각각 독립인 기저 함수(basis function) ϕ_k의 선형 결합(linear combination)으로 함수를 표현하는 방법이다.

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t).$$

$x(t)$의 평활화(smoothing) 정도는 기저 함수의 수 K에 의해서 결정된다. 그리고 기저 함수를 이용한 가장 단순한 선형 평활화(linear smoother)는 상수 c_k가 오차제곱합(Sum of Squares Error) 식 (1)을 최소로 하는 최소제곱법을 적용시켜 얻을 수 있다.

$$SMSSE(y|c) = \sum_{j=1}^{n} [y_j - \sum_{j=1}^{K} c_k \phi_k(t_j)]^2 \qquad (1)$$

식 (1)을 행렬식으로 표현하면,

$$SMSSE(y|c) = (y - \Phi c)'(y - \Phi c) = \| y - \Phi c \|^2 \qquad (2)$$

이 되고, 이를 최소로 하는 c는 $c = (\Phi'\Phi)^{-1}\Phi' y$로 구해진다.

기저 함수의 바람직한 특성은 모 함수와 추정한 함수가 비슷한 형태를 가지도록 하는 것이다. 기저 함수를 사용함으로써 함수에 대한 정보를 잘 적용시킬 수 있으며 이상적으로는 비교적 적은 수 K의 기저를 사용하면 계산을 더 적게 할 뿐만 아니라 잘 근사하게 된다.

데이터의 형태에 따라 기저를 선택하는 방법도 달라진다. 주기적인 데이터

(periodic data)라면 일반적으로 퓨리에 기저(Fourier basis)를 사용하고, 비주기적인 데이터(non-periodic data)라면 B-스플라인 기저(B-spline basis)를 사용한다.

일반적으로 잘 알려진 퓨리에 급수를 이용한 퓨리에 기저의 표현은 $\phi_0(t)=1$, $\phi_{2k-1}(t)=\sin kwt$ 그리고 $\phi_{2k}(t)=\cos kwt$로 정의하며 다음과 같이 나타낸다.

$$\hat{x}(t)=\sum_{k=0}^{\infty} c_k\phi_k(t)=c_0+c_1\sin wt+c_2\cos wt+c_3\sin 2wt+c_4\cos 2wt+\cdots \qquad (3)$$

이 기저는 주기적이며 모수 w는 주기 $2\pi/w$를 결정하고, 그것은 기저가 적용되는 구간의 길이와 동일하다. 퓨리에 급수는 상당히 안정된 함수로 유명하며 지역적 성질이 적고 곡률의 차수가 동일한 함수에서 잘 적용된다.

B-스플라인 기저는 스플라인 함수(spline function)를 사용하여 정의된다. 스플라인은 근사되어야 할 함수의 간격을 중단점(breakpoints) 또는 노트(knots)라고 하는 $\tau_l, l=1,\cdots,L-1$값으로 구분된 L개의 부분구간(subinterval)으로 나눈다. 스플라인 함수를 사용한 B-스플라인 기저 함수 $\phi_k(t)$는 다항식 부분의 차수인 m과 τ개의 간격으로 이루어진 노트로 결정된다. $B_k(t,\tau)$는 τ개의 노트로 정의된 B-스플라인 기저함수의 t 시점에서의 값을 나타내기 위해 사용된다. 여기서 k는 t값의 바로 왼쪽에 있거나 가장 큰 노트의 개수를 말한다. 초기의 노트에 더한 $m-1$개의 노트가 계산되고, 첫 번째 m차 B-스플라인 기저 함수는 왼쪽 경계에서 시작하는 모든 토대(support)를 가진다. 이러한 표기법은 모든 내부 노트가 이산형인 일반적인 경우에 $m+L-1$개의 기저 함수를 나타낸다. 이 표기법에 따라 이산적인 내부 노트를 가진 스플라인 함

수 $S(t)$는 다음과 같이 나타낸다.

$$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau)$$

2.2 데이터 조정(data registration)

함수적 데이터 분석에서 두 번째로 수행할 단계인 데이터 조정은 $x(t)$뿐만 아니라 t에서 관측된 x에서의 조정을 의미한다. 이 과정을 통해 어떤 곡선들을 재조정하여 분석을 용이하게 하기 위함이다. 이러한 조정을 하는 방법에는 이동 조정(shift registration), 경계표(landmarking), 일반적 변환(general transformation)등이 있다. 그 중에서 가장 많이 쓰는 이동 조정법에 대해 살펴보자. 이동 조정은 정렬하는 방법의 한 형태로 적당한 상수를 더하여 이동하는 방법으로 아래와 같이 나타낼 수 있다.

$$x_i^* = x_i(t + \delta_i)$$

이러한 이동 조정에서 가장 중요한 것은 적당한 상수가 고정된 효과(fixed effects), 장애 효과(nuisance effects), 임의효과(random effects)인지 구별하는 것이 중요하다. 고정된 효과는 이동이 우리의 관심에서 조금 벗어난 것을 말하며, 장애 효과는 이동에 대해 설명되어져야 하지만 우리의 관심 밖에 있는 것을 말한다. 임의효과는 이동이 곡선의 형태에 영향을 주는 것을 의미한다.

2.3 함수적 선형 모형(functional linear models)

독립변수(independent variable)를 g개의 효과 그룹으로 나누어 종속변수(dependent variable)에 대한 효과를 알아보고 그룹별로 차이가 유의한지에 대해 알아보기 위해 함수적 선형 모형을 적합 시키고자 한다. 종속변수는 함수적 관측치인 y이고, g개 효과 그룹이 유의하게 차이가 있는지에 대한 분산분석의 문제 이므로 함수적 분산분석 (functional analysis of variance; *FANOVA*)이라고 할 수 있다. g그룹에서 m번째 관측치를 y_{mg}라 하면 함수적 회귀 모형을 다음과 같이 나타낼 수 있다.

$$y_{mg}(t) = \mu(t) + \alpha_g(t) + \epsilon_{mg}(t) \tag{4}$$

함수 $\mu(t)$는 전체 평균 함수를 나타내며 α_g는 g그룹에서 y에 대한 효과 함수를 나타낸다. $\epsilon_{mg}(t)$는 잔차 함수를 의미하며 그룹 g에서 m번째 설명되지 않은 변동성을 나타낸다. 각 그룹의 효과를 파악하기 위해서 아래와 같은 제약조건이 필요하다.

$$\sum_g \alpha_g(t) = 0 \text{ for all } t$$

식 (4)를 이용하여 전체 평균 함수와 각각의 열이 각각의 그룹을 나타내는 $k \times (g+1)$인 계획행렬(design matrix) Z를 사용한 식으로 표현할 수 있다. β_j를 $\beta_1 = \mu, \beta_2 = \alpha_1, \cdots, \beta_{g_{+1}} = \alpha_g$ 로 정의하면 식(7)를 다음과 같이 나타낼 수 있다.

$$y_{mg}(t) = \sum_{j=1}^{g+1} z_{mg}\beta_j(t) + \epsilon_{mg}(t) \tag{5}$$

식 (5)를 다음과 같은 행렬식으로 나타낼 수 있다.

