ECON 322: Econometric Analysis 1

Assignment 5: Topics 7, 8 and 9

For this assignment, we use a survey on cannabis consumption in Canada. The survey was conducted in 2017 by Statistics Canada. The dataset is saved into the file A5datai.rda, where i is the number assigned to you in Quiz 6. The dataset A5data included in the file contains 17 variables. The data was obtained through the University Library website. You click on the "file & resources" tab and choose "Statistics & numerical data". You then click on "ODESI Data Retrieval". ODESI stands for "Ontario Data Documentation, Extraction Service and Infrastructure". If you like data analysis, it is a great source. To retreive the raw data, you expand "Health", "Canada", "Canadian Tobacco, Alcohol and Drugs Survey", "2017" and "Dataset: Canadian Tobacco, Alcohol and Drugs Survey, 2017: Person file", which is the survey used in this assignment. The data are in "Metadata". Once selected, you can click on the download button on the top right of the right window. For R, you select the CSV format. It will come with a PDF in which all variables are described. You can also get the description on the website by clicking on "Variable Description" below "Metadata". Any survey data need a little cleaning before you can use them in a regression. In this course, we don't have time to cover it, but if you are interested, the document cleanData.R uploaded to Learn shows how to create the whole dataset. Your file is a subset of this dataset. For example, he variable CAN_010 is the answer to: "During your lifetime, have you ever used or tried marijuana?". The possible answers are: 1- Yes, 2- No, 6- Valid Skip, 7- Don't know, 8- Refusal, and 9- Not stated. I used that variable to create the variable "EverUsed", which is 1 if the answer was 1 and 0 if the answer was 2. Individuals who provided other answers were removed from the sample.

The variables are:

- "EverUsed": A dummy variable equals to 1 if the individual has tried marijuana at least once in his life, and 0 otherwise.
- "moreThanOnce": A dummy variable equals to 1 if the individual has tried marijuana more than once, and 0 if he has tried only once. Since this question is conditional on having tried at least once, a missing value appears when the individual has never tried.
- "past12Months": A dummy variable equals to 1 if the individual has used marijuana at least once in the past 12 months, and 0 otherwise. Here, we put a 0 for those who have never tried.
- "ageFirst": How old was the individual when he started using marijuana. The observation is a missing value if the individual has never tried.
- "famsize": The family size of the individual's household.
- "size_0_14": The number of family members who are between 0 and 14 years old.
- "size_15_24": The number of family members who are between 15 and 24 years old.
- "size_25_44": The number of family members who are between 25 and 44 years old.
- "size_45plus": The number of family members who are older than 44 years old.

- "urban": A dummy variable equals to 1 if the individual lives in an urban area, and 0 if he lives in a rural area.
- "province": The province in which the individual lives (in character format).
- "age": The individual age in years.
- "male": 1 for males and 0 for females.
- "married": 1 for married and 0 is unmarried.
- "degree": A numeric value that indicates the individual's level of education: 0 for no high school degree, 1 for high school degree, 1.5 for something between high school and college, 2 for a college degree, 3 for a bachelor's degree, and 4 for any graduate degree. If you prefer, you can create a dummy variable for each degree.
- "drinkingAge18": A dummy variable equals to 1 if the individual lives in a province in which the minimum drinking age is 18 years old, and 0 if it is 19.
- "weights": Survey weights.

Notice that these are not iid observations. For example, some minority groups are over sampled to make sure ther appear in the survey. Survey weights are there to adjust for the sampling method used. For example, in my dataset, we have the following entries:

```
head(A5data$weight)

## [1]   210  1447 28465   312   214   898

sum(A5data$weight)

## [1] 29108806
```

The weights mean that the first entry counts for 210 individuals and fourth for 312 individuals. We would replicate the population by repeating the first individual 210 times, the second individuals 1447 times and so on. That would lead to a sample of 29.108806 million individuals. Of course, we never do it, because it would lead to too large datasets. I will show in one of the lectures on topic 8 how to use weighted least squares to run a regression using the weights without repeating the entries. I will guide you on a small exercise at the end of the assignment. For the moment, ignore that issue but keep in mind that without the weights, the sample may not be representative of the population.

Part I

The objective of the first part of the project is to compare cannabis consumption between different groups of the population. We want to use "EverUsed" as dependent variable, so you will be running linear probability models (LPM).

