Examen final

N. Saunier

14 décembre 2020

Please

- **consider that this is a english translation of the final exam: please consider that the french version is the original version and refer to it in case of ambiguous or inacurrate wording**;
- note the scale (the total score is out of 20) and the indicative time to devote to each exercise;
- clearly indicate the numbers of the questions you are dealing with and your corresponding answers (and underline or frame the numerical results);
- pay particular attention to the wording and definition of the notations you use;
- note that some exercises require files available on Moodle ("Examen Final" Section) (text files are provided in a version with a period and a comma for decimal numbers, if necessary). Statistical tables are available on Moodle if necessary.

Exercise 1 (modèle de régression) 40 min (/5.5 pts)

A multivariate linear regression model yield the results presented in the tables 1 et 2.

Table 1: Global results

R^2	1.000
R^2 adjusted	1.000
Statistic F	3309
Prob (>F)	0.0128
Observations	5

Table 2: Coefficients of the model

	coef	**std err**	**t**	**P$>$ |t|**	**[0.025**	**0.975]**
Intercept	3.5347	0.162	21.759	0.029	1.471	5.599
x1	7.3775	0.239	30.870	0.021	4.341	10.414
x2	-4.9703	0.132	-37.551	0.017	-6.652	-3.289
x3	0.4180	0.165	2.531	0.240	-1.681	2.517

1. Is it a good idea to add a fourth variable to the model, knowing that its correlation with the dependent variable is very strong? To justify. (1 pt)
2. Describe the quality of the model, if it is significant, and indicate the significant variables of the model. Justify. (1 pt)
3. A new data collection is done including the fourth variable and providing 100 observations provided in the file `exercise1.csv`.
   (a) Based on a visualization of the distribution of the dependent variable y, describe the shape of the distribution (0.5 pt)
   (b) Propose a model of the variable y as a function of the four independent variables $x1$, $x2$, $x3$ and $x4$.
      i. Describe the quality of your model and indicate the significant variables. (2 pts)
      ii. By relying on visualization of the residuals, check and comment if the estimation conditions of the model are verified. (1 pt)

Exercise 2 (analyse et fouille de données) 70 min (/9.5 pts)

This exercise is based on a set of traffic data collected on a highway in the Portland metro area for seven consecutive days in September 2011, available in the file `portland-1395.csv`. The data is aggregated at 20 s intervals and the attributes are as follows:

- detectorid: detector identifier
- starttime: date and time of the start of the interval
- volume: number of vehicles detected in the interval of 20 s
- speed: average vehicle speed
- occupancy: occupancy rate (proportion of the time that the sensor is occupied by a vehicle)
- status: detector status (not used)
- dqflags: quality indicator (not used)
- date: date deducted from the starttime attribute

Please answer the following questions:

1. Describe a sensor technology to collect the three attributes volume, speed and occupancy, and indicate one advantage and one disadvantage. (1 pt)
2. Choose two days and do the following analyzes:
   (a) calculate the 95 % confidence intervals of the speeds for each day; (1 pt)
   (b) compare the average speeds using a statistical test; (1 pt)
   (c) test the adequacy of the distribution of the speeds of one of these days to the normal distribution. (1.5 pts)
3. Explain (without doing it) how to compare, using a statistical test, the means of the variable volume according to the days and the conditions to apply the test. (1 pt)
4. Using a segmentation method, group the traffic conditions (described by the three attributes volume, speed and occupancy) and describe the resulting groups. Choose a small number of groups (2 to 4). (3 pts)
5. Describe (without doing it with the data) a graphic visualization of the groups and the three attributes used to create the groups. (0.5 pt)
6. Describe (without doing it) a supervised learning method for identifying important variables in groups obtained by the segmentation method. (0.5 pt)

Exercise 3 (regression model) 25 min (/3 pts)

The results of a discrete choice model of the logit type are presented in the table 3. The variable to predict is the choice of the plane for a trip (the alternatives are the train, the bus and the car). The explanatory variables of the model are as follows:

- invt: travel time (min)
- hinc: household income (1000$)
- psize: number of people traveling together

Dep. Variable:	car		**No. Observations:**		210	
Model:	Logit		**Df Residuals:**		206	
Method:	MLE		**Df Model:**		3	
Date:			**Pseudo R-squ.:**		0.8240	
Time:			**Log-Likelihood:**		-21.775	
converged:	True		**LL-Null:**		-123.76	
	coef	**std err**	**z**	**P$>\|$z$\|$**	**[0.025**	**0.975]**
const	9.7860	2.312	4.232	0.000	5.254	14.318
invt	-0.0469	0.009	-5.185	0.000	-0.065	-0.029
hinc	0.0291	0.020	1.460	0.144	-0.010	0.068
psize	-1.0656	0.536	-1.988	0.047	-2.116	-0.015

Table 3: Results of a logit model

Please answer the following questions:

1. Describe the quality of the model, if it is significant, and indicate the significant variables of the model. To justify. (1 pt)
2. Explain how to compare the effects of different independent variables. (0.5 pt)
3. Which model would allow us to study the factors associated with the choice of the mode of transport among at least three modes. (0.5 pt)
4. Discuss a survey method for collecting such data (reference population, type of survey and survey technique). (1 pt)

Exercise 4 (spatiale analysis) 15 min (/2 pts)

A model with a spatial component per zone was estimated and we want to validate whether the residuals are well distributed spatially using a spatial autocorrelation analysis. Moran's I for the residuals by zone is -0.2315, with a p-value (obtained by simulation) of 0.04.

1. Are the residuals well distributed spatially (without spatial autocorrelation)? Justify. (1 pt)
2. What would be the interval of the Geary index C for the residuals? (0.5 pt)
3. What measurement would make it possible to determine the places where the residues are most concentrated? (0.5 pt)