ECONOMETRICS PROJECT

TSE (2021 – 2022)

Problem 1

1) After initializing a library called “project” with the instruction LIBNAME, we find the total number of observations in the dataset, which is equivalent to the number of rows in the dataset :

Number of rows/observations in dataset
2555

There are 2555 observations in the dataset.

2) **For qualitative variables / dummies**, we find the following simple statistics by using PROC FREQ :

La procédure FREQ

1 if woman is in neighborhood that received treatment (one or two)				
targeted_neighborhood	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	424	16.59	424	16.59
1	2131	83.41	2555	100.00

1 if woman received treatment 1				
treated1	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	1535	60.08	1535	60.08
1	1020	39.92	2555	100.00

1 if woman received treatment 2				
treated2	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	1771	69.32	1771	69.32
1	784	30.68	2555	100.00

1 if the woman has any schooling				
any_school	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	2088	81.72	2088	81.72
1	467	18.28	2555	100.00

1 if the woman is married				
d_married	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	499	19.53	499	19.53
1	2056	80.47	2555	100.00

Do you have access to a TV				
tv_access	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	781	30.57	781	30.57
1	1774	69.43	2555	100.00

Do you have access to a TV with cable server or a dish				
cable_access	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	1785	69.86	1785	69.86
1	770	30.14	2555	100.00

Does your HH follow the advice of a pir/murshid in important decisions of HH				
advice_pir	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	879	34.40	879	34.40
1	1676	65.60	2555	100.00

1 if women self-reports that she voted & had a verifiable ink mark on her thumb				
voted	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	941	36.83	941	36.83
1	1614	63.17	2555	100.00

Let’s explain briefly some numbers but not all. For example :

- For the **variable “d_married”** (which is =1 if the woman is married), we can see that “d_married”=1 for 2056 observations. This means that 2056 (80.47%) women among the 2555 are married, and that 499 (19.53%) women are not married.

- For the **variable “treated1”** (which is =1 if the woman was treated by treatment 1), we can see that “treated1”=1 for 1020 observations. This means that 1020 (39.92%) women among the 2555 have been treated by treatment 1, and that 1535 (60.08%) women haven’t been treated by treatment 1.

For the quantitative variables, we find the following simple statistics by using PROC MEANS which allows us to get a summary of continuously distributed data (it reports the number of observations, the mean, the standard variation, and the min and max values) :

La procédure MEANS

Variable	Libellé	N	Moyenne	Ec-type	Minimum	Maximum
age	Age in Completed years	2555	37.9956947	15.9068962	1.0000000	99.0000000
land	total land owned	2555	3.0614160	8.2141950	0	110.0250015
hhsize	Total number of household members	2555	11.8708415	5.8193699	2.0000000	46.0000000
num_women	Total number of women in the household	2555	3.5142857	1.6465512	1.0000000	11.0000000
house_quality	House quality index from pca	2555	0.1810044	1.4192232	-1.8611678	4.0527663

Let’s explain briefly some numbers but not all. For example :

- For the variable “age”, we can see that among the 2555 women observed, they have an average age of 37.99. The youngest girl is 1 year old, and the oldest woman is 99 year old.

- For the variable “hhsize”, we can see that for the 2555 women observed, they have, on average, 11.87 members in their household. The smallest household has 2 people in it, and the largest household has 46 members.

3) We are looking for the percentage of women 30 years old or younger who were taught both about the importance of voting and the balloting process, i.e. for **the percentage of women who are both 30 years old or younger AND who are treated 2**. As a consequence of that, we have decided to create a frequency table with 2 variables : “age” (of the women) and “treated2” (=1 if the woman was treated by treatment 2). We have the following table :

