선형회귀분석, SVR, 시계열 모형을 통한 서울 어린이대공원 입장객 수 예측

Predicting the number of visitors at Children's grand park using linear regression, SVR and time series model

2010019913 박정훈
2012020623 이명준
2014018940 고형권
2014019425 최건호

2017년 11월 16일

목 차

표 목차

그림 목차

그림 목차

Ⅰ. 서론

1.1 주제

날로 빅데이터와 AI에 대한 관심이 증가하는 시대에 이들의 쓰임은 여러 곳에서 폭발적인 수요를 동반하며 증가하고 있다. 국내 메이저 통신사 중 하나인 KT에서 야간버스 노선을 만드는 데에 사람들의 통화 데이터를 분석해 사용했다는 것은 이미 너무나 유명한 예이다. 본 팀은 주제를 정하는 가운데 여러 가지 아이디어를 떠올렸고, 다음과 같다.

1) 서울 롯데월드 입장객 수 예측
2) 스키장 일별 입장객 수 예측
3) 한양대 사랑방 학교식당 시간별 수요 예측
4) 서울시 버스 필요/불필요 노선 파악 및 개선 방향 제시
5) 공공 와이파이 부족 지역 파악

그러나 각각의 경우마다 필수적 데이터의 결여라는 문제가 발생했고, 제시된 아이디어는 현실적인 탐구주제로는 적합하지 않았다. 다른 적합한 주제를 찾던 중 서울 어린이 대공원의 입장객 수가 공공데이터 포털에 올라와 있었고, 이를 예측 하는 모델을 만드는 것으로 주제를 잠정적으로 정하였다. 처음 어린이 대공원의 입장객수 분석 모델을 만드는 것에 약간 회의적이었던 이유는 이용객 수가 적을 것으로 생각해 분석 모델의 설명력이 떨어지고 모델의 효용성이 낮을 것이라는 의견이 컸기 때문이었다. 그러나 실제 데이터를 본 결과, 서울 어린이 대공원은 일평균 28600명에 달하였으며 또한 2006년 10월 무료 개방 이후 점점 더 많은 사람들이 찾고 있는 서울의 가장 큰 공원 중 하나로서 일별 예측 모델을 만드는 것이 적절하다고 판단되어 최종적으로 이 주제를 선택하게 되었다.

1.2 분석 모델

본 팀은 첫 번째 방법으로 선형회귀분석을 예측 모델로 사용하였는데, 그 이유는 통계분석 프로그램 사용 전에 충분한 사전 지식이 필요하다고 생각했기 때문이었다. 또한 팀원들 구성이 2학년 1명, 3학년 2명, 4학년 1명의 학부생으로 구성되어 기초적인 통계지식만 가지고 있었으므로, 회귀분석을 사용하는 것은 하나의 도전이었다. 또한 회귀분석은 로지스틱 회귀분석, 인공 신경망, 딥러닝 등 수많은 예측 모델의 기본으로서 회귀분석을 제대로 알고 있는 것이 미래에 더 깊은 학문을 공부하는데 도움을 줄 것이라 생각하였다. 후에 모델의 예측력을 높이기 위해 인공신경망을 사용하였으나 과적합의 문제로 인해 SVR을 사용하여 예측을 시도 해보았고 또한 일별 분석의 한계를 보완하고자 시계열을 이용해 월별 분석을 시행해 보았다.

1.3 글의 구성

본 레포트는 현재 서술된 1장을 포함하여 4개의 장으로 구성된다. 2장에서는 모델 형성 및 분석에 사용된 회귀분석 기법과 변수 변환 및 선택법, SVR, 시계열의 이론에 대하여 다루었고, 3장에서는 만든 예측 모델들에 대해 설명하고, 예측을 실시하여 모델들 간의 성능을 비교하였다. 4장에서는 결론과 함께 부족하고 더 연구가 필요한 부분에 대해 논의하였다.

Ⅱ. 본론

2.1 회귀분석

독립변수와 종속변수 사이의 함수관계를 추구하는 통계적 방법을 회귀분석이라고 부른다. 일반적으로 식은 $= \beta + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + ... + \widehat{\beta_p} x_p + \epsilon$와 같이 표기하며, 이 때 ϵ은 평균이 0, 표준편차가 σ인 정규분포를 따르며 상호 독립적인 오차항을 의미한다. ($\epsilon \sim \quad (0, \sigma^2)$)

$X = \begin{bmatrix} x_{11} \dots x_{1p} \\ \dots \\ x_{n1} \dots x_{np} \end{bmatrix}, Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \dots \\ \beta_p \end{bmatrix}$로 표기하고 전체 식은 $Y = X\beta + \epsilon$와 같이 쓴다.

$\hat{\beta} = (X'X)^{-1} X'y$으로 $\hat{\beta}$를 추정하고, $Y = X\hat{\beta}$으로 $\hat{Y}$를 구한다.

총 제곱 합(Sum of squares of total): $SST = \quad Y - \overline{Y} \|^2$

회귀 제곱 합(Sum of squares of regression): $SSE = \| Y - \hat{Y} \|^2$

오차 제곱 합(Sum of squares error): $SSR = \| \hat{Y} - \overline{Y} \|^2$

회귀 모델을 구한 다음에는 식이 적합한지 여러 가지 통계량을 고려해서 선택한다.

1) 결정계수 $R^2 = \frac{SSR}{SST}$: 결정계수는 총 변동 중 회귀선에 의해서 설명이 되는 변동의 비율로 구하는 것으로 0과 1사이의 값이며 1에 가까울수록 유용성이 높다고 애기할 수 있다.

2) 수정 결정계수 $R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{n-p-1}$: 결정계수는 변수의 개수가 늘어날수록 값이 증가하므로 자유도를 고려한 수정 결정계수를 사용하여 보정을 한다.

3) $AIC(Akaike\ Information\ Criterion) = 2k - 2\ln(L)$, k는 모델 내의 number of estimated parameter. L은 예측 모델 로그우도 함수의 극대 값이다. AIC값은 낮을수록 좋다.

4) $BIC(Bayesian\ Information\ Criterion) = -2^*\ln(L) + kln(n)$, n은 관측 데이터 수(sample size), k는 free parameter의 개수, L은 예측 모델 로그우도 함수의 극대 값이다. BIC값은 낮을수록 좋다.

5) $MAE(Mean\ Absolute\ Error) = \frac{1}{n} \sum_{i=1}^{n} |Y - \hat{Y}|$, 이 때 Y는 실제 값, $\hat{Y}$는 예측 값이다.

6) $RMSE(Root\ Mean\ Square\ Error) = \quad \frac{1}{n} \sum_{i=1}^{n} (Y - \hat{Y})^2$, MAE와 RMSE는 낮을수록 좋다.

2.2 변수 변환

변수 변환은 잔차분석을 통해 진행한다. 잔차의 산점도는 잔차를 축으로 놓고 Y축으로는

1) Y에 대하여

2) X_i $i=1,2,...,k$에 대하여

3) Y_i가 시계열인 경우에 시간에 대하여

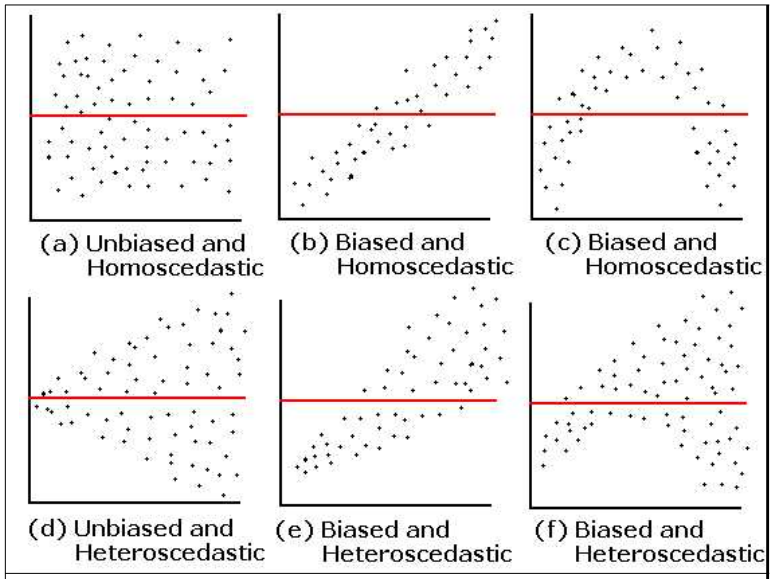

그림1. 잔차 예시

(a): 등분산 가정을 만족하므로 변수 변환의 필요가 없다.
(b): 절편이 필요한데 사용되지 않은 경우이다.
(c): 독립변수의 제곱항 등이 필요한 경우이다.
(d): 가중회귀를 쓰거나 Y_i를 변환하여 써서 회귀 분석함이 바람직하다.

2.3 변수 선택법

1) 전문가적 선택법(작위적 선택): 이 방법은 전문가가 자신의 과거 경험과 해당 문제에 대한 지식을 바탕으로 변수를 선택하는 방법을 뜻한다. 구한 모델이 주관적일 수 있으며 또한 최적 모델이라는 보장을 못한다는 단점이 있다.

2) Stepwise Method(휴리스틱 알고리즘): 회귀분석의 경우에 Stepwise방법을 사용하면 굉장히 빠른 시간내에 모델을 만들어 낼 수 있다. Stepwise의 경우, Forward 방법을 개선한 것으로 순서는 다음과 같다.

(1) 모든 독립변수 중에서 종속변수 Y와 가장 상관관계가 높은 변수를 선택한다. R^2를 가장 크게 하여 주는 독립변수를 고를 수도 있다.

(2) $Y=f(X_p,X_i)$, i p인 두 번째 변수(X_i)를 적합시키고 F검정으로 이 변수의 추가선택이 유의한지 검정한다.

(3) X_p,X_q가 모두 유의하여 남아 있는 경우에 다음으로 들어올 변수를 선택해 주기

위해 $f(X, X_q, X_i)$, $i \neq p, q$를 각각 적합시키고, R^2를 가장 크게 하는 변수 X_i를 구하고 이를 X_r이라고 한다. X_r이라 하고, F검정을 하여 유의하지 않으면 유의하면 계속 남겨 놓고 유의하지 않으면 변수 선택을 중단하고 X_p, X_q만 뽑아 준다. X_r이 유의한 경우, X_p, X_q에 대한 F검정을 하여 유의하지 않은 변수가 있으면 제거시키고 다음 순서로 넘어간다.

위와 같은 절차를 계속 밟아가며 새로이 선택된 변수가 유의하지 않을 때까지 선택 절차가 계속된다.

3) 유전 알고리즘: 자연계의 진화 원리를 모방해 만들어낸 최적해 탐색기법이다. 임의로 만들어진 초기 개체는 집단의 구성원으로서 더 적합한 다음 세대를 만들어 내며 진화하고, 개체 내 차이가 일정 수준 이하 일 때 진화를 멈추어 최적의해를 찾아낸다.

<유전 알고리즘을 통한 변수 선택의 순서>

(1) 모집단 생성: 크게 두 가지 방법이 있는데, 무작위 초기화 혹은 하나의 사전지식이나 경험을 기반으로 하는 유도된 초기화법을 사용할 수 있다.

