Statistical Analysis on Italian Trips

A project presented by:

Salma Chabib [872519]

Lilia Grasso [813210]

Alice Schiavone [872388]

Data Science

University of Milano-Bicocca

Milan

25/12/2020

Abstract

In 2020, the whole world stopped from travelling for a year, due to a global pandemic. This lead to a critically negative effect on the economy and on the relationships of people that relied on holidays to visit their loved ones. In order to plan a new start for Italian people and allow them to travel again, this paper aims to analyse how Italians moved in the country or abroad in 2019, and sets objectives to reach again in the following years, when we hope to go back to our normal way of living.

Using micro-data made available by ISTAT, it is possible to describe where Italian people spend their time (for vacancy or business reasons) and how much they are willing to spend for it, assisted by statistical tools and methodologies. These can also estimate values for the entire population, starting from a sample of less than 5000 units, and find correlations between the different characteristics of each person's trip. The results presented at the end of this research paper were somehow expected but it is always better to highlight some known problems of the Italian population.

Keywords: travelling, trips, Italy, statistics

I. Introduction

A. Data Set Description

The following analysis is based on the open data set provided by ISTAT [1] (Italian National Institute of Statistics), the main producer and provider of high quality statistics in Italy. The original survey [2] target was to collect information about Italians Households Budget, that tracks how Italian people spend money, from January 1 to December 31 of every year. To analyze Italian people trips in 2019, the used data set of micro-data is a focus of the original survey, that intends to put the attention on this specific topic. This is an annual survey, starting from 2014, but ISTAT has been collecting this type of data since 1964, adapting the survey format to the evolving standards (currently following Europeans standards of ECoicop, which stands for European Classification of Individual Consumption by Purpose). Istat consideres a trip anything that move people out of their 'usual envoironment', for a vacation or for business reasons, with or without overnight stays. People moving daily or weekly for business, study or personal reasons are not considered tourists. The international standard classifies trips by reason (business or vacation) and by the number of stays (short trip from 1 to 3 overnight stays, long trips for more than 3 nights).

The reference population is made by resident families. By family we mean a set of people living in the same household, bound by marriage or other type of family connection, or that contribute to the family household budget. Therefore, are excluded from the reference population people living permanently in communities (military barracks, hospitals, orphanages, religious institutes, etc.) and those staying but not resident in the national territory.

The data set provided by ISTAT is made by a wide range of information through 4393 observations and 52 attributes. While having a sufficient set of values to do statistical analysis, the number of observations of the population doesn't come anywhere near the real population of Italy (more than 60 million people).

B. Sampling Description

The survey was designed to provide data at regional and national level, as Italy was divided into 5 geografical macroareas, made by its 20 administrative subdivisions:

- Nord-ovest : Piemonte, Valle d'Aosta, Lombardia, Liguria;
- Nord-est: Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna;
- Centro: Toscana, Umbria, Marche, Lazio;
- Sud: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria;
- Isole: Sicilia, Sardegna

Istat also estimated data at municipalities level, but this data was not included because of issues about privacy.

The sampling base adopted, meaning the selection list of the sample units, is based on the LAC (Liste Anagrafiche Comunali), a single archive of families residing in Italian municipalities, coming from the municipal registry lists.

The process was designed as a two-stage selection (municipalities, families), because the interview is of the type CAPI (Computer-assisted personal interviewing), a interviewing technique in which the respondent uses an eletronic device to answer the questions, in presence of the interviewer as a guide. This is usually done when the questionnaire is complex and/or long, and would be difficult to be done over a phone call. In this case, for both cost and organizational reasons, the sample units come from a limited number of municipalities.

Stratas were formed based on municipality and the 4 trimesters of the year: each municipality is involved in the survey every month (and the number of sample families it is divided into months), while municipalities with a lower number of residents (NAR - Non Auto Rappresentativi) participates in the survey four months in the year, three months apart.

The goal of stratification is to form groups (or layers) of units characterized, in relation to the variables under investigation, by maximum homogeneity in the layers and maximum heterogeneity between the layers. Achieving this goal translates into precision of the estimates, that is to say a reduction of the sampling error with the same sample size.

C. Attributes Selection

List of chosen attributes:

- ESPE_CO - continuous numeric: unit's total medium expense;
- ESPE_GIO - continuous numeric: unit's medium expense per day;
- SESSO - binary nominal categorical: unit's gender;
- DURATA - discrete numeric: unit's number of nights per holiday;
- ETA10 - discrete numeric: unit's age;
- MESE - ordinal categorical: trips' month of the unit
- TIPOMARE - binary nominal categorical: unit chose a seaside location;
- TIPOMONT - binary nominal categorical : unit chose a mountain location;
- TIPOCITTA - binary nominal categorical: unit chose to visit a city;
- TIPOCAMP - binary nominal categorical: unit chose to visit countryside;
- TIPOCRO - binary nominal categorical: unit chose to go on a cruise;
- TIPOALTRO - binary nominal categorical: unit chose another destination not mentioned above;
- ALLOGG - nominal categorical: unit chose type of accommodation;
  - rental house;
  - B&B, hotel sleeping accommodation with breakfast included;
  - rental room;
  - timeshare house, property with a divided form of ownership based different time slots, typically in one-week increments;
  - agritourism;
  - friend's house;
  - other private house;
- npart - number of family members actually participating in the trip;
- PIATTALL1 - binary nominal categorical: unit answered yes or no to the usage of the platforms "Booking, Expedia, Tripadvisor, Trivago, Kajak";
- PIATTALL2 - binary nominal categorical: unit

answered yes or no to the usage of the platforms ”Aibnb, Homeaway, ScambioCasa, HomeToGo”.