Answer the following questions by running the appropriate regression (only include the regressors that are needed to answer the questions) and performing the appropriate hypothesis tests. For all tests, do not use the F nor the t distribution, because exact tests do not exist in practice. Only use the asymptotic distributions.

1. Compare the probabilities of marijuana consumption for males and females. Are they significantly different? First, perform a BP test. Then, use a robust test if you reject the homoscedasticity assumption and a non-robust if you don't.

2. Compare the probabilities of marijuana consumption for married and non-married individuals. Are they significantly different? First, perform a BP test. Then, use a robust test if you reject the homoscedasticity assumption and a non-robust if you don't.

3. Do you see an impact on marijuana consumption to have a lower minimum drinking age? First, perform a BP test. Then, use a robust test if you reject the homoscedasticity assumption and a non-robust if you don't.

4. Is the effect of being married different for males and females? First, perform a BP test. Then, use a robust test if you reject the homoscedasticity assumption and a non-robust if you don't.

5. Test the joint hypothesis that the probability of having tried marijuana in all provinces are the same. Perform a short-White test and use a robust test if needed.

6. Using the model from the previous question, test the hypothesis that the probability of having tried marijuana in Ontario and British Columbia is the same against the alternative that it is not the same. Use the short-White test from the previous question to choose between robust and non-robust test.

   Hint: the "province" variable is a factor with 10 levels (one for each province). We have learned that you can add the variable in **lm**, and R will add 9 dummy variables, and exclude one to avoid perfect multicollinearity. The omitted dummy, which we call the base, is the first in the **levels** list (the order may be different in your dataset):

```
levels(A5data$province)

##  [1] "Newfoundland and Labrador" "Prince Edward Island"
##  [3] "Nova Scotia"               "New Brunswick"
##  [5] "Quebec"                    "Ontario"
##  [7] "Manitoba"                  "Saskatchewan"
##  [9] "Alberta"                   "British Columbia"
```

   If you want to change the base group to facilitate the test that you want to perform, you can change it using **relevel**. Suppose, for example, that I want Manitoba to be the first level, then I would redefine the province variable as follows:

```
A5data$province <- relevel(A5data$province, "Manitoba")
levels(A5data$province)[1]

## [1] "Manitoba"
```

   You can also replace "Manitoba" by the integer 7, because Manitoba was the seventh in the **levels** list before changing it.

7. Do you see any difference between individuals with different level of education? Perform a short-White test and use a robust test if needed.

   Hint: for that question, you may want to transform "degree" into factor. It is now a numerical variable, which is not appropriate to compare the different groups. You can create a new variable in A5data as follows (or replace the existing variable degree if you prefer):

```
A5data$degreeF <- as.factor(A5data$degree)
```

Do not only compare groups with the base group (degree=0). Compare also other groups (the choice is yours)

Part II

In this part, we want to choose a model. The dependent variable is EverUsed and the regressors are age, male, married, famsize, married and urban.

1. The first step is to choose the functional form for age. Consider the following models:

$$EverUsed = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 age^3 + \beta_4 male + \beta_5 married + \beta_6 famsize + \beta_7 urban + u$$

$$EverUsed = \beta_0 + \beta_1 log(age) + \beta_2 male + \beta_3 married + \beta_4 famsize + \beta_5 urban + u$$

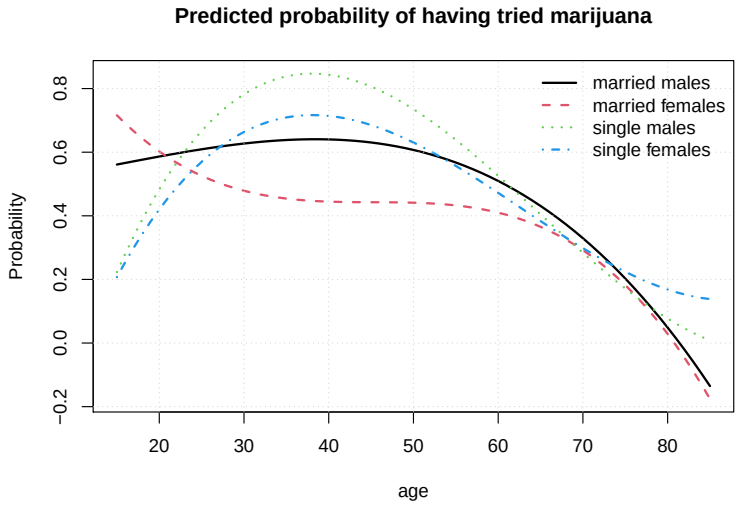
   Choose the best model using the J-test at 5%. If both models are rejected or none of them is rejected, choose the one that is the least rejected. Use the selected model for the following questions. Make sure you use the robust covariance matrix if the errors are heteroscedastic.

