STA356L (Statistical Computing-II Lab)

Lab Assignment-I

Lab Tasks

The Obesity Dataset (ObesityData.csv) by Palechor & de la Hoz Manotas (2019)[1] for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico includes the following variables:

Gender, Age, Height, Weight, Family history with overweight, Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH2O), Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Consumption of alcohol (CALC) and Transportation used (MTRANS).

For each of the following tasks, you must provide the R-codes, Results (answer of the questions only; not all the R outputs) and Interpretation of the results.

1. Read the given dataset (ObesityData.csv) into R.

2. Calculate frequencies, percentages and cumulative frequencies of all the categorical variables.

3. Present the categorical variables using suitable graphs (e.g., horizontal and vertical bar diagram, pie chart, etc.). [Figures must be clear and contain all the necessary components such as title of the figure, titles of the horizontal and vertical axis, values in each axis, legend, etc.]

4. Calculate appropriate summary statistics (e.g., minimum, maximum, mean, median, mode, 1st quartile, 3rd quartile, standard deviation, variance, coefficient of variation, interquartile range, coefficient of quartile deviation, etc.) for each of the numerical variables.

5. Present the numerical variables using suitable graphs (e.g., histogram, frequency curve, histogram with normal probability curve, box plot, stem and leaf plot, etc.).

6. Explain which summary measures and graphs are appropriate to present each individual variable and why. [Hints: test normality for each of the numerical variables. If it follows normal distribution, present it using mean, standard deviation/variance, coefficient of variation, histogram, etc. and if it does not follow normal distribution, present it using median, interquartile range, coefficient of quartile deviation, box plot, etc.]

7. Perform exploratory subgroup analysis: calculate summary statistics and draw graphs for each numerical variable by every categorical variable. For example, calculate minimum, maximum, mean, median, mode, 1st quartile, 3rd quartile, standard deviation, variance,

coefficient of variation, interquartile range, coefficient of quartile deviation of age, height, weight, FCVC, CH2O, FAF, TUE separately for every level of the categorical variables such as separately for male and female, smoker and non-smoker, … so on.

8. Draw a scatter diagram between (i) age and weight, (ii) age and height and (iii) height and weight. Draw respective regression lines on the scatterplots. Also, draw a scatterplot matrix of all the numerical variables. Comment on your results.

9. Recode the variable MTRANS into MTRANS_RC in which both the 'Walking' and 'Bike' categories will be 'Ownself' and all other categories will be 'Car'. Also, convert the variable FCVC to a factor (name as FCVC_factor) by labeling 1 as 'Never', 2 as 'Sometimes' and 3 as 'Always'.

10. Calculate body mass index (BMI) by using the formula, BMI = Weight/(Height)^2. Make a categorical variable (name as BMI_cat) using the following categorization of BMI values:
    Less than 18.5 as 'Underweight'; 18.5 to 24.9 as 'Normal'; 25.0 to 29.9 as 'Overweight'; Greater than 30 as 'Obesity'.

11. Calculate BMI for respondents (i) whose age > 30 years, (ii) who are non-smokers, have physical activity of 2 days and drink more than 1 liter of water daily.

12. Create a new dataset (name as: obesity_sub) by taking the respondents whose height is more than 1.8 meter and who eat high caloric food frequently. Calculate mean and standard deviation of BMI using the obesity_sub dataset.

13. Calculate correlation between (i) age and weight, (ii) age and height and (iii) height and weight. Compare and contrast these results with the results of task number 8.

14. Calculate correlation between age and BMI and comment on the relationship. Does this correlation significantly differ from zero? Calculate correlation matrix of all the numerical variables. Comment on your results.

15. Using appropriate method, test whether the respondent's (i) average age is equal to 30 years, (ii) average height is greater than to 1.7 meters, (iii) average consumption of water daily (CH2O) is equal to 2 liters and (iv) average BMI is less than 30. [Hints on appropriate method: among parametric and non-parametric tests, which is applicable in each individual question? If parametric, then z-test or t-test? Why?]

16. Using appropriate methods, test whether the average age, height and BMI of the respondents significantly differ between (i) male and female, and (ii) smoker and non-smoker. Also, test whether the BMI value is higher for the respondents with family history of overweight and lower for those who monitor their calorie consumption. [Hints on appropriate method: among parametric and non-parametric tests, which is applicable in each individual question? If parametric, then z-test or t-test? Why?]

17. Using appropriate methods, test whether there are significant variations in the average age, height and BMI of the respondents among different groups of CAEC, CALC and MTRANS. Also, identify which pairs of the group means of each variable significantly differ (perform multiple comparison). [Hints on appropriate method: among parametric and non-parametric tests, which is applicable in each individual question and why?]

18. Using appropriate method, test whether there is significant association between BMI and (i) gender, (ii) family history with overweight, (iii) smoking, (iv) FAVC, (v) CAEC, (vi) SCC, (vii) CALC and (viii) MTRANS. Report the contingency tables, row/column/total percentages and table of expected cell count. [Hints on appropriate method: among asymptotic and exact tests, which is applicable in each individual question and why?]

19. Create a new dataset (give name as: obesity_small) from the obesity dataset keeping the variables only: Gender, Age, Height, Weight, family history with overweight, FAVC, CH20, SCC, FAF, TUE and CALC. Create another dataset (namely: obesity_new) by dropping the variables FAF, TUE and CALC from the obesity_small dataset. Using the obesity_new dataset, fit a multiple linear regression model of Weight on Gender, Age, Height, family history with overweight, FAVC, CH20 and SCC. Interpret the outputs of the model including estimates of the parameters. How much variation of Weight is explained by the explanatory variables of your model (Multiple coefficient of determination R^2)? Predict the weight of a 23 years old male respondent having family history with overweight, Height = 1.77, FAVC = yes, CH20 = 1 and SCC = no.

20. (i) From the Obesity Dataset (ObesityData.csv), identify the significant factors affecting BMI. To do this, first, perform the model building/variable selection using a suitable method. Then, present the final fitted model including estimates of the parameters, their confidence intervals, p-values. Comment on your findings. Comment on the goodness of fit of the final model/How well the model fitted with the data (F statistics with its p-value, R^2, adjusted R^2, And AIC Value)? Find the predicted values and the residuals.

    (ii) Perform model adequacy checking (check model assumptions) and examine whether each of the model assumptions is satisfied or not. Give justification of your answer. Propose alternative solutions if any model assumption is not satisfied.

    (iii) Find if there is any outlier in the data. Also, discuss whether the outliers (if any) are influential or not.

21. Draw a random sample of size n = 50 from the Obesity Dataset (ObesityData.csv) using the last three digits of your registration number as seed number. Repeat the tasks 10-14 and the task 16 using this sample. Discuss if any different methods you had to apply to perform the assigned tasks on this sample. Also, compare the results of the analysis of sampled data and original data.

*For all the tasks involving tests of hypotheses, write the null hypothesis and alternative hypothesis, the value of test statistic, 95% confidence intervals (CI), P-value and comments.

[1] Palechor, F. M., & de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in brief, 25, 104344.