(2) 다음 세대로의 전달: 크게 세 가지 방법에 의해 이루어진다. 또한 변이의 비율에 따라 모든 과정에서 영향을 받는다.

1) 무성생식(asexual reproduction): 부모 세대에서 확률적으로 뽑아져 그대로 내려온다. 다만 mutation rate에 의해서 고유 chromosome간에 자리(locus)가 조금씩 바뀔 수 있다. 우수한 개체가 다음 세대로 이어지도록 하는 역할을 한다.

2) 유성생식(sexual reproduction): 부모 간 결합(crossover)으로 생성된다.

3) 이민(immigration): 무작위적인 배열로 이뤄짐. 끊임없이 다양한 해집합을 통한 전체적인 관점의 접근으로 local optima에서 벗어나 global optima를 찾을 수 있게 도와준다.

+) 변이(mutation): chromosome내 자리 변화. 지역 최적해(local optimal)나 사점(dead corner)에 빠지는 것을 벗어나게 하는 메커니즘을 가지고 있다.

(3) 정지조건: 두 가지 방법이 주로 사용된다. 하나는 현재의 세대 또는 반복횟수가 미리 지정한 값에 도달한 경우 멈추는 것이고, 다른 하나는 수렴된 유전자의 수(비율 등)를 조사해 수렴성을 판단하는 것이다.

2.4 SVR

SVR이란 SVM(Support Vector Machine)의 원리를 회귀문제에 이용한 머신 러닝의 한 기법이다. SVR에서는 비선형 회귀 문제를 해결하기 위하여 먼저 입력 값이 고차원의 형상공간에 사상되고 그 후 결과 값과 연관된 함수를 찾는다. SVR의 선형 추정 함수는 $(x) = w\,x + b$로 표현 가능하며, ϵ-무감도 손실함수 $_e$는 일반적으로 SVR에 사용되는 비용함수는 다음과 같다.

$$L_e(F(x), q) = \max(0, |f(x) - q| - \epsilon) = \max(0, |w^r x + b - q| - \epsilon)$$

이 때 ϵ은 튜브의 반지름을 나타내는 정밀모수로서 사용자의 지정에 따라 조정가능하다.

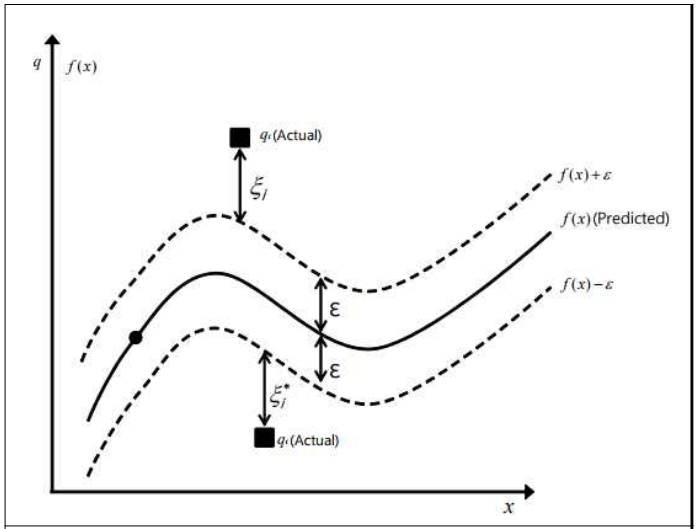

그림2. The Notation of ϵ-Insensitive Loss Function in SVR(Lu et al., 2009)

따라서 SVR은 실제 값 q_i와 예측 값 $f(x) = w^r x + b$의 차이를 가능한 한 ϵ이내로 유지하면서 마진을 최대화하여 구하게 된다. 이 때 실제 문제에서 과적합의 문제를 방지하기 위하여 소프트 마진을 사용하는데 그를 위해 사용하는 여유 변수 c를 이용하면 다음과 같은 문제로 변환이 가능하다.

$$Min : R_{reg}(f) = \frac{\|w\|^2}{2} + c \sum_{i=1}^{n} (\zeta_i + \zeta_j^{\phi})$$

$$s.t. \begin{cases} q_i - w^r x - b \le \epsilon + \zeta_i \\ w^r x + b - q_i \le \epsilon + \zeta_j^{\phi} \\ \zeta_i, \zeta_j^{\phi} \ge 0 \end{cases}$$

c값 역시 사용자의 지정에 따라 마진의 정도(hardness, softness)를 결정 가능하며 여러 값을 대응시켜본 후 최적의 c를 선택한다. Lagrangian 승수와 KKT 조건을 위 식에 적용한 SVR 기반 회귀 함수의 일반적 형태는 다음과 같다.

$$f(x, v) = f(x, \alpha, \alpha^*) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x_j) + b$$

x, x_j)는 커널 함수이고 이는 $K(x_i, x_j) = \phi(x_i)\phi(x_j)$로 표현 가능하다.

대표적인 커널에는

1) 다항 커널: $K(x_i, x_j) = (x_i \cdot x_j)^d$

2) RBF 커널: $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, 이 때 γ는 $\frac{1}{\sigma^2}$로도 표현 가능하다.

3) 시그모이드 커널: $K(x_i, x_j) = \tanh(\kappa x_i \cdot x_j + c)$ 등이 있다.

2.5 시계열

기존의 방법은 인과관계를 기반으로 만들어진 모델로 예측을 진행한다. 시계열 분석은 과거의 자료에 기반으로 모델을 생성한다. 시계열 자료는 일반적으로 동일한 시간간격을 이루어진 벡터로서, 그 특징은 시계열 자료들이 독립이 아니라 상호 연관되어 있다는 점이다. 시계열 자료를 대상으로 하는 시계열 분석은 시계열 자료의 구조를 파악하고, 분석 대상인 시계열 자료의 구조와 특성을 토대로 미래의 값을 예측하고, 생성된 시스템을 제어하는 데 그 목적이 있다.

시계열 분석의 주요한 개념은 다음 3가지로 구분된다.

1) 경향성(Trend): 대부분 시계열 자료는 Trend가 없는 독립적인 것으로 가정한다. 시계열 자료가 일정하게 상승, 또는 하강하는 Trend를 가지면, 분석하기 전에 이를 제거하도록 한다.

2) 연속 의존성(Serial Dependence): 대부분 시계열 자료는 인접한 자료들 사이에 상관성이 없는 것으로 가정한다. 시계열 자료가 인접한 자료들 사이에 상관관계가 있으면, 분석하기 전에 이를 통제한다.

3) 정상성(Stationary): 일반적으로 시계열은 어디에서 살펴보든지 같은 성질을 가지는 것으로 가정한다. 정상성은 다음과 같은 조건을 만족하는 시계열을 의미한다.

- 시계열 자료의 모든 시간 에 대하여, 평균이 일정하다.
- 시계열 자료의 모든 시간 t에 대하여, 분산이 일정하다.
- 시계열 자료 $x_{\ 1}$, x_{t2}의 자기상관함수 (CF: Autocovariance Function) 및 편자기상관함수(PACF : Patial ACF)는 시간 $t1, t2$에만 의존한다.

시계열 자료의 정상성 가정을 검증하기 위한 지표는 자기상관함수와 편자기상관함수 등 2개의 지표가 적용된다.

2.6 AR, MA, ARIMA, seasonal ARIMA

1) 자기회귀 모형(AR Model : Auto Regressive Model)

자기회귀 모형은 현재의 관측자료 x_t가 현재의 관측자료를 설명하여 주는 과거 자료(x_{t-1}, x_{t-2},)들과 설명하여 주지 못하는 부분(α_t)의 선형 결합으로 표시되는 모형이다. 현재자료를 설명 못하는 (α_t)는 White Noise로서, $\mu = 0$,

$Var = \sigma^2$, $Cov(y_{t,} y_{t+\gamma}) = \gamma(\tau) = \sigma_Z^2(when, \tau = 0)$, $Cov(y_t, y_{t+Cov(y+y_{t+\gamma})=0,(\gamma \quad 0)})$을 나타낸다. 자기회귀 모형은 시계열의 관측자료와 평균과의 편차($y_t = x_t - \mu$)가 다음과 같이

시차변수(Lagged Variables)를 회귀분석과 같은 형태의 선형 결합으로 구성될 수 있다는 가정에 기반을 두는 모형이다.

$$R\ p): y\ = a_1 y_{t-1} + a_2 y_{t-2} + \quad + a_p t_p + \alpha_t\ = \sum_{i=1}^{p} a_{iy_{-i}} + \alpha_t$$

2) 이동평균 모형(MA Model : Moving Average Model)

이동평균 모형은 현재와 과거의 설명하여 주지 못하는 부분(α_t)의 선형결합으로 표시되는 모형이다. 모형에서 설명 못하는 (α_t)는 백색잡음으로 $\mu = 0$, $Var = \sigma^2$, $Cov(y_t, y_{t+\gamma}) = \gamma(\tau) = \sigma_Z^2(when, \tau = 0)$, $Cov(y_t, y_{t + Cov(y_t + y_{t+\gamma}) = 0, (\gamma \quad 0)})$을 나타낸다. 이동평균 모형은 자기회귀모형과 유사하게 시차변수를 선형결합 하는 형태로 구성되며, 일반적으로 q차의 이동평균 모형의 수식은 다음과 같이 표현된다.

$$MA(q)\ : y_t = \alpha_t - \theta_1 \alpha_{t-1} - \quad - \theta_q \alpha_{t-q}$$

수식에서 α_i는 평균 0, 분산σ^2의 백색잡음이며, θ_i는 이동평균 계수이다. 그리고 유한한 시차변수를 가진 이동평균 모형은 정상성을 가진다.

3) 자기회귀 누적 이동평균 모형(ARIMA :Auto Regressive Integrated Moving Average Model)

현실적으로 당면하는 시계열 자료는 평균, 분산, 자기공분산 등이 시간의 변화에 따라 일정하지 않는 비정상 시계열이 대부분이다. 비정상 시계열 자료를 정상 시계열로 변환하기 위한 방법은 두 가지로 구분된다. 첫 번째 변환 방법은 동질적인 비정상 시계열에 대하여 차분을 통한 정상 시계열 변환방법이다.

1차 차분 : $\triangle y_t = y_t - y_{t-1}$, d차 차분 : $\triangle^d y_t = \triangle y_t - \triangle y_{t-d}$

두 번째 변환방법은 차분을 수행하여도 분산의 동질성이 없이 시간의 변화에 따라 정상성이 없는 시계열로서, 관측자료에 대하여 log변환, Root변환, Box-Cox 변환 등을 수행하는 경우이다. 비정상 시계열 자료를 d차 차분하여, 정상성을 갖는 새로운 시계열 자료인 $\triangle^d y_t = \triangle y_t - \triangle y_{t-d}$ 는 정상 확률 과정으로 표현되는데, 이는 ARMA모형과 차분연산을 통합한 다음의 식으로 표현된다.

$$ARIMA(p,d,q)\ : a_p(\triangle)(1-\triangle)^d y_t\ = \theta_0 + \theta_q(\triangle) f_t$$

수식에서 f_t는 평균 0, 분산 σ_f^2인 White Noise이며, θ_0눈 추세모수로서 0이다.

Ⅲ. 예측 모델 및 결과 분석

3.1 데이터 전처리

모델을 만들면서 고려한 예측 변수와 설명변수는 다음과 같다.

