# RohiniVenkitaramanIyer_Rcode.R

rohiniiyer

2024-03-31

```r
setwd("~/OneDrive - Harvard University/Resume/International Organizations/World Bank/McNamara/Rohini_Ven

if (!requireNamespace("haven", quietly = TRUE)) {
  install.packages("haven")
}
if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}
if (!requireNamespace("labelled", quietly = TRUE)) {
  install.packages("labelled")
}
if (!requireNamespace("tidyr", quietly = TRUE)) {
  install.packages("tidyr")
}
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}
if (!requireNamespace("estimatr", quietly = TRUE)) {
  install.packages("estimatr")
}

library(haven)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(labelled)
library(tidyr)
library(ggplot2)
library(estimatr)
```

```r
GEM_Baseline <- read_dta("data/GEM_Baseline.dta")
GEM_Treatment_Status <- read_dta("data/GEM_Treatment_Status.dta")

############################ Task 1 - Data cleaning #########################

# summary(GEM_Baseline)
# Commenting this function out so it doesn't print in the final output.
# But I used this function to get a sense of the data.

# Converting character variables that have only numbers into numeric
char <- c("HHID", "q122_father_attend_school", "w20_hours_unpaid_job99",
          "s02_cash_savings")
GEM_Baseline[char] <- lapply(GEM_Baseline[char], as.numeric)
```

```
## Warning in lapply(GEM_Baseline[char], as.numeric): NAs introduced by coercion

## Warning in lapply(GEM_Baseline[char], as.numeric): NAs introduced by coercion

## Warning in lapply(GEM_Baseline[char], as.numeric): NAs introduced by coercion
```

```r
# HHID has 3 NAs, dropping these observations
GEM_Baseline <- GEM_Baseline[!is.na(GEM_Baseline$HHID), ]

# Detecting duplicates
GEM_Baseline$is_duplicate <- duplicated(GEM_Baseline) |
  duplicated(GEM_Baseline, fromLast = TRUE)
GEM_Baseline$is_duplicate_HHID <- duplicated(GEM_Baseline$HHID) |
  duplicated(GEM_Baseline$HHID, fromLast = TRUE)

table(GEM_Baseline$is_duplicate, useNA = "always")
```

```
##
## FALSE  TRUE  <NA>
##  1286     6     0
```

```r
table(GEM_Baseline$is_duplicate_HHID, useNA = "always")
```

```
##
## FALSE  TRUE  <NA>
##  1286     6     0
```

```r
# This indicates that the duplicates have the same data throughout and
# so 1 of the entries of each pair can be dropped
GEM_Baseline <- unique(GEM_Baseline)

GEM_Baseline <- GEM_Baseline %>%
  select(-c("is_duplicate", "is_duplicate_HHID"))

table(GEM_Baseline$w02_paid_in_cash_job1, useNA = "always")
```

```
##
```

```
##            .  FOOD HAWKER          H11          H12          H13          H14
##            2            1          160            6            8            3
##          H15  HOT WAITERS          N31          N32          N33          N34
##            3            1            8           13           15            5
##          N35          N37 No Paid Work          P44          P45          P47
##           34           21          789            5            2            4
##          P52          P53          R11          R21          R22          R23
##            1            3            1            3           53           26
##          R24          R25          R26          R27          R28          R29
##            2           26           21            9           34           30
##         <NA>
##            0
```

```
GEM_Baseline <- GEM_Baseline %>%
  mutate(w02_paid_in_cash_job1 = replace(w02_paid_in_cash_job1,
                                         w02_paid_in_cash_job1 == "FOOD HAWKER",
                                         "N33"),
         w02_paid_in_cash_job1 = replace(w02_paid_in_cash_job1,
                                         w02_paid_in_cash_job1 == "HOT WAITERS",
                                         "R23"),
         w02_paid_in_cash_job3 = replace(w02_paid_in_cash_job3,
                                         w02_paid_in_cash_job3 == "..", "."))