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{6}$$

$\boldsymbol{\beta}$는 잔차제곱합을 최소로 만드는 최소제곱법의 벡터 표현 $\| \boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta} \|^2$로 나타낼 수 있으며 함수적 데이터의 경우에도 최소 제곱법은 제곱된 노름(norm)의 형태로 나타낼 수 있다. 일반적인 최소제곱법은

$$LMSSE(\boldsymbol{\beta}) = \int [\boldsymbol{y}(t) - \boldsymbol{Z}\boldsymbol{\beta}(t)]' [\boldsymbol{y}(t) - \boldsymbol{Z}\boldsymbol{\beta}(t)] \, dt \tag{7}$$

이 된다. 제약조건 $\sum_{j=2}^{g+1} \beta_j = 0$ 하에서 $LMSSE(\boldsymbol{\beta})$를 최소로 만드는 것은 함수적 모수 μ와 α_g의 최소제곱 추정치 $\hat{\boldsymbol{\beta}}$이다. 이러한 방식으로 $\hat{\boldsymbol{\beta}}$를 구함으로써 각 그룹에 대한 효과를 추정하고, 그룹간의 효과가 유의한지 파악할 수 있다.

3. 함수적 데이터 분석 기법의 적용

3.1 옥션 데이터(auction data)

현재 우리나라의 옥션 중고장터(http://used.auction.co.kr)라는 사이트에서 다양한 중고 물품들이 거래되고 있다. 경매 형식으로 이루어지는 방식으로 판매자는 물품에 대한 상세 내용과 함께 경매 시작 금액을 정하여 거래를 한다. 경매 시작 금액은 1,000원부터 원하는 대로 설정할 수 있으며, 경매에 참여하지 않고 즉시 구매할 의사가 있는 구매자를 위한 즉시구매 가격을 등록할 수 있다. 경매에 올라온 물품들은 판매자의 설정에 의해 3/5/7일 간의 진행 기간 동안 이루어진다. 이 기간 동안 구매를 원하는 사람들이 입찰 가격을 매기고, 경매 진행 기간 동안에 최고 매입 가격을 매긴 입찰자에게 최종 가격으로 판매가 이루어지는 방식으로 운영되고 있다. 이러한 옥션 데이터는 시간에 따라 이산형으로 관측되는 형태로 이를 함수적 데이터 분석에 적용하고자 한다.

옥션 중고장터에서 CANON EOS600D 카메라에 대한 경매 중 총 8개의 입찰 기록을 수집하였다. 다음의 <그림 1>은 옥션 중고장터 사이트에서 수집한 CANON EOS600D 카메라 중 하나에 대한 상위 경매기록이다.

입찰자 ID	입찰일자	입찰가격	수량	누적수량
jjd17***	2014-10-30 21:34:33.370	620,000 원	1 개	1 개
sung8***	2014-10-30 19:17:29.647	615,000 원	1 개	-
seyangc***	2014-10-30 17:34:12.473	610,000 원	1 개	-
jjd17***	2014-10-30 16:16:28.240	605,000 원	1 개	-
abj7***	2014-10-30 13:56:05.640	600,000 원	1 개	-
seyangc***	2014-10-30 13:15:10.447	595,000 원	1 개	-
kjs6jac***	2014-10-30 07:58:01.570	590,000 원	1 개	-
abj7***	2014-10-29 23:00:36.150	580,000 원	1 개	-
hanssem2***	2014-10-29 17:32:28.110	575,000 원	1 개	-
VUTHIHUYE***	2014-10-29 15:15:49.010	570,000 원	1 개	-
seyangc***	2014-10-29 10:28:14.697	565,000 원	1 개	-
kjs6jac***	2014-10-29 07:40:31.260	560,000 원	1 개	-
VUTHIHUYE***	2014-10-28 23:56:48.027	550,000 원	1 개	-
abj7***	2014-10-28 09:55:26.390	540,000 원	1 개	-
jins4***	2014-10-28 08:38:09.873	535,000 원	1 개	-

<그림 1> 경매 기록

2014년 10월 23일 12시부터 진행된 이 경매는 2014년 10월 30일 23시까지 진행 되었다. 시작가는 1,000원으로 하여 총 입찰 수는 51회이고, 최종 입찰가격은 620,000원에 형성되었다.

다음의 <표 1>과 <표 2>는 8개의 경매 자료에 대한 요약 통계량(summary statistics)을 연속형 변수(continuous variable)와 범주형 변수(categorical variable)으로 나눈 것이다.

<표 1> 연속형 변수에 대한 요약 통계량

변수명	개수	평균	중앙값	최소값	최대값	표준편차
시작가(원)	8	3250	1000	1000	10000	4166.19
입찰가(원)	8	538125	515000	420000	710000	90787.72
입찰횟수(회)	8	41.16	43	18	54	11.67
판매자점수(점)	8	98.88	100	94	100	2.23

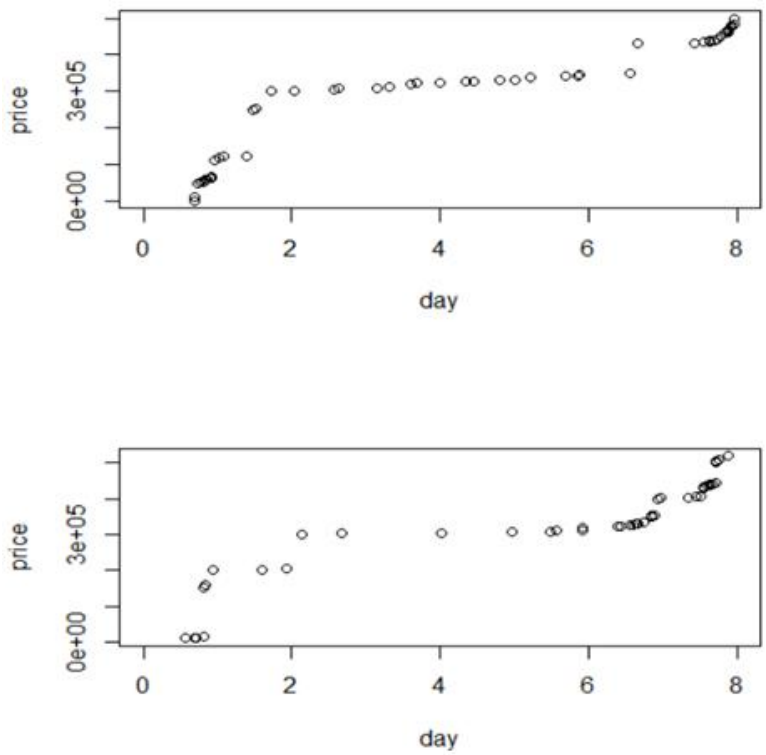
<표 2> 범주형 변수에 대한 요약 통계량

변수명	범주	개수	비율
상태	새제품	2	25%
(condition)	중고제품	6	75%
점프비딩	있다	5	62.5%
(jump bidding)	없다	3	37.5%

