DATA ANALYSIS 2. Exam

Mock EXAM - 2023

Your Student ID:

This is intended for a 90 minute paper and pen, closed book exam.

The maximum score is 50p.

Part I. Short questions - rules and advice

- Please be very brief and answer the question only. Please do not answer questions that are not asked.
- If the question asks for a specific answer (yes/no, which one, list advantages/disadvantages etc) then the answer has to contain the answer to the question (yes/no, which one, list advantages/disadvantages etc) in an explicit way.
- It is good practice to give the answer first and the argument next.

1 Part I. Short questions

1. What are the differences of nonparametric and parametric regressions? Give one example for a nonparametric regression and one for a parametric regression. Give a reason for each kind of method. (5p)

2. The following table shows regression results of log hotel price on distance to the city center from a European city. Hotels within 5 km are kept in the data. Prices refer to a weekend night in the fall of 2017. Interpret the constant, the slope coefficient and the R^2 numbers in the table.

Table 1: Hotel regression practice table

VARIABLES	(1) lnprice
Distance	-0.20^{***}
	(0.015)
Constant	5.00^{***}
	(0.05)
Observations	205
R-squared	0.291

3. Using the results from the table of the previous question construct a 95% CI for the coefficient on distance and interpret it. (3p).

   Would an 80% CI be wider (larger) or tighter (smaller), and why? (2p)

4. In a North-American cross-section data set with 100,000 of families, log income is regressed on a binary variable that is one if a family lives in Canada (D_Canada) and zero if they live in USA. The slope coefficient is 0.07.

   (a) Interpret this number. (2p)

   (b) When we add to the regression another binary variable that is one if the family has kids under 18 (D_Kids), and zero otherwise, the slope coefficient on the $D_C$anada binary variable is 0.15. Interpret this number. (2p)

   Now estimate the following model with the interaction:

   $(\ln \text{ income })^E = \alpha + \beta_1 D_C\text{anada} + \beta_2 D_K ids + \beta_3 D_C\text{anada} * D_2\text{kids}$

   We find $\beta_1 = 0.08, \beta_2 = -0.1, \beta_3 = 0.2$

   (c) Is having kids associated with increasing or decreasing earnings? (2p)

   (d) How does the association of having kids in the US vs Canada differ with earnings? (2p)

   (e) Do Americans without kids make more or less than Canadians, and by how much? (2p)

5. What is the advantage and disadvantage of the linear probability model (LPM) and the logit model? (3p)

   For which kind of task shall we use LPM and when shall we use logit? (2p)

6. You have 15 years of daily data. In daily time series data the average afternoon temperature (in degree Celsius) in Madrid, Spain is regressed on the log number of tourists landing at the airport.

   (a) The slope coefficient is -0.05. Interpret this coefficient and try to make sense of it. (3p).

   (b) You now add monthly dummies to the regression and find a slope coefficient of -0.01. How would you interpret this number? (2p)

2 Part II. Multiple choice

Pick one answer for each question. The right answer is 3p. Wrong or missing answer is 0p.

Q1: Which of the following statements is ALWAYS TRUE about the coefficient on a binary x variable in a simple regression $y^E = a + bx$?

- It shows the difference in average y between the $x = 1$ group and the $x = 0$ group.
- It shows average y in the $x = 1$ group.
- It shows average y in the x = 0 group.
- It shows the effect of increasing x from 0 to 1 on average y.

Q2: Which of the following makes the SE of the slope coefficient estimate in a linear regression of y on x smaller?

- The smaller the R-squared of the regression.
- The larger the standard deviation of x.
- The smaller the number of observations.
- The larger the standard deviation of the regression residuals.

Q3: You are estimating a linear regression of y on x. What is a false positive of a test with the null hypothesis that the slope coefficient is zero?

- I accept that the slope is zero and it is not zero in reality.
- I reject that the slope is zero and it is not zero in reality.
- I reject that the slope is zero and it is zero in reality.
- I accept that the slope is zero and it is zero in reality.

Q4: In a linear probability model using cross-sectional data, smoking is the y variable, the x variables are age, gender (1 if female, 0 if male), and their interaction. The estimated coefficient on age is -0.01, the estimated interaction coefficient is 0.03, the estimated coefficient on female is -0.15. Interpret this age coefficient estimate of -0.01.

- Men who are one year older are less likely to smoke, by 1 percentage point.
- People who are one year older are less likely to smoke, by 1 percentage point.
- Men are less likely to smoke than women of the same age, by 1 percentage points.
- People who are one year older are less likely to smoke, by 0.01 percentage point.

Q5: Consider two time series, x(t) and y(t). If both have a positive time trend...

- regressing y(t) on x(t) may give misleading results.
- instead of regressing y(t) on x(t), the analyst should model first difference instead.
- trends may give rise to spurious correlation.
- All the above are true.