# Merging baseline data with treatment status data
merged <- merge(GEM_Baseline, GEM_Treatment_Status, by="HHID", all.x=TRUE)
merged <- merged %>%
  set_variable_labels(HHID = "Household ID", treatment = "Treatment Status")

############################# Task 2A - Table #############################

# Converting savings from Kenyan Shillings to US Dollars
merged <- merged %>%
  mutate(cash_savings_usd = s02_cash_savings/135,
         jewellery_savings_usd = s04_jewellery_savings_value/135)

merged$total_savings_usd <- rowSums(merged[, c("cash_savings_usd",
                                               "jewellery_savings_usd")],
                                    na.rm = TRUE)

merged <- merged %>%
  mutate(total_savings_usd = ifelse(is.na(cash_savings_usd) &
                                      is.na(jewellery_savings_usd),
                                    NA, total_savings_usd))
merged <- merged %>%
  set_variable_labels(cash_savings_usd = "Cash savings (in USD)",
                      jewellery_savings_usd = "Jewellery savings (in USD)",
                      total_savings_usd = "Total savings (in USD)")

# Summary statistics for savings
get_var_label <- function(var_name, data) {
  var_label <- attr(data[[var_name]], "label")
  return(var_label)
}
```

```r
summary_stats <- function(data, variables) {
  stats <- sapply(data[variables], function(x) {
    c(
      "Number of households" = sum(!is.na(x)),
      Mean = round(mean(x, na.rm = TRUE), 2),
      Median = round(median(x, na.rm = TRUE), 2),
      SD = round(sd(x, na.rm = TRUE), 2),
      Min = round(min(x, na.rm = TRUE), 2),
      Max = round(max(x, na.rm = TRUE), 2)
    )
  })
  return(stats)
}

summary_table <- summary_stats(merged,
                               c("cash_savings_usd", "jewellery_savings_usd",
                                 "total_savings_usd"))
summary_df <- as.data.frame(t(summary_table))
rownames(summary_df) <- sapply(c("cash_savings_usd", "jewellery_savings_usd",
                                 "total_savings_usd"),
                               get_var_label, data = merged)
print(summary_df)
```

```
##                             Number of households   Mean Median       SD     Min
## Cash savings (in USD)                        840  52.35  41.52   510.68  -59.26
## Jewellery savings (in USD)                   640 423.27  18.53 10247.42 -222.22
## Total savings (in USD)                      1288 244.46  29.48  7234.74 -218.52
##                                  Max
## Cash savings (in USD)       14814.81
## Jewellery savings (in USD) 259259.26
## Total savings (in USD)     259265.93
```

```r
# Number of households represents the NON-MISSING observations for each variable
write.csv(summary_df, "output/savings.csv")

# These descriptive statistics indicate the presence of
# (1) negative values (which doesn't make sense wrt savings) and
# (2) extreme values that impact the mean (which is highly different from the
# median in all 3 cases).
# While considering total savings instead of cash savings or jewellery savings
# individually helps us reduce the number of missing values in the savings variables,
# the wide range of the values for the total_savings variable could have
# implications while performing regression analyis below.

############################# Task 2B - Graph ############################

# 1. Creating a job type level dataset

merged <- merged %>%
  rename(w02_paid_in_cash_job4 = w16_unpaid_job99)

keep <- c("HHID", "treatment", "w02_paid_in_cash_job1", "w02_paid_in_cash_job2",
          "w02_paid_in_cash_job3", "w02_paid_in_cash_job4", "w07_hours_job1",
```

```r
          "w07_hours_job2", "w07_hours_job3", "w20_hours_unpaid_job99")
graph <- subset(merged, select=keep)

# Reshaping data
graph_long <- pivot_longer(graph, cols = starts_with("w02_paid_in_cash_job"),
                           names_to = "job", values_to = "job_type")

graph_long <- graph_long %>%
  mutate(hours = case_when(
    job == "w02_paid_in_cash_job1" ~ w07_hours_job1,
    job == "w02_paid_in_cash_job2" ~ w07_hours_job2,
    job == "w02_paid_in_cash_job3" ~ w07_hours_job3,