D. Methodologies and Tools

To ensure a solid base on a statistical analysis about Italy, this paper research was done on data from the largest and most reliable source of data in the country. The systematic approach to the topic of trips of Italians was carried by the analysis of qualitative and quantitative data, looking for correlations between attributes of a single unit and describing the general habits of the population based on their characteristics. The survey data is the best data to answer to questions about this topic, because it gives a lot of information directly from the source (excluding response/non-response bias). In order to better describe the sample characteristics, it was deployed a descriptive analysis approach, selecting the right attributes to show the most relevant information about the topic, like the month in which Italians make most trips or the preferred type of chosen destination. Predictive analysis was done to estimate values based on the collected data and to look for correlations between different attributes, developing regression models. Data is reviewed with the assistance of the R programming language (which also plotted the presented charts) and the code was shared between the group members on GitHub, where it can be examined by the reader.

II. Development

A. Descriptive Analysis

This section will describe the sample composition and summarize the correlations between some of its attributes, all the graphs and statistical analysis are made through R [3] and with useful manuals [4]. This preliminary analysis will be the foundation of the further inferential investigation. Figure 1 shows the sample's age intervals. The classes are 8, made by the following number of years intervals:

- Less than 14 years;
- 15-24 years;
- 25-34 years;
- 35-44 years;
- 45-54 years;
- 55-64 years;
- 65-74 years;
- More than 75 years.

The age composition is heterogeneous: it is worth to notice that the sample has a very high number of units aged 14 or less: 519 units out of 4939. Of course, this will affect further analysis, especially about the expense analysis and correlations, because we assume that minors (young people in general) will lower the expected results, assuming they do not have the same budget adults have. It is also clear that the sample distribution of values is slightly negative skewed: this was expected given that young working people are assumed to go on trips more than older people, who may not be able to meet the physical requirements needed to travel.

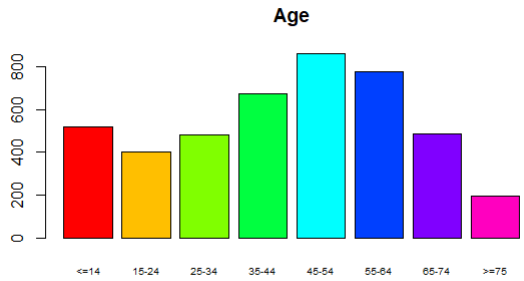

Fig. 1. Age of the sample's observations

It is interesting to point out that the vast majority of people that travel in Italy do not travel far, instead 80% of Italians prefer to stay in their home country (see Figure 2). Is it because Italy is a wonderful country, and its inhabitants don't have to move too much to reach a vacancy location, or because travelling far is more expensive than staying in the countries borders?

This will be investigated and (partially) answered later.

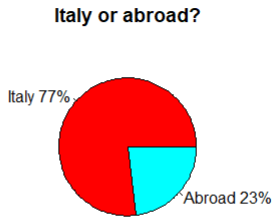

Fig. 2. Destination chosen by Italian people

The investigation has been carried out asking people which type of places they went, letting them choose between:

- seaside location;
- mountain location;
- city;
- countryside;
- cruise;
- another destination not mentioned above;

Figure 3 shows that the favourite place to travel for Italians is the seaside, close second to cities.

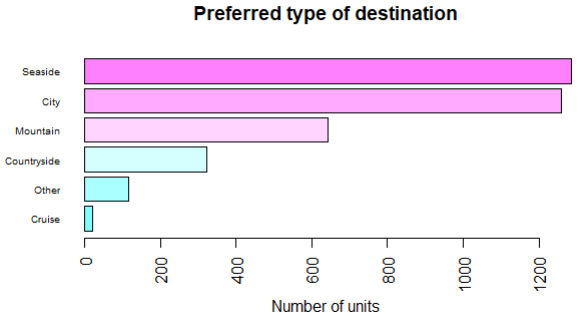

Fig. 3. Destination chosen by Italian people

Considering that the overall mean of the variable ESPE_CO is $\mu_{expense} = 382.4887$, we tried to check whether this expense is influenced by the type of destination or not. The boxplot in Figure 4 shows the total medium expense relative to the types of destination preferred by Italians. The black line corresponds to the median of the expenses incurred throughout the trip. The box shows the 50% of the observations, large as the $IQR = Q_3 - Q_1$ (InterQuartile Range) where Q_1 is the first quartile and Q_3 the third quartile. This index is also utilised to identify outliers as the observations that falls below the two whiskers - i.e $Q_1 - 1.5IQR$ and above $Q_3 + 1.5IQR$ and they're represented in the graph by white dots. Even though the white dots stand out, they're not significant to the sample analysis so they don't need to be excluded. The notch displays the confidence interval around the median which is normally based on

$$median \pm 1.57 \cdot \frac{IQR}{\sqrt{n}}.$$

Although it is not a formal test, if two boxes' notches do not overlap there is "strong evidence" (95% confidence) their medians differ. Furthermore, the shape of the boxplots shows that the data are skewed and the majority of observations fall between an expense between 0€ and 1000€. The number of units that go on a cruise ("TIPOCROC" attribute) is very low, in fact this reflects on the bloxplot which observations fall in the IQR. The boxplot of units which goes somewhere different from the other destinations ("TIPOALTRO" attribute) shows a slight uncertainty on the median too, due to the high variability shown in the standard deviation: this means that the total expense value vary a lot for every destination. Furthermore, the greater expense is relative to trips at the sea, in fact they have the greater IQR between all the destination.

