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. We may not be able to capture more patterns in this sample to represent the whole population. The third thing is that I think this data is missing for a reason. Some student may never heard or use ChatGPT before. This is either they do not have computer at home or just no one tells them. If we delete these students, our sample maybe bias on overestimating the impact of ChatGPT on middle school students.

Part 2

The corrected code is:

```
# Create new column to store the imputations value
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. we then decided to use a linear regression model to impute the values of missing data. In this model, we chose the count number to be our response variable and we tried to parameterize the relationship between count and student's age and whether his home has a computer. After we have our trained model, we use the model to obtain predictions for the missing values and fill that predicted value into row 3.