    job == "w02_paid_in_cash_job4" ~ w20_hours_unpaid_job99,
    TRUE ~ NA_real_
  ))

graph_long <- graph_long %>%
  select(-c("w07_hours_job1", "w07_hours_job2", "w07_hours_job3",
            "w20_hours_unpaid_job99"))

graph_long <- graph_long %>%
  mutate(jobn = case_when(
    job == "w02_paid_in_cash_job1" ~ "Paid job 1",
    job == "w02_paid_in_cash_job2" ~ "Paid job 2",
    job == "w02_paid_in_cash_job3" ~ "Paid job 3",
    job == "w02_paid_in_cash_job4" ~ "Unpaid work"))

graph_long <- graph_long %>%
  select(-"job")
graph_long <- graph_long %>%
  rename(job = jobn, work_code = job_type)

graph_long <- graph_long[, c("HHID", "job", "work_code", "hours", "treatment")]
haven::write_dta(graph_long, "output/GEM_job_type_R.dta")

# 2. Creating the graph

# Consolidating work codes
table(graph_long$work_code, useNA = "always")
```

```
##
##                            . 1 ,2 ,3, 4 ,7       1 2 3 4   1 2 3 4 6 7
##            16            81             1             3             2
##     1 2 3 4 7   1 2 3 4 7 8     1 2 3 6 7   1 2 3 6 7 8       1 2 3 7
##            18             2             2             1            17
##     1 2 3 7 8       1 2 4 7  1, 2 ,3 ,4,7    1, 2, 3, 7           1,2
##             1             1             1             1             1
##         1,2,3       1,2,3,4 1,2,3,4,5,6,7   1,2,3,4,5,7   1,2,3,4,6,7
##             3             4             2             3             9
##     1,2,3,4,7   1,2,3,4,7,8   1,2,3,5,6,7     1,2,3,5,7     1,2,3,6,7
##           166             5             1             4            11
##   1,2,3,6,7,8    1,2,3,6,7.       1,2,3,7     1,2,3,7,8       1,2,4,7
##             1             1           977             5             2
```

```
##         1,3,4     1,3,4,7,8         1,3,7       1,3,7,8           1,7
##             1             1             3             1             1
##             2           2 3       2,3,4,7     2,3,6,7,8         2,3,7
##             1             1             1             1             6
##       2,3,7,8           2,7             3         3 7 8           3,7
##             1             1             1             1             1
##         6 7 8             7           H11           H12           H13
##             1             5           319             7             8
##           H14           H15           N31           N32           N33
##             3             3             8            14            16
##           N34           N35           N37  No Paid Work           P44
##             5            35           176          2638             5
##           P45           P47           P52           P53           R11
##             2             4             1           167             1
##           R21           R22           R23           R24           R25
##             3            53            28             2            27
##           R26           R27           R28           R29          <NA>
##            21           173            36            31             0
```

```
graph_long <- graph_long %>%
  mutate(work_code = case_when(
    work_code == "" | work_code == "." ~ NA_character_,
    work_code == "No Paid Work" ~ "Reported as no paid work under paid jobs",
    substr(work_code, 1, 1) == "H" ~ "Household Services and Cleaning",
    substr(work_code, 1, 1) == "N" ~ "Nonformal and Other",
    substr(work_code, 1, 1) == "P" ~ "Professionals",
    substr(work_code, 1, 1) == "R" ~ "Retail, Food, Service",
    TRUE ~ "Unpaid work"
  ))
table(graph_long$work_code, useNA = "always")
```

```
##
##           Household Services and Cleaning
##                                       340
##                       Nonformal and Other
##                                       254
##                             Professionals
##                                       179
## Reported as no paid work under paid jobs
##                                      2638
##                     Retail, Food, Service
##                                       375
##                               Unpaid work
##                                      1273
##                                      <NA>
##                                        97
```

