**Exam of statistical methods for geographical and social environmental (simulation)**

1. In 2018 was presented a research about multiple regression analysis of main economic indicators in tourism in a European country (during the period 2010-2017). The variables involved into the study were:

*Y -* is travel and tourism total contribution to GDP (in billion Euro)

*X*1 - is capital investment (in billion Euro)

*X*2 - is arrivals of foreigners by means of transport (in thousands)

*X*3 - is overnights of foreigners and native in hotels (in thousands)

The author provided the correlation matrix (Tab.1) and the results for the multiple regression model (Tab.2).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Y | *X*1 | *X*2 | *X*3 |
| Y | 1 |  |  |  |
| *X*1 | 0.824 | 1 |  |  |
| *X*2 | 0.834 | 0.623 | 1 |  |
| *X*3 | 0.712 | 0.521 | 0.621 | 1 |

*Tab.1 The correlation matrix*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *X*1 | *X*2 | *X*3 |
| Coefficient | 2.112 | 0.0854 | 0.0721 |
| SSE | 26,543.401 |  |  |
| TSS | 50,867.87 |  |  |

*Tab.2 The results for multiple regression model.*

1. Comment in general the values in correlation matrix and in particular the correlation between *X1* and *X3*.
2. Given that the constant is equal to -4.1882, write the estimated multiple linear regression model and explain the estimated parameters.
3. How well do the explanatory variables in the model predict *y,* using the prediction equation?

2) The contingency table and the comparison using ratios

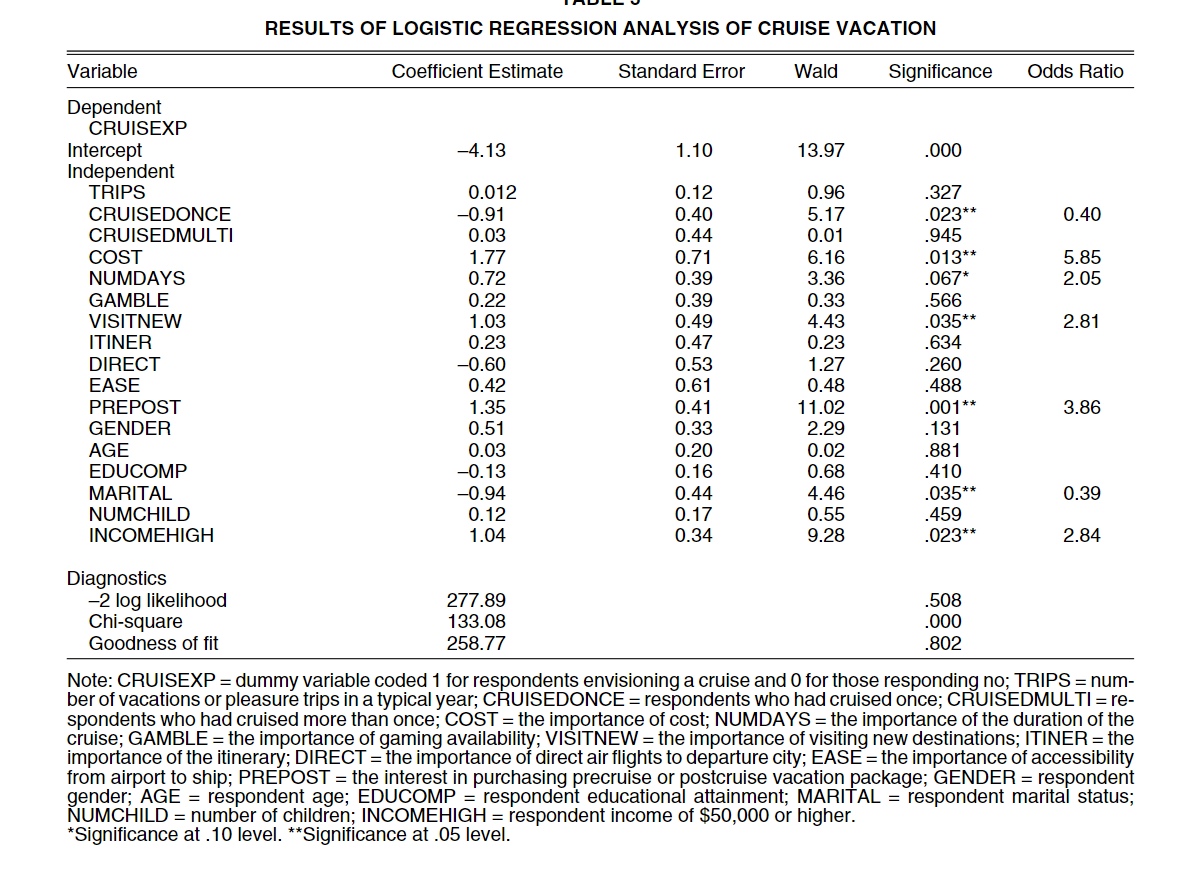
The following table reports the distribution of a sample of Italian population (aged more than 18 years old) by the Use of Train for Travel (Y) and Gender (X).

|  |  |  |
| --- | --- | --- |
| **X**  **Y** | **Male** | **Female** |
| Yes Train | 152 | 115 |
| No Train | 319 | 423 |

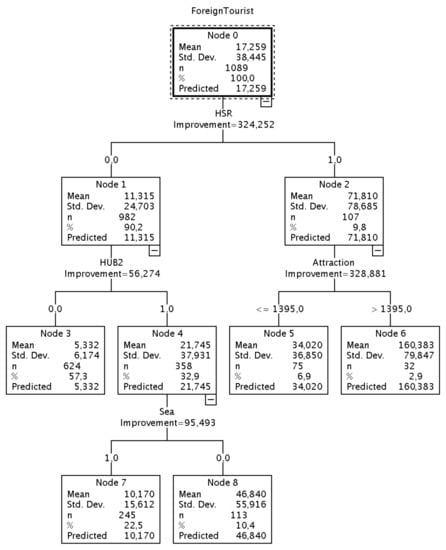
1. Are X and Y stochastically independent? Justify the answer.
2. Compare proportions who report being *No Train* respect to Gender.
3. Considering the Success: Y=”Yes Train” and Failure: Y=’No Train’, compute the odds ratio and comment the result.

3) The logistic regression and the interpretation of the parameters.

Comment the results of the logistic regression for the following real case, paying attention only on the significant independent variables.



4) The regression tree and the steps in the hierarchical segmentation.



Comment the regression tree.

The real case is based on a dataset dealing with information regarding the tourists’ arrivals as well as transport modes for the 99 Italian provinces, observed during the 2006–2016 time period. Therefore, 1089 observations (99 provinces x 11 years) have been collected.

The dependent variable considered is:

ForeignTourist: no. of Foreign Tourists; number of arrivals from other countries travelling for both tourism and business tourism purposes

The independent variables are:

* Transportation systems variables
* HSR is a dummy variable assuming Value 1 if the HSR is present, 0 if otherwise;
* HUB2 is a dummy variable assuming Value 1 if the airport is not a first level hub; 0 if otherwise;
* LowCost: no. of operating bases of low-cost airlines.
* Attractiveness variables
* GDP is the Gross Domestic Product of the province
* Attraction: is the no. of activities in a given province
* Sea is a dummy variable assuming Value 1 if the province is close to the sea; 0 if otherwise;
* POP is the number of inhabitants in a given province
* Unemployment: percentage of unemployed in a given province