이 성 건 교수지도

석사학위 청구논문

앙상블(ensemble) 기법을 이용한 일반화가법모형(GAM) 연구

- 연속형 반응변수를 중심으로 -

2013

성신여자대학교 대학원

통 계 학 과

김 하 나

앙상블(ensemble) 기법을 이용한 일반화가법모형(GAM) 연구

- 연속형 반응변수를 중심으로 -

이 성 건 교수지도

이 논문을 석사학위논문으로 제출함

2012년 11월

성신여자대학교 대학원

통계학과

김 하 나

인 준 서

김하나의 석사학위 논문으로 인준함.

심사위원 _____인

심사위원 _____인

심사위원 _____인

성신여자대학교 대학원

논 문 개 요

통계학에서 주로 사용되는 일반화선형모형(Generalized Linear Model; GLM)은 변수 간 인과관계를 설명하는데 좋은 도구로서 많이 이용된다. 그러나 많은 현실 문제에서는 선형가정으로 설명될 수 없는 비선형 효과들이 존재한다. 이러한 비선형 효과를 효율적으로 설명할 수 있는 방법론으로 일반화가법모형이 이용된다. 일반화가법모형(Generalized Additive Models; GAM)에서 평활함수를 사용하여 비선형 효과를 확인할 수 있으며, 이 평활함수(smoothing function)의 추정은 로컬스코어링 알고리즘(local scoring algorithm)을 통해서 이루어진다.

한편, 자료에 대한 예측에서 배깅(bagging), 랜덤포레스트(random forest) 등의 앙상블 예측 기법이 단일 예측보다 더 좋은 성능을 가질 수 있음이 많은 연구를 통해 증명되었다. 앙상블 기법은 붓스트랩 표본에서 단일 예측자를 결합하여 변동이 작고 예측력이 좋은 결합 예측자를 구하는 것이다.

이항 반응변수에서 일반화가법모형을 단일 분류자로 하는 앙상블 분류자는 De Bock(2010)에 의해 제안되었으며, 또한 이 앙상블 분류자가 단일 분류자나 다른 앙상블 분류자보다 더 좋은 수행능력을 보이는 것을 추후 연구를 통해 증명하였다(De Bock & Van den Poel, 2012).

본 논문에서는 연속형 반응변수에서의 일반화가법모형을 단일 예측자로 하는 앙상블 기법을 제안하고자 한다. 또한 모의실험과 국내 요양보험자료를 통해 제안한 앙상블 기법의 우수성을 확인하고자 하였다.

목 차

제 1 장 서 론

많은 자료 분석에서 선형회귀분석은 반응변수에 대한 예측과 분류, 그리고 설명변수와의 연관관계를 이해하는데 좋은 도구로서 자주 사용된다. 그러나 현실 문제에서는 선형회귀분석으로 설명되는 선형관계만 존재하지 않는다. 따라서 선형관계를 가정한 선형모형은 자료가 말하고자 하는 전부를 읽어내기 어려울 수 있다. 이렇게 선형관계가 아닌 비선형적인 관계를 설명할 수 있는 좋은 통계방법으로서 일반화가법모형(Generalized Additive Models; GAM)이 있다(Hastie & Tibshirani, 1990).

일반화가법모형은 일반화선형모형의 확장 형태로서 비선형 함수의 결합을 통해서 반응변수의 비선형성을 모형화하고 해석하는 것을 가능하게 하는 방법론이다. 이러한 비선형성을 추정하기 위해서 평활함수(smoothing function)를 이용하며, 추정은 로컬 스코어링 알고리즘(local scoring algorithm)을 통해 가능하다.

한편, 반응변수에 대한 예측과 분류에서 단일 예측·분류자에 비해 다중의 예측·분류자를 결합하는 결합 예측 방법이 더 좋은 수행력을 보이는 것이 선행의 연구를 통해 밝혀졌다(Bauer & Kohavi, 1999; Bühlmann, 2002; Skurichina & Duin, 2002). 그러한 결합 예측 방법에는 배깅(bagging; Breiman, 1996), 랜덤포레스트(random forest; Breiman, 2001)등이 있으며 이러한 방법들은 모두 의사결정나무(decision tree)를 단일 분류자로 한다. 최근 이항 반응변수를 가지는 일반화가법모형과 배깅을 결합한 분류 방법론이 De Bock(2010)에 의해 제안되었다. 일반화가법모형을 이용한 결합 예측 방법론은 기존의 비모수적인 방법인 의사결정나무를 이용한 방법론과는 달리 자료의 비

선형성을 모형에 포함하여 반응변수에 대한 예측의 성능을 향상시킬 수 있다 (De Bock, 2012).

본 논문에서는 기존의 이항 반응변수에서 나아가 연속형 반응변수에 대한 일반화가법모형을 단일 예측자로 하여 결합한 예측 방법을 제안하고자 한다. 제안한 방법론이 기존의 일반화가법모형에 의한 단일 예측이나, 의사결정나무를 단일 예측자로 하는 배깅, 랜덤포레스트에 비해 좋은 수행력을 보이는지 모의실험을 통해 확인하고자 한다. 2장에서는 일반화가법모형과 결합 예측 방법론인 배깅과 랜덤포레스트에 대해 간략하게 설명하였다. 3장에서는 모의실험을 통한 예측 방법론들을 비교 분석하였다. 또한 국내 요양보험 자료에 각 방법론을 적용한 결과를 요약하였다. 모든 분석은 통계 분석 프로그램인 R(R Development Core Team, 2012)을 이용하였다. 방법론들 간의 예측력의 비교 기준은 제곱근평균제곱오차(Root Mean Square Error; RMSE)를 사용하였다. 마지막으로 4장에서는 모의실험과 실제자료를 통한 분석 결과를 정리하고 본 연구의 한계점과 추후 연구에 대해 논의하였다.

제 2 장 일반화가법모형

일반화가법모형(Generalized Additive Models; GAM)은 일반화선형모형(Generalized Linear Models; GLM)을 포괄하는 모형으로서, 변수들의 선형모형과 평활함수(smoothing function)를 함께 고려하는 보다 확장된 방법론이다.

2.1. 소개

가법모형은 선형모형의 좋은 특징을 많이 가지면서 동시에 더 유동적이다. 선형모형의 장점은 결과의 해석에 있어서 보다 직관적이라는 것이다. 만약 선형모형에서 설명변수 x_j의 변화가 있을 때 반응변수에 대한 예측 값의 변화를 알고 싶을 때는 회귀계수 β_j만 알면 된다. 반면, 가법모형에서는 부분반응함수(partial response function) f_j가 회귀계수 β_j와 같이 모형에서의 반응변수와 설명변수간의 비선형성을 나타내는 역할을 한다.

일반화가법모형을 회귀식의 형태로 표현하면 아래 식(2.1)과 같이 나타낼 수 있다.

$$E(Y|X_1,X_2,\cdots,X_p)=\alpha+f_1(X_1)+f_2(X_2)+\cdots+f_p(X_p), \qquad (2.1)$$

여기서, $X_1,X_2,\cdots,X_p$는 설명변수이며, Y는 반응변수, f_j는 X_j에 대한 평활함수이다. 만약 모형의 각 함수를 선형기저함수(linear basis function)로 하였다면 모형 결과는 단순최소제곱법으로 적합한 결과와 동일하다. 그러나 가법모형에서는 삼차평활스플라인(cubic smoothing spline)이나 커널평활자(kernel smoother)등의 산점도평활자(scatterplot smoother)를 사용하여 적합하고, 모든 p개의 함수에 대해서 반복적으로 추정하는 것이 일반적이다. 또한 비모수

(non-parametric)적인 형태의 함수 f_j는 모형을 좀 더 유동적으로 만들어 준다. 반응변수 Y에 대한 조건부 평균 $\mu(X_1, X_2, \cdots, X_p)$는 연결함수(link function)인 g를 통한 가법함수와 관련이 있으며 이것을 식으로 정리하면 아래의 식 (2.2)와 같이 정의할 수 있다.

$$g[\mu(X_1, X_2, \cdots, X_p)] = \alpha + f_1(X_1) + \cdots + f_p(X_p). \tag{2.2}$$

이와 같은 연결함수는 일반적으로 아래 [표 1]와 같다.

[표 1] 지수족 분포에서의 연결함수

반응변수의 분포	연결함수
정규분포	$g(\mu) = \mu$
이항분포	$g(\mu) = \log it(\mu) = \log\left\{\frac{\mu}{(1-\mu)}\right\}$ $g(\mu) = probit(\mu) = \Phi^{-1}(\mu)$
포아송분포, 음이항분포	$g(\mu) = \log(\mu)$
감마분포	$g(\mu) = \mu^{-1}$
역가우스분포	$g(\mu) = \mu^{-2}$

반응변수가 지수족(exponential family)인 분포에서는 여러 형태의 연결함수가 가능하다. 일반화선형모형에서 지수족은 널리 알려져 있는데, 이는 일반화가법모형에서도 동일하게 이용될 수 있다. 또한 함수 f_j는 산점도 평활자를 기초로 하는 알고리즘으로 추정이 가능하며 추정된 함수 $\hat{f}_j$는 X_j의 비선형성 효과를 나타낼 수 있다. 그러나 모든 f_j가 반드시 비선형일 필요는 없으며 선형의 형태와 다른 모수적인 형태를 혼합할 수도 있다. 이러한 모형을 준모수적 모형(semi-parametric model)이라고 한다. 따라서 반응변수에 적절한 연결

함수를 이용하면 일반화가법모형을 적합할 수 있다. 위의 지수족과 연결함수에 대한 보다 자세한 이론은 Dobson & Barnett (2008), McCullagh & Nelder, (1989) 등을 참고하기 바란다.

다음 절에서는 일반화가법모형의 적합방법에 대해서 알아보도록 하겠다.

2.1.1. 가법모형의 적합

이번 절에서는 가법모형의 적합을 위한 알고리즘과 일반화에 대해서 알아보고자 한다. 일반화가법모형의 적합을 위해서는 크게 세 가지 알고리즘을 이용할 수 있는데, 평활스플라인 알고리즘(smoothing spline algorithm), 백피팅 알고리즘(backfitting algorithm), 로컬스코어링 알고리즘(local scoring algorithm)이 대표적인 방법이다.

일반화가법모형의 적합을 하기에 앞서 가법모형에 대해 알아보면 다음과 같다. 가법모형은 아래와 같은 식(2.3)으로 표현될 수 있다.

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

여기서 Y는 반응변수, $X_1, X_2, \cdots, X_p$의 설명변수이며, $f_j(X_j)$는 j번째 설명변수의 평활함수, ϵ은 평균이 0 분산이 σ^2를 가지는 정규분포를 따르는 오차항이다.

(1) 평활스플라인(smoothing spline)

가법모형에서 비선형 효과를 적합하기 위한 방법으로 산점도 평활자(scatterplot smoother)를 이용한다. 산점도 평활자란 2차원의 도표를 통해서 설명변수와 반응변수간의 관계를 모형화 하는 것으로, 대표적인 방법에는 단

순선형회귀(simple linear regression), 국소회귀(local regression), 평활스플라인(smoothing spline) 등이 있다. 여기서는 가법모형의 평활함수 f_j를 추정하기 위한 기저함수를 평활스플라인(smoothing spline)으로 하는 방법에 대해서 설명하고자 한다.

가법모형에서는 자료의 비선형성을 설명하고자 너무 많은 모수를 사용하여 과적합(overfitting)을 하는 경우가 있다. 이때의 해결책으로 모수가 추가될 때마다 모형에 벌점(penalty)을 주는 벌점화 추정(penalized estimation)을 사용한다. 평활스플라인 모형(smoothing spline model)은 최소제곱법으로 추정하기 때문에 선형회귀모형과 많은 유사한 특성이 있다. 결국 평활함수의 추정량 $\hat{f}_j(x_j)$는 반응변수 y와 비모수 추정량 $f_j(x_j)$간의 제곱합, SS을 최소화 하는 것이다. 제곱합을 식으로 정리하면 아래의 식(2.4)와 같다.

$$SS(f_j) = \sum_{j=1}^{p} [y - f_j(x_j)]^2 \tag{2.4}$$

위의 식(2.4)을 최소화하는 $f_j(x_j)$를 추정할 때 많은 모수가 사용되기 때문에 벌점화 추정은 $f_j(x_j)$를 추정하는데 사용된 모수의 개수만큼 벌점(penalty)을 더하는 방법으로 해결한다. 스플라인 모형에서의 벌점을 아래의 식(2.5)과 같이하면 평활함수에 대한 추정은 아래와 같은 식(2.6)을 이용하여 할 수 있다.

$$\lambda \int_{x_1}^{x_p} [f_j''(x_j)]^2 dx \tag{2.5}$$

$$SS(f_j, \lambda) = \sum_{j=1}^{p} [y - f_j(x_j)]^2 + \lambda \int_{x_1}^{x_p} [f_j''(x_j)]^2 dx \tag{2.6}$$

여기서, 반응변수 Y, 설명변수 $X_1,X_2,\cdots,X_p$ 이며, 함수 $f_j(x_j)$는 1차, 2차 도함수를 가진다. 벌점화 최소제곱(penalized least square)법은 평활 추정량 $\hat{f}_j(x_j)$를 구하기 위하여 위의 식(2.6)을 최소화 하는 해를 구하는 것이다. 이때, 평활모수 λ에 의해서 모형의 평활정도가 결정되는데 평활모수에 대한 자세한 내용은 2.1.2.절에서 다시 알아보겠다.

(2) 백피팅 알고리즘(backfitting algorithm)

선형회귀모형의 일반적인 형태는 아래의 식(2.7)과 같다.

$$E[Y|X_1,\cdots,X_p] = \beta_0 + \sum_{j=0}^{p}\beta_j X_j, \tag{2.7}$$

또한 X_0는 항상 상수 1이다. 모든 X_j에 대한 정보를 알 수 없으나 적어도 하나의 성분인 X_k에 대해서는 가정할 수 있으므로 이에 대한 조건부 기댓값은 아래와 같이 구할 수 있다.

$$\begin{aligned} E[Y|X_k = x_k] &= E[E[Y|X_1, X_2, \cdots X_k, \cdots X_p]|X_k = x_k] \\ &= E\left[\sum_{j=0}^{p}\beta_j X_j | X_k = x_k\right] \\ &= \beta_k x_k + E\left[\sum_{j\neq k}\beta_j X_j | X_k = x_k\right], \end{aligned}$$

여기서,

$$\begin{aligned} \beta_k x_k &= E[Y|X_k = x_k] - E\left[\sum_{j\neq k}\beta_j X_j | X_k = x_k\right] \\ &= E\left[Y - \left(\sum_{j\neq k}\beta_j X_j\right) | X_k = x_k\right], \end{aligned} \tag{2.8}$$

위의 식(2.8)에서 기댓값은 k번째 편잔차(partial residual)이다. 전체 잔차는 Y

와 기댓값의 차이이며 편잔차는 Y와 X_k의 기여도를 무시한 기댓값의 차이이다. 이 Y와 X_k의 기여도를 제외한 차이를 $Y^{(k)}$라고 하면, 편잔차는 아래의 식(2.9)와 같이 다시 정리할 수 있다.

$$\beta_k x_k = E[Y^{(k)}|X_k = x_k]. \tag{2.9}$$

편잔차를 얻기 위해서는 다른 β_j들을 알 필요가 있다. 선형모형에서 이 추정법은 반복 근사 알고리즘인 가우스-자이델 알고리즘(Gauss-Seidel algorithm), 더 흔하게는 백피팅 알고리즘(backfitting algorithm)으로 알려져 있다. 그러나 선형모형의 적합에서는 백피팅 알고리즘을 많이 이용하지 않는데 그 이유는 $(X^TX)^{-1}X^Ty$를 계산하는 것이 더 빠르고 단순하기 때문이다. 따라서 백피팅 알고리즘은 선형모형이 아닌 가법모형에서 주로 이용된다.

