Introduction to data mining with the tips dataset

Lab 1: Important issues illustrated from a case study

- Data and Objective understanding
- Descriptive statistics
- Visualisation tools
- Regression

Please send comment to (version 1.0 – July 21, 2016):

Romain Billot *romain.billot@telecom-bretagne.eu*
Yannis Haralambous *yannis.haralambous@telecom-bretagne.eu*
Philippe Lenca *philippe.lenca@telecom-bretagne.eu*
Sorin Moga *sorin.moga@telecom-bretagne.eu*

1 Data and objective understanding

–The Tips dataset– Food server's tips in restaurants may be influenced by many factors (e.g. the nature and location of the restaurant, the size of the party, the table location and the day of the week…). Restaurant managers need to know which factors matter when they assign tables to food servers. Indeed, for the sake of staff morale, they usually want to avoid either the substance or the appearance of unfair treatment of the servers, for whom tips (at least in restaurants in the United States) are a major component of pay.

In one restaurant, a food server recorded some data on all customers they served during an interval of two and a half months in early 1990[1]. The restaurant, located in a suburban shopping mall, was part of a national chain and served a varied menu. In observance of local law the restaurant offered seating in a non-smoking section to patrons who requested it. Each record includes a day and time, and thus taken together, they show the server's work schedule. The food server provided a comma-separated-value file `tips.csv` containing 244 records, described by 7 variables (`total bill`, `tip`, `sex`, `smoker`, `day`, `time` and `size`; see Table 1).

Variable	Meaning
`total bill`	Total bill, including tax, in US dollars
`tip`	Tip (gratuity) in US dollars
`sex`	Sex of person paying for the bill
`smoker`	Smoker in party?
`day`	from Thursday to Sunday
`time`	Dinner or Lunch
`size`	Size of the party

Table 1: Variables of the `tips.csv` dataset

Question 1

What do you know from the text above and what information is missing?

[1] Source: Bryant, P. G. and Smith, M. A. (1995), Practical Data Analysis: Case Studies in Business Statistics, Richard D. Irwin Publishing, Homewood, IL.

Question 2

Do you have some idea about the objectives of the study and the knowledge you could extract from the data? Could you suggest a list of questions of interest?

Question 3

Load the dataset into R studio or SAS (depending on your group) and have a look at it using the str() function. Describe the data (the format of the data, the quantity of data –number of example/records and variable/fields–). What are the expected values and role of each variable?

Question 4

Tip is usually referred to by percentage points, or as a rate. This enables a normalization over the total bill and a comparison of values across other variables. The question is now to create a "tip rate" variable and to add it to the original dataset.

Home work

Explore the notion of scale of measurement. Provide a short note with meaningful definitions and examples. Explain why it is important to consider the right scale for each variable. What is the scale for each of the eight variables?

2 Descriptive statistics and visualisation

Question 5

Explore univariate summaries with the R *summary* function.

Question 6

Plot a representation of the days distribution in the dataset and comment.

Question 7

Prepare a plot of the amount of tips against the total bill. What can you see ? Test the correlation between the two variables.

Question 8

Draw and interpret three boxplots :

1. the distribution of the total bill,

2. the distribution of tips;

3. the distributions of tips vs. days.

Question 9

Draw an histogram of tips. What can you say about the shape of the data ? Is this restaurant expensive ? Split the plotting window into 6 subplots (function *mfrow*) and plot 6 histograms with increasing numbers of breaks.

Question 10

Display the counts (proportions) for Gender of the Bill Payer and Smoking Parties. Do the same for time of the day (dinner or lunch) and day of the week

Question 11

Display the counts (proportions) for Gender of the Bill Payer and Smoking Parties. Do the same for time of the day (dinner or lunch) and day of the week

Question 12

Who pay mostly the bills ? men or women ? and when ? Try to visualise the conditional distributions of Sex given the day of the week, with a mosaic plot

3 Regression

Question 13

Before starting with the regression, we will learn how to build dummy variables, which is sometimes useful. Create four new variables, named `thu`, `fri`, `sat`, `sun`, that take 1 if the dining party was held on that day, 0 otherwise. Use the function *with* of R and force the variable to the R factor type with the *factor* function

Question 14

Fit a general linear model with `tip rate` as a response variable against all the other variables of interest : `sex`, `smoker`, `time`, `size`, `thu`, `fri`, `sat`, `sun`

Question 15

Fit a model with only the `size` as an explanatory variable

Question 16

Use a stepwise algorithm with the AIC statistic as a variable selection process to select a good model. Start from the full model of question 13. What do you remark?

Home work

Explore the notion of interaction between the Gender and the smoking habit by including explicitely this interaction into a model with `size`, `sex`, `smoke`

Question 17

Check the linear relationship between the tip and the total bill, seen at question 7, with a linear model and interpret the quality of this model