

Assignment 1: Build a predictive model based on Galton Family Data

YANG Can

2020/2/14

Data description: The Galton Families Data Set

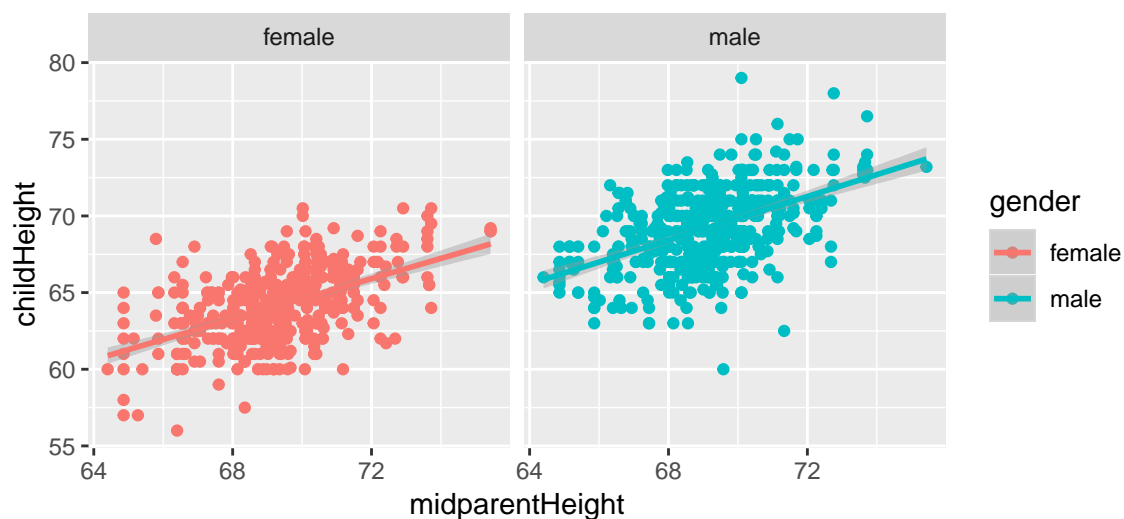
The Galton family height data is accessed by R package `HistData`, where the mid-parent heights are calculated by $(\text{father} + 1.08 * \text{mother}) / 2$. You may use `?GaltonFamilies` to check more details of the data set.

```
library(HistData)
data(GaltonFamilies)
head(GaltonFamilies)
```

```
##   family father mother midparentHeight children childNum gender
## 1    001   78.5   67.0         75.43         4         1  male
## 2    001   78.5   67.0         75.43         4         2 female
## 3    001   78.5   67.0         75.43         4         3 female
## 4    001   78.5   67.0         75.43         4         4 female
## 5    002   75.5   66.5         73.66         4         1  male
## 6    002   75.5   66.5         73.66         4         2  male
##   childHeight
## 1          73.2
## 2          69.2
## 3          69.0
## 4          69.0
## 5          73.5
## 6          72.5
```

The relationship between the mid-parent heights and children's heights can be visualized as

```
library(ggplot2)
ggplot(data=GaltonFamilies, aes(x=midparentHeight, y=childHeight, color=gender)) +
  facet_wrap(~gender) + geom_point() + geom_smooth(method="lm")
```



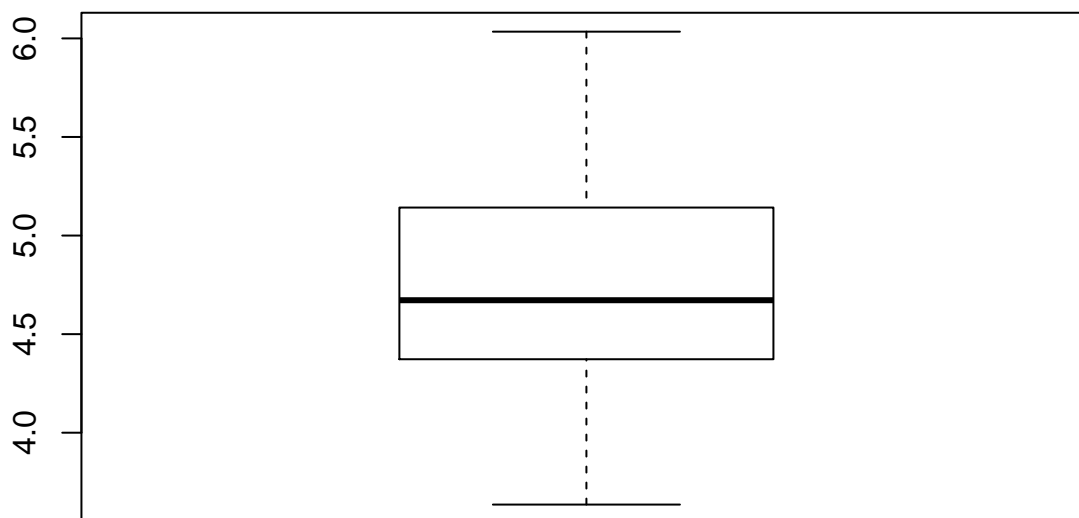
An example of the predictive model

In this assignment, you are asked to fit a model to predict `childHeight` based on the `GaltonFamilies` data set. Specifically, you need to partition the whole data set into **training set** and **testing set** (Suppose you randomly select 200 children as the test set and the remaining samples are used for training). Let's take a **standard linear regression model** as an example. The linear model is fitted using the **training set**, and its performance is evaluated on the **testing set**. Due to the randomness of partition, you should summarize your comparison based on 50 replications. (Hint: the prediction performance can be evaluated by mean-square error $MSE = \frac{1}{200} \sum_{i=1}^{200} (y_i - \hat{y}_i)^2$, where y_i is a childheight in the test set and \hat{y}_i is the predicted value from your model). The code is given as follows:

```
nrep <- 50
n <- dim(GaltonFamilies)[1]
ntest <- 200
MSE <- data.frame(matrix(0,nrep,1))
names(MSE) <- c("lm")
for (i in 1:nrep){
  # partition training data and testing data
  train <- sample(1:n, n-ntest, replace = FALSE)
  traindata <- GaltonFamilies[train,]
  testdata <- GaltonFamilies[-train,]

  # Fit a baseline model
  fit_baseline <- lm(childHeight~gender + midparentHeight, traindata)
  # make prediction based on the fitted model
  height_baseline_pred <- predict(fit_baseline,testdata)
  # Evaluate model performance
  MSE[i,1] <- mean((testdata$childHeight - height_baseline_pred)^2)

  # Warning: You should not use childNum as a predictor in your model
  # because children within a family are listed in decreasing order of height
  # for boys followed by girls
}
boxplot(MSE)
```



```
head(MSE) #the mean squared errors of the baseline model
```

```
##          lm
```

```
## 1 6.034165
## 2 4.904788
## 3 5.263207
## 4 5.478551
## 5 4.147113
## 6 5.055878
```

Your task

- Try your best to build your own model such that it can be better than the standard linear regression model in terms of the prediction accuracy (measured by the mean squared error and evaluated in the same way as the above code). Note that you should **NOT** use the variable `childNum` in your prediction model because children within a family are listed in decreasing order of height for boys followed by girls.

Requirement

- You need to submit a report, in which you should clearly describe your method and explain your idea. The code should also be included.
- You can use R or Python for coding. You are also **allowed to call packages** in R or Python to do this project as long as you understand the chosen method.
- Your report should be in the **pdf** format, which is automatically generated by either R markdown or Jupyter notebook.
- The report is due to March 4 (Mid night), 2020.