연속형 변수들 중 시작가는 판매자가 경매를 시작하기 위해 정한 가격을 말하며 입찰가는 최종적으로 낙찰된 가격이다. 입찰횟수는 경매 기간 동안 이루어진 입찰 기록의 수이고 판매자점수는 경매를 올린 판매자의 별점 평가 점수를 의미한다. 범주형 변수들 중 상태(condition)는 물품이 새 제품인지 중고제품인지를 의미한다. 비록 중고장터이지만 간혹 새 제품으로 명시하고 중고장터에 경매로 올라오는 경우가 있다. 점프비딩(jump bidding)은 어떤 시점에서 현재가 대비 50% 이상 더 높은 가격으로 그 다음 시점에 입찰을 매긴 것의 유무를 말한다.

3.2 옥션 데이터의 함수적 데이터 분석 수행 단계

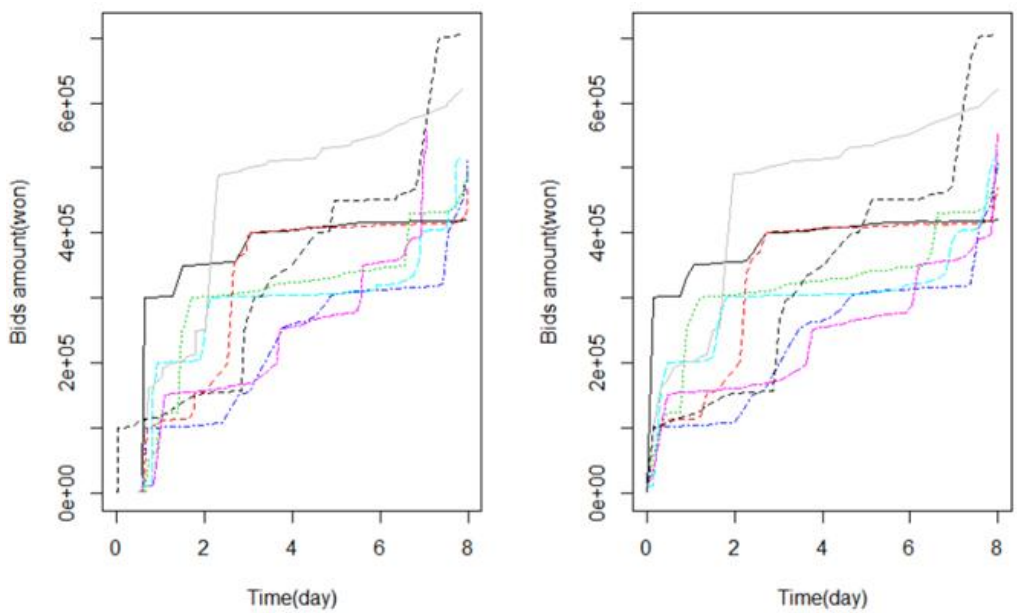
다음의 <그림 2>는 옥션 자료 중 2개의 산점도를 그린 결과이다.



<그림 2> 두 개의 옥션 데이터 자료 산점도

<그림 2>의 상단의 산점도에 해당하는 옥션은 총 54회의 입찰이 이루어졌고, 아래의 산점도는 47회의 입찰이 이루어졌다. 두 개의 산점도를 비교해보면, 각 입찰이 이루어진 시간(day)과 입찰가격(price)이 서로 다름을 확인할 수 있다. 모든 경매에서는 경매 참여자의 참여 시간이 각기 다르기 때문에 시

간을 동일한 간격으로 조정해 줄 필요가 있다. 시간 t_j를 0일 부터 8일까지 동등하게 54개로 나누어서 입찰가격 값을 조정해 주었다. 이 과정은 선형보간법(linear interpolation)을 사용하였다. 이 과정을 통해 새로 지정된 시간에 대한 입찰 가격을 $y^{(j)} = (y_1^j, \cdots, y_{54}^j)$로 표현할 수 있다. <그림 3>은 선형보간법을 적용하기 전과 후의 그래프이다.



<그림 3> 선형보간법 전과 후의 그래프

8개의 옥션 데이터의 경매 시작 시간과 끝나는 시점이 각각 다른 것을 <그림 3>의 왼쪽 그래프를 통해 확인할 수 있다. 이에 대해 모든 옥션의 시간을 동일한 구간으로 맞춰 주기 위해 선형보간법을 취하여 얻어진 새로운

$y^{(j)}$와 시간 축을 조정하여 시간에 대한 그래프를 그린 결과가 오른쪽이다.

조정된 시간과 가격에 대한 데이터를 함수형태로 나타내기 위해서 페널티를 부과한 평활화 스플라인(penalized smoothing spline)을 적용시켜보았다. 데이터를 함수형태로 나타내기 위해서 B-스플라인 기저를 사용하였고, 다항색의 스플라인 함수는 다음과 같이 나타낸다.

