ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Part Time M.Sc. in Business Analytics

Course: Statistics for Business Analytics I
Semester: Fall 2015
Instructor: Panagiotis Tsiamyrtzis (pt@aueb.gr)

FINAL PROJECT

Deadline: 07 March 2016

The final project counts for 30% in forming the grade of this course. The deliverable should be a report, where for each of the 5 problems given below you need to write a short description of your solution, provide supporting plots, statistics, tests etc. all coming from R and a final conclusion with your findings. Along with the report, you need also to submit your .R code that was used to answer the questions (allowing to reproduce your detailed analysis).

1. In a lottery we have 30 balls numbered $1-30$ that are well mixed in an urn. We draw without replacement 5 balls from the urn and a ticket is winning when it guessed all 5 numbers. In the spreadsheet "Lottery_Data" of the file "Final_Project_BA_PT_2015.xlsx" (available on the class's portal) we provide the most recent 10,000 draws from the urn. Using these data answer the following questions:

   (a) Is this a fair lottery? Provide as much of graphical and numerical evidence as you can for or against this argument.

   (b) If you would play this lottery, would you have some favorable number(s)?

(c) How your answers in the previous two questions would change if you use only the first 100 recorded draws? How about if you used only the first 1,000 recorded draws?

2. In the spreadsheets named "W", "X", "Y" and "Z" of the file "Final_Project_BA_PT_2015.xlsx" there are data sets from various random variables. Each student will take one column of data from each of these four spreadsheets (use the column that has your ID number "BAPT1xxx"). For each of these four univariate samples, that you were assigned, you need to:

(a) Provide various forms of graphical representation of the data.

(b) Describe the type of data along with various descriptive statistics that could provide insight on which is the generating random variable.

(c) Based on **(a)** and **(b)** search for the random variable that fits "best" the observed data.

(d) Provide point estimates of the unknown parameter(s) of the "best" fitted model.

3. In the spreadsheet named "Data3" of the file "Final_Project_BA_PT_2015.xlsx" you will find the recorded variables Y, X_1, X_2, X_3 (continuous) and W (categorical with three levels) on 150 cases. Using these data answer the following questions:

(a) Run the parametric one-way ANOVA of each of the continuous variables (Y, X_1, X_2, X_3) on the categorical variable (W). Specifically,

(i) provide a graphical representation of each of the continuous versus the categorical variable

(ii) provide the ANOVA output

(iii) check the assumptions and provide alternatives when the assumptions are violated

(iv) if significant mean differences exist, then use two different ad-hoc methods to identify the mean grouping versus the levels of the categorical variable

(b) Run the non-parametric one-way ANOVA of each of the continuous variables (Y, X_1, X_2, X_3) on the categorical variable (W).

(c) Provide a scatter-plot matrix of Y, X_1, X_2, X_3

(d) Provide a scatter-plot matrix of Y, X_1, X_2, X_3, annotating the different levels of W in each plot using a different color.

(e) Run the regression model of Y on all the continuous variables (X_1, X_2, X_3).

(f) Examine the regression assumptions and provide alternatives if any of them fails. Also provide a discussion of potential outliers and influential points (if any).

(g) Use the ANOVA F-test to examine which is the "best" sub-model.

(h) Using the model found in **(g)** provide a point estimate and a 95% confidence interval for the prediction of Y when: $(X_1, X_2, X_3) = (3.1, 3.75, 1.2)$.

4. In the spreadsheet named "Data4" of the file "Final_Project_BA_PT_2015.xlsx" you will find the recorded variables Y (continuous) and W, Z (categorical with two levels each) on 84 cases. Using these data answer the following questions:

(a) Provide a plot of Y versus the W and Z and comment on it.

(b) Provide the interaction plot of Y versus W and Z and comment on it.

(c) Run the parametric two-way ANOVA of Y on the categorical variables W and Z (including the interaction term) . Provide the fit, examine the assumptions and comment on the significance of the terms.

5. In the spreadsheet named "Data5" of the file "Final_Project_BA_PT_2015.xlsx" you will find the recorded variables Y, T and D on 578 cases. The Y expresses the height of a tree (in cm), T is the time in months after we plant the tree and D is the type of

the soil used to plant these trees (categorical with 4 levels). Using these data answer the following questions:

(a) Provide a plot of Y versus T annotating the different levels of D using a different color.

(b) Run a regression model of Y on T for each of the four levels of D (i.e. run four regression models).

(c) Examine the regression assumptions and provide alternatives if any of them fails.

(d) Comment on the regression coefficient estimates found at the four different levels of D.