```
graph_long$hours <- ifelse(graph_long$hours < 0, NA, graph_long$hours)

graph_long <- graph_long %>%
  mutate(treat_hours = case_when(treatment == 1 ~ hours),
         control_hours = case_when(treatment == 0 ~ hours))
graph_long$treatment <- as.factor(graph_long$treatment)
```

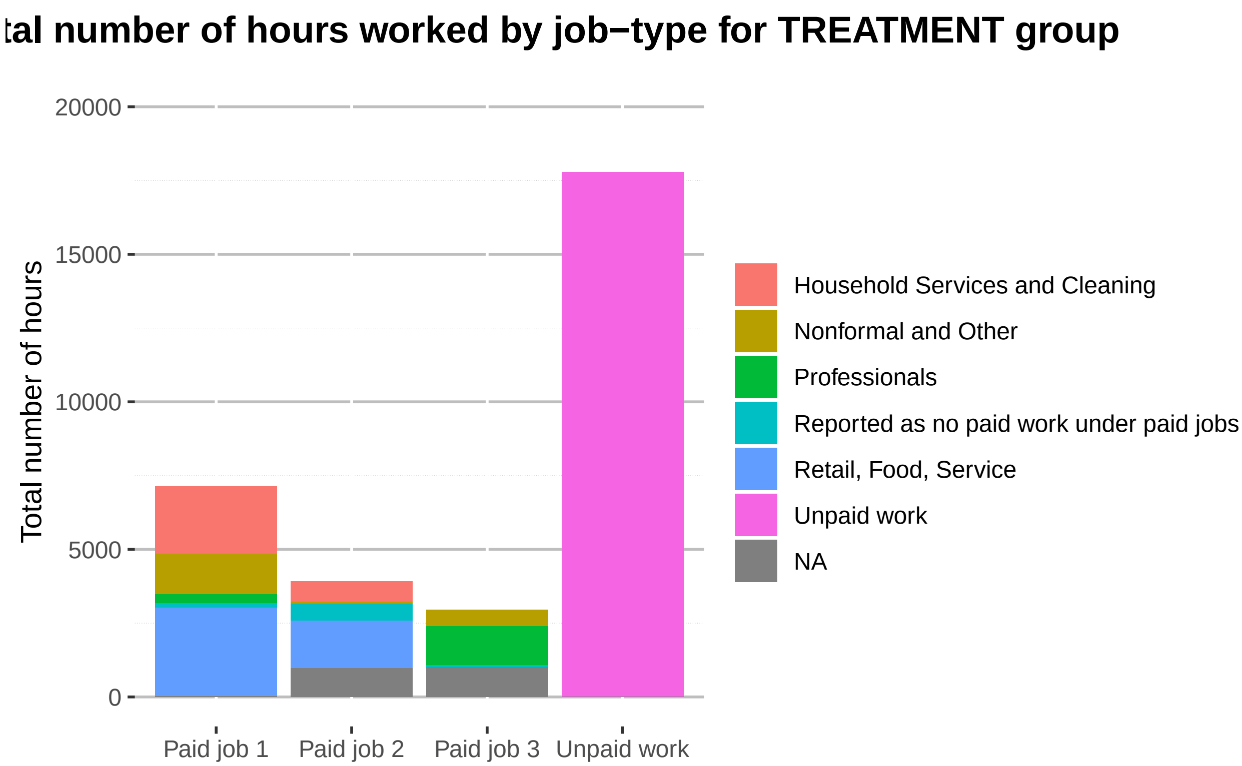
```r
treat <- graph_long %>%
  filter(treatment == 1) %>%
  group_by(job, work_code) %>%
  summarise(total_hours = sum(treat_hours, na.rm = TRUE))
```

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

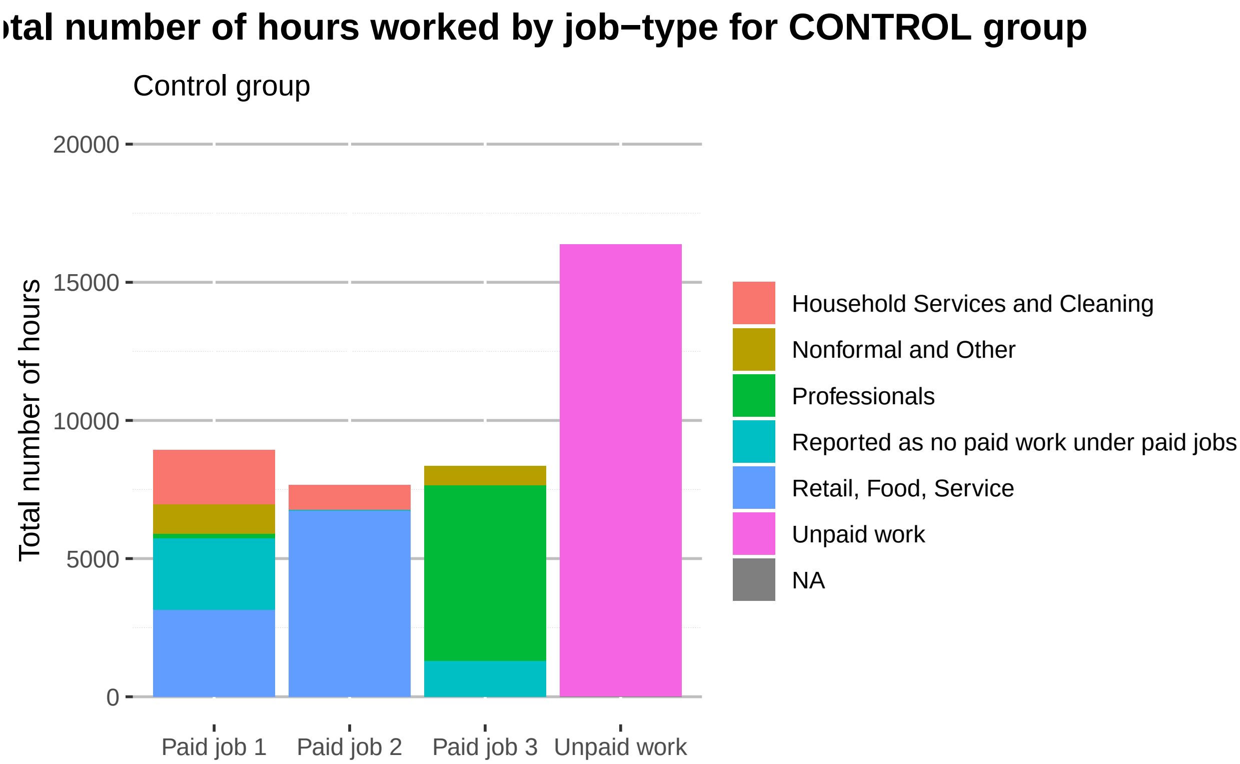
```r
control <- graph_long %>%
  filter(treatment == 0) %>%
  group_by(job, work_code) %>%
  summarise(total_hours = sum(control_hours, na.rm = TRUE))
```

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

```r
graph1_treatment <- ggplot(treat, aes(x = job, y = total_hours,
                                      fill = work_code)) +
  geom_bar(stat = "identity", position = "stack") +
  theme(panel.grid.major.y = element_line(color = "gray", linewidth = 0.5),
        panel.grid.minor.y = element_line(color = "gray", linewidth = 0.1,
                                          linetype = "dotted"),
        panel.background = element_rect(fill = "white"),
        axis.text.x = element_text(angle = 0, vjust = 1, hjust = 0.5),
        plot.title = element_text(face = "bold", margin = margin(t = 20, b = 10),
                                  size = 14, hjust = 0.3),
        legend.position = "right",
        legend.margin = margin(t = 0, r = 0, b = 0, l = 0)) +
  scale_x_discrete(labels = function(x) stringr::str_wrap(x, width = 15)) +
  labs(title = "Total number of hours worked by job-type for TREATMENT group",
       x = "",
       y = "Total number of hours",
       fill = "") +
  coord_cartesian(ylim = c(0, 20000))

print(graph1_treatment)
```




```
ggsave("output/graph1_treatment.png", plot = graph1_treatment, width = 8,
       height = 6, dpi = 300)

graph2_control <- ggplot(control, aes(x = job, y = total_hours,
                                      fill = work_code)) +
  geom_bar(stat = "identity", position = "stack") +
  theme(panel.grid.major.y = element_line(color = "gray", linewidth = 0.5),
        panel.grid.minor.y = element_line(color = "gray", linewidth = 0.1,
                                          linetype = "dotted"),
        panel.background = element_rect(fill = "white"),
        axis.text.x = element_text(angle = 0, vjust = 1, hjust = 0.5),
        plot.title = element_text(face = "bold", margin = margin(t = 20, b = 10),
                                  size = 14, hjust = 0.3),
        legend.position = "right",
        legend.margin = margin(t = 0, r = 0, b = 0, l = 0)) +
  scale_x_discrete(labels = function(x) stringr::str_wrap(x, width = 15)) +
  labs(title = "Total number of hours worked by job-type for CONTROL group",
       subtitle = "Control group",
       x = "",
       y = "Total number of hours",
       fill = "") +
  coord_cartesian(ylim = c(0, 20000))

print(graph2_control)
```




```r
ggsave("output/graph2_control.png", plot = graph2_control, width = 8,
       height = 6, dpi = 300)

# Note - Paid jobs 1, 2, and 3 report data of most recent jobs held since
# beginning of 2012. However, unpaid work is reported for the past 7 days before
# the date of the survey.
# From the graphs, 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.
# Moreover, inferring from the shading based on work code, women in the control
# group devote more hours to "retail, food, and service" and "professional" jobs
# as their 2nd and 3rd paid jobs respectively.
# The distribution within paid job 1 is similar for treatment and control groups,
# except that women report more unpaid hours even within this category in the
# control group.
# Unpaid hours in the last 7 days ranks highest and almost similar across both groups.

############################## Task 3 - Regression #############################