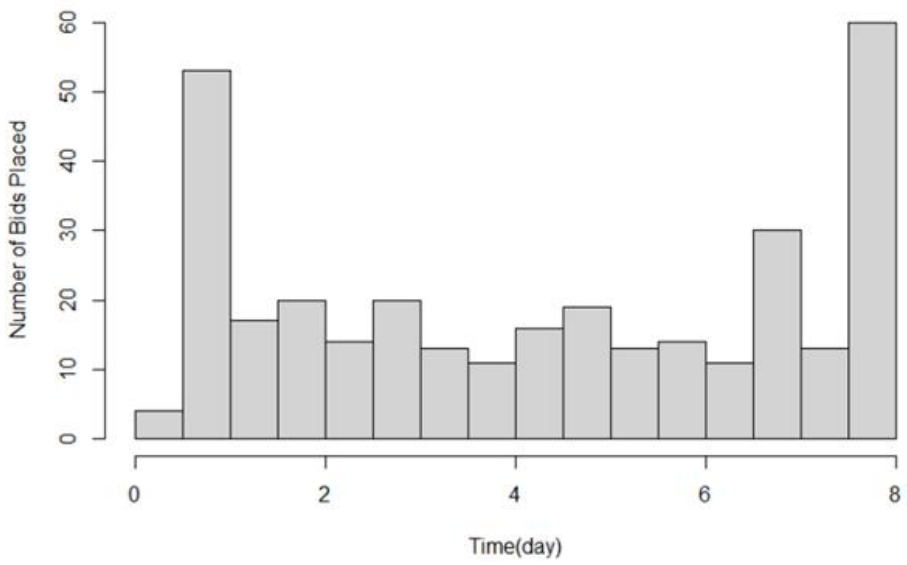
$$y_j = x(t) + \epsilon_j = \sum_{k}^{K} c_k \phi_k(t) + \epsilon_j \tag{8}$$

이를 통해서 페널티를 부과한 오차제곱합(PENSSE)을 최소로 하는 해의 값을 찾기 위함이다. 페널티를 부과한 경우가 페널티를 부과하지 않은 경우에 비해 계산적으로 더 효율적이고 유용한 결과를 주기 때문에 페널티를 적용하여 평활법(smoothing)을 적용하였다.

$$PENSSE_{\lambda}(x|y) = [y - x(t)]' W[y - x(t)] + \lambda PEN_2(x) \tag{9}$$

여기서 $PEN_2(x)$는 페널티 함수로써 t에 대한 함수의 2차 미분의 제곱을 적분한 것으로 $PEN_2(x) = \int [D^2 x(s)]^2 ds$ 이다. λ는 평활화 모수(smoothing parameter)로 데이터에 대해서 적합 시킨 값 간의 교환율(rate of exchange)를 의미하며 λ값이 커질수록 함수는 더 부드러워 진다. 평활화 모수 값을 찾기 위해서 일반화된 근사모델의 정확도 평가방법(Generalized Cross Validation; GCV)을 이용하여 GCV값이 최소가 되는 값을 찾은 결과 $\lambda = 0.1$로 선택하였다.

스플라인의 노트(knots)를 결정하기 위해 시간에 따른 입찰의 빈도수를 <그림 4>의 히스토그램으로 확인하였다.

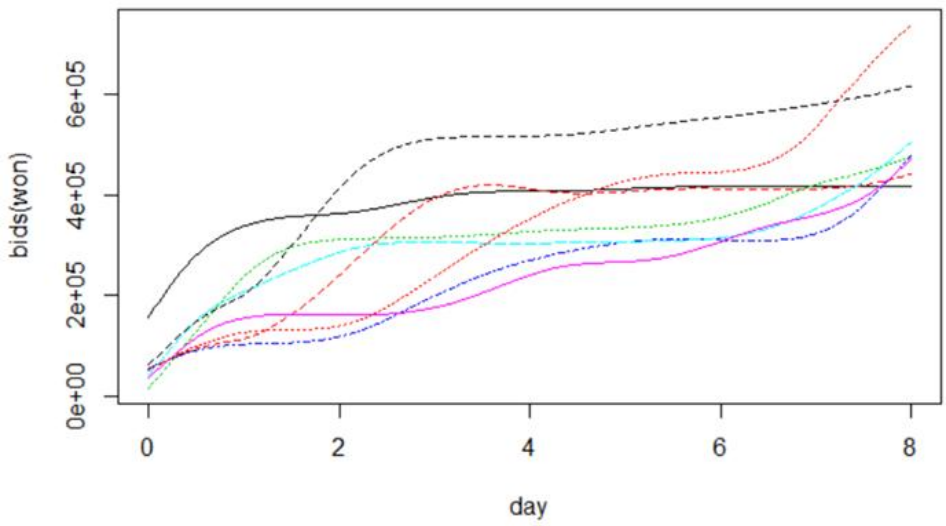


<그림 4> 입찰 횟수에 대한 히스토그램

입찰 초기인 1일 동안 상당히 많은 수의 입찰이 이루어졌고, 1일부터 6.5일 까지는 상대적으로 저조한 입찰이 이루어졌다. 입찰 마감시간이 다가옴에 따라 6.5일부터 7일, 특히 7.5일부터 8일 사이에 입찰이 급격히 증가한 것을 확인할 수 있다. <그림 4>의 히스토그램을 근거로 하여 스플라인의 노트를 다음과 같이 설정하였다.

$$\{0.0, 0.3, 0.7, 1.0, 2.1, 3.2, 4.3, 5.4, 6.5, 6.8, 7.1, 7.4, 7.7, 8.0\} \quad (10)$$

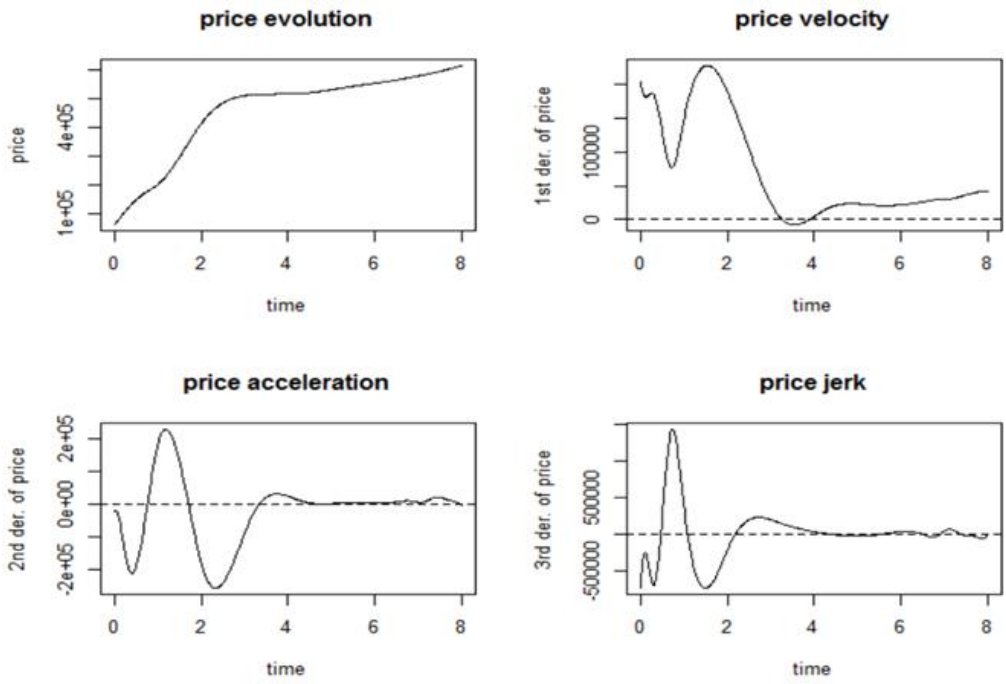
0일부터 1일까지는 3개, 1일부터 6.5일까지는 5개, 6.5일부터 8일까지는 5개로 총 13개의 노트로 구성하였다. 또한 함수의 3차 미분계수까지의 그래프를 살펴보기 위해 스플라인의 차수는 4차로 정하였다. 이러한 조건으로 데이터를 부드러운 함수형태로 변환한 결과는 <그림 5>와 같다.