: 과거 입장객수 (일)

X_1: 이벤트(축제) (일)

X_2: 나들이 지수 (일) (만든 값)

X_3: 불쾌지수(일)

X_4: 구글검색수 (일) (상대 값)

X_5: 통합대기지수 (일)

X_6: 휴일 (일)(4 factor)

X_7: 계절 (일)(4 factor)

X_8: 구제역 (일)

X_9: AI(일)

X_{10}: 메르스-기사수 (일)

X_{11}: 세월호-기사수 (일)

X_{12}: 금값 (일)

X_{13}: 코스피 (일)

X_{14}: WTI유가 (일)

X_{15}: 원달러환율 (일)

X_{16}: 달러인덱스 (일)

X_{17}: 생활물가지수 (월별)

X_{18}: 경제심리지수 (월별)

X_{19}: 실업률 (월별)

X_{20}: 인구수 (월별)

X_{21}: 인근 학교의 소풍 (월별)

X_{22}: 유동인구수 (일) (제외)

X_{23}: 지하철 이용객수- 아차산,세종대(일) (제외)

X_{24}: 에버랜드 입장객수 (월별) (제외)

1) X_1: 이벤트 데이터(4 factors)

③: 대규모(축제, 기간이 길다, 참여행사&음악행사가 동시에, 대규모 외부행사(연예인, 게임대회))

②: 중규모(보고 듣는 큰 행사)

①: 소규모(보고 듣는 작은 행사(공연), 어린이 참여 행사, 기타 행사, 대부분의 전연령)

⓪: 무규모(시설 관리 공단 내부행사, 위촉식, 행사 없음)

2) X_2 X_3: 나들이 지수 및 불쾌지수: 불쾌지수는 알려져 있는 공식을 사용하였으며, 나들이 지수의 경우 사기업인 케이웨더에서 제공하고 있지만 기업 기밀로 공식이 알려져 있지는

않다. 따라서 자의적으로 만든 공식을 이용하였다. 공식은 다음과 같다.

불쾌지수: $\frac{9}{5}T-0.55(1-\frac{H}{100})(\frac{9}{5}T-26)+32$

나들이 지수: $K_2=\frac{100}{2(T-14.5)^2+50R*P*S+H}$, T=온도, R=강수여부, P= 수량, S=하늘상태, H=습도

3) X_4: 구글검색수: 네이버와 구글에서 검색어에 따른 기간별 빈도수를 제공하고 있지만, 네이버의 경우 2016년부터 서비스를 제공하기 때문에 구글의 검색빈도 데이터를 사용하였다. 데이터는 0부터 100까지 값의 상대도수를 제공하였다.

4) X_5: 통합대기지수: 6개의 미세먼지(PM10, PM2.5, SO2, CO, NO2, O3)에 대한 데이터와 기상청에서 제공하는 통합대기지수 공식을 사용하여 통합대기지수를 만들어 사용하였다.

5) X_6, X_7: 휴일, 계절: 휴일의 경우 하루만 쉬는 공휴일 등은 특별한 경우라고 생각하여 A, 일반적인 주말(토, 일)의 경우 B, 공휴일 등이 추가되어 주말 이틀과 더불어 3일을 쉬는 경우 주말은 그대로 B, 추가적 휴일은 A로 표기하였고 마지막으로 4일 이상의 휴일인 경우 주말 여부와 상관없이 C로 표기하였다. 계절 데이터의 경우 3,4,5/6,7,8/9,10,11/12,1,2로 4등분하여 봄, 여름, 가을, 겨울로 나누어 표기하였다.

6) X_8, X_9, X_{10}, X_{11}: 구제역, AI, 메르스, 세월호: 어린이 대공원에는 동물원이 있으며, 이는 서울에서는 거의 유일하다고 얘기할 수 있다. 또한 이는 과천 서울대공원이나 에버랜드의 동물원과 달리 공짜이기 때문에 사람들은 동물원 때문에도 어린이 대공원을 찾을 것이라고 예상하였다. 구제역과 AI의 경우는 수시로 발생하며, 정도에 따라 등급을 4단계로 나누는데 평시(관심, 1단계), 의사환축 발생시(주의, 2단계), 인접 또는 타지역 전파(경계, 3단계), 여러 지역 발생 및 전국 확산 우려(심각, 4단계)이며, 이에 따라서 일별 등급을 매겼다. 메르스와 세월호의 경우는 약간 다르게 네이버 뉴스의 관련 일별 기사수를 체크하여 데이터로 사용하였다.

7) $X_{12}, X_{13}, X_{14}, X_{15}, X_{16}$: 금값, 코스피지수, 유가, 원달러환율, 달러인덱스: 장이 열리는 날과 열리지 않는 날이 있어 열리지 않는 날의 데이터는 전날의 마감 가격으로 집어넣어 전처리하였다.

8) $X_{17}, X_{18}, X_{19}, X_{20}$: 생활물가지수, 경제심리지수, 실업률, 인구수: 이 데이터들은 월별 데이터 밖에 얻을 수 없어서 부득이하게 월별로 다 같은 값을 집어넣어 사용하게 되었다.

9) X_{21}: 인근 학교의 소풍: 광진구에 있는 유치원, 초등학교를 대상으로 조사하였으며, 정확한 데이터를 구할 수 없어 일일이 전화를 걸어 조사하였다. 유치원 등은 가는 빈도가 매우 높고 정확한 일자를 기록해놓지 않은 곳이 많았으며, 초등학교 역시 비슷한 문제로 인해 인원 및 일자에 대한 정확한 데이터를 얻을 수 없었다. 따라서 소풍을 가는 월을 조사하였고, 초등학교는 학년별로 갈 것으로 가정하여 1/6을 곱해주어 유치원 인원과 합쳐 월별로 평일에 같은 값을 넣는 것으로 대체하였다.

10) $_3$: 지하철 이용객수: 공공데이터포탈에서 제공하는 지하철 이용객수 데이터를 사용하였으며, 어린이대공원에 인접해있는 역 두 곳(5호선 아차산역, 7호선 어린이대공원역)의 승차 하차수의 합을 변수로 사용하려 하였으나, 데이터를 2015년 이후로 밖에 구할 수 없어 할 수 없이 제외시켰다.

11) X_{22} X_{24}: 유동인구 데이터, 에버랜드 입장객수: 유동인구 데이터의 경우 시간별 해당 지점의 유동인구의 주별 평균값이 기록된 데이터를 구하여 사용하려고 하였으나, 해당 년의 주별 평균이었기 때문에 유의하지 않을 것이라 생각하여 배제하였다. 에버랜드 입장객수의 경우 여러 정황을 고려할 때 어린이대공원 입장객수와 비례할 것이라 생각되어 사용하려 하였으나, 월별 데이터 밖에 구할 수 없었으며, 다른 데이터와 충돌하여 배제하였다.

이와 같이 전처리된 설명변수 $X_1, X_2 \ldots, X_{21}$등 21개의 변수를 모델을 만드는데 사용하였다.

3.2 변수 선택법에 따른 모델 생성

1) 전문가적 방법에 의해 결정된 모델: 이 방법에서는 처음에 전체 변수를 대입하여 모델을 만들고 변수마다 잔차를 비교분석하여 변환하며 모델에 포함 / 불포함을 반복 시행하여, Validation dataset에 대하여 최고의 예측률을 보이는 변수들을 선정하였다.

변수 변환

① 전체 잔차 그래프: 깔때기 모양을 보이고 있으므로 Y값에 로그를 취해 오른쪽과 같이 변환시켰다.

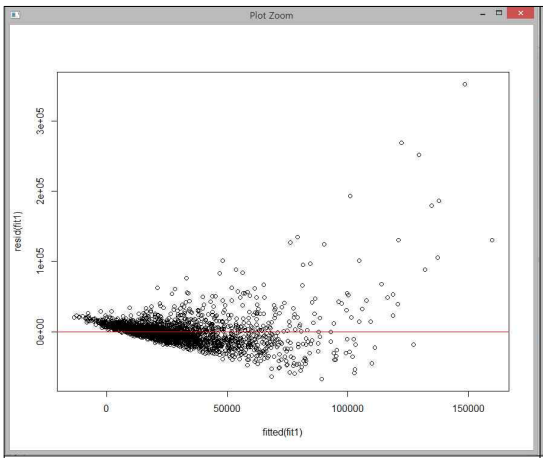

그림3. 전체 잔차 변환 전

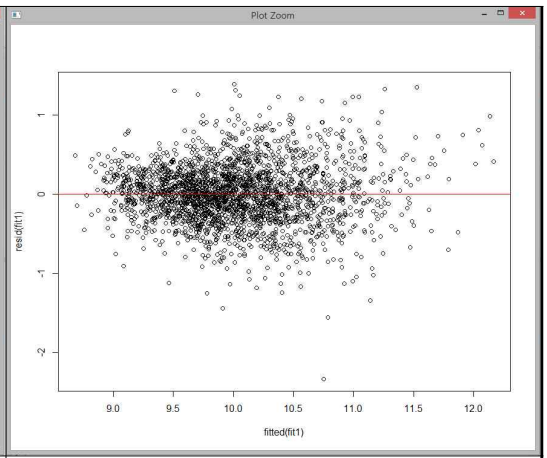

그림4. 로그를 취한 후

② 통합대기지수: 역 깔때기 모양으로 값이 증가할수록 폭이 줄어들어 역수 변환을 하였고 오른쪽 그림과 같이 잔차 그래프가 변하였다.

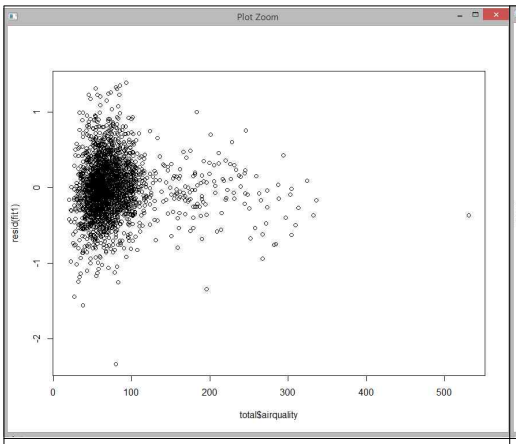

그림5. 통합대기지수 변환 전

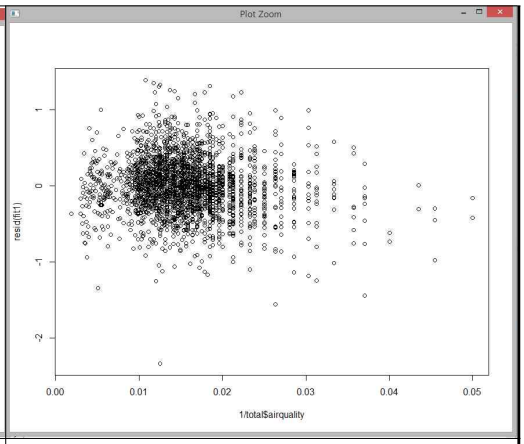

그림6. 역수를 취한 후

③ X_2 나들이 지수: 역 깔때기 모양이고, 로그를 취해 잔차에 이상이 없도록 조정하였다.