가법모형을 회귀식으로 표현하면 아래의 식(2.10)과 같다.

$$E[Y|X_1,\cdots,X_p] = \alpha + \sum_{j=1}^{p} f_j(x_j), \tag{2.10}$$

이 식은 $f_j(x_j) = \beta_j x_j$인 선형모형의 특별한 형태를 가진다. 여기서 f_j는 임의의 비선형 함수이다. 백피팅 알고리즘의 아이디어는 반응변수에 각 설명변수의 기여도를 분리하여 단순하게 추가하는 것이며, 이 기여도는 반드시 변수에 비례할 필요는 없다. 가법모형의 좋은 특성 중 하나는 선형모형에서의 편잔차를 이용 할 수 있다는 것이다. 가법모형에서의 편잔차는 아래의 식(2.11)과 같이 정의할 수 있다.

$$Y^{(k)} = Y - \left(\alpha + \sum_{j \neq k} f_j(x_j)\right). \tag{2.11}$$

이를 위의 식(2.9)과 같이 정리하면 다시 아래의 식(2.12)와 같다.

$$E\left[Y^{(k)}|X_k = x_k\right] = f_k(x_k). \tag{2.12}$$

만약, 임의의 1차원 회귀를 추정하는 것이 가능하다면 가법모형에서 백피팅 알고리즘을 이용한 추정이 가능하다. 백피팅 알고리즘을 정리하면 아래 [표 2]와 같다.

[표 2] 가법모형에서의 백피팅 알고리즘

1. 초기화 : $\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} y_i,\ \hat{f}_i \equiv 0,\ all\ \text{for}\ i, j$

2. 반복 : $j = 1, 2, \cdots, p, \cdots, 1, 2, \cdots, p, \cdots,$

$$\hat{f}_j \leftarrow S_j\left[\left\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\right\}_1^N\right]$$

$$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N}\sum_{i=1}^{N} \hat{f}_j(x_{ij}),\ \text{여기서}\ S_j\text{는 삼차평활스플라인}$$

3. $\hat{f}_j$가 수렴할 때까지 위의 단계를 반복하여 실시한다.

위의 [표 2]와 같은 단순 반복 알고리즘을 통해서 해를 추정할 수 있다. $\hat{\alpha} = average(y_j)$ 즉, y의 평균으로 초기 값을 설정하고, 식(2.12)의 조건부 기댓값을 사용하여 x_{ij}의 함수로서 새로운 추정치 $\hat{f}_j$를 얻을 수 있다. 이 방법을 $\hat{f}_j$가 일정한 값으로 수렴하여 안정화 될 때까지 반복함으로서 모형에 대한 추정이 가능하다. 여기서 S_j와 같은 평활연산자(smoothing operators)는 삼차평활스플라인을 사용하는데 국소다항회귀(local polynomial regression)나 커널방법(kernel method)으로 대체가 가능하다. 그런데 이 백피팅 알고리즘은 가법모형

에서는 쉽게 적용할 수 있으나, 일반화가법모형에서는 적용하기가 복잡하다. 때문에 일반화가법모형에서는 가중 백피팅 알고리즘으로 추정한다. 이 가중 백피팅 알고리즘이 바로 아래에 설명할 로컬스코어링 알고리즘(local scoring algorithm; Hastie & Tibshirani, 1990)이라고 할 수 있다.

(3) 로컬스코어링 알고리즘(local scoring algorithm)

위의 백피팅 알고리즘은 가법모형에서는 적합이 용이하나 일반화가법모형에서 백피팅 알고리즘은 좀 더 복잡하다. 일반화가법모형은 일반화선형모형을 확장한 형태이므로 가법모형은 선형모형을 확장한 것이라고 할 수 있다. 그러므로 선형모형에서의 $\alpha+\sum_j x_j\beta_j$를 가법모형에서의 $\alpha+\sum_j f_j(x_j)$로 대체가능하다.

일반화선형모형에서 회귀계수 β는 다음의 식(2.13)의 스코어 방정식(score equations)으로 정의할 수 있다.

$$\begin{aligned}&\sum_{i=1}^{n} x_{ij}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)V_i^{-1}(y_i-\mu_i),\\ &\eta_i=\alpha+\sum_{j=1}^{p}\beta_j x_{ij},\\ &\mu_i=g(\eta_i),\quad j=0,1,\cdots,p,\ i=1,\cdots,n\ ,\end{aligned}\tag{2.13}$$

여기서 V_i는 Y_i에 대한 분산행렬(variance matrix)이다. 또한 x_{ij}는 j번째 설명변수의 i번째 관측치이며, η_i는 i번째 관찰치에 대한 선형예측자, μ_i는 i번째 관찰치에 대한 적합값이며, 여기서 g는 연결함수(link funtion)이다. 피셔의 점수화 방법(fisher scoring method)은 위의 스코어 방정식을 푸는 일반적인 방

법으로 현재의 회귀계수벡터(coefficient vector)를 β^0라 하고, 그에 대한 선형예측자를 η^0, 적합값(fitted value)을 μ^0를 이용하여 아래와 같은 식(2.14)로 보정반응변수를 구한다.

$$z_i = \eta_i^0 + (y_i - \mu_i^0)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_0 . \tag{2.14}$$

그리고 가중치 w_i를 아래의 식(2.15)와 같이 구한다.

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)_0^2 (V_i^0)^{-1}. \tag{2.15}$$

이 알고리즘은 회귀모형을 통해서 보정된 종속변수 z_i에 대해서 가중치 w_i를 추가하여 개정된 회귀계수 β를 추정하는 방법이다. 반복을 통해서 새로운 μ^0와 η^0가 계산되면 새로운 z_i가 추정된다. 이때, 편차(deviance)의 변화량이 일정하게 작아질 때 까지 위의 과정을 반복한다.

일반화가법모형이 일반화선형모형과 다른 점은 선형예측자(linear predictor)를 가법예측자(additive predictor)로 대체하였다는 것이다. 로컬스코어링 알고리즘이란 일반화가법모형에서 가법성분(additive term)에 대한 추정을 위해서 가중 백피팅 알고리즘(weighted backfitting algorithm)을 이용하는 것이다. 로컬스코어링 알고리즘은 아래의 [표 3]와 같다.

[표 3] 로컬 스코어링 알고리즘

1. 초기화 : $\hat{\alpha}=\frac{1}{N}\sum_{i=1}^{N}y_i,\ \hat{f}_i\equiv 0,\ all\ \text{for}\ i,\quad t=0$

2. 반복 : (1) 보정된 반응변수

$$z_i=\eta_i^t+(y_i-\mu_i^t)\left(\frac{\partial\eta_i}{\partial\mu_i}\right)_{t,}\quad i=1,\cdots,n$$

이때, $\eta_i^t=\alpha+\sum_{j=1}^{p}f_j{}^t(x_{ij}),\quad \mu_i^t=g^{-1}(\eta_i^t),$

(2) 가중값

$$w_i=\left(\frac{\partial\mu_i}{\partial\eta_i}\right)_0^2(V_i^{0})^{-1},\quad i=1,\cdots,n$$

(3) 보정된 반응변수 z_i에 가중치 w_i를 이용하여 가중가법모형을 적합 시킨다. 이때, $\hat{f}_j$는 앞서 소개한 백피팅 알고리즘을 통해서 추정한다.

3. 추정된 함수의 변화가 적어질 때까지 앞의 2단계를 반복하여 수행한다.

2.1.2. 평활함수

우리가 많이 사용하는 선형모형은 모수의 추정이 비교적 용이하고 결과 해석이 쉽다는 장점이 있다. 그러나 실생활에서 얻어지는 대부분의 자료는 선형성을 만족한다고 보기 어렵기 때문에 비선형적인 자료의 성격을 잘 설명할 수 있는 평활함수는 통계적 기법에서 매우 유용하다고 할 수 있다. 비선형적 관계를 추정할 수 있는 평활 알고리즘은 커널 함수(kernel function), 국소다항회귀(local polynomial regression), 조각별 회귀(piecewise regression), 평활스

플라인(smoothing spline) 등 여러 가지로 종류가 다양한데, 그 중 평활스플라인에 대해서 앞의 2.1.1.절에서 살펴보았다. 이번 절에서는 평활함수의 추정과 평활모수(smoothing parameter)에 대해서 알아보도록 하겠다.

(1) 평활(smoothing)

일반화가법모형은 비모수적 회귀와 평활을 기반으로 하여 만들어진다. 따라서 일반화가법모형을 적합하기 위해서는 평활과 비모수적 회귀방법의 원리를 알아야할 필요가 있다. 일반적으로 두 변수 x와 y의 관계를 요약하는데 사용되는 통계적 기법은 회귀나 상관계수 등의 모수적(parametric)인 방법이다. 만약, 모수적인 방법으로 설명변수 x와 반응변수 y 두 변수 간의 상관관계를 구하려면 단순하게 선형모형에 적합할 수 있다. 이때, 추정해야 할 모수는 선형회귀모형의 회귀계수 β만 존재한다. 그러나 평활함수나 비모수적 회귀에 대해서는 단순히 하나의 모수만이 아닌 비선형함수에 대한 여러 가지 사전 가정이 필요하다. 이러한 평활함수를 이용한 비선형모형의 단점은 선형모형에 비해서 시간과 비용이 많이 든다는 것과 결과의 해석이 어렵다는 것이다. 그러나 요즈음의 컴퓨터 성능은 예전에 비해 발전되어 계산에 걸리는 시간과 비용이 많이 줄었으며, 결과의 해석의 어려움을 수용할 만큼의 자료에 대한 설명력을 얻을 수 있다. 이러한 이유로 평활 함수를 사용한 비선형 모형이 많이 이용되고 있다.

(2) 평활 추정법과 평활 모수(smoothing parameter)

평활스플라인 모형(smoothing spline model)에서 평활 모수의 추정은 앞선 2.1.1.절에서 벌점화 최소제곱법을 이용하는 것을 확인하였다. 평활함수 $f_j(x_j)$의 추정량을 구하기 위해서는 위에서 설명한대로 식(2.6)을 최소화하는 해를

구하면 된다.

식(2.6)에서 λ는 평활모수로 두 번째 항에 영향을 주게 되며, 두 번째 항은 모형에 벌점을 추가하여 모형의 적합도를 결정하게 된다. 만약 평활모수 λ가 무한대로 커지면 많은 벌점이 모형에 추가되어 평활 추정치 $\hat{f}_j(x_j)$는 일반 선형최소제곱의 형태가 된다. 그러나 평활모수 λ가 0이 되면 모형에 벌점이 추가되지 않아 평활 추정치 $\hat{f}_j(x_j)$는 자료에 과적합하는 형태가 될 것이다.

이와 같이 평활모수 λ는 평활함수에서 매우 중요한 역할을 하며, 이 평활모수의 결정에 대해서 많은 연구가 진행되었다. 대표적으로 많이 사용되는 방법으로 교차타당성(cross-validation) 평가가 있다. 교차타당성 평가는 관찰치의 반복추출을 기반으로 하여 모형의 적합을 평가하는 일반적인 방법으로 많은 통계모형에서 사용된다. 교차타당성 평가는 잔차제곱합(RSS, residual sum of squares)을 근거로 하여 모형의 적합도를 측정하며, 모형에 적합된 반응변수 y가 그 예측에 이용되지 않도록 자료를 나누는 것이다. 교차타당성 평가의 가장 단순한 경우를 설명하면 자료를 단순히 둘로 나누어 첫 번째 자료는 훈련데이터(train data)로 하여 모형을 적합하고 두 번째 자료를 검증데이터(test data)로 하여 만들어진 모형에 대하여 예측 값을 구하는 것이다(Efron & Tibshirani, 1993).

이 교차타당성 평가를 이용해서 평활모수 λ를 선택하는 방법은 아래와 같다. 만약 n개의 관찰치가 있다면, i번째 관찰치를 제외하고 평활모수를 λ로 하여 추정한 평활함수는 $\hat{f}_{\lambda}(x_{-i})$이다. 추정한 평활함수 $\hat{f}_{\lambda}(x_{-i})$와 반응변수 y_i 간의 차이의 제곱합 즉 잔차제곱합을 구하여 횟수 n번으로 나눈 값이 교차타당성 평가을 이용한 평활모수의 값이라고 할 수 있다. 이 값을 CV점수(cross validation score)라고 하며 수식으로 나타내면 아래의 식(2.16)와 같다.

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}_{\lambda}(x_{-i})]^2 \tag{2.16}$$

이 과정을 실시한 뒤 구해진 CV점수 중 가장 작은 값을 평활 모수로 고려한다. 그런데 이 교차타당성 평가는 반복적인 계산방법이기 때문에 변수의 개수와 관찰치의 수가 많아지면 계산의 횟수가 매우 커지는 단점이 있다. 또한 스플라인 모형에서는 평활함수가 자료를 변환시키는 역할을 하므로 교차타당성 평가를 실시할 때 마다 구해지는 최적의 λ가 매번 바뀌게 된다(Wood, 2006). 평활스플라인 모형에서 이러한 문제를 해결하기 위하여 제안된 방법이 일반화 교차타당성(generalized cross-validation) 평가이다.

일반화 교차타당성 평가를 이용하여 구한 값을 GCV점수라고 하며 아래의 식(2.17)과 같다.

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}[y_i + \hat{f}_{\lambda}(x_i)^2]}{[1 - n^{-1}tr(\boldsymbol{S})]^2}, \tag{2.17}$$

$$\boldsymbol{S} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda^{2p}\boldsymbol{D})^{-1}\boldsymbol{X},$$

여기서, $\boldsymbol{S}$는 평활자행렬(smoother matrix)이며, $\boldsymbol{X} = X_1, \cdots, X_p$는 설명변수, p는 다항기저함수의 차수로 λ^{2p}는 벌점이다. $\boldsymbol{D}$는 벌점행렬(penalty matrix)의 행렬연산자로 매듭(knot)의 개수를 k로 하여 다음과 같은 행렬을 가진다.

$$\boldsymbol{D} = \begin{bmatrix} 0_{2\times 2} & 0_{2\times k} \\ 0_{k\times 2} & \boldsymbol{I}_{k\times k} \end{bmatrix}$$

식(2.17)에서 $\hat{f}_{\lambda}(x_i)$는 위에서와 마찬가지로 주어진 λ에 대한 모든 자료에 대한 평활스플라인의 추정량이며, 이를 통해 계산된 GCV점수는 CV점수보다 우

월하다. 또한 이 GCV점수를 이용하여 적합한 평활스플라인에서는 벌점이 모형의 과적합을 막아 국소다항회귀에서 GCV점수를 이용하였을 때보다 더 좋은 것으로 알려져 있다(Luke, 2008).