<그림 5> 페널티를 부과한 평활화 스플라인

평활화를 통해 데이터가 부드러운 함수 형태인 $f(t)$로 표현 되었다. 이 함수 자체를 가지고 분석을 하고, 또한 함수의 미분계수간의 관계를 확인해보는 작업을 할 것이다.

함수적 데이터 분석에서는 함수 자체에 대한 그래프뿐만 아니라 함수에 대한 미분계수에 대한 그래프와 함께 동태를 파악하는 것도 중요하다. <그림 6>은 함수형태로 변환한 입찰가격 중 어느 하나의 옥션에 대한 플롯(plot)과 1차, 2차, 3차 미분계수에 대한 함수, 즉 $f'(t), f''(t), f'''(t)$ 의 플롯이다. $f'(t)$는 $f(t)$의 1차 미분으로 입찰가격의 속도(velocity)를 의미하고 $f''(t)$는 $f(t)$의 2차 미분인 입찰 가격의 가속도(acceleration)이다. 그리고 $f'''(t)$는 $f(t)$의 3차 미분인 입찰 가격의 가속도의 변화율로 저크(jerk)라고 표현된다.



<그림 6> 입찰가격에 대한 플롯과 미분계수에 대한 플롯

<그림 6>에서 왼쪽 상단에 있는 플롯은 시간에 대한 입찰가격의 플롯이다. 0일부터 2일까지는 거의 같은 기울기로 가격이 상승하는 것으로 보이며 2일부터 4일 사이에서는 증가가 거의 이루어지지 않는 것으로 보인다. 4일 이후부터 경매가 종료되는 시점까지는 꾸준히 증가하기는 하나 그 기울기가 입찰 초기인 0일부터 2일까지 보다는 작은 것으로 관측된다. 하지만 함수의 증감에 대한 변화율을 보기 위해서는 $f(t)$를 한 번 미분한 $f'(t)$의 플롯을 확인하여야 한다. 오른쪽 상단에 있는 플롯이 $f'(t)$, 즉 입찰가격의 변화 속도이다. 0일부터 1일까지는 $f(t)$의 기울기인 속도가 감소하다가 1일부터 2일까지는 증가하는 형태를 보인다. 그리고 이 이후부터 4일까지는 계속 감소하다가 4일

부터는 증가하는 추세를 가진다. 3일에서 4일 사이에서는 속도 값이 음의 값으로 관측된 것을 확인할 수 있다. 이는 평활화 스플라인 과정에서 오차가 발생하여 단조 증가하는 형태를 완벽하게 나타내지 못한 것이다. 평활화 모수의 값을 현재 설정한 값보다 크게 한다면 전 시점에서 단조 증가하며 더 부드러운 평활화 스플라인을 얻을 수 있다. 입찰가격의 속도에 대한 변화율인 $f''(t)$에 대한 플롯은 왼쪽 하단에 있는 그림이다. 0일부터 1일까지는 음수 값을 가지며 1일부터 2일까지는 양의 값을 가지는 것으로 확인된다. 그 이후부터 4일까지는 음의 값을 가지다가 4일 이후부터는 입찰가격의 속도에 대한 변화율이 거의 0을 유지하는 것으로 보인다. 마지막으로 오른쪽 하단에 있는 입찰가격의 가속도에 대한 변화율인 $f'''(t)$는 대부분의 t시점에서 음수 값 또는 0의 값을 가지지만 1일 전후로 양의 값을 가지며 1일이 되는 시점에서 가속도의 변화율이 가장 큰 것으로 보인다. 함수 자체인 $f(t)$만의 플롯을 보고 확인할 수 없었던 함수의 변화율, 속도 및 가속도에 대한 변화율을 이와 같이 확인함으로써 함수에 대한 동태를 파악할 수 있다.

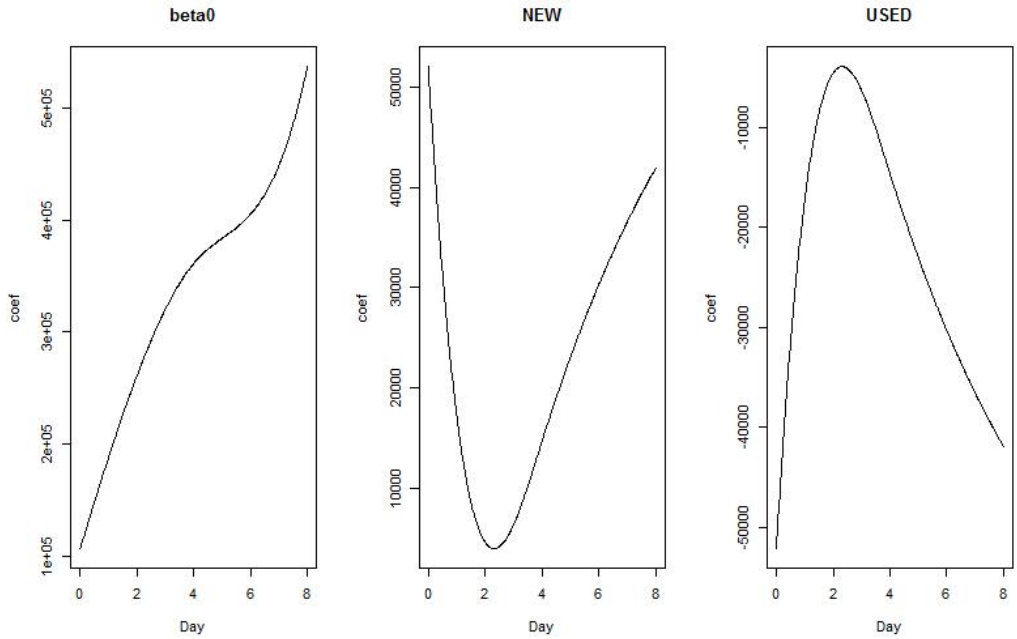
3.3 함수적 회귀 분석

3.2절에서 관측된 데이터를 함수 형태로 변환하여 기본적인 함수적 데이터 분석을 수행하였다. 변환된 함수적 데이터를 이용하여 함수적 회귀 분석을 수행해 보았다. 옥션 데이터에서 얻을 수 있는 변수들 중 <표 2>에서 언급한 범주형 변수를 사용하여 각 변수에 대한 그룹의 효과의 유의성을 알아보기 위해 함수적 선형 모형(functional linear models)을 세웠다.