# Creating a variable that indicates total number of people living in a household
merged$hh_members <- rowSums(merged[, c("q130_a_husband", "q130_b_boyfriend",
                                        "q130_c_father", "q130_d_mother",
                                        "q130_e_stepfather", "q130_f_stepmother",
                                        "q130_g_father_in_law", "q130_h_mother_in_law",
                                        "q130_i_own_children", "q130_j_grandparents",
                                        "q130_k_brothers",
```

```r
                                        "q130_l_sisters")], na.rm = TRUE)

# Creating a variable that indicates total hours of paid labor from all 3 paid jobs
merged$total_hours_paid <- rowSums(merged[, c("w07_hours_job1", "w07_hours_job2",
                                              "w07_hours_job3")],
                                   na.rm = TRUE)

reg1 <- lm_robust(total_savings_usd ~ q102_age + q105_attend_school +
                    q120_a_vocational_training + q134_i_water_piped +
                    q134_a_electricity + q134_c_television +
                    q134_l_sewing_machine + w02_paid_in_cash +
                    w20_hours_unpaid_job99 + m901_b_currently_married +
                    m912_a_spouse_attend_school + m912_spouse_years_education +
                    treatment + total_hours_paid + hh_members, data = merged,
                  clusters = HHID, se_type = "stata")
summary(reg1)
```

```
##
## Call:
## lm_robust(formula = total_savings_usd ~ q102_age + q105_attend_school +
##     q120_a_vocational_training + q134_i_water_piped + q134_a_electricity +
##     q134_c_television + q134_l_sewing_machine + w02_paid_in_cash +
##     w20_hours_unpaid_job99 + m901_b_currently_married + m912_a_spouse_attend_school +
##     m912_spouse_years_education + treatment + total_hours_paid +
##
## Standard error type:  stata
##
## Coefficients:
##                              Estimate Std. Error  t value  Pr(>|t|)   CI Lower
## (Intercept)                 -79.78372   52.92016 -1.50762 0.1340403 -184.46503
## q102_age                      3.99508    2.38777  1.67315 0.0966667   -0.72815
## q105_attend_school           29.71466   11.02876  2.69429 0.0079708    7.89868
## q120_a_vocational_training   -5.14332    6.27069 -0.82022 0.4135717  -17.54736
## q134_i_water_piped            4.91147    4.62586  1.06174 0.2902914   -4.23894
## q134_a_electricity           -5.02022    5.59535 -0.89721 0.3712391  -16.08837
## q134_c_television            -4.76332    4.61578 -1.03196 0.3039760  -13.89378
## q134_l_sewing_machine        13.55456    5.75994  2.35325 0.0200870    2.16083
## w02_paid_in_cash             -0.96438    5.39539 -0.17874 0.8584145  -11.63701
## w20_hours_unpaid_job99        0.03661    0.14900  0.24573 0.8062736   -0.25812
## m901_b_currently_married      2.20060    4.76129  0.46219 0.6447093   -7.21770
## m912_a_spouse_attend_school   0.31028    0.08390  3.69841 0.0003172    0.14433
## m912_spouse_years_education   0.03589    0.04703  0.76309 0.4467698   -0.05714
## treatment                    -2.36884    4.03501 -0.58707 0.5581595  -10.35049
## total_hours_paid              0.03000    0.04985  0.60184 0.5483106   -0.06861
## hh_members                    0.05745    1.35395  0.04243 0.9662178   -2.62080
##                             CI Upper  DF
## (Intercept)                  24.8976 132
## q102_age                      8.7183 132
## q105_attend_school           51.5306 132
## q120_a_vocational_training    7.2607 132
## q134_i_water_piped           14.0619 132
## q134_a_electricity            6.0479 132
## q134_c_television             4.3671 132
```

```
## q134_l_sewing_machine        24.9483 132
## w02_paid_in_cash              9.7082 132
## w20_hours_unpaid_job99        0.3313 132
## m901_b_currently_married     11.6189 132
## m912_a_spouse_attend_school   0.4762 132
## m912_spouse_years_education   0.1289 132
## treatment                     5.6128 132
## total_hours_paid              0.1286 132
## hh_members                    2.7357 132
##
## Multiple R-squared:  0.06647 ,   Adjusted R-squared:  -0.05321
## F-statistic:  1405 on 15 and 132 DF,  p-value: < 2.2e-16
```

