Problem Set 2 Solutions

Exercise 1: Do Microfinance Loans Affect Expenditure?

Most of the world's poor have limited access to formal credit. Traditionally, they have had to resort to their social networks (family, friends) or local moneylenders that charge exorbitant interest rates (upwards of 100%). This often prevents individuals from making investments in potential businesses or productive assets. In the last 15 years, microfinance institutions (MFI) have emerged all across the developing world to address this problem. The basic model of a microfinance institution (such as the Grameen Bank) is to provide small loans to a group of potential borrowers at much lower interest rates. The number of very poor families with a microloan has grown exponentially: from 7.6 million in 1997 to 137.5 million in 2010. Micro-credit has been heralded as a major advance in the reduction of global poverty. However, in recent years, critiques have emerged accusing micro-finance institutions of acting irresponsibly by holding the poor to very strict repayment schedules and charging unreasonably high interest rates. So where does this leave us? Does micro-credit help or hinder? In the January 2015 issue of the American Economic Journal: Applied Economics, six papers evaluating the merits of micro-credit were released. We will explore the results from one of these papers by Augsburg et al., which evaluates a microfinance institution in Bosnia and Herzegovina.

Data Description

The data for this exercise come from a study conducted in Bosnia and Herzegovina investigating the effects of a small loan on access to liquidity, self-employment, income, labor supply, expenditure, and savings. These were individual liability loans with monthly repayments and an interest rate of 22%. The sample consists of potential borrowers (who were just marginally eligible for loans). Approximately half the sample was randomly selected to receive the loan (the treatment group), while the other half did not receive anything (control group). You have a subsample of individuals (both in the treatment and control group) that the researchers used for their analysis. The respondent (= the loan applicant) answered questions about the household they belonged to as well as about their health and personal outcomes. The PS2_MFI.dta file includes the following variables (along with some others that we will not ask you to analyze):

- `treatment` : dummy equal to 1 if the respondent is in the treatment group (which received a loan)
- `resp_female` : dummy equal to 1 if the respondent is female
- `resp_age` : the respondent's age
- `hhmem` : number of household members
- `hhmem_adults` : number of adults in household (≥ 14)
- `hhmem_children` : number of children in household (<14)
- `hhmem_elderly` : number of elderly in household (>64)
- `total_exp` : total annual household expenditure in Bosnian Convertible Marks (BAM)
- `food_exp` : annual household expenditure on food in BAM
- `nondur_exp` : annual household expenditure on nondurables (rent, fuel, transport, clothes, insurance,...) in BAM
- `dur_exp` : annual household expenditure on durables (education, furniture, vehicle...) in BAM
- `temp_exp` : annual household expenditure on temptation goods (cigarettes, tobacco, alcohol) in BAM

The prevailing exchange rate at the time of the study was approximately USD 1 = 1.63 BAM.

Question 1

Load the dataset `PS2_MFI.dta`. Notice that this is a `.dta` file. So you will need to use the `haven` package.

```
In [1]: library(haven)
ps2<-read_dta("PS2_MFI.dta")
head(ps2)
```

A tibble: 6 × 15

interest	treatment	resp_female	resp_age	resp_primary	resp_secondary	resp_tertiary	hhmem	...
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	...

a) How many respondents are in your data set? How many respondents are unmarried? What is the mean age among the respondents in the sample? What is the mean number of children in respondents' households in your sample?

Note there are some missing values of respondents' age. What argument do you have to add to `mean()` to get around this?

```
In [2]: paste("Number of respondents", nrow(ps2))
paste("Number of unmarried respondents", sum(ps2$resp_married==0)) #number of unmarried respondents
paste("Mean age",mean(ps2$resp_age, na.rm=TRUE))
paste("Mean # kids", mean(ps2$hhmem_children))
```

'Number of respondents 539'

'Number of unmarried respondents 421'

'Mean age 37.3439861710837'

'Mean # kids 0.866474582560297'

There are 539 respondents in the data set. 421 of them are unmarried. The average age of respondents is 37.3 (noting that we are missing age information on 1 respondent). On average, a respondent's household has 0.87 children in it.

b) Construct a variable `total_exp_pc` equal to total expenditures per capita in BAM. Plot a histogram (Hint: use the `hist()` command) of this constructed variable. What is the range of household total expenditures per capita? (You may want to refer to US Dollars in the discussion, so as to make sense of the income level of these MFI clients).

