How to work with your data sets

Biodiversity and resource economics: Seminar Choice Experiments

2022-12-20

Task

The final product is a short paper (2500 to 3500 words including tables) which reports results from the choice experiment from the dataset you have selected. The paper should have the following structure:

1. Introduction: Briefly explain the topic and the aim of your analysis. Look into the additional material (codebook, paper). You find this material in the folder where you downloaded the data (max. 500 words).
2. Methods: Describe the method you use. You need not explain each step to derive the model you use. It is more important that your approach becomes clear and how it is specifically differs from the main modelling approaches. This includes writing down your utility function with all parameters and variables used (max 1000 words).
3. Results: Present your results in a comprised manner, use one or two tables and figures, but not more (max. 1500 words).
4. Discussion and conclusion: Briefly discuss your results and draw some general, maybe policy relevant conclusions (max. 500 words).

Everything related to this paper should be saved in one folder. The folder should be an RStudio Project (it contains an .Rproj file). The paper should be written in RMarkdown or Quarto and the final product should be a pdf. To submit your work, bring the printed PDF to the exam, and upload the whole RStudio Project to Github (or another Git Hosting website). Put the link on Moodle somewhere in the announcements. Make sure you have already finished all courses on Datacamp before you start working on the paper. Please always try to work with `dplyr` and other `tidyverse` packages for data cleaning and with `apollo` for modelling. Also use google to find answers to your questions.

Some Guidance

1. Create a new R project

To start with, make a new RStudio project, which you can call for example: *Project_ResourceEco* or something else. In this project create a folder called *Data* where you put the data set you have downloaded from nextcloud. Then, create a new folder called "Scripts". Also create a new .Rmd or .qmd file. This file will be your main file from which you call all your scripts using the `source` function and where you write the content of your paper. It is your choice how much code you put directly in the rmd or qmd file and how much code you source from external scripts. However, it is not advisable to have all code embedded in the main file.

Before you read in the data you should load the packages you want to use later. We recommend you to work with these packages:

```
library(apollo)
library(tidyverse)
library(haven)
library(labelled)
```

Note that you must install the packages before you can load them you can do this with:

```
# Note: Hashtags start a comment and must be removed if you want to excecute code

# install.packages(c("apollo", "tidyverse", "haven", "labelled"))
```

Now we can read in the data like this[1]:

```
#If you have a csv file:

# database <- read.csv("data/filename.csv")

#If you have a rds file:

# database <- readRDS("data/filename.rds")
```

It is part of your task to create a well structured, reproducible and readable code. You can use several resources in the internet to help you structure a project. Here are some resources:

- https://libscie.github.io/rmarkdown-workshop/handout.html
- https://bookdown.org/thea_knowles/dissertating_rmd/
- https://kdestasio.github.io/post/r_best_practices/
- https://resources.github.com/github-and-rstudio/

[1] It is useful to directly read in the data with the name *database* as the `apollo` package later requires a data frame called *database*.

2. Inspect the Datasets

After you have successfully read in the data you should start inspecting the dataset. Identify how the dataset is structured and which variables are included. Some useful commands for this are `View`, `names`, `summary`, `table`, `boxplot`, `arrange`, and `hist` to name a few. Remember that if you don't know how a command work or have some issues excecuting commands you can always use:

```
?names
?summary
```

which will give you information about the command and its syntax.

Inspecting the dataset you should identify the choice variable, the identifier variable, the attributes of the choice experiment and the socio-economic variables. To help you with this you find specific guidance for your dataset below.

2.1 Land Use Datasets

After loading the dataset let's take a look at the variables:

```
View(database) # gives you an overview of the data in spreadsheet-style
names(database)
```

```
##  [1] "Kreiskennzahl"   "CountyName"      "Respondent_ID"   "Choicesituation"
##  [5] "Duration"        "dDatum"          "f001"            "f002"
##  [9] "f003"            "f004"            "f006"            "f007"
## [13] "f0081"           "f0082"           "f009"            "f033"
## [17] "f034"            "f035"            "f037"            "f038"
## [21] "f039"            "PINC"            "choice"          "ASCsq_1"
## [25] "ASCsq_2"         "ASCsq_3"         "wald_1"          "wald_2"
## [29] "wald_3"          "groe_1"          "groe_2"          "groe_3"
## [33] "prei_1"          "prei_2"          "prei_3"          "FoUnt_1"
## [37] "FoUnt_2"         "FoUnt_3"         "ConSh_1"         "ConSh_2"
## [41] "ConSh_3"         "HaAge_1"         "HaAge_2"         "HaAge_3"
## [45] "SQwald_1"        "SQwald_2"        "SQwald_3"        "SQwald2_1"
## [49] "SQwald2_2"       "SQwald2_3"
```

choice is the **choice variable** here, by using `table` we get information on how often each alternative has been choosen.

```
table(database$choice)
```

```
##
##    1    2    3
## 2450 2669 6284
```

Respondent_ID is the **identifier variable**. With:

```
length(unique(database$Respondent_ID))
```

```
## [1] 1267
```

we get the numbers of respondents.