이번 절에서는 일반화가법모형의 추정과 적합, 비선형성을 추정하는데 이용되는 평활함수에 대해서 정리하였다. 다음 절에서는 결합 예측 방법론에 대해서 알아보도록 하겠다.

2.2. 결합 예측 방법론

본 절에서는 결합 예측 방법론에 관하여 설명하고자 한다. 결합 예측 방법론이란 단일 예측, 분류자를 결합하여 예측 또는 분류하는 것이며, 단일 예측, 분류자에 비해 이 경우의 예측, 분류력이 더 좋은 것으로 많은 선행 연구를 통해서 알려져 있다. 지금까지 다양한 방법이 제시되어 왔는데 그 중 대표적인 방법이 배깅(Breiman, 1996), 랜덤포레스트(Breiman, 2001) 등이 있다.

2.2.1. 소개

붓스트랩 방법(bootstrap method; Efron, 1979)이 소개된 이후 컴퓨터 성능의 발달로 붓스트랩은 모수의 추정이나 예측에 정확성을 높이는 방법으로 많이 이용되어 왔다. 그 중 결합 예측 방법론은 붓스트랩 표본에서 여러 개의 단일 예측 분류자를 결합하여 반응변수에 대한 예측이나 분류에서 더 좋은 수행력을 가지는 방법이다. 이는 많은 선행 연구를 통해 수행력의 우수함이 증명되어 왔다(Bauer & Kohavi, 1999; Bühlmann, 2002; Skurichina & Duin, 2002). 여러 결합 예측 방법 중에 배깅과 랜덤포레스트에 대해서 다음 절에서 소개하도록 하겠다.

2.2.2. 배깅, 랜덤포레스트

(1) 배깅(bagging)

배깅은 결합 예측 방법론의 하나로 Breiman에 의해서 1996년에 발표되었다. 일반적으로 훈련데이터(train data)의 크기가 작을 때 단일 예측자(single predictor)는 좋은 수행력을 보이기 어렵다. 적은 훈련데이터에 대해서 예측자를 개발하면 편향(biased)되고 큰 분산을 가지는 좋지 않은 모수를 추정하는 경우가 많다. 또한, 예측자가 의사결정나무(decision tree)와 같이 불안정한 방

법에 의해서 개발된 경우 편향되고 큰 분산을 가지는 모수를 추정하게 된다. 이때의 단점을 보완할 수 있는 좋은 방법이 붓스트랩(bootstrap) 방법이다. 붓스트랩 방법을 이용해 반복적으로 무작위 복원 붓스트랩 추출법을 추출하여 다중의 예측자를 만들어 낼 수 있다.

배깅 예측자(bagging predictor)는 다중의 예측자를 결합하여 예측자로 하는 방법이다. 이 결합은 연속형 반응변수에서는 예측값의 평균을 이용하며 범주형 반응변수에서는 다중투표(plurality vote) 방식을 이용한다. 앞서 말한 의사결정나무와 같이 불안정한 예측자의 경우 배깅 예측자는 단일 예측자에 비해 더 좋은 정확도를 가진다. 배깅의 알고리즘은 아래와 같다.

[표 4] 배깅 알고리즘

1. 훈련데이터(train data), $L=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$,
   여기서 y 는 연속형 반응변수, 범주형 반응변수 모두 가능
2. $\phi(x,L)$: 분석용 데이터 L에서 x를 입력하였을 때 y를 예측하는 단일 예측자
3. 붓스트랩 표본(bootstrap data),
   $L^k(k=1,2,\cdots,B)$: 분석용 데이터 L과 동일한 분포 하에서 독립인 n개의 개체를 가지고 있는 k번째 붓스트랩 표본, 무작위 복원 추출로 B번의 붓스트랩을 실시
4. $\phi^{(k)}(x,L^{(k)})$: 붓스트랩 표본 $L^{(k)}$에서의 예측자
5. 배깅 결합 예측자(bagging predictor)
   반응변수가 연속형인 경우의 배깅 예측자 ;
   $\phi_B(X)=average\{\phi^{(k)},(x,L^{(k)})\}$

반응변수가 범주형인 경우의 배깅 예측자 ;

$$\phi_B = \text{arg}max_j \sum_{k=1}^{B} \{\phi^{(k)}, (x, L^k) = j\}$$

위의 알고리즘에서 확인할 수 있듯이 연속형 반응변수에 대한 배깅 예측자를 구하는 방법을 다시 정리하면, B개의 붓스트랩 표본에서 B개의 단일 예측자를 구하고, 그 평균을 이용하여 다중 예측자를 결합한다. 또한, 범주형 반응변수에서의 배깅 예측자는 역시 B개의 붓스트랩 표본에서의 B개의 예측자에서 다중투표(plurality vote) 방식을 이용하여 특정 범주(j)에 대한 분류가 최대로 많은 범주로 분류하는 것이다.

(2) 랜덤포레스트(random forest)

랜덤포레스트는 2001년 Breiman에 의해 발표되었으며, 단일 예측자로 이용되는 의사결정나무 간의 상관관계를 줄이기 위한 배깅의 수정된 방법이라고 할 수 있다. 배깅은 추정된 예측함수의 분산을 줄이는 방법이며, 특히 의사결정나무와 같이 큰 분산(variance)과 낮은 편향(bias)을 가지는 예측자에서 좋은 수행력을 보인다. 배깅에서 이용되는 B개의 나무는 모두 동일한 분포에서 추출된 표본를 통해서 만들어 졌다. 다시 말하면 B개 나무의 평균의 기댓값은 그 나무 중 하나의 기댓값과 동일하다고 볼 수 있다는 것이다. 즉 배깅 예측자의 편향은 각각의 나무의 편향과 동일하며 오직 분산의 감소에 대해서만 기대할 수 있다. 이러한 특징이 부스팅과 같은 앙상블(ensemble) 기법과 배깅의 다른 점이다(Hastie, Tibshirani, Friedman, 2008).

독립적이고 같은 분포를 가진 임의의 확률변수들인 붓스트랩 표본 B의 평균에 대한 분산이 $\frac{1}{B}\sigma^2$이며, 각 B에 대한 분산은 σ^2라면 결합 예측자에 대

한 분산은 아래 식(2.18)와 같다.

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{2.18}$$

만약 여기서 단순하게 독립이 아닌 동일한 분포 하에서 ρ는 양의 상관관계를 나타낼 것이다. B가 증가하면 위 수식(2.18)에서의 두 번째 항이 0으로 수렴하게 된다. 그러나 첫 번째 항은 남게 되며, 배깅에서 이용되는 단일 예측자간의 상관관계의 크기가 결합 예측방법을 통해 분산을 감소시키는 배깅의 장점을 제약하게 된다. 랜덤포레스트는 단일 예측자인 의사결정나무간의 상관관계를 감소시켜 배깅에서 분산을 감소시키는 아이디어에서 제안되었다. 이러한 목표는 나무의 가지를 늘리는 과정에서 설명변수를 무작위로 선택하는 방법을 통해서 이룰 수 있다. 랜덤포레스트의 알고리즘을 살펴보면 아래와 같다.

[표 5] 랜덤포레스트 알고리즘

1. 붓스트랩 반복 횟수 $b=1,\cdots,B$

(a) 훈련데이터(크기N)으로부터 Z^*인 붓스트랩 표본을 뽑는다.

(b) T_b : 붓스트랩 표본로부터 만들어진 랜덤포레스트 나무, 아래와 같은 반복적인 과정을 통해서 각 나무의 끝마디를 만든다. 여기서 마디의 크기($n_{\min}$)를 최소로 한다.

(가) p개의 설명변수로부터 m개의 변수를 무작위로 추출한다.

(나) m개의 변수 중 가장 좋은 분리점을 가지는 변수를 선택한다.

(다) 두 개의 자식마디로 나눈다.

2. 앙상블 나무 $\{T_b\}_1^B$를 얻는다.

새로운 값 x에 대한 예측은 아래와 같다.

연속형 반응변수의 경우

$\widehat{f_{rf}^{B}}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

범주형 반응변수의 경우

$\widehat{C_{rf}^{B}}(x) = majority\ vote\left\{\widehat{C}_b(x)\right\}_1^B$, 여기서 $\widehat{C}_b(x)$는 b번째 랜덤포레스트 나무의 범주에 대한 분류예측자

지금까지 결합 예측 방법론 중에서 배깅과 랜덤포레스트에 대해서 알아보았다. 배깅과 랜덤포레스트 모두 의사결정나무를 단일 예측, 분류자로 하는 결합 방법을 이용하였다. 다음 절에서는 일반화가법모형을 이용한 결합 예측 방법론을 알아보도록 하겠다.

2.2.3. 앙상블 기법을 이용한 일반화가법모형

일반화가법모형은 일반화선형모형에 설명변수와 반응변수간의 비선형 관계를 확인할 수 있는 좋은 통계적 기법으로 확인되었다. 선행 연구에서 이항 반응변수에서 일반화가법모형을 앙상블 기법의 기본 분류자로 이용하는 방법이 제안되었다(De Bock, 2010). 본 논문에서는 연속형 반응변수에 대하여 붓스트랩 표본을 이용한 앙상블 기법의 기본 예측자로 일반화가법모형을 이용하는 방법을 제안하고자 한다. 다음 절에서 기존에 연구된 이항 반응변수에서의 붓스트랩을 이용한 일반화가법모형을 설명하고, 연속형 반응변수에서의 붓스트랩을 이용한 일반화가법모형에 대해서 정리하고자 한다.

(1) 이항 반응변수

기존의 배깅이나 랜덤포레스트 등의 앙상블 기법에서 주로 이용되던 의사결정나무와 같은 불안정한 분류자(weak classifier)를 붓스트랩 표본의 기본 분류자로 이용하는 대신에 일반화가법모형을 이용하고자 하였다. 이를 통해, 설명변수와 반응변수 간의 비선형 관계를 파악하여 모형을 개발하고 예측변수에 대한 분류의 정확도를 향상시키는 것을 제안하였다. 이항 반응변수에서 앙상블 기법을 이용한 일반화가법모형은 아래와 같은 과정을 가진다.

이항반응변수에 대한 일반화가법모형은 아래의 식(2.19)과 같다.

$$\text{logit}(P(Y=1|X)) \equiv \log\left\{\frac{P(Y=1|X)}{1-P(Y=1|X)}\right\} = \sum_{j=1}^{p_c} f_j(X_j) + \sum_{k=1}^{p_b} \beta_k X_k \quad (2.19)$$

여기서, X_j, $j=1,\cdots,p_c$는 연속형 설명변수이며, X_k, $k=1,\cdots,p_b$ 가변수로 코딩된 범주형 설명변수이다. 여기서 연속형 변수인 X_j에 대한 비선형 관계는 평활함수인 $f_1(x_1), f_2(x_2), \cdots, f_{p_c}(X_{p_c})$ 를 통해 적합이 가능하다.

위와 같은 일반화가법모형을 기본 분류자로 이용하는 앙상블 기법의 과정은 아래 [표 6]과 같다.

[표 6] 앙상블 기법을 이용한 일반화가법모형 (이항반응변수에 대하여)

1. 입력값

$D=\{(x_i,y_i)\}_{i=1}^{n}$; $x_i \in X \subset R^p$; $y_i \in Y=\{0,1\}$, D는 훈련데이터,

설명변수 x_i는 p개, 이항 반응변수 y,

m ; 붓스트랩 표본의 개수

r ; 무작위로 선택되는 변수의 수 $r \leq p$

2. 모형적합 과정

$l = 1,2,\cdots,m,$

$R_{l,c}$는 연속형 설명변수, 설명변수의 개수는 $p_{l,c}$

$R_{l,b}$는 이분형 설명변수, 설명변수의 개수는 $p_{l,b}$

l번째 기본분류자를 C_l로 하는 준모수적 일반화가법모형

$$C_l : P_l(Y=1|X) = 1/\left\{1+\exp\left(-\left(\sum_{j=1}^{p_{l,c}} f_{l,j}(X_j) + \sum_{k=1}^{p_{l,b}} \beta_{l,k} X_k\right)\right)\right\},$$

여기서 $X_j \in R_{l,c}$, $X_k \in R_{l,b}$

3. 예측 과정

$C(x) = \frac{1}{m}\sum_{l=1}^{m} C_l(x)$: 관찰치 x가 범주 1에 포함된다고 분류할 앙상블 예측자에 대한 확률

위의 과정을 통해서 이항 반응변수에 대한 일반화가법모형을 기본 분류자로 하는 앙상블 결합 분류자를 얻을 수 있다.

(2) 연속형 반응변수

본 논문에서는 이항 반응변수에 적용하였던 앙상블 기법을 이용한 일반화가법모형을 연속형 반응변수에 관해서 적용해 보고자 한다. 위에서의 설명과 마찬가지로 연속형 반응변수를 가지는 자료에 대해서 주된 관심은 모형의 적합과 그를 이용한 예측확률의 추정일 것이다. 결합 예측 방법론에서 많은 선행연구를 통해서 증명된 것처럼 반응변수의 예측에 있어서도 단일 예측자에 비해 결합 예측자가 우수한 성능을 보이는 것을 알 수 있다. 따라서 이항 반응변수에 그쳤던 이전 선행연구를 확장해 보고자 한다.

연속형 반응변수에 대한 일반화가법모형은 아래와 같은 식(2.20)으로 나타낼 수 있다.

$$P(Y|X) = \sum_{j=1}^{p} f_j(X_j). \qquad (2.20)$$

[표 7] 앙상블 기법을 이용한 일반화가법모형(연속형 반응변수에 대하여)

1. 입력값

$D = \{(x_i, y_i)\}_{i=1}^{n}$; $x_i \in X \subset R^p$; $y_i \in Y$, D는 훈련데이터,

설명변수 x_i는 p개, 연속형 반응변수 y,

m : 붓스트랩 표본의 개수

r : 무작위로 선택되는 변수의 수 $r \leq p$

2. 모형적합 과정

$l = 1, 2, \cdots, m$,

R_l는 연속형 설명변수, 설명변수의 개수는 무작위로 선택된 r_l

l번째 기본예측자를 C_l로 하는 일반화가법모형

$C_l : P_l(Y|X) = \sum_{j=1}^{r_l} f_{l,j}(X_j)$, 여기서 $X_j \in R_l$

3. 예측 과정

$C(x) = aggriation\{C_l(x)\}$: 관심있는 관찰치 x에 대한 예측확률 C_l을 결합하여 앙상블 예측자 $C(x)$를 얻는다.

여기서 결합 방법은 세 가지로 제안한다.

