NIFB19000U: Qualification Course in Econometrics

2021/2022

Take-home mock exam

23rd August, 10:00 – 26th August, 10:00

All aids allowed

The mock exam contains 8 pages

~~Your assignment should be handed in as a `.pdf` file through Digital Eskamen at KUnet no later then 10:00 AM on Friday the 27th of August.~~ Mock exam answers should only be provided in English. Remember that any code used throughout the assignment should be provided as an appendix or a separate file and not included in your assignment. Your answers will solely be judged on the provided answers, not your `R`-scripts. ***Please remember to read each question carefully and thoroughly.*** The exam consists of 8 pages with 3 parts, 14 questions, and a more open-ended assignment in part 3. Each part will count toward your grade. In addition to the mock exam questions, you have been provided with some data in the included file: `mock_examdata.RData` or `mock_examdata.txt`, which you will be using for the empirical questions.

Credence Goods: Demand for Organic Vegetables

The Economic Setting

Credence goods are defined as having certain characteristics that cannot be seen or experienced directly by the consumer. Therefore, consumers lack important information about attributes of the product - even after consumption. Many organic products are referred to as credence products because information about their core attribute (their organic or non-organic status) is asymmetric. That is, consumers may struggle in identifying whether or not the bought product is organic, if this isn't explicitly stated. Therefore, consumers form individual perceptions about the characteristics contained in the organic good, and these perceptions may vary between consumers. Some consumers perceive organic goods are healthier; others perceive them as being better for the environment. Labelling is one way of providing information for consumers to make them distinguish between products of varying quality when these variations are of a credence nature, hence organic-labels help consumers in distinguishing the organic products from the non-organic ones. Trust in the label is an important factor since people who distrust the message, might not consider labelled products as different from other products and they might even feel cheated and therefore prioritize unlabelled products. Denmark is one of the few countries where the organic certification and labelling system is governmental. The state organic label, the 'Ø' label, was introduced in 1989 and is the sole national organic label. The organic label is recognized by almost all Danish consumers. The high level of confidence and salience associated with the organic label in Denmark, is likely linked to the fact that it is governmental, but it is also influenced by the Danes' relatively high level of trust in public authorities. The demand for organic products has increased considerably in Denmark in the last decades, but the Danish Ministry of Agriculture wants to further increase the demand for organic vegetables.

Therefore, policymakers and politicians have a direct interest in better understanding why some consumers prefer organic produce. The Danish Ministry of Agriculture has therefore asked you to answer a bunch of questions on econometrics, having

armed you with a data set from GfK ConsumerTracking Scandinavia in the file `examdata.RData`. Given the limited funding available to The Danish Ministry of Agriculture, they are interested in knowing which persons they should be targeting, when trying to increase the consumption of organically produced vegetables in Denmark. In particular, they are interested in knowing whether or not the demand for organic vegetables is driven primarily by opinions on the credence characteristics associated with organic goods, or by household characteristics.

Data Description

The supplied data is procured by GfK ConsumerTracking Scandinavia, who maintain a panel of Danish households who report their day-to-day food purchases. It's important to note that the data only covers within-household consumption. The households select themselves into the panel voluntarily, but GfK try to maintain a balanced panel with respect to geography, education level, income, and so on. The incentive to participate is receiving prizes and gift cards after having participated in the panel for a certain amount of time. The data set you are working with contains information on each household's average monthly consumption of organic and non-organic vegetables, respectively. Each observation is therefore an individual household, the share of their vegetable purchases that are organic, and their associated characteristics.

