Stata Assessment

Minjae Seo

April 17, 2024

Part 2: Exploratory Analysis

Please look at the figure section for the Figures of Exploratory Data Analysis

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

Q1: Explore the relationship between depression and:

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

Present the results from your exploration through tables, plots, a write-up, or anything else you think would be useful. Given using wave 1 data only

```
keep if wave == 1
```

Depression and household(Household wealth), proxied by total asset value

- Summary statistics for total asset value and depression score:
  The average total asset value is approximately 966 with a wide variation, as indicated by the high standard deviation. Asset values range from very low (1.5) to quite high (41,685), suggesting economic diversity among the individuals. Moreover, the average depression score is around 20(mild depression), and the range is from 0 to 50, suggesting that there are individuals with no depressive symptoms as well as individuals with high levels of depressive symptoms.
- Correlation between depression score and total asset value:
  There is a very low positive correlation between total asset value and depression scores, which means that as total asset value increases, there is a slight tendency for depression scores to increase as well, but this relationship is not strong.
- Scatter plot of depression score(kessler score) and total asset value:
  The scatter plot of depression scores versus total asset value shows a dense cluster of data at the lower total asset value. There are a few outliers with high asset values but the plot does not indicate a clear pattern between assets and depression scores, which corresponds with the low correlation coefficient.

Depression and demographic characteristic

Demographic characteristic: Age

Age is a continuous variable but ignores the outlier age ≤ 0(-999), which I didn't drop due to the effect of dropping other important information for other columns.

- Summary statistics for age and depression score:
  The average age is approximately 45 but with a very large range (nan to 107).
  As I indicated earlier, ignore age smaller than 0.

Demographic characteristic: Gender

Gender is an indicator variable where 1 - Male, 5 - Female

- Frequncy of Gender:
  In total 7,456 individuals, with a higher number of females (55.61%) compared to males (44.39%).

- Summary statistics of depression scores by gender:
  On average, females report higher depression scores (mean 20.7) compared to males (mean 19.6). The standard deviation is similar for both genders, indicating similar variability in depression scores within each gender.

- Box Plot of Depression Scores by Gender:
  The boxplot shows a higher median depression score for females, and both genders have a similar range of scores. There is a notable number of outliers for both genders, especially males, which could be interesting to investigate further about outliers.

- Histogram of Depression Scores by Gender:
  This histogram visually indicates the greater number of female individuals compared to male individuals in the RCT.

- Difference in depression scores between genders:
  The result shows a statistically significant difference in depression scores between genders, with females having a higher average score (20.69706) than males (19.56918).

Household: treatment_hh

treatment_hh is an indicator variable where 1 - treated household, 5 - not treated household

- Summary statistics of depression scores by treatment indicator:
  The average depression scores are very similar between control (20.19753) and treatment (20.19518) households, suggesting that the treatment, at this stage, has not resulted in a significant difference in depression score

- Box Plot of Depression Scores by Treatment Status:
  The boxplot indicates that the median depression score is roughly the same for both control and treatment households. Both groups also display a similar range of depression scores and the presence of outliers.

- Histogram of Depression Scores by Treatment status:
  The histogram shows that the distribution between control and treatment households is nearly even in Wave 1, which is good for comparative analysis.

- Difference of Kessler scores between treatment groups:
  A two-sample t-test reveals that there is no statistically significant difference in the mean depression scores between the control and treatment groups in Wave 1 (p-value ¿ 0.05).

Regression Analysis

- Regress depression scores on total asset value:
  There is a statistically significant negative relationship between total asset value and depression scores. For every unit increase in total asset value, depression scores decrease by 0.0000607, holding other factors constant. The p-value (¡0.01) indicates this effect is highly significant.

- Regress depression scores on age:
  p-value not being below common thresholds for significance, this relationship is not statistically reliable with the given data

- Regress depression scores on gender:
  Gender shows a significant positive effect on depression scores, with a coefficient of 0.199. This indicates that being male(Yes(1) male) is associated with an average increase of 0.199 points in depression scores compared to females, holding other factors constant. This effect is highly significant (p¡0.01).

- Regress depression scores on treatment status:
  Treatment status is significantly associated with a decrease in depression scores. Being in a treatment household is associated with a 0.247 point decrease in depression scores, with other factors held constant. This effect is significant at the 0.05 level.