- 평균 : $C(x) = \frac{1}{m} \sum_{l=1}^{m} C_l(x)$

- 절사평균 : $C(x)=\frac{1}{n-2j}\sum_{i=j+1}^{n-j}C_l(x)$, 여기서 절사비율은 j/m
- 중위수 : $C(x)=\begin{cases}C_l(x), & m=2l-1\\ \frac{C_l(x)+C_{l+1}(x)}{2}, & m=2l\end{cases}$

결합의 방법은 평균, 절사평균, 중위수의 세 가지로 한다. 과적합의 문제를 피하고자 붓스트랩 표본의 크기는 단일 분류자의 훈련데이터의 크기보다 절반으로 줄여서 적합하였다.

제 3 장 모의실험

본 절에서는 연속형 반응변수에 대하여 앞의 2.2.3.절에서 제안한 앙상블 기법을 이용한 일반화가법모형의 효율성을 모의실험을 통해서 확인하려고 한다.

3.1. 모의실험설계

현실에 존재하는 다양한 형태의 자료에서 제안한 방법론의 효율성을 확인하기 위하여 여러 경우의 모의실험 자료를 생성하였다. 또한 Friedman(1991)이 논문에서 사용한 3개의 모의실험 자료 역시 동일한 확률분포와 수식을 통해서 생성하여 제안한 방법론의 효율성을 비교하였다.

모의실험에서 앙상블(ensemble) 기법을 이용한 일반화가법모형과 기존에 제안된 일반화가법모형, 배깅, 랜덤포레스트를 함께 적용하여 비교하도록 한다. 이 비교에 이용된 통계량은 제곱근평균제곱오차(RMSE: Root Mean Square Error)이다.

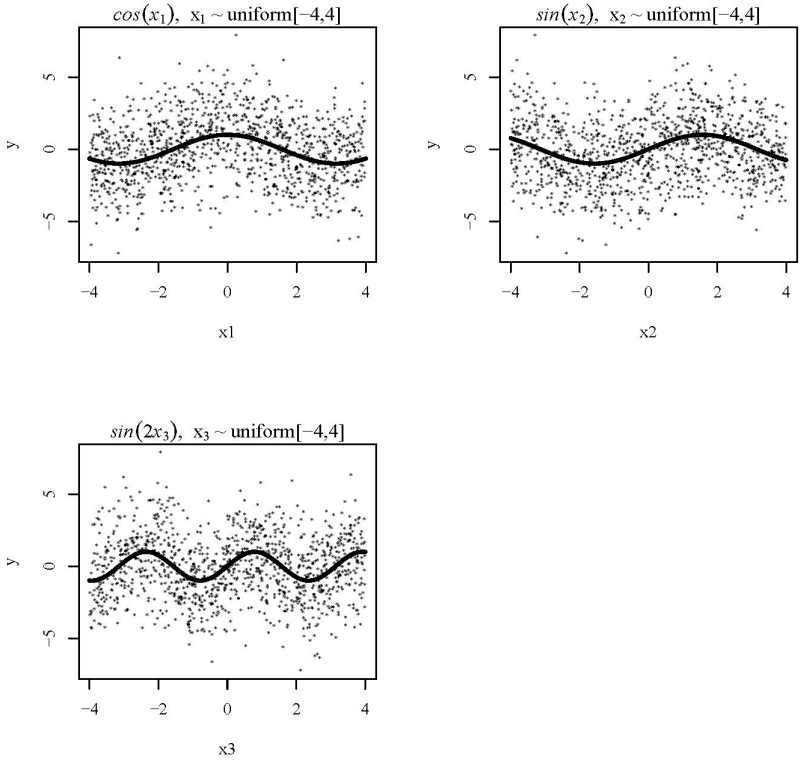
모의실험자료는 1,200개의 관찰치를 가지며 각 훈련데이터(train data)의 크기는 200개, 검증데이터(test data)의 크기는 1,000개로 하였다. 그리고 과적합(overfitting)의 문제를 방지하기 위해서 앙상블 기법을 이용한 일반화가법모형에서의 훈련데이터의 크기는 100개로 하였다.

모의실험의 설계는 아래와 같으며 자료의 생성은 통계프로그램 R을 이용하였다.

1. 삼각함수

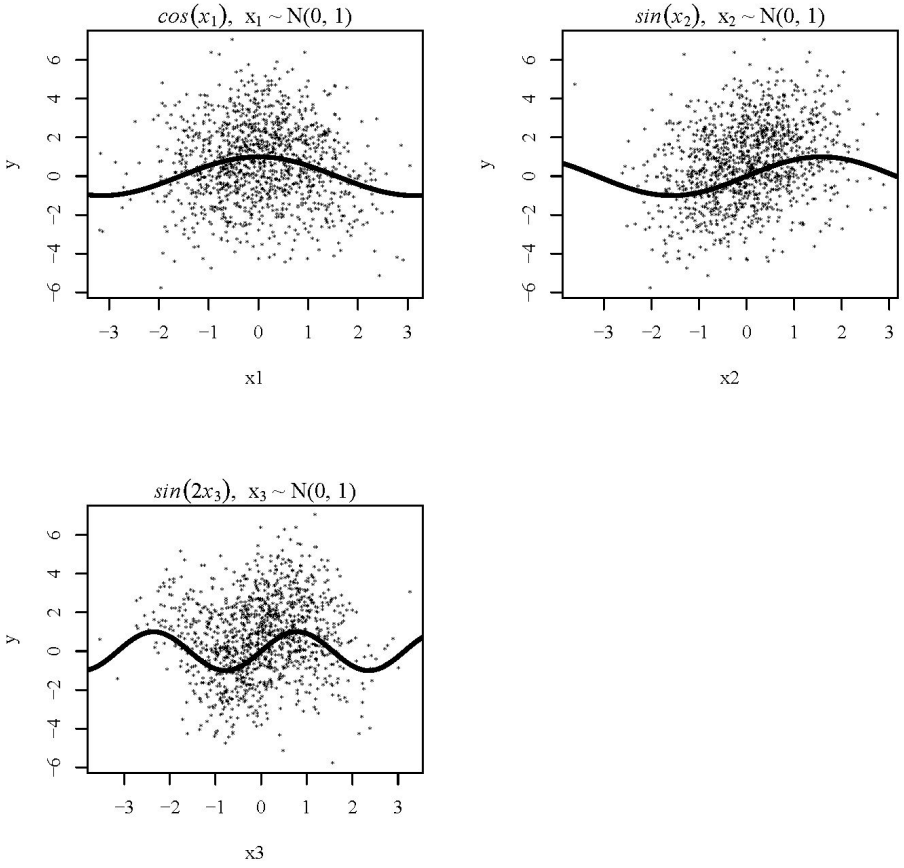
삼각함수 자료1은 x_1, x_2, x_3의 세 개의 설명변수와 연속형 반응변수를 가지는 자료로 [그림 1]과 같은 형태의 실험 자료이다. 각 설명변수 x_1, x_2, x_3는 -4에서 4의 범위를 가지는 균일분포로부터 생성되었다. 또한 오차항 $\epsilon_1, \epsilon_2, \epsilon_3$은 평균이 0, 표준편차 1을 따르는 정규분포로부터 생성되었다. 위의 설명변수와 오차항을 이용하여 아래의 식(2.21)과 같이 연속형 반응변수를 생성하였다.

$$\begin{aligned} y_1 &= \cos(x_1) + \epsilon_1, \\ y_2 &= \sin(x_2) + \epsilon_2, \\ y_3 &= \sin(2x_3) + \epsilon_3, \\ y &= y_1 + y_2 + y_3 \end{aligned} \tag{2.21}$$



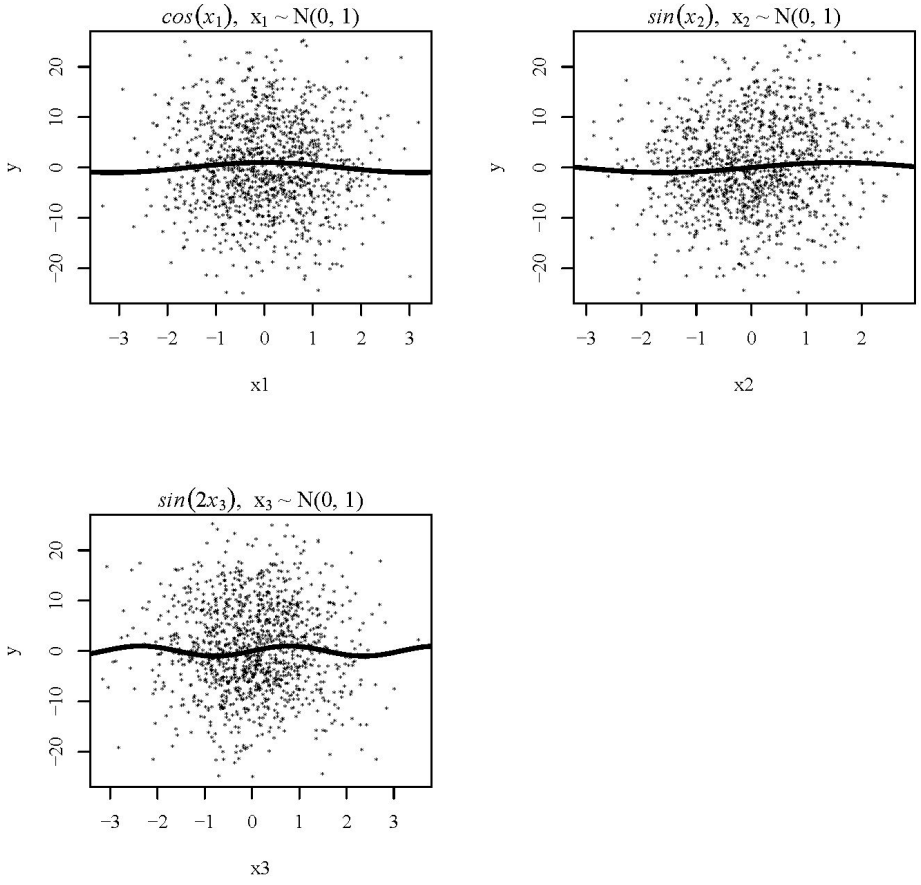
[그림 1] 삼각함수를 이용한 자료1에 대한 산점도

삼각함수 자료2는 x_1, x_2, x_3의 세 개의 설명변수와 연속형 반응변수를 가지는 자료로 [그림 2]과 같은 형태의 실험 자료이다. 각 설명변수 x_1, x_2, x_3와 오차항 $\epsilon_1, \epsilon_2, \epsilon_3$은 평균 0, 표준편차 1을 따르는 정규분포로부터 생성되었다. 위의 설명변수와 오차항을 이용하여 위의 식(2.21)과 같이 연속형 반응변수를 생성하였다.



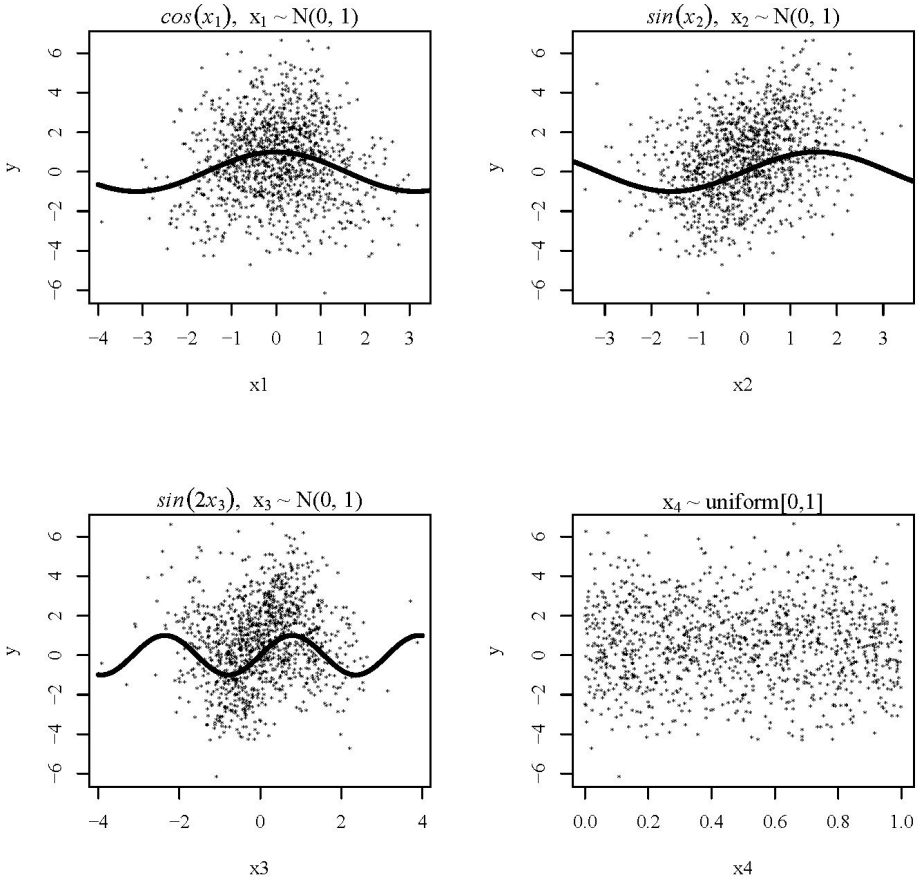
[그림 2] 삼각함수를 이용한 자료2에 대한 산점도

삼각함수 자료3은 x_1, x_2, x_3의 세 개의 설명변수와 연속형 반응변수를 가지는 자료로 [그림 3]과 같은 형태의 실험 자료이다. 각 설명변수 x_1, x_2, x_3는 평균 0, 표준편차 1을 따르는 정규분포로부터 생성되었으며, 오차항 $\epsilon_1, \epsilon_2, \epsilon_3$은 평균 0, 표준편차 5를 따르는 정규분포로부터 생성되었다. 위의 설명변수와 오차항을 이용하여 위의 식(2.21)과 같이 연속형 반응변수를 생성하였다.