The table in Figure 6 represents the relation between two quantitative variables that are the medium expense per day and the duration of the trips expressed in number of nights. The mean of duration of Italians trip is $\mu_{duration} = 5$ days. The scatter-plot shows a cloud of points in the bottom left part of the diagram. Observing the distribution of data, the intuitive assumption that can be made is that if people spend more

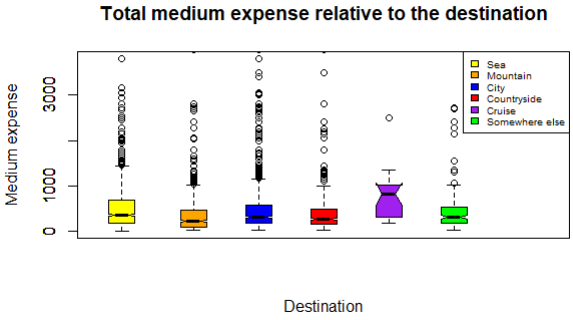

Fig. 4. Total expense based on the destination

Total Expense per	Mean	SD	IQR	Q1-25%	Q3-75%
TIPOMARE	507.51	481.02	510	180	690
TIPOMONT	367.44	447.1	373	102	475
TIPOCITTA	485.66	537.63	394.75	177	495
TIPOCAMP	440	556.06	345	150	400
TIPOCROC	866.3	661.35	706	319	1025
TIPOALTRO	485.71	570.53	353	181	534

Fig. 5. Index Position of total expense per destination

days on trips, the medium expense per day is going to decrease logarithmically, but it can't be said for sure, so it has been decided to search for the appropriate function to fit data. Since the relation modelled by the regression linear model doesn't describe in a good way the bound between the two variables, a logarithmic regression better fits the data, for the parameters α, β in this equation:

$$expense\_per\_day = \alpha + \beta \cdot \log(duration) + \epsilon,$$

obtaining, once fitted, the following formula:

$$Medium\_Expense\_Per\_Day =$$

$$132.62 - 30 \cdot \log(duration).$$

By using this formula is possible to calculate the value of medium expenses per day of the units in the sample, but since it is not a linear relation, the goodness of the model cannot be verified needs to be verified with more complex tools. Furthermore, the observations plotted are identified by their type of employment, as shown in the legend: the red plots refer to CEOs and employees; the green ones to the workers, the light blue ones to the entrepreneurs and the violet ones to the freelancers. Diving deeper in the analysis of correlation, it's possible to calculate the index relative to the dependency of the two variables ESPE_GIO and DURATA. First, it has been computed the correlation's index that is equal to $\rho = -0.293$. Then, we've checked the null hypothesis stating that correlation is equal to 0, and since p-value resulted to be smaller than $\alpha = 0.05$, the null hypothesis has been rejected with a 95% confidence interval of $[-0.319, -0.265]$. The negative correlation could imply a repulsive relation between the two variables, as the graph shows the increasing of one variable as the other one decreases. Since the correlation's index is very low, it has been decided to calculate medium dependency's index η^2,which resulted to be equal to $\eta^2_{expense|duration} = 0.291$. Further developments on the estimations of this indexes for the population, will be analysed in the inferential part of the report.

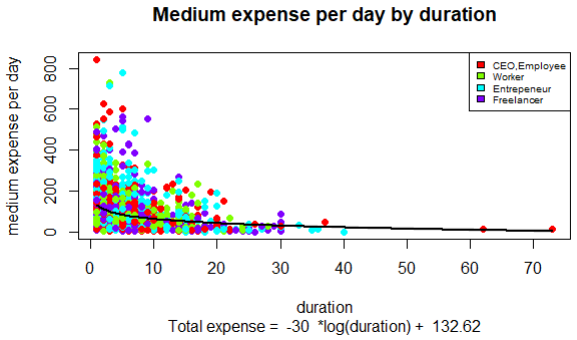

Fig. 6. Relation between total medium expense and number of nights

To verify where Italian people booked their trips, the data set provides two 2 columns where the units answered 'yes' or 'no', depending on their experience. The first one is related to "Booking, Expedia, Tripadvisor, Trivago, Kajak" (platforms which are similar to each other) that give users the opportunity to book accommodation provided by hotels and means of transport. As the pie chart in Figure 7 shows, the red slice is the percentage of people that used one of this platform to book their trip and it corresponds

to the 53%; the light blue one shows the percentage of people (47%) who did not. The second column is related to "Aibnb, Homeaway, ScambioCasa, HomeToGo", platforms that allow users to book only the accommodation provided by private citizens. As pie chart in Figure 8 shows, the people who used these platforms are the 7% of the sample, a very low percentage compared to the people who booked on the other type of platforms.