2. Using the selected model, test the null hypothesis that the model is correctly specified at 5% using the RESET test. Interpret the meaning of your result. Make sure you use the robust covariance matrix if the errors are heteroscedastic.

3. Interact male and married with each other and with the function of age (age, age^2 and age^3 or log(age) depending of the selected model). Then, perform a RESET test on the model. Is your model less rejected than the one without interactions? Make sure you use the robust covariance matrix if the errors are heteroscedastic.

4. Using the model with interactions from the previous question, compare the predicted probability of having used marijuana between married males, married females, non-married males and non-married females, by plotting the average predictions with respect to age for the four groups on the same chart. You can set urban and famsize to their sample average. Interpret your results.

   This is an example of the type of chart I am expecting:



Part III

As mentioned above, the observations from the dataset are not iid. In order to get proper estimates of the population coefficients, you need to take into account that some observations are more important than others in terms of representativity. One way to do it is to repeat each observation a number of times equals to its associated survey weight. Consider the following simple example. Suppose we have a sample $x = \{1, 10, 4, 8, 12\}$ with survey weights $w = \{100, 12, 230, 500, 56\}$. If we want to estimate the population mean, we can use $\bar{x}$:

```
x <- c(1,10,4,8,12)
mean(x)

## [1] 7
```

That's not a good estimates, because it does not take into account that some observations are more important than others. We can think of the above x as being a cheap way to store a large vector with many equal values. A proper estimate can be obtained by repeating the first observation 100 times, the second 12 times and so on.

```
x2 <- c(rep(1,100), rep(10, 12), rep(4, 230), rep(8, 500), rep(12, 56))
length(x2)

## [1] 898
```

Then the sample mean of the new vector is a better estimate of the population mean:

```
mean(x2)

## [1] 6.47216
```

This approach is not very efficient, because it implies creating very big vectors. In our dataset, for example, that would imply creating a new dataset with 29 million of rows. If you tried to do it, you would probably freeze your computer. Instead, we compute a weighted sum and divide by the sum of the weights. See by yourself that it is indeed identical:

```
w <- c(100,12,230,500,56)
sum(w*x)/sum(w)

## [1] 6.47216
```

The same can be done with OLS. Instead of repeating observations and minimizing the squared residuals, we minimize the weighted sum of the squared residuals:

$$\min_{\beta} \sum_{i=1}^{n} w_i u_i^2$$

By doing it, we are giving more importance to observations with larger weights. This method is called weighted least squares (WLS). It is like the one we saw in class for solving the problem of heteroscedasticity, but the purpose is different. In the case of heteroscedasticity, it is not recommended to use WLS because it is impossible to know for sure what is the true process for the conditional variance. Also we have to rely on feasible-WLS which produces biased estimators. In the case of survey with weights, we have no choice if we want a good estimate of the population coefficient. In R, all you have to do is to set the argument "weight" of `lm` to the weight variable. See the difference:

```r
coef(fit1 <- lm(EverUsed~male, A5data, weight=weight))

##  (Intercept)         male
##  0.494687052 -0.009052084

coef(fit2 <- lm(EverUsed~male, A5data))

## (Intercept)        male
##  0.42478814  0.05838494
```

The WLS estimate suggests that males are 0.91% less likely to have used marijuana at least once, while it is 5.84% more likely if we use OLS. We should trust the WLS results more than OLS.

Consider the following model:

$$EverUsed = \beta_0 + \beta_1 male + \beta_2 married + \beta_3 (male \times married) + u$$

Estimate the model by OLS and WLS and compare your results (interpret the coefficients in both regressions). If you reject homoscedasticity, use robust standard errors to evaluate the significance of the coefficients.