- Wrap-up:
  The regression analysis indicates that total asset value and gender have significant effects on depression scores, with total asset value having a protective effect and female gender being associated with higher scores. Since we only looked at a few of variables, there might be other unobserved factors that might contribute to variations in depression scores.

Regression Analysis : Impact of each variable on depression scores

	(1)	(2)	(3)	(4)
Covariates	kessler_score	kessler_score	kessler_score	kessler_score
total_asset_value	-6.07e-05***			
	(4.73e-06)			
age		0.00121		
		(0.00152)		
gender			0.199***	
			(0.0262)	
treatment				-0.247**
				(0.105)
Constant	19.03***	18.78***	18.19***	18.96***
	(0.0552)	(0.0883)	(0.0990)	(0.0740)
Observations	13,822	13,842	13,842	13,842
R-squared	0.010	0.000	0.004	0.000

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Evaluating the RCT

Using Wave 2 data to measure outcomes, answer the following questions, explain 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.
Given using wave 2 data only

```
keep if wave == 2
```

Assumption for my RCT evaluation

- No Spillover Effects: The treatment-assigned group does not affect the control group. In other words, no interaction between treated and untreated households could influence the outcomes of the latter.
- Stable Unit Treatment Value Assumption (SUTVA): This assumes that the treatment effect is consistent across all units, and there is only one version of the treatment.
- No Selection Bias Post-Treatment: No systematic difference in dropout rates or survey response rates between the treated and untreated groups after the intervention.
- Measurement accuracy/error: The method for measuring depression (Kessler score) remains consistent and reliable across both waves. However, kessler score is likely measured inconsistently with both waves due to the survey question.
- No Unmeasured Confounder: All other variables that affect the outcomes are due to randomization.
- Other assumptions: Compliance(given perfect attendance of gt), Randomization(randomized control setting), Temporal Stability(any other external factors than intervention(GT) are consistent with both wave)

Q2: Were the GT sessions effective at reducing depression?
To check whether Group Therapy sessions in wave 2 are effective for decreasing depression, I will do a statistical analysis to compare depression scores(kessler score) between treatment(treated household) and control(controlled household) at wave 2(after intervention).

- Randomization between treated and not treated groups:
  Similar with Wave 1, the distribution between control and treatment households in Wave 2 remains nearly even.
- Randomization between Gender:
  Frequency Table of Gender (Wave 2, Figure 20): The distribution of gender remains consistent, with females representing a higher percentage of the sample.
- Summary statistics of Kessler score and treatment status:
  There's a reduction in the mean depression score in Wave 2 for both groups compared to Wave 1, with treatment households showing a slightly lower mean score than control households.
- Summary statistics of depression scores by treatment status:
  Summary Statistics of Depression Scores by Treatment Status (Wave 2, Figure 18): Similar to Wave 1, Wave 2 data also shows that the mean depression scores between control and treatment households are very close, with overlapping standard deviations.
- Regression Analysis: average treatment effect for treated wave 2(after intervention)

There is a statistically significant negative effect of treatment on depression scores, with the coefficient for treat_hh being -0.539. This suggests that, on average, being in a treated household is associated with a reduction of approximately 0.539 points in the depression score as measured by the Kessler psychological distress scale (kessle_score), controlling for other factors. The three asterisks indicate that this result is significant at the 1% level (p¡0.01), implying strong evidence against the null hypothesis of no effect. Given the statistical significance of the treatment effect, we can conclude that the Group Therapy was effective at reducing depression scores on average among participants in the treatment group during wave 2 of the study. However, we still have to keep an eye on validation of our assumptions such as omitted variable bias and external validity.

Table 1: Impact of Group Therapy on Depression Scores(Wave 2)

	(1)
VARIABLES	kessler_score
treatment	-0.539***
	(0.137)
Constant	17.52***
	(0.0990)
Observations	6,386
R-squared	0.002

Robust standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Q3: Did the effect of the GT sessions on depression vary for men and women? To answer this question perform a linear regression of the Kessler Score against a ”Woman” binary variable, a ”Treated Household” binary variable, and an 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.

Present your results as you see fit and add a write-up with your interpretation and conclusions.