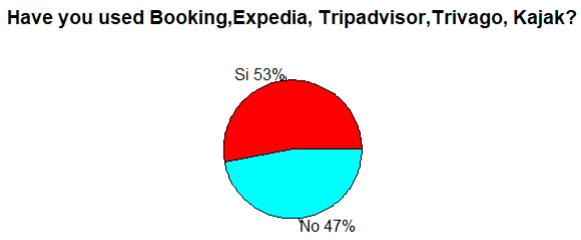

Fig. 7. Percentage of Italian people using Booking, Expedia, TripAdvisor, Trivago and Kajak to book whole trip.

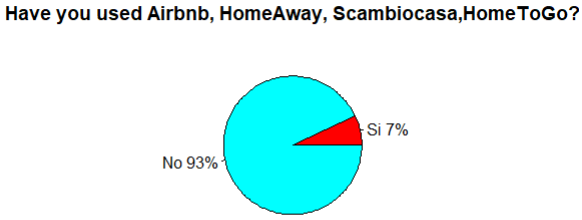

Fig. 8. Percentage of italian people using Airbnb, HomeAway, ScambioCasa, HomeToGo to book accommodation.

Another attribute that is relevant to the analysis is the month, because it shows which is the favourite and most frequent period of time when Italians go on trips. We have plotted a bar chart in Figure 10 in which months are represented on the x-axis and frequencies related to each month (provided by the occurrences of each unit) on the y-axis. By looking at the graph, it is possible to say that the mode is 8, that corresponds to the $8th$ month of the year, while the median μe cannot be determined, since the number of months is even and the two median positions, that fall in the middle, are different from each other.

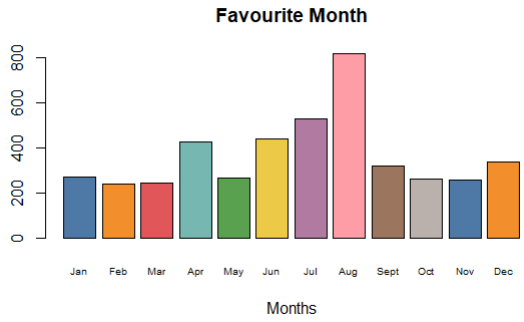

Fig. 9. Number of Italian people going on holiday per month

The relation between the variables "ESPE_GIO" and "ALLOG" is shown in the graph 11. What really stands out is that all the distributions are skewed, some positively and some negatively. The distributions of the medium expense per day, based on B&B, Rental Room, TimeShare House and Friend's House, are all skewed to the right and this implies that the mean is larger than the median, while the distribution of the medium expense per day of people housed in agritourism is skewed to the left. Since the median is a positional index with both mean and mode, by looking at the plot it is possible to say that the lowest medium expense per day is related to the accommodations "TimeShared House" and "Friend's House", given that 75% of the observations stay below 100. Furthermore, the boxes of these distributions are very tiny and this suggests small variability of the data, but even if they look similar, the number of outliers of "Friend's house" is significantly greater than "TimeShared House"'s outliers. This huge quantity of outliers could suggest that the medium expense per day of people staying at friends' house is related to something else, hence correlations with other variables could occur. In conclusion, the plot

shows that the highest medium expense per day is given by the accommodations ”Rental Room” and ”B&B”, according to the hypothesis that the daily expense depends on the type of accommodations.

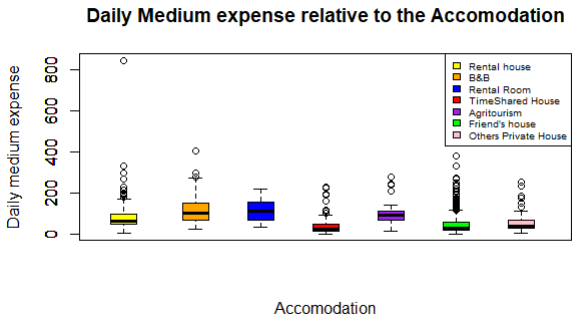

Fig. 10. Relation between daily medium expense and the type of accommodation

B. Predictive Analysis

Regarding the predictive analysis, in first place we’ve calculated the confidence’s interval where population’s parameters are most likely to be included. For a fixed sample, higher degrees of confidence require a wider (less precise) confidence interval. In second place, we’ve verified some assumptions made for the population through statistical hypothesis testing on the sample. Finally, we carried out regression analysis - i.e. multiple regression and linear regression, in order to estimate relationships between a dependent variable and one or more independent variables.

1) Estimation: Our goals is to estimate the values of parameters based on measured empirical data that has a random component - i.e. through a procedure used to calculate the value of some property of a population from observations of the sample. There are two types of estimation: point estimate and interval estimate. We used the interval estimate that defines a range within which the value of the property can be expected (with a certain degree of confidence) to fall. At first we estimated the range of the attributes: duration, total medium expense and favourite month. This range, within which the value of the mean is most likely to fall, is the interval of confidence ”IC”, computed with a 95% degree of confidence.