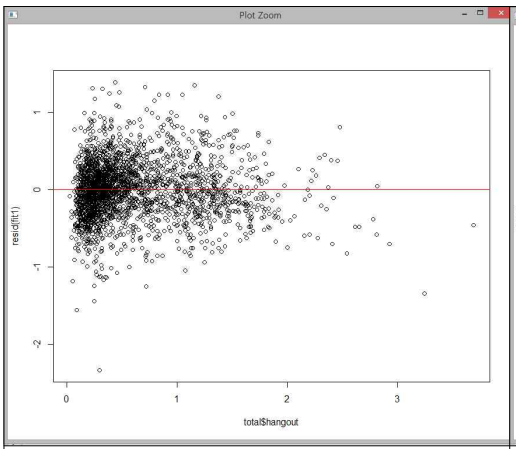

그림7. 나들이 지수 변환 전

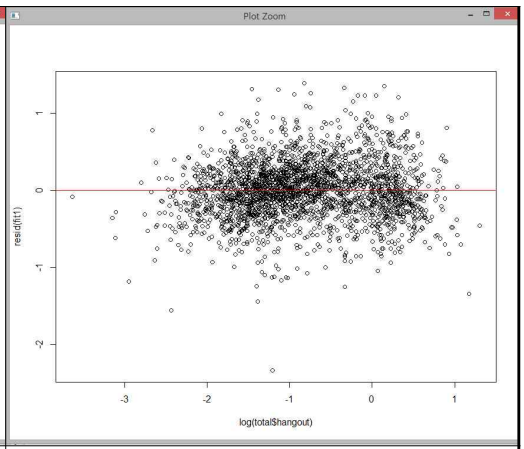

그림8. 로그를 취한 후

④ X_{10} 메르스: 역시 역 깔때기 모양으로 로그를 취하여 잔차를 조정하였다.

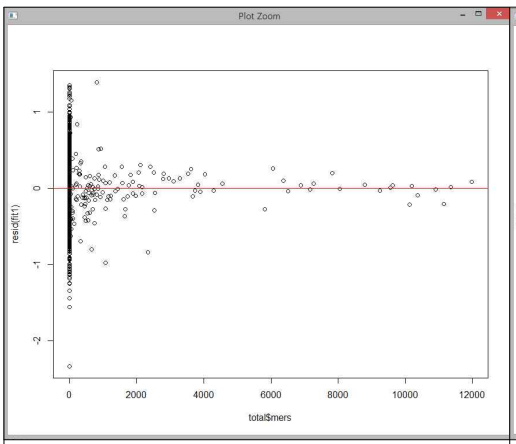

그림9. 메르스 데이터 변환 전

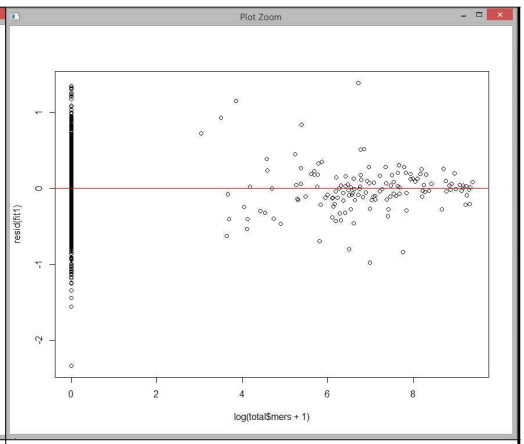

그림10. 로그를 취한 후

⑤ $_1$ 세월호: 루트를 씌워 사용하였다.

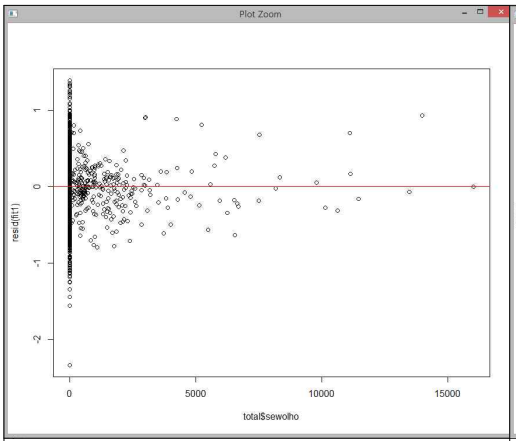

그림11. 세월호 데이터 변환 전

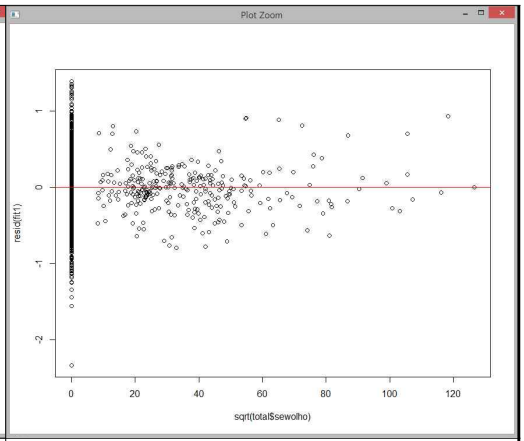

그림12. 루트를 취한 후

나머지 데이터는 잔차에 이상이 없다고 판단하여 그대로 사용하였다.
최종적으로 선택된 모델은 다음과 같다.

$$MODEL \ : \log(Y) = 5.953e^{-1} + 1.033e^{-1} * X_1^1 + 3.947e^{-2} * X_1^2 + 2.751e^{-1} * X_1^3 + 2.673e^{-1} * \log(X_2) + 1.721e^{-2} * X_3 + 9.719e^{-1} * X_6 + 6.189e^{-1} * X_6^B + 5.242e^{-1} * X_6^C + 2.886e^{-1} X_3 * X_6^A + 1.740e^{-1} X_3 * X_6^B + 2.642e^{-1} X_3 * X_6^C + 7.296e^{-3} X_4 + 2.033e^{-1} X_7^{Spring} - 1.622e^{-1} X_7^{Summer} + 1.652e^{-1} X_7^{Winter} - 4.590e^{-4} X_{13} + 1.591e^{-3} X_{14} + 2.915e^{-6} X_{20} + \epsilon$$

2) Stepwise에 의해 결정된 모델
최초 선택된 모델의 변수는 다음과 같다.

$$Y \sim X_4 + X_2 + X_6 + X_{20} + X_1 + X_3 + X_7 + X_{21} + X_8 + X_9 + X_{14} + X_{12} + X_{19} + X_5 + X_{15} + X_{11} + X_{13}$$

잔차변환
① 전체 잔차: 로그변환 하였다.

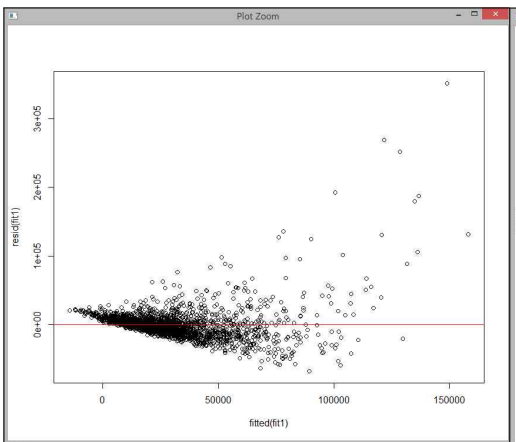

그림13. 세월호 데이터 변환 전

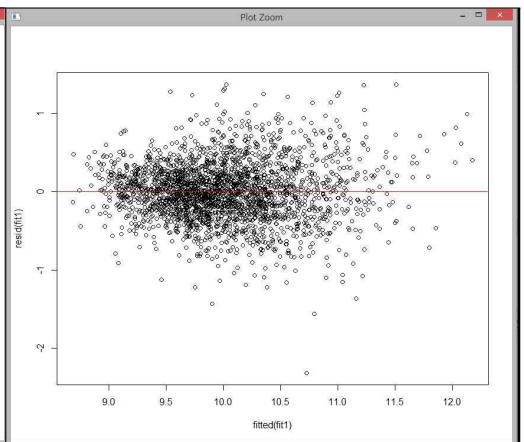

그림14. 로그를 취한 후

② 통합대기지수: 역수 취하였다.

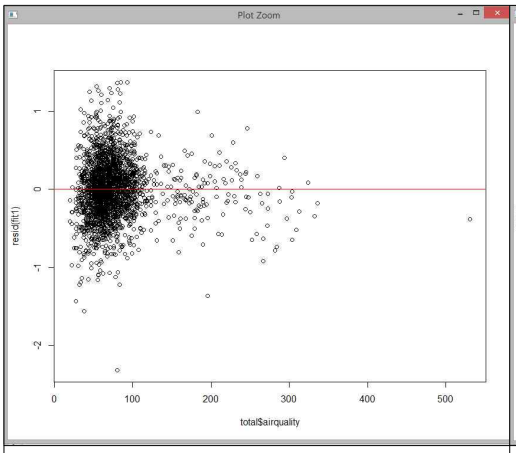

그림15. 세월호 데이터 변환 전

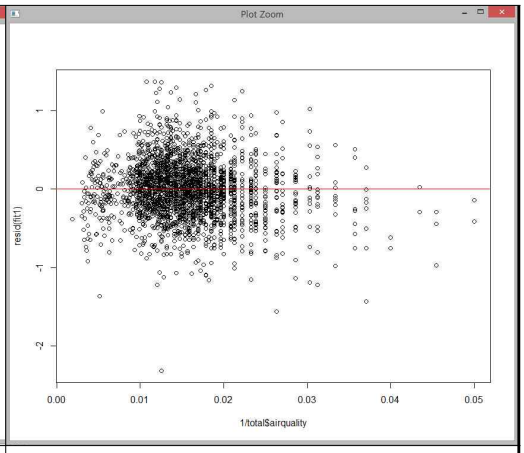

그림16. 역수를 취한 후

③ X_2 나들이 지수: 로그 취하였다.

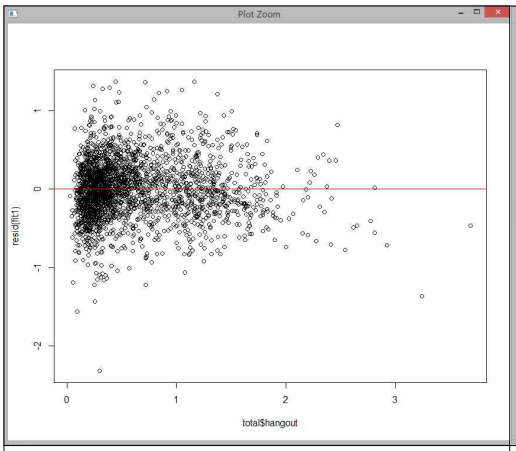

그림17. 나들이 데이터 변환 전

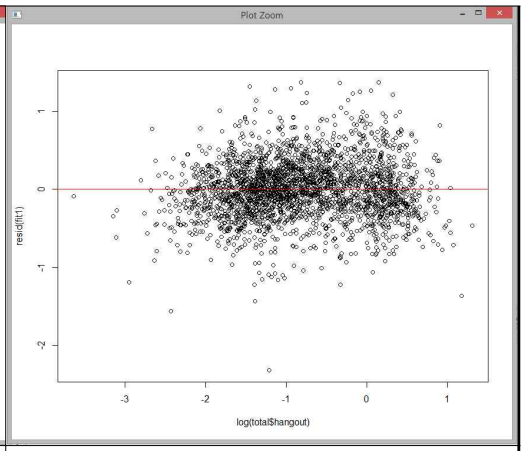

그림18. 로그를 취한 후

④ X_{11} 세월호: 루트 취하였다.