In addition to reporting their purchases, a requirement for partaking in the GfK panel is filling out a survey on household characteristics such as: income, education level, household composition, and so on. These survey results pertain to the *primary shopping responsible* of the household. There are 2215 participating households aggregated across the entirety of 2020. The following is a description of the included variables:

Variable	Description
kids06	number of kids in the age group 0-6 years
kids714	number of kids in the age group 7-14 years
kids1520	number of kids in the age group 15-20 years
hh_inc	income group for the household, where 14 is the highest
organic_tastes_better	$= 1$ if the household believes organic products taste better, 0 otherwise
organic_buyer	$= 1$ if the household perceives itself as a primarily organic household, 0 otherwise
pay_more_organic	$= 1$ if the household is willing to pay more for organic produce, 0 otherwise
organic_salience	$= 1$ if the household believes organic products should be more salient, 0 otherwise
female	$= 1$ if the primary shopping responsible is a woman, 0 otherwise
age	age in years for the primary shopping responsible
rural	$= 1$ if the household lives in a rural area, 0 otherwise
urban	$= 1$ if the household lives in an urban area, 0 otherwise
capital	$= 1$ if the household lives in the capital area, 0 otherwise
low_educ	$= 1$ if you have a vocational or no education, 0 otherwise
med_educ	$= 1$ if you have a short- or medium-cycle higher education, 0 otherwise
high_educ	$= 1$ if you have a long-cycle higher education, 0 otherwise
org_vol_share	volume share of total vegetables bought in 2020 that are organic.

**Based on survey results issued to the participating panelists. The answers belong to that of the primary shopping responsible. In the survey, they are asked to which degree they agree with some statements about organic products.*

Part I: Working with Economic Data

1. Provide a description of the data we are working with, along with a table of some key descriptive statistics. At minimum this description should reflect the following:

    - A discussion of the data collection method. Is it likely error prone? And how could this influence our results later?
    - Whether or not *self-selection* seems to be a problem. Furthermore, how could this affect the external validity of your study?
    - Would you consider the data to be representative of the general Danish population?

    Hint: *Remember that you can use the package* `stargazer` *to easily produce tables with descriptive statistics directly from* `R` *to both .tex and .txt formats.*

2. What is the mean organic share of vegetables consumed in rural, urban, and capital areas? Does this imply that living in certain areas increases the purchased share of organic vegetables? Comment.

3. Do those that identify as an organic buyer ($organic\_buyer = 1$):

   (a) have a higher willingness to pay for organic goods?

   (b) believe organic products taste better?

   (c) believe organic goods should be more visible in supermarkets?

4. How many of the participating households have reported that they have not purchased any organic vegetables throughout the entire year $org\_vol\_share = 0$?

5. Compare the densities of $\log(org\_vol\_share)$ and org_vol_share. What are some of the (possible) advantages of taking the $\log$ of dependent variables in multiple linear regressions?

   Hint: *Seeing as there are multiple observations of $org\_vol\_share = 0$ and the natural logarithm hereof is undefined you may want to use the following snippet of code when taking logs:*

```
log(data$org_vol_share[data$org_vol_share > 0])
```

Part II: Econometric Analysis

6. Estimate the two *linear probability models*[1]:

$$\begin{aligned} organic\_buyer = \beta_0 + \beta_1 hh\_inc + \beta_2 female + \beta_3 age \\ + \beta_4 urban + \beta_5 capital + \beta_6 edu\_med + \beta_7 edu\_long + u, \end{aligned} \quad (1)$$

1 You can read more about the linear probability model in section 7.5 on pages 239-244 and section 8.5 on pages 284-285 in Wooldridge (2018).

$$
\begin{aligned}
organic\_buyer = \beta_0 &+ \beta_1 hh\_inc + \beta_2 female + \beta_3 age \\
&+ \beta_4 urban + \beta_5 capital + \beta_6 edu\_med + \beta_7 edu\_long \\
&+ \beta_8 organic\_tastes\_better + \beta_9 pay\_more\_organic \\
&+ \beta_{10} organic\_salience + u, \qquad (2)
\end{aligned}
$$

and report your findings in a regression table. Discuss differences in the sign, magnitude, and statistical significance of the variables in the two models. Based on this discussion, which factors are the most important determinants of being a buyer of organic products? Remember to carefully state relevant null- and alternative hypotheses.

7. Discuss the advantages and disadvantages of using *linear probability models*. In addition, explain why heteroskedasticity is unavoidable in model (1) and (2).