As we can see there are some variables which appear three times. These are the attributes from the choice experiment, in this example there are **six attributes**: *wald_*, *groe_*, *prei_*, *FoUnt_*, *ConSh_* and *HaAge_*. Every land use data set has six attributes. *ASCsq_* is not an attribute, but a dummy variable for the status quo.

Important: Every land use dataset includes the variables *SQwald_* and *SQwald2_* which are modified attributes of the forest (*wald_*) attribute including the individual status quo of the respondents (share of forest within 15km of the place of residence) and the squared share (*SQwald2_*)[2]. If you include *SQwald_* in your logit model you cannot include the attribute *wald_*, you must decide for one of the two attributes.

The number behind an attribute is indicating the alternative, *prei_2* is for instance the value of the price attribute for the second alternative, which of course varies among the different choices, what we can see using the following command:

```
summary(database$prei_2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   25.00   50.00   72.45  110.00  160.00
```

The remaining variables are socio-economic and other variables. To find out what the variable *f006* refers to you can use:

```
var_label(database$f003) # gives you the label of a variable
```

```
## [1] "Schule"
```

```
val_labels(database$f003) # gives you the levels of the variable
```

```
##  Sekundarabschluss I Sekundarabschluss II       Hochschulreife
##                    1                     2                    3
##   Hochschulabschluss       kein Abschluss         keine Angabe
##                    4                     5                    7
```

[2]For more details see the uploaded paper of Sagebiel et al. (2017).

This works because the variable is labeled, for **unlabeled variables it won't work**. You also see the label of a variable underneath in the data viewer when you execute `View`. Unfortunately, we can see that *f003* is labeled in German, just use a translator tool to translate the names for yourself to English to make sure that you know what each variable is referring too. If you want you could also rename the variable like this:

```
database <- database %>% rename(education = "f003")
```

2.2 Urban Green Dataset

After loading the dataset let's take a look at the variables:

```
View(database) # gives you an overview of the data in spreadsheet-style
names(database)
```

```
##  [1] "id"              "cset"            "choosen1"        "Naturnähe1"
##  [5] "Erreichbarkeit1" "Miete1"          "choosen2"        "Naturnähe2"
##  [9] "Erreichbarkeit2" "Miete2"          "choosen3"        "Naturnähe3"
## [13] "Erreichbarkeit3" "Miete3"          "choice"          "h"
## [17] "City"            "Birthyear"       "Kaltmiete"       "NK"
## [21] "FlatSize"        "Balcony"         "Kitchen"         "Garden"
## [25] "GreenBackyard"   "Elevator"        "GuestWC"         "ParkingLot"
## [29] "Kleingarten"     "WalkingDistance" "VisitFrequency"  "RelaxingUGS"
## [33] "SafeUGS"         "CleanUGS"        "ImportanceUGS"   "Gender"
## [37] "IncomePresent"   "HousholdSize"    "KidsNo"          "Graduation"
## [41] "GraduationOther" "Work"            "WorkOther"       "WorkingTime"
## [45] "FamilyStatus"    "HealthCondition" "NaturalnessUGS"
```

choice is the **choice variable** here, by using `table` we get information on how often each alternative has been choosen.

```
table(database$choice)
```

```
##
##     1     2     3
##  8099  8881 31980
```

id is the **identifier variable**. With:

```
length(unique(database$id))
```

```
## [1] 4913
```

we get the numbers of respondents.

As we can see there are some variables which appear three times. These are the attributes from the choice experiment, in this example there are **three attributes**: *Naturnähe*, *Erreichbarkeit* and *Miete*. The variables *choosen* are dummy variables indicating whether an alternative has been choosen (1) or not(0). This variables could be used to include an attribute specific constant (ASC) in the model. The number behind an attribute is indicating the alternative, *Miete2* is for instance the value of the price attribute for the second alternative, which of course varies among the different choices, what we can see using the following command:

```
summary(database$Miete2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   157.2   544.5   698.9   796.8   945.0  6930.0
```

The remaining variables are socio-economic and other variables, which you should be able to identify via their names and with the help of the additional material related to your dataset.

