Homework 1: Smoothers, Generalized Additive Models, and Storytelling

Harvard CS 109B, Spring 2017

Danqing Wang

2/13/2017

Problem 1: Heart Disease Diagnosis

In this problem, the task is to build a model that can diagnose heart disease for a patient presented with chest pain. The data set is provided in the files `dataset_1_train.txt` and `dataset_1_test.txt`, and contains 6 predictors for each patient, along with the diagnosis from a medical professional.

- By visual inspection, do you find that the predictors are good indicators of heart disease in a patient?

```
train <- read.csv("./CS109b-hw2_q1_datasets/dataset_1_train.txt")
test <- read.csv("./CS109b-hw2_q1_datasets/dataset_1_test.txt")

str(train)
```

```
## 'data.frame':    210 obs. of  7 variables:
##  $ Age         : int  67 37 59 54 58 50 52 54 57 57 ...
##  $ Sex         : int  1 1 1 1 0 0 1 0 1 1 ...
##  $ ChestPain   : Factor w/ 4 levels "asymptomatic",..: 1 2 2 2 1 3 4 2 2 1 ...
##  $ RestBP      : int  160 130 126 150 100 120 118 108 150 132 ...
##  $ ExAng       : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ Thal        : Factor w/ 3 levels "fixed","normal",..: 2 2 1 3 2 2 1 2 3 3 ...
##  $ HeartDisease: Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
```

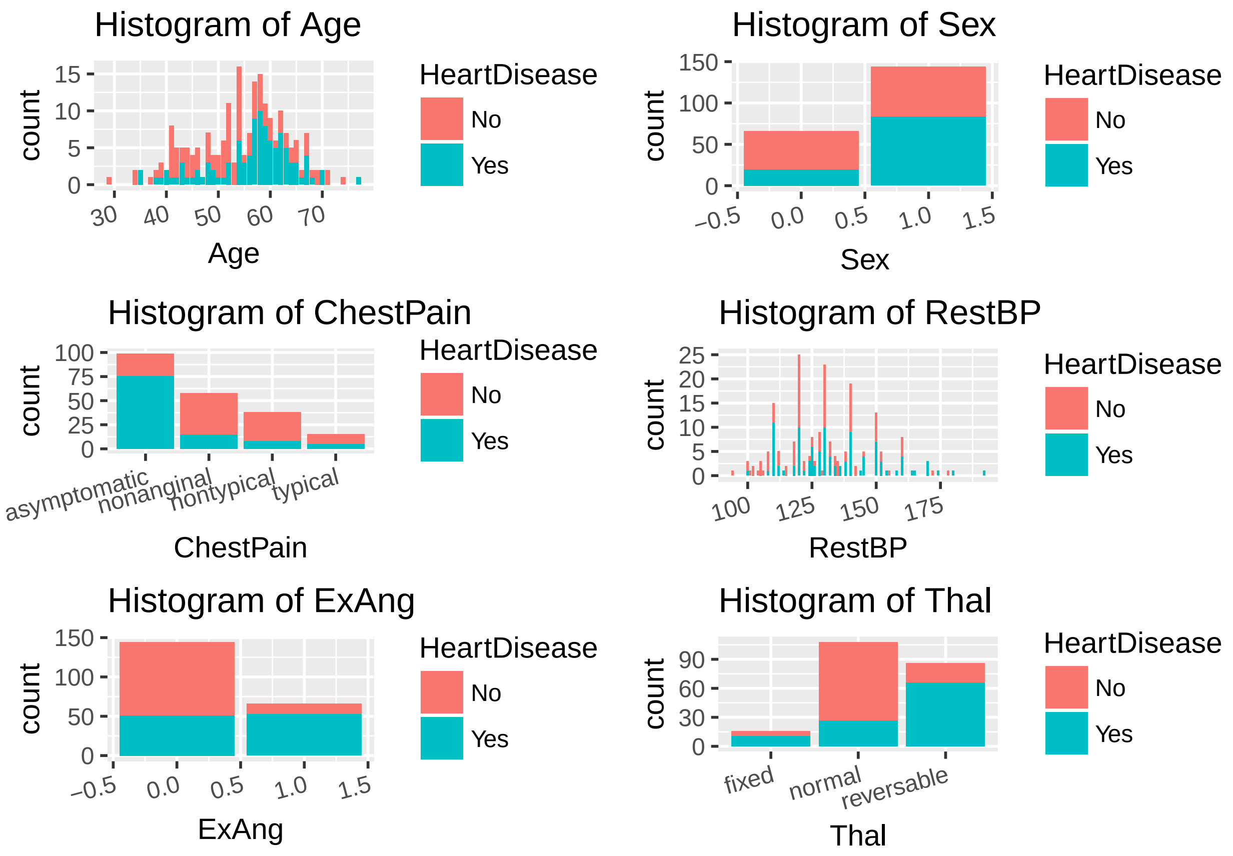
```
str(test)
```

```
## 'data.frame':    91 obs. of  7 variables:
##  $ Age         : int  63 67 67 56 56 48 58 60 66 43 ...
##  $ Sex         : int  1 1 1 1 1 1 1 1 0 1 ...
##  $ ChestPain   : Factor w/ 4 levels "asymptomatic",..: 4 1 1 3 2 3 3 1 4 1 ...
##  $ RestBP      : int  145 160 120 120 130 110 120 130 150 150 ...
##  $ ExAng       : int  0 1 1 0 1 0 0 1 0 0 ...
##  $ Thal        : Factor w/ 3 levels "fixed","normal",..: 1 2 3 2 1 3 2 3 2 2 ...
##  $ HeartDisease: Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 2 1 1 ...
```

