GAM

2024-03-21

Contents


```
knitr::opts_chunk$set(
  collapse = TRUE,
  warning = FALSE,
  message = FALSE,
  fig.dim = c(10, 5),
  fig.format = "png")
```

Load Data and Package

```
library(tidyverse)
library(caret)
## Load the the training/test set & control method

# Load the training and test sets
train_data <- read.csv("./Data/train_data.csv")
test_data <- read.csv("./Data/test_data.csv")

# Load the control method
ctrl1 <- readRDS("./Data/train_control.rds")

# change variables to be factors again
train_data <- train_data %>%
  mutate(gender = as_factor(gender),
         diabetes = as_factor(diabetes),
         hypertension = as_factor(hypertension),
         vaccine = as_factor(vaccine),
         severity = as_factor(severity))

test_data <- test_data %>%
  mutate(gender = as_factor(gender),
         diabetes = as_factor(diabetes),
         hypertension = as_factor(hypertension),
         vaccine = as_factor(vaccine),
         severity = as_factor(severity))
```

GAM

- **Family and Link Function**: The model assumes a Gaussian family for the error distribution and uses an identity link function. This is typical for a regression problem where the outcome is continuous.

Formula: The outcome is modeled as a linear combination of several covariates, including gender, hypertension, diabetes, and others. Additionally, smooth terms for age, SBP, LDL, BMI, height, and weight are included to capture non-linear effects.

GCV Score: The Generalized Cross-Validation (GCV) score is 0.1915381, which assesses the model's predictive performance, with lower values indicating better fit.

Parametric Coefficients: The table lists the estimated coefficients for each predictor, with their standard errors, t-values, and p-values. For instance, the coefficient for 'gender1' is -0.081585, and it is statistically significant at the 0.001 level (indicated by ***), suggesting a strong relationship with the outcome.

Smooth Terms Significance: This section presents the significance of smooth terms, with 's(LDL)' and 's(BMI)' being highly significant ($p < 0.001$), indicating important non-linear relationships with the outcome.

Model Performance Metrics: - The adjusted R-squared value is 0.195, indicating that approximately 19.5% of the variance in the outcome is explained by the model. - Deviance explained is 20.3%, which is another way of measuring model fit, similar to R-squared. - The scale estimate is about 0.18951. - The size of the dataset used for the model is 2,402 observations.

In summary, this GAM model includes both linear and non-linear relationships between predictors and the outcome. Some predictors show significant effects on the outcome, with non-linear relationships for variables like BMI and LDL being notably significant. Overall, the model explains a modest proportion of the variance in the outcome, suggesting there may be other factors not included in the model that also influence the outcome.

```r
set.seed(1)

x_train = train_data[1:14]
y_train = train_data$recovery_time

x_test = test_data[1:14]
y_test = test_data$recovery_time

model.gam <- train(x = x_train,
                   y = y_train,
                   method = "gam",
                   #metric = "RMSE", by default
                   trControl = trainControl(method = "cv", number = 10))



summary(model.gam$finalModel)
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ gender + hypertension + diabetes + vaccine + severity +
##     study + smoking + race + s(age) + s(SBP) + s(LDL) + s(bmi) +
##     s(height) + s(weight)
##
## Parametric coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.724572   0.027478 135.547  < 2e-16 ***
## gender1       -0.081585   0.017836  -4.574 5.03e-06 ***
## hypertension1  0.064807   0.027607   2.347    0.019 *
## diabetes1     -0.023631   0.024989  -0.946    0.344
## vaccine1      -0.140549   0.018228  -7.711 1.83e-14 ***
## severity1      0.151900   0.028654   5.301 1.26e-07 ***
## studyB        -0.106678   0.019017  -5.610 2.26e-08 ***
## smoking        0.054644   0.012995   4.205 2.71e-05 ***
## race          -0.006137   0.008316  -0.738    0.461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                 edf Ref.df      F p-value
## s(age)    2.140e-08      9  0.000  0.5768
## s(SBP)    1.096e+00      9  0.214  0.1608
## s(LDL)    6.970e+00      9  1.664  0.0365 *
## s(bmi)    6.761e+00      9 34.624  <2e-16 ***
## s(height) 3.339e-01      9  0.042  0.1622
## s(weight) 1.301e+00      9  0.446  0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.195   Deviance explained = 20.3%
## GCV = 0.19154  Scale est. = 0.18951   n = 2402
#plot(model.gam$finalModel)

# Calculate training RMSE of optimal model
gam_train_RMSE = sqrt(mean((y_train - predict(model.gam))^2))
gam_train_RMSE
## [1] 0.4330114

# Calculate test RMSE of optimal model
test_predictions = predict(model.gam, x_test)

gam_test_RMSE = sqrt(mean((y_test - test_predictions)^2))
gam_test_RMSE
## [1] 0.4298968

# Total Sum of Squares
TSS <- sum((train_data$recovery_time - mean(train_data$recovery_time))^2)

# Residual Sum of Squares
RSS <- sum((train_data$recovery_time - predict(model.gam))^2)

# R^2
R2 <- 1 - (RSS/TSS)
R2
## [1] 0.2034079
```