Group therapy effect on alleviating depression in Ghana (RCT)

Oscar Cuadros

Instructions

We have provided you with survey data collected from households in Ghana in two waves. These households were part of a Randomized Controlled Trial (RCT), where participants in treated households received Group Therapy to alleviate symptoms of depression.1

To explore whether Group Therapy (GT) alleviates depression or not, an RCT was conducted in partnership with a local organization that provides GT sessions designed to improve mental health through guided discussions led by a trained facilitator.

After Wave 1 of data collection was complete, half of the households in the sample were randomly selected to receive the opportunity to attend weekly GT sessions for a period of three months at no cost to participants. In each treatment household, the household head and their spouse were invited to GT sessions. For the purposes of this analysis, assume that there was perfect attendance at these sessions. Once all GT sessions were completed, Wave 2 data were collected (6 months after Wave 1).

Through the exercises below, you will prepare the data, conduct exploratory analysis, and present findings for the RCT.

There are three datasets, which include data collected at Wave 1 (before the intervention), and Wave 2 (after the intervention). The datasets are as follows:

1. Demographics. This dataset includes treatment assignment at the household level and demographic information for each member in the sampled households.
2. Assets. This dataset includes the quantity and monetary value of assets owned by the household, under three categories of assets: Animals, Tools, and Durable Goods.
3. Depression information. This dataset includes information for the Kessler Psychological Distress Scale collected for household heads and their spouses.

PART I

Demographics

1. Import the demographics dataset, and calculate a variable proxying for household size, based on how many members were surveyed in each household in Wave 1.

I grouped the variable "hhid", filtered the "wave" variable that equals 1, and found the maximum value in "hhmid". I created the temporary dataset "max_values" with this information. Later, I joined the "max_values" dataset with "demographics" to have the expected answer. Please feel free to check the following outcomes.

```
demographics <- read_dta(file.path(path, "demographics.dta"))

max_values <- demographics %>%
  group_by(hhid) %>%
  filter(wave == 1) %>%
  summarize(max = max(hhmid))

max_values
```

```
## # A tibble: 5,009 x 2
##         hhid   max
##        <dbl> <dbl>
##  1 101001002     5
##  2 101001003     1
##  3 101001004     1
##  4 101001009     2
##  5 101001010     2
##  6 101001012     3
##  7 101001013     2
##  8 101001015     6
##  9 101001022     3
## 10 101001023     3
## # ... with 4,999 more rows
```

```
final_demographics <- demographics %>%
  left_join(max_values, by = "hhid")

final_demographics %>%
  select(wave, hhid, max)
```

```
## # A tibble: 34,427 x 3
##     wave      hhid   max
##    <dbl>     <dbl> <dbl>
##  1     1 101001002     5
##  2     1 101001002     5
##  3     1 101001002     5
##  4     1 101001002     5
##  5     1 101001002     5
##  6     2 101001002     5
##  7     2 101001002     5
```

```
##  8     2 101001002     5
##  9     2 101001002     5
## 10     1 101001003     1
## # ... with 34,417 more rows
```

Assets

2. To calculate the monetary value of all assets, you should use the 'currentvalue' variable. However, you will notice that this variable is often missing (usually for durable goods). Please use the median of "currentvalue" for each type of asset (by type we mean, for example, "Room Furniture", "Radio", "Cell (mobile) Phone handset", etc.) to impute the missing values.