The data sets contain the response variable `HeartDisease` and six predictors: `Age`, `Sex`, `ChestPain`, `RestBP`, `ExAng`, and `Thal`. Among which, `Age` and `RestBP` are continuous predictors, while the rest are categorical predictors.

We examine the data by plotting `HeartDisease` against every predictor:

```
library(ggplot2)
library(gridExtra)

predictors <- colnames(train)[1:6]
p <- list()

for(i in predictors){
  p[[i]] <- ggplot(train, aes_string(x = i, fill = 'HeartDisease'))+
    geom_bar()+
    ggtitle(paste("Histogram of", i))+
    theme(axis.text.x = element_text(angle = 15, hjust = 1))
}
do.call(grid.arrange, p)
```



ANSWER From the bar charts, we see that the Age distribution of heart disease patients is fairly normal. However, those with Sex marker 1 are more likely to suffer from heart disease thank those with Sex marker 0. People who have asymptomatic chest pain are more likely to suffer from heart disease, and people with reversable thallium scan result tend to suffer from heart disease. However, the RestBP and the presence of exercise induced angina(ExAng) are not good indicators of whether a person is likely to suffer from heart disease.

- Apply the generalized additive model (GAM) method to fit a binary classification model to the training set and report its classification accuracy on the test set. You may use a smoothing spline basis function wherever relevant, with the smoothing parameter tuned using cross-validation on the training set. Would you be able to apply the smoothing spline basis to categorical predictors? Is there a difference in the way you would handle categorical attributes in `R` compared to `sklearn` in `Python`?

We first transform the response variable `HeartDisease` into 0 and 1 in both the `train` and the `test` sets:

```
levels(train$HeartDisease) <- c(0, 1)
levels(test$HeartDisease) <- c(0, 1)
```

ANSWER Unlike the case in `sklearn` of python, we do not need to convert categorical variables to dummy variables of 1 and 0 here, R recognizes them as categorical (here: `ChestPain`, `Thal`). However, the R is recognizing `Sex` and `ExAng` as integers instead of factors, we should make them into factors using the `factor()` function.

Cross Validation function:

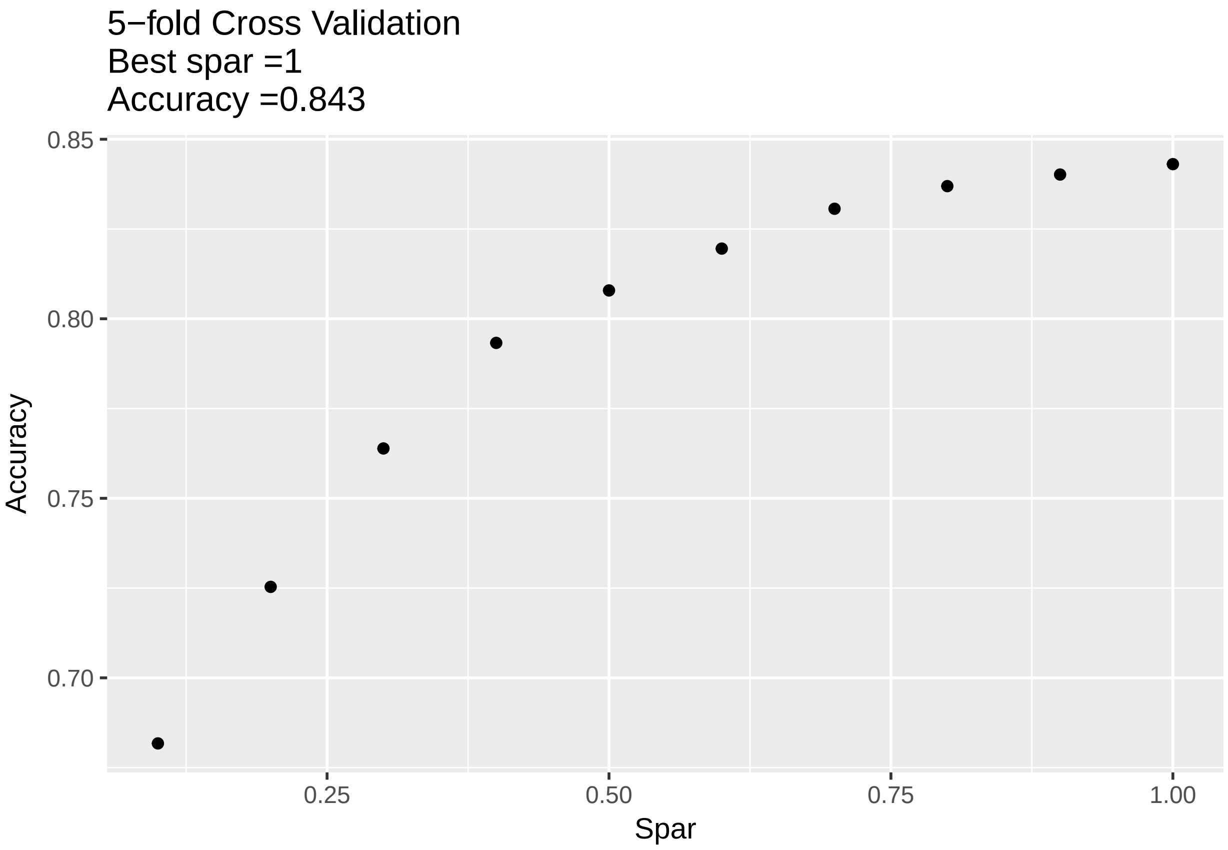
```
library(boot)
# Function to compute k-fold cross-validation accuracy for a given classification model
cv_accuracy = function(model, data, k) {
  # Input:
  #   'model' - a fitted classification model
  #   'data' - data frame with training data set used to fit the model
  #   'k' - number of folds for CV
  # Output:
  #   'cv_accuracy' - cross-validation accuracy for the model
  set.seed(109)
  acc <- 1 - cv.glm(data, model, K = k)$delta[1]
  return(acc) }
```

