ICA4

Jonathan Thompson

03/11/2024

Question 1

You are given data on people watching streaming tv. You know the age of the person, their income, and if they received a coupon for the service. You can also know the number of streaming hours for each person during a given week. Use the dataset "streaming_data.RData" for this question.

a) Find the descriptive statistics of the dataset.(i.e use data summary)

```
load("C:\\Users\\catho\\OneDrive\\Documents\\AAT\\streaming_data.RData")

library(modelsummary)

library(AER)

stream <- data.frame(age, coupon, income, streaminghours)
View (stream)
names(stream)
```

```
## [1] "age"            "coupon"         "income"         "streaminghours"
```

```
names(stream) <-c("Age", "Coupon", "Income", "Streaming Hours")

stream$`Income per 10K` <-stream$Income/10000

library(modelsummary)
datasummary_skim(stream)
```

b) Run OLS on the dataset.Estimate a model of streaming hours on the explanatory variables.

	Unique	Missing Pct.	Mean	SD	Min	Median	Max	
Age	1567	0	44.0	15.0	−3.7	43.6	92.6	
Coupon	2	0	0.3	0.5	0.0	0.0	1.0	
Income	1567	0	56 101.7	10 210.6	18 590.1	56 074.2	94 364.1	
Streaming Hours	585	0	3.2	5.8	0.0	0.0	42.6	
Income per 10K	1567	0	5.6	1.0	1.9	5.6	9.4	

	OLS	Tobit
(Intercept)	6.504***	4.491*
	(0.896)	(2.203)
Age	0.028**	0.080***
	(0.010)	(0.024)
Coupon	0.159	0.454
	(0.311)	(0.769)
Income per 10K	−0.828***	−2.161***
	(0.141)	(0.355)
Num.Obs.	1567	1567
R2	0.026	
R2 Adj.	0.024	
AIC	9906.6	5654.8
BIC	9933.4	5681.6
Log.Lik.	−4948.287	
F	13.959	
RMSE	5.69	9.24

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
reg1 <- lm(`Streaming Hours` ~ Age + Coupon + `Income per 10K`, data = stream)
```

c) Notice that the number of streaming hours cannot be negative. Run a Tobit model to estimate streaming hours.

```
library(AER)
reg2 <-tobit(`Streaming Hours` ~ Age +Coupon +`Income per 10K`, data=stream)

modelsummary(list("OLS"=reg1, "Tobit"= reg2), stars=TRUE)
```

d) What are the differences between c and a? Answer: Question A presents basic descriptive statistics such as mean and standard deviation while Question C presents two specific models in comparison to each other, focusing mainly on Tobit. Some of these measures for OLS & Tobit comparisons include R2, Log Likelihood and P Values. Tobit models also estimate the impact of independent variables while summary measures from Question A do not. Additionally, the Tobit Model focuses on measures of the linear effect of the latent variable that isn't censored while Question A only outputs summary measures. The difference in outputs lie in the depth of analysis and information provided. In conclusion, there can be more statistical significance that can be explained for the output in Question C vs. in the output in Question A.

e) What is the marginal effect of coupons?

```
library(censReg)
```

```
## Loading required package: maxLik
```

```
## Loading required package: miscTools
```

```
##
```

```
## Please cite the 'maxLik' package as:
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. C
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum o
## https://r-forge.r-project.org/projects/maxlik/

##
## Please cite the 'censReg' package as:
## Henningsen, Arne (2017). censReg: Censored Regression (Tobit) Models. R package version 0.5. http://
##
## If you have questions, suggestions, or comments regarding the 'censReg' package, please use a forum
## https://r-forge.r-project.org/projects/sampleselection/
```

```
estResult <- censReg(`Streaming Hours` ~ Age + Coupon + `Income per 10K`, data=stream)
knitr::kable(margEff(estResult))
```

	x
Age	0.02981923
Coupon	0.16826383
`Income per 10K`	-0.80122102

Question 2

—Selection Model–

In this analysis, you will use customer level data on travel expenditures. You will observe the following variables. Use the dataset "tour_data.RData" for this question.

Variables	Description
income	Household income
education	Education level of the household head
health	Health status index of the household members
tripweather	Weather quality in the destination of the trip
participation	Dummy variable for tourism participation
expenditure	Total household tourism expenditure

a) Estimate a regression model of expenditures on income, education, and tripweather.

