IE266 Engineering Statistics
Case Study II

Due date: 26.06.2022 Sunday, 23:59

In answering all questions, please state your assumptions clearly.

Cigarette smoking is the leading cause of preventable death in the world. More than 7 million people die every year worldwide due to tobacco use. Smoking among youth is an issue that affects countries worldwide, mainly developing countries. Approximately 90% of smokers begin smoking before the age of 18. In recent years, several actions were taken to discourage people from smoking such as high taxation, banning the advertisement of cigarettes, placing warning images on cigarette packs, and banning smoking indoors and in public places.

In an attempt to better understand how smoking at a young age affects lung health and cigarette addiction later in life, a survey and a medical examination are conducted on 1853 people between the ages of 30-35 who smoke, which includes 8 questions. The survey questions are:

- How long did you smoke before you were 18 years old? (in years)
- When you were in high school, what was your family's monthly household income? (in today's Turkish liras)
- What is your biological sex? (male or female)
- When you were in high school, how many of your family members smoked? (Mother, father, siblings) (Zero, One, Two, More than 3)
- When you were in high school, how many of your friends in your close friend group smoked? (Zero, One, Two, More than 3)
- Had anyone in your family warned you about the hazards of smoking? (0: No, 1: Yes)
- Have you ever been informed about the hazards of smoking in any of your courses in high school? (0:No, 1: Yes)
- Have you seen anti-tobacco messages on TV/on billboards/in newspapers? (0: No, 1: Yes)

To quantify the level of addiction and harm caused by smoking, the lung capacity of the responders was measured by the medical examination and recorded.

A researcher wishes to test the effects of duration of smoking, income, biological sex, smoking habits of family and close friends, and exposure to warnings about the hazards of smoking on addiction and lung health. To do this, he wishes to construct a linear regression model for lung capacity as a response variable and the survey questions as independent variables.

1. Test the following claims, using contingency tables and chi-square tests, at $\alpha=0.05$ by clearly stating the hypotheses, performing the test, and clearly stating the conclusions with the p-values: **(Bonus 5 points)**

(a) Duration of smoking and gender are independent.

(b) The number of smokers in the family and the number of smokers in the friend group are independent.

2. Construct the linear regression model using all variables. Then perform the following steps:

(a) Plot the response against the continuous variables with groups of significant categorical variables to see if any interaction terms are necessary and add them to the model.

(b) Construct the linear regression model using the stepwise regression methods (backward elimination, forward selection, etc). Explaining every selection/elimination step clearly and at every step provide the value of adjusted R^2.

(c) Check whether the transformation of the response variable is necessary.

<u>At every step make sure the assumptions of the model are verified.</u>

3. Construct a 95% confidence interval on the mean response for a respondent :

- who smoked for 1.65 years during high school
- whose family had a monthly household income of 12500 liras
- is a male
- had 1 family member who smoked when s/he was a student
- had 2 close friends who smoked
- had been warned about the hazards of smoking by family
- had not taken a course that explained the hazards of smoking
- had seen anti-tobacco messages on TV/on billboards/in newspapers

4. Construct a 95% prediction interval for the response of the following respondent :

- who smoked for 3.35 years during high school
- whose family had a monthly household income of 10350 liras
- is a female
- had 2 family members who smoked when s/he was a student
- had more than 3 close friends who smoked
- had not been warned about the hazards of smoking by family
- had taken a course that explained the hazards of smoking
- had seen anti-tobacco messages on TV/on billboards/in newspapers

Format and Organization:

- Please write in proper font size 12 and 1.5 paragraph spacing with reasonable margins.
- You do not need to include an introduction, conclusion, and appendix section in the report.
- Number and title report sections properly.
- The format and organization of the report will be considered in grading.
- This homework is for a group of four. Working in collaboration with other groups is not allowed. You can discuss the problem with your partners only.
- Submit your work under Case Study 2 on ODTÜClass. Upload a single zip file, including your report and all the files you use to answer the questions (R, and Excel files).
- You should use R to carry out your statistical analysis.
- Please use comments in your R scripts to make your codes readable.
- Prepare your report so that the reader wouldn't have to run the code. Put all graphs, necessary analysis in the report (do not include the code in the report)