I created three variables ("median_durable", "median_animal", and "median_tool") which represented the median value of every single category. Median values are present according to their asset type. Please feel free to check the following outcomes.

```
assets <- read_dta(file.path(path, "assets.dta"))

assets <- assets %>%
  group_by(durablegood_code) %>%
  mutate(median_durable = median(currentvalue, na.rm = TRUE),
         median_durable = ifelse(is.na(durablegood_code) == TRUE, NA, median_durable)) %>%
  ungroup(durablegood_code) %>%
  group_by(animaltype) %>%
  mutate(median_animal = median(currentvalue, na.rm = TRUE),
         median_animal = ifelse(is.na(animaltype) == TRUE, NA, median_animal)) %>%
  ungroup(animaltype) %>%
  group_by(toolcode) %>%
  mutate(median_tool = median(currentvalue, na.rm = TRUE),
         median_tool = ifelse(is.na(toolcode) == TRUE, NA, median_tool))

assets %>%
  select(Asset_Type, median_durable, median_animal, median_tool)
```

```
## # A tibble: 164,693 x 5
## # Groups:   toolcode [51]
##        toolcode          Asset_Type median_durable median_animal median_tool
##       <dbl+lbl>           <dbl+lbl>          <dbl>         <dbl>       <dbl>
##  1 NA           1 [1-Animals]                 NA              60          NA
##  2  4 [Hoe]     2 [2-Tools]                   NA              NA           9
##  3 14 [Cutlass] 2 [2-Tools]                   NA              NA          10
##  4 NA           3 [3-Durable Goods]          200              NA          NA
##  5 NA           3 [3-Durable Goods]           20              NA          NA
##  6 NA           3 [3-Durable Goods]           80              NA          NA
##  7 NA           3 [3-Durable Goods]            3.5            NA          NA
##  8 NA           3 [3-Durable Goods]            7.5            NA          NA
##  9 NA           3 [3-Durable Goods]           20              NA          NA
## 10 NA           3 [3-Durable Goods]           20              NA          NA
## # ... with 164,683 more rows
```

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

Here I created three new variables ("total_durable", "total_animal", and "total_tool") that reflected the outcome of the asset's median times the quantity. After that, I added those values and stored them in the "total_value" variable. Please feel free to check the following outcomes.

```
assets_total <- assets %>%
  mutate(total_durable = quantity * median_durable,
         total_animal = quantity * median_animal,
         total_tool = quantity * median_tool)

assets_total$total_value <- rowSums(assets_total[, c("total_durable", "total_animal", "total_tool")], na
```

4. 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: household ID, wave ID, total value of animals, total value of tools, and total value of durable goods. Then, also create a total asset value variable.

I grouped the data by "hhid" and "wave" variables to have data sorted through households and waves. Then, I presented data regarding the total sum of durable, animal, and tool values. Finally, I added the three columns to have a general value variable (total_asset). Please feel free to check the following outcomes.

```
final_assets <- assets_total %>%
  group_by(hhid, wave) %>%
  summarize(sum_durable = sum(total_durable, na.rm = TRUE),
            sum_animal = sum(total_animal, na.rm = TRUE),
            sum_tool = sum(total_tool, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'hhid'. You can override using the
## `.groups` argument.
```

```
final_assets <- final_assets %>%
  mutate(total_asset = sum_durable + sum_animal + sum_tool,
         hhid = as.numeric(hhid))

final_assets %>% head(10)
```

```
## # A tibble: 10 x 6
## # Groups:   hhid [5]
##         hhid  wave sum_durable sum_animal sum_tool total_asset
##        <dbl> <dbl>       <dbl>      <dbl>    <dbl>       <dbl>
##  1 101001002     1         786         300       19        1105
##  2 101001002     2        3704         180       29        3913
##  3 101001003     1        1336.          0      168        1504.
##  4 101001003     2       25474           0     4075       29549
##  5 101001004     1        1146         600       53        1799
##  6 101001004     2       17106           0     3406       20512
##  7 101001009     1         758.          0       10         768.
##  8 101001009     2       20180.          0    12356.      32536
##  9 101001010     1        1173           0       19        1192
## 10 101001010     2       11000.          0       29       11028.
```

Mental health

5. A Kessler-10 scale is a measure of mental health that uses 10 questions that identify 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.

I created four variables to face NAs. (1) "pre-score": sums the Kessler sub scores without NAs. (2) "nas": counts how many NAs were in the row. (3) "add_weight": weightes the missing NAs according to the observed average. This variable is equal to pre_score / (10 - nas).