Table 2: Regression Analysis: Interaction Effects of Treatment and Gender on Depression Score

	(1)
VARIABLES	kessler_score
treatment	-0.130
	(0.205)
gender	0.0224
	(0.0477)
treatment:gender	0.723***
	(0.275)
Constant	17.04***
	(0.155)
Observations	6,386
R-squared	0.005

Robust standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Linear Regression of Differential Treatment Effect: Measuring the Average Treatment Effect of GT by Gender

In the regression model:

$$\text{kessler\_score\_i} = \beta_0 + \beta_1 \text{gender\_i} + \beta_2 \text{treat\_hh\_i} + \beta_3 (\text{gender\_i} \times \text{treat\_hh\_i}) + \epsilon\_i$$

- i = index for each individual in each household at wave 2
- kessler_score_i = the depression score of individual i
- gender_i = binary indicator of each individual's gender (1 = male, 5 = female)
- treat_hh_i = binary indicator of each household's treatment status (0 = control, 1 = treated)

Based of the regression table

$$\text{kessler\_score\_i} = 17.04 - 0.13 \times \text{gender\_i} + 0.02 \times \text{treat\_hh\_i} + 0.723(\text{gender\_i} \times \text{treat\_hh\_i}) + \epsilon\_i$$

Interpretation of the coefficients:

- $\beta_0 = 17.04$, The average baseline depression score for female in the control (non-treated) households is 17.04 points.
- $\beta_1 = -0.13$, The avaerage difference in depression score between male and female within household is not treated is -0.13 points, but the effect is insignificant(p ¿ 0.05, standard error of 0.205).
- $\beta_2 = 0.02$, The average difference in depression score between treated and control household for female individual is 0,02 points, also the effect is not statistically significant (standard error of 0.0477).
- $\beta_3 = 0.723$, The treatment's effect on reducing depression scores is more present for males than for females. Specifically, for males, being in a treated household is associated with an additional decrease of 0.723 points in depression scores over the effect observed in females. And, the interaction term is highly significant.
- ϵ_i, the error term, capturing random influences on depression score for individual i

For the differential effect

Interpretation and Conclusion:
The regression analysis suggests that the impact of GT sessions on depression does indeed vary significantly by gender:

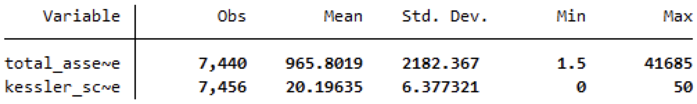
For Female:
The treatment does not have a statistically significant effect on reducing depression scores.

For Males:
The effect of the treatment is both positive and significant, implying that GT sessions are more beneficial for male individual, helping to reduce their depression score significantly more than for female.

Conclusion:
Therefore, the GT sessions' effectiveness in reducing depression scores varies between men and women, with a notably greater benefit observed for men.

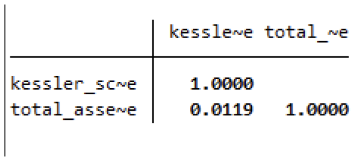
Exploratory Data Analysis Figures

Due to the size of the figures, I inserted all of the data analysis figures into the last pages



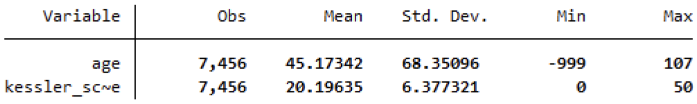
```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
total_asse~e |      7,440    965.8019    2182.367        1.5      41685
kessler_sc~e |      7,456    20.19635    6.377321          0         50
```

Figure 1: Summary statistics for total asset value and depression score(Wave 1)



```
             | kessle~e total_~e
-------------+------------------
kessler_sc~e |   1.0000
total_asse~e |   0.0119   1.0000
```

Figure 2: Correlation between depression scores and total asset values(Wave 1)



