université de BORDEAUX

MAG2
MAGEVAL2

---

11 April 2024 10h50 to 12h05

Econometrics

*Please answer **four out of five** questions in Stata. If you complete all five exercises, then your **best four** answers will be counted for the final grade. There are also a few **bonus tasks** (marked by an *) which are **not mandatory** to answer, but if you provide the answer, then these will be counted towards the final grade. Each question is worth 20 points, so you can get 80 points in total without the bonus questions.*

*Please answer the questions **as comments** (for example, using *) in your do-file. Your do-file should contain necessary codes and comments on results to fully answer the questions. In case your code does not work for some reason, please simply explain how you could have answered if the code worked well. Where appropriate, please round your results with two decimal places.*

*You have until 12h15 to send your do-file by email to **niclas.knecht@u-bordeaux.fr**. Please put your **first and last name** in the name of do-file (e.g. Niclas_Knecht.do).*

Good luck!

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator.

STATIONERY REQUIREMENTS
None

SPECIAL REQUIREMENTS
None

1 Heteroskedasticity. Omitted variable bias.

We are interested in the returns to education. Using **exam_data1.dta**, estimate the following model:

$$\begin{aligned} log(wage) = &\alpha + \beta_1\, educ + \beta_2\, exper + \beta_3\, tenure + \\ &\beta_4\, married + \beta_5\, south + \beta_6\, urban + u \end{aligned}$$

where *wage* is monthly earnings in USD (*lwage* the log version of *wage* in the data), *educ* years of education, *exper* years of professional experience, *tenure* years of tenure with the current employer, *married* a dummy for marital status, *south* a dummy for living in the south, and *urban* a dummy for those living in urban areas.

(a) Create a scatter plot of residuals against fitted values. What can you infer from the pattern you observe in the graph? Do you observe heteroskedastic or homoskedastic error terms?

Additionally, run White's test for heteroskedasticity. What do you conclude?

[6 marks]

(b) Re-estimate the model with heteroskesdacity-robust standard errors. Comment on the standard errors of the coefficients. Interpret the relevant coefficient.

[4 marks]

(c) Do you suspect that there is omitted variable bias in this model? If yes, explain how it can impact the coefficient estimate for *educ*. [5 marks]

(d) Re-estimate the model above by adding the variables IQ and KWW as a proxy for ability. What happened to the coefficient on *educ*? Comment and interpret.

[5 marks]

(e) * Bonus question: If we cannot control for an important variable because it is not observed or cannot be accurately measured, how we can generally fix the issue of omitted variable bias? [2 marks]

2 Non-linear functions. Interaction terms.

We are interested in the effect of house- and lot size on the price of the house. Estimate the following model of house prices (*price*) as a function of lot size (*lotsize*) and the size of the house (*sqrft*) with **exam_data2.dta**:

$$log(price) = \alpha + \beta_1 lotsize + \beta_2 sqrft + u$$

(*a*) Interpret the relevant coefficients. [4 marks]

(*b*) Now estimate the model with log-transformed versions of *lotsize* and *sqrft* (log-log model). Interpret the coefficients. [4 marks]

From now on continue with the model from (*b*).

(*c*) Add the variable *colonial* (a dummy for colonial style house) to the regression. How can you interpret the coefficient for colonial? [3 marks]

(*d*) Now add an interaction term between *log(sqrft)* and *colonial*. How can you interpret the coefficient on the interaction term (ignoring the statistical significance for now)? [5 marks]

(*e*) What is the total estimated effect of *log(sqrft)* for $colonial = 0$ and $colonial = 1$? [4 marks]

(*f*) * Bonus question: In some models, we need to add squared terms. Do you think we need to add squared terms here? If yes, estimate the model from (*b*) with squared terms. [1 mark]

(*g*) * Bonus question: In which cases it does not make sense to do log-transformation of variables? [1 mark]