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{2} x_{ij}\beta_j(t) + \epsilon_i(t) \qquad (11)$$

범주형 변수로 상태와 점프비딩이 있는데 각각 새제품(NEW)/중고제품(USED), 유(YES)/무(NO)로 되어있으므로 그룹을 두 개로 나눌 수 있다. 이에 따라 각 변수의 효과를 확인하기 위해서 $\sum_{j=1}^{2}\beta_j(t) = 0$ for all t 의 제약조건을 주었다.

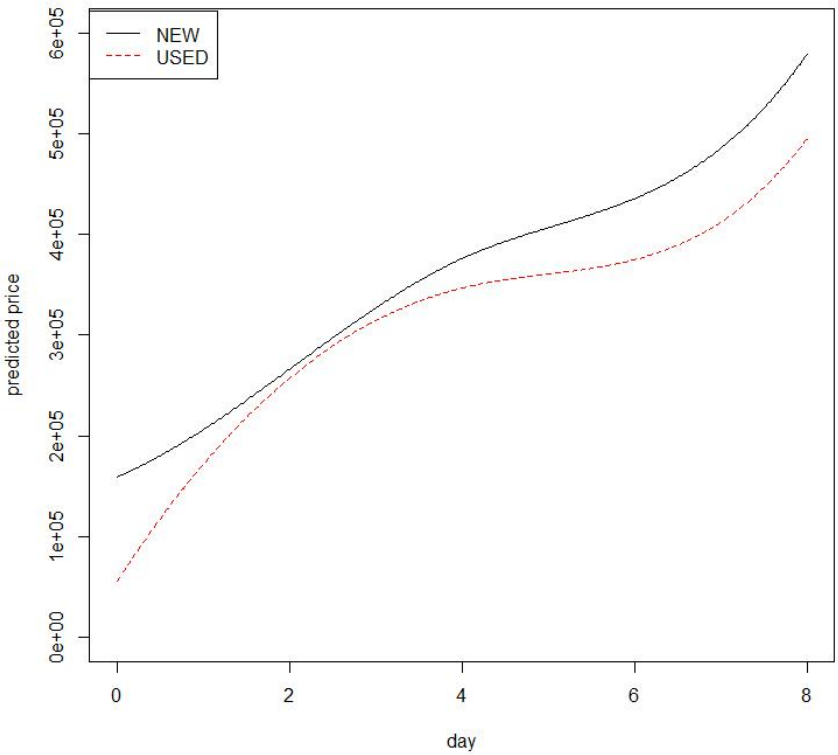
다음의 <그림 7>은 상태변수인 새제품과 중고제품에 대한 입찰가격을 함수적 선형 모형으로 적합한 회귀계수(coefficients)에 대한 결과이다.



<그림 7> 상태에 대한 입찰가격의 함수적 선형 모형의 회귀 계수

첫 번째 플랏은 상태에 대한 입찰가격의 함수적 선형 모형을 적합 시킨 결과의 β_0, 즉 절편의 기울기이자 평균함수이다. 시간이 지남에 따라 입찰가격의 평균은 꾸준히 증가하는 추세를 보이고 있다. 두 번째 플랏은 β_1, 즉 새제품에 대한 회귀계수에 대한 플랏이다. 경매가 시작된 입찰 초기에는 계수 값이 크지만 경매가 시작되고 중간 시점에는 상대적으로 값이 떨어져있다. 그리고 경매 마감시간이 다가옴에 따라 점차 증가한다. 이를 통해서 상태가 새제품일 경우 경매 초기와 경매 종료 시간에 가격에 대한 예측 값이 높은 값으로 매겨지지만 경매가 진행되고 중간 쯤 되는 시점에는 그 값이 상대적으로 낮음을 알 수 있다. 반면에 중고제품의 경우, 회귀계수 값이 모두 음수 값을 가지는 것으로 보아 가격에 대한 중고제품의 효과는 음의 관계를 가지고 있음을 알 수 있다.

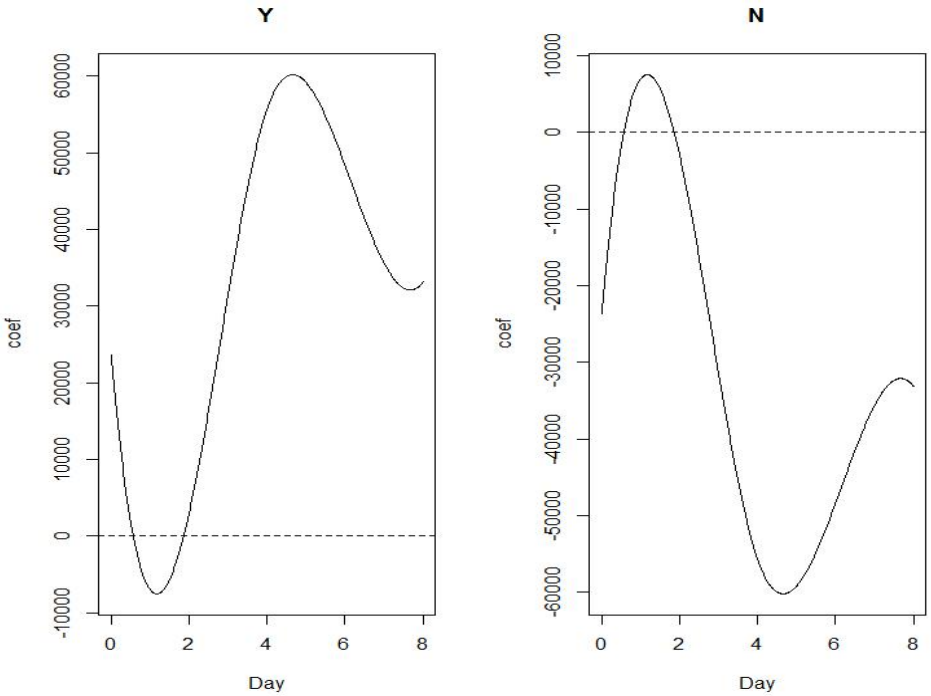
다음의 <그림 8>은 새제품과 중고제품에 대한 가격 예측 값을 그린 결과이다.