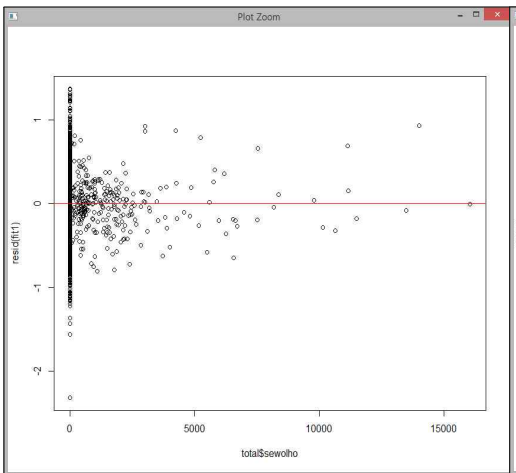

그림19. 세월호 데이터 변환 전

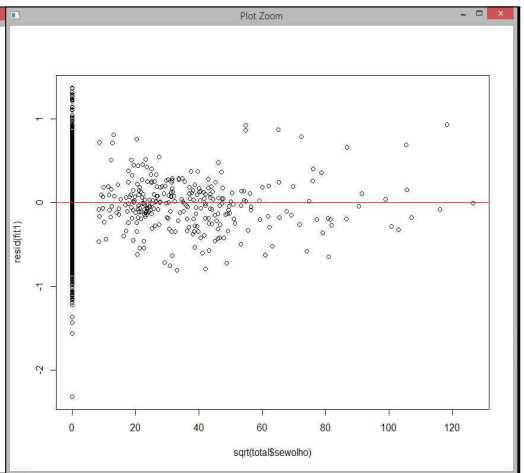

그림20. 로그를 취한 후

최종적으로 선택된 모델은 다음과 같다.

$$ODEL \ : \log(Y) = -3.397e*1 - 7.737e^{-3} * X_4 + 3.267e^{-1} * \log(X_2) + 7.236e^{-1} * X_6 + 4.428e^{-1} * X_6^B + 2.659e^{-1} * X_6^C + 5.338e^{-6} * X_{20} + 2.659e^{-1} * X_1^1 + 6.599e^{-2} * X_1^2 + 3.091e^{-1} * X_1^3 + 1.932e^{-2} * X_3 + 3.474e^{-2} * X_7^{Spring} - 1.153e^{-1} * X_7^{Summer} + 2.412e^{-1} * X_7^{Winter} - 1.348e^{-5} * X_{21} + -1.360e^{-1} * X_8^1 - 2.985e^{-1} * X_8^2 - 1.645e^{-1} * X_8^3 + 7.618e^{-2} * X_9^1 + 3.571e^{-2} * X_9^2 + 3.829e^{-1} * X_9^3 - 6.419e^{-3} * X_{14} + 3.614e^{-4} * X_{12} - 4.909e^{-2} * X_{19} - 5.380e^{-0} * 1/X_5 - 2.687e^{-3} * X_{15} - 3.222e^{-3} * \quad X_{11} - 5.687e^{-4} * X_{13} + \epsilon$$

3) Genetic Algorithm에 의해 결정된 모델

유전알고리즘에 의해 최초 선택된 모델은 다음과 같다.

$$Y \sim X_6 + X_7 + X_1 + X_8 + X_9 + X_3 + X_2 + X_4 + X_{20} + X_{21} + X_{14} + X_{12} + X_{19}$$

이후 변수 변환을 시행하였다.

변수변환

① 전체잔차: 로그 취하였다.

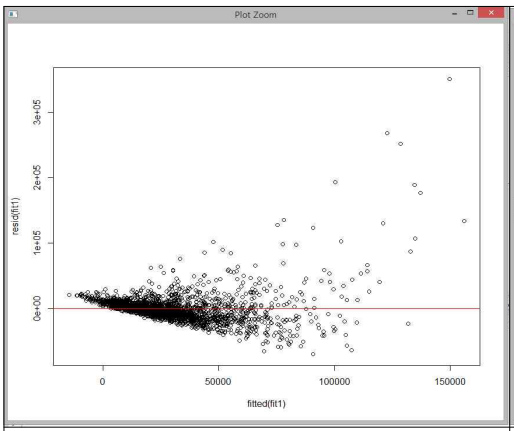

그림21. 전체 잔차 변환 전

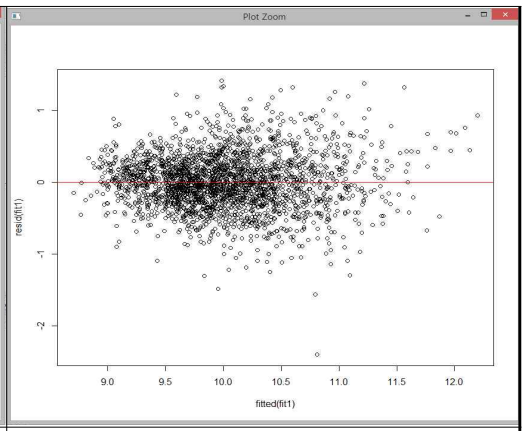

그림22. 로그를 취한 후

② X2, 나들이 지수: 로그를 취하였다

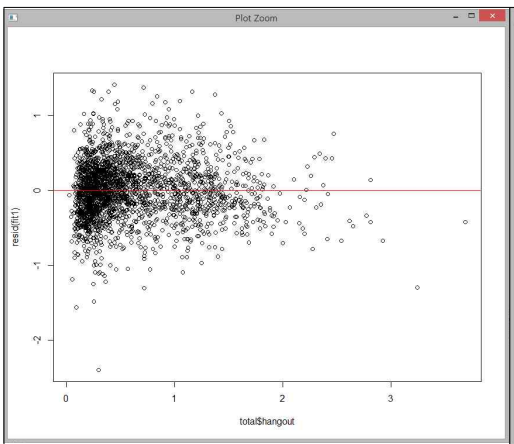

그림23. 나들이 지수 변환 전

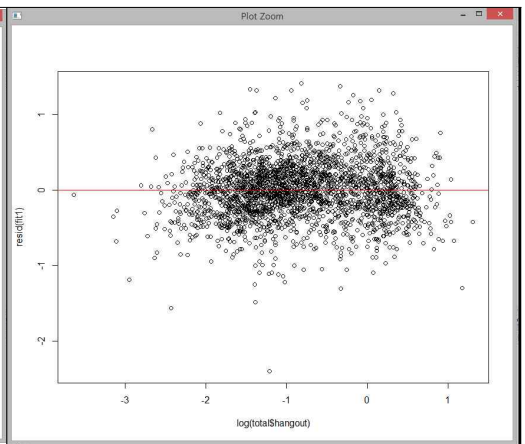

그림24. 로그를 취한 후

변수변환을 한 후 최종 선택된 모델은 다음과 같다.

$$ODEL \ : \log(Y) = \ -6.980e\ *1 + 7.032e^{-1} * X_6 \ + 4.366e^{-1} * X_6^B + 2.850e^{-1} * X_6^C +$$
$$2.899e^{-1} * X_7^{Spring} - 1.477e^{-1} * X_7^{Summer} + 2.316e^{-1} * X_7^{Winter} + 1.071e^{-1} * X_1^1 + 8.217e^{-1} * X_1^2 +$$
$$2.865e^{-1} * X_1^3 - 1.644e^{-1} * X_8^1 - 2.648e^{-1} * X_8^2 - 2.289e^{-1} * X_8^3 + 1.013e^{-1} * X_9^1 +$$
$$9.904e^{-2} * X_9^2 + 3.207e^{-1} * X_9^3 + 1.961e^{-2} * X_3 + 3.318e^{-1} * \log(X_2) + 8.014e^{-3} * X_4 +$$
$$5.015e^{-1} * X_{20} - 1.672e^{-1} * X_{21} - 3.428e^{-1} * X_{14} + 4.254e^{-1} * X_{12} - 3.969e^{-1} * X_{19} + \epsilon$$

따라서 최종 모델 3가지는 다음과 같다.

번호	모델	변수 개수
①	$MODEL\,1: \log(Y) =$ $5.953e^{-1} + 1.033e^{-1} * X_1^1 + 3.947e^{-2} * X_1^2 + 2.751e^{-1} * X_1^3 +$ $2.673e^{-1} * \log(X_2) + 1.721e^{-2} * X_3 + 9.719e^{-1} * X_6^A + 6.189e^{-1} * X_6^B +$ $2.886e^{-1} X_3 * X_6^A + 1.740e^{-1} X_3 * X_6^B + 2.642e^{-1} X_3 * X_6^C +$ $7.296e^{-3} X_4 + 2.033e^{-1} X_7^{Spring} - 1.622e^{-1} X_7^{Summer} + 1.652e^{-1} X_7^{Winter} -$ $4.590e^{-4} X_{13} + 1.591e^{-3} X_{14} + 2.915e^{-6} X_{20} + \epsilon$	9
②	$MODEL\,2: \log(Y) =$ $-3.397e^{0} * 1 - 7.737e^{-3} * X_4 + 3.267e^{-1} * \log(X_2) + 7.236e^{-1} * X_6^A +$ $4.428e^{-1} * X_6^B + 2.659e^{-1} * X_6^C + 5.338e^{-6} * X_{20} + 2.659e^{-1} * X_1^1 +$ $1.932e^{-2} * X_3 + 3.474e^{-2} * X_7^{Spring} - 1.153e^{-1} * X_7^{Summer} +$ $2.412e^{-1} * X_7^{Winter} - 1.348e^{-5} * X_{21} - 1.360e^{-1} * X_8^1$ $-2.985e^{-1} * X_8^2 - 1.645e^{-1} * X_8^3 + 7.618e^{-2} * X_9^1 + 3.571e^{-2} * X_9^2 +$ $3.829e^{-1} * X_9^3 - 6.419e^{-3} * X_{14} + 3.614e^{-4} * X_{12} - 4.909e^{-2} * X_{19}$ $-5.380e^{-0} * 1/X_5 - 2.687e^{-3} * X_{15} - 3.222e^{-3} * \quad X_{11} - 5.687e^{-4} * X_{13} + \epsilon$	17
③	$MODEL\,3: \log(Y) =$ $-6.980e^{0} * 1 + 7.032e^{-1} * X_6^A + 4.366e^{-1} * X_6^B + 2.850e^{-1} * X_6^C +$ $2.899e^{-1} * X_7^{Spring} - 1.477e^{-1} * X_7^{Summer} + 2.316e^{-1} * X_7^{Winter} + 1.071e^{-1} * X_1^1 +$ $8.217e^{-1} * X_1^2 + 2.865e^{-1} * X_1^3 - 1.644e^{-1} * X_8^1 - 2.648e^{-1} * X_8^2$ $9.904e^{-2} * X_9^2 + 3.207e^{-1} * X_9^3 + 1.961e^{-2} * X_3 + 3.318e^{-1} * \log(X_2)$ $+ 8.014e^{-3} * X_4 + 5.015e^{-1} * X_{20} - 1.672e^{-1} * X_{21} - 3.428e^{-1} * X_{14}$ $+ 4.254e^{-1} * X_{12} - 3.969e^{-1} * X_{19} + \epsilon$	13

표1: 최종 모델

만들어진 모델을 분석해보고 각각의 적합성을 고려하기 위해

① 표준화 회귀계수(변수별 중요도),
② 결정계수 , AIC $alue$(설명력)
③ F값(적합성),
④ 더빈-왓슨 테스트(자기상관)
⑤ 분산팽창계수(다중공선성),
⑥ QQ norm, 히스토그램(잔차의 정규성)
등을 살펴보았다.

