Stata Exam 2025

Minjae Seo

March 6, 2025

Part 1

Q1: At what level is each dataset uniquely identified (i.e., what does each row represent, and which variables are unique identifiers)

Solution:

For ***Demographics dataset***, each row represents one individual in a given wave, uniquely identified by a household ID and a household member ID(and the wave indicator if both waves are in one file with different village ID). In other words, the combination of household identifier, member identifier (and wave, if applicable) is unique for each observation. Morevoer, this dataset includes household-level treatment assignment and individual demographics

For ***Assets dataset***, each row represents one asset record for a household in a given wave, uniquely identified by household ID, the asset type, and the wave. Households can have multiple asset entries (one per asset type owned) per wave with different instance numbers. Thus, the identifiers are household ID, asset type, and wave indicator(and instance numbers of durable household goods.)

For ***Depression dataset***, each row represents one individual's Kessler survey in a given wave, uniquely identified by household ID and person ID (and wave). The survey was administered only to household heads and their spouses, so only those individuals appear in this dataset (each at most twice, once per wave).

Demographics

Q2: Import the demographics dataset and calculate a proxy variable for household size based on the number of members surveyed in each household in Wave 1. Assume the household size for Wave 2 remains the same.?

Assets

Asset data is useful as a proxy for wealth. This dataset contains information on the quantity and monetary value of each individual asset owned by surveyed households. Through the questions below, you will aggregate this information to estimate a monetary value for three asset categories.

Q3: To calculate the monetary value of all assets, you should use the 'current value' variable,

which reports the monetary value of a single unit of the asset. However, you will notice that this variable is often missing. Please use the median of "current- value" for each type of asset (by type we mean, for example, "chickens", "Cutlass", "Room Furniture", "Radio", "Cell (mobile) Phone handset", etc.) to impute the missing values.?

Q4: Create a variable that contains the total monetary value for each observation, by multiplying quantity and the imputed current value.

Q5: Produce a dataset at the household-wave level (for each household, there should be at most two observations, one for each wave) which contains the following variables.

Mental Health / Depression Data

Q6: A Kessler-10 scale is a measure of mental health that uses 10 questions that iden tify how often people experience symptoms associated with depression. Using this reference, construct the kessler score (name it kessler score) and a categorical variable named kessler categories with 4 categories: no significant depression, mild depression, moderate depression, and severe depression.

Be careful: sometimes variables in the Kessler section are missing, which can complicate construction of a score. Please justify, using comments in your code, how you handle missing values.

Constructing a single dataset

Q7: At this point you have created three datasets: demographics, assets, and mental health. Please combine all three of these datasets to create a single dataset that you will use for data exploration and analysis. The unit of observation in this dataset should be an individual in a given survey round. (There should be at most two observations per individual, one for Wave 1 and another for Wave 2).

Solution (Q2-Q7): All analyses have been conducted in the `Part1.do` file. Please refer to the code for detailed implementation.

How to Execute the Code: To run the analysis, update the global file paths in the `.do` file in my zip folders accordingly. You may execute the code either line by line or run the entire script at once, depending on your preference.

Part 2: Exploratory Analysis

Using Wave 1 data, conduct exploratory analysis to understand the relationship between depression and household and demographic characteristics among individuals in Ghana. Specifically, do the following:

Q1: Explore the relationship between depression and:

1. Household wealth, proxied by total asset value.
2. A household or demographic characteristic that seems interesting to you.

Table 1: Summary Statistics: Kessler Score and Total Assets

Variable	Observations	Mean	Std. Dev.	Min	Max
Kessler Score	7,428	19.60	7.10	1	50
Total Assets	7,440	11,152.39	119,049.8	1.5	5,791,358

Kessler scores measure depression levels, while total assets serve as a proxy for household wealth.

Table 2: Correlation Matrix: Kessler Score and Total Assets

Variable	Kessler Score	Total Assets
Kessler Score	1.000	-0.0051
Total Assets	-0.0051	1.000

The correlation between depression and wealth is nearly zero, indicating no clear relationship.