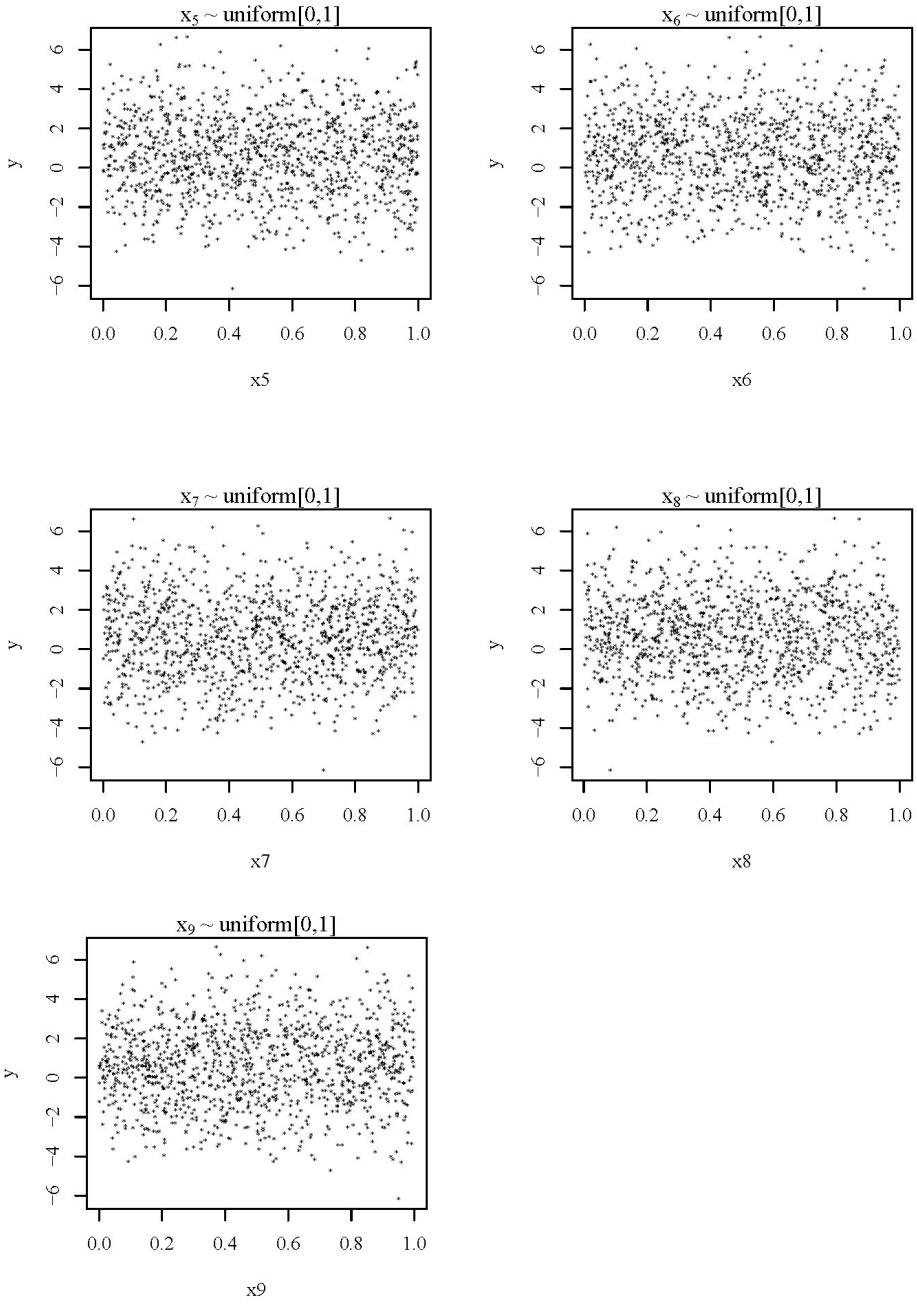


[그림 3] 삼각함수를 이용한 자료3에 대한 산점도

삼각함수 자료4은 $x_1, \cdots, x_9$의 아홉 개의 설명변수와 연속형 반응변수를 가지는 자료로 [그림 4]과 같은 형태의 실험 자료이다. 각 설명변수 x_1, x_2, x_3는 평균 0, 표준편차 1을 따르는 정규분포로부터 생성되었으며, $x_4, \cdots, x_9$는 0에서 1의 범위를 가지는 균일분포로부터 생성되었다. 또한 오차항 $\epsilon_1, \epsilon_2, \epsilon_3$은 평균 0, 표준편차 1를 따르는 정규분포로부터 생성되었다. 위의 설명변수와 오차항을 이용하여 위의 식(2.21)과 같이 연속형 반응변수를 생성하였다.



[그림 4] 삼각함수를 이용한 자료4에 대한 산점도 계속

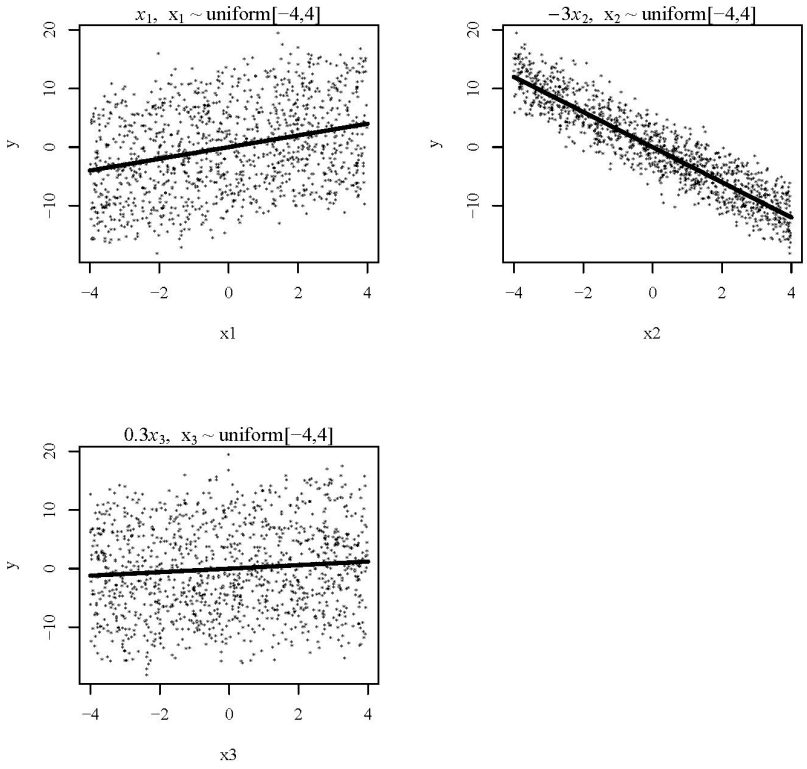


[그림 4] 삼각함수를 이용한 자료4에 대한 산점도

2. 선형함수

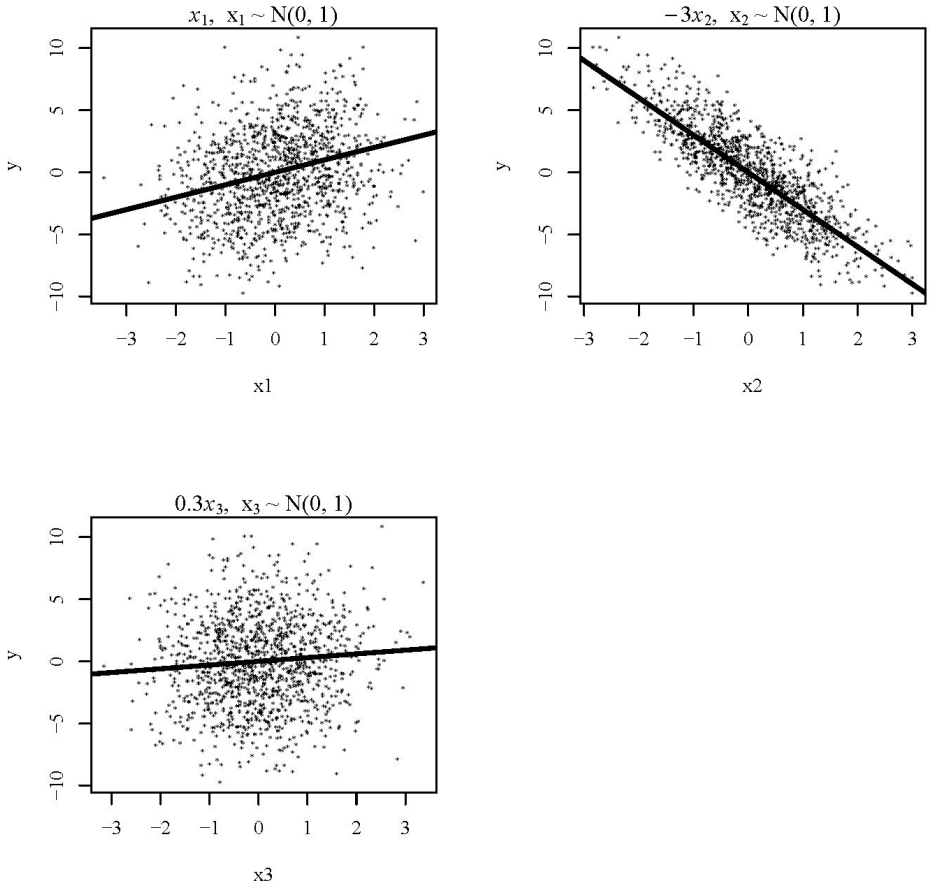
선형함수 자료1는 x_1, x_2, x_3의 세 개의 설명변수와 연속형 반응변수를 가지는 자료로 [그림 5]과 같은 형태의 실험 자료이다. 각 설명변수 x_1, x_2, x_3는 -4에서 4의 범위를 가지는 균일분포로부터 생성되었다. 또한 오차항 $\epsilon_1, \epsilon_2, \epsilon_3$은 평균이 0, 표준편차 1을 따르는 정규분포로부터 생성되었다. 위의 설명변수와 오차항을 이용하여 아래의 식(2.22)과 같이 연속형 반응변수를 생성하였다.

$$\begin{aligned} y_1 &= x_1 + \epsilon_1, \\ y_2 &= -3x_2 + \epsilon_2, \\ y_3 &= 0.3x_3 + \epsilon_3 \\ y &= y_1 + y_2 + y_3 \end{aligned} \tag{2.22}$$



[그림 5] 선형함수를 이용한 자료1에 대한 산점도

선형함수 자료2는 x_1, x_2, x_3의 세 개의 설명변수와 연속형 반응변수를 가지는 자료로 [그림 6]과 같은 형태의 실험 자료이다. 각 설명변수 x_1, x_2, x_3와 오차항 $\epsilon_1, \epsilon_2, \epsilon_3$은 평균 0, 표준편차 1을 따르는 정규분포로부터 생성되었다. 위의 설명변수와 오차항을 이용하여 위의 식(2.22)과 같이 연속형 반응변수를 생성하였다.

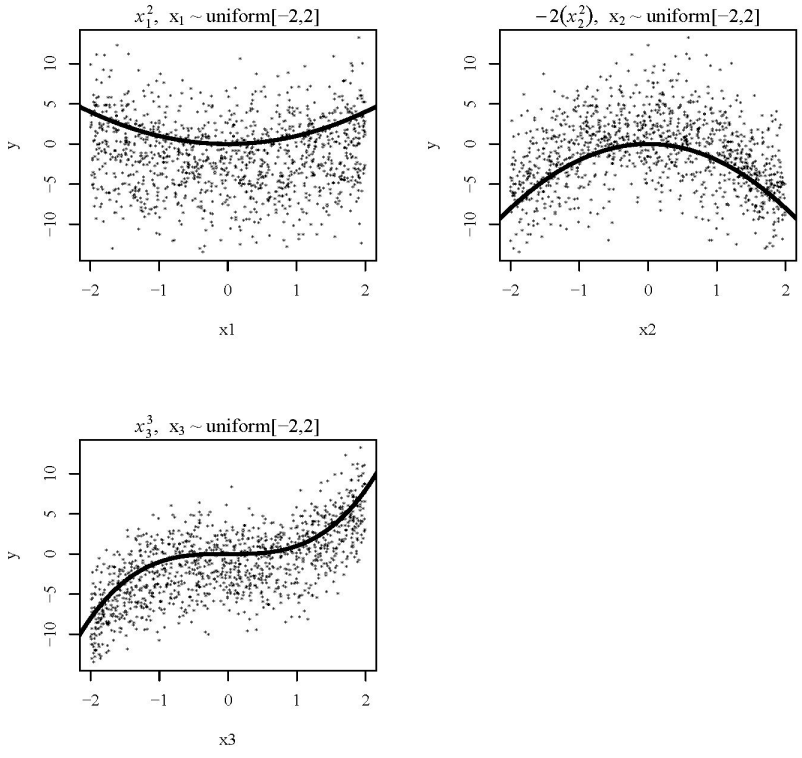


[그림 6] 선형함수를 이용한 자료2에 대한 산점도

3. 다항함수

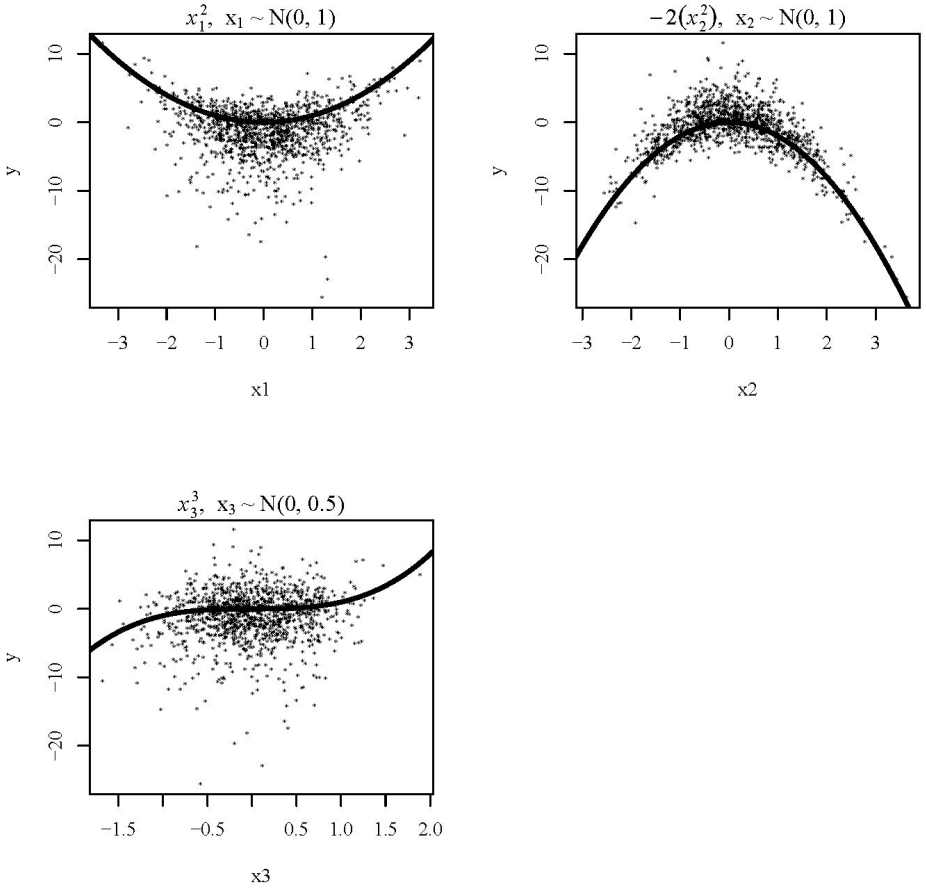
다항함수 자료1은 x_1, x_2, x_3의 세 개의 설명변수와 연속형 반응변수를 가지는 자료로 [그림 7]과 같은 형태의 실험 자료이다. 각 설명변수 x_1, x_2, x_3는 -2에서 2의 범위를 가지는 균일분포로부터 생성되었다. 오차항 $\epsilon_1, \epsilon_2, \epsilon_3$은 평균이 0, 표준편차 1을 따르는 정규분포로부터 생성되었다. 위의 설명변수와 오차항을 이용하여 아래의 식(2.23)과 같이 연속형 반응변수를 생성하였다.

$$\begin{aligned} y1 &= x1^2 + \epsilon_1, \\ y2 &= -2x_2^2 + \epsilon_2, \\ y3 &= x_3^3 + \epsilon_3 \\ y &= y_1 + y_2 + y_3 \end{aligned} \tag{2.23}$$



[그림 7] 다항함수를 이용한 자료1에 대한 산점도

다항함수 자료2는 x_1, x_2, x_3의 세 개의 설명변수와 연속형 반응변수를 가지는 자료로 [그림 8]과 같은 형태의 실험 자료이다. 각 설명변수 x_1, x_2와 오차항 $\epsilon_1, \epsilon_2, \epsilon_3$은 평균 0, 표준편차 1을 따르는 정규분포로부터 생성되었으며, x_3는 평균 0, 표준편차 1을 따르는 정규분포로부터 생성되었다. 위의 설명변수와 오차항을 이용하여 위의 식(2.23)과 같이 연속형 반응변수를 생성하였다.