La procédure FREQ

Fréquence
Pourcentage

treated2(1 if woman received treatment 2)	Age																			
	1	2	10	12	13	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0	1 0.04	1 0.04	1 0.04	1 0.04	0 0.00	2 0.08	4 0.16	70 2.74	30 1.17	103 4.03	12 0.47	59 2.31	27 1.06	17 0.67	166 6.50	22 0.86	19 0.74	84 3.29	6 0.23	166 6.50
1	0 0.00	0 0.00	0 0.00	0 0.00	1 0.04	0 0.00	1 0.04	33 1.29	12 0.47	42 1.64	5 0.20	31 1.21	10 0.39	7 0.27	76 2.97	8 0.31	11 0.43	28 1.10	2 0.08	96 3.76
Total	1 0.04	1 0.04	1 0.04	1 0.04	1 0.04	2 0.08	5 0.20	103 4.03	42 1.64	145 5.68	17 0.67	90 3.52	37 1.45	24 0.94	242 9.47	30 1.17	30 1.17	112 4.38	8 0.31	262 10.25

Thus, if we add up the numbers in yellow (which represent the percentage of women who are treated 2, for EACH possible age), we obtain : 0.04 + 0.04 + 1.29 + 0.47 + 1.64 + 0.2 + 1.21 + 0.39 + 0.27 + 2.97 + 0.31 + 0.43 + 1.1+ 0.08 + 3.76 = 14.2. **To conclude, we see that 14.2% of the women observed are both 30 years old or younger AND treated2.**

However : The question is unclear. It may be asking the percentage of women that are treated2, AMONG THOSE who are 30 years old or younger. In this case, the answer is different. On the previous table, we can see that we have :

- 1+1+1+1+1+2+5+103+42+145+17+90+37+24+242+30+30+112+8+262 = **1154 women 30 years old or younger**.

- 1+1+33+12+42+5+31+10+7+76+8+11+28+2+96 = **363 women who received treatment 2 AMONG THOSE who are 30 years old or younger.**

- Thus, (363/1154)*100 = **31,45 % of women received treatment 2 AMONG THOSE who are 30 years old or younger.**

4) We are looking for the percentage of neighborhoods that were targeted, for each village. By using the instruction PROC FREQ, we obtain :

La procédure FREQ

Village code=1

1 if woman is in neighborhood that received treatment (one or two)

targeted_neighborhood	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	49	21.78	49	21.78
1	176	78.22	225	100.00

Village code=3

1 if woman is in neighborhood that received treatment (one or two)

targeted_neighborhood	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	51	19.92	51	19.92
1	205	80.08	256	100.00

Village code=5

1 if woman is in neighborhood that received treatment (one or two)

targeted_neighborhood	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	64	19.88	64	19.88
1	258	80.12	322	100.00

La procédure FREQ

Village code=2

1 if woman is in neighborhood that received treatment (one or two)

targeted_neighborhood	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	50	16.72	50	16.72
1	249	83.28	299	100.00

La procédure FREQ

Village code=4

1 if woman is in neighborhood that received treatment (one or two)

targeted_neighborhood	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	38	14.39	38	14.39
1	226	85.61	264	100.00

La procédure FREQ

Village code=6

1 if woman is in neighborhood that received treatment (one or two)

targeted_neighborhood	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	44	16.60	44	16.60
1	221	83.40	265	100.00

Village code=9

1 if woman is in neighborhood that received treatment (one or two)

targeted_neighborhood	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	41	11.95	41	11.95
1	302	88.05	343	100.00

La procédure FREQ

Village code=10

1 if woman is in neighborhood that received treatment (one or two)

targeted_neighborhood	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	39	12.70	39	12.70
1	268	87.30	307	100.00

Village code=11

1 if woman is in neighborhood that received treatment (one or two)

targeted_neighborhood	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	48	17.52	48	17.52
1	226	82.48	274	100.00

We can see that :
- for the village 1, 78.22% of the women were living in a neighborhoods with treated individuals.
- for the village 2, 83.28% of the women were living in a neighborhoods with treated individuals.
- for the village 3, 80.08% of the women were living in a neighborhoods with treated individuals.
- for the village 4, 85.61% of the women were living in a neighborhoods with treated individuals.
- for the village 5, 80.12% of the women were living in a neighborhoods with treated individuals.
- for the village 6, 83.40% of the women were living in a neighborhoods with treated individuals.
- for the village 9, 88.05% of the women were living in a neighborhoods with treated individuals.
- for the village 10, 87.30% of the women were living in a neighborhoods with treated individuals.
- for the village 11, 82.48% of the women were living in a neighborhoods with treated individuals.