Attribute	Sample’s Mean	IC-95%
DURATA	5	[5.36,5.69]
ESPE_CO	382.48	[369.55,395.43]
MESE	6.75	[6.66,6.84]

After computing these estimations, we dove deeper and we estimate the interval within the population’s mean of the daily medium expense, relative to the type of accommodation, is most likely to fall:

Daily Expense per accommodation	Sample’s Mean	IC-95%
Rental House	79.84€	[73.27,76.41]
Bed&Breakfast	119€	[111.23,126.26]
Rental Room	112.82€	[93.84,131.81]
TimeShared House	39.26€	[34.25,44.26]
Agritourism	96.80€	[82.27,111.33]
Friend’s house	44.68€	[42.55,46.82]
Other’s private house	57.69€	[49.05,66.33]

The estimations of the interval where population’s mean of total medium expense related to the type of destination can be expected to fall, are presented in the following table:

Total Expense per Destination	Sample’s Mean	IC-95%
Seaside	507€	[490,525]
Mountain	367€	[349,386]
City	485€	[466,505]
Countryside	440€	[415,465]
Cruise	866€	[834,898]
Somewhere else	485€	[459,512]

Moreover we computed the proportion of both people

who goes abroad and who spend their holiday in Italy using the attribute DEST_IE, computing the range where the population's proportion is likely to fall at a 95% degree of confidence. Furthermore we computed the proportion of people using "Booking, Expedia, Tripadvisor, Trivago, Kajak" (Piattall1) to book their trip or "Airbnb, HomeAway, ScambioCasa, HomeToGo" (Piattall2) to book their accommodation.

Attribute	Sample's proportion	IC-95%
DEST_IE (Italy)	77%	[76%,78%]
DEST_IE (Abroad)	23%	[22%,24%]
Piattall1 (Yes)	53%	[51%,55%]
Piattall2 (Yes)	7%	[6%,8%]

2) Hypothesis Testing: A statistical hypothesis is a hypothesis that is testable on the basis of observed data modeled. After computing the medium expense daily's interval of confidence for the population, the first hypothesis that we are going to verify is that the average total expense of the considered year (2019) is greater than the previous year (2018). In order to obtain these means and using them just as a form of comparison, we used the data set that ISTAT has provided for the year 2018 with the same attributes.

$$\begin{cases} H_0 : \mu_{2019} > \mu_{0,2018} \\ H_1 : \mu_{2019} \leq \mu_{0,2018} \end{cases}$$

H_0 is the null hypothesis that summarize our prediction of the increase in the medium expense, while H_1 is the alternative hypothesis that indicates the opposite trend. We verified the statistical hypothesis mentioned above per each type of destinations included in our data set using the following means of the previous year per each type:

$$\begin{cases} \mu_{0,2018} = 484.63, & TIPOMARE \\ \mu_{0,2018} = 348.82, & TIPOMONT \\ \mu_{0,2018} = 428.05, & TIPOCITTA \\ \mu_{0,2018} = 414.78, & TIPOCAMP \\ \mu_{0,2018} = 875.23, & TIPOCROC \\ \mu_{0,2018} = 557.38, & TIPOALTRO. \end{cases}$$

Since we obtained the same results per each type of destination (accepting the null hypothesis), we analyze the procedure just for one type: TIPOMARE. Being the size of our sample $n > 30$, we assume that the distribution could be approximated by a Gaussian -i.e. follows a normal distribution- and to verify the null hypothesis we have made a Z-test that is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large. Nuisance parameters such as standard deviation should be known in order to perform an accurate z-test, but in our case, we computed the estimation of standard deviation's sample that is $s = 481.02$. We computed the z-score that is a number representing how many standard deviations above or below the mean population a score derived from a z-test is.

$$Z_{score} = \frac{\mu_{2019} - \mu_{0,2018}}{\frac{s_{2019}}{\sqrt{n}}} = -1.36$$

While considering a level of confidence of $1-\alpha$ where $\alpha = 0.05$, we computed the theoretical $-Z_\alpha$ that is equal to 1.64. Considering that to reject the null hypothesis we need to have $Z_{score} < -Z_{(\alpha)}$ where the rejection's region is: Since $Z_{score} > -Z_{(\alpha)}$ we accept the null hypothesis that the total medium expense of holiday to the seaside has increased over the last year.

3) Regression's Analysis: Using the regression model to dive deeper in the relations among variables, we had verified five key assumptions before starting:

- Linear relationship.

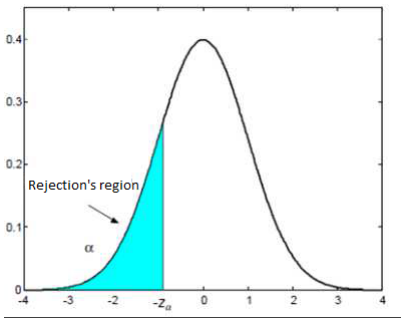

Fig. 11. Region of rejection of the null hypothesis

- Multivariate normality.
- No or little multicollinearity.
- No auto-correlation.
- Homoscedasticity.