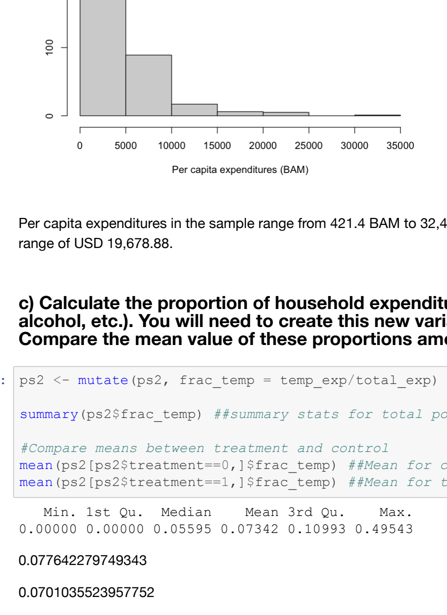
```
In [3]: library(dplyr)
#Create variable
ps2 <- mutate(ps2, total_exp_pc = total_exp/hhmem)

#Plot a histogram
hist(ps2$total_exp_pc,
     main = "Per capita expenditures",
     xlab = "Per capita expenditures (BAM)")

#Calculate the range of per capita expenditures in BAM and USD
summary(ps2$total_exp_pc) #summarize values in BAM
summary(ps2$total_exp_pc/1.63) #summarize values in USD
paste("Range", range(ps2$total_exp_pc)[2]-range(ps2$total_exp_pc)[1])
paste("Range", (range(ps2$total_exp_pc)[2]-range(ps2$total_exp_pc)[1])/1.63)
```

```
-- Attaching packages --
-- Conflicts --
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
[illegible]
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
[illegible]
```

'Range 32076.5774786287'

'Range 19678.8787...'



Per capita expenditures in the sample range from 421.4 BAM to 32,498.0 BAM, which is approximately USD 258 to USD 19,937, for a range of USD 19,678.88.

c) Calculate the proportion of household expenditures spent on "temptation" goods (cigarettes, alcohol, etc.). You will need to create this new variable. What is the mean? What is the median? Compare the mean value of these proportions among treatment and control households.

```
In [4]: ps2 <- mutate(ps2, frac_temp = temp_exp/total_exp)

summary(ps2$frac_temp) ##summary stats for total population

#Compare means between treatment and control
mean(ps2[ps2$treatment==0,]$frac_temp) ##Mean for control
mean(ps2[ps2$treatment==1,]$frac_temp) ##Mean for treatment
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.05595 0.07342 0.10993 0.49543
```

0.077642279749343

0.0701035523957752

In the full population, an average of 7.34 $\%$ and a median of 5.60 $\%$ of household expenditures are spent on temptation goods. The mean is 0.7 percentage points lower in the treatment group than in the contorl group.

Question 2

We will now explore the role of household size in food consumption. Consider these two models:

Model (1): $\ln($ food_exp_pc $)= \beta_0 + \beta_1 \ln($ nondur_exp_pc $)+\beta_2$ treatment $+u$

Model (2): $\ln($ food_exp_pc$)= \beta_0 + \beta_1 \ln($nondur_exp_pc$)+\beta_2$ treatment$+\beta_3 \ln($hhmem$)+u$

a) Estimate equations (1) and (2).

```
In [5]: #First, create a new variable that is food consumption expenditures per capita
ps2 <- mutate(ps2, food_exp_pc = food_exp/hhmem, nondur_exp_pc = nondur_exp/hhmem)

reg1<-lm(log(food_exp_pc)~log(nondur_exp_pc)+treatment, data=ps2)
summary(reg1)

reg2<-lm(log(food_exp_pc)~log(nondur_exp_pc)+treatment +log(hhmem), data=ps2)
summary(reg2)
```

```
Call:
lm(formula = log(food_exp_pc) ~ log(nondur_exp_pc) + treatment,
    data = ps2)

Residuals:
     Min       1Q   Median       3Q      Max
-3.06478 -0.46000  0.02476  0.47317  2.11368

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.82703    0.15848  36.767   <2e-16 ***
log(nondur_exp_pc)  0.25387    0.02466  10.293   <2e-16 ***
treatment          -0.10013    0.06060  -1.652    0.099 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6973 on 536 degrees of freedom
Multiple R-squared:  0.1712,	Adjusted R-squared:  0.1682
F-statistic: 55.38 on 2 and 536 DF,  p-value: < 2.2e-16

Call:
lm(formula = log(food_exp_pc) ~ log(nondur_exp_pc) + treatment +
    log(hhmem), data = ps2)

Residuals:
     Min       1Q   Median       3Q      Max
-3.04594 -0.40743  0.03019  0.43025  2.23424

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.94847    0.18545  37.468  < 2e-16 ***
log(nondur_exp_pc)  0.18010    0.02394   7.523 2.27e-13 ***
treatment          -0.08494    0.05586  -1.521    0.129
log(hhmem)         -0.57568    0.05868  -9.811  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6426 on 535 degrees of freedom
Multiple R-squared:  0.2976,	Adjusted R-squared:  0.2937
F-statistic: 75.57 on 3 and 535 DF,  p-value: < 2.2e-16
```

b) Interpret each of the estimated parameters of equation (2). The results you are finding on the role of household size may a priori seem suprising. Try to think about a scenario where two households of the same size, with the same treatment status and per capita expenditures on nondurable goods, decide to move in together. What does your estimation predict about food consumption per capita?