Moreover, the question being quite unclear, we can provide a different answer. By using

```
proc freq data = project.gw2018;
 tables neighborhood_code;
 by village_code;
run;
```

, we can also easily see the number of neighborhoods in each village.

Finally, we can also see that :
- for the village 1, among its 6 neighborhoods, 1 is not targeted. So (5/6)*100 = 83.33% (of neighborhoods) is targeted.
- for the village 2, among its 8 neighborhoods, 1 is not targeted. So (7/8)*100 = 87.5% is targeted.
- for the village 3, among its 7 neighborhoods, 1 is not targeted. So (6/7)*100 = 85,71% is targeted.
- for the village 4, among its 12 neighborhoods, 2 are not targeted. So (10/12)*100 = 83.33% is targeted.
- for the village 5, among its 6 neighborhoods, 1 is not targeted. So (5/6)*100 = 83.33% is targeted.
- for the village 6, among its 7 neighborhoods, 1 is not targeted. So (6/7)*100 = 85,71% is targeted.
- for the village 9, among its 7 neighborhoods, 1 is not targeted. So (6/7)*100 = 85,71% is targeted.
- for the village 10, among its 7 neighborhoods, 1 is not targeted. So (6/7)*100 = 85,71% is targeted.
- for the village 11, among its 7 neighborhoods, 1 is not targeted. So (6/7)*100 = 85,71% is targeted.

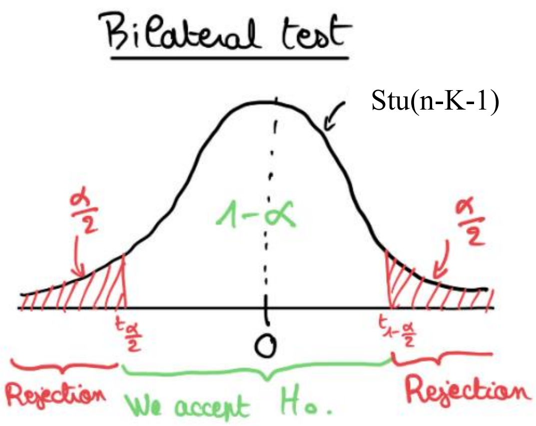
Problem 2

1) <u>I predict that β1 and β2 will be equal to (or close to) zero.</u> Indeed, there seems to be no connection between *treated1* or *treated2* AND *x* (which is a woman or household characteristic, such as if the woman has any schooling or not, or the size of owned land for example). This is due to the fact that neighborhoods are **<u>randomly</u>** assigned to be targeted or not, and targeted neighborhoods **<u>randomly</u>** receive either treatment 1 or treatment 2, and even in targeted neighborhoods some women may not receive the treatment but it is **<u>randomly</u>** decided. As the subject states it : “As the treatment was randomly assigned, there should be no systematic differences between the women in the different groups at this stage”. There is no link between the initial situation of the women (educated or not) and whether they receive treatment or not. Post-treatment, we can imagine that the treatments have an effect on voting, but not on women or household characteristics *x*.

2) a) We want to test that β_2 is statistically different from zero at α=10% significance level. A test is a decision rule concerning the null hypothesis H_0 against the alternative hypothesis H_1. Here we have :
$H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$ (bilateral test).

2) b) In this case, we use the test statistic (t-ratio) : t = $\frac{\widehat{\beta_2} - \beta_2}{\hat{\sigma}_{\widehat{\beta}_2}}$ ~ Stu(n-K-1) where n is the sample size and K is the number of explanatory variables. **<u>With the H_0 hypothesis</u>**, we have : t = $\frac{\widehat{\beta_2} - 0}{\hat{\sigma}_{\widehat{\beta}_2}}$ ~ Stu(n-K-1) where n-K-1 = 2555-2-1 = 2552 (it is the degrees of freedom of the Student's t-distribution). We also know that when n -> + ∞, the Student distribution tends toward a N(0,1) distribution. Then, let’s compute the value of the test statistic for this test. By estimating the model by OLS, we find : $\widehat{\beta_2} = 0.02582$ and $\hat{\sigma}_{\widehat{\beta}_2} = 0.01974$. Thus we have **<u>t = 1.308</u>**.

