Statistical and Predictive Modeling I (DATA 1204)
Final Project (30% of Total Grade)
Professor: Fatma Tetikoglu

Background

Mr. John Hughes has been collecting data on the effect of personal attributes on household expenses. He has put together a dataset (**MultiRegDataset.csv**) with contains 1338 observations (rows) and 7 features (columns). The details of the features are as follows:

Independent (Input) variables:

- Age
- Sex
- BMI
- Children
- Smoker
- Region

Dependent (Output) variable:

- Expenses

The Ask from Mr. John Hughes

He would like to understand the following:

a) The effect of **smoking** on expenses by creating a linear regression model
b) The effect of **all input variables** on expenses by creating a multivariate regression model

You will create the following:

1. A power point deck to report your findings and state your conclusion based on your results. (See Appendix A for details)

2. Copy of the R code that you used to generate the results (cut and paste into Word Document)

Please post your PowerPoint (.ppt) and Word Document (.doc or .docx) containing all your R code under Final Project by 11:59 pm on Tuesday, December 13th, 2022.

Appendix A

PowerPoint Requirements:

Cover Slide

- Title: Final Project (DATA 1204)
- Name (First and Last)
- Student Number

Slide 1

- Description of the research requirements (i.e., the ask from Mr. John Hughes)

Slide 2-4

- Compute and state the basic statistics (i.e., Mean, SD, Min/Max). Please explain your findings (Hint: Don’t forget to use actual numbers).
- Create and show a fully labeled Histogram of the dependent variable(expenses). Please explain your findings.

Slides 5-7

- Conduct a T-test that the mean for expenses is equal to 10,000:
  - ✓ State the hypotheses related to the test
  - ✓ State and explain the results of your T-Test

Slides 8-9

- Perform a simple linear regression **using smoker** as your independent variable and expenses as your dependent variable
  - ✓ State the simple linear regression model
  - ✓ Interpret the simple linear regression model
  - ✓ Evaluate the simple linear regression model

Slides 10-11

- Perform a multiple linear regression on all variables and report the results
  - ✓ State the multiple linear regression model
  - ✓ Interpret the multiple regression model
  - ✓ Evaluate the multiple regression model

Slide 12-13

- State your conclusions based on evidence from your analysis

Word Document Requirements:

1. All R code used in report

Final Term Project Rubric

Slides	Exemplary	Proficient	Incomplete	Incorrect or Unacceptable
1	Clear description of the research question is given.	Mostly Clear description of the research question is given.	Incomplete description of the research question is given.	Description of research problem is incorrect or missing.
2-4	Histogram is correct, properly labeled and explained. Statistics computed are correct and meaningful.	Histogram is correct, properly labeled and explained. Statistics computed are mostly correct and meaningful.	Histogram is is mostly correct. Statistics computed are mostly correct and meaningful.	Histogram and some statistics are not correct.
5-7	Results or the t-test are reported correctly. Assumptions that need to be satisfied is clearly stated along with whether they were satisfied. Hypotheses are clearly stated and correct.	Results or the t-test are reported correctly. Assumptions stated are correct and an explanation of whether they were satisfied is mostly correct. Hypotheses are clearly stated and mostly correct	Results or the t-test are reported correctly. Assumptions are incomplete and the explanation is also incomplete. Hypotheses are incomplete	Results of the t-test are incorrect. Hypotheses are missing or incorrect
8-9	Simple linear regression is performed correctly and reported correctly. Coefficients and evaluations are interpreted correctly with detail.	Simple linear regression is performed correctly and reported correctly. Coefficients and evaluations are interpreted correctly with limited detail	Simple linear regression is performed correctly and reported correctly. There are some issues with coefficient and evaluation interpretation	Simple linear regression is incorrect and subsequently all other answers are also incorrect
10-11	Multiple linear regression is performed correctly and reported correctly. Coefficients and evaluations are interpreted correctly with detail.	Multiple linear regression is performed correctly and reported correctly with limited detail.	Multiple linear regression is performed correctly and reported correctly.	Multiple linear regression is incorrect and subsequently all other answers are also incorrect
12-13	Conclusions about the research question are clearly stated and correct. Evidence for the conclusions is presented clearly.	Conclusions about the research question are clearly stated and correct. Evidence for the conclusions is mostly presented clearly.	Conclusions about the research question are clearly stated and correct. Evidence for the conclusions is incomplete.	Conclusions are incorrect or poorly stated
R Code	All R code is clearly stated and correct.	All R code are clearly stated and mostly correct	R code is correct but is incomplete	R code missing or incorrect