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      7,456    45.17342    68.35096       -999        107
kessler_sc~e |      7,456    20.19635    6.377321          0         50
```

Figure 3: Summary statistics for age and depression score(Wave 1)

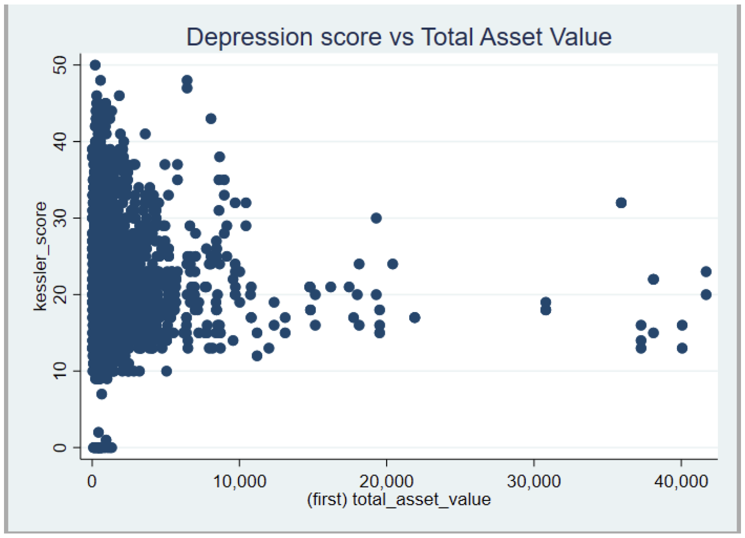
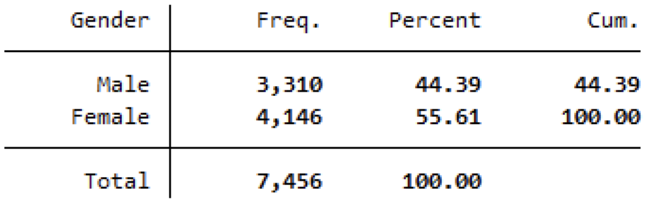
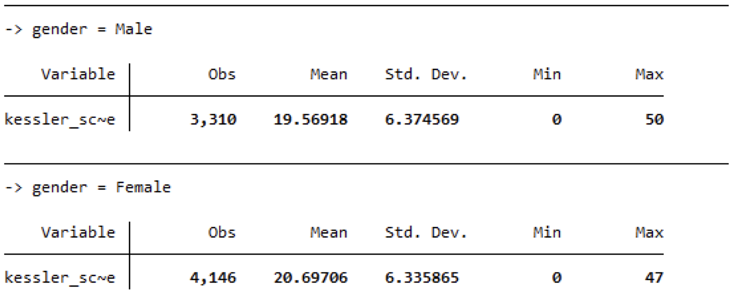


Figure 4: Scatter plot of depression scores(depression scores) and total asset value(Wave1)



```
     Gender |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |      3,310       44.39       44.39
     Female |      4,146       55.61      100.00
------------+-----------------------------------
      Total |      7,456      100.00
```

Figure 5: Frequencies table of Gender(Wave 1)



```
------------------------------------------------------------------------
-> gender = Male

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kessler_sc~e |      3,310    19.56918    6.374569          0         50

------------------------------------------------------------------------
-> gender = Female