2) c) Decision rule : we reject H_0 if $|t| > t_{n-K-1\ ,\ 1-\frac{\alpha}{2}}$. Otherwise, wo do not reject H_0.



So we have : $t_{1-\frac{\alpha}{2}} = t_{1-\frac{0.1}{2}} = t_{1-0.05} = t_{0.95} = 1.645$.

2) d) We have 2 methods to conclude :

- **First, with the test statistic** : $|t| = 1.308 < t_{1-\frac{\alpha}{2}} = 1.645$. <u>So we do not reject H_0. So we can say that β_2 is statistically equal to 0</u>.

- **Secondly, with the p-value** : the p-value is the probability, under H_0, to observe a test statistic that is further away from H_0 than the one we actually observe. If p-value < α : we reject H_0 ; otherwise, we do not reject H_0. With SAS, we find, with the REG procedure, that the **p-value = 0.1909** > α = 0.1. <u>So we do not reject H_0. So we can say that β_2 is statistically equal to 0</u>.

Problem 3

1) In this regression, we want to explain variations in the variable *voted* that can be attributed to changes in 2 explanatory variables, *treated1* and *treated2*. This will enable us to quantify the strength of the relationship between these variables. More particularly, here, we want to test the impact, *ceteris paribus*, of receiving each treatment on the decision of voting (or not) for the woman observed. We want to see how much receiving the first or second treatment influences the decision of voting or not.

2) When running M2, we predict that both β_1 and β_2 will be positive as the campaigns/treatments are meant to reduce the gap in voting behavior between men and women (i.e. to make women vote more) by giving them information. Indeed, women with information about the voting process are more likely to go to vote in elections.

Moreover, we can give an additional piece of information : since treatment 1 gives information about the importance of voting and treatment 2 gives information about the importance of voting **<u>but also</u>** information regarding the balloting process, we can conclude that women who received the treatment 2 received more information about elections than the women who received the treatment 1. As a consequence of it, we can think that the women who received the treatment 2 will be more likely to vote in elections than the women who received the treatment 1 (*ceteris paribus* ; this condition is also verified because a woman cannot receive both treatment 1 and treatment 2). So we can think that β_2 will be higher than β_1.

3) By estimating the model by OLS, we observe that $\widehat{\beta_1} = 0.08753$ and $\widehat{\beta_2} = 0.13091$. Then we create the variable age2 = age². Then we estimate the new model, and we obtain the following estimated coefficients :

Paramètres estimés			
Variable	Libellé	DDL	Valeur estimée des paramètres
Intercept	Intercept	1	0.10846
treated1	1 if woman received treatment 1	1	0.09075
treated2	1 if woman received treatment 2	1	0.12758
age	Age in Completed years	1	0.01984
age2		1	-0.00016722
any_school	1 if the woman has any schooling	1	0.05817
advice_pir	Does your HH follow the advice of a pir/murshid in important decisions of HH	1	-0.05430
house_quality	House quality index from pca	1	0.01483

<u>Interpretation of the estimated coefficients **ceteris paribus** and on average :</u>
- A woman treated with treatment 1 increases her chances of voting by 9.075% (significant at 1%).
- A woman treated with treatment 2 increases her chances of voting by 12.758% (significant at 1%).
- A one-year increase in the woman's age increases her chances of voting by 1.984% (significant at 1%).
- The estimated coefficient associated with the variable age2 is slightly negative which means that there is a negative quadratic link between the age and the chances of voting. The older the women are, the less an age increase increases their chances to go to vote (significant at 1%).
- Receiving any schooling increases the woman's chances of voting by 5.817% (significant at 5%).
- A woman being a member of a household following the advice of a spiritual guide in important decisions decreases her chances of voting by 5.430% (significant at 1%).
- An augmentation of the house quality index from pca increases the woman's chances of voting.

