Mid-term Paper Assignment

Tadao Hoshino （星野匡郎）

ver. 2019 Spring Semester

Mid-term Paper Assignment

Mid-term Paper Assignment:
Write a short empirical paper using discrete choice methods.[a]

[a] It is OK (but not necessary) to use some advanced methods not taught in this course.

Requirements for the term paper

- The paper must be typed in **English**,[1] **3-5 pages** in length (not including the title page), **11-12 pt. font**, and **single-authored**.
- **The R script** used in your analysis must be attached to the paper as Appendix (just copy-and-paste the commands from the script file).
- The paper must have its **Title** and **Conclusions**. The title page must contain the **title** of your paper, your **name** and **ID**.

[1] 日本語プログラムの学生は日本語でも OK だが，英語が望ましい.

Mid-term Paper Assignment

Available datasets

- Choose **one dataset** you want to analyze from the following list:[2]
  - **BalcombeFraser2009.csv**[3]
  - **FokPaapDijk2012.csv**
  - **LiLee2009.csv**
  - **Mroz2012.csv**

  where these csv files can be downloaded from **Course Navi**. A short description about each dataset is provided below.

- This assignment accounts for 40% of your grade.
- Submission deadline: 23:59 07/02/2019 JST.
  Submit the **pdf** or **docx** file of your paper through **Course Navi**.

[2] These datasets are taken from the **Journal of Applied Econometrics Data Archive**.

[3] The same survey data used in the **R** exercise in Lecture #6. This one is the original (i.e., raw) dataset.

BalcombeFraser2009

- Type of analysis: Dichotomous choice contingent valuation
- CVM question:

> Imagine an area with a scenic overlook in a nearby federal or state public forest. In the past, this area was free with only picnic tables and a dirt parking lot. This year the area is the same as always, but it is part of the Fee Demonstration Program (described in the cover letter), so you must buy a permit or face a fine of $100 if caught without a permit. Permits are sold at a visitor's center that you pass on the way to the site.
>
> If a permit to use this area costs $_____ per visitor per day, would you buy it, keeping in mind your household income and other financial commitments?

- Bid variable: **Variable7a** $\in \{\$3, \$5, \$10\}$
- Response variable: **Variable7c** $\in \{0, 1, 2\}$, where 0 = No, 1 = Yes, 2 = Not sure
- For details about the survey and the other variables, see the included pdf files.

FokPaapDijk2012

- Type of analysis: Binary choice and/or ordered choice models
- Dependent variables:
  - **v1**: ranking of the six gaming platforms, Xbox, PS, PSP, Game Cube, Game Boy and PC. 1 to 6, 1 being most preferred, 6 being least preferred.
  - **v2**: which of the six platforms do the respondents own? 0 = not own, 1 = own.
- For details about the other variables, see the included readme file.

LiLee2009

- Type of analysis: Binary choice models
- Dependent variable:
  - **Dep_Var**: 1 if voted/intended to vote for Dole; -1 if voted/intended to vote for Clinton[4]
- A key independent variable:
  - **Sub_Expectation**: The respondent's subjective expectation of discussion group's average choice.

> *From time to time, people discuss important matters (or government, elections and politics) with other people. Looking back over the last few months, I'd like to know the people you talked with about these matters. These people might or might not be relatives. Can you think of anyone?*

> *As things currently stand, how do you think your jth discussant will vote in the 1996 presidential election? Do you think he/she will vote for Bill Clinton, Bob Dole, some other candidate, or do you think he/she probably won't vote?*

- For details about the other variables, see the included readme file.

[4] This needs to be re-coded as a dummy (i.e., 0 or 1) variable.

Mroz2012

- Type of analysis: Ordered choice models
- Dependent variables:
  - **health**: self-reported health status. (xcellent - verygood - good - fair)
- For details about the other variables, see the included readme file.

Tips

- It is advisable to report several different models and compare the results, rather than focus on only one model.
  - e.g., estimating models with different sub-samples (e.g., male vs. female, young vs. adult).
- In **R**, there are many packages that can be used to export estimation results into tables; e.g., **stargazer**.