```
load("C:\\Users\\catho\\OneDrive\\Documents\\AAT\\tour_data.RData")

View (tourexp)

names(tourexp)
```

```
## [1] "income"        "education"     "health"        "tripweather"
## [5] "participation" "expenditure"
```

```
names(tourexp) <-c("Income", "Education", "Health", "Trip Weather", "Participation", "Expenditure")

reg3 <-lm(Expenditure ~ Income + Education +`Trip Weather`, data=tourexp)
```

	Seletion Probit	Selection Linear
Intercept	−16.545***	−22 092.620***
	(0.961)	(916.811)
XSIncome	0.051***	
	(0.004)	
XSEducation	0.524***	
	(0.037)	
XSHealth	0.521***	
	(0.037)	
XOIncome		80.666***
		(3.702)
XOEducation		873.198***
		(40.301)
XOTrip Weather		451.601***
		(32.081)
imrData$IMR1		796.628***
		(182.830)
Num.Obs.		500
R2		0.851
R2 Adj.		0.849
AIC	672.4	8707.4
BIC	692.0	8732.7
Log.Lik.		−4347.693
RMSE	0.33	1445.84

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

b) Estimate a probit model on participation using income, education, and health

```
reg4 <-glm(Participation ~ Income+Education+Health, data=tourexp, family=binomial(link ="probit"))
```

c) Estimate a sample selection model of tourism expenditures.

```
library(sampleSelection)
```

```
## Warning: package 'sampleSelection' was built under R version 4.3.3
```

```
reg5 <-heckit(Participation ~ Income + Education + Health, Expenditure ~ Income+Education +`Trip Weathe
```

```
modelsummary(list("Seletion Probit"=reg5$probit, "Selection Linear"=reg5$lm),stars=TRUE, coef_rename = 
```

```
modelsummary(list("Expenditure Probit"=reg3, "Participation Probit"=reg4, "Selection Probit"=reg5$probi
```

d) Compare your estimates in a) and c). Do you detect sample selection bias? How do you know? Answer: There is evidence of sample selection bias and its positive. The p value is also significant and is very far from 0. Additionally, the Mills Ratio value at 796.6281 is very far from 0 which indicates positive bias.

	Expenditure Probit	Participation Probit	Selection Probit	Selection Linear
(Intercept)	−19 731.179***	−16.545***		
	(752.760)	(0.967)		
Income	71.948***	0.051***		
	(3.171)	(0.003)		
Education	788.164***	0.524***		
	(35.894)	(0.038)		
Trip Weather	449.997***			
	(32.656)			
Health		0.521***		
		(0.037)		
XS(Intercept)			−16.545***	
			(0.961)	
XS (Income)			0.051***	
			(0.004)	
XS (Education)			0.524***	
			(0.037)	
XS(Health)			0.521***	
			(0.037)	
XO(Intercept)				−22 092.620***
				(916.811)
XO(Income)				80.666***
				(3.702)
XO (Education)				873.198***
				(40.301)
XO (Trip Weather)				451.601***
				(32.081)
IMR				796.628***
				(182.830)
Num.Obs.	500	1000		500
R2	0.651			0.851
R2 Adj.	0.649			0.849
AIC	8724.2	672.4	672.4	8707.4
BIC	8745.3	692.0	692.0	8732.7
Log.Lik.	−4357.103	−332.179		−4347.693
F	308.426	98.925		
RMSE	1473.31	0.33	0.33	1445.84

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001