Table 3: Regression: Household Wealth and Depression

Variable	Coefficient	Std. Error	p-value
Total Assets	-3.05e-07	6.91e-07	0.658
Constant	19.59	0.08	0.000***
Observations	7,412		
R-squared	0.0000		
Adj. R-squared	-0.0001		
Root MSE	7.094		

Regression results suggest that household wealth does not significantly predict depression levels ($p = 0.658$).

Table 4: Summary Statistics: Age and Depression

Variable	Observations	Mean	Std. Dev.	Min	Max
Age	7,426	49.39	16.36	4	107
Kessler Score	7,398	19.58	7.09	1	50

The average respondent is around 49 years old, with a wide age range from 4 to 107.

Table 5: Correlation Matrix: Age and Kessler Score

Variable	Kessler Score	Age
Kessler Score	1.000	0.119
Age	0.119	1.000

A weak positive correlation (0.119) suggests older individuals may experience slightly higher depression.

Table 6: Regression: Impact of Age on Depression

Variable	Coefficient	Std. Error	p-value
Age	0.0517	0.0050	0.000***
Constant	17.03	0.26	0.000***
Observations	7,398		
R-squared	0.0142		
Adj. R-squared	0.0141		
Root MSE	7.039		

Age significantly predicts depression ($p < 0.001$), though with a small effect size.

Table 7: Descriptive Statistics: Kessler Score by Gender

Gender	Mean	Std. Dev.	Min	Max
Male	19.06	6.98	1	50
Female	20.02	7.16	1	47

Females report slightly higher depression scores (20.02) than males (19.06).

Table 8: Regression: Gender Differences in Depression Scores

Variable	Coefficient	Std. Error	p-value
Gender (Female = 1)	0.96	0.165	0.000***
Constant (Male)	19.06	0.12	0.000***
Observations	7,428		
R-squared	0.0045		
Adj. R-squared	0.0044		
Root MSE	7.08		

Women have significantly higher depression scores than men ($p < 0.001$), though the effect size is small.

Interpretation:

Table 1 presents summary statistics for Kessler scores and total household assets. The mean Kessler score is 19.60, with a standard deviation of 7.10. Household wealth, proxied by total assets, has a highly skewed distribution, with a mean of 11,152.39 but a maximum value exceeding 5.7 million. Table 2 and Table 3 show that there is no significant relationship between total assets and depression. The correlation coefficient is approximately zero (-0.0051), and the regression coefficient is also insignificant ($p = 0.658$). This suggests that wealth is not a strong predictor of depression within this dataset. Figure 1 further supports this conclusion, showing a near-flat trend in the scatter plot of Kessler scores against household wealth. Additionally, the other figure displays box plots of depression scores across different wealth quartiles, showing minimal variation across the groups.

Table 4 and Table 5 reveal a weak positive correlation (0.119) between age and depression. The regression results in Table 6 confirm that age is a significant predictor of depression ($p < 0.001$), though the effect size is small. Figure 1 illustrates the distribution of Kessler scores across different age categories. It suggests that older individuals tend to report slightly higher depression scores. Figure 1 presents a scatter plot with a fitted regression line, further confirming this positive but weak relationship. Figure 1 presents a box plot of Kessler scores by gender, visually demonstrating the slightly higher depression scores among women compared to men.

Evaluating the RCT

Using Wave 2 data to measure outcomes, answer the following questions, explaining any decisions and assumptions you make, and interpret your results. There is no need for you to address the validity of the random assignment of the intervention.

Q2: Were the GT sessions effective at reducing depression?

Table 9: Regression: GT Treatment and Depression

Variable	Coefficient	Std. Error	p-value
Treatment Household	-0.54	0.14	0.000***
Constant	17.52	0.10	0.000***
Observations	6,386		
R-squared	0.0024		
Adj. R-squared	0.0023		
Root MSE	5.47		

GT sessions had a statistically significant effect on reducing depression, as indicated by the negative coefficient (-0.54) on the treatment household variable. The p-value of 0.000 confirms that this effect is statistically significant at conventional levels. However, the R-squared value (0.0024) indicates that the treatment only explains a very small portion of the variation in depression scores. While the reduction in depression is statistically significant, the effect size is relatively small, suggesting that while GT sessions may have had some beneficial impact, other factors likely play a much larger role in influencing depression levels. This is further supported by Figure 2, which shows similar distributions of depression scores between the treatment and control groups. The GT sessions appear to have had a slight effect in reducing depression, the magnitude of the effect is relatively modest, and additional interventions may be needed to achieve more substantial mental health improvements.