2.3 Energy Dataset

After loading the dataset let's take a look at the variables:

```
View(database) # gives you an overview of the data in spreadsheet-style
names(database)
```

```
##  [1] "gid"              "choice"           "rate1"            "hold1"
##  [5] "tech1"            "loc1"             "admin1"           "admin_11"
##  [9] "admin_21"         "admin_31"         "admin_41"         "profit1"
## [13] "rate2"            "hold2"            "tech2"            "loc2"
## [17] "admin2"           "admin_12"         "admin_22"         "admin_32"
## [21] "admin_42"         "profit2"          "rate3"            "hold3"
## [25] "tech3"            "loc3"             "admin3"           "admin_13"
## [29] "admin_23"         "admin_33"         "admin_43"         "profit3"
## [33] "id"               "country_sample"   "age"              "gender"
## [37] "employment"       "rural"            "current_address"  "nationality"
## [41] "number_residents" "children"         "under_14"         "education"
## [45] "social_status"    "income_dist"      "invest"           "country"
```

choice is the **choice variable** here, by using `table` we get information on how often each alternative has been choosen.

```
table(database$choice)
```

```
##
##     1     2     3
##  6275  7112 10821
```

id is the **identifier variable**. With:

```
length(unique(database$id))
```

```
## [1] 3026
```

we get the numbers of respondents.

As we can see there are some variables which appear three times. These are the attributes from the choice experiment, in this example there are **four attributes**: *admin*, *rate*, *loc* and *hold*. The variables *admin_* are dummy variables indicating the admin for the alternative and are basically just a recoding of the *admin* variable. The variable *profit* is the profit each alternative would give which was calculated based on the variables *rate* and *hold* and could be also used in your logit model as an attribute. The variable *tech* is constant between alternatives, therefore it is not an attribute of the choice experiment. The number behind an attribute is indicating the alternative, *rate2* is for instance the value of the rate attribute for the second alternative, which of course varies among the different choices, what we can see using the following command:

```
summary(database$rate2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   10.00   15.97   20.00   50.00
```

The remaining variables are socio-economic and other variables, which you should be able to identify via their names and with the help of the additional material related to your dataset.

2.4 Transportation

After loading the dataset let's take a look at the variables:

```
View(database) # gives you an overview of the data in spreadsheet-style
names(database)
```

```
##  [1] "RID"                 "DESIGN_ROW"          "SCENARIO"
##  [4] "pref1"               "ASC.1"               "COST.1"
##  [7] "WIDTH.1"             "GREEN.1"             "FACIL.1"
```

```
## [10] "ASC.2"                 "COST.2"               "WIDTH.2"
## [13] "GREEN.2"               "FACIL.2"              "ASC.3"
## [16] "COST.3"                "WIDTH.3"              "GREEN.3"
## [19] "FACIL.3"               "Status"               "Kindofsample"
## [22] "FrequencyNearRadbahn" "ModePref_1"           "BikeFreqSummer"
## [25] "BikeFreqWinter"        "ModePref_3"           "ModePref_4"
## [28] "ClosestStation"        "HeardofRadbahn"       "ShouldBeBuilt"
## [31] "q45"                   "switch"               "Age"
## [34] "Gender"                "Occupation"           "Income"
## [37] "Postcode"              "Children"             "STATUS_fac"
## [40] "devicetype_fac"        "ModePref_1_fac"       "ModePref_2_fac"
## [43] "ModePref_3_fac"        "ModePref_4_fac"       "q185_1_fac"
## [46] "q185_2_fac"            "q185_3_fac"           "q185_4_fac"
## [49] "q62_fac"               "q169_fac"             "q45_fac"
## [52] "q47_fac"               "P2_fac"               "RBarea_fac"
## [55] "q184_fac"              "switch_fac"           "Age_fac"
## [58] "Gender_fac"            "Occupation_fac"       "Income_fac"
## [61] "Postcode_fac"          "children_fac"         "Age2"
## [64] "Gender2"
```

pref1 is the **choice variable** here, by using `table` we get information on how often each alternative has been choosen.

```
table(database$pref1)
```

```
##
##   1   2   3
## 248 232 180
```

RID is the **identifier variable**. With:

```
length(unique(database$RID))
```

```
## [1] 110
```

we get the numbers of respondents.

As we can see there are some variables which appear three times. These are the attributes from the choice experiment, in this example there are **four attributes**: *COST.*, *WIDTH.*, *GREEN.* and *FACIL.*. The variables *ASC.* are dummy variables which are always 1 if an alternative is not the status quo and 0 otherwise. This variables could be used to include an attribute specific constant (ASC) in the model. The number behind an attribute is indicating the alternative, *COST.2* is for instance the value of the cost attribute for the second alternative, which of course varies among the different choices, what we can see using the following command:

```
summary(database$COST.2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    30.0    50.0   120.0   126.2   200.0   300.0
```

The remaining variables are socio-economic and other variables, which you should be able to identify via their names and with the help of the additional material related to your dataset.

3. Estimate a model using `apollo`

With the help of the material from the seminar you should then be able to estimate a logit model using the `apollo` package in R. If you have any questions do not hesitate to contact us.