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kessler_sc~e |      4,146    20.69706    6.335865          0         47
```

Figure 6: Summary Statistics by Gender(Wave 1)

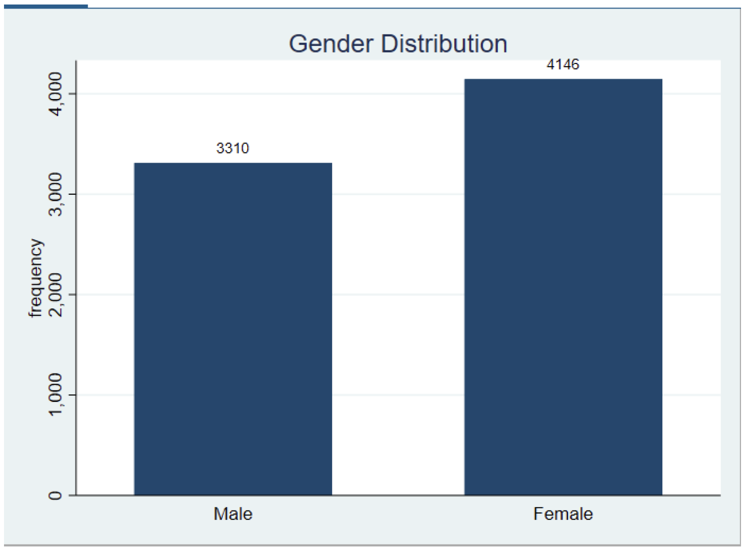


Figure 7: Histogram of gender Distribution(Wave 1)

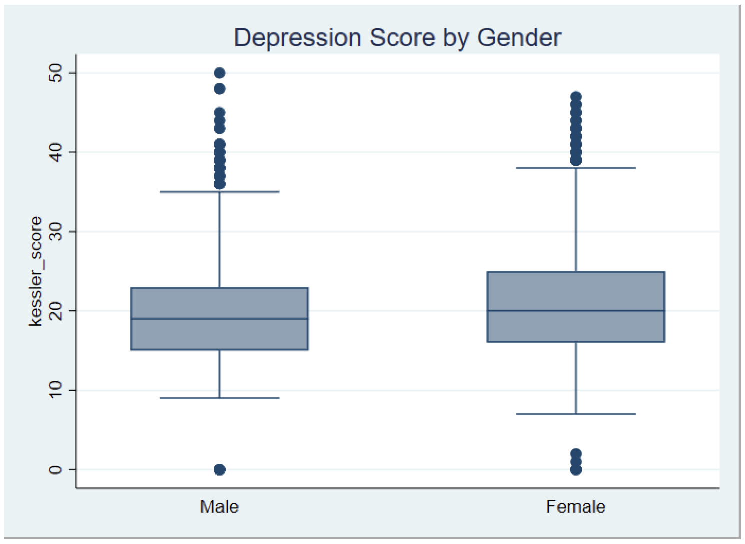


Figure 8: Boxplot of depression scores by Gender(Wave 1)

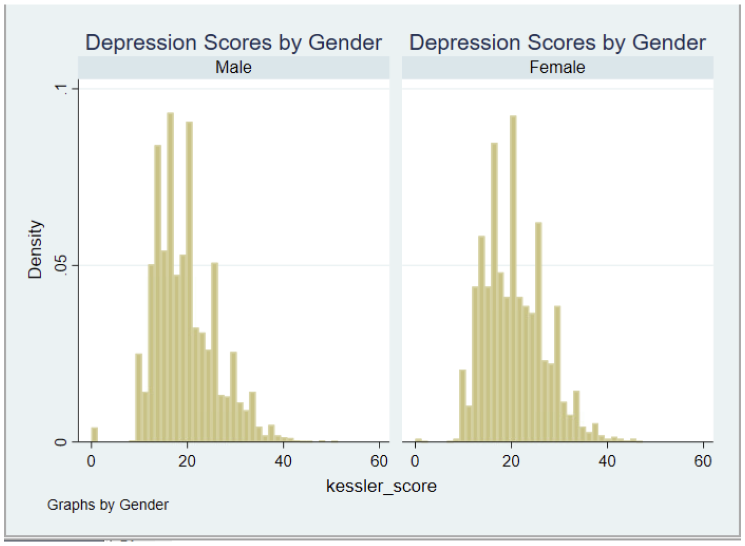
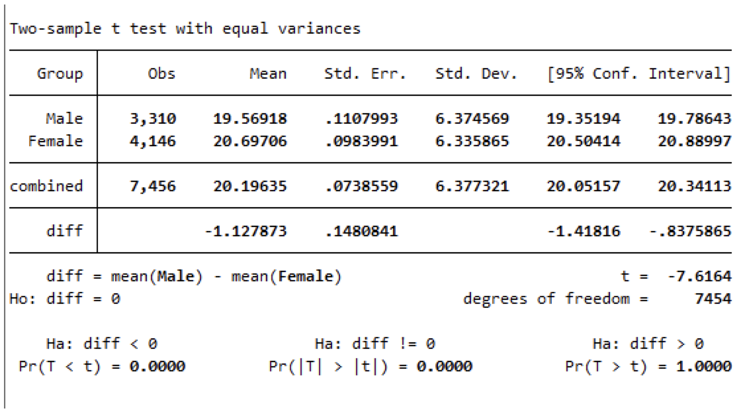
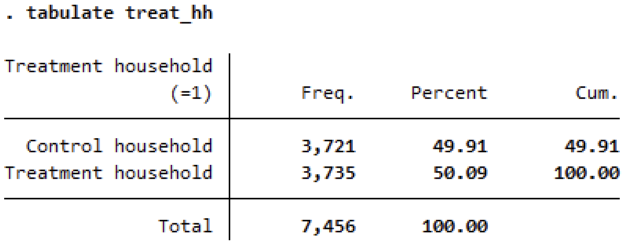


Figure 9: Histogram of depression scores by Gender(Wave 1)



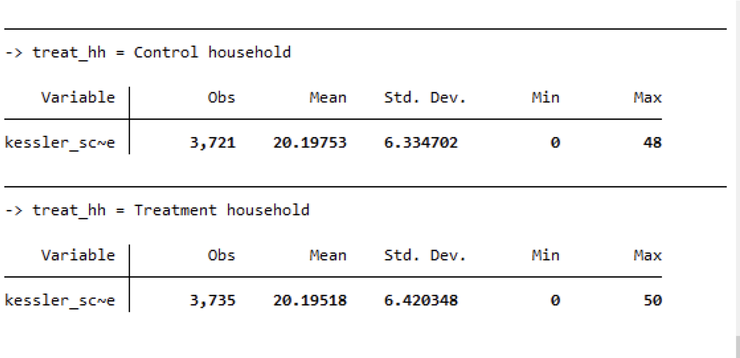
```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |   3,310    19.56918    .1107993    6.374569    19.35194    19.78643
  Female |   4,146    20.69706    .0983991    6.335865    20.50414    20.88997
---------+--------------------------------------------------------------------
combined |   7,456    20.19635    .0738559    6.377321    20.05157    20.34113
---------+--------------------------------------------------------------------
    diff |           -1.127873    .1480841               -1.41816   -.8375865
------------------------------------------------------------------------------
    diff = mean(Male) - mean(Female)                              t =  -7.6164
Ho: diff = 0                                     degrees of freedom =     7454

