```
library(ElemStatLearn)
library(splines)
data(SAheart)

SAheart$famhist <- c()
regressors <- names(SAheart[, -9])

# Creates the regression formula for these exercises
formulaCreator <- function(regs, df = NULL)
{
  regText <- paste("chd ~ ns(", regs[1], ", df=", df, ")", sep = "")
  for(i in 2:length(regs))
  {
    regText <- paste(regText, "+ ns(", regs[i], ", df=", df,")", sep="")
  }
  return(as.formula(regText))
}

# a)
set.seed(324)
train <- sort(sample(nrow(SAheart), 300))

cubic.spline.model <- glm(formulaCreator(regressors, 4), data = SAheart[train, ], family
= binomial)
summary(cubic.spline.model)
# only few basis functions have a measurable influence on the probability of the coronary
heart disease
# The variables that have at least one basis function to be significant are
# tobacco(+) ("+" means increasing probability), typea(+), obesity(-) and age(+)

# predictions are 0 if log-ratio < 0 and 1 else
predictions <- ifelse(predict(cubic.spline.model, newdata = SAheart[-train,]) < 0, 0, 1)

# misclassification rate
(res.table <- table(predictions, SAheart$chd[-train]))
1 - sum(diag(res.table)) / sum(res.table)
# 28.40 %

# b) stepwise logistic spline regression

# first, we get a starting point by estimating the whole model and looking for the
significants
cubic.spline.full <- glm(formulaCreator(regressors, 4), data = SAheart, family = binomial)
summary(cubic.spline.full)
# tobacco(+), ldl(+), adiposity(+), typea(+), obesity(-), age(+) have at least one
significant basis function

initial.names <- c("tobacco", "ldl", "adiposity", "typea", "obesity", "age")
cubic.spline.init <- glm(formulaCreator(initial.names, 4), data = SAheart, family =
binomial)

# Two approaches:
# First, I start "in the middle" using the model with all variables as regressors that
had at least one significant basis function.
# Second, I start with the full model and see if the two approaches differ
result.init <- step(cubic.spline.init, scope = cubic.spline.full)
result.full <- step(cubic.spline.full)
c(formula(result.init), formula(result.full))
# We obtain the same result with both methods
# c)
reduced.names <- c("tobacco", "ldl", "typea", "obesity", "age")
cubic.spline.reduced <- glm(formulaCreator(regs = reduced.names, df = 4), data =
SAheart[train, ], family = binomial)

predictions <- ifelse(predict(cubic.spline.reduced, newdata = SAheart[-train,]) < 0, 0, 1)

# misclassification rate
(res.table <- table(predictions, SAheart$chd[-train]))
1 - sum(diag(res.table)) / sum(res.table)
# 26.54%

# d)
cubic.spline.reduced <- glm(formulaCreator(regs = reduced.names, df = 4), data = SAheart,
family = binomial)

tobacco.estim <- ns(SAheart[, "tobacco"], df=4) %*% cubic.spline.reduced$coef[2:5]
ldl.estim <- ns(SAheart[, "ldl"], df=4) %*% cubic.spline.reduced$coef[6:9]
typea.estim <- ns(SAheart[, "typea"], df=4) %*% cubic.spline.reduced$coef[10:13]
obesity.estim <- ns(SAheart[, "obesity"], df=4) %*% cubic.spline.reduced$coef[14:17]
age.estim <- ns(SAheart[, "age"], df=4) %*% cubic.spline.reduced$coef[18:21]

plotdata <- data.frame(c(SAheart[,"age"], SAheart[,"ldl"], SAheart[,"obesity"],
SAheart[,"tobacco"], SAheart[,"typea"]))
names(plotdata) <- "ActualValues"
plotdata$Estimates <- c(age.estim, ldl.estim, obesity.estim, tobacco.estim, typea.estim)
plotdata$Grouping <- as.factor(sort(rep(reduced.names, nrow(SAheart))))

library(ggplot2)
qplot(ActualValues, Estimates, data = plotdata, group = Grouping, color = Grouping, geom
= "line", main = "Estimate of Impact on diseases")

# Interpretation:
# All variables are starting from a certain point with impact zero. As the actual values
grow, the impact on the probability changes
# AGE: the lowest observation is at 15 and from then on, the probability to get chd rises
by the age. It is nearly all the time
#       monotonously growing and just slightly decreasing from ~35 to ~45 which is not
necessarily significant.
# LDL, TOBACCO, TYPEA: All three variables increase the probability for this disease by
higher numbers.
# OBESITY: This one is probably the most interesting variable. I suppose - as the values
vary between ~16 and ~47 - that obesity is
#           measured by the BMI. The plot suggests that a very small BMI increases the
probability for CHD more than a very large BMI. The
#           optimal value for the BMI is around 30, which is in fact not healthy, but
given this data, this method and/or this disease,
#           I would suppose that.
```