**Stata software test – Results – Rohini Venkitaraman Iyer**

**Task 1: Data cleaning**

Some cleaning notes –

* Dealing with duplicates
  + There are 3 observations with missing HHID and 3 observations with duplicate HHID.
  + The 3 observations with duplicate HHID are identical across all variables.
  + All 6 of these observations have been dropped because the ones missing HHID cannot be matched to treatment status and for the other 3, we are retaining 1 set of their data.
* Question “w16\_unpaid\_job99”
  + This is a select all that apply kind of multiple-choice question.
  + Ideally, I would have broken this variable down into a series of 0/1 type dummy variable - one for each option.
  + And then created a separate "rowtotal" column that adds all the 1s for the different unpaid tasks that the women undertook over the past 7 days.
  + However, since that is time consuming, I am proceeding further with the remaining test and will revisit this at the end if I have time.
  + This is also something I would consult with my supervisor about

**Task 2A: Table**

* Missing values
  + There is 1 household that has missing values for both cash and jewellery savings. Hence, I modified total\_savings to show a missing value for this household rather than 0 because there are 16 other households that actually report 0 savings.
  + “Number of households’ column in the output table represents the NON-MISSING number of observations for each variable.

Table 1: Descriptive Statistics (in USD)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Number of Households | | Mean | Median | SD | Min | Max |
| Cash Savings | 840 | 52.4 | | 41.5 | 510.7 | -59.3 | 14814.8 |
| Jewellery Savings | 640 | 423.3 | | 18.5 | 10247.4 | -222.2 | 259259.3 |
| Total Savings | 1288 | 244.5 | | 29.5 | 7234.7 | -218.5 | 259265.9 |

**Task 2B: Graph**

* For this graph, job types 1, 2, and 3 are PAID jobs since BEGINNING OF 2012.
* While job type 4 is UNPAID work in the LAST 7 DAYS.
* I make this note and proceed due to shortage of time. Under normal circumstances, I would discuss this decision with my Supervisor.

A graph showing a number of hours worked by a patient

Description automatically generated

From this graph, women in the control group seem to engage in higher number of “paid” labor hours as indicated in the difference in bar heights particularly for jobs 2 and 3.

**The “job type” level dataset can be found in the “output” folder labelled “GEM\_Job\_Type.dta”**

**Task 3: Regression**

1. What output that would be, and what information would be included in it?
2. What specification you would use and why?
3. What variables you would consider and why? You can mention variables that are not included in the dataset but that you consider relevant.

🡪 The regression uses total savings in USD as the outcome variable and factors like

* age of the respondent
* whether they attended school
* received vocational training (before this intervention)
* their household characteristics like having piped water
* assets like sewing machine, electricity, television
* whether the house is rented
* how many household members (eating from the same pot)
* whether the women received cash/kind payment for any work they did
* the total no. of paid hours of labor they engaged in
* the total no. of unpaid hours of labor they engaged in
* their current marital status
* their spouses' education
* and their treatment status in the experiment

as independent variables to understand what affects household savings.

🡪 Based on the regression, the women's age, whether they went to school, whether their spouse went to school, and whether they own a sewing machine are factors that positively and significantly influence total savings. The results are significant because p < 0.05 (except for age when it is < 0.1). The regression is also clustered at the household level to account for within household correlation.

🡪 If I had more time, I would –

* Give a more detailed interpretation of the results (like using the magnitudes of the significant results to comment about the dependent variable)
* Perform regressions with each independent variable separately because the fact that factors like hours of paid labor, or household members, or hours of unpaid labor are not significant influencers of household savings is puzzling and seems counterintuitive.
* The sample size for regressions is only 133 which is only about 10% of the dataset. This is due to the missing values, but I would spend some time figuring how to deal with them and get my supervisor's advice on this.
* Format my table (like editing the independent variable names)
* Other than the variables available in this dataset, a few other indicators that I would be interested in observing as factors that influence household savings from the broader survey would be –
  + Migration indicators - like if they have ever lived outside Nairobi, especially in an urban setting.
  + where they accumulate their savings
  + whether they have a bank account
  + age at start of marriage
  + no. of children, etc.

Table 2: Regression Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (1) |  |  |  |
|  | Total Savings |  |  |  |
|  | Coefficient | Robust Standard Error | t | p value |
| 102. How old are you? | 3.995084 | 2.387766 | 1.673147 | .0966666 |
| 105. Did you attend school in Kenya? | 29.71466 | 11.02876 | 2.694288 | .0079708 |
| 120. a. Are you currently attending any vocational training? | -5.143318 | 6.270688 | -.8202159 | .4135717 |
| i. WATER PIPED INTO YOUR COMPOUND | 4.911471 | 4.62586 | 1.061742 | .2902914 |
| a. Electricity | -5.020215 | 5.595348 | -.8972122 | .3712392 |
| c. Television | -4.763317 | 4.615779 | -1.031964 | .303976 |
| l. Sewing machine | 13.55456 | 5.759938 | 2.353247 | .020087 |
| W02.a. Have you done any work for which you were paid in cash or in kind since | -.9643825 | 5.395394 | -.1787418 | .8584145 |
| W20. How many hours did you do unpaid work in the last seven days? | .0366127 | .1489963 | .2457291 | .8062736 |
| b. Are you currently married? | 2.200597 | 4.761287 | .4621854 | .6447093 |
| 912. a. Did your spouse attend school in Kenya? | .3102814 | .083896 | 3.698405 | .0003172 |
| c. How many years of formal education did your spouse complete? | .0358881 | .0470298 | .7630927 | .4467698 |
| Treatment Status | -2.368835 | 4.035014 | -.5870698 | .5581596 |
| Total Hours worked in paid jobs | .0300014 | .0498491 | .6018436 | .5483106 |
| Total HH members eating from one pot | .0574519 | 1.353953 | .0424327 | .9662178 |
| Constant | -79.78372 | 52.92016 | -1.507624 | .1340403 |
| Observations | 133 |  |  |  |

**Optional Task 4 – Randomization evaluation**

1. What output would you create to test whether the randomization is balanced? And what information would be included in it?

🡪 I would create balance tables that compare basic socio-demographics like

* Age
* Education
* marital status
* no. of children
* no. of household members
* household savings
* assets
* hours spent on paid v/s unpaid labor, etc.

for the treatment and the control groups to check if the randomization has resulted in 2 groups that are similar on observable and unobservable indicators prior to the beginning of the intervention. This would help isolate and attribute any differences between the groups post intervention to the treatment alone.

1. What statistical process would you follow?

🡪 I would start with some summary statistics like mean, SD, proportions for each characteristic broken down by treatment status. Then I would create balance tables using 2 sample t-tests to compare sample means of the two groups on the factors listed in the previous point.

Code - ttest var, by(treatment)

The expectation is that the t-test results for each variable would NOT BE SIGNIFICANT indicating that treatment and control means are not significantly different from each other for that variable.

1. What variables you would consider and why? You can mention variables that are not included in the dataset but that you consider relevant.

🡪 Variables that I would use would be similar to the ones I used and outlined in the regression section.