HW 1

Zack Shin

2022-09-11

#Question 2  
  
#here, we assign the value of reading the CSV dataset to the object called hw1  
  
hw1 <- read.csv("HW1data.csv")  
  
#here, the str function is used to gather the answers to the first question, those being observation and variable count, variable names, and variable types.  
  
str(hw1)

## 'data.frame': 32 obs. of 3 variables:  
## $ Wgt : int 2940 3130 2420 2450 2760 2440 3226 3301 2729 3410 ...  
## $ Gest : int 38 38 36 34 39 35 40 42 37 40 ...  
## $ Smoke: chr "Y" "N" "Y" "N" ...

#In the dataset HW1data we have 32 observations and 3 variables. These are Wgt, Gest, and Smoke. The former two variables are both quantitative, while the latter variable, Smoke, is categorical in nature.

#Question 3  
  
#We should not use the same observations to fit the models as we should to compare the quality of the fitted models because if we were to do so, the model would always look perfect. Using different sets of testing data will ensure that the model works for all types of data.

#Question 4  
  
#Prediction is used when an input is readily available, but an output for that input is not. In this scenario, an example of prediction would be using either model to estimate the number of dollars that would be received in sales for some given dollar amount spent on TV ads.  
  
#Inference involves the estimation of the relationship between the variables in an experiment. In the context of this example, inference would look like examining the results of the experiment and seeing how the number of dollars spent on advertising affects the number of dollars received in sales for that same day. For instance, if the data followed a positive, linear relationship, it could be surmised that an increase in dollars spent on advertising is correlated with an increase in dollars received from sales for a particular day.   
  
#Reducible error is error which arises from the fact that the model is merely an estimate of the function, and not the function itself. It is called reducible because there are other models that could more accurately estimate the true function. In the context of this example, reducible error would occur since it is almost impossible within the confines of the study to match the model to the true function 100%.   
  
#Irreducible error is error which exists regardless of how accurate the model is to the true function. It is an "upper limit" on the accuracy of the model. In the mathematical model, it is represented as c, or sometimes as ε. In the context of this example, irreducible error would exist in the model inherently.

#Question 5  
  
#(a) In this example, we have 100 irises, and we are interested in finding out if we can predict sepal length from petal lengths. This would be an example of a supervised learning method. This is because we are interested in using the independent variables to try and find a relationship to some dependent variable, in this case, the sepal length.  
  
#(b) In this example, using the same irises, we are now measuring both sepal length and petal length, and simply grouping the observations into iris type. Because there is no dependent variable, and we are only working on organizing the independent variables, this is an example of unsupervised learning methods.

#Question 6  
  
#Variance and bias of a model are two terms that go into calculating the estimated value of a particular variable. Bias reflects the ability for the functional form of the model to get close to the true relationship between the predictors and the outcome. In the context of the above example, bias would mean how well is the predictive model about to get close to the true relationship between the predictor and response variables.   
#Variance, on the other hand, is how much the model varies from theoretical sample to sample. In the context of the dollars example, variance would denote how much the model changes between samples.