Universidad de Talca, FEN

Discrete Choice Models: Problem Set 2

Professor: Mauricio Sarrias

2020

1 Applications: Estimation of ML Estimator in R

1. Assume that you have a sample of n independent observations from a Poisson distribution with density function.:

$$f(y;\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} \quad y = 0, 1, 2, ...$$

where $\mathbb{E}(y) = \text{Var}(y) = \lambda$.

In this section, you have to create your own likelihood function (including the gradient and hessian) to estimate λ using `maxLik` package. Please, include your code and show that it works using a simulated dataset (Hint: See the solutions in Homework 1 to program the likelihood, gradient and hessian. For the simulated data use `rpois` using $n = 1000$ and set the seed using `set.sed(1)`).

2. In this section, we use a cleaned version of CASEN 2013 survey and run a wage (Mincer) equation. The first thing to do is to load the dataset. Since it is in STATA format, we need the **foreign** library to load it.

To load the data, we use the following codes:

```
# Load packages and data
library("foreign") # Package to load Stata dataset
setwd("~/Dropbox/Mis Clases/Discrete Choice Models/Homeworks/2020/HW2")
casen2009 <- read.dta("clean_casen2009.dta",
                      convert.factors = FALSE,
```

```
                        convert.underscore = FALSE)
```

The function `read.dta` allow us to read a .dta (Stata) file. If we want to know the name of the variables in the dataset we can type:

```
# Names of variables
names(casen2009)

##  [1] "region" "age"    "sch"    "wage"   "hours"  "wageh"  "lwageh" "male"
##  [9] "tenure" "exper"
```

We can get information about the number of individuals and variables using the `dim` function:

```
# Number of observations and variables
dim(casen2009)

## [1] 246925     10
```

The first element indicates the number of rows (individuals) in the dataset, while the second element indicates the number of columns (variables). Thus, our sample has 246,925 individuals and 10 variables.

Now we can make a summary of the variables using the command `summary`:

```
# summary statistics of variables
summary(casen2009)

##      region            age              sch             wage
##  Min.   : 1.000   Min.   :  0.00   Min.   : 0.00   Min.   :    1105
##  1st Qu.: 6.000   1st Qu.: 16.00   1st Qu.: 6.00   1st Qu.:  165750
##  Median : 8.000   Median : 33.00   Median :10.00   Median :  198900
##  Mean   : 8.396   Mean   : 35.29   Mean   : 9.08   Mean   :  304562
##  3rd Qu.:12.000   3rd Qu.: 52.00   3rd Qu.:12.00   3rd Qu.:  331500
##  Max.   :15.000   Max.   :107.00   Max.   :20.00   Max.   :14378000
##  NA's   :1        NA's   :1        NA's   :53162   NA's   :160916
##      hours            wageh               lwageh            male
##  Min.   :   4.0   Min.   :     1.03   Min.   : 0.03    Min.   :0.0000
##  1st Qu.: 160.0   1st Qu.:   920.83   1st Qu.: 6.83    1st Qu.:0.0000
##  Median : 180.0   Median :  1151.04   Median : 7.05    Median :0.0000
##  Mean   : 266.2   Mean   :  2006.26   Mean   : 7.16    Mean   :0.4904
##  3rd Qu.: 192.0   3rd Qu.:  1925.62   3rd Qu.: 7.56    3rd Qu.:1.0000
##  Max.   :3996.0   Max.   :302965.00   Max.   :12.62    Max.   :1.0000
```

```
##  NA's   :156315   NA's   :160916      NA's   :160916   NA's   :1
##      tenure           exper
##  Min.   : 0.00    Min.   : -7.00
##  1st Qu.: 0.00    1st Qu.:  8.00
##  Median : 3.00    Median : 26.00
##  Mean   : 7.21    Mean   : 27.87
##  3rd Qu.:10.00    3rd Qu.: 43.00
##  Max.   :75.00    Max.   :101.00
##  NA's   :158234   NA's   :53162
```

The sample has individual information about wage, hourly wage, gender, age, schooling, and region of residence, tenure and potential experience. For example, the summary statistics show that the average age of the individuals is 35 years old, while the average hourly wage is $2006 Chilean pesos. Furthermore, almost half of the sample is composed of men. It is also important noting that there exist missing values—indicated by `NA's`— in all the variables, and more striking is that we have individuals with negative potential experience (why?).

To clean for missing values, we can use the following command:

```
# Removing missing observations
clean_casen <- casen2009[complete.cases(casen2009), ]
dim(clean_casen)

## [1] 84251    10
```

Note that once we have clean for missing values the new sample `clean_casen` comprises 10 variables for 84,251 individuals.

In what follows, we will estimate different versions of the following regression model:

$$\log(\texttt{wageh}_i) = \beta_1 + \beta_2 \texttt{sch}_i + \boldsymbol{x}_i'\boldsymbol{\delta} + \epsilon_i, \tag{1.1}$$

where $\texttt{wageh}_i$ is the hourly wage per month for individual i.

a) Estimate model 1.1 using MLE (not OLS). To do so, please use **maxLik** package and `linear.mlc` function from Lecture 3. Additional to schooling , `sch`, use as controls, $\boldsymbol{x}_i = (\texttt{male}, \texttt{exper}, \texttt{tenure})$. Use Newton-Raphson algorithm for the optimization (`method = 'nr'`). Comment the results focusing on the return to schooling and include your code.

b) Estimate the model using OLS (`lm()` function), and convince yourself that the ML estimates are very close to OLS estimates. Obtain $\widehat{\sigma}^2_{OLS}$ and compare it with $\widehat{\sigma}^2_{ML}$. Are the similar? Why? Please, include your codes.

c) Use the LR statistic to test the null hypothesis that $H_0 : \beta_{\texttt{exper}} = \beta_{\texttt{tenure}} = 0$. Use your own code.

d) Use the Wald statistic to test the null hypothesis that $H_0 : \beta_{\texttt{exper}} = \beta_{\texttt{tenure}} = 0$. Use your own code.