① 표준화 회귀계수(변수별 중요도 파악)

1	Standardized Coefficients:: (Intercept) discomfort log(hangout) sortA sortB sortC 0.00000000 0.39469614 0.31900787 0.20741397 0.42196882 0.11912868 seasonspring seasonsummer seasonwinter factor(event)1 factor(event)2 factor(event)3 0.13466949 -0.10749771 0.10874549 0.04474297 0.01211817 0.08949174 google population oil kospi log(hangout):sortA log(hangout):sortB 0.17388434 0.20288745 0.05781887 -0.07261698 0.07000802 0.14428467 log(hangout):sortC 0.06960313
2	dized Coefficients:: (Intercept) google I(log(hangout)) sortA sortB 0.00000000 0.18437231 0.38989759 0.15442822 0.30189707 sortC population factor(event)1 factor(event)2 factor(event)3 0.06042150 0.37153853 0.04867554 0.02025939 0.10055469 discomfort factor(season)spring factor(season)summer factor(season)winter school 0.44315131 0.23020276 -0.07636507 0.15878595 -0.03907518 factor(cow)1 factor(cow)2 factor(cow)3 factor(bird)1 factor(bird)2 -0.09607787 -0.13979073 -0.05093176 0.05512491 0.02494009 factor(bird)3 oil gold unemploy I(1/airquality) 0.12602610 -0.23323909 0.10941930 -0.03655448 -0.04660287 usdkrw I(sqrt(sewolho)) kospi -0.19407480 -0.06919623 -0.08996781
3	Standardized Coefficients:: (Intercept) sortA sortB sortC factor(season)spring 0.00000000 0.15007578 0.29766232 0.06476898 0.19204906 factor(season)summer factor(season)winter factor(event)1 factor(event)2 factor(event)3 -0.09785771 0.15247380 0.04638982 0.02522549 0.09320612 factor(cow)1 factor(cow)2 factor(cow)3 factor(bird)1 factor(bird)2 -0.11617009 -0.12400497 -0.07090334 0.07331047 0.06917175 factor(bird)3 discomfort I(log(hangout)) google population 0.10556005 0.44993461 0.39599699 0.19097510 0.34906948 school oil gold unemploy -0.04844073 -0.12456291 0.12879055 -0.02955978

표2: 표준화 회귀계수

1번 모델의 경우 휴일에 따른 분류와 불쾌지수, 나들이 지수 등이 높은 변수로 측정되었고 반면 국제 유가와 코스피 지수, 행사 등이 낮은 변수로 선택 되었다. 2번 모델은 인구, 기름 등의 중요도가 1번에 비해 크게 늘었고, AI가 변수로 채택되었지만 양수로 나오는 이상한 현상도 발견할 수 있었다. 휴일과 나들이 지수, 불쾌지수는 여전히 중요한 변수로 측정되었다. 3번은 2번과 대체로 비슷했는데, 소풍은 음수로, AI가 양수로 나오는 등 상식과 반하는 오류가 있고, 인구, 휴일, 나들이, 불쾌지수 등이 여전히 중요하게 측정되었다.

② 결정계수 R^2, AIC value 비교하기(적합성, 설명력)

	AIC $value$	BIC $value$	$adj.R^2$
1번 모델	2222.375	2339.307	0.6775
2번 모델	2031.398	2200.949	0.7017
3번 모델	2126.492	2272.656	0.6899

표3: R^2, AIC 비교

3개 값 모두 2번 모델이 가장 좋게 나왔으며, 1번 모델은 가장 안 좋게 나왔다.

③ 값(적합성)

1	Residual standard error: 0.3721 on 2538 degrees of freedom Multiple R-squared: 0.6797, Adjusted R-squared: 0.6775 F-statistic: 299.3 on 18 and 2538 DF, p-value: < 2.2e-16
2	Residual standard error: 0.3579 on 2529 degrees of freedom Multiple R-squared: 0.7049, Adjusted R-squared: 0.7017 F-statistic: 223.7 on 27 and 2529 DF, p-value: < 2.2e-16
3	Residual standard error: 0.3649 on 2533 degrees of freedom Multiple R-squared: 0.6927, Adjusted R-squared: 0.6899 F-statistic: 248.3 on 23 and 2533 DF, p-value: < 2.2e-16

표4: F 값 비교

세 모델 모두 p-value가 상당히 낮으므로 전체 모델은 상당히 유의하다고 할 수 있다.

④ 더빈 왓슨 테스트(자기상관)

1	Durbin-Watson test data: fit1 DW = 1.2138, p-value < 2.2e-16 alternative hypothesis: true autocorrelation is greater than 0
2	DW = 1.3578, p-value < 2.2e-16 alternative hypothesis: true autocorrelation is greater than 0
3	DW = 1.3135, p-value < 2.2e-16 alternative hypothesis: true autocorrelation is greater than 0

표5: 더빈 왓슨 테스트

모델 1의 경우 1.2로 상대적으로 2에 가까우므로 큰 문제가 있다고 말할 수는 없다. 모델 2와 3은 각각 1.3578, 1.31로 더 가까워져 문제가 위 모델보다는 줄었다고 말할 수 있다.

⑤ 분산팽창계수(다중공선성)

1		GVIF	Df	GVIF^(1/(2*Df))
	discomfort	4.728588	1	2.17453
	log(hangout)	2.611572	1	1.61603
	sort	13.668644	3	1.54627
	season	10.645009	3	1.48317
	factor(event)	1.683499	3	1.09069
	google	1.380705	1	1.17503
	population	2.646285	1	1.62674
	oil	2.207051	1	1.48561
	kospi	1.567696	1	1.25207
	log(hangout):sort	13.731355	3	1.54745
2		GVIF	Df	GVIF^(1/(2*Df))
	google	1.383177	1	1.176085
	I(log(hangout))	2.586383	1	1.608223
	sort	2.005560	3	1.122982
	population	4.457775	1	2.111344
	factor(event)	1.936269	3	1.116420
	discomfort	5.853735	1	2.419449
	factor(season)	24.644881	3	1.705903
	school	1.916195	1	1.384267
	factor(cow)	4.397370	3	1.279967
	factor(bird)	10.415433	3	1.477791
	oil	5.812269	1	2.410865
	gold	2.415401	1	1.554156
	unemploy	2.565073	1	1.601584
	I(1/airquality)	1.317743	1	1.147930
	usdkrw	3.727395	1	1.930646
	I(sqrt(sewolho))	1.846771	1	1.358960
	kospi	2.630422	1	1.621858
3		GVIF	Df	GVIF^(1/(2*Df))
	sort	1.963683	3	1.119039
	factor(season)	21.193230	3	1.663539
	factor(event)	1.829650	3	1.105931
	factor(cow)	3.609036	3	1.238507
	factor(bird)	7.020065	3	1.383748
	discomfort	5.783327	1	2.404855
	I(log(hangout))	2.444265	1	1.563415
	google	1.365727	1	1.168643
	population	3.279297	1	1.810883
	school	1.883497	1	1.372406
	oil	3.809856	1	1.951885
	gold	2.248982	1	1.499661
	unemploy	2.560010	1	1.600003

표6: 다중공선성

모델 1,2,3 모두에서 자유도를 고려한 오른쪽 값을 봤을 때 값들이 10 이하이고, 꽤 작으므로 다중공선성이 없다고 말할 수 있다.

⑥ QQ norm, 히스토그램(잔차의 정규성)

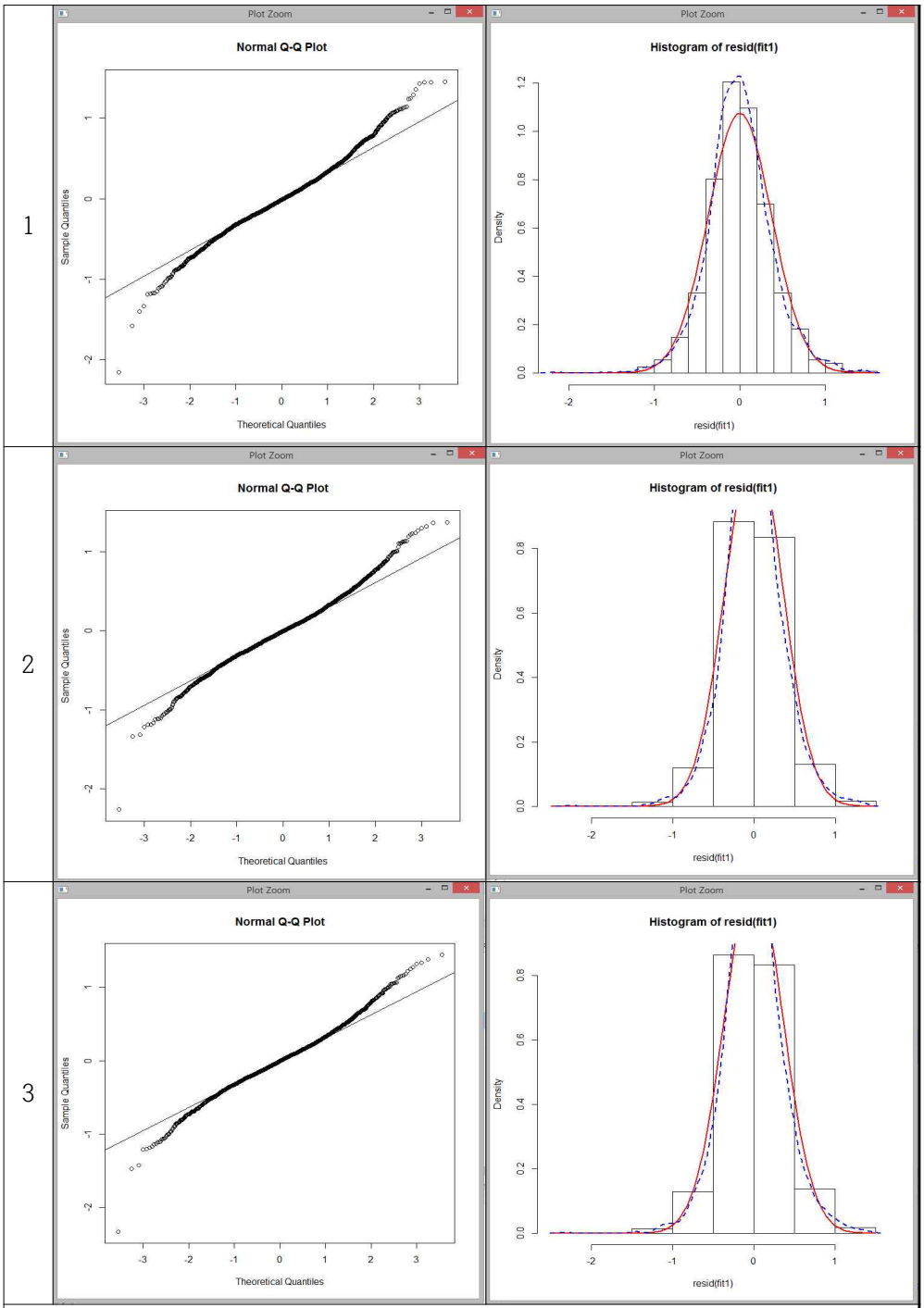

그림25: QQ norm과 히스토그램

QQ Plot의 경우 세 모델 모두 양 끝에서 선을 벗어나므로 조금 문제가 있다고는 얘기할 수 있다. 1번 모델의 히스토그램의 경우 실제 분포(파란선)가 이상적 정규 분포(빨간선)와 크게 다르지 않으므로 잔차의 정규성을 잘 만족한다고 얘기할 수 있고, 2,3번 모델은 1번 모델보다 잔차그래프가 조금 이상하긴 하지만 큰 이상은 없는 것으로 보인다.