In that way, we are not dramatically affecting the final Kessler score and preserving its nature. Please feel free to check the following outcomes.

```
depression <- read_dta(file.path(path, "depression.dta"))

depression$pre_score <- rowSums(depression[, c("tired", "nervous", "sonervous", "hopeless", "restless",

depression$nas <- rowSums(is.na(depression))

final_depression <- depression %>%
  mutate(add_weight = if_else(nas > 0, pre_score/(10 - nas),0),
         kessler_score = pre_score + add_weight,
         kessles_categories = ifelse(kessler_score >= 10 & kessler_score <= 19,
                                     "no significant depression",
                                     ifelse(kessler_score >= 20 & kessler_score <= 24,
                                            "mild depression",
                                            ifelse(kessler_score >= 25 & kessler_score <= 29,
                                                   "moderate depression",
                                                   "severe depression")
                                            )
                                     )
)

final_depression %>%
  select(pre_score, add_weight, kessler_score, kessles_categories)
```

```
## # A tibble: 13,842 x 4
##    pre_score add_weight kessler_score kessles_categories
##        <dbl>      <dbl>         <dbl> <chr>
##  1        21          0            21 mild depression
##  2        30          0            30 severe depression
##  3        13          0            13 no significant depression
##  4        26          0            26 moderate depression
##  5        27          0            27 moderate depression
##  6        26          0            26 moderate depression
##  7        15          0            15 no significant depression
##  8        23          0            23 mild depression
##  9        10          0            10 no significant depression
## 10        10          0            10 no significant depression
## # ... with 13,832 more rows
```

Constructing a single dataset

6. 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).

I joined the data considering three main variables: "wave", "hhid", and "hhmid". Considering the RCT objective, I opted to join datasets so that household heads (i.e., husband and wife) were the only household members in the data. By including other family members, we could have measured spillover effects, but I'm assuming that is not the purpose of the experiment. The final dataset is saved as "data_rct_final.dta". Please feel free to check the following outcomes.

```
data_rct <- left_join(final_depression, final_demographics, by = c("wave", "hhid", "hhmid"))
data_rct_join <- left_join(data_rct, final_assets, by = c("wave", "hhid"))

data_rct_join %>% head(10)
```

```
## # A tibble: 10 x 39
##     wave      hhid hhmid   tired nervous soner~1 hopel~2 restl~3 sores~4 depre~5
##    <dbl>     <dbl> <dbl> <dbl+l> <dbl+l> <dbl+l> <dbl+l> <dbl+l> <dbl+l> <dbl+l>
##  1     1 101001002     1 2 [A l~ 2 [A l~ 2 [A l~ 2 [A l~ 2 [A l~ 1 [Non~ 1 [Non~
##  2     1 101001002     2 3 [Som~ 3 [Som~ 3 [Som~ 3 [Som~ 3 [Som~ 2 [A l~ 2 [A l~
##  3     2 101001002     1 2 [A l~ 1 [Non~ 1 [Non~ 2 [A l~ 1 [Non~ 1 [Non~ 2 [A l~
##  4     2 101001002     2 1 [Non~ 2 [A l~ 1 [Non~ 5 [All~ 3 [Som~ 1 [Non~ 3 [Som~
##  5     1 101001003     1 3 [Som~ 3 [Som~ 2 [A l~ 3 [Som~ 3 [Som~ 2 [A l~ 2 [A l~
##  6     2 101001003     1 3 [Som~ 3 [Som~ 1 [Non~ 3 [Som~ 3 [Som~ 3 [Som~ 3 [Som~
##  7     2 101001003     2 1 [Non~ 1 [Non~ 1 [Non~ 1 [Non~ 3 [Som~ 1 [Non~ 1 [Non~
##  8     1 101001004     1 3 [Som~ 2 [A l~ 2 [A l~ 2 [A l~ 2 [A l~ 2 [A l~ 1 [Non~
##  9     2 101001004     1 1 [Non~ 1 [Non~ 1 [Non~ 1 [Non~ 1 [Non~ 1 [Non~ 1 [Non~
## 10     2 101001004     2 1 [Non~ 1 [Non~ 1 [Non~ 1 [Non~ 1 [Non~ 1 [Non~ 1 [Non~
## # ... with 29 more variables: everythingeffort <dbl+lbl>,
## #   nothingcheerup <dbl+lbl>, worthless <dbl+lbl>, pre_score <dbl>, nas <dbl>,
## #   add_weight <dbl>, kessler_score <dbl>, kessles_categories <chr>,
## #   villageid <dbl>, treat_hh <dbl+lbl>, gender <dbl+lbl>, age <dbl>,
## #   relationship <dbl+lbl>, maritalstatus <dbl+lbl>, spouseinhouse <dbl+lbl>,
## #   agemarried <dbl>, religion <dbl+lbl>, religionother <chr>,
## #   fatherinhouse <dbl+lbl>, fathereduc <dbl+lbl>, fathereducother <chr>, ...
```

PART II

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:

