Statistical Modelling and Inference

Barcelona GSE, Fall 2020

Prof: Om

CONTENT:
This exam contains 60 points in total. These are added to those from the projects to form your final mark.
The exercises are organised in three parts. Those in the first are expected to be answered quite quickly, they are test-type questions on fundamentals. Those in the second require some statistical modelling. Those in the third require modelling and some semi-advanced statistical computing.
There is a label marking the difficulty of the question, 1 means basic (necessary to pass), 2 intermediate and 3 harder.
Read carefully what each question asks you to do.
You have **3 hours** to complete this exam and you work in your prechosen group of 4 people.

DELIVERABLES:
One or two Jupyter notebooks, R or Python, that contains all solutions, figures, code etc. I am giving the possibility of two notebooks in case you want to use R for some questions and Python for others. Organize the cells in the notebook so that the questions appear **in the same order as in this exam**. If this is not the case, 20% will be deducted from the final mark.
You should return your notebook by email at 13.15, submission past that time will be considered invalid.

1 The quickies

1. **[1, 10 points]** A large study for the identification of genes associated with a particular cancer screened 300 hundred genes using logistic regression analysis. The analysis returned a p-value for each gene, more precisely for the coefficient in front of the expression level for that gene as it appears in the regression model. The p-values for each of the 300 genes are contained in the file `group3_data1.txt`. On the basis of these results one analyst involved in the project suggests that genes with labels

   $$20, 22, 92, 163, 177, 197, 215, 226, 230, 231, 246, 253, 255, 256, 275, 277, 278$$

   should be investigated for further analysis since they appear to be associated with the cancer. Another analyst, however, recommends that from this study there is no evidence that any gene among those 300 is associated with the cancer.

   1. Explain why their opinions differ
   2. According to your analysis of the data, which analyst appears more credible to you?

2. **[1, 10 points]** Suppose that you get ten independent noisy measurements of the same quantity, that is $y_j \sim N(\mu, \sigma_j^2)$ for $j = 1, 2, \ldots, 10$ and for known variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_{10}^2$, with $\sigma_1^2 < \sigma_2^2 < \cdots < \sigma_{10}^2$. Provide an estimator for μ whose mean square error is less than $(\sigma_1^2 + \sigma_2^2 + \cdots \sigma_{10}^2)/100$. I am not expecting you to prove mathematically that your estimator has the desired property, this is why the exercise is level 1. Only to give me an estimator that has the property and explain the procedure you followed to find it. You can convince yourself that it does maybe by numerical experimentation (or by doing the maths, if you feel like it).

3. **[1, 10 points]** The municipality of Barcelona is worried about fraud evasion in real estate transactions. In particular that the prices reported officially in the sales contracts are a percentage of the actual prices since the parties prefer to exchange the remaining amount in cash and without declaring it in order to save up on the associated transaction tax. In exchange of some amnesty and tax benefits, the municipality secured the collaboration of a large real estate company that manages a large portfolio of fairly representative apartments in the city and collected data on their transactions for which the actual price was revealed. The municipality also has a list of recent transactions on apartments from other real estate companies in the city, whose real prices are uncertain and could be subject to under-reporting due to tax evasion. The data is contained in file `group3_data2.txt` and consists of reported transaction prices for 100 flats that were recently sold in the city together with a list of apartment characteristics that the municipality believes that typically affect the commercial price in a proportional way. Specifically, the columns in the data file consist of: the reported transaction price of the property, the square meters of the apartment, the number of bedrooms in the apartment, the floor the apartment is situated in the building, and an indicator whether the property belongs to the collaborating real estate firm (1) or not (0); the point is that in transactions for which the label is 1 the reported price is the actual transaction price, but

the rest are suspected for tax evasion. Using this information develop a sound statistical methodology to:

1. Assess if there is evidence that there is tax evasion
2. If you find that there is, provide an estimate of the percentage by which on average true prices differ from reported ones.

2 The smoothies

4. [**2-3, 15 points**] The file `group3_data3.txt` contains the data for this problem, which is about black-box prediction. No context is given other than that the data consists of a sequence of 0s and 1s and they are generated sequentially. The first 10,000 entries are the training data set. The next 500 are test data. The final 500 values are probabilistic forecasts under the oracle for the test data, i.e., the true model that has generated the dataset; for example the entry 10,501 is the predictive probability that the 10,001 test data point is 1. This probability has been computed using the whole past history of the data, under the true model, which is unknown to you. Your task is to come up with a model or algorithm to provide probability forecasts for the 500 test data. You can train your algorithm or model on the basis of the first 10,000 training data. As an output I want you to show how your approach compares with the oracle. Make sure to report **only** what you consider as your single best answer and not things you might have tried as intermediates.

3 The tough cookie

5. [**3, 15 points**] “Delaying”, a well-known airline company, is experimenting with a new pricing scheme to provide incentive to its customers to purchase advance online checkin.The company wants to learn the function that expresses the demand for such advance checkins as a function of the price this service is offered, in order to decide how to set this price in the future. The company has gathered data on 6000 flights. In each of these flights it has experimented with a slightly different price offered. However, for reasons that have to do with the overall marketing and operational strategy of the company, the number of such advance online checkins offered per flight varied and customers bought them at a first-come/first-serve basis. As a result of the constraint, in several flights the actual demand for the service offered exceeded the number of items offered, and in these cases the company cannot directly observe the actual demand, it can only observe that it exceeded the offer. The data set is included in `group3_data4.txt`; the first column is the price offered for the online checkin on the given flight; the second the number of advance online checkins offered for the flight; the third the demand recorded for the flight. Use this data to estimate and report the pricing function that expresses the demand as a function of the price.