In figure 12 there is a summary of the most relevant coefficients estimated computing the multiple linear regression of total expense during trip and type of destination. The model computed is: $Y = \beta_0 + \beta_1 \cdot TIPOMARE + \beta_2 \cdot TIPOMONT + \beta_3 \cdot TIPOCAMP + \beta_4 \cdot TIPOCITTA + \beta_5 \cdot TIPOCROC + \beta_6 \cdot TIPOALTRO + \epsilon$.

```
Residuals:
    Min      1Q  Median      3Q     Max
-1019.0  -259.4  -141.4    90.6  4552.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2131.530    255.605   8.339  < 2e-16 ***
TIPOMARE    -186.536     21.252  -8.777  < 2e-16 ***
TIPOMONT     -39.006     25.038  -1.558   0.1194
TIPOCAMP      -8.109     28.094  -0.289   0.7729
TIPOCITTA   -142.873     20.227  -7.063 2.05e-12 ***
TIPOCROC    -439.537    105.972  -4.148 3.46e-05 ***
TIPOALTRO   -116.434     45.439  -2.562   0.0104 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 471.7 on 2754 degrees of freedom
  (1632 observations deleted due to missingness)
Multiple R-squared:  0.04441,   Adjusted R-squared:  0.04233
F-statistic: 21.33 on 6 and 2754 DF,  p-value: < 2.2e-16
```

Fig. 12. Summary of multiple linear regression of total expense during trip and types of destination

The values of the coefficients are:

- $\beta_0 = +2131.53$
- $\beta_1 = -186.54$
- $\beta_2 = -39.01$
- $\beta_3 = -8.11$
- $\beta_4 = -142.87$
- $\beta_5 = -439.54$
- $\beta_6 = -116.43$

The binary nominal categorical variables used in the model are managed as dummy variables in fact, for example: to predict the total medium expense for a unit "TIPOMARE", the total medium expense is equal to $Y = -2131.53 + 1 * (-186.54) = 1944.99$ since $\beta_0 = -2131.53$, $\beta_1 = -186.54$ and "TIPOMARE" assumes value 1. Diving deeper if the qualitative variable has 2 categories, only one dummy variable is needed in order to represent both categories. In other words:

- $X_n = 0$ in case of absence of the observed qualitative variable;
- $X_n = 1$ in case of presence of the observed qualitative variable.

For a given predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the beta coefficient of the predictor is significantly different from zero. It can be seen that, changing in TIPOMARE, TIPOCITTA, TIPOCROC, TIPOALTRO are significantly associated to changes in ESPE_CO while changes in TIPOMONT, TIPOCAMP are not significantly associated with ESPE_CO. The coefficient (β_0) can be interpreted as the average effect on y of one unit increase in predictor, holding all other predictors fixed. For the purpose of deciding whether to accept or reject the null hypothesis, we compute the $F_{statistics}$ that is 21.33 which is smaller than the $F_{value} = 2.102$, so we reject the null hypothesis stating that there is independence.

```
 TIPOMARE  TIPOMONT  TIPOCAMP TIPOCITTA  TIPOCROC TIPOALTRO
 1.394576  1.389673  1.011714  1.259566  1.002110  1.039678
> sqrt(vif(reg3)) > 2 # problem?
 TIPOMARE  TIPOMONT  TIPOCAMP TIPOCITTA  TIPOCROC TIPOALTRO
    FALSE     FALSE     FALSE     FALSE     FALSE     FALSE
```

Fig. 13. Evaluating VIF of multiple linear regression

Multicollinearity occurs when there's correlation between predictors (i.e. independent variables) in a model and its presence can adversely affect regression results. To evaluate it in this regression analysis

we compute the variance inflation factor (VIF) that estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model. The numerical value for VIF tells us (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, looking at the table in figure 13, TIPOMARE, TIPOMONT and TIPOCITTA have a VIF respectively 1.39, 1.38 and 1.25 telling us that the variances of these coefficients are, respectively, 39%, 38% and 25% bigger than what we would expect if there was no multicollinearity — if there was no correlation with other predictors. If VIF is higher than 5 could indicate multicollinearity in our analysis, but after computing it for every variable as shown in the table it is possible to conclude that our model is not affected by multicollinearity.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1386.8  -169.9   -67.1    90.4  4261.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -140.618     18.093  -7.772 9.57e-15 ***
DURATA        24.789      1.075  23.056  < 2e-16 ***
DEST_IE      312.935     13.998  22.356  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 385.1 on 4390 degrees of freedom
Multiple R-squared:  0.2258,    Adjusted R-squared:  0.2255
F-statistic: 640.3 on 2 and 4390 DF,  p-value: < 2.2e-16
```

Fig. 14. Summary of multiple linear regression of total expense during trip, duration and type of country (Italy or abroad)

The summary of the coefficients estimated computing the multiple linear regression of total medium expense, duration and type of destination (Italy/abroad) are presented in 14. The obtained model is the following: $Y = \beta_0 + \beta_1 \cdot DURATA + \beta_2 \cdot DEST\_IE + \epsilon$. The coefficients assume the following values:

- $\beta_0 = -140.61$,
- $\beta_1 = 24.79$,
- $\beta_2 = 312.93$.

Furthermore for each coefficient, t-value was computed to evaluate whether or not there is significant association between the predictor and the outcome variable. It can be seen that, changing in DURATA and DEST_IE are significantly associated to changes in ESPE_CO. For the purpose of deciding whether to accept or reject the null hypothesis, we compute the $F_{statistics}$ that is 640.3 which is smaller than the $F_{value} = 2.99$, so we reject the null hypothesis stating that there is independence.