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Figure 10: Difference in depression scores between genders(Wave 1)



```
. tabulate treat_hh

Treatment household |
              (=1) |      Freq.     Percent        Cum.
-------------------+-----------------------------------
  Control household |      3,721       49.91       49.91
Treatment household |      3,735       50.09      100.00
-------------------+-----------------------------------
              Total |      7,456      100.00
```

Figure 11: Frequency table of treatment status indicator(Wave 1)



```
-------------------------------------------------------------------------------
-> treat_hh = Control household

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kessler_sc~e |      3,721    20.19753    6.334702          0         48

-------------------------------------------------------------------------------
-> treat_hh = Treatment household

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kessler_sc~e |      3,735    20.19518    6.420348          0         50
```

Figure 12: Summary Statistics of depression scores by treatment status(Wave 1)

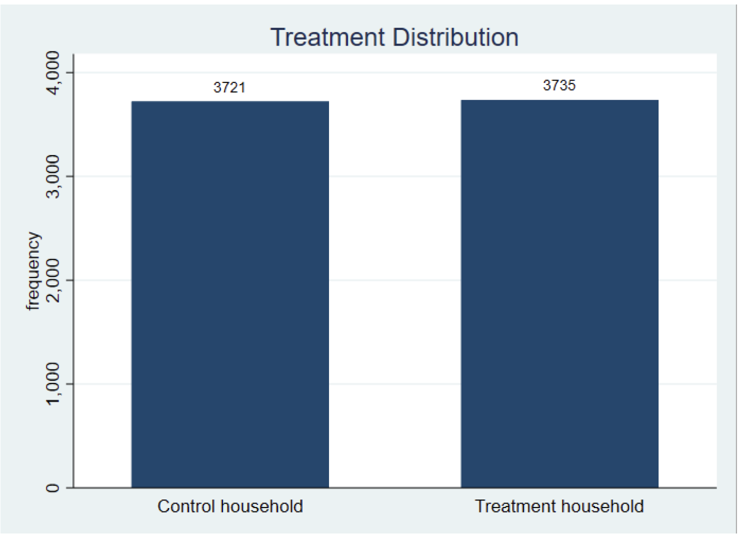


Figure 13: Histogram of treatment household distribution(Wave 1)