Using cross validation, we determine the best spar value to use using a GAM model, with smooth function applied onto the continuous variables `Age` and `RestBP`.

```
library(gam)
```

```
## Loading required package: splines

## Loading required package: foreach

## Loaded gam 1.14
```

```
# Set of spar values
param_val <- seq(0.1, 1, by = 0.1)

# Perform 5 fold cross validation to find the best spar
num_param <- length(param_val)
cv_score <- rep(0., num_param)

for(i in 1:num_param){
  mod_formula = as.formula(paste0('HeartDisease ~ s(Age, spar =', param_val[i],') +
                      factor(Sex) +
                      ChestPain +
                      s(RestBP, spar =', param_val[i],')+
                      factor(ExAng) +
                      Thal'))

  mod.gam <- gam(mod_formula, data = train, family = binomial(link="logit"))

  cv_score[i] = cv_accuracy(mod.gam, train, 5)
}

# Write cv_score as a dataframe
cv_score <- data.frame(Spar = param_val, Accuracy = cv_score)
spars.best <- cv_score$Spar[which(cv_score$Accuracy == max(cv_score$Accuracy))]
spars.best.score <- max(cv_score$Accuracy)

# Visualization of cv_score
ggplot(cv_score, aes(x = Spar, y = Accuracy))+
  geom_point()+
  ggtitle(paste0("5-fold Cross Validation \nBest spar =", spars.best,
                 "\nAccuracy =", round(spars.best.score, 3)))
```



ANSWER The best spar value from 5-fold cross-validation is 1, with an accuracy of 0.843. The high spar value means we are 'smoothing' over all of the datapoints, which produces a very broad smooth.

Using this value, we predict on the test dataset:

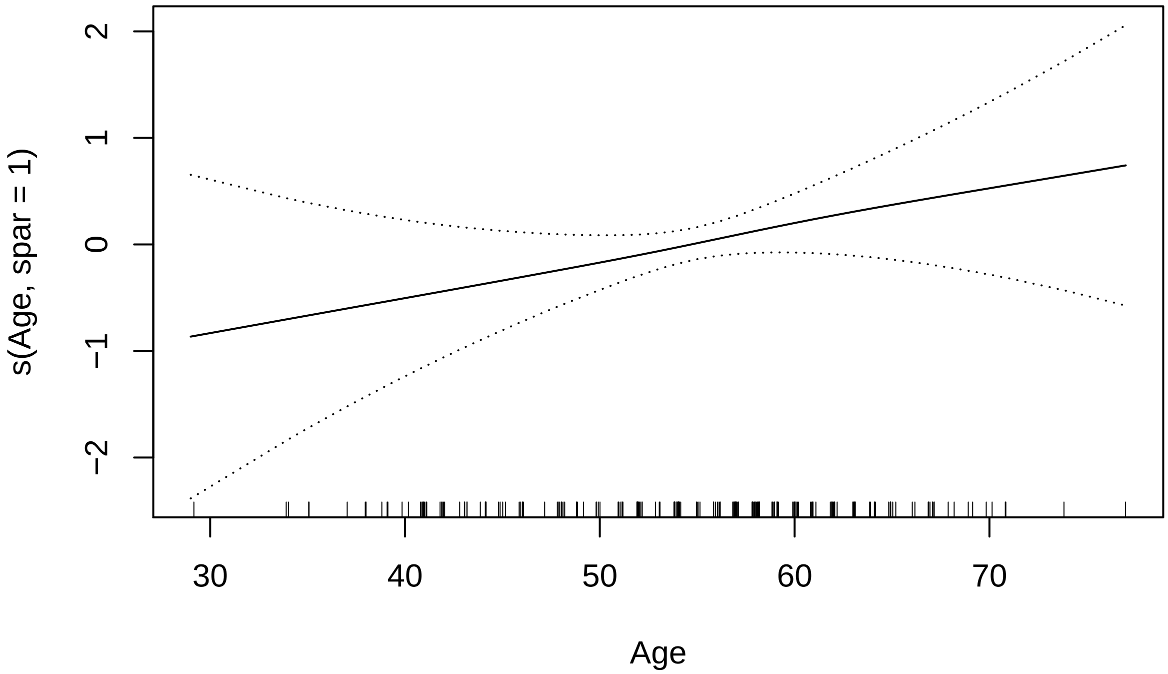
```
# train model
mod_formula <- as.formula(paste0('HeartDisease ~ s(Age, spar =', 1,') +
                          factor(Sex) +
                          ChestPain +
                          s(RestBP, spar =', 1,')+
                          factor(ExAng) +
                          Thal'))

mod.gam <- gam(mod_formula, data = train, family = binomial(link="logit"))

# predict on test set
pred <- round(predict(mod.gam, newdata = test, type="response"))
test.accuracy <- mean(test$HeartDisease == pred)

print(paste('The test accuracy is', test.accuracy))
```

