STA363_Client_Challenge_1

Alex Zhang

2024-02-02

Part 1

I will not recommend just removing 71 students from the data and proceed. The first thing is that by deleting these students, we also reduced our sample size by about 11%. Generally we should keep every student because everyone in this sample could mean certain pattern in the population. The second thing is that we may lose the representative nature of our sample. It is possible that these student miss this question for a reason. They may not heard ChatGPT before either because they do not have a computer at home or they are too young to know technology. We cannot delete these students because then we assume every student in this sample knows ChatGPT and eventually overestimate the impact of ChatGPT on middle school students.

Part 2

The corrected code is:

```r
# Create new column to store the imputations values
middleschool[, "CountUsedChatGPTonHW_New"] <- middleschool[ , "CountUsedChatGPTonHW"]

# Train the regression model
model <- lm( CountUsedChatGPTonHW ~ ComputerHome + Age, data = middleschool)

# Replace the missing value on row 3 with our predicted value
middleschool[3, "CountUsedChatGPTonHW_New"] <- predict( model, newdata = middleschool[3, ])
```

The Colleague B is trying to use what calls regression imputation on handling the missing data. Basically we try to train a statistical model we have chosen based on the information we know. In this case, we think there is a linear relationship between the count of using ChatGPT on homework and student's age and whether his home has a computer. We would use these two features to create a linear model and use this model to predict the missing number of count in row 3.