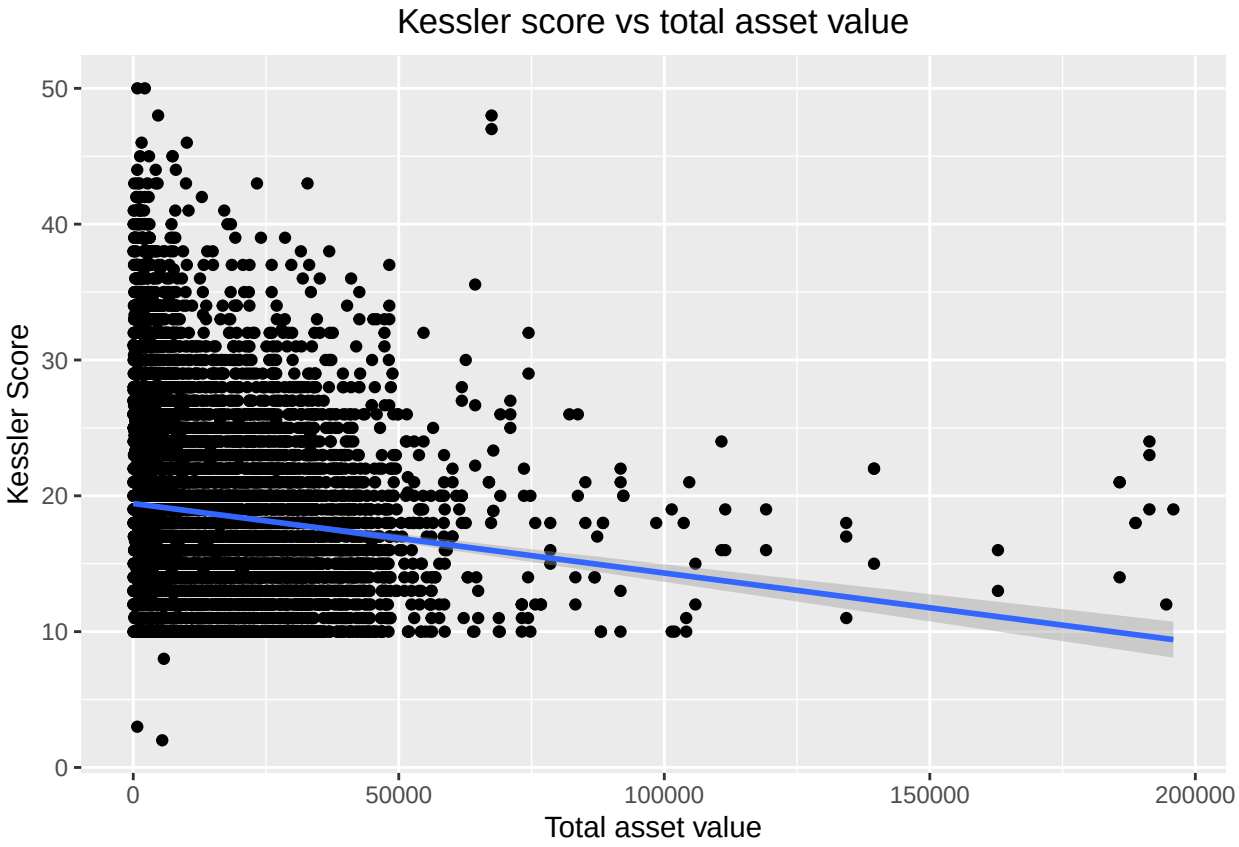
1. Explore the relationship between depression and household wealth, proxied by total asset value. Present the results from your exploration through tables, plots, a write-up, or anything else you think would be useful.

The scatterplot suggests a negative relationship between the total asset value and the Kessler score. Thus, we might expect that as income increases, depression levels decrease, and vice versa. This statement is confirmed by a linear regression, where we appreciate a 1% change in the asset value is associated with a 0.07 change in the Kessler score at a 5% level.