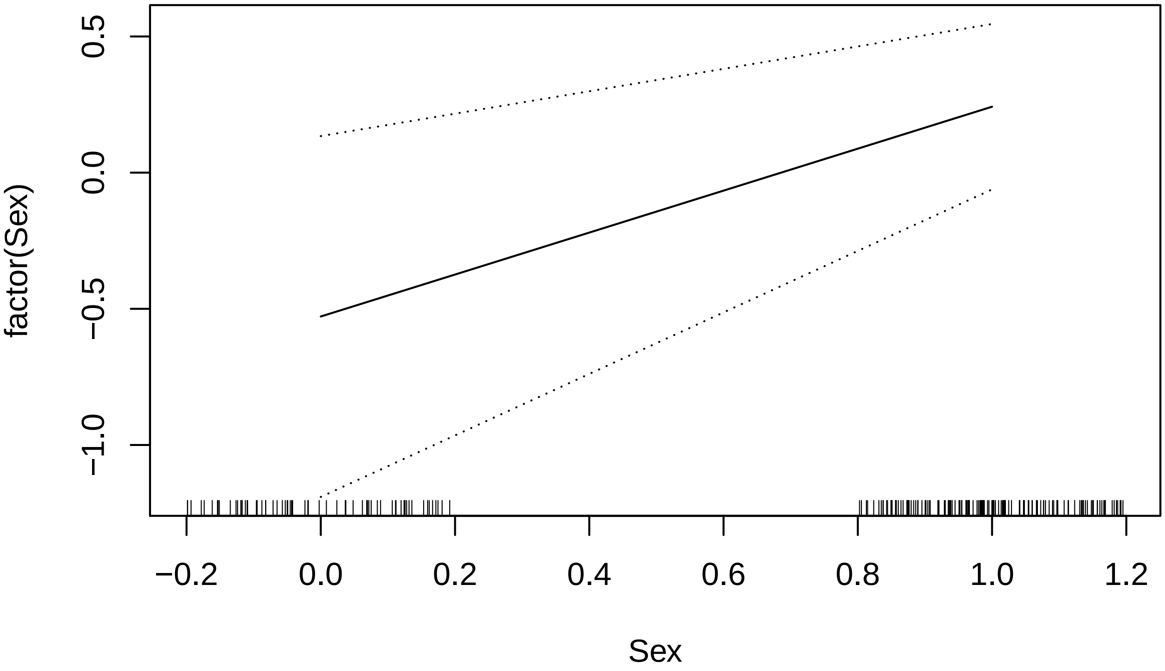
```
## [1] "The test accuracy is 0.813186813186813"
```

