NIFB22001U Introduction to Econometrics

2022/2023

72 hour take-home exam

22nd August, 10:00 – 25th August, 10:00

All aids allowed

The exam contains 9 pages

Your assignment should be handed in as a `.pdf` file through the Digital Exam portal at KUnet no later than 10:00 AM on Friday, the 25th of August. Exam answers should be provided in English. Remember to **not** include any `R`-code in your assignment, but provide it as an appendix or a separate file. Your answers will solely be judged on the provided answers and tables, not your `R`-scripts. Keep in mind that econometrics should be independent of the statistical software used. The point is that any readers of your work should be able to replicate your work, from start to finish, using any statistical software. Therefore, only carrying out your analysis in `R` is not sufficient. You should provide **formal answers** to each asked question.

Please remember to read each question carefully and thoroughly.

The exam consists of 9 pages with two parts. Your exam as a whole will determine your final grade. In addition to the exam questions, you have been provided with some data in the included files: `examdata2023_01.RData` and `examdata2023_02.RData` (or the equivalent `.txt` versions of the same data), which you will be using for the empirical questions.

Remember to avoid plagiarism! You can find the official university policy on plagiarism and exam cheating on KUnet.

Examples of plagiarism and exam cheating are if you:

- Copy other people's texts without making use of quotation marks and reference to the source, so that it may appear to be your own text
- Use the ideas or thoughts of others without making use of source referencing, so it may appear to be your own idea or your thoughts
- Reuse parts of a written paper that you have previously submitted and for which you have received a passing grade without making use of quotation marks or source references (self-plagiarism)

Part 1: The Returns to Education

In the following, you will be studying the effect of an individual's years of education on their hourly wage. However, we are concerned that unobserved factors, such as motivation or innate ability, may simultaneously affect education and hourly wages. The data is available from `examdata2023_01.RData` and its included variables are described further below:

Data: For each individual $n = 1,000$, we have the following variables:

- *wage*: Hourly wage in USD
- *educ*: Years of education
- *library*: Number of public libraries within a 5-km radius of the individual's childhood residence

1. Consider the following model:

$$wage = \alpha + \beta educ + u. \tag{1}$$

   Which assumptions are required in order for β to be a consistent and unbiased estimator of the causal effect of education on hourly wages? Explain each assumption in detail.

2. Estimate model (1) and interpret your results. Do the assumptions you presented in the previous question seem reasonable? Discuss.

3. Why might model (1) suffer from endogeneity? Provide a few examples.

4. What makes a good instrumental variable? Explain how the number of public libraries within a 5-km radius of the individual's childhood residence, *library*, could be a valid instrument for *educ*.

5. Implement a two-stage least squares (2SLS) regression to estimate the supposed causal effect of years of education on hourly wages. Interpret your results in comparison to model (1).

6. Discuss whether the assumptions needed for *library* to be a valid instrument for *wage* seem reasonable.

Part 2: Global Food Consumption Patterns

2.1 Data description and preliminary analysis

The data in `examdata2023_02.RData` contains information on consumption patterns and various macroeconomic indicators for a range of countries in the year 2018. The data stems from various sources like The World Bank's World Development Indicators and the FAO's Food Balance Sheets. The data is described in table 1 below:

Variable	Description
`Country`	Name of the Country
`Year`	Year
`CalTotal`	Total per-capita consumption (kcal/capita/day)
`CalAnimal`	Consumption of animal products (kcal/capita/day)
`CalVegetal`	Consumption of vegetal products (kcal/capita/day)
`PropCalAnimal`	Proportion of animal products in total consumption (%)
`PropCalVegetal`	Proportion of vegetal products in total consumption (%)
`AgLandShare`	Agricultural land (% of land area)
`AgGdpShare`	Agriculture, forestry, and fishing, value added (% of GDP)
`ArableLandShare`	Arable land (% of land area)
`GdpPerCapita`	GDP per capita, PPP (current international $)
`GiniCoef`	Gini coefficient
`HCI`	Human capital index (HCI) (scale 0-1)
`LifeExp`	Life expectancy at birth, total (years)
`PopYoung`	Population ages 0-14 (% of total population)
`PopMiddle`	Population ages 15-64 (% of total population)
`PopOld`	Population ages 65 and above (% of total population)
`PopFemale`	Population, female (% of total population)
`PopMale`	Population, male (% of total population)
`PopRural`	Rural population (% of total population)
`PopUrban`	Urban population (% of total population)

Table 1: *Variables in the data set* `examdata2023_02.RData`

In this assignment, you will be using the data in `examdata2023_02.RData` to investigate

the relationship between the proportion of calories from animal products and various macroeconomic and sociodemographic indicators for the included countries. You are only expected to use the supplied data and should not spend time gathering data from other sources.

1. Describe the supplied data. This description should include a table of descriptive statistics for all variables in the data set.