Nonetheless, the negative relationship between asset value and Kessler score is not preserved in all cases. When analyzing asset value per percentiles as clusters, we can observe a positive relationship in the highest ones, especially between 81th to 100th percentiles, who gather the highest asset value in the data.

```
data_rct_final <- data_rct_join  %>%
  mutate(percentile_20 = ntile(total_asset, 5),
         percentile_name = ifelse(percentile_20 == 1, "1st - 20th",
                                  ifelse(percentile_20 == 2, "21th - 40th",
                                         ifelse(percentile_20 == 3, "41th - 60th",
                                                ifelse(percentile_20 == 4, "61th - 80th",
                                                       ifelse(percentile_20 == 5, "81th - 100th",NA))
```

```
data_rct_final %>%
  filter(total_asset < 200000) %>%
  ggplot(aes(total_asset, kessler_score)) +
  geom_point() +
  geom_smooth(method="lm", formula = y ~ x)  +
  labs(title = "Kessler score vs total asset value",
       x = "Total asset value",
       y = "Kessler Score") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5),
        legend.position = c(0.93, 0.78))
```

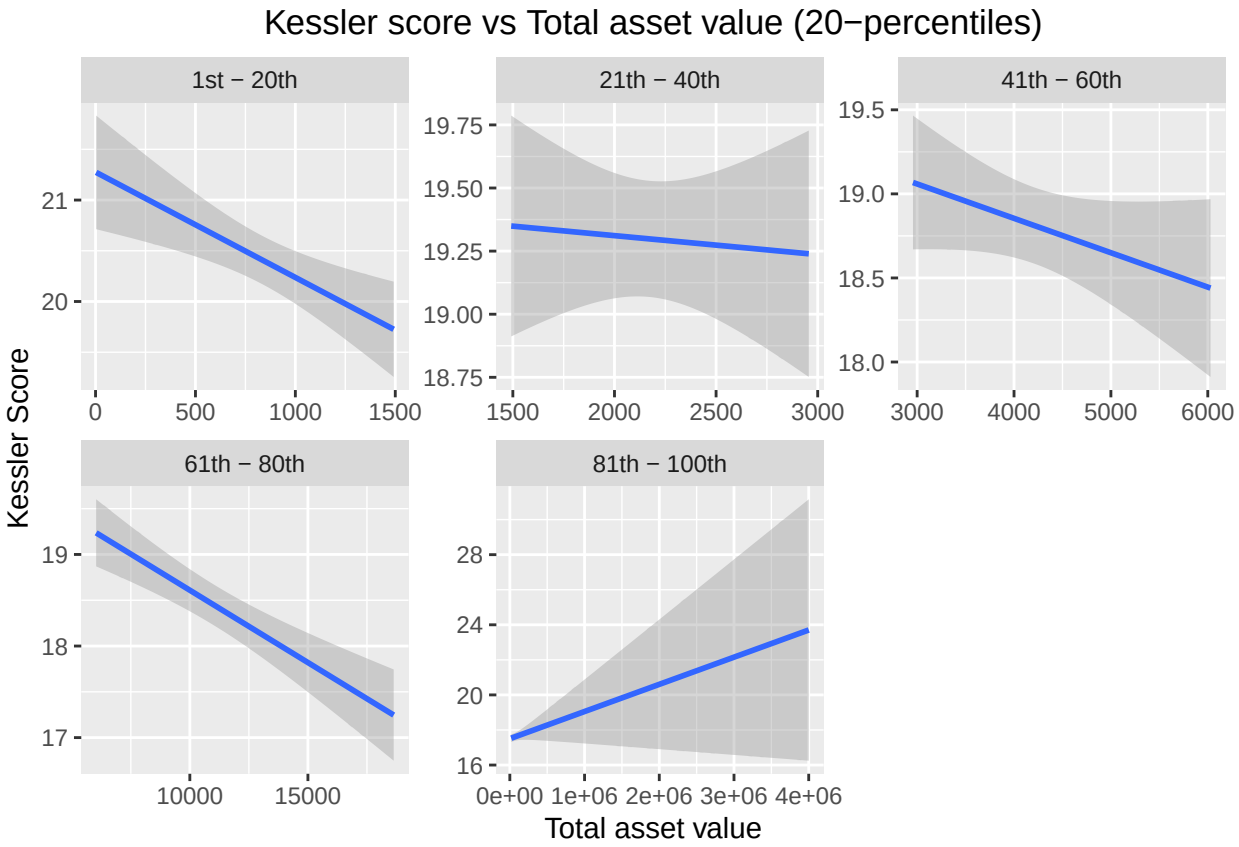



```
lm_1a <- lm_robust(data = data_rct_final, kessler_score ~ log(total_asset))
summary(lm_1a)
```

```
##
## Call:
## lm_robust(formula = kessler_score ~ log(total_asset), data = data_rct_final)
##
## Standard error type:  HC2
##
## Coefficients:
##                  Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)       24.9798    0.33398   74.79 0.000e+00  24.3251  25.6344 13798
## log(total_asset)  -0.7208    0.03868  -18.63 1.512e-76  -0.7966  -0.6449 13798
##
## Multiple R-squared:  0.02548 ,   Adjusted R-squared:  0.02541
## F-statistic: 347.2 on 1 and 13798 DF,  p-value: < 2.2e-16
```

```
data_rct_final %>%
  filter(!is.na(percentile_20)) %>%
  ggplot(aes(total_asset, kessler_score)) +
  geom_smooth(method="lm", formula = y ~ x)  +
  labs(title = "Kessler score vs Total asset value (20-percentiles)",
       x = "Total asset value",
       y = "Kessler Score") +
  theme(plot.title = element_text(hjust = 0.5),
```

```
        plot.caption = element_text(hjust = 0.5),
        legend.position = c(0.93, 0.78)) +
  facet_wrap(vars(percentile_name), scales = "free", strip.position = "top", nrow = 2)
```