$\hat{\beta}_0$ tells us that if a household's per capita expenditures on nondurable goods is equal to 1 BAM (such that $log(nondur_exp_pc) = 0$), treatment status is equal to control, and household members are equal to 1 ((such that $log(hhmem) = 0$)), we would predict that their food consumption expenditures would be 6.95 BAM.

$\hat{\beta}_1$ tells us that holding number of household members and treatment status constant, 1$\%$ higher per capita expenditures on nondurable goods is associated with 0.18$\%$ higher per capita food expenditures by 0.18$\%$. This is highly statistically significant ($t = 7.523$).

$\hat{\beta}_2$ tells us that being in the treatment group (holding per capita expenditures on nondurable goods and number of household members fixed) decreases per capita food expenditures by 8.4$\%$, but is not statistically significant at even a 0.1 significance level ($t = -1.521$). This suggests that receiving a microfinance loan does not affect per capita food expenditure.

Finally, $\hat{\beta}_3$ tells us that holding treatment status and per capita expenditures on nondurable goods fixed, 1$\%$ larger households have on average 0.576$\%$ lower food expenditures per capita. This is highly statistically significant ($t = -9.811$).

If these two households move in together, the household size doubles (increases by 100$\%$) while per capita expenditures on nondurable goods and treatment status stay the same. Hence we expect per capita food expenditures to decrease by 57.6$\%$.

c) How did your estimate of $\hat{\beta}_1$ change between equation (1) and equation (2)? Without performing any calculations, what information does this give you about the correlation between expenditure per capita on nondurable goods and household size? (Explain your reasoning in no more than 4 sentences.)

$\hat{\beta}_1$ goes from 0.26 in Model 1 to 0.18 in Model 2, meaning we had an upward bias before including a control for total household members. Given the difference in the estimates, it seems that model 1 suffered from omitted variable bias (a violation of MLR 4). We see in model 2 that $\log(hhmem)$ is negatively correlated with $\log(food\_exp\_pc)$. Hence we can infer that the correlation between expenditures per capita on nondurable goods and household size is negative. Formally, we could calculate:

$\widehat{\beta_{1,model1}} - \widehat{\beta_{1,model2}} = \rho\hat{\beta}_3 \rightarrow 0.26 - 0.18 = \rho(-0.576) \rightarrow \rho = -0.139$ where ρ is the correlation between log nondurable good expenditures per capita and log household size.

d) Predict the expected value of food expenditure per capita of a treatment household with 3 members and per capita expenditures on nondurable goods of BAM 1000 using your estimates from equation (2).

```
In [6]: exp(reg2$coefficients[1]+reg2$coefficients[2]*log(1000)+reg2$coefficients[3]*1+reg2$coefficients[4]*log(3))
```

(Intercept): 1763.70195835973

For this household, we would predict per capita food expenditures to be 1763.7 BAM.

Question 3

A country's dependency ratio is the ratio of old and young dependents (dependents are those not in the labor force) to the working-age population. A similar measure could be constructed for the household:

$$hhdr = \frac{\text{hh members under 14 or over 64}}{\text{hh members aged 14 to 64}}$$

Model 2 (as well as Model 1) does not capture how the composition of a household, i.e. the characteristics of the members, is associated with food consumption per capita. You suspect that the structure of the family affects food expenditure per capita controlling for the log of household size and the log of expenditure per capita on nondurable goods (think about how children and older people might consume less food than adults; and how larger households might have more children). Specifically you hypothesize that a higher dependency ratio is associated with lower food expenditure per capita, holding other factors constant.

(a) Write an equation you could estimate that would allow you to test this hypothesis.

$\ln($ food_exp_pc$)= \beta_0 + \beta_1 \ln($nondur_exp_pc$)+\beta_2$treatment$+\beta_3 \ln($hhmem$)+\beta_4$hhdr$+u$

(b) Estimate the equation in part (a). What can you conclude about the hypothesis?