[그림 8] 다항함수를 이용한 자료2에 대한 산점도

4. 프리드먼 데이터

프리드먼 데이터는 시뮬레이션 데이터로 MARS(Friedman, 1991) 논문에 설명된 자료를 이용하였다. 자료는 모두 두 가지가 있는데 자료의 구성은 아래와 같다.

1) 데이터1

10개의 연속형 설명변수 $x_1,\cdots,x_{10}$와 연속형 반응변수 y를 가지며, 설명변수 $x_1,\cdots,x_{10}$은 모두 0에서 1까지의 균일분포로부터 생성되었다. 오차항 ϵ은 평균이 0, 표준편차는 1을 따르는 정규분포로부터 추출하였으며, 설명변수와 오차항을 이용한 반응변수는 아래와 같은 식(2.24)을 통해 만들어졌다. 프리드먼 데이터는 200개의 관찰치를 가지지만 본 논문에서는 실험의 용이함을 위하여 1,200개의 자료를 생성하였다.

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon \qquad (2.24)$$

2) 데이터2

4개의 연속형 설명변수 $x_1,\cdots,x_4$와 연속형 반응변수 y를 가지며, 오차항 ϵ은 평균이 0, 표준편차는 1을 따르는 정규분포로부터 추출하였다. 설명변수는 각각 다음과 같은 분포를 따른다. x_1은 0부터 100까지의 범위를 가지는 균일분포, x_2은 40π부터 560π의 범위를 가지는 균일분포, x_3은 0부터 1까지의 범위를 가지는 균일분포, x_4는 1부터 11의 범위를 가지는 균일분포로부터 생성되었다. 설명변수와 오차항을 이용한 반응변수를 프리드먼 데이터2는 식(2.25), 1,200개의 자료를 생성하였다.

$$y = \left(x_1^2 + \left(x_2 x_3 - \left(\frac{1}{x_2 x_4} \right) \right)^2 \right)^{1/2} + \epsilon \qquad (2.25)$$

위의 데이터를 제안한 방법과 기존의 방법에 적용하여 비교한다. 모형 비교는 다음과 같이 이루어졌다. 각 모의실험 자료에서의 훈련데이터(train data)의 크기는 200개, 검증데이터(test data)의 크기는 1,000개로 한다. 결합 예측을 위한 단일 분류자의 개수는 각 100개로 하고, 결합 예측을 위한 붓스트랩 표본의 크기는 과적합(overfitting)을 막기 위하여 100개로 하였다. 또한, 반복실험은 각 알고리즘 마다 1,000번 실시하였으며, 그 결합의 방법과 붓스트랩 표본추출 방법에 따라 아래의 [표 8]에서 설명한 6가지 방법으로 모의실험을 실시한다.

[표 8] 결합 예측자의 종류

1. GAMbag1 : 붓스트랩 표본에서 배깅 방법을 이용한다. 여기서 단일 예측자를 일반화가법모형으로 하며, 결합 예측자는 100개의 단일 예측자의 평균(mean)으로 한다.
2. GAMbag2 : 붓스트랩 표본에서 배깅 방법을 이용한다. 여기서 단일 예측자를 일반화가법모형으로 하며, 결합 예측자는 100개의 단일 예측자의 절사평균(trimed mean)으로 한다. 여기서, 절사의 비율은 10%로 한다.
3. GAMbag3 : 붓스트랩 표본에서 배깅 방법을 이용한다. 여기서 단일 예측자를 일반화가법모형으로 하며, 결합 예측자는 100개의 단일 예측자의 중위수(median)로 한다.
4. GAMrnf1 : 붓스트랩 표본에서 랜덤포레스트 방법을 이용한다. 여기서 단일 예측자를 일반화가법모형으로 하며, 결합 예측자는 100개의 단일

예측자의 평균(mean)으로 한다. 무작위로 선택되는 변수의 개수는 $\sqrt{p}$ 개로 한다.

5. GAMrnf2 : 붓스트랩 표본에서 랜덤포레스트 방법을 이용한다. 여기서 단일 예측자를 일반화가법모형으로 하며, 결합 예측자는 100개의 단일 예측자의 절사평균(trimed mean)으로 한다. 여기서, 절사의 비율은 10%로 한다. 또한, 무작위로 선택되는 변수의 개수는 $\sqrt{p}$ 개로 한다.
6. GAMrnf3 : 붓스트랩 표본에서 랜덤포레스트 방법을 이용한다. 여기서 단일 예측자를 일반화가법모형으로 하며, 결합 예측자는 100개의 단일 예측자의 중위수(median)로 한다. 무작위로 선택되는 변수의 개수는 $\sqrt{p}$ 개로 한다.

위의 6가지 결합 예측자와 단일 일반화가법모형, 배깅, 랜덤포레스트를 모의실험 자료를 이용하여 모형에 적합하고, 검증데이터에 대한 예측값으로 구한 제곱근평균제곱오차(RMSE)를 통해 모형의 우월성을 비교한다.

3.2. 적용결과 및 해석

3.1.절에서 제안한 일반화가법모형을 이용한 앙상블 기법에 대한 모의실험 결과는 아래와 같다.

모형에 포함된 설명변수에 대한 평활함수의 자유도가 높을수록 훈련자료에 잘 적합되는 모형이 된다. 전체데이터 1,200개에서 200개의 훈련데이터로 모형을 만들고 1,000개의 검증데이터로 제곱근평균제곱오차(RMSE)를 구하였다. 모든 알고리즘에 대하여 반복실험 1,000번을 실시하여 각 RMSE의 평균, 또는 가장 작은 RMSE를 가지는 횟수를 비율로 표시하였다.

[표 9]와 [표 10]의 결과를 통해서 삼각함수 자료1에서의 제안한 방법론과 기존의 방법론을 비교할 수 있다. 자유도가 낮은 모형에서는 배깅이 다른 방법론에 비해 더 좋은 성능을 가지는 것을 확인 할 수 있다. 그러나 자유도가 커지면 커질수록 즉, 데이터에 잘 적합되는 모형에서는 단일 일반화가법모형이 좋은 성능을 가짐을 알 수 있다. [표 9]를 통해서 자유도 1, 2에서는 배깅이 가장 작은 RMSE를 가지며, 자유도 3~7에서는 단일 일반화가법모형이 가장 작은 RMSE를 가짐을 알 수 있다. 이 결과를 통해 [그림 1]과 같은 비선형성을 가지는 자료에서는 일반화가법모형 단일 예측자로도 충분한 예측력을 얻을 수 있는 것을 알 수 있다.

다음의 [표 11]와 [표 12]의 결과를 통해 삼각함수 자료2에서의 제안한 방법론과 기존의 방법론을 비교할 수 있다. 이 자료에서는 근소한 차이지만 제안한 방법론이 더 좋은 성능을 가지는 것을 확인 할 수 있다. 자유도 2, 3에서는 배깅을 이용한 일반화가법모형이 가장 작은 RMSE를 가지며, 자유도 6, 7에서는 랜덤포레스트를 이용한 일반화가법모형이 가장 작은 RMSE를 가진다. 두 방법론 모두 다중 예측자에 대한 결합 방법은 절사평균으로 한 것이다. 이 결

과를 통해서 [그림 2]와 같이 선형이나 비선형을 가정하기 어려운 경우에는 자유도를 높여 자료에 대한 적합을 강하게 하고, 그로부터 얻은 단일 예측자를 결합한 앙상블 기법이 더 좋은 성능을 얻음을 알 수 있다.

[표 13]와 [표 14]의 결과를 통해서 삼각함수 자료3에서의 제안한 방법론과 기존의 방법론을 비교할 수 있다. 삼각함수 자료3은 기존의 삼각함수 자료2과 동일한 분포의 설명변수에 오차항의 표준편차를 크게 한 자료로 이러한 편차가 큰 자료에서는 제안한 방법론이 기존의 방법론 보다 좋은 성능을 가지는 것을 확인 할 수 있다. 자유도 1~2까지는 랜덤포레스트를 이용하여 절사평균으로 결합한 모형이 좋은 성능을 가지며, 자유도 3~7에서는 랜덤포레스트를 이용하여 중위수로 결합한 모형이 가장 좋은 성능을 가지는 것을 알 수 있다.

[표 15]와 [표 16]의 결과를 통해서 삼각함수 자료4에서의 제안한 방법론과 기존의 방법론을 비교할 수 있다. 삼각함수 자료4은 기존의 삼각함수 자료2과 동일한 분포의 설명변수에 반응변수와 상관없는 설명변수를 추가한 자료로 이러한 자료에서는 제안한 방법론이 기존의 방법론 보다 좋은 성능을 가지는 것을 확인 할 수 있었다. [표 15]를 통해서 자유도 1에서는 랜덤포레스트가 좋은 성능을 가지며, 자유도 2~7에서는 제안한 방법론이 기존의 방법론보다 좋은 성능을 가짐을 알 수 있다. 또한 자유도 2~5까지는 배깅을 이용한 일반화가법모형이 더 좋은 예측력을 가지며, 자유도 6~7에서는 랜덤포레스트를 이용한 일반화가법모형이 더 좋은 예측력을 가짐을 확인할 수 있다.

[표 17]와 [표 18]에서는 선형함수 자료1에서 제안한 방법론과 기존의 방법론을 비교할 수 있다. 이 자료에서 제안한 방법론이 더 좋은 성능을 가지는 것을 확인 할 수 있다. [표 17]을 통해서 자유도 1~4에서는 단일 일반화가법모형이 가장 작은 RMSE를 가지며, 자유도 5~7에서는 배깅을 이용한 일반화가법모형이 가장 작은 RMSE를 가짐을 확인할 수 있다. 여기서 다중 예측자들

의 결합방법은 평균을 사용한 것이다. 그러나 그 평균 RMSE의 차이가 아주 근소하여 [표 18]의 1000번의 실험에서는 모든 자유도에서 단일 일반화가법모형이 40%의 비율로 다른 방법론에 비해 좋은 성능을 보임을 알 수 있다.

[표 19]와 [표 20]의 결과를 통해서 선형함수 자료2에서 제안한 방법론과 기존의 방법론을 비교할 수 있다. 이 자료에서는 근소한 차이로 단일 일반화가법모형이 가장 작은 RMSE를 가지는 것을 확인할 수 있다.

[표 21]와 [표 22]의 결과를 통해서 다항함수 자료1에서 제안한 방법론과 기존의 방법론을 비교할 수 있다. 이 자료에서는 자유도 7에서 배깅을 이용한 일반화가법모형이 가장 작은 RMSE를 가지는 것을 확인할 수 있다. 여기서 다중 예측자의 결합 방법은 평균을 사용한 것이다. 그러나 [표 22]의 결과를 살펴보면 자유도 1, 2를 제외한 나머지 결과에서 거의 대부분 단일 일반화가법모형이 다른 방법론에 비해 좋은 성능을 가지는 것을 확인할 수 있다.

[표 23]와 [표 24]의 결과를 통해서 다항함수 자료2에서 제안한 방법론과 기존의 방법론을 비교할 수 있다. 이 자료에서는 자유도 1, 2에서는 랜덤포레스트가 나머지에서는 일반화가법모형이 가장 작은 RMSE를 가지는 것을 확인할 수 있다.

[표 25]와 [표 26]의 결과를 통해서 프리드먼 데이터1에서 제안한 방법론과 기존의 방법론을 비교할 수 있다. 이 자료에서 본 논문에서 제안한 방법론이 더 좋은 성능을 가지는 것을 확인 할 수 있다. 자유도 1에서는 랜덤포레스트, 자유도 2에서는 단일 일반화가법모형이 가장 작은 RMSE를 가진다. 자유도 3~7에서는 배깅을 이용한 일반화가법모형이 가장 작은 RMSE를 가진다. 여기서 다중 예측자들의 결합방법은 중위수를 사용한 것이다.

[표 27]와 [표 28]의 결과를 통해서 프리드먼 데이터2에서 방법론들을 비교하는 것이 가능하다. 이 자료에서는 랜덤포레스트가 가장 좋은 성능을 가지는

것을 확인할 수 있다.

10개의 데이터를 통해서 제안한 방법론과 기존의 방법론을 비교한 결과 일부 데이터에서만 제안한 방법론이 좋은 것을 확인 할 수 있었다. 주로 데이터의 분포가 선형이나 비선형의 어떤 패턴을 가정하기 어려운 경우의 자료에서 제안한 방법의 성능이 다른 방법론에 비해 좋음을 확인할 수 있었다. 따라서 자유도를 높여 자료에 잘 적합되는 단일 예측자를 결합하는 경우의 예측의 성능을 향상시키는 것을 확인할 수 있었다. 그리고 삼각함수 자료3과 같이 자료의 오차가 큰 경우에서 제안한 앙상블 방법론이 기존의 방법론에 비해 예측에 대한 성능이 더 좋은 것을 확인할 수 있었다. 또한 프리드먼 데이터1나 삼각함수 자료4와 같이 모형에 고려되는 설명변수의 개수가 많은 경우 제안한 앙상블 방법론이 유용함을 확인 할 수 있었다.