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.

2. Were the GT sessions effective at reducing depression?

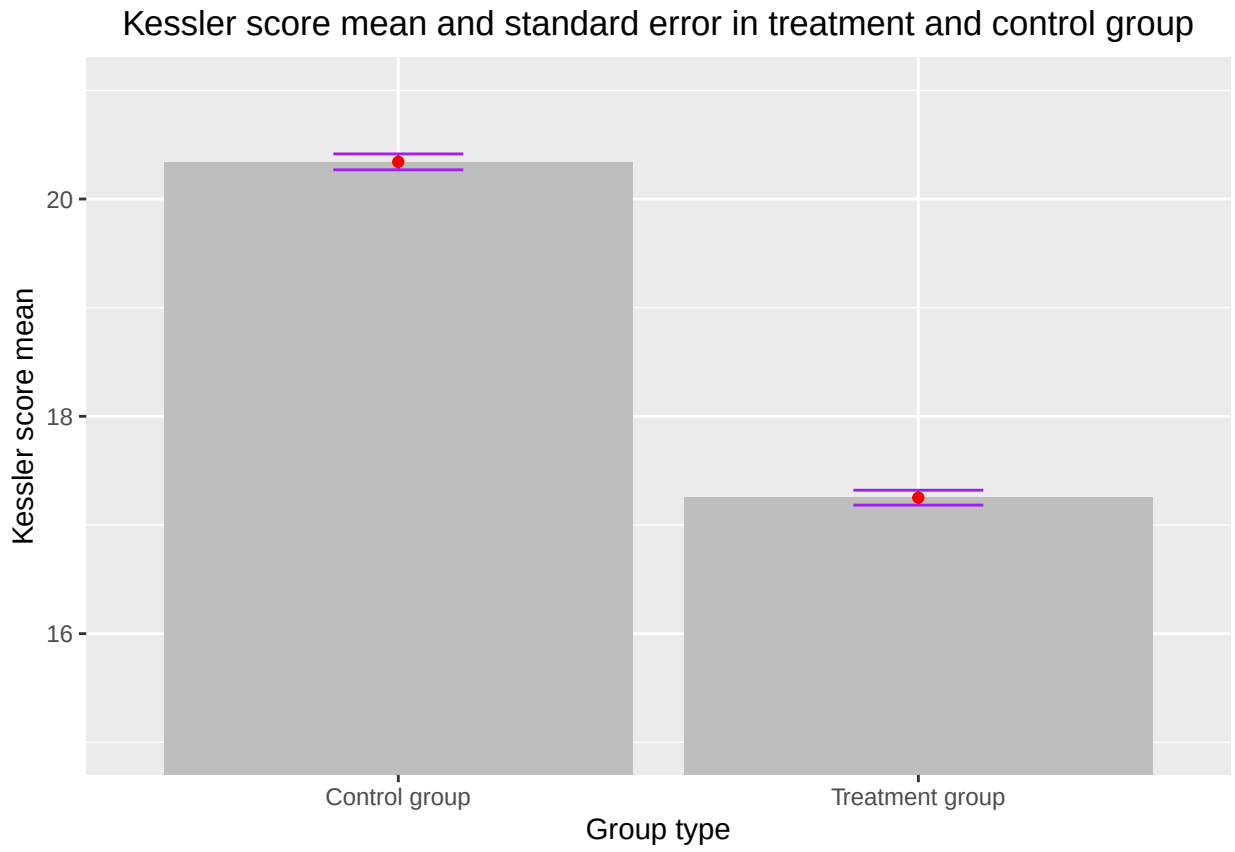
Assuming randomization, SUTVA, no attrition, and perfect compliance, I appreciate an economical and statistically significant (p-value < 2e-16) treatment effect. The OLS suggests that Group Therapy decreases by 3.09 points [CI:-3.287, -2.894] the Kessler depression score at a 5% level, which represents an improvement of 15% compared with the people who didn't receive the treatment.

```
data_rct_final <- data_rct_final %>%
  mutate(treatment = wave - 1)

lm_2a <- lm_robust(kessler_score ~ treatment, data = data_rct_final)
summary(lm_2a)
```

```
##
```

```
## Call:
## lm_robust(formula = kessler_score ~ treatment, data = data_rct_final)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)    20.34    0.07326   277.7  0.000e+00   20.198   20.485 13818
## treatment      -3.09    0.10033   -30.8 1.383e-201   -3.287   -2.894 13818
##
## Multiple R-squared:  0.06298 ,   Adjusted R-squared:  0.06291
## F-statistic: 948.9 on 1 and 13818 DF,  p-value: < 2.2e-16
```

```
data_rct_final %>%
  group_by(treatment) %>%
  summarise(mean = mean(kessler_score, na.rm = TRUE),
            se = sd(kessler_score, na.rm = TRUE)/sqrt(length(kessler_score))) %>%
  mutate(group = if_else(treatment == 1, "Treatment group", "Control group")) %>%
  ggplot(aes(x = group, y = mean)) +
  geom_bar(stat = "identity",
                   fill = "gray") +
  geom_errorbar(aes(ymax = mean + se,
                    ymin = mean - se),
                position = position_dodge(width=0.9),
                width = 0.25,
                color = "purple") +
  geom_point(color = "red") +
  coord_cartesian(ylim = c(15, 21)) +
  labs(title = "Kessler score mean and standard error in treatment and control group",
       x = "Group type",
       y = "Kessler score mean") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5),
        legend.position = c(0.93, 0.78))
```