Q3: Did the effect of GT sessions on depression differ by gender? Perform a linear regression of the Kessler Score on:

1. Woman (binary variable)
2. Treated Household (binary variable)
3. Interaction term: Treated Household * Woman, using only Wave 2 observations.

Note: In your write-up for this question, please make sure to explain and interpret all coefficients in your specification, keeping in mind units and reference groups.

Table 10: Regression: Gender and Treatment Interaction

Variable	Coefficient	Std. Error	p-value
Treatment Household	-0.13	0.21	0.531
Woman (Female=1)	0.81	0.20	0.000***
Treatment Household × Woman	-0.72	0.28	0.009**
Constant	17.06	0.15	0.000***
Observations	6,386		
R-squared	0.0052		
Adj. R-squared	0.0047		
Root MSE	5.46		

Interpretation of Each Coefficient:

Constant: Represents the mean depression score (Kessler score) for men in the control group (i.e., untreated men). The average depression score for untreated men is 17.06.
Treatment Household: The coefficient for treatment is not statistically significant ($p > 0.05$). This suggests that, for men, being in a treated household did not significantly impact depression levels.
Woman: Women in the control group (untreated households) have depression scores that are 0.81 points higher than men, which is statistically significant.
Interaction Term (Treatment Household × Woman): This coefficient tells us how GT sessions affected women compared to men. A negative coefficient means that the treatment effect was weaker for women than for men. The effect of treatment reduces the additional depression burden women have by 0.72 points but does not completely eliminate the gender difference. Women in treated households still have slightly higher depression than men but experience a smaller gap than untreated women.

Table 11: Regression: Gender and Treatment Interaction adding age variable

Variable	Coefficient	Std. Error	p-value
Age	0.0627	0.0043	0.000***
Treatment Household	-0.1016	0.2036	0.618
Woman(Female=1)	0.9812	0.1922	0.000***
Treatment × Woman	-0.7804	0.2713	0.004**
Constant	13.9522	0.2570	0.000***
Observations	6,385		
R-squared	0.0374		
Adj. R-squared	0.0368		
Root MSE	5.3762		

Interpretation of Each Coefficient:

Constant: The average depression score for untreated men is 13.95.
age: Older individuals tend to have higher depression scores. Each additional year of age is associated with a 0.06 point increase in depression.
Treatment Household: Treatment does not significantly affect depression levels when controlling for age.
Woman: The gender gap in depression remains significant. Women still report higher depression scores than men.
Interaction Term (Treatment Household × Woman): The coefficient is slightly larger in magnitude (-0.78 vs. -0.72 in the first model), suggesting that age does not fully explain the gendered differences in treatment effects.

Conclusion

The analysis finds that GT sessions did not significantly reduce depression for either men or women. The treatment coefficient remains statistically insignificant in both models, even after controlling for age. This suggests that GT sessions were ineffective in lowering overall depression scores. Additionally, women consistently report higher depression scores than men. In both models, women in the control group have significantly higher depression scores (0.8 - 0.98 points higher on average) compared to men. This finding aligns with previous research indicating that women tend to experience higher levels of depression. The effect of GT sessions on depression differed by gender. The negative and significant interaction term (-0.72 to -0.78) suggests that the treatment was less effective for women. While the intervention did slightly reduce the gender gap in depression, it did not fully eliminate it. For men, GT sessions had no statistically significant effect. Lastly, age is positively associated with depression. After controlling for age, the results indicate that older individuals tend to have slightly higher depression scores, with an increase of approximately 0.06 points per year. However, the gender gap in depression and the ineffectiveness of GT treatment persist even after accounting for age.