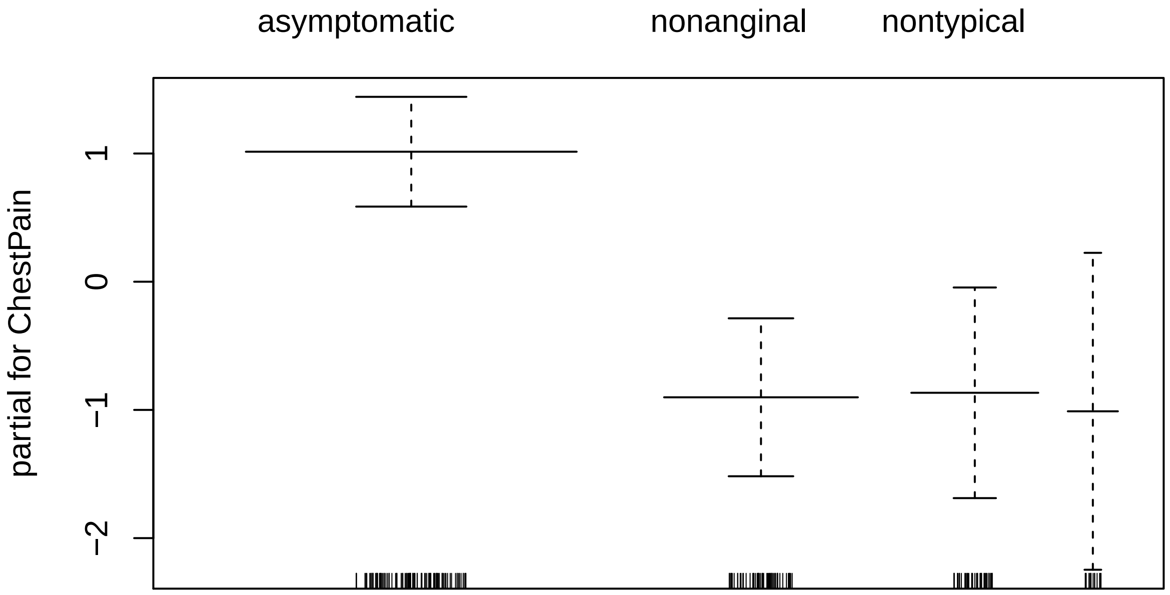
ANSWER The accuracy on the test set is 0.813 with a best spar value of 1.

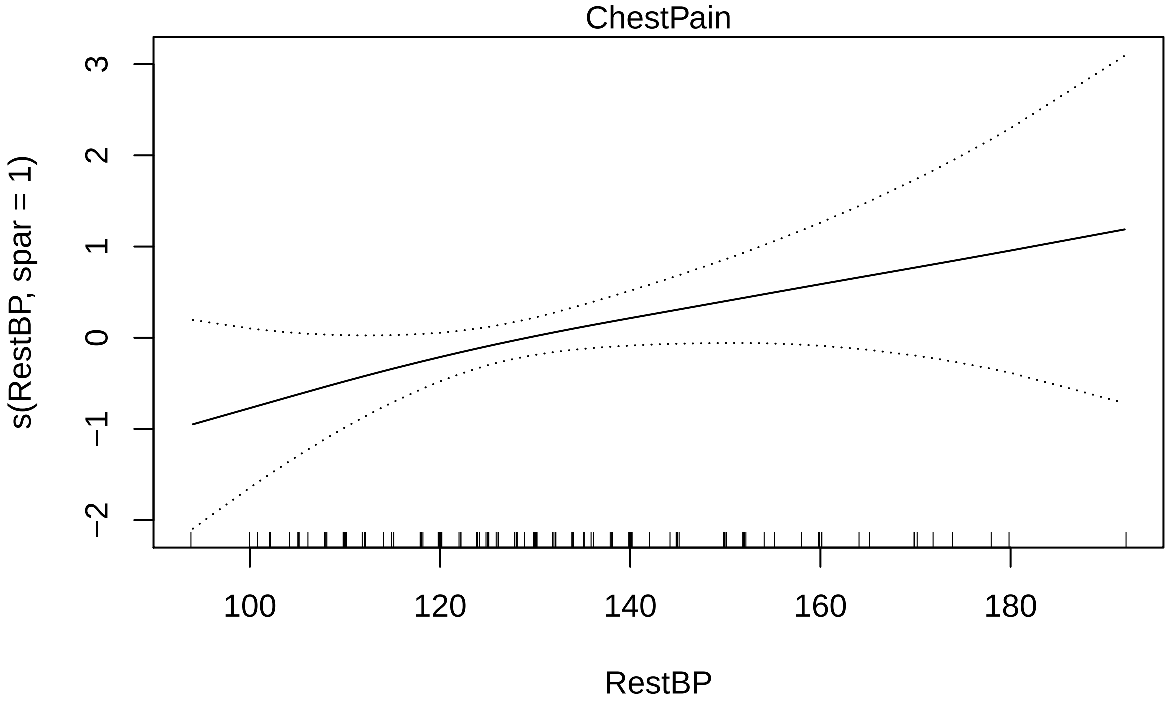
- Plot the smooth of each predictor for the fitted GAM. By visual inspection, do you find any benefit in modeling the numerical predictors using smoothing splines?