3. 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. In your write-up please interpret the results. Present your results as you see fit and add a write-up with your interpretation and conclusions.

Yes, there is an heterogeneous treatment effect between men and women. On average, women who didn't participate in the Group Therapy have slightly higher depression score levels than men (20.70 vs. 20.08) at a 5% level. Nonetheless, therapies have a better impact on women than men. Men who receive treatment decrease their depression score by 14% (from 20.08 to 17.26), while women's depression decreases by 16.8% (from 20.70 to 17.24). In conclusion, on average, therapies reduce "mild" depression to "no significant" depression but have a higher impact on women.

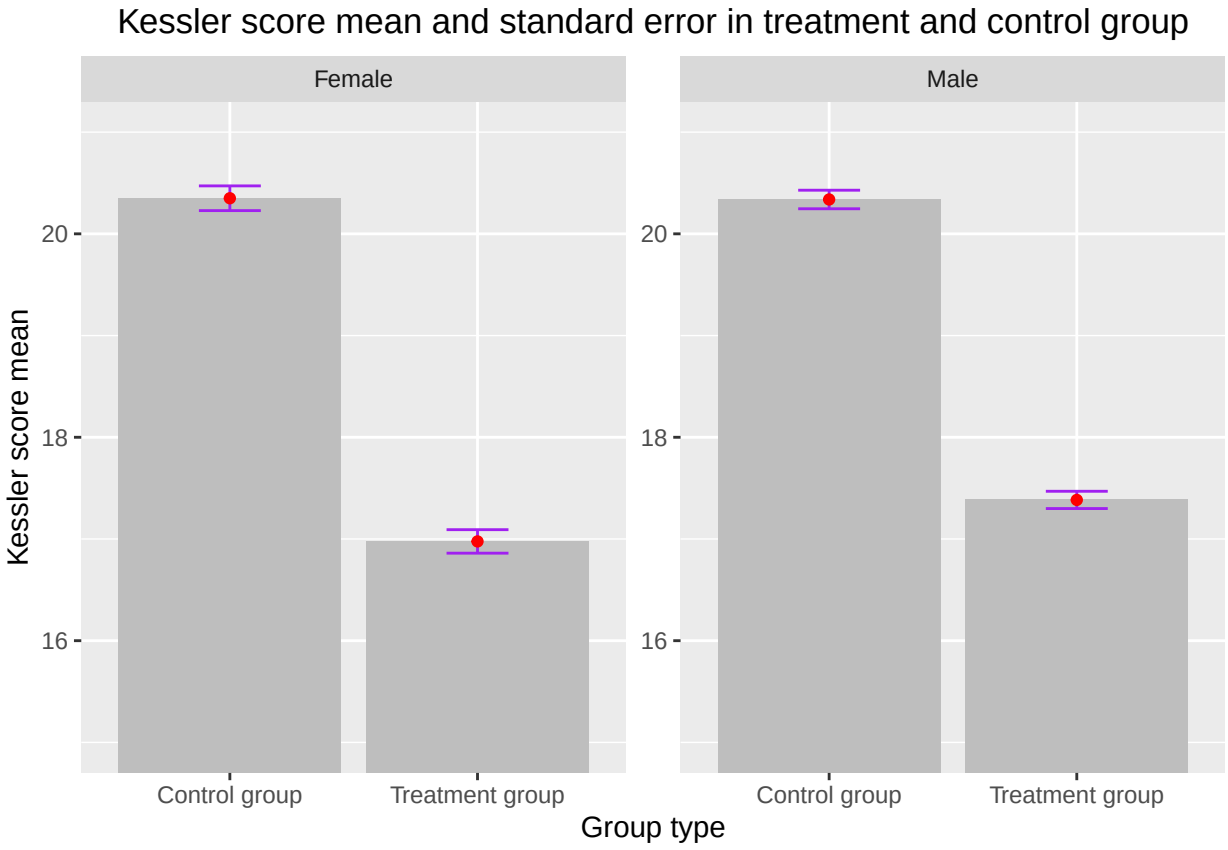
```
data_rct_final_2 <- data_rct_final %>%
  mutate(women = hhmid - 1,
         sex = if_else(women == 1, "Female", "Male"))

lm_3a <- lm_robust(kessler_score ~ treatment + women + treatment*women, data = data_rct_final_2)
summary(lm_3a)
```

```
##
## Call:
## lm_robust(formula = kessler_score ~ treatment + women + treatment *
##     women, data = data_rct_final_2)
##
```

```
## Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)      20.0846    0.08335 240.954  0.000e+00  19.9212  20.2480 13816