Moreover, we find $\widetilde{\beta_1} = 0.09075$ and $\widetilde{\beta_2} = 0.12758$.
So $\widehat{\beta_1} < \widetilde{\beta_1}$ and $\widehat{\beta_2} > \widetilde{\beta_2}$. We observe that the estimated β_1 increased and the estimated β_2 decreased.

4) a) We know that *treated1*, *treated2*, and *treated* are dummies.
We know that :

$$\underset{\text{treated}}{\mathbb{1}_{A\cup B}} = \underset{\text{treated1}}{\mathbb{1}_{A}} + \underset{\text{treated2}}{\mathbb{1}_{B}} - \mathbb{1}_{A\cap B}$$

= 0 because treatments 1 and 2 are independent (a woman who receives treatment 1 cannot have treatment 2 at the same time ; a woman who receives treatment 2 cannot have treatment 1 at the same time).

So we can conclude that : **<u>treated = treated1 + treated2</u>**.

4) b) We want to evaluate the impact (on the decision to vote) of living in a targeted neighborhood and not being treated. Two different explanatory variables must be taken into account : *treated* and *targeted_neighborhoods*. We have the following new model :

$$voted_i = \beta_0 + \beta_1 treated_i + \beta_2 targeted\_neighborhood_i + u_i$$

Let's estimate this new model, in the case of a woman who didn't receive the treatment (i.e. treated = 0). We obtain :

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

treated =	0

Paramètres estimés

Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	Intercept	1	0.51415	0.02404	21.38	<.0001
treated		0	0	.	.	.
targeted_neighborhood	1 if woman is in neighborhood that received treatment (one or two)	1	0.09747	0.03644	2.67	0.0076

We can see here that $\widehat{\beta_2} = 0.09747$. **This means that a non-treated woman living in a targeted neighborhood has 9.747% more chances to vote than a non-treated woman in a non-targeted neighborhood (in average, ceteris paribus).**

(Moreover : If we estimate the model M3 : $voted_i = \beta_0 + \beta_1 treated_i + u_i$*, we find that* $\widehat{\beta_1}$ *= 0.10638. So a treated woman (who is in a targeted neighborhood by definition) has 10.63% more chances to vote that a non-treated woman. So :*

- a woman in a targeted neighborhood who is <u>non-treated</u> raises her chances to vote by 9.747% VS a woman non treated and in a non-targeted neighborhood.

- a woman in a targeted neighborhood who is <u>treated</u> raises her chances to vote by 10.63% VS a woman non treated ***but either in a targeted neighborhood OR NOT*** *=> <u>this means that we can't compare directly the two percentages</u>).*

5) a) We want a 95% confidence interval :

- To have this, we regress the model M3. We obtain the estimated coefficient $\widehat{\beta_1}$ of the variable *treated*, which is equal to 0.10638.
- We also find that $\hat{\sigma}_{\widehat{\beta}_1} = 0.02085$.
- Then the level of confidence being 95% (~ 0.95 ; i.e. 1-α=0.95 i.e. α=1-0.95=0.05), we need the quantile of order $1 - \frac{\alpha}{2} = 0.975$ read on the Student law table Stu(2555-1-1) = Stu (2553) : $q_{1-\frac{\alpha}{2}} = 1.960$.

Then we conclude that the coefficient β_1 is with a probability of 95% in the interval :
[0.10638 - 1.960*0.02085 ; 0.10638 + 1.960*0.02085] = [0.0655 ; 0.1472].
<u>This means that being treated increases the chances of voting of a woman by between 6.55% and 14.72% with 95% level confidence.</u>

We can also find it automatically on SAS :

Paramètres estimés

Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à95%	
Intercept	Intercept	1	0.55659	0.01752	31.77	<.0001	0.52224	0.59094
treated		1	0.10638	0.02085	5.10	<.0001	0.06550	0.14726

Both bounds are positive and quite high : the program seems efficient.