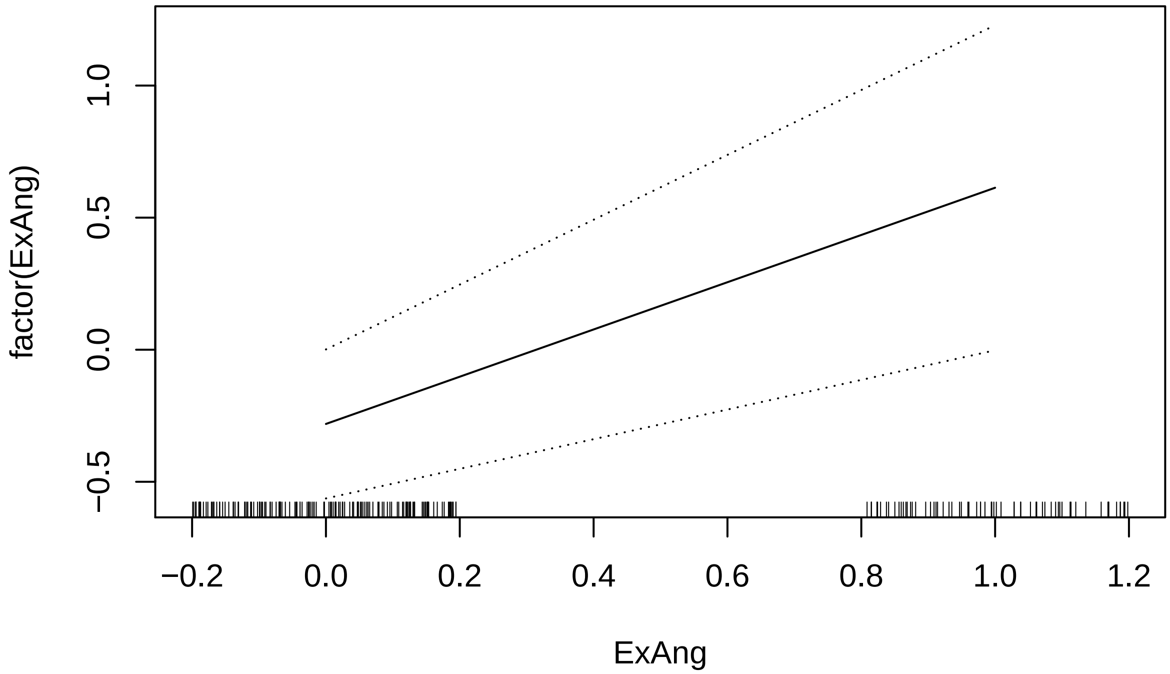
```
plot(mod.gam, se=TRUE)
```

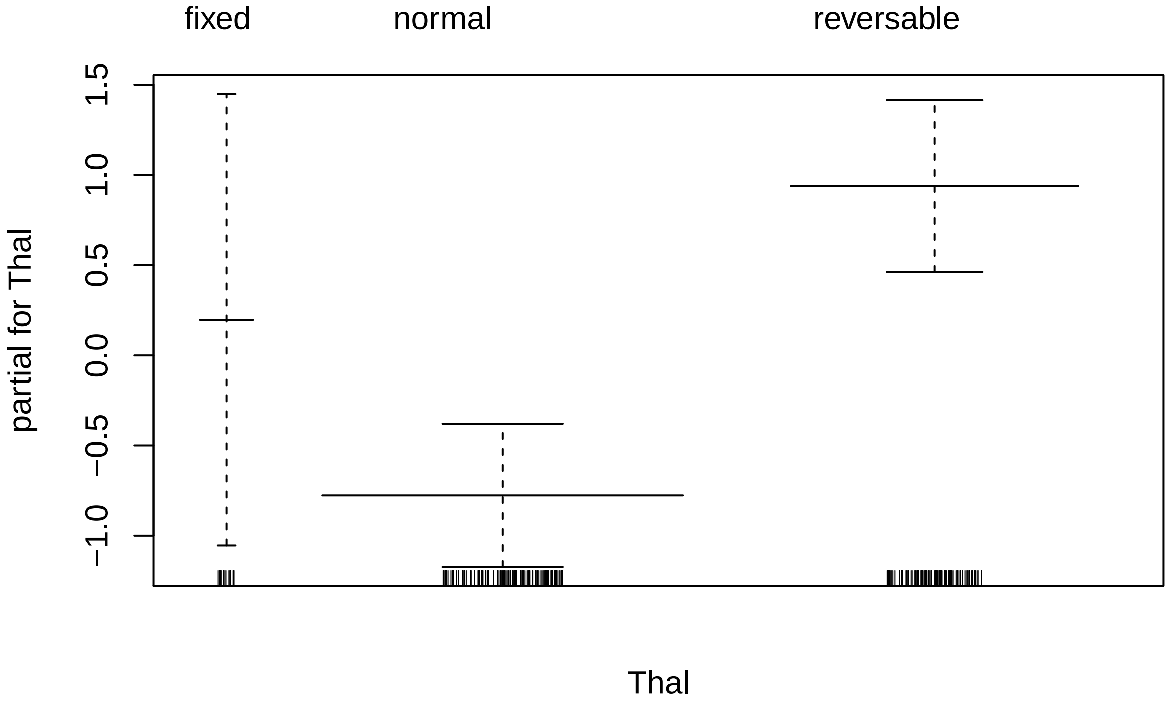

s(Age, spar = 1)
2
1
0
−1
−2
30
40
50
60
70
Age


factor(Sex)
0.5
0.0
−0.5
−1.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Sex


asymptomatic
nonanginal
nontypical
partial for ChestPain
1
0
−1
−2


ChestPain
s(RestBP, spar = 1)
3
2
1
0
−1
−2
100
120
140
160
180
RestBP





ANSWER Smooth spline cannot be applied to categorical variables, but only quantitative variables, which in this case are `Age` and `RestBP`. If the fitted smooth is nonlinear in a GAM, then there is a clear advantage to using smoothing spline terms as opposed to linear terms. However, here the quantitative variables `Age` and `RestBP` have linear relationships with the response variable, there is no spline. This indicates that we do not need to apply smooth functions on these two variables either.