2. The supplied data contains missing data points. Does this affect the representativeness of the sample?

3. How may a nonrandom sample affect any of our OLS estimates? Discuss based on the missing observations you identify. Finally, you should justify whether the missing observations are *missing completely at random*, *missing at random*, or *missing not at random*.

Regardless of your answer to the previous two questions, you should proceed with removing observations with missing values in the variable `PropCalTotal` from the data set.

4. Assume that the consumption of vegetal and animal products are normal goods. Furthermore, suppose that the consumption of animal products is a luxury good. Oppositely, let us assume that the consumption of vegetal products is a necessity. How would you expect the proportion of calories from animal and vegetal products to be related to the GDP per capita? Explain your reasoning.

5. Investigate the relationship between the proportion of calories from animal and vegetal products against the GDP per capita graphically. What do you find? Explain whether people seem to be substituting vegetal products with animal products as their income increases.

6. Another hypothesis is that due to pro-environmental and pro-climate preferences, people in countries with high human capital at a certain point decide to consume

fewer animal products in favor of vegetal products. Given the available data, how would you examine this hypothesis in further detail?

2.2 Regression analysis

1. Compare the distributions of $PropCalAnimal$ and $\log(PropCalAnimal)$. Which of the two variables would you prefer to use in a regression analysis? Explain your reasoning.

2. Estimate the following model using OLS

$$\begin{aligned} PropCalAnimal = \beta_0 &+ \beta_1 AgLandShare + \beta_2 AgGDPShare \\ &+ \beta_3 ArableLandShare + \beta_4 GdpPerCapita \\ &+ \beta_5 HCI + \beta_6 LifeExp + \beta_7 PopMiddle \\ &+ \beta_8 PopOld + \beta_9 PopFemale + \beta_{10} PopUrban + u. \end{aligned} \tag{2}$$

   Do you find it likely that the Gauss-Markov assumptions apply to model (2)? Furthermore, in terms of model (2), what would this imply about the estimated parameters?

3. Re-estimate model (2) with $\log(PropCalAnimal)$ as the new dependent variable:

$$\begin{aligned} \log(PropCalAnimal) = \beta_0 &+ \beta_1 AgLandShare + \beta_2 AgGDPShare \\ &+ \beta_3 ArableLandShare + \beta_4 GdpPerCapita \\ &+ \beta_5 HCI + \beta_6 LifeExp + \beta_7 PopMiddle \\ &+ \beta_8 PopOld + \beta_9 PopFemale + \beta_{10} PopUrban + u. \end{aligned} \tag{3}$$

   How does the interpretation of the estimated parameters change when going from model (2) to model (3)? For instance, how would you interpret the estimated coefficient on $AgLandShare$ and $LifeExp$ in model (3) as compared to model (2)?

4. Compare the distribution of residuals from model (2) and model (3). Which model do you prefer and why? Are there any issues related to taking the logarithm of the dependent variable in model (3)?

5. Carry out a test for conditional heteroskedasticity in model (2) and model (3). What do you conclude?

6. Using model (2), test for the joint significance of $AgLandShare$ and $AgGdpShare$. State the relevant null- and alternative hypotheses carefully. Why might carrying out the same test in model (3) be problematic?

7. Now, estimate the model:

$$\begin{aligned} PropCalAnimal = \beta_0 &+ \beta_1 ArableLandShare + \beta_2 \log(GdpPerCapita) \\ &+ \beta_3 HCI + \beta_4 LifeExp + \beta_5 PopMiddle \\ &+ \beta_6 PopOld + \beta_7 PopFemale + \beta_8 PopUrban + u. \end{aligned} \tag{4}$$

   In model (4), how do you interpret the estimated coefficient on $\log(GdpPerCapita)$? What does this imply about the relationship between the proportion of calories from animal products and the GDP per capita?

8. Formally test the hypothesis that the proportion of calories from animal products is increasing until a certain level of human capital (HCI), at which point it decreases by using and eventually extending model (4). Remember to state the relevant null- and alternative hypotheses carefully. Discuss your findings. What do you conclude about the initial hypothesis from question five in the preliminary analysis?

2.3 Model selection

In this part, you are asked to discuss the performance of the three different models you should have estimated in the previous part. Your discussion should, at minimum, include the following:

- A table with the estimated coefficients, standard errors, significance levels, and adjusted R^2 values. Furthermore, the table should include the results of your tests for conditional heteroskedasticity and the test statistic from a RESET test of each model.

- A discussion of the results based on the table. For instance, which model do you prefer and why? Furthermore, following the *general-to-specific* approach, you should consider whether any variables can be removed from your model of choice.

- Finally, you should discuss any limitations you came across during your analysis in parts 2.1 and 2.2 and conclude upon your results. Are there any variables you would have liked to include in your analysis? If so, why? And, are the missing observations likely to influence your results?