```r
summary_output <- capture.output(summary(reg1))
write.table(summary_output, "output/regression_summary.txt")

# The above regression uses total savings in USD as the outcome variable and factors
# like age of the respondent, whether they attended school or received vocation
# training (before this intervention), their household characteristics like having
# piped water, and assets like sewing machine, electricity, television, and how
# many household members (eating from the same pot), whether the women received
# cash/kind payment for any work they did and the total no. of paid and
# unpaid hours of labor they engaged in, their current marital status and their
# spouses' education and their treatment status in the experiment as
# independent variables to understand what affects household savings.

# Based on the above 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 i < 0.1).
# The regression is also clustered at the household level to account for within
# household correlation. I have used se.type=stata to adjust for
# heteroscedasticity and potential serial correlation in the errors.

# The fact that hours of paid labor, or household members, or hours of unpaid labor
# are not significant predictors of household savings seemed counter-intuitive.
# Hence, I regress each variable individually below to check their impact of
# household savings with standard errors clustered at the household level.

dep_var <- c("q102_age", "q105_attend_school", "q120_a_vocational_training",
             "q131_residence", "q134_a_electricity", "q134_c_television",
             "q134_i_water_piped", "q134_l_sewing_machine", "w02_paid_in_cash",
             "w20_hours_unpaid_job99", "m901_b_currently_married",
             "m912_a_spouse_attend_school", "treatment", "hh_members",
             "total_hours_paid")

for (var in dep_var) {
  formula <- as.formula(paste("total_savings_usd ~ ", var))
  reg2 <- lm_robust(formula, data = merged, cluster = HHID)
  print(summary(reg2))
}
```

```
##
## Call:
```

```
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)    -4704     4660.7  -1.009   0.3135 -13867.1   4459.1 390.6
## q102_age         270      265.2   1.018   0.3094   -251.4    791.4 410.4
##
## Multiple R-squared:  0.001764 ,  Adjusted R-squared:  0.0009864
## F-statistic: 1.036 on 1 and 1284 DF,  p-value: 0.309
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)           32.62      12.28   2.657   0.1172    -20.2    85.44 2.00
## q105_attend_school   217.28     207.09   1.049   0.4037   -669.7  1104.26 2.01
##
## Multiple R-squared:  2.098e-06 , Adjusted R-squared:  -0.0007934
## F-statistic: 1.101 on 1 and 1258 DF,  p-value: 0.2943
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|) CI Lower
## (Intercept)                  267.6      223.6   1.197   0.2316   -171.1
## q120_a_vocational_training  -234.5      223.6  -1.048   0.2961   -676.5
##                           CI Upper     DF
## (Intercept)                  706.4 1160.0
## q120_a_vocational_training   207.5  147.6
##
## Multiple R-squared:  9.012e-05 , Adjusted R-squared:  -0.0006905
## F-statistic: 1.099 on 1 and 1282 DF,  p-value: 0.2946
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)       -69.7      103.6 -0.6725   0.5029   -275.5    136.1 93.53
## q131_residence    162.9      155.6  1.0464   0.2980   -146.1    471.8 95.62
##
## Multiple R-squared:  4.635e-05 , Adjusted R-squared:  -0.0007318
## F-statistic: 1.095 on 1 and 1286 DF,  p-value: 0.2956
##
```

```
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                     Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper
## (Intercept)            30.75      1.277   24.08 2.462e-74    28.24    33.26
## q134_a_electricity   288.25    271.878    1.06 2.895e-01  -245.77   822.26
##                        DF
## (Intercept)        326.0
## q134_a_electricity 564.8
##
## Multiple R-squared:  0.0003004 , Adjusted R-squared:  -0.0004806
## F-statistic: 1.124 on 1 and 1281 DF,  p-value: 0.2893
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)          484.6      429.9   1.127   0.2601   -359.6   1328.8  603