```
   DURATA  DEST_IE
1.039399 1.039399
> sqrt(vif(reg2)) > 2 # problem?
 DURATA DEST_IE
  FALSE   FALSE
```

Fig. 15. Evaluating VIF of multiple linear regression

In this case (figure 15) both VIF of DURATA and DEST_IE are 1.03 telling us that the variances of these coefficients are 3% bigger than what we would expect if there was no multicollinearity. Considering that our $VIF < 5$, our multiple linear regression is not affected by multicollinearity.

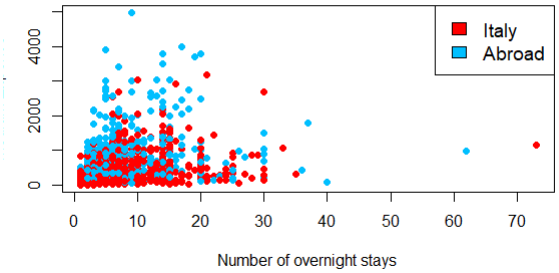

Fig. 16. Scatter Plot of Country (Italy or abroad), duration of the trip and total medium expense

The scatter plot shown in 16 is a representation of the previous multiple linear regression model; the red dots refer to ”Italy” according to the pie chart in the descriptive part, while the blue ones refer to ”Abroad”. It can be noticed that there is a cloud of points in bottom left part of the plot and this behaviour suggests that most of the units falls under 2000 or less and their trip lasts less than 20 nights. It is also worth to notice the presence of outliers: these are the units that deviate from the principal cloud previously described. An explanation to this outliers could be that some travellers spend more than 2000 in total per holiday and their trip last more than 20 nights.

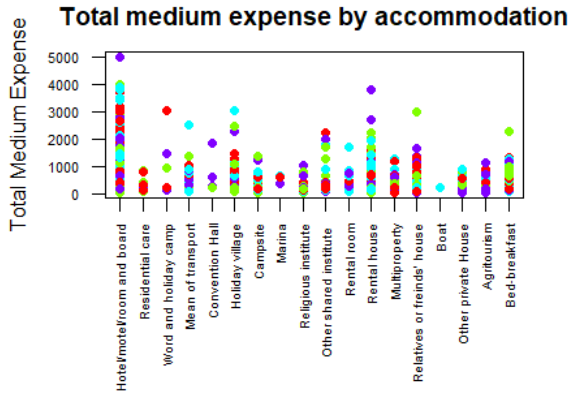

Fig. 17. Scatter Plot of Type of Accommodations versus Total Medium Expense

The plot in 17 is a simple scatter plot showing for each of the 18 levels of the variable "ALLOG_Fact", the total medium expense. The levels of the variable "ALLOG_Fact" are the following one:

- "Hotel/motel/room and board";
- "Residential care";
- "Word and holiday camp";
- "Mean of transport";
- "Convention Hall";
- "Holiday village";
- "Campsite";
- "Marina";
- "Religious institute";
- "Other shared institute ";
- "Rental room";
- "Rental house";
- "Multiproperty";
- "Relatives or freinds' house";
- "Boat";
- "Other private House";
- "Agritourism";
- "Bed-breakfast".

Note that, as defined, the residuals are on the y-axis and the fitted values are on the x-axis. This plot is not a classical example of a well-behaved residuals vs

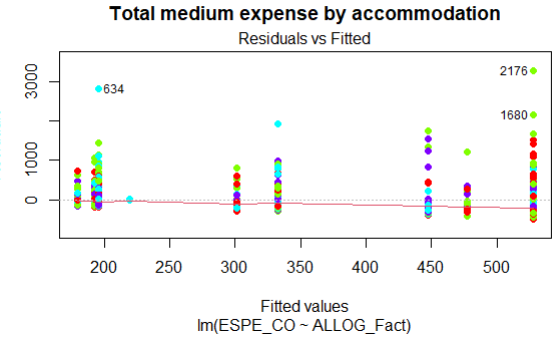

Fig. 18. Residuals of total medium expense vs Fitted values of the type of accommodations

fits plot though. Here are the reasons why the plot in figure 18 of residual vs fits plot cannot be considered appropriate:

- the residuals do not "bounce randomly" around the 0 line showing that the assumption that the relationship is linear is not reasonable.
- the residuals do not roughly form a "horizontal band" around the 0 line and this suggests that the variances of the error terms are not equal.
- residuals "stands out" from the basic random pattern of residuals and this suggests that there are a lot of outliers.

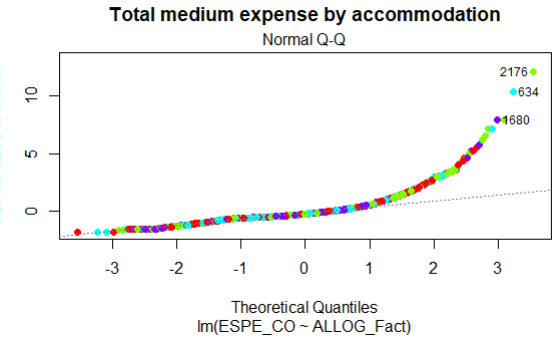

Fig. 19. QQ-Plot comparing the two distributions (observed vs theoretical)

In order to deepen our analysis of the type of distribution that describes the relation between ESPE_CO and ALLOG_Fact, we've plotted a Q-Q plot in figure 19 (graphical tool that helps us on assessing from

which type of distribution our set data come from), assuming that theoretical quantiles come from a normal distribution. It's just a visual check, not an airtight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation. Since the points forming a line that's roughly straight we can state that our sets of quantiles come from the same distribution.

