Exercise 1. (5 points) Execute the code below to create the temporary dataset ***extract***. The ID variable contains a state abbreviation as the first two digits and a sex code in the 5th position.

data extract;

input id;

datalines;

MA40M99

CO76F4512

TX91M5

;

run;

a. We want to extract the state code and the sex code. Write a SAS data step that includes the SUBSTR function to accomplish these tasks. What is the length assigned to the new variables? (4 points)

b. Repeat step a. but use a length statement to define a length of 2 for the state code and a length of one for the sex code. (1 point)

Exercise 2. (6 points) The dataset ***air.sas7bdat*** contains the series of monthly numbers of international airline travelers from January 1949 to December 1960.

|  |  |  |
| --- | --- | --- |
| Variable | Values | Notes |
| Date | Date formatted as MONYY5. | Month and year |
| Air | Numeric | international airline travel (in thousands) |

1. Format the variable date as MONYY7 and print the first 5 observations. (1 point)
2. Add two new variables month and year in the airline dataset by extracting the month and year from the date variable. (1 point)
3. Test if there is evidence of a difference in the mean number of travelers across months. As usual, you will report the test statistic, distribution under the null with df, p value, decision, and conclusion. (2 points)
4. Test if there is evidence of a difference in the mean number of travelers across years. As usual, you will report the test statistic, distribution under the null with df, p value, decision, and conclusion. (2 points)

Exercise 3. (9 points) In the National Health and Nutrition Examination Survey (NHANES), adult respondents self-reported if they have diabetes and if a close relative – including father, mother, sisters or brothers— had diabetes. In addition, measured weight and height were used to calculate BMI. We are interested in assessing if BMI and diabetes in a close relative are associated with the probability of having diabetes. The data are in the SAS dataset ***diabetes.sas7bdat***.

|  |  |  |
| --- | --- | --- |
| Variable | Values | Notes |
| Id | Numeric | Unique identifier |
| Bmi | Numeric | Body mass index (BMI) |
| Diabetes | 1 = Yes  0 = No | Indicator of diabetes |
| family\_risk | 1 = Yes  0 = No | Indicator of close relative with diabetes |

1. Fit the adequate multivariable model to assess the association between BMI and diabetes in a close relative with the probability of having diabetes. Interpret the odds ratio associated with BMI in this model. (2 points)
2. Use PROC STANDARD to standardize the BMI variable (mean of 0 and standard deviation of 1) and output the new variable into a temporary dataset called ***diabetes2***. (1.5 point)
3. Fit the model again to assess the association between standardized BMI and diabetes in a close relative with the probability of having diabetes. Interpret the odds ratio associated with standardized BMI. (2 points)
4. Use PROC RANK with the dataset ***diabetes*** to create quartile groups based on BMI. Name the new rank variable bmigroup and output the new variable into a temporary dataset called ***diabetes3***. (1.5 point)
5. Fit the model again to assess the association between categorized BMI and diabetes in a close relative with the probability of having diabetes. You will use the lowest quartile group of BMI as the reference level. Interpret the odds ratios associated with categorized BMI.(2 points)