The key findings are **1. GT sessions did not significantly reduce depression. 2. Women reported higher depression scores than men. 3. GT sessions slightly reduced the gender gap in depression but were not effective overall. 4. Older individuals tend to have slightly higher depression scores**

Appendix:Figures

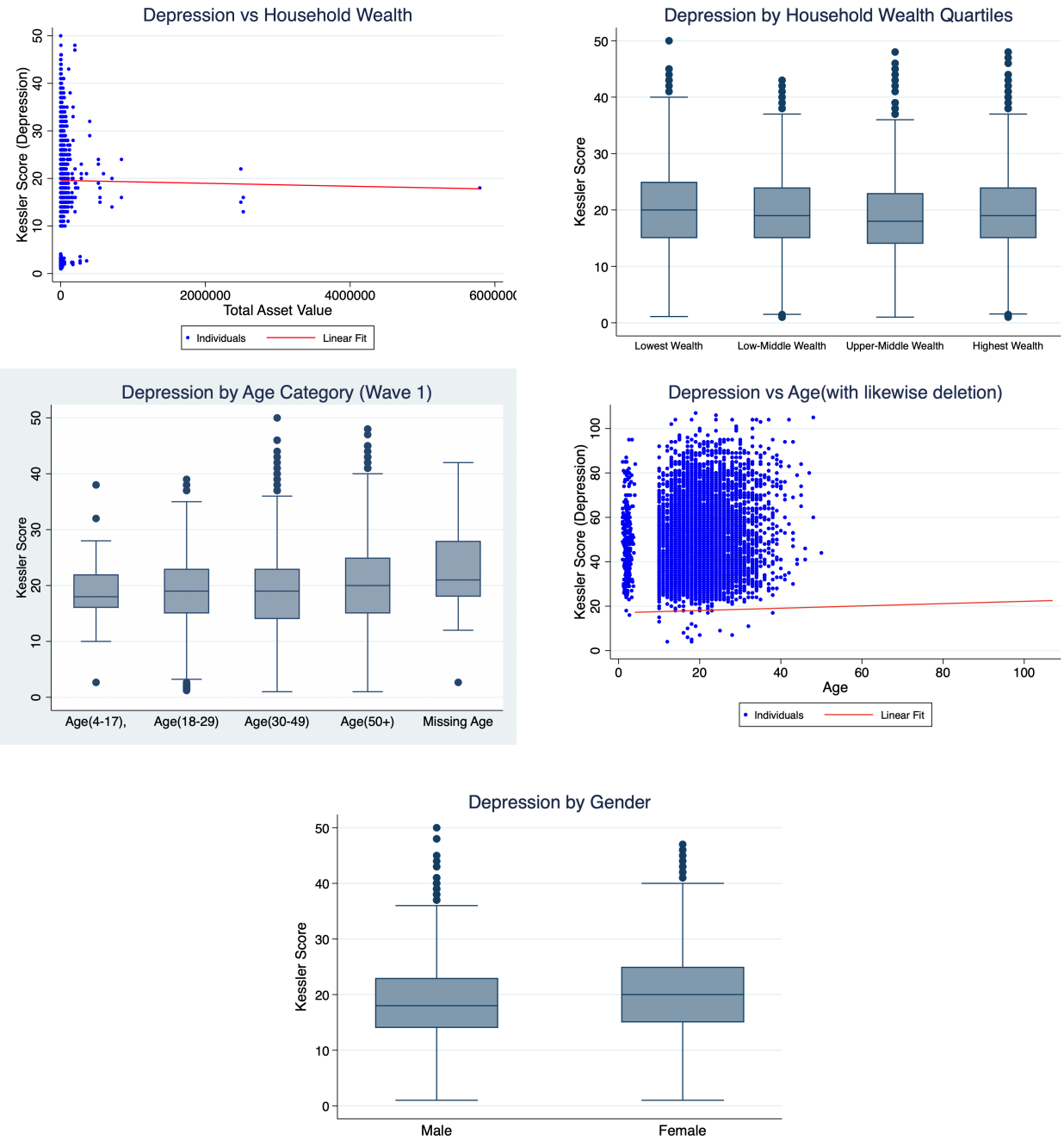

Figure 1: **Exploratory Data Analysis (Part 2-1)**

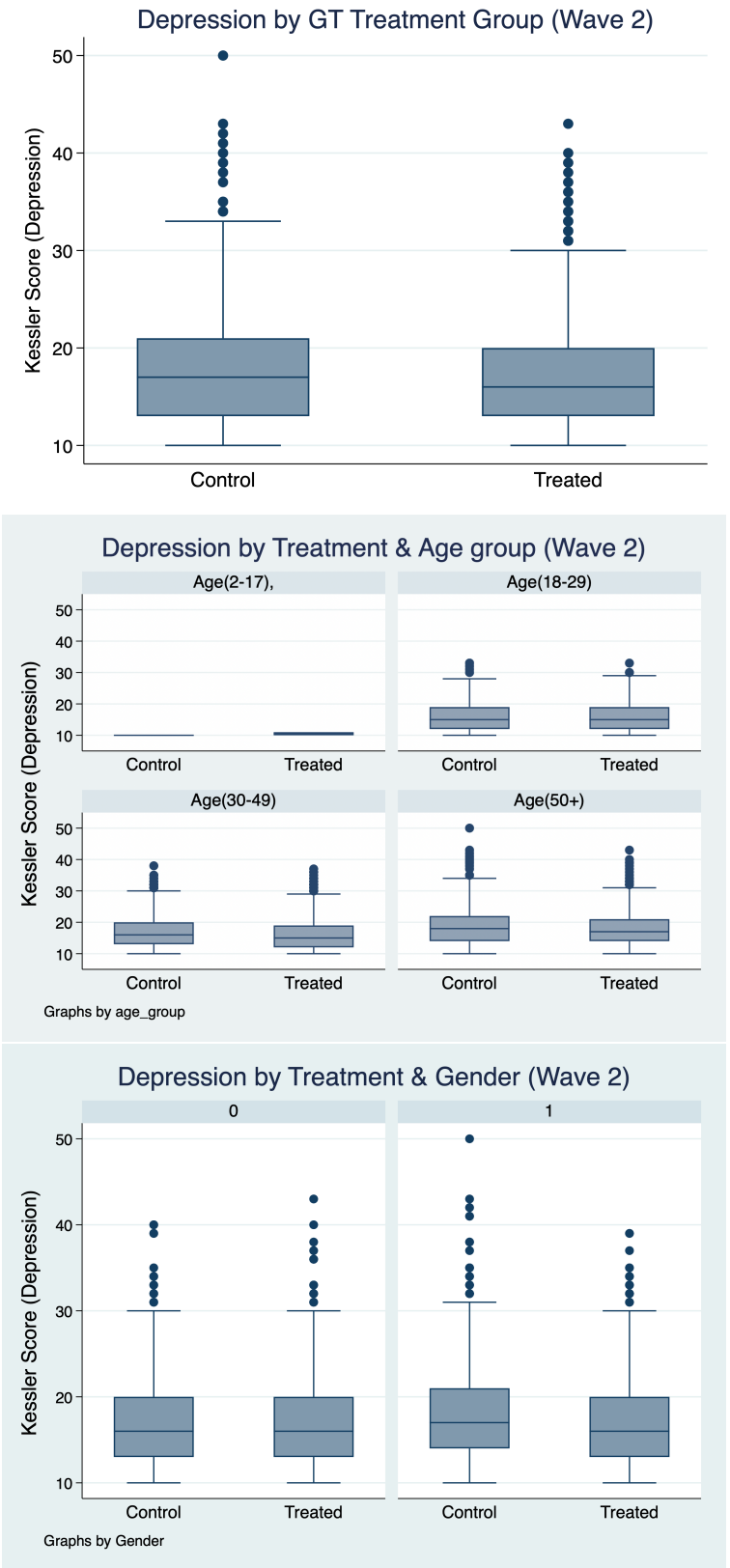

Figure 2: **Exploratory Data Analysis (Part 2-2)**