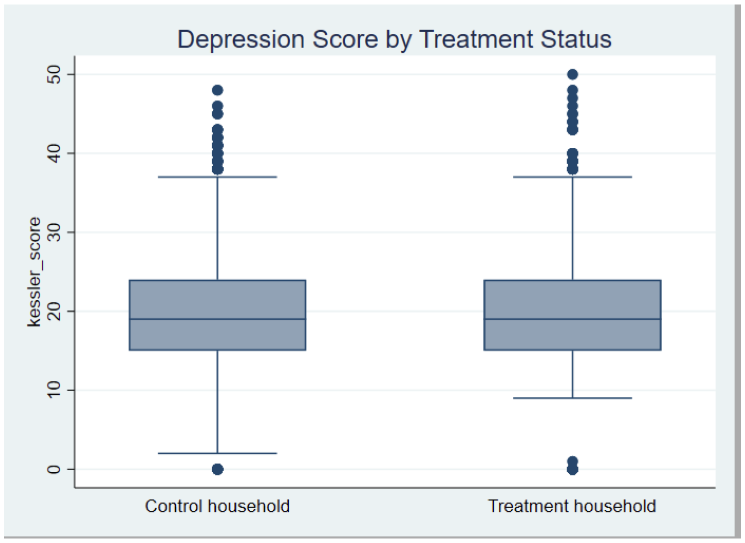


Figure 14: Boxplot of depression scores by treatment status(Wave 1)

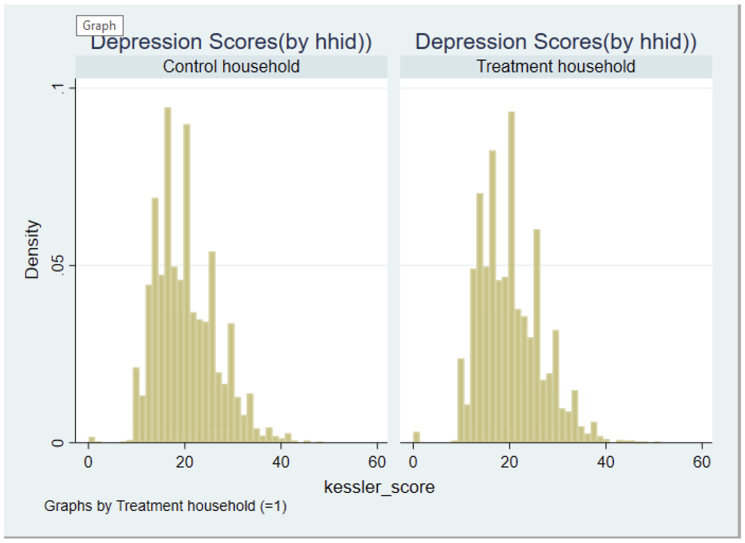
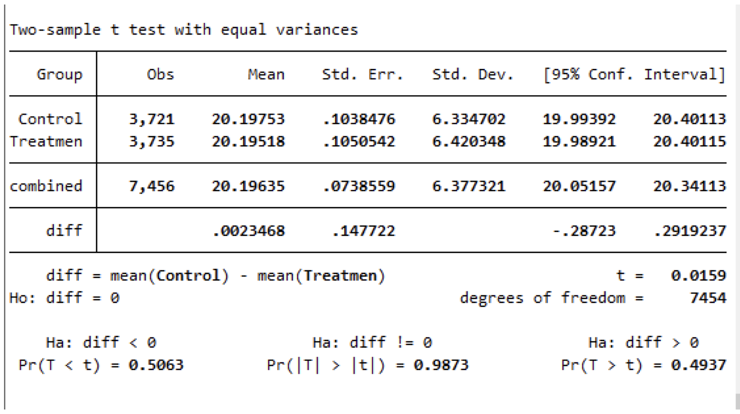
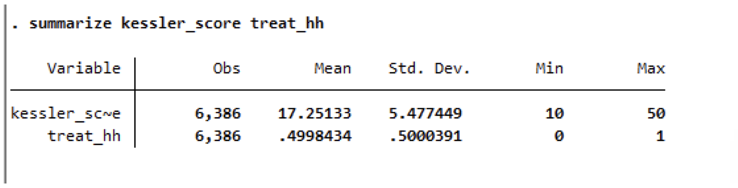


Figure 15: Histogram of depression scores by treatment assigned household(Wave 1)