8. Consider the multiple linear regression:

$$
\begin{aligned}
org\_vol\_share = \beta_0 &+ \beta_1 hh\_inc + \beta_2 female + \beta_3 age \\
&+ \beta_4 urban + \beta_5 capital + \beta_6 edu\_med + \beta_7 edu\_long \\
&+ \beta_8 organic\_tastes\_better + \beta_9 pay\_more\_organic \\
&+ \beta_{10} organic\_salience + \beta_{11} kids06 + \beta_{12} kids714 \\
&+ \beta_{13} kids1520 + u, \qquad (3)
\end{aligned}
$$

estimated by OLS. Are the assumptions that ensure the unbiasedness and consistency of the OLS estimator reasonable in the proposed model? Reflect on your answer.

9. Under which assumptions can we carry out standard inference (t-tests and F-tests)? Do these assumptions seem reasonable in our specified model? Discuss.

10. Test for the joint significance of all the dummy variables related to the education level of the primary shopping responsible in (3). Remember to carefully state your null- and alternative hypothesis. Based on your result, would you choose to omit either edu_med or edu_long (or both)? Comment briefly.

11. Why are the variables $rural$ and edu_short omitted from the proposed model in (3)? Moreover, how does this affect the way we can interpret the estimated intercept $\widehat{\beta}_0$? How would we interpret the coefficients $\widehat{\beta}_4$, $\widehat{\beta}_5$, $\widehat{\beta}_6$, and $\widehat{\beta}_7$?

Note: Removing observations

Seeing as households that do not purchase organic vegetables at all, likely have little explanatory power on the sociodemograhpic indicators associated with being a frequent purchaser of organic vegetables, we will exclusively be estimating on data where $org\_vol\_share > 0$ for the remainder of this assignment. In R you can do this using the following code:

```
data = subset(data, org_vol_share > 0)
```

12. Now, consider a model with $\log(org\_vol\_share)$ as the dependent variable:

$$\begin{aligned}\log(org\_vol\_share) = \beta_0 &+ \beta_1 hh\_inc + \beta_2 female + \beta_3 age \\ &+ \beta_4 urban + \beta_5 capital + \beta_6 edu\_med + \beta_7 edu\_long \\ &+ \beta_8 organic\_tastes\_better + \beta_9 pay\_more\_organic \\ &+ \beta_{10} organic\_salience + \beta_{11} kids06 + \beta_{12} kids714 \\ &+ \beta_{13} kids1520 + u, \end{aligned} \tag{4}$$

also estimated by OLS. Compare the distribution of the estimated residuals from models (3) and (4). Is this model closer to fulfilling the assumption a normal error term ($u \sim \mathcal{N}(0, \sigma^2)$) or that of asymptotic normality?

13. Carry out a test for heteroskedasticity on models (3) and (4). Use the F-statistic (or χ^2) and report the p-value. You should also briefly present the basic idea behind the test. Does heteroskedasticity in either of the two models seem to be a problem? What are some of the consequences of having persistent heteroskedasticity?

14. Test for functional form misspecification in models (3) and (4) with the RESET-test. How do your results compare?

Part III: Reporting and Discussing Results

Before moving on to discussing your results, you should present them in a regression table. Remember to include relevant test statistics and follow the general framework we have discussed when reporting regression results in econometrics. Having done so, you can move on to discussing and analyzing your results on the basis of your regression table. This should include:

- Reporting whether or not any of the included variables can be omitted from your model of choice. When carrying out any hypothesis test, you should always remember to state the relevant null hypothesis, test statistic, and ensuing p-value. Then you should re-estimate the model and report it in your new regression table. Why is this model superior to the others?
- The limitations of your analysis and the issues you may have come across in determining the correct functional form and the persistence of heteroskedasticity in your model. In particular, this should focus on how this may affect the validity of your results.
- Are there any data issues (outliers, influential observations, missing data, and so on) that may have influenced your results? If so, was there anything you could have done differently to remedy those problems?
- Lastly, when you are asked to present an executive summary to the Danish Ministry of Agriculture; what are the main takeaways of your analysis? What characterizes a frequent consumer of organic vegetables? You can also choose to include the results from the *linear probability models* (1) and (2).