## q134_c_television   -452.1      429.9  -1.052   0.2931  -1295.5    391.2 1266
##
## Multiple R-squared:  0.0009734 , Adjusted R-squared:  0.0001966
## F-statistic: 1.106 on 1 and 1287 DF,  p-value: 0.2931
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                     Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper
## (Intercept)            30.85      1.043   29.56 9.260e-122    28.80    32.89
## q134_i_water_piped    24.80     23.177    1.07  2.849e-01   -20.67    70.26
##                      DF
## (Intercept)         640
## q134_i_water_piped 1278
##
## Multiple R-squared:  0.0008975 , Adjusted R-squared:  0.0001158
## F-statistic: 1.145 on 1 and 1279 DF,  p-value: 0.2849
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)               256.7      213.2   1.204   0.2287   -161.5    675.0
## q134_l_sewing_machine   -225.7      213.2  -1.059   0.2931   -650.2    198.8
```

```
##                              DF
## (Intercept)           1217.00
## q134_l_sewing_machine   77.14
##
## Multiple R-squared:  5.006e-05 , Adjusted R-squared:  -0.0007275
## F-statistic: 1.121 on 1 and 1287 DF,  p-value: 0.29
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                  Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)         30.27      1.862  16.259 7.891e-35    26.59    33.95 149.0
## w02_paid_in_cash   521.12    519.512   1.003 3.168e-01  -502.15  1544.39 245.4
##
## Multiple R-squared:  0.0004668 , Adjusted R-squared:  -0.001078
## F-statistic: 1.006 on 1 and 648 DF,  p-value: 0.3162
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)            260.3319   218.6431  1.1907   0.2388 -177.581  698.245
## w20_hours_unpaid_job99  -0.3862     0.5054 -0.7641   0.5124   -2.238    1.466
##                            DF
## (Intercept)            56.476
## w20_hours_unpaid_job99  2.417
##
## Multiple R-squared:  7.089e-06 , Adjusted R-squared:  -0.000791
## F-statistic: 0.5838 on 1 and 1254 DF,  p-value: 0.445
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)                 33.07      13.81   2.395  0.02065    5.295    60.85
## m901_b_currently_married    81.90      76.77   1.067  0.29174  -72.730   236.54
##                             DF
## (Intercept)              46.92
## m901_b_currently_married 44.93
##
## Multiple R-squared:  0.001375 ,  Adjusted R-squared:  -0.003206
## F-statistic: 1.138 on 1 and 219 DF,  p-value: 0.2872
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
```

```
##
## Standard error type:  CR2
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|) CI Lower
## (Intercept)                 86.1169     56.588  1.5218   0.1295   -25.40
## m912_a_spouse_attend_school  -0.4437      1.206 -0.3679   0.7752   -15.49
##                            CI Upper      DF
## (Intercept)                   197.6 222.329
## m912_a_spouse_attend_school     14.6   1.008
##
## Multiple R-squared:  8.471e-06 , Adjusted R-squared:  -0.003794
## F-statistic: 0.1353 on 1 and 264 DF,  p-value: 0.7133
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)    55.25      23.52  2.3488  0.01914    9.059    101.4  628
## treatment     369.80     394.08  0.9384  0.34821 -403.306   1142.9 1283
##
## Multiple R-squared:  0.0006533 , Adjusted R-squared:  -0.0001238
## F-statistic: 0.8806 on 1 and 1287 DF,  p-value: 0.3482
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)    574.1      531.0  1.0811   0.2800   -468.5   1616.6 672.5
## hh_members    -114.5      114.6 -0.9994   0.3184   -340.0    110.9 322.1
##
## Multiple R-squared:  0.001268 ,  Adjusted R-squared:  0.0004911
## F-statistic: 0.9988 on 1 and 1287 DF,  p-value: 0.3178
##
## Call:
## lm_robust(formula = formula, data = merged, clusters = HHID)
##
## Standard error type:  CR2
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)  CI Lower CI Upper     DF
## (Intercept)      241.4085   196.9755  1.2256   0.2207 -145.2834 628.1005 746.39
## total_hours_paid   0.1015     0.1969  0.5157   0.6202   -0.3534   0.5564   7.91
##
## Multiple R-squared:  1.493e-06 , Adjusted R-squared:  -0.0007761
## F-statistic: 0.266 on 1 and 1287 DF,  p-value: 0.6061
```