## treatment        -2.8200    0.11339 -24.869 1.352e-133  -3.0422  -2.5977 13816
## women             0.6138    0.08975   6.839  8.311e-12   0.4379   0.7897 13816
## treatment:women  -0.6380    0.10836  -5.888  4.008e-09  -0.8504  -0.4256 13816
##
## Multiple R-squared:  0.0664 ,    Adjusted R-squared:  0.0662
## F-statistic:   337 on 3 and 13816 DF,  p-value: < 2.2e-16
```

```
data_rct_final_2 %>%
  group_by(treatment, sex) %>%
  summarise(mean = mean(kessler_score, na.rm = TRUE),
            se = sd(kessler_score, na.rm = TRUE)/sqrt(length(kessler_score))) %>%
  mutate(group = if_else(treatment == 1, "Treatment group", "Control group")) %>%
  ggplot(aes(x = group, y = mean)) +
  geom_bar(stat = "identity",
                   fill = "gray") +
  geom_errorbar(aes(ymax = mean + se,
                    ymin = mean - se),
                position = position_dodge(width=0.9),
                width = 0.25,
                color = "purple") +
  geom_point(color = "red") +
  coord_cartesian(ylim = c(15, 21)) +
  labs(title = "Kessler score mean and standard error in treatment and control group",
       x = "Group type",
       y = "Kessler score mean") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5),
        legend.position = c(0.93, 0.78)) +
   facet_wrap(vars(sex), scales = "free", strip.position = "top")
```

```
## `summarise()` has grouped output by 'treatment'. You can override using the
## `.groups` argument.
```


Kessler score mean and standard error in treatment and control group
Female
Male
Kessler score mean
20
18
16
Control group
Treatment group
Group type