10장

Generalized Additive Model

Regression Model

- For a response variable and predictor variables $x_1, x_2, \ldots, x_k$ can be modeled using a mean function $f(\cdot)$ as follows: $$y_i = f(x_{1i}, x_{2i}, \ldots, x_{ki}) + \varepsilon_i$$

- $f(\cdot)$ would be a parametric / nonparametric regression or a smoothing spline regression.

- Nonparametric regression / smoothingspline models tend to require many samples for accurate estimation results.

- If the number of predictors is large, more samples are required (curse of dimensionality).

- To avoid this problem, an additive model (GAM) is suggested by Stone (1985).

Generalized Additive Model

- An additive model is as follows:

$$y_i = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \ldots + f_k(x_{ki}) + \varepsilon_i$$

- $f_i(\cdot)$ would be a parametric/ nonparametric regression or a smoothing spline regression.

- If the response variable is allowed to have many types of distributions (a normal, Poisson, Logistic, etc.), it is referred to as a generalized additive model (GAM).

- Hastie and Tibshirani (1990) suggested the GAM.

- For example,

  if the response variable has a Logistic distribution, the GAM can be expressed as

$$ln\frac{P(y_i=1)}{P(y_i=0)} = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \ldots + f_k(x_{ki})$$

Case Study 1 : Bankruptcy Prediction

- Data
  - Bankrupt & non-bankrupt companies during IMF bail-out were examined.
  - Companies were selected from Korea Stock Exchange.
  - 30 bankrupt and 54 non-bankrupt companies were investigated.
  - Three predictors: long-term fixed adopt rate (D3), cash-flow versus deb ratio (D8), gross capital investment efficiency (F6) were considered.
  - Two class of response variable: Bankrupt versus non-bankrupt is considered.

- Model

$$ln\frac{P(y_i=1)}{P(y_i=0)} = \beta_0 + f_1(D_{3i}) + f_2(D_{8i}) + f_3(F_{6i})$$

  - Mean functions are modeled via smoothing spline methods.

Case Study 1 : R code

```
library(mgcv);
y0=read.delim("c:/predict/bankrupt.txt",header=T); # read data set
names(y0);
###########################################################
class=as.numeric(y0$class);
D3=as.numeric(y0$D3);
D8=as.numeric(y0$D8);
F6=as.numeric(y0$F6);
z0=cbind(class, D3, D8, F6);
z1=data.frame(z0);
############### glm ################################################
out0=glm(class~D3+D8+F6, family="binomial", data=z1);
summary(out0);
################ gam ####################################################
out1=gam(class~ s(D3)+s(D8)+s(F6), family="binomial", data=z1);
summary(out1);
```

Case Study 2 : Call volume forecasting

- Data
  - 5-min call arrival rates at a call center.
  - One day is composed of 169 observations.
  - Intraday and intraweek patterns are suspected.

- Transform

$$N_t \sim Poisson \quad \Rightarrow \quad Y_t = \sqrt{N_t + 0.25} \sim Normal$$

- Model

$$y = \beta_0 + factor(week\_day) + S_2(time) + \varepsilon$$

- $\varepsilon \sim N(0, \sigma^2)$
- $S_i(\cdot)$: smoothing spline

Case Study 2 : Call volume forecasting

- Comparison with a seasonal linear model

Seasonal linear model

$$y_{ij} = \mu + \alpha_{d_i} + \beta_j + \delta_{d_{ij}} + \varepsilon_{ij}$$

- $\varepsilon_{ij} \sim N(0, \sigma^2)$

- d_i : weekday
- α_{d_i} : day of week
- β_j : time of day
- $\delta_{d_{ij}}$: time of day x day of week

Case Study 2 : R code

```
library(mgcv); library(lubridate);
y0=read.delim("c:/predict/newcall5.txt",header=T); # read data set
#######################################################
call=as.numeric(y0$call);
y=as.numeric(y0$adj_call);
time=as.numeric(y0$time);
wd=wday(as.Date(y0$date)); #wd=weekdays(as.Date(y0$date));
z0=cbind(call, y, time, wd);
z1=data.frame(z0);
############## glm #############################################
out0=glm(y~factor(time)+factor(wd)+factor(time):factor(wd), data=z1);
summary(out0);
############### gam #################################################
out1=gam(y~s(time)+factor(wd), data=z1);
summary(out1);
```

Case Study 3 : flight arrival time forecasting

- arrival time (N_i) = actual arrival time – scheduled arrival time

  If the arrival time is positive, it is the delayed case
  If the arrival time is negative, it is the early arrival case