5) b) We can recommend this policy which seems to be good. With a 95% level confidence, the program will increase the chances of voting of a treated woman by between 6.6% and 14.7%. This is a good policy because it has a positive impact on the number of women voters, which will limit abstention. If more people vote, it is more democratic. In addition, it will increase the equality between men and women voters and empower women to act for their future. Finally, we can recommend this policy because it is not subject to the two main concerns stated by some members of the government : with a 95% level confidence, the policy will not lower the chances of voting for women, and it is not going to increase it too much (as it will increase the chances of voting of a woman by between 6.6% and 14.7% with 95% confidence).

Problem 4

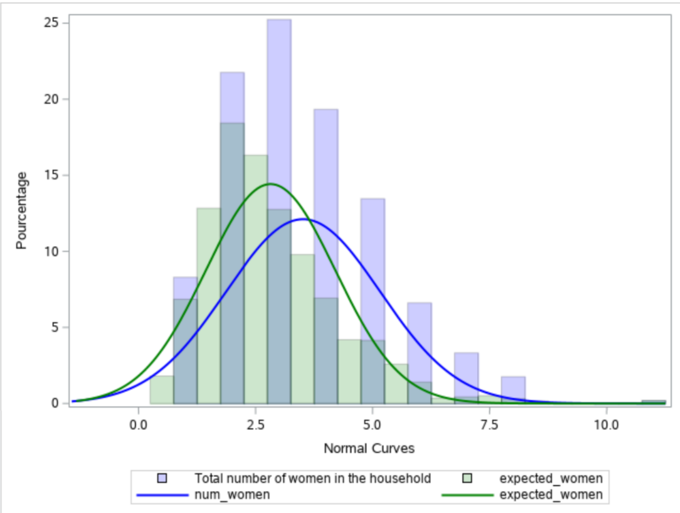
1) a) According to the World Bank's data on Pakistan in 2011, we know that :
- 48.5 % of the total population is female ;
- 49 % of this total female population is adult.

Thus, we can create the variable "expected_women", which is the number of adult females (woman = adult female) that each household should have, in the following way :

```
expected_women = (hhsize*0.485)*0.49
```

This new variable gives the THEORETICAL number of adult females (= women) in each household, whereas the variable "num_women" gives the EMPIRICAL number of women in each household.

1) b) We have to plot a single histogram with both the distribution of the expected number of women and the actual number of women. The actual number of women (in each household) is represented by the variable "num_women". The expected number of women (in each household) is represented by the variable "expected_women". Let's compare the distribution of "num_women" and "expected_women". We obtain:



On average, we can see that there are a little bit more women than expected.

2) We construct a new variable : more_women = num_women – expected_women. Then we provide the summary statistics for more_women (with PROC MEANS) :

La procédure MEANS

Variable d'analyse : more_women

N	Moyenne	Ec-type	Minimum	Maximum
2555	0.6931802	1.3714473	-5.9319000	7.4352500

The number of expected women in each household is quite different from that in the sample. The mean of the variable "more_women" is equal to 0.69 and it is positive. This means that **ON AVERAGE** there are more women than expected in each household (which is also the results of the histograms of the question 1) – b)). More precisely, **ON AVERAGE**, there are 0.69 women more than expected in each household.

3) First, let's construct the gender ratio variable : interact_treated_mw = treated * more_women.

Then, let's consider the new model M4 : $voted_i = \beta_0 + \beta_1 treated_i + \beta_2 more\_women + \beta_3 interact\_treated\_mw + u_i$

By estimating the model by OLS, we obtain :

Paramètres estimés				
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type
Intercept	Intercept	1	0.56664	0.01900
treated		1	0.10216	0.02294
more_women		1	-0.01817	0.01333
interact_treated_mw		1	0.01041	0.01561

Now let's test at the α = 5% significance level if both β_2 and β_3 are different from 0. A test is a decision rule concerning the null hypothesis H_0 against the alternative hypothesis H_1. Here we have :

$H_0 : \beta_2 = \beta_3 = 0$ vs. $H_1 : \beta_2 \neq 0$ or **$\beta_3 \neq 0$.**