```
    Min      1Q  Median      3Q     Max
-1432.2  -191.9   -81.9    81.2  4405.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  350.364     13.775   25.44   <2e-16 ***
DURATA        30.906      1.101   28.06   <2e-16 ***
npart        -64.401      5.321  -12.10   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 399.9 on 4390 degrees of freedom
Multiple R-squared:  0.1655,    Adjusted R-squared:  0.1652
F-statistic: 435.4 on 2 and 4390 DF,  p-value: < 2.2e-16
```

Fig. 20. Summary of multiple linear regression of total expense during trip, duration and family members

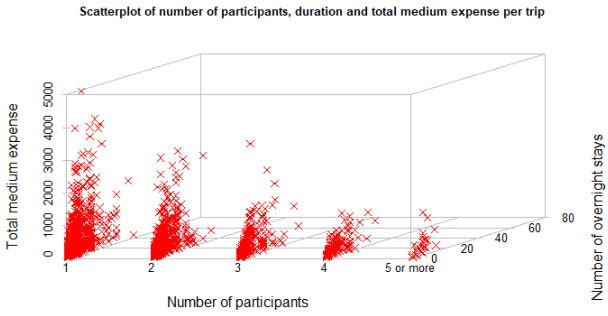

Fig. 21. Three-Dimensional Scatter-plot which highlights the role of the number of participants on the final duration and total expense of the trip

The diagram plotted in figure 21 is a three dimensional scatter plot where on the x-axis are the number of participants, on the y-axis the number of overnight stays and on the z-axis the total medium expense. The summary of the coefficients estimated computing the multiple linear regression of total medium expense, duration and member of the family participating to the trip, are presented in 20. The obtained model is the following: $Y = \beta_0 + \beta_1 \cdot DURATA + \beta_2 \cdot npart + \epsilon$. The coefficients assume the following values:

- $\beta_0 = +350.36$,
- $\beta_1 = +30.906$,
- $\beta_2 = -64.40$.

It's worth to notice that greater is the number of family's members, the smaller is the total medium expense of the holiday and the relative duration. The null hypothesis is that the variables are independent among them and for the purpose (deciding whether to accept or reject the null hypothesis), the $F_{statistics}$ has been computed and it turned out to be equal to 435.4, smaller than the $F_{value} = 2.99$, leading us to reject the null hypothesis.

```
  DURATA    npart
1.011761 1.011761
> sqrt(vif(reg5)) > 2 # problem?
DURATA  npart
 FALSE  FALSE
```

Fig. 22. Evaluating VIF of multiple linear regression

In this case (figure 22) both VIF of duration of trip and member of the family (npart) are 1.01, telling us that the variances of these coefficients are 1% bigger than what we would expect if there was no multicollinearity. Considering that our $VIF < 5$, our multiple linear regression is not affected by multicollinearity.

III. Conclusions

To sum up, the investigation aims to draw some clear lines on what are the main aspects that characterise the tourist demand of the Italian population. The estimations concerns the number of tourists and the number of days per trip, considering the medium expense, both total and daily, and if their destination was in Italy or a foreign country. The medium amount of money that Italian people spend for their trips is given by the type of accommodation they choose and the type of destination they visit. With different types of accommodation and destinations the overall expense increases, actually low daily expense

is associated to Time-Shared House while high daily expense is associated to B&B and rental room (the average sums to 382.48). Furthermore the highest expense corresponds to a trip to an abroad destination and to the holiday on cruise, in fact the medium expense per day is 866, but it is worth to notice that the variability of this attribute is very high. The Italian's favourite type of places to visit on holidays are seaside and city. The modal class 45 years - 54 years represents the overall age of people that goes on trip, while the favourite month to travel is August for summer holidays and December for winter ones. Italian people travel overall in Italy, indeed just a small percentage of Italian travellers go Abroad for holidays and their trips lasts about 5 nights. Last but not least, it is worthwhile to notice that Italians are not comfortable with booking accommodations on online marketplaces, in fact just a small percentage of people stated that they had booked using platforms. The discrepancy is incredibly visible comparing the usage of platform where companies offers theirs services and the usage of platforms where privates share their rooms (just the 7% uses this type of booking website). In conclusion, the entire analysis shows, according to theirs expenses for trips, a huge discrepancy among the Italian's way of experience holidays.

REFERENCES

[1] ISTAT, "Viaggi e Vacanze," 2019. [Online]. Available: https://www.istat.it/it/archivio/178695.

[2] ——, "Nota Metodologica di Viaggi e Vacanze," 2019. [Online]. Available: https://www.istat.it/microdata/download.php?id=/import/fs/pub/wwwarmida/264/2019/01/Nota.pdf.

[3] R. delevopers, "Sito Ufficiale R," 2020. [Online]. Available: https://cran.r-project.org/.

[4] A. Vardanega, "R per l'Analisi dei Dati," in *Ricerca Sociale con R*, Oct. 2020. [Online]. Available: https://www.agnesevardanega.eu/wiki/r/start.