- Using a likelihood ratio test, compare the fitted GAM with the following models: (i) a GAM with only the intercept term; (ii) a GAM with only categorical predictors; and (iii) a GAM with all predictors entered linearly.

(i) GAM with only the intercept term

DO WE NEED TO USE CROSS VALIDATION FOR THESE?

```
mod.intercept <- gam(HeartDisease ~ 1, data = train,
                     family = binomial(link = "logit"))
anova(mod.intercept, mod.gam, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: HeartDisease ~ 1
## Model 2: HeartDisease ~ s(Age, spar = 1) + factor(Sex) + ChestPain + s(RestBP,
##     spar = 1) + factor(ExAng) + Thal
##   Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
## 1    209.00     291.10
## 2    199.26     181.08 9.7418   110.02 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANSWER The p-value of 2.2e-16 is smaller than 0.001, this means the model with all predictors (with smooth spline basis functions) is better at significance level of 0.001.

(ii) GAM with only categorical predictors

```
mod.cat <- gam(HeartDisease ~ Sex + ChestPain +
                 ExAng + Thal, data = train,
               family = binomial(link="logit"))

anova(mod.cat, mod.gam, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: HeartDisease ~ Sex + ChestPain + ExAng + Thal
## Model 2: HeartDisease ~ s(Age, spar = 1) + factor(Sex) + ChestPain + s(RestBP,
##     spar = 1) + factor(ExAng) + Thal
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    202.00     189.75
## 2    199.26     181.08 2.7418   8.6727  0.02729 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANSWER The p-value of 0.027 is smaller than 0.05, this means the model with all predictors (with smooth spline basis functions) is better at significance level of 0.05.

(iii) GAM with all predictors entered linearly.

```
mod.linear <- gam(HeartDisease ~ Age + Sex + ChestPain +
                    RestBP + ExAng + Thal, data = train,
                  family = binomial(link="logit"))

anova(mod.linear, mod.gam, test = "Chi")
```

```
## Analysis of Deviance Table
```

```
##
## Model 1: HeartDisease ~ Age + Sex + ChestPain + RestBP + ExAng + Thal
## Model 2: HeartDisease ~ s(Age, spar = 1) + factor(Sex) + ChestPain + s(RestBP,
##     spar = 1) + factor(ExAng) + Thal
##   Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
## 1    200.00     181.62
## 2    199.26     181.08 0.74182  0.53772   0.3557
```

ANSWER The p-value of 0.3557 is larger than 0.05, we fail to reject the null hypothesis and conclude that the GAM model with all predictors entered linearly is better. This is also consistent with our earlier linear plots that there isn't any smooth splines present.

Problem 2: The Malaria Report

You work for the Gotham Times media organization and have been tasked to write a short report on the World Health Organisation's (WHO) fight against malaria. The WHO Global Malaria Programme (http://www.who.int/malaria/en/) has been working to eliminate the deadly disease over the past several decades, your job is to discuss their work and spotlight the impact they've had. Your writing and graphics should be easily understood by anyone interested in the topic, and not necessarily just physicians and experts.

Key Facts and Quotes on Malaria

Here are some informative key facts and quotes about Malaria that you may want to include in your report:

- RISK: About 3.2 billion people – almost half of the world's population – are at risk of malaria.
- CASES: 214 million malaria cases reported worldwide in 2015.
- INCIDENCE: 37% global decrease in malaria incidence between 2000 and 2015. (Malaria incidence is defined as number of new cases reported in a given time period / the number of people at risk.)
- MORTALITY: 60% decrease in global malaria mortality rates between 2000 and 2015.
- "Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected female mosquitoes."
- "Young children, pregnant women and non-immune travelers from malaria-free areas are particularly vulnerable to the disease when they become infected."
- "Malaria is preventable and curable, and increased efforts are dramatically reducing the malaria burden in many places."

Many of these facts were pulled from the WHO website, where you can find many more.

The Data

The datasets consist of country-level information for 2015, estimated malaria cases over time, and funding values and sources over time.

Dataset 1: data/global-malaria-2015.csv This dataset contains observed and suspected malaria cases as well as other detailed country-level information **for 2015** in 100 countries worldwide. The CSV file consists of the following fields:

- WHO_region, Country, Country Code, UN_population
- At_risk - % of population at risk
- At_high_risk - % of population at high risk
- Suspected_malaria_cases
- Malaria_cases - actual diagnosed cases

Dataset 2: data/global-malaria-2000-2013.csv This dataset contains information about suspected number of malaria cases in the same 100 countries for the years 2000, 2005, 2010, 2013.