In this case, we use the Fisher test statistic :

$$F = \frac{\frac{SSR_0 - SSR}{q}}{\frac{SSR}{(n-K-1)}} \sim F(q; n-K-1) \quad \textbf{(under } \mathbf{H_0}\textbf{)}$$

Where :
- q = number of tested restrictions = 2
- n = sample size = 2555
- K = number of explanatory variables = 3

We also have :

$$F = \frac{\frac{(1-R_0^2)\,SST - (1-R^2)\,SST}{q}}{\frac{(1-R^2)\,SST}{(n-K-1)}} = \frac{\frac{(R^2-R_0^2)}{q}}{\frac{(1-R^2)}{(n-K-1)}} \sim F(q; n-K-1)$$

We know (from the OLS estimation of the model M4) that **R^2 = 0.0112.**

Moreover, with restrictions (i.e. under H_0), we estimate the following model :

$voted_i = \beta_0 + \beta_1 treated_i + u_i$

And we find : **R_0^2 = 0.0101.**

Thus **F = 1.4189 ~ F(2;2551).**

Decision rule : we reject H_0 at the level α if F > $F_{q\ ,\ n-K-1} = F_{2,2551}$. Otherwise, wo do not reject H_0.

So we have : $F_{2,2551} = 3$.

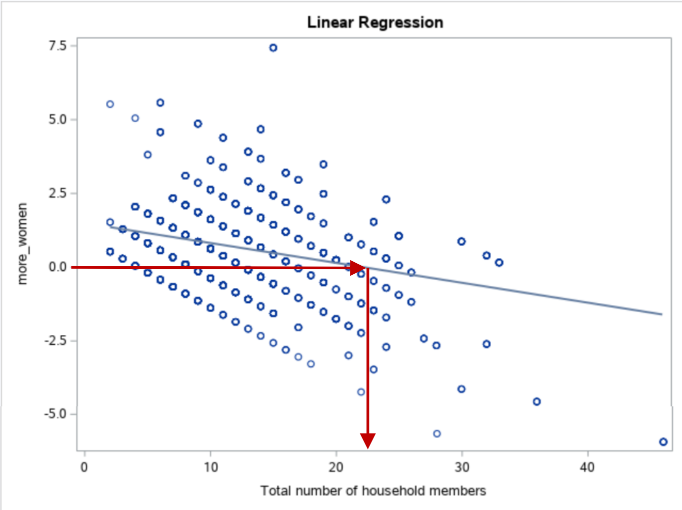
Thus, we have : $F < F_{2,2551}$. **We don't reject H_0 : we can say that β_2 and β_3 are statistically equal to 0.**

Finally, the interpretations :

- In this problem, we want to test if family characteristics and the gender ratio in Pakistan matter to interventions that try to change voting turnout. If $\beta_2 = \beta_3 = 0$, this means that *more_women* and *interact_treated_mw* are not significant variables, so we can remove them from the model. This means that in the end, the gender ratio has no effect on voting decision.

- If only $\beta_3 = 0$, then the variable *interact_treated_mw* is not a significant variable, which means that we can remove it from the model. However, the variable *more_women* can be significative. This means that the difference in within household gender ratio can matter to voting.

4) Let's plot a scatterplot for the relationship between *more_women* (y axis) and *hhsize* (x axis), with the linear regression plotted in the plot. We have :



We can see on the scatterplot that the surplus of women in a household reverses sign for a household size of 23. It can be interpreted the following way: when the number of people in the household exceeds 23, the difference between the number of women in the household and the number of women expected is negative. So, when the number of people in the household exceeds 23, there are less women in the household than expected knowing the size of the household.

5) When we observe the plot of the predicted values of the residuals, we see that they are not in the shape of a cloud around 0 which is what we usually expect from residuals. Instead, we observe two lines slightly tilted. However, we see that they are symmetrical relative to 0. To be able to estimate the model, we need to have the mean equal to zero which is the case here thanks to the symmetry of the two lines relative to 0.