3.3 회귀분석 모델의 예측 및 분석

1) 예측 결과: 예측의 경우 2016.06.01. - 2017.05.31.의 입장객수 데이터를 가지고 하였다. 회귀문제이고, 예측하려는 변수 입장객 수의 값이 큰 편이므로 오차가 생길 수밖에 없어 $= \begin{matrix} |\ \ -Y| < 5000, & \text{은 예측} \\ |Y-\hat{Y}| > 5000, & \text{틀린 예측} \end{matrix}$ 으로 본 팀의 기준을 제시하고 임의 분리하여 옳은 예측의 데이터 개수를 구하여 모델의 예측력을 구하고 비교하였다.

	옳은 예측	틀린 예측	비율	RMSE	MAE
1번 모델	**272**	**93**	**0.7452**	7241.452	**4361.901**
2번 모델	242	123	0.6630	7647.774	5103.721
3번 모델	253	112	0.6932	**7129.443**	4689.315

표7: 예측 결과

옳게 예측한 비율과 MAE의 관점에서는 1번모델이 가장 좋았고, RMSE의 관점에서는 3번 모델이 가장 좋게 나왔다.

3.4 SVR을 사용한 분석

본 레포트에서는 다항 커널과, RBF 커널을 이용하였으며, 그 전의 연구들을 참고하여 다항 커널의 경우 차원 d =1, 2, 3을 사용하였고 RBF 커널의 경우 =1, 0.04, 0.005를 사용하였으며 공통 모수인 c =0.0001, 0.001, 0.01, 0.08, 0.1로 5가지 경우 그리고 ϵ= 0.1에서 0.1단위로 1까지 10개를 사용하였다.

표에서 제시한 바와 같이 다항커널 모델의 경우 ϵ=0.8, d=1, C=0.08인 모델이 최적이었고, RBF 커널을 사용한 모델의 경우 ϵ=0.4, γ=0.005, C=1인 모델이 가장 좋게 나왔다.

	C	d,	ϵ									
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
다항 커널	0.0001	1	0.32	0.32	0.30	0.29	0.28	0.27	0.25	0.24	0.24	0.22
		2	0.32	0.32	0.30	0.29	0.27	0.27	0.24	0.24	0.23	0.23
		3	0.32	0.32	0.30	0.29	0.27	0.27	0.24	0.24	0.24	0.23
	0.001	1	0.38	0.37	0.36	0.35	0.33	0.30	0.28	0.27	0.25	0.22
		2	0.33	0.32	0.31	0.30	0.28	0.28	0.25	0.24	0.25	0.22
		3	0.33	0.33	0.32	0.30	0.29	0.27	0.25	0.27	0.24	0.22
	0.01	1	0.70	0.70	0.69	0.68	0.67	0.62	0.6	0.51	0.42	0.32
		2	0.42	0.44	0.41	0.41	0.38	0.35	0.32	0.29	0.27	0.25
		3	0.51	0.50	0.46	0.44	0.41	0.37	0.32	0.30	0.27	0.23
	0.08	1	0.71	0.71	0.71	**0.72**	**0.72**	**0.72**	**0.72**	*0.73*	**0.72**	0.69
		2	0.56	0.56	0.55	0.53	0.52	0.48	0.42	0.40	0.33	0.28
		3	**0.72**	0.71	0.71	0.71	0.69	0.65	0.64	0.58	0.49	0.41
	1	1	0.70	0.71	0.70	0.71	0.70	0.70	0.70	0.69	0.69	0.70
		2	0.46	0.47	0.47	0.48	0.49	0.48	0.48	0.49	0.47	0.46
		3	0.63	0.59	0.62	0.62	0.59	0.64	0.6	0.62	0.61	0.65
RBF 커널	0.0001	1	0.32	0.32	0.29	0.29	0.27	0.27	0.24	0.24	0.23	0.22
		0.04	0.32	0.32	0.30	0.29	0.28	0.27	0.24	0.24	0.24	0.22
		0.005	0.32	0.32	0.30	0.29	0.27	0.27	0.24	0.24	0.24	0.22
	0.001	1	0.32	0.32	0.29	0.29	0.27	0.27	0.24	0.24	0.24	0.22
		0.04	0.36	0.35	0.34	0.32	0.30	0.27	0.26	0.26	0.25	0.23
		0.005	0.33	0.33	0.32	0.30	0.29	0.28	0.25	0.26	0.24	0.22
	0.01	1	0.32	0.31	0.28	0.27	0.25	0.25	0.24	0.23	0.22	0.20
		0.04	0.62	0.59	0.56	0.53	0.50	0.44	0.39	0.34	0.28	0.24
		0.005	0.43	0.42	0.40	0.36	0.35	0.32	0.30	0.28	0.25	0.22
	0.08	1	0.26	0.26	0.24	0.21	0.21	0.19	0.19	0.19	0.18	0.16
		0.04	0.66	0.66	0.68	0.66	0.65	0.65	0.64	0.6	0.56	0.51
		0.005	0.70	**0.71**	0.70	0.70	0.68	0.62	0.61	0.53	0.43	0.34
	1	1	0.18	0.17	0.16	0.16	0.15	0.14	0.14	0.14	0.15	0.14
		0.04	0.56	0.60	0.60	0.60	0.64	0.66	0.66	0.70	0.68	0.69
		0.005	**0.73**	**0.73**	*0.74*	*0.742*	**0.73**	**0.73**	**0.73**	**0.73**	**0.72**	0.69

표8: 전체 모수 및 커널에 따른 분석 결과

	옳은 예측	틀린 예측	비율	RMSE	MAE
Best Poly	266	99	0.7287	7693.312	4688.414
Best RBF	271	94	**0.7424**	**6893.838**	**4330.099**

표9: 최적모델의 예측 결과

3.5 시계열을 사용한 월별 분석

1) 데이터 성질 확인: 2010년 6월부터 2016년 5월까지 자료를 월별로 변환 후 데이터의 성질을 확인해보았다.

① 무작위성 확인

```
        Runs Test

data:  z
Standard Normal = -3.7981, p-value = 7.291e-05
alternative hypothesis: less
```

그림26. Runs Test

입력 자료는 Runs Test 입력조건에 맞추어 조사자료 값이 조사자료 Median보다 작으면 0, 크면 1로 변환하였다. 주어진 자료는 Standard Normal = -3.7981, p-value = 7.291e-05로 자료가 한쪽으로 치우치지 않고 무작위성이 있다는 귀무가설을 채택할 수 없다.

② 추세 검정에 대한 분석: spearman test

```
        Spearman's rank correlation rho

data:  tt and rr1
S = 79196, p-value = 0.02045
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2733295
```

그림27. Spearman Test

시간과 순위가 독립이라는 귀무가설을 기각하며, 시계열 자료에 추세가 존재한다고 판단된다.

③ 독립성 검정 : Box Test(Box-Pierce, Ljung-Box)

```
        Box-Pierce test

data:  v_last.ts
X-squared = 16.161, df = 1, p-value = 5.817e-05

        Box-Ljung test

data:  v_last.ts
X-squared = 16.844, df = 1, p-value = 4.058e-05
```

그림28. Box Test

'자기상관성이 없다.'라는 귀무가설을 채택할 수 없다.

④ 정상성 (단위근) 검정

```
        Augmented Dickey-Fuller Test

data:  (log(v_last.ts))
Dickey-Fuller = -4.453, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test((log(v_last.ts)), alternative = "stationary", k = 0) :
  p-value smaller than printed p-value
```

그림29. Unit Root Test

Stationary이므로 ARIMA모델 사용이 가능하다.

2) 요소분해 및 그래프 확인

시계열 자료는 우연변동, 추세변동, 계절 변동 등 다양한 변동들의 성분들이 중복적으로 중첩되어 있다. 여러 방법들이 있지만 여기에서는 가법모형(additive)만 이용한다.

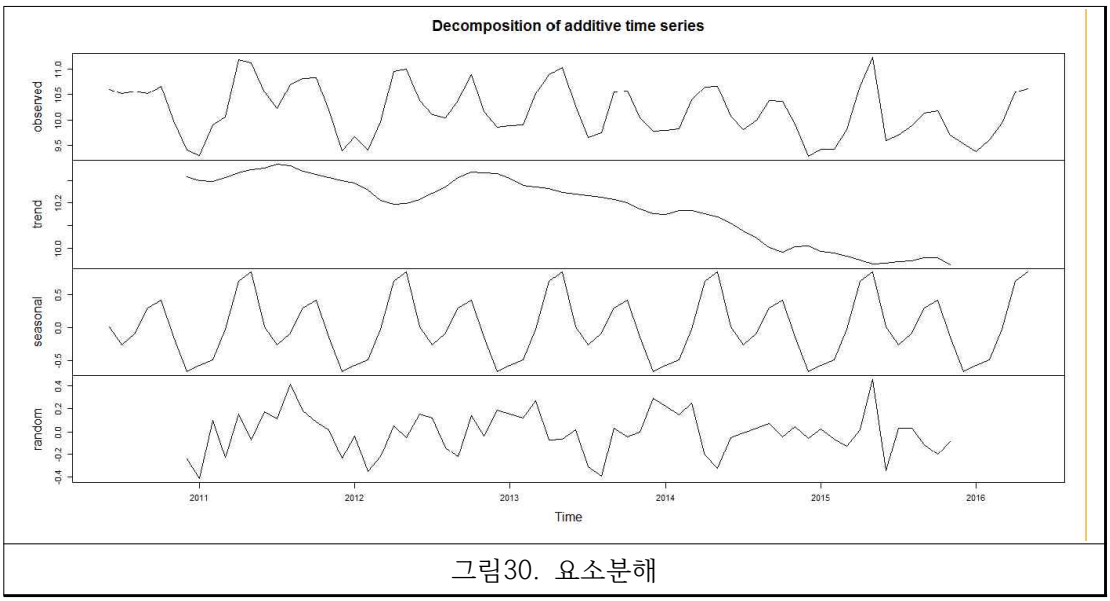

그림30. 요소분해

그림30은 맨 위부터 실제 월별 그래프, 추세 변동 그래프, 계절 변동 그래프, 잔차 변동 그래프가 그려져 있다. 점점 감소하는 추세 그래프를 확인할 수 있고, 12개월 단위로 두 개의 봉우리가 있는 계절 그래프를 확인할 수 있다.