Dataset 3: data/global-funding.csv This dataset contains the total funding for malaria control and elimination (in millions USD) provided by donor governments, multilateral organizations, and domestic sources between 2005 and 2013.

Dataset 4: data/africa.topo.json The TopoJSON file (extension of GeoJSON) contains the data of the boundaries for the African countries.

You can also explore the very large database provided by the WHO, though a bit of manual processing may be needed.

Exploratory Data Analysis

Your first task is to use this data for exploratory data analysis (EDA) and to create several visualizations (e.g., bar graphs, line charts, scatterplots, maps) using Tableau or ggplot2. It may also be useful to reshape the data in R before visualizing it – take a look at the R code at the end of this `Rmd` document.

You will notice some regional discrepancies, keep these in mind as you explore the data and gain an understanding of what the data are saying. Try to identify a few key messages that can be supported by the data.

Planning

In planning your report, think about the many facets of information contained in the data, and explore which of the visualziations are most effective to illustrate some key messages. You are free to decide what data you want to use and what kind of story you would like to tell. You may use statistical modeling techniques (e.g., linear regression) if you think they are appropriate, but must explain and justify their use. Consider the visualization and storytelling principles we discussed in class.

Your Report

Your report should have a catchy title and be 500-600 words in length (not more!) with at least two different visualizations, including titles and captions. Structure your report using balanced design and make sure to start with a global overview with context, need, task, and message. The text and visualizations should mutually reinforce your messages through effective redunancy.

Your visualizations must be effective and well designed. Push yourself a little. If you are a Tableau/ggplot2 beginner edit them carefully based on the principles we discussed in class. If you are a Tableau/ggplot2 expert try to do something special that goes beyond the usual graphs and charts.

Submission

Your report can be written in your tool of choice (Word, Google Docs, Markdown, etc.) and should look professionally designed. Upload the report as a PDF into the homework folder. In addition, submit a separate PDF with **all** of the EDA visualizations you created and any code you used.

Grading Criteria

We will grade your report using the following criteria:

- Is the report informative, accurate, and engaging?

- Are the messages clearly expressed?
- Is the writing appropriate for the target audience, the readers of the Gotham Times?
- Are the visualizations effective, accurate, and easy to understand?
- Are the scales, axis, labels, and color maps appropriate?
- Do the titles and captions for the visualizations tell the reader what the message of each visualization is?
- Does the report accurately highlight some interesting facts about the data?
- Does the report follow the storytelling principles discussed in class?

After we reorganize the data, we load them in the following:

```r
m_byyear <- read.csv('./CS109b-hw2_q2_datasets/global-malaria-byYear.csv')
f_byyear <- read.csv('./CS109b-hw2_q2_datasets/global-funding-byYear.csv')
m_2015 <- read.csv('./CS109b-hw2_q2_datasets/global-malaria-2015.csv')

str(m_byyear)
```

```
## 'data.frame':    490 obs. of  6 variables:
##  $ X                        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Code                     : Factor w/ 98 levels "AFG","AGO","ARG",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Country.x                : Factor w/ 98 levels "Afghanistan",..: 1 3 4 5 14 8 13 6 7 10 ...
##  $ WHO_region               : Factor w/ 6 levels "African","Eastern Mediterranean",..: 2 1 4 3 1 1 1 5
##  $ year                     : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ Estimated_Malaria_Counts: int  1500000 4800000 240 1700 2800000 2700000 7200000 3100000 1700 4800
```

```r
str(f_byyear)
```

```
## 'data.frame':    63 obs. of  4 variables:
##  $ X     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Source: Factor w/ 7 levels "All Other Sources",..: 3 6 2 5 7 1 4 3 6 2 ...
##  $ year  : int  2005 2005 2005 2005 2005 2005 2005 2006 2006 2006 ...
##  $ Amount: num  308.2 110.4 435.7 NA 15.4 ...
```

```r
str(m_2015)
```

```
## 'data.frame':    98 obs. of  8 variables:
##  $ X                      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ WHO_region             : Factor w/ 6 levels "African","Eastern Mediterranean",..: 2 1 4 3 1 1 1 5
##  $ Country                : Factor w/ 98 levels "Afghanistan",..: 1 3 4 5 14 8 13 6 7 10 ...
##  $ Code                   : Factor w/ 98 levels "AFG","AGO","ARG",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ UN_population          : int  31627506 24227524 42980026 9629779 10816860 10598482 17589198 15907
##  $ At_risk                : num  75.6 100 NA NA 100 ...
##  $ At_high_risk           : num  26.9 100 NA NA 100 ...
##  $ Suspected_malaria_cases: int  390000 3100000 NA NA 1400000 3200000 7000000 8400 50 9900 ...
```

We examine the rest in Tableau.