```r
# None of the variables produce significant results here either.
# My hypothesis is that missing data for key indicators like
# whether the respondent was paid in cash, whether they received vocational
# training, etc. in underestimating the impact of these factors on total savings.
# Moreover, as noted in part 2A, the wide range of values in total_savings
# (including negative values and outliers/extreme values) could also be
# contributing to the insignificant results. i.e. the outliers in total_savings
# could suppress the impact of other variables on total_savings.
# This kind of result and missing values issue are things I would discuss with my
# supervisor to understand how researchers deal with them and navigate next steps.

# Further, other than the variables available in this dataset,
# a few other indicators that I would be interested in observing as factors that
# influece 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.

################ Optional Task 4 - Randomization Evaluation #################

# 1. I would create balance tables that compare basic sociodemographics 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 all other 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.

# I would create balance tables using t-tests to compare sample means of the two
# groups on the factors listed in the previous point.
# 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.
# Variables that I would use would be similar to the ones I used and
# outlined in the regression section.

# A short example of using t-test for creating balance tables is coded below.

test_var <- c("q102_age", "q105_attend_school", "q120_a_vocational_training",
              "q131_residence", "q134_l_sewing_machine", "w02_paid_in_cash",
              "w20_hours_unpaid_job99", "m901_b_currently_married",
              "m912_a_spouse_attend_school")

for (var in test_var) {
  formula <- as.formula(paste(var, "~ treatment"))
  t_test <- t.test(formula, data = merged)
  print(t_test)
}
```

```
##
##  Welch Two Sample t-test
##
## data:  q102_age by treatment
## t = -0.47143, df = 1267.7, p-value = 0.6374
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.15328710  0.09389075
## sample estimates:
## mean in group 0 mean in group 1
##        18.31529        18.34498
##
##
##  Welch Two Sample t-test
##
## data:  q105_attend_school by treatment
## t = 0.53803, df = 1177.6, p-value = 0.5907
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.003883413  0.006818035
## sample estimates:
## mean in group 0 mean in group 1
##       0.9983713       0.9969040
##
##
##  Welch Two Sample t-test
##
## data:  q120_a_vocational_training by treatment
## t = 1.2383, df = 1255.5, p-value = 0.2158
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.01187690  0.05252675
## sample estimates:
## mean in group 0 mean in group 1
##      0.10543131      0.08510638
##
##
##  Welch Two Sample t-test
##
## data:  q131_residence by treatment
## t = 0.36026, df = 1285.1, p-value = 0.7187
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.02696252  0.03909247
## sample estimates:
## mean in group 0 mean in group 1
##        1.933227        1.927162
##
##
##  Welch Two Sample t-test
##
## data:  q134_l_sewing_machine by treatment
## t = -0.64703, df = 1286.4, p-value = 0.5177
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.03312901  0.01669600
## sample estimates:
## mean in group 0 mean in group 1
##      0.05087440      0.05909091
##
##
##  Welch Two Sample t-test
##
## data:  w02_paid_in_cash by treatment
## t = -0.036401, df = 644.63, p-value = 0.971
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.06632547  0.06391125
## sample estimates:
## mean in group 0 mean in group 1
##       0.7682540       0.7694611
##
##
##  Welch Two Sample t-test
##
## data:  w20_hours_unpaid_job99 by treatment
## t = 0.15975, df = 1013.9, p-value = 0.8731
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -5.077491  5.977453
## sample estimates:
## mean in group 0 mean in group 1
##        26.71126        26.26128
##
##
##  Welch Two Sample t-test
##
## data:  m901_b_currently_married by treatment
## t = 1.3188, df = 212.35, p-value = 0.1886
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.0398150  0.2007975
## sample estimates:
## mean in group 0 mean in group 1
##       0.8155340       0.7350427
##
##
##  Welch Two Sample t-test
##
## data:  m912_a_spouse_attend_school by treatment
## t = -1.0155, df = 147.1, p-value = 0.3115
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.9613474  0.6298621
## sample estimates:
## mean in group 0 mean in group 1
##        0.982906        1.648649
```

```
# As expected, none of the p-values are significant which means that we can
# assume that there are no significant differences between the treatment and
# control groups wrt to these variables that have been tested for.
# But they could be significant for other variables so it is important to conduct
# these balance tests on as many comparison variables as possible and relevant
# for the specific analysis
```