New York University Shanghai

Problem Set - Week 1

BUSF-SHU 210: Business Analytics (Spring 2019)

1. Calculate the R^2 for the linear regression models with different covariates in the Wine Analytics case. Specifically, consider the following models:

- Price∼Harvest Rain
- Price∼Age
- Price∼Winter Rain
- Price∼Population

Which model has the best in-sample fit?

2. Student Assessment Scores

In this problem, you will predict the reading scores of students on the 2009 Assessment.

The dataset `Assessment_train.csv` contains information about the students taking the exam.

Each row in the dataset `Assessment_train.csv` represents one student taking the Assessment. The dataset has the following variables:

- *grade*: The grade in school of the student (most 15-year-olds in America are in 10th grade)
- *male*: Whether the student is male (1/0)
- *raceeth*: The race/ethnicity composite of the student
- *preschool*: Whether the student attended preschool (1/0)
- *expectBachelors*: Whether the student expects to obtain a bachelor's degree (1/0)
- *motherHS*: Whether the student's mother completed high school (1/0)
- *motherBachelors*: Whether the student's mother obtained a bachelor's degree (1/0)
- *motherWork*: Whether the student's mother has part-time or full-time work (1/0)
- *fatherHS*: Whether the student's father completed high school (1/0)
- *fatherBachelors*: Whether the student's father obtained a bachelor's degree (1/0)
- *fatherWork*: Whether the student's father has part-time or full-time work (1/0)
- *selfBornUS*: Whether the student was born in the United States of America (1/0)

- $motherBornUS$: Whether the student's mother was born in the United States of America (1/0)
- $fatherBornUS$: Whether the student's father was born in the United States of America (1/0)
- $englishAtHome$: Whether the student speaks English at home (1/0)
- $computerForSchoolwork$: Whether the student has access to a computer for schoolwork (1/0)
- $read30MinsADay$: Whether the student reads for pleasure for 30 minutes/day (1/0)
- $minutesPerWeekEnglish$: The number of minutes per week the student spend in English class
- $studentsInEnglish$: The number of students in this student's English class at school
- $schoolHasLibrary$: Whether this student's school has a library (1/0)
- $publicSchool$: Whether this student attends a public school (1/0)
- $urban$: Whether this student's school is in an urban area (1/0)
- $schoolSize$: The number of students in this student's school
- $readingScore$: The student's reading score, on a 1000-point scale

There are some missing values in this data set, which we need to remove before building the linear regression model. Use the function `na.omit()` to remove the missing values in both the training data and the testing data. To apply this function, you may type: `AssesmentTrain = na.omit(AssessmentTrain)` in the R console.

(a) Factor variables are variables that take on a discrete set of values. An ordered factor has a natural ordering between the levels (an example would be the classifications "large," "medium," and "small"). Which of the variables in the Assessment data set are unordered factors with at least three levels? Which of the variables are ordered factors with at least three levels?

(b) To include unordered factors in a linear regression model, we define one level as the "reference level" and add a binary variable for each of the remaining levels. In this way, a factor with n levels is replaced by $n-1$ binary variables. The reference level is typically selected to be the most frequently occurring level in the dataset.

As an example, consider the unordered factor variable "place", with levels "Shanghai", "Beijing", and "Guangzhou". If "Beijing" were the reference level, then we would add binary variables "placeShanghai" and "placeGuangzhou" to a linear regression problem. All Shanghai examples would have placeShanghai=1 and placeGuangzhou=0. All Guangzhou examples would have placeShanghai=0 and placeGuangzhou=1. All Beijing examples would have placeShanghai=0 and placeGuangzhou=0.

Now, consider the variable "raceeth" in our problem, which has levels "American Indian/Alaska Native", "Asian", "Black", "Hispanic", "More than one race", "Native Hawaiian/Other Pacific Islander", and "White". Because it is the most common in the data set, we will select White as the reference level. Which binary variables should be included in the regression model? For a student who is Asian, which of the binary variables you created above should be set to 0, which of them should be set to 1?

(c) Because the race variable takes on text values, by default R selects the first level alphabetically ("American Indian/Alaska Native") as the reference level, instead of the most common level ("White"). To reset the reference level of the factor, we can use the following two commands in the R console: `AssessmentTrain$raceeth = relevel(AssessmentTrain$raceeth, "White")`. Build a linear regression model to predict the reading score of students using all remaining variables as covariates.

Please show a screen shot of your linear regression model using the "summary" function. Find the independent variables that are significant (with $p-$value, $Pr(>|t|)$, smaller than 0.05). Consider two students A and B. They have all variable values the same, except that student A is in grade 11 and student B is in grade 9. What is the predicted reading score of student A minus the predicted reading score of student B? Please interpret the estimated coefficient of the variable "Asian".

(d) If we remove all the independent variables that are not significant (with $p-$value, $Pr(>|t|)$, no smaller than 0.05), we can build a new linear regression model. Report the (in-sample) R^2's for both the original model and the model with insignificant variables removed.