- Response variable: in order to make a positive value in log transform

$$y_i = \ln(N_i + 150)$$

- Explanatory variables
  - Departing airport: departure delay time
  - Airborne state: scheduled airborne time
  - Arriving airport:
    1. Airport capacity
       (seasonal factor: time of day, day of month, month of year)
    2. Weather condition
    3. Airline

Case Study 3 : flight arrival time forecasting

- Model

  S( ): smoothing spline function

$$y_i = \beta_0 + s(t_i) + \mathrm{s}(d_i) + \mathrm{s}(m_i) + \mathrm{s}(dep_i) + \mathrm{s}(h_i) + \sum_{k=1}^{6} \gamma_k air_k + k_1 \omega_i + \varepsilon_i$$

- Benchmark models: Linear regression, median regression

$$y_i = \beta_0 + \beta_1 t_i + \mathrm{factor}(d_i) + \mathrm{factor}(m_i) + \beta_3 dep_i + \beta_4 h_i + \sum_{k=1}^{6} \gamma_k air_k + \beta_5 \omega_i + \varepsilon_i$$

- Computer software: R packet “mgcv”, function “gam”

Case Study 3 : R code

```
library(mgcv);
y0=read.delim("c:/predict/new2010.txt",header=T); # read data set
##################################################################
arr=y0$ARR_DELAY      # arrival delay time: 도착지연시간
dep=y0$DEP_DELAY    # departure delay time: 출발지연시간
h=y0$CRS_ELAPSED_TIME # airborne time (flying time)
atime=y0$CRS_ARR_H    # planed arriving time: 도착예정시각
ar_day=y0$DAY_OF_MONTH      # day of month of flight: 비행 예정일
ar_mon=y0$MONTH             # month of flight: 비행 달
w3=y0$w3;  # weather condition
p1=y0$A1;  p2=y0$A2;  p3=y0$A3;  p4=y0$A4;  p5=y0$A5;  p6=y0$A6; # airline dummies
parr=arr+150  #min(arr)=-81
logparr=log(parr);
z0=cbind(logparr, parr, arr, dep, h, atime, ar_day, ar_mon, p1, p2, p3, p4, p5, p6, w3);
z1=data.frame(z0);
```

Case Study 3 : R code

```
############## glm ##########################################
out0=glm(logparr~dep+h+atime+factor(ar_day)+factor(ar_mon)+p1+p2+p3+p4+p5+p6+w3, data=z1);
summary(out0);
newdataset=z1[1:100,];
pred0=predict.glm(out0, newdata=newdataset);
newpred0=exp(pred0)-150
diff0=(newpred0-newdataset$arr)^2;
(mse0=mean(diff0));
(rmse0=sqrt(mse0));

############### gam ############################################
out1=gam(logparr~s(dep)+s(h)+s(atime)+s(ar_day)+s(ar_mon)+p1+p2+p3+p4+p5+p6, data=z1);
summary(out1);
newdataset=z1[1:100,];
pred1=predict.gam(out1, newdata=newdataset);
newpred1=exp(pred1)-150
diff1=(newpred1-newdataset$arr)^2;
(mse1=mean(diff1));
(rmse1=sqrt(mse1));
```

Case Study 3 : R code

```
############## graph #############################################
g1=gam(logparr~s(dep), data=z1);
g2=gam(logparr~s(h), data=z1);
g3=gam(logparr~s(atime), data=z1);
g4=gam(logparr~s(ar_day), data=z1);
g5=gam(logparr~s(ar_mon), data=z1);

x11(); par(mfrow=c(2,3));
plot(dep, fitted(g1));
plot(h, fitted(g2));
plot(atime, fitted(g3));
plot(ar_day, fitted(g4));
plot(ar_mon, fitted(g5));
```

Reading lists

1) Hastie, T.J., Tibshirani, R.J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.*SAS/ETS Software: Applications Guide 1*.

2) Stone, C.J. (1985), "Additive Regression and Other Nonparametric Models, "Annals of Statistics, 13, 689-705.

3) Kim, M.S. (2011). A comparison of seasonal linear models and seasonal ARIMA models for forecasting intra-day call arrivals, *The Korean Communications in Statistics*, 18, 237-244.

4) Ruppert, D., Wand, M.P., Carroll, R.J. (2003) *Semiparametric Regression. Cambridge University Press, UK.*

5) Kim, M.S. (2016). Analysis of short-term forecasting for flight arrival time