[표 9] 삼각함수 자료1에 대한 1,000번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	2.2134	2.0867	**2.0098**	**1.9591**	**1.9269**	**1.9147**	**1.9140**
배깅	**2.0221**	**2.0200**	2.0198	2.0218	2.0198	2.0201	2.0205
랜덤포레스트	2.0253	2.0241	2.0222	2.0268	2.0237	2.0233	2.0235
GAMbag1 (mean)	2.2151	2.7044	2.1395	1.9633	1.9311	1.9174	1.9153
GAMbag2 (trimed mean)	2.2149	2.0882	2.0127	1.9628	1.9309	1.9174	1.9154
GAMbag3 (median)	2.2149	2.0881	2.0125	1.9629	1.9313	1.9180	1.9163
GAMrnf1 (mean)	2.3556	2.1739	2.0396	1.9911	1.9583	1.9394	1.9275
GAMrnf2 (trimed mean)	2.2062	2.1088	2.0388	1.9911	1.9589	1.9405	1.9289
GAMrnf3 (median)	2.2054	2.1078	2.0398	1.9947	1.9647	1.9473	1.9364

[표 10] 삼각함수 자료1에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	0.0000	0.0080	**0.3260**	**0.6620**	**0.6600**	**0.5400**	**0.3880**
배깅	**0.5780**	**0.5690**	0.2870	0.0150	0.0000	0.0000	0.0000
랜덤포레스트	0.4220	0.4170	0.1950	0.0020	0.0000	0.0000	0.0000
GAMbag1 (mean)	0.0000	0.0020	0.0220	0.0690	0.0750	0.1240	0.1740
GAMbag2 (trimed mean)	0.0000	0.0000	0.0470	0.0590	0.0690	0.0850	0.0860
GAMbag3 (median)	0.0000	0.0030	0.0940	0.1360	0.1160	0.1120	0.0910
GAMrnf1 (mean)	0.0000	0.0000	0.0020	0.0270	0.0480	0.1000	0.2220
GAMrnf2 (trimed mean)	0.0000	0.0000	0.0190	0.0240	0.0300	0.0380	0.0390
GAMrnf3 (median)	0.0000	0.0010	0.0080	0.0060	0.0020	0.0010	0.0000

[표 11] 삼각함수 자료2에 대한 1,000번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	1.9402	1.8776	1.8361	**1.8256**	**1.8275**	1.8409	1.8496
배깅	**1.8989**	1.8996	1.9002	1.8998	1.8983	1.8998	1.8992
랜덤포레스트	1.9245	1.9246	1.9262	1.9247	1.9237	1.9252	1.9261
GAMbag1 (mean)	2.1188	1.8779	1.8352	1.8260	1.8300	1.8446	1.8561
GAMbag2 (trimed mean)	1.9430	1.8771	**1.8346**	1.8257	1.8295	1.8442	1.8556
GAMbag3 (median)	1.9428	**1.8766**	1.8347	1.8259	1.8294	1.8439	1.8552
GAMrnf1 (mean)	1.9453	1.8941	1.8545	1.8373	1.8320	1.8364	1.8391
GAMrnf2 (trimed mean)	1.9442	1.8942	1.8552	1.8379	1.8321	**1.8358**	**1.8379**
GAMrnf3 (median)	1.9445	1.8978	1.8602	1.8430	1.8367	1.8394	1.8403

[표 12] 삼각함수 자료2에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	0.0340	0.2620	**0.2850**	**0.3590**	**0.3310**	**0.2680**	**0.2550**
배깅	**0.8260**	0.1640	0.0010	0.0000	0.0010	0.0080	0.0140
랜덤포레스트	0.0600	0.0060	0.0000	0.0000	0.0000	0.0000	0.0000
GAMbag1 (mean)	0.0010	0.0560	0.1900	0.1800	0.1290	0.0860	0.0650
GAMbag2 (trimed mean)	0.0040	0.1100	0.1480	0.0830	0.0580	0.0430	0.0240
GAMbag3 (median)	0.0120	**0.2890**	0.2660	0.1750	0.1320	0.1020	0.0600
GAMrnf1 (mean)	0.0060	0.0500	0.0580	0.0940	0.1470	0.1500	0.1570
GAMrnf2 (trimed mean)	0.0400	0.0570	0.0400	0.0920	0.1590	0.2380	0.2520
GAMrnf3 (median)	0.0170	0.0060	0.0120	0.0170	0.0430	0.1050	0.1730

[표 13] 삼각함수 자료3에 대한 1,000번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	8.9068	8.9422	8.9933	9.0489	9.1177	9.1969	9.2750
배깅	9.2484	9.2456	9.2464	9.2428	9.2408	9.2462	9.2504
랜덤포레스트	9.4122	9.4130	9.4108	9.4105	9.4115	9.4185	9.4182
GAMbag1 (mean)	8.9086	8.9420	8.9909	9.0455	9.1076	9.1822	9.2490
GAMbag2 (trimed mean)	8.9087	8.9429	8.9927	9.0491	9.1125	9.1894	9.2585
GAMbag3 (median)	8.9097	8.9451	8.9965	9.0555	9.1224	9.2033	9.2773
GAMrnf1 (mean)	8.8724	8.8916	8.9105	8.9302	8.9611	8.9973	9.0250
GAMrnf2 (trimed mean)	**8.8719**	8.8902	8.9078	8.9266	8.9565	8.9917	9.0193
GAMrnf3 (median)	8.8724	**8.8890**	**8.9047**	**8.9203**	**8.9487**	**8.9821**	**9.0074**

[표 14] 삼각함수 자료3에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	0.0860	0.0230	0.0030	0.0000	0.0000	0.0000	0.0000
배깅	0.0000	0.0000	0.0000	0.0010	0.0010	0.0010	0.0040
랜덤포레스트	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMbag1 (mean)	0.0500	0.0420	0.0080	0.0050	0.0000	0.0020	0.0000
GAMbag2 (trimed mean)	0.0270	0.0110	0.0030	0.0010	0.0010	0.0010	0.0000
GAMbag3 (median)	0.0600	0.0250	0.0090	0.0010	0.0000	0.0000	0.0000
GAMrnf1 (mean)	0.2340	0.2150	0.1770	0.1560	0.1540	0.1620	0.1620
GAMrnf2 (trimed mean)	0.1850	0.1810	0.1850	0.1490	0.1250	0.1240	0.1240
GAMrnf3 (median)	**0.3580**	**0.5030**	**0.6150**	**0.6870**	**0.7190**	**0.7100**	**0.7100**

[표 15] 삼각함수 자료4에 대한 1,000번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	1.9992	1.9368	1.9404	1.9863	2.0573	2.1498	2.2301
배깅	1.9070	1.9102	1.9039	1.9077	1.9084	1.9091	1.9050
랜덤포레스트	**1.8791**	1.8839	1.8836	1.8791	1.8809	1.8843	1.8803
GAMbag1 (mean)	3.1232	1.8856	2.4255	**1.8352**	**1.8541**	1.8799	1.9097
GAMbag2 (trimed mean)	1.9401	**1.8656**	**1.8314**	1.8355	1.8545	1.8805	1.9107
GAMbag3 (median)	1.9401	**1.8656**	1.8319	1.8364	1.8557	1.8821	1.9129
GAMrnf1 (mean)	2.2111	1.9274	2.0543	1.8910	1.8742	**1.8747**	**1.8766**
GAMrnf2 (trimed mean)	1.9645	1.9298	1.9013	1.8836	1.8762	1.8764	1.8779
GAMrnf3 (median)	1.9737	1.9374	1.9072	1.8885	1.8805	1.8805	1.8816

[표 16] 삼각함수 자료4에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	0.0000	0.0160	0.0070	0.0050	0.0000	0.0000	0.0000
배깅	0.1670	0.0990	0.0270	0.0380	0.0500	0.0880	0.1110
랜덤포레스트	**0.7440**	0.2630	0.0840	0.1040	0.1820	**0.2200**	**0.3150**
GAMbag1 (mean)	0.0220	0.1540	**0.3860**	**0.4110**	**0.2880**	0.2190	0.1090
GAMbag2 (trimed mean)	0.0270	0.1670	0.2270	0.1770	0.1090	0.0670	0.0370
GAMbag3 (median)	0.0340	**0.2900**	0.2640	0.1810	0.1890	0.1020	0.0460
GAMrnf1 (mean)	0.0000	0.0110	0.0050	0.0660	0.1570	0.1990	0.1950
GAMrnf2 (trimed mean)	0.0030	0.0000	0.0000	0.0110	0.0160	0.0370	0.0940
GAMrnf3 (median)	0.0030	0.0000	0.0000	0.0070	0.0090	0.0680	0.0930

[표 17] 선형함수 자료1에 대한 1,000번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	**1.8228**	**1.8276**	**1.8397**	**1.8477**	1.8604	1.8711	1.8835
배깅	2.5210	2.5203	2.5281	2.5166	2.5269	2.5324	2.5273
랜덤포레스트	2.0947	2.0994	2.0992	2.0952	2.1015	2.0998	2.0997
GAMbag1 (mean)	1.9674	1.9372	1.8984	1.8609	**1.8601**	**1.8705**	**1.8823**
GAMbag2 (trimed mean)	1.8231	1.8281	1.8402	1.8480	1.8602	1.8707	1.8825
GAMbag3 (median)	1.8232	1.8282	1.8404	1.8484	1.8607	1.8715	1.8834
GAMrnf1 (mean)	3.2783	3.3850	3.1004	3.0833	3.0969	3.0804	3.1029
GAMrnf2 (trimed mean)	2.8411	2.8374	2.8281	2.8682	2.8698	2.8428	2.8566
GAMrnf3 (median)	2.3277	2.3340	2.3382	2.3518	2.3664	2.3560	2.3602

[표 18] 선형함수 자료1에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	**0.4460**	**0.4330**	**0.4890**	**0.4770**	**0.4420**	**0.4070**	**0.3760**
배깅	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
랜덤포레스트	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMbag1 (mean)	0.1780	0.1920	0.2060	0.2290	0.2270	0.2760	0.3230
GAMbag2 (trimed mean)	0.1180	0.1060	0.0940	0.1130	0.1270	0.1610	0.1140
GAMbag3 (median)	0.2580	0.2690	0.2110	0.1810	0.2040	0.1560	0.1870
GAMrnf1 (mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf2 (trimed mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf3 (median)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

[표 19] 선형함수 자료2에 대한 1,000번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	**1.7044**	**1.7059**	**1.7126**	**1.7209**	**1.7316**	**1.7427**	**1.7549**
배깅	1.8904	1.8905	1.8897	1.8890	1.8887	1.8914	1.8891
랜덤포레스트	1.8671	1.8645	1.8655	1.8668	1.8650	1.8650	1.8655
GAMbag1 (mean)	1.7050	1.7066	1.7133	1.7223	1.7324	1.7435	1.7555
GAMbag2 (trimed mean)	1.7050	1.7066	1.7134	1.7226	1.7329	1.7441	1.7563
GAMbag3 (median)	1.7051	1.7068	1.7138	1.7231	1.7337	1.7452	1.7576
GAMrnf1 (mean)	2.0253	2.0195	2.0286	2.0329	2.0345	2.0459	2.0474
GAMrnf2 (trimed mean)	1.9804	1.9735	1.9816	1.9854	1.9856	1.9957	1.9966
GAMrnf3 (median)	1.8412	1.8382	1.8461	1.8527	1.8577	1.8681	1.8746

[표 20] 선형함수 자료2에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	**0.5170**	**0.5140**	**0.4950**	**0.5580**	**0.5380**	**0.5290**	**0.5180**
배깅	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010	0.0000
랜덤포레스트	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010
GAMbag1 (mean)	0.1710	0.1990	0.2490	0.2530	0.2810	0.2860	0.3030
GAMbag2 (trimed mean)	0.0940	0.0770	0.0860	0.0560	0.0510	0.0510	0.0610
GAMbag3 (median)	0.2180	0.2100	0.1700	0.1310	0.1260	0.1310	0.1100
GAMrnf1 (mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf2 (trimed mean)	0.0000	0.0000	0.0000	0.0010	0.0000	0.0000	0.0010
GAMrnf3 (median)	0.0000	0.0000	0.0000	0.0010	0.0040	0.0020	0.0060

[표 21] 다항함수 자료1에 대한 1,000번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	3.3736	2.2385	**1.9053**	**1.8347**	**1.8259**	**1.8295**	1.8382
배깅	2.4499	2.4497	2.4457	2.4447	2.4473	2.4467	2.4494
랜덤포레스트	**2.2094**	**2.2067**	2.2041	2.2048	2.2071	2.2067	2.2056
GAMbag1 (mean)	3.5918	2.2526	1.9145	1.8391	1.8283	1.8305	**1.8380**
GAMbag2 (trimed mean)	3.3870	2.2503	1.9135	1.8388	1.8282	1.8306	1.8382
GAMbag3 (median)	3.3847	2.2485	1.9130	1.8391	1.8286	1.8312	1.8391
GAMrnf1 (mean)	3.5480	2.7838	2.4489	2.3588	2.3335	2.3233	2.3247
GAMrnf2 (trimed mean)	3.4983	2.7329	2.4359	2.3425	2.3155	2.3044	2.3047
GAMrnf3 (median)	3.3836	2.6804	2.3955	2.3044	2.2795	2.2703	2.2710

[표 22] 다항함수 자료1에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	0.0000	0.2480	**0.8900**	**0.7530**	**0.6220**	**0.5180**	**0.4300**
배깅	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
랜덤포레스트	**1.0000**	**0.7130**	0.0000	0.0000	0.0000	0.0000	0.0000
GAMbag1 (mean)	0.0000	0.0020	0.0150	0.0710	0.1380	0.2200	0.2820
GAMbag2 (trimed mean)	0.0000	0.0030	0.0160	0.0610	0.0820	0.1040	0.1210
GAMbag3 (median)	0.0000	0.0340	0.0790	0.1150	0.1580	0.1580	0.1670
GAMrnf1 (mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf2 (trimed mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf3 (median)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

[표 23] 다항함수 자료2에 대한 1,000번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	3.6323	2.2766	**1.9211**	**1.8677**	**1.8565**	**1.8639**	**1.8759**
배깅	2.3736	2.3753	2.3703	2.3768	2.3666	2.3688	2.3776
랜덤포레스트	**2.2172**	**2.2200**	2.2153	2.2172	2.2092	2.2098	2.2200
GAMbag1 (mean)	3.7014	2.3041	1.9443	1.8851	1.8701	1.8746	1.8839
GAMbag2 (trimed mean)	3.6954	2.2972	1.9394	1.8815	1.8672	1.8725	1.8823
GAMbag3 (median)	3.6906	2.2922	1.9360	1.8793	1.8661	1.8722	1.8834
GAMrnf1 (mean)	3.6875	2.6910	2.3417	2.2438	2.2027	2.1956	2.2017
GAMrnf2 (trimed mean)	3.6836	2.6735	2.3127	2.2135	2.1718	2.1663	2.1738
GAMrnf3 (median)	3.6849	2.5606	2.1590	2.0710	2.0403	2.0460	2.0648

[표 24] 다항함수 자료2에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	0.0000	0.1780	**0.9070**	**0.8060**	**0.7160**	**0.6400**	**0.5850**
배깅	0.0060	0.0020	0.0000	0.0000	0.0000	0.0000	0.0000
랜덤포레스트	**0.9940**	**0.7830**	0.0000	0.0000	0.0000	0.0000	0.0000
GAMbag1 (mean)	0.0000	0.0000	0.0040	0.0260	0.0780	0.1420	0.2060
GAMbag2 (trimed mean)	0.0000	0.0010	0.0060	0.0270	0.0560	0.0630	0.0760
GAMbag3 (median)	0.0000	0.0360	0.0830	0.1410	0.1500	0.1530	0.1300
GAMrnf1 (mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf2 (trimed mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf3 (median)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0020	0.0030