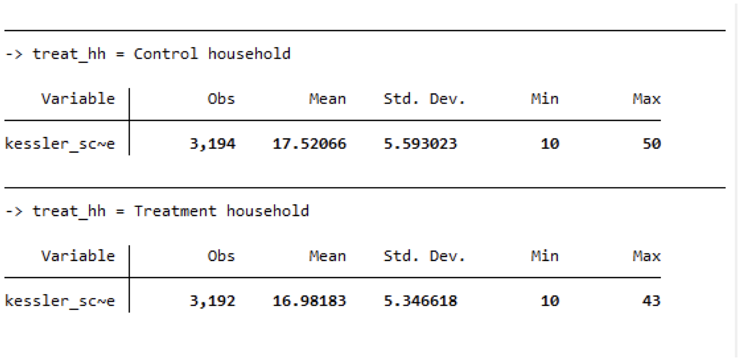
```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Control |   3,721    20.19753    .1038476    6.334702    19.99392    20.40113
Treatmen |   3,735    20.19518    .1050542    6.420348    19.98921    20.40115
---------+--------------------------------------------------------------------
combined |   7,456    20.19635    .0738559    6.377321    20.05157    20.34113
---------+--------------------------------------------------------------------
    diff |            .0023468     .147722                -.28723    .2919237
------------------------------------------------------------------------------
    diff = mean(Control) - mean(Treatmen)                         t =   0.0159
Ho: diff = 0                                     degrees of freedom =     7454

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.5063         Pr(|T| > |t|) = 0.9873          Pr(T > t) = 0.4937
```

Figure 16: Difference in depression scores between treatment assigned(Wave 1)



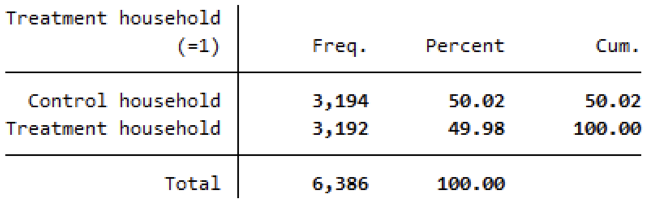
```
. summarize kessler_score treat_hh

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kessler_sc~e |      6,386    17.25133    5.477449         10         50
    treat_hh |      6,386    .4998434    .5000391          0          1
```

Figure 17: Summary Statistics of depression scores and treatment status(Wave 2)



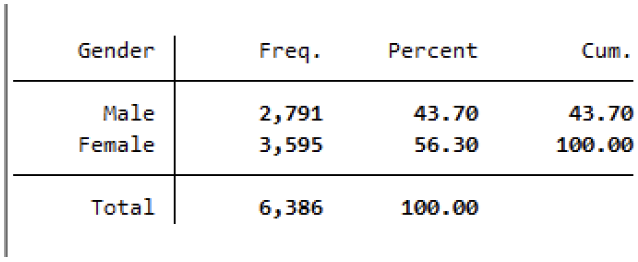
```
-> treat_hh = Control household

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kessler_sc~e |      3,194    17.52066    5.593023         10         50

-> treat_hh = Treatment household

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kessler_sc~e |      3,192    16.98183    5.346618         10         43
```

Figure 18: Summary Statistics of depression scores by treatment status(Wave 2)



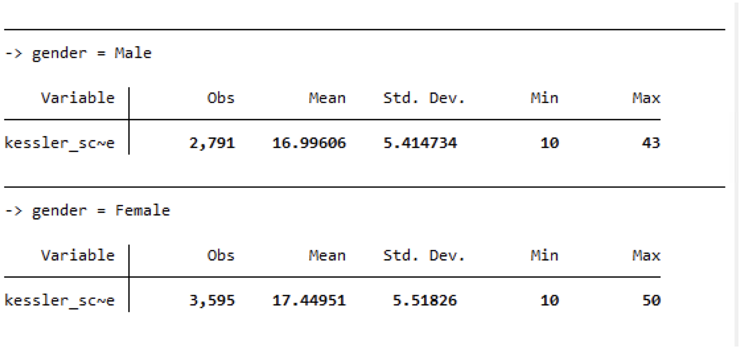
```
Treatment household |
               (=1) |      Freq.     Percent        Cum.
--------------------+-----------------------------------
  Control household |      3,194       50.02       50.02
Treatment household |      3,192       49.98      100.00
--------------------+-----------------------------------
              Total |      6,386      100.00
```

Figure 19: Frequency table of treatment status and controlled status(Wave 2)



```
     Gender |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |      2,791       43.70       43.70
     Female |      3,595       56.30      100.00
------------+-----------------------------------
      Total |      6,386      100.00
```

Figure 20: Frequency table of treatment status and controlled status(Wave 2)



```
-> gender = Male

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kessler_sc~e |      2,791    16.99606    5.414734         10         43

-> gender = Female

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
kessler_sc~e |      3,595    17.44951     5.51826         10         50
```

Figure 21: Summary Statistics of depression scores by Gender(Wave 2)