Note that some households don't have members aged 14 to 64, which means that their $hhdr$ would be undefined. Replace $hhdr$ with NA for these observations.

```
In [7]: #Calculate the dependency ratio for each household
ps2 <- mutate(ps2, hhdr = (hhmem_children + hhmem_elderly)/(hhmem_adults - hhmem_elderly))

#Some of the denominators are 0, replace hhdr in these households with NAs
ps2[which(ps2$hhdr==Inf),]$hhdr<-NA

#Estimate the new regression
model3<-lm(log(food_exp_pc)~log(nondur_exp_pc)+treatment+log(hhmem)+hhdr, data=ps2)
summary(model3)
```

```
Call:
lm(formula = log(food_exp_pc) ~ log(nondur_exp_pc) + treatment +
    log(hhmem) + hhdr, data = ps2)

Residuals:
     Min       1Q   Median       3Q      Max
-3.04244 -0.41084  0.03477  0.44006  2.23218

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.95594    0.18840  36.921  < 2e-16 ***
log(nondur_exp_pc)  0.17946    0.02424   7.403 5.24e-13 ***
treatment          -0.08858    0.05636  -1.572    0.117
log(hhmem)         -0.56860    0.06439  -8.830  < 2e-16 ***
hhdr               -0.02110    0.06104  -0.346    0.730
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6446 on 531 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.2948,	Adjusted R-squared:  0.2895
F-statistic:  55.5 on 4 and 531 DF,  p-value: < 2.2e-16
```

Our null hypothesis is that $\beta_4 = 0$. Note that with a t-statistic of -0.346, we fail to reject the null hypothesis, even at the 10\% significance level.

Exercise 2: Perceptions of Global Warming

Gallup Polling collects information about a variety of topics (health, environment, political attitudes, education). In March 2018, Gallup conducted a poll to gather information about the perceived onset of global warming among Californian voters. Question: "Do you believe that global warming is caused by human activities?"

Group	Number of Observations	Yes
All voters	1408	697
Republicans	482	133
Democrats	521	375
Independents	405	189

Consider first the overall result (all voters). Let p be the fraction of all voters in California (the population of interest) that believe that global warming is caused by human activities.

Note: You answer to the following questions should not necessarily require any R code (unless you would like to use R as a calculator). Some of your answers will require that you type in equations. LaTeX is a typsetting language that makes nicely formatted equations. At the beginning and end of an equation you will need to type a \$ to tell Jupyter that you are typing a Latex equation. The following website has some simple examples showing you how to format your equations in Latex: http://www.personal.ceu.hu/tex/cookbook.html .

(a) Use the survey results to estimate p for the whole population.

```
In [8]: p_hat<-697/1408
p_hat
```

0.495028409090909

$\hat{p} = 0.495$

(b) Construct a 95% confidence interval for p. Interpret.

```
In [9]: se_p_hat<-sqrt((p_hat*(1-p_hat))/1408)

left_side<-p_hat-1.96*se_p_hat
right_side<-p_hat+1.96*se_p_hat

print(paste0("[", left_side, ",", right_side, "]"))
```

```
[1] "[0.468912612426549,0.521144205755269]"
```

Using the z-table, for two-sides, with 95% confidence level, $c = 1.96$. Hence we can calculate the standard error of $\hat{p}$, and plug it into our confidence interval formula to get the confidence interval displayed above. This means we believe there is a 95\% probability that p is between 0.4689 and 0.5211.

(c) Suppose you want to test the null hypothesis that 65% of Democrats believe that global warming is caused by human activities against the alternative hypothesis that more than 65% of Democrats believe that global warming is caused by human activities. Write down the null and alternative hypotheses. Is this a one-sided or two-sided test?

The hypotheses are:

$$H_0 = p\_dem = 0.65$$
$$H_1 = p\_dem > 0.65$$

This is a one-sided test.

(d) Generate a test statistic which will allow you to test the null hypothesis that more than 65% of Democrats believe that global warming is caused by human activities, and identify a critical value which will give you 99% confidence.

```
In [10]: p_dem_hat<-375/521

t_stat<-(p_dem_hat-0.65)/sqrt((0.65*(1-0.65))/521)
p_dem_hat
t_stat
```

0.719769673704415

3.33883419751349

Noting that $\hat{p} = 375/521 = .7198$ (the number of Democrats saying yes divided by the total number of Democrats sampled), we calculate our t-statistic using the formula:

$$t = \frac{.7198 - .65}{\sqrt{.65(1-.65)/521}} = 3.3388$$

Additionally, using a z-table, we note that the critical value for a 99% one-sided (positive) test is 2.33.

(e) Implement your test and interpret your results.

Comparing the test statistic and the critical value:

$$3.34 > 2.33$$

Therefore, we reject the null. We have statistical evidence at the 1% level that more than 65% of democrats believe that global warming is caused by human activities.