<그림 8> 상태에 대한 가격 예측 값

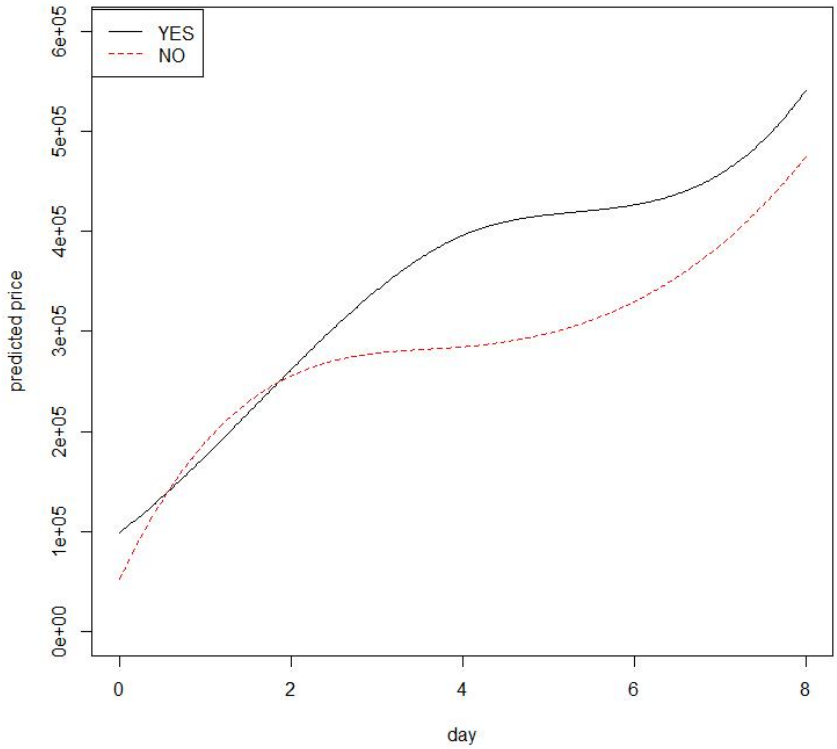
실선은 새제품의 가격 예측 값이고, 점선은 중고제품의 가격 예측 값이다. 전체적으로 새제품이 중고제품보다 가격이 더 높게 예측되는 것을 확인할 수 있다. 특히 경매 초기와 경매 마감시간 전에 새제품과 중고제품의 가격 예측 값 차이가 중간 시점에 비해 크게 나타나는 것을 확인할 수 있다.

<그림 9>는 점프비딩의 유무에 대한 입찰가격의 함수적 선형 모형을 적합시킨 회귀 계수에 대한 플랏이다.



<그림 9> 점프비딩에 대한 입찰가격의 함수적 선형 모형의 회귀 계수

점프비딩이 있는 그룹의 회귀계수인 왼쪽 플랏을 보면 경매 초기에 계수 값이 음수인 부분이 존재하고, 경매 중간 시점부터 계수 값이 증가하여 4일에서 6일 사이에 높은 값으로 나타난다. 그리고 경매 후기에 오히려 값이 낮아지는 것으로 보아 점프비딩이 있는 경우, 경매 초기나 경매 마감 때의 가격 예측 값보다 경매 중반에 가격이 더 높게 예측 되는 것을 알 수 있다. 반면에 점프비딩이 없는 그룹에 대한 회귀계수인 오른쪽 플랏을 보면 알 수 있듯이 입찰 초기에 잠깐 양수 값을 가지나 전반적으로는 음수 값을 가진다.



<그림 10> 점프비딩에 대한 가격 예측 값

<그림 10>은 점프비딩의 유무에 따른 가격 예측 값에 대한 플랏이다. 실선은 점프비딩이 있는 경우의 가격 예측 값을 의미하고, 점선은 점프비딩이 없는 경우의 가격 예측 값을 의미한다. 전체적으로 보았을 때, 점프비딩이 있는 경우의 가격 예측 값이 점프비딩이 없는 경우의 가격 예측 값보다 대체적으로 높게 예측 되었다. 점프비딩은 3.1절에서 언급한 것과 같이 어떤 시점에서 현재가 대비 50% 이상 더 높은 가격으로 그 다음 시점에 입찰을 매기는 것을 의미한다. 일반적으로 점프비딩은 경매 초기나 경매 마감 전이 아닌 중간시점에 입찰이 활발히 일어나지 않고, 정체된 시점에 갑작스럽게 입찰가격

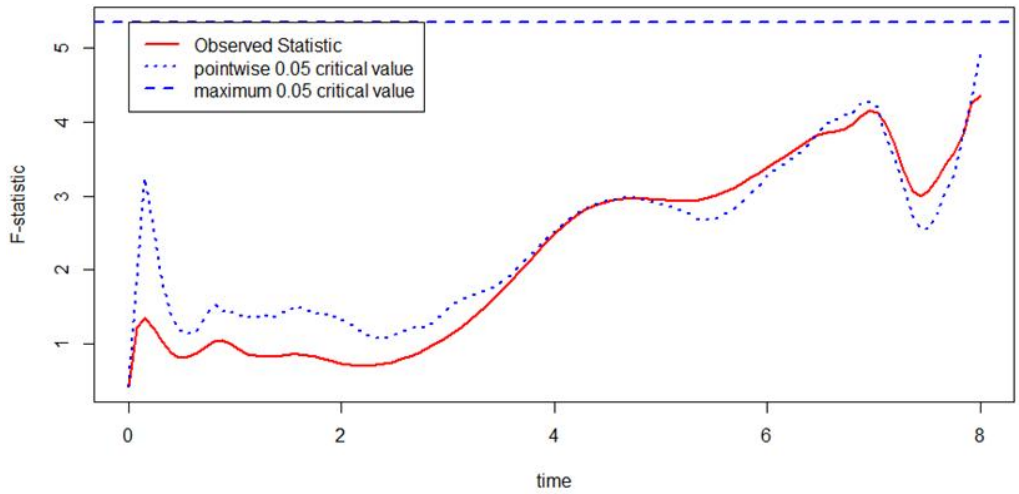
을 올리는 경우가 많다. 그렇기 때문에 경매 초기나 마감 전 보다 경매 중간 시점인 4일부터 6일 간에 점프비딩의 유무에 따른 가격 예측 값의 차이가 큰 것으로 보인다.

3.3 함수적 F 검정

함수적 선형 회귀 분석의 일반적인 경우에서 단변량 F 통계량의 함수형(functional) 버전으로 F 통계량을 변형하여 함수적 F 검정(functional F -test)를 시행하였다.