3) 모델 생성

```
 ARIMA(2,0,2)(1,1,1)[12] with drift         : Inf
 ARIMA(0,0,0)(0,1,0)[12] with drift         : 22.07796
 ARIMA(1,0,0)(1,1,0)[12] with drift         : 14.82652
 ARIMA(0,0,1)(0,1,1)[12] with drift         : 12.50316
 ARIMA(0,0,0)(0,1,0)[12]                    : 24.99233
 ARIMA(0,0,1)(1,1,1)[12] with drift         : Inf
 ARIMA(0,0,1)(0,1,0)[12] with drift         : 19.62171
 ARIMA(0,0,1)(0,1,2)[12] with drift         : Inf
 ARIMA(0,0,1)(1,1,2)[12] with drift         : 15.70185
 ARIMA(1,0,1)(0,1,1)[12] with drift         : 12.38041
 ARIMA(1,0,0)(0,1,1)[12] with drift         : 10.79407
 ARIMA(2,0,1)(0,1,1)[12] with drift         : 14.43362
 ARIMA(1,0,0)(0,1,1)[12]                    : 15.94096
 ARIMA(1,0,0)(1,1,1)[12] with drift         : Inf
 ARIMA(1,0,0)(0,1,0)[12] with drift         : 18.92646
 ARIMA(1,0,0)(0,1,2)[12] with drift         : Inf
 ARIMA(1,0,0)(1,1,2)[12] with drift         : 14.37903
 ARIMA(0,0,0)(0,1,1)[12] with drift         : 15.39053
 ARIMA(2,0,0)(0,1,1)[12] with drift         : 11.95987

 Best model: ARIMA(1,0,0)(0,1,1)[12] with drift

Series: log(v_last.ts)
ARIMA(1,0,0)(0,1,1)[12] with drift

Coefficients:
          ar1      sma1    drift
      0.3312  -0.6448  -0.0075
s.e.  0.1221   0.2297   0.0021

sigma^2 estimated as 0.05726:  log likelihood=-1.03
AIC=10.07   AICc=10.79   BIC=18.44
```

그림31. 시계열 모델

2010년 6월부터 2016년 5월까지 월별 자료를 log변환을 하여 모델을 생성하였다. 각각의 모델에 대해서 우측에 AIC 값이 적혀 있다. 이 중 값이 가장 적은 ARIMA(1,0,0)(0,1,1) with drift가 선택 되었다. 'with drift'라는 뜻은 추세가 존재한다는 뜻이다. 위에서 확인한 요소분해 그래프와 유사하게 추세(drift)가 음수로 추정되었다.

4) 예측 및 분석

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jun 2016	17129.12	12598.475	23289.07	10707.578	27401.79
Jul 2016	14955.66	10821.553	20669.09	9118.114	24530.47
Aug 2016	17700.19	12785.121	24504.78	10762.656	29109.60
Sep 2016	25062.71	18099.732	34704.34	15235.015	41229.97
Oct 2016	27358.58	19757.348	37884.23	16630.090	45008.29
Nov 2016	16018.96	11568.272	22181.98	9737.196	26353.29
Dec 2016	10954.64	7911.015	15169.25	6658.825	18021.83
Jan 2017	10985.48	7933.285	15211.96	6677.570	18072.56
Feb 2017	12216.71	8822.434	16916.89	7425.980	20098.10
Mar 2017	18458.13	13329.747	25559.57	11219.857	30366.04
Apr 2017	35354.51	25531.731	48956.38	21490.494	58162.51
May 2017	42325.61	30566.863	58607.82	25729.033	69627.85

그림32. 시계열 모델 예측 결과

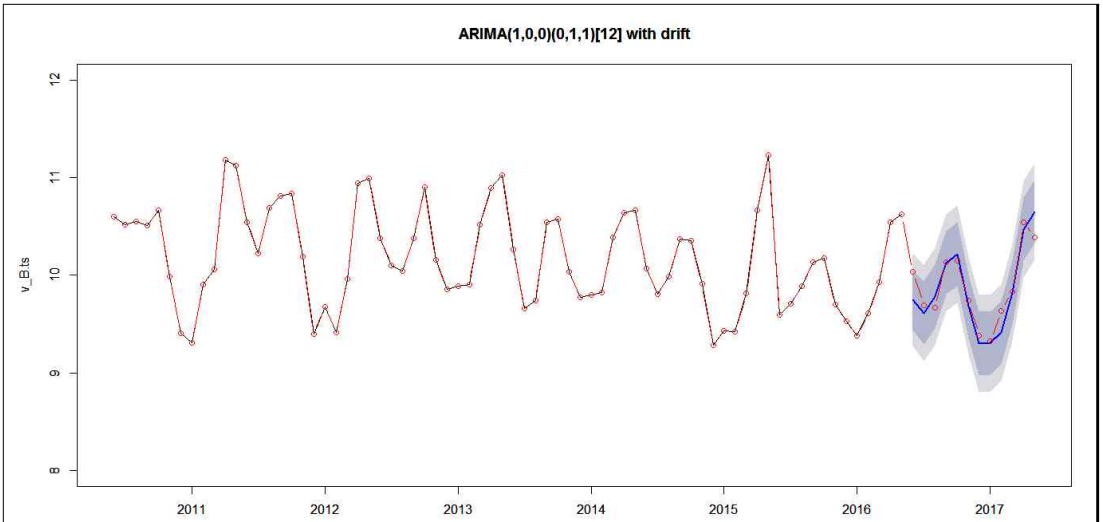

그림33. 시계열 모델 예측 결과

2010년 6월부터 2016년 5월까지 자료를 이용해 2016년 6월부터 2017년 5월까지 자료를 예측하였다. 그림33 에서 짙은 영역은 80% 신뢰 범위를 나타내고 옅은 영역은 95% 신뢰 범위를 나타낸다. 위 그림의 빨간 점은 실제 값들을 나타낸다. 그림과 같이 모든 점들이 짙은 영역 안에 들어갔다.

ME	RMSE	MAE	MPE	MAPE
136.6658	3599.035	2385.216	2.300416	10.49283

표10: 시계열 모델의 오차

위에서 보는 바와 같이 일별 모델을 만들었을 때보다 훨씬 적은 MAE와 RMSE값을 도출하여 월별모델이 더 좋은 결론을 낸다는 것을 알 수 있다.

Ⅳ. 결론 및 추후 연구

'어린이대공원 입장객수'을 주제로 정하고 예측 모델을 만들면서 여러 가지 어려운 점이 많이 있었지만 그중 가장 힘들었던 것은 데이터를 모으고 전처리 하는 과정이었다. 본 팀은 학술활동 프로그램을 시작하기 이전에 별다른 데이터 전처리 방식에 대한 학습이 부족하였기 때문에 이 과정에서 시간과 노력을 많이 쏟아야 했다. 예를 들어, 실제로 가장 중요한 변수라고 생각했던 '인근 학교의 소풍날' 데이터의 경우 일별 자료를 구할 수 없어 월별 자료로 반영하였다.

이후 일별 예측에는 회귀분석, 인공신경망, SVR 모델을 사용하였다. 일별 예측에서는 회귀분석 1번 모형이 가장 좋은 예측률(74.52%)을 보였고 RBF 커널을 사용한 SVR 모델이 예측률은 근소하게 낮지만 가장 작은 MAE 값(4330.099)을 보였다. 따라서 현재의 데이터를 가지고 어떤 분석 방법을 택하는지에 대한 논의는 더 필요할 것으로 보이며 동시에 이 데이터의 한계라고도 얘기할 수 있을 듯하다. 실제로 가장 중요한 변수일 것으로 생각되는 인근 학교의 '소풍날' 데이터의 경우 일별 자료를 구할 수 없어 월별로 같은 값을 집어넣었고 이와 같은 작업이 4월과 5월의 많은 예측 오류를 야기하였을 것으로 생각된다. 그리고 모델의 변수를 줄이는 과정에서 오히려 예측률이 증가하였으므로 단순히 많은 데이터를 사용하기보다는 쓸모 있는 변수를 잘 택하는 것이 돈과 시간의 낭비를 줄이고 좋은 결론을 도출하는데 유의할 것이라고 생각한다. 또한 본 팀은 회귀분석 직후 인공신경망을 통해 회귀 모델을 만들고 예측을 시도하였는데 오히려 예측률이 떨어지는 현상이 발생하였고 우리는 이를 '부족한 데이터에 대한 과적합이 발생하였다'라고 결론지었다. 과적합에 강한 SVR의 경우 회귀분석과 비슷한 예측률을 보였기 때문에 타당한 전개라고 생각한다.

모델의 정확도를 높이기 위하여 과거의 입장객 수만 가지고도 유의한 결과를 끌어낼 수 있을 것으로 생각되는 시계열 모델을 사용해 보았다. 시계열 모델의 경우 일별 분석의 오차가 너무 커지게 되어 월별 예측을 시도하였다. 요소 분석을 통해 감소하는 추세를 알 수 있었고, 봄과 가을에 입장객이 증가하고 여름 특히 겨울에 입장객이 감소하는 계절적 용인을 가지는 것을 확인할 수 있었다. 시계열 분석으로 채택 모형은 MAE 값을 2385.216, MAPE 값을 10.49283으로 가졌다. 비록 월별 예측이었지만 좀 더 만족스러운 결과를 얻을 수 있었다.

따라서 본 팀은 높은 예측률을 위해서는 경우에 맞게 여러 기법을 시도해보고 장단점을 비교하는 것이 문제의 오류를 파악하고 다양한 시각을 키우는데 도움이 됨을 확인하였다. 특히, 일별 예측에 사용된 요인 분석의 경우 유의한 데이터를 얻는 것이 가장 중요하였다. 앞으로 데이터 분석에 우리가 학술활동 프로그램을 통해 배운 지식과 데이터를 보는 안목이 힘이 되리라 믿는다.

Reference

[1] 진강규 (2000), 유전 알고리즘과 그 응용, 교우사, p.61-62, 93-94

[2] 문병로 (2003), 유전 알고리즘, 두양사

[3] 김동일 (2009), 선형회귀모델의 변수선택을 위한 다중목적 유전 알고리즘과 그 응용

[4] 홍승현 (1999), 유전자 알고리즘을 활용한 인공신경망 모형 최적입력변수의 선정: 부도예측 모형을 중심으로

[5] 김혁주, 김예형 (2015), 변수선택 기법을 이용한 한국 프로야구의 득점과 실점 설명

[6] 이재길 (2017), R프로그램에 기반한 시계열 자료 분석, 황소걸음아카데미

[7] 김성진 등 (2012), 감정 예측 모형의 성과 개선을 위한 Support Vector Regression 응용

[8] 박성현 (2007), 회귀분석 3판, 민영사

[9] Acosta-Gonzalez, E. and Fernandez-Rodriguez, F. (2004), Model Selection Via Genetic Algorithms

[10] Kutner, M. et al. (2004), Applied regreesion models 4^{th} edition, McGraw Hill education

[11] Paterlini, S. and Minerva, T (2010), Regression Model Selection Using Genetic Algorithms

[12] Sofge, D. A. (2002), Using Genetic Algorithm Based Variable Selection to Improve Neural Network Models for Real-World Systems

[13] Smola, A. J. and Scholkopf, B. (2003), A Tutorial on Support Vector Regression