Sample code:

```
library(stargazer)

data <- read.csv("flabor.csv")
logit_all <- glm(labor ~ ., data, family = binomial(link = "logit"))
logit_u40 <- glm(labor ~ ., data, subset = (age < 40), family = binomial(link = "logit"))
logit_o40 <- glm(labor ~ ., data, subset = (age >= 40), family = binomial(link = "logit"))

stargazer(logit_all, logit_u40, logit_o40, type = "html", out = "table1.doc",
     intercept.bottom = F, omit.stat = "aic",
     title            = "Estimation Results",
     dep.var.labels   = "Labor force participation",
     column.labels    = c("All observations", "Age < 40", "Age >= 40"))
```

Tips

The above code produces the following table in **doc** (Word) format.

Estimation Results

	Dependent variable:		
	Labor force participation		
	All observations	Age < 40	Age >= 40
	(1)	(2)	(3)
Constant	1.479*	-0.283	1.883
	(0.838)	(2.069)	(1.386)
kids6	-1.506***	-1.493***	-1.507***
	(0.199)	(0.238)	(0.467)
kids18	-0.091	-0.004	-0.208**
	(0.067)	(0.108)	(0.101)
age	-0.041*	0.055	-0.086***
	(0.022)	(0.060)	(0.033)
educ	0.275***	0.359***	0.231***
	(0.047)	(0.085)	(0.059)
husage	-0.026	-0.071**	0.009
	(0.022)	(0.034)	(0.030)
huseduc	-0.055	-0.112*	-0.013
	(0.035)	(0.060)	(0.045)
huswage	-0.057***	-0.102**	-0.041*
	(0.021)	(0.044)	(0.024)
Observations	753	298	455
Log Likelihood	-457.470	-163.123	-288.374
Note:			*p**p***p<0.01

Tips

Elimination of observations with missing values:

- We can use the `na.omit()` function for this purpose.
- In **R**, blank cells are automatically filled with NA (NA = not available):

	A	B	C	D	E
1	ID	Y	X1	X2	X3
2	1	0.09	0.77	0.33	0.06
3	2	0.74	0.23	0.49	
4	3			0.15	0.94
5	4	0.5	0.42	0.36	0.25
6	5	0.5	0.89		0.17
7	6	0.78	0.44	0.02	0.41

```
> data <- read.csv("data.csv")
> data
  ID    Y   X1   X2   X3
1  1 0.09 0.77 0.33 0.06
2  2 0.74 0.23 0.49   NA
3  3   NA   NA 0.15 0.94
4  4 0.50 0.42 0.36 0.25
5  5 0.50 0.89   NA 0.17
6  6 0.78 0.44 0.02 0.41
>
```

Tips

- The `na.omit()` function removes all rows with one or more NAs.

```
> data <- read.csv("data.csv")
> data
  ID    Y   X1   X2   X3
1  1 0.09 0.77 0.33 0.06
2  2 0.74 0.23 0.49   NA
3  3   NA   NA 0.15 0.94
4  4 0.50 0.42 0.36 0.25
5  5 0.50 0.89   NA 0.17
6  6 0.78 0.44 0.02 0.41
> data1 <- na.omit(data)
> data1
  ID    Y   X1   X2   X3
1  1 0.09 0.77 0.33 0.06
4  4 0.50 0.42 0.36 0.25
6  6 0.78 0.44 0.02 0.41
```

- If, for instance, `X2` is not used in the analysis, the 5th observation should not be discarded (the larger the sample size the more accurate the estimate).

Tips

- Before removing observations with missing values, select the variables used in the analysis using the `subset()` function.

```
> data  <- read.csv("data.csv")
> data1 <- subset(data, select = c(ID, Y, X1, X3))
> data1
  ID    Y   X1   X3
1  1 0.09 0.77 0.06
2  2 0.74 0.23   NA
3  3   NA   NA 0.94
4  4 0.50 0.42 0.25
5  5 0.50 0.89 0.17
6  6 0.78 0.44 0.41
> data2 <- na.omit(data1)
> data2
  ID    Y   X1   X3
1  1 0.09 0.77 0.06
4  4 0.50 0.42 0.25
5  5 0.50 0.89 0.17
6  6 0.78 0.44 0.41
```