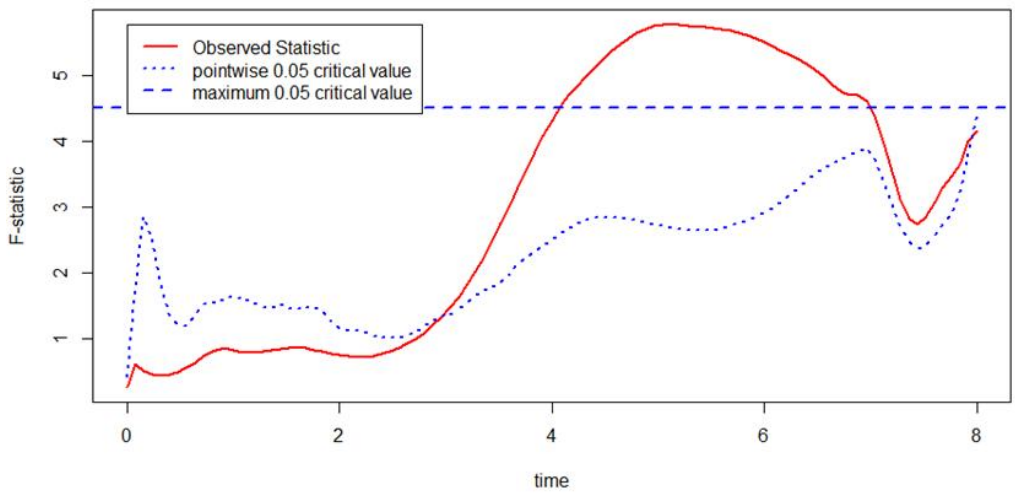
$$F(t) = \frac{Var[\hat{y}(t)]}{\frac{1}{n}\sum(y_i(t) - \hat{y}(t))^2} \tag{12}$$

식 (12)를 이용하여 $F(t)$의 최대값을 계산함으로써 하나의 수로 계산을 하고 이를 순열검정법(permutation test)으로 수행한다. 이에 대한 보다 자세한 내용은 Ramsay와 Hooker(2009)의 Functional Data Anlaysis with R and MATLAB을 참고하면 된다.



<그림 11> 상태 변수에 대한 순열검정

<그림 11>은 상태변수의 새제품과 중고제품의 효과가 유의하게 차이가 있는지 확인하기 위하여 순열검정을 한 결과이다. 실선은 상태 변수에 대한 값을 식(16)으로 계산하여 최대값을 계산한 F 통계량 값이다. 이 선이 최대 0.05 임계치인 파선을 넘어가는 시점에 한해서 두 그룹의 효과가 유의하게 차이가 난다고 할 수 있다. 하지만 <그림 11>에서 볼 수 있듯이 전 시점에서 실선이 파선을 넘는 곳이 없으므로 시간에 따라 상태 변수의 두 그룹에 대한 효과가 유의하게 차이가 나는 시점이 없다고 할 수 있다. 이는 옥션의 중고장터 특성상 중고 제품을 거래하는 곳이므로 새제품인지 중고제품인지가 가격을 결정하는 데에 중요한 요소는 아닌 것으로 보인다.



<그림 12> 점프비딩 변수에 대한 순열검정

<그림 12>는 점프비딩 변수의 유/무 효과가 유의하게 차이가 있는지 확인하기 위하여 순열검정을 한 결과이다. 실선은 점프비딩 변수에 대한 값을 식(16)으로 계산하여 최대값을 계산한 F 통계량 값이다. 이 선이 최대 0.05 임계치인 파선을 넘어가는 시점에 한해서 두 그룹의 효과가 유의하게 차이가 난다고 할 수 있다. <그림 12>의 4일에서 7일까지의 실선을 보면 이 시점에서만 실선이 파선을 넘어가는 것을 확인할 수 있다. 즉, 4일에서 7일까지의 시간 동안에는 점프비딩이 있는 경우와 없는 경우에 대한 입찰 가격의 효과가 유의하게 차이가 있다고 할 수 있다. 점프비딩이 주로 이루어지는 시점은 경매 기간 중 중간 시점인 4일에서 6일, 7일 사이 이므로 이 기간동안에는 점프비딩의 유무에 따라 가격이 유의한 차이를 나타내는 것으로 해석할 수 있다.

6. 결론 및 토의과제

본 연구에서는 시간의 흐름에 따라 관측되는 데이터를 기존의 분석 방법으로는 제한된 분석만을 할 수 있다는 단점을 보완하는 해결책으로 비모수적 방법 중 하나인 함수적 데이터 분석을 이용하여 함수 자체로 변환하고 분석하여 실제 데이터를 적용시켜 보았다.

본 연구의 결과를 요약하면 다음과 같다.

(1) 함수의 동태 변화를 고려하지 않는 기존 방식과는 달리 함수적 데이터로 변환하여 함수 자체뿐만 아니라 미분간의 관계에 대한 동태를 파악할 수 있었다.

(2) 함수 형태로 함수적 회귀분석을 적합하여 설명변수에 대한 종속변수의 변동 및 예측을 할 수 있었다.

따라서 시간에 의존하는 데이터의 경우 함수형태로 변환하여 함수적 데이터 분석을 하여 데이터에 대한 정보의 손실을 줄일 수 있고, 함수의 동태를 파악하는데 더 이점이 있다고 판단된다.

함수적 회귀 분석에서 예측 모형에 대한 정확성을 더 높이기 위해서는 관측 데이터의 수를 증가시키는 것이 요구된다. 또한 옥션 데이터에서 보다 정확하고 시간에 따른 입찰 가격에 대한 동태를 예측하기 위해 시간에 따라 변하는 설명변수(time-varying predictor variable)를 적용하여 함수적 회귀 분석을 수행하는 연구가 이루어질 필요가 있다.

참고 문헌

<1> 안홍세 (2005). *함수적 데이터 분석의 실제 적용에 관한 연구*. 한국외국어대학교 석사논문, pp. 1-10

<2> Ramsay, J.O. and Silverman, B.W. (2005), Functional Data Analysis, 2nd ed., Springer, New York

<3> Ramsay, J.O. and Silverman, B.W. (2002), Applied Functional Data Analysis: Methods and Case Studies, Springer, New York

<4> Ramsay, J.O., Hooker, G. and Graves, S. (2009), Functional Data Analysis with R and MATLAB, Springer, New York

<5> Wang, S., Jank, W. and Shmueli, G. (2008). Explaining and Forecasting Online Auction Prices and Their Dynamics Using Functional Data Analysis. *Journal of Business & Economic Statistics*, Vol.26 No.2, pp. 144-153