3 Dummy variables. Linear probability models.

Use the data in **exam_data3.dta** for this question. The binary variable to be explained is *approve*, which is equal to one if a mortgage loan to an individual was approved. The key explanatory variable is *white*, a dummy variable equal to one if the applicant is White. The other applicants in the data set are Black and Hispanic. To test for discrimination in the mortgage loan market, a linear probability model can be used:

$$approve = \beta_0 + \beta_1\, white + \text{other factors.}$$

(a) If there is discrimination against non-whites, and the appropriate factors have been controlled for, what is the sign of β_1? [4 marks]

(b) Regress *approve* on *white* and report the results. Interpret the coefficient on *white*. Is it statistically significant? [4 marks]

(c) As controls, add the variables *obrat*, *loanprc*, *unem*, *male*, and *married*. What happens to the coefficient on white? Is there still evidence of discrimination against non-Whites? [4 marks]

(d) Now allow the effect of race to interact with the variable measuring other obligations as a percent of income (*obrat*). Interpret the interaction term. [4 marks]

(e) Using the model from (d), what is the effect of being white on the probability of approval when $obrat = 12$ and $obrat = 40$ (the others are at their respective means)? Obtain a 95% confidence interval for these effects. [4 marks]

(f) * Bonus question: Give one example of another method for estimating the equation. Replicate (e) with this method. [2 marks]

4 Panel data regression.

The file **exam_data4.dta** includes panel data on House of Representative elections in 1988 ($time = 1$) and 1990 ($time = 2$). Only winners from 1988 who are also running in 1990 appear in the sample; these are the incumbents (i.e., the current members of the House). Incumbent identifier variable is *id*, and time identifier is *time*. We suggest the following model that explains the share of the incumbent's vote in terms of their share of total election campaign expenditures and characteristics of the incumbent:

$$vote_{it} = \alpha + \beta_1 incshr_{it} + \beta_2 democ_i + \beta_3 prtystr_i + \delta_1 d90_t + u_{it}$$

where $vote_{it}$ is the share of incumbent's vote, $incshr_{it}$ is the incumbent's share of total campaign spending, $democ_i$ is a dummy indicating whether incumbent is Democrat, $prtystr_i$ is the strength of the competing party (the share of vote in the previous election), and $d90_t$ dummy for 1990.

(a) Which incumbent had the biggest positive change in share of votes? Who had the biggest negative change in share of votes? [4 marks]

(b) Estimate this model using pooled OLS. Interpret the coefficient estimates for β_1 and β_2. [4 marks]

(c) What could be problematic with this estimated effect? [4 marks]

(d) Now estimate the model using first difference (between two periods). Explain why some variables dropped out. Compare the coefficient estimates for β_1 with estimates from (b) and explain if there are any differences. [8 marks]

(e) * Bonus question: Can you think of one or more potentially important, time-varying variables that have been omitted from the analysis? [2 marks]

5 Instrumental variable regression.

Using **exam_data5.dta**, consider the following model of returns to education:

$$log(wage) = \alpha + \beta_1\, educ + \beta_2\, exper + \beta_3\, black\ + \beta_5\, north + \beta_6\, city + u$$

where *wage* is monthly earnings in USD (*lwage* the log version of *wage* in the data), *educ* years of education, *exper* years of professional experience, *black* a dummy for Black individuals, *north* a dummy for living in the north, and *city* a dummy for those living in urban areas.

(a) Estimate the model with simple OLS. Interpret the coefficients. [4 marks]

(b) Explain why β_1 might be biased and how an instrumental variable may help. In particular, talk about the underlying assumptions of instrumental variable regressions. [6 marks]

(c) Estimate the regression of *educ* on *brthord* (birth order, i.e., $brthord_i = 1$ if i is the first born, $brthord_i = 2$ if i is the second child, etc.). Do you think *brthord* is a good candidate for an instrumental variable for *educ*? Explain. [4 marks]

(d) Estimate the model from (a) but using *brthord* as an instrumental variable for *educ*. Compare the IV estimates of *educ* with the OLS estimates from (a). Comment. [6 marks]

(e) * Bonus question: Can you suggest (hypothetically) any other instrumental variable for *educ*? Why might this variable be a good candidate for an instrumental variable? [2 marks]

END OF PAPER