[표 25] 프리드먼 데이터1에 대한 1,000번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	2.4923	**1.6267**	1.4963	1.5105	1.5411	1.5747	1.6032
배깅	2.5746	2.5735	2.5663	2.5709	2.5600	2.5657	2.5691
랜덤포레스트	**2.4853**	2.4848	2.4740	2.4737	2.4661	2.4757	2.4757
GAMbag1 (mean)	2.5080	2.0518	1.5004	1.7398	1.5194	1.5397	1.5554
GAMbag2 (trimed mean)	2.5064	1.6642	1.4964	1.4950	1.5153	1.5356	1.5514
GAMbag3 (median)	2.5057	1.6620	**1.4936**	**1.4925**	**1.5129**	**1.5336**	**1.5497**
GAMrnf1 (mean)	8.1671	3.6645	3.6089	3.5985	3.6075	3.6149	3.6225
GAMrnf2 (trimed mean)	3.8765	3.6963	3.6365	3.6209	3.6250	3.6278	3.6316
GAMrnf3 (median)	3.9390	3.7514	3.6857	3.6631	3.6615	3.6580	3.6575

[표 26] 프리드먼 데이터1에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	0.3770	**0.9950**	0.4090	0.0700	0.0180	0.0050	0.0000
배깅	0.0750	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
랜덤포레스트	**0.4820**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMbag1 (mean)	0.0030	0.0000	0.0010	0.0240	0.0210	0.0290	0.0500
GAMbag2 (trimed mean)	0.0060	0.0000	0.0580	0.1430	0.2040	0.2550	0.2820
GAMbag3 (median)	0.0570	0.0050	**0.5320**	**0.7630**	**0.7570**	**0.7110**	**0.6680**
GAMrnf1 (mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf2 (trimed mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf3 (median)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

[표 27] 프리드먼 데이터2에 대한 1,000번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	131.778	131.952	133.023	134.167	135.000	136.148	137.103
배깅	87.797	87.730	88.075	87.793	87.812	87.550	87.353
랜덤포레스트	**75.887**	**75.977**	**76.066**	**75.864**	**75.810**	**75.795**	**75.780**
GAMbag1 (mean)	132.170	132.499	133.635	134.999	1236.766	137.187	138.167
GAMbag2 (trimed mean)	131.823	131.973	133.093	134.464	135.368	136.683	137.663
GAMbag3 (median)	131.590	131.594	132.824	134.235	135.189	136.577	137.654
GAMrnf1 (mean)	222.052	222.613	224.661	225.847	224.705	225.828	226.423
GAMrnf2 (trimed mean)	222.094	222.469	224.360	225.134	223.729	224.581	224.805
GAMrnf3 (median)	232.326	232.397	233.660	233.366	231.641	231.582	230.869

[표 28] 프리드먼 데이터2에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
배깅	0.0310	0.0350	0.0270	0.0350	0.0330	0.0490	0.0340
랜덤포레스트	**0.9690**	**0.9650**	**0.9730**	**0.9650**	**0.9670**	**0.9510**	**0.9660**
GAMbag1 (mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMbag2 (trimed mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMbag3 (median)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf1 (mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf2 (trimed mean)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GAMrnf3 (median)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

3.3. 실제자료의 적용

본 논문에서 제안한 방법론을 실제 자료에 적용하여 그 결과를 다른 방법론들과 비교해 보았다.

국내 요양보험에 관한 자료로 청결에 소요되는 서비스시간을 반응변수로 하였다. 또한 설명변수는 각 영역별 제약에 대한 100점 환산 점수로 하였으며, 결측값을 제외한 총 3,744개의 관찰치로 이루어져 있다. 실제자료를 제안한 방법론에 적용시키고 그 예측의 효용성을 비교하기 위하여 훈련데이터는 500개 검증데이터는 3,244개의 관찰치로 하고, 500번 반복 수행한 결과를 비교하였다. 자료에 대해서 간단히 정리하면 아래 [표 29]과 같다. 또한 자료의 기초통계량을 [표 30]을 통해서 확인 할 수 있다. [표 30]에 의하면 반응변수인 청결에 소요되는 서비스 시간은 0부터 39시간의 범위를 가지며, 평균 11.12시간이다. 설명변수는 모두 100점 환산 점수로 0점부터 100점의 범위를 가지나, 간호욕구영역 100점 변수는 최대값이 76.11점으로 100점을 가지는 관찰치는 없는 것을 알 수 있다.

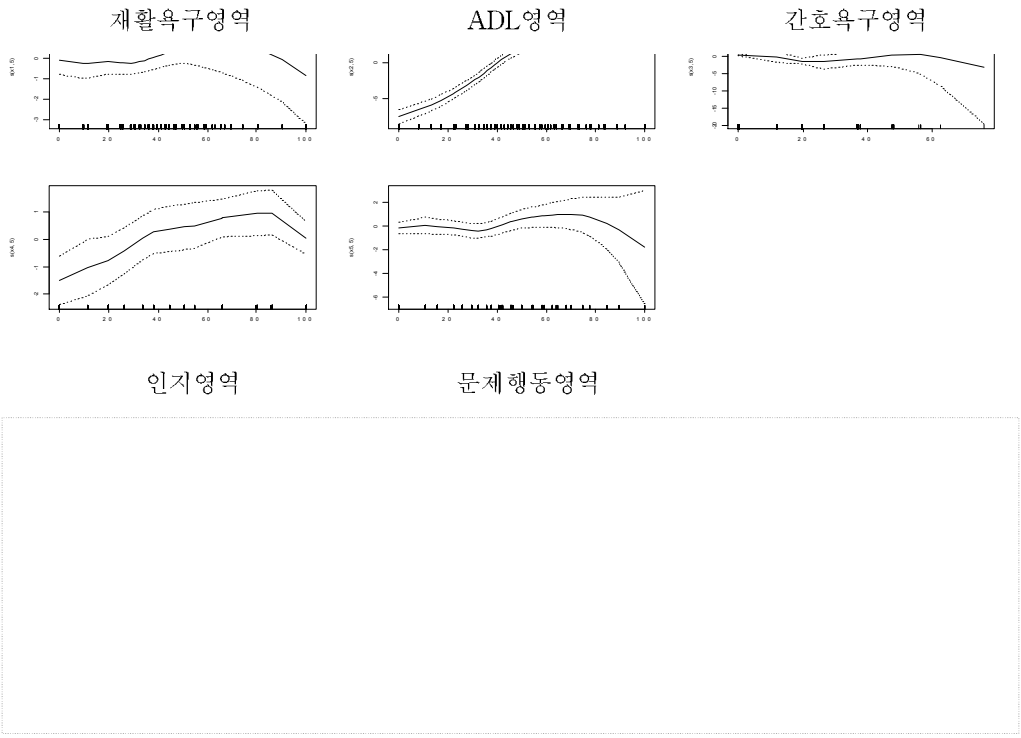
[표 29] 자료의 구성

반응변수	청결에 소요되는 서비스시간
설명변수	재활욕구영역 100점 득점
	ADL(일상상활동작평가)영역 100점 득점
	간호욕구영역 100점 득점
	인지영역 100점 득점
	문제행동영역 100점 득점

[표 30] 자료의 기초 통계량

변수	청결서비스 시간	재활욕구 영역	ADL영역	간호욕구 영역	인지영역	문제행동 영역
평균	11.08	29.25	46.31	4.49	58.13	23.83
표준편차	11.12	23.01	30.78	9.85	36.08	22.13
중위수	7.31	28.93	45.40	0.00	65.71	22.20
최소값	0.00	0.00	0.00	0.00	0.00	0.00
최대값	39.00	100.00	100.00	76.11	100.00	100.00

자료에 대한 이해를 돕기 위하여 단일 일반화가법모형을 적합시켜 얻은 결과를 아래 [그림 9]과 같이 그래프로 나타내었다. [그림 9]을 통해서 재활욕구영역, 인지영역, 문제행동영역에서 모두 비선형 효과를 가짐을 알 수 있다. 일반화가법모형에서의 평활함수의 자유도는 df=5로 하였다.



[그림 9] 단일 일반화가법모형 적합결과

장기요양보험 자료에 대한 방법론의 적용결과는 아래의 [표 31]과 [표 32]을 통해서 알 수 있다. 그 결과 랜덤포레스트 방법을 적용하여 다중 예측자를 결합하는 것으로 제안한 방법론이 다른 방법론에 비해 경쟁력을 가지는 것을 확인할 수 있다. 여기서 결합 방법은 중위수를 이용한 것이다.

[표 31] 실제 데이터에 대한 500번 반복의 평균 RMSE

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	29.7988	29.2881	28.5880	27.1467	28.8815	29.6699	28.2673
배깅	10.2477	10.2428	10.2414	10.2449	10.2553	10.2483	10.2464
랜덤포레스트	10.3029	10.3033	10.3069	10.3030	10.3125	10.3061	10.3048
GAMbag1 (mean)	25.4754	23.9948	26.5504	26.8968	26.4901	27.5096	43.7885
GAMbag2 (trimed mean)	17.2810	16.6938	17.2737	17.7157	18.5126	17.4885	23.1285
GAMbag3 (median)	13.1995	13.4629	13.6103	13.6311	14.5255	13.9429	15.3070
GAMrnf1 (mean)	16.9993	16.6709	18.8617	17.2268	16.8272	17.6011	26.1325
GAMrnf2 (trimed mean)	11.2079	11.0881	11.0510	11.0620	11.1323	11.2450	12.0074
GAMrnf3 (median)	**10.2206**	**10.2213**	**10.2232**	**10.2164**	**10.2230**	**10.2193**	**10.2200**

[표 32] 실제 데이터에 대한 가장 좋은 예측자의 비율

자유도	df=1	df=2	df=3	df=4	df=5	df=6	df=7
GAM	0.000	0.000	0.000	0.000	0.000	0.000	0.000
배깅	0.358	0.384	0.404	0.368	0.354	0.374	0.390
랜덤포레스트	0.034	0.058	0.042	0.038	0.044	0.032	0.032
GAMbag1 (mean)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
GAMbag2 (trimed mean)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
GAMbag3 (median)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
GAMrnf1 (mean)	0.000	0.000	0.000	0.002	0.000	0.000	0.000
GAMrnf2 (trimed mean)	0.016	0.032	0.018	0.030	0.028	0.018	0.018
GAMrnf3 (median)	**0.592**	**0.526**	**0.536**	**0.562**	**0.574**	**0.576**	**0.560**

제 4 장 결론

본 논문에서는 일반화가법모형에 앙상블 기법을 적용하였다. 앙상블 기법은 다중의 예측자를 결합하여 예측의 성능을 높이는 기법으로 많이 이용되고 있다. 본 논문에서는 기존의 이항 반응변수의 예측에서 나아가 연속형 반응변수를 가지는 자료에서 예측의 성능을 향상시키는 앙상블 기법을 제안하였다. 또한, 모의실험과 실제자료를 이용하여 제안한 방법론이 기존의 방법론들보다 연속형 반응변수의 예측에 있어서 더 좋은 성능을 가지는 것을 확인하고자 하였다.

모의실험결과 일부 자료에서 제안한 방법론이 더 좋은 성능을 가지는 것을 확인할 수 있었다. 주요 결과는 다음과 같다.

첫째, 삼각함수 자료3과 같이 편차가 큰 자료에서 일반화가법모형을 이용한 앙상블 기법이 결과의 예측에 좋은 성능을 가지는 것을 확인 할 수 있었다.

둘째, 프리드먼 데이터1과 삼각함수 자료4와 같이 설명변수의 개수가 많은 자료에서 제안한 앙상블 기법이 더 좋은 성능을 가지는 것을 확인 할 수 있었다.

셋째, 실제자료에서도 기존의 방법론에 비해 제안한 방법론이 더 좋은 성능을 가짐을 알 수 있었다.

넷째, 삼각함수 자료1이나 다항함수 자료1, 다항함수 자료2와 같은 편차가 비교적 작고 설명변수가 적은 자료에서는 단일 일반화가법모형이 예측에 좋은 성능을 가짐을 알 수 있었다.

다섯째, 일반화가법모형에서 평활함수의 자유도가 낮아 선형 적합과 유사한 경우에는 제안한 앙상블 기법에 비해 배깅이나 랜덤포레스트가 더 좋은 성능을 가짐을 확인하였다.

추후 연구에서는 반응변수가 이산형이나 연속형이 아닌 계수형자료(count data)이거나 다변량자료(multivariate data)인 경우에 대해서도 앙상블 기법을 이용한 일반화가법모형을 살펴볼 수 있겠다.

참 고 문 헌

[1] Bauer, E., Kohavi, R., (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning.* 36, 105-139.

[2] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

[3] Breiman, L. (2001). Random forest. *Machine Learning*, 45, 5-32.

[4] Bühlmann, P. (2002). *Bagging, subagging and Bragging for improving some prediction algorithms.* In: Akritas, M.G., Politis, D.N. (Eds.), Recent Advances and Trends in NonParametric Statistics. Elsevier, Amsterdam.

[5] De Bock, K. W., Coussement, K., Van den Poel, D. (2010). Ensemble classification based on generalized additive models. *computational Statistics&Data Analysis.* 54(6), 1535 - 1546.

[6] De Bock. K. W., Van den Poel, D. (2012). Reconciling performance and interpretability in custor chun prediction using ensemble learning based on generalized additive models. *Expert Systems with Applications.* 39 (8), 6816-6826.

[7] Dobson, A. J., Barnett. A. G., (2008). *An Introduction to Generalized Linear Models.* Chapman&Hall/CRC, London.

[8] Efron. B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics.* 7(1), 1-26.

[9] Efron, B., Tibshirani, R. (1993). *An Introduction to the Bootstrap,* Chapman&Hall/CRC, London.

[10] Friedman, J. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*. 19(1), 1-67.

[11] Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models*. Chapman&Hall/CRC, London.

[12] Hastie, T., Tibshirani, R., Friedman, J. (2008). *The Elements of Statistical Learning*, 2nd Edition. Springer, NewYork.

[13] Keele, L. (2008). *Semiparametric Regression for the Social Sciences*. Wiley, NewYork.

[14] McCullagh, P., Nelder, J. (1989). *Generalized Linear Models*, 2nd Edition. Chapman&Hall/CRC, London.

[15] Skurichina, M., Duin, RPW. (2002). Bagging, Boosting and the Random Subspace Method for Linear Classifiers. *Pattern Analysis & Applications*. 5, 121-135.

[16] Wood, S. N. (2006). *Generalized Additive Models* : An Introduction with R. Chapman&Hall/CRC, London.

ABSTRACT

A Study on the Generalized Additive Models Using Ensemble Methods

HANA KIM

Department of Statistics

The Graduate School

Sungshin Women's University

The GLMs(Generalized Linear Models) have been one of the common ways for demonstrating the interactions of cause and effect between parameters. The detailed methodology of GLMs is effectively able to describe the effects of explanatory variables. GLMs cover almost situations for response variables that belongs to exponential family of distributions but it can not capture nonlinear effects. However, nonlinear effects can be verified by smooth function which is computed by local scoring algorithm in the aspect of GAMs(Generalized Additive Models).

Many studies have shown that ensemble method such as bagging and random forests is better than the single predictor by making an accurate estimate (Breiman, 1996).

This study provides the ensemble method related to the base predictor derived from GAMs in with the continuous response variable. Also, the superiority of ensemble method is illustrated with simulation study and real data of korea care insurance.