Lam Wun Ming 1155049536
Asgn5

1. As this is not a random assignment, it is not advised to compare participants to non-participants to see the effect of microfinancing, which can cause endogeneity problem, causation effect raised from other omitted variables. To be specific, participants of microfinancing program are more likely to have financial burden, so they participate in the microfinancing program. It's not the microfinancing increase their financial burden. In order to establish their causal relationship, we need to control all of the variables except microfinancing by matching the profile of participants and non-participants.

2. $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$ In Ordinary Least Squares (OLS), R^2 is Often used as a goodness-of-fit measure. The idea of R^2 is simple, it is the collection of difference between actual value and predicted value. There are several approaches to thinking about R^2 in continuous outcome variables, but being difficult to calculate in categorical outcome variables. We cannot simply collect the difference between 0 and 1 in logistic regression because it may cause over-estimation or under-estimation. However, to evaluate the goodness-of-fit of logistic regression, several pseudo R^2 have been developed. These are "pseudo" R^2 because they look like R^2 in the sense that they are on a similar scale,

ranging from 0 to 1 (though some pseudo R^2 never achieve 0 or 1) with higher values indicating better model fit. I will apply "rcompanion" library to evaluate $R^2_{McFadden}$, $R^2_{Cox\ and\ Snell}$ and $R^2_{Nagelkerke}$. The result is 0.511, 0.419 and 0.640. Moreover, AIC is also a measure to compare two different logistic model, model with lower AIC would be the preferred one. AIC of this model is 2317.We will also compare the model by AIC in the next following question.

The first six records of propensity score are shown as following:

```
         1          2          3          4          5          6
0.18469024 0.23084124 0.58621144 0.03550805 0.03072494 0.02001465
```

3. The $R^2_{McFadden}$, $R^2_{Cox\ and\ Snell}$ and $R^2_{Nagelkerke}$ are 0.435, 0.322 and 0.545 respectively. The AIC is 2662. Both of the R^2 and AIC pointed out that the previous model is better than this model. The first six records of propensity are shown as following:

```
          1           2           3           4           5           6
0.103485241 0.399095674 0.116560730 0.204586574 0.112423051 0.002592949
```

One of the reason of model I outperform model II is that variable used in model I have more predictive power. The difference between model I and model II is that model I use those "Parents of household head own land" variables, while model II does not. When we go back and consider the original purpose of logistic regression, it's not difficult to find that the phenomenon account for those variables are better in classifying the participants and non-participants.

4. To conduct the post-matching tests, we need to determine the dependent variable to do matching first. Among those 108 variables in the dataset, only "village", "thana", "percaexpenditure", and "HHlandirrigated" are not being used by the above model. "percaexpenditure" and "HHlandirrigated" indicated the per-capita income of a particular group/family and how much land they irrigated. It seems that they are suitable to analyze the impact of microfinancing. The matching results shown in the Table I and Table II in appendix I. The result of both models is not significant in both per capita income and amount of land irrigated. It means that microfinance do not have significant impact on these two fields. The summary of match balance statistics is shown as Table III and IV in appendix II.

Lam Wun Ming 1155049536
Asgn5

Appendix I

Table I: Impact of microfinancing on per capita expenditure

	Model I	Model II
Estimate...	17.261	-26.694
AI SE......	39.989	46.57
T-stat.....	0.43164	-0.57319
p.val......	0.66601	0.56652
Original number of observations.............	4126	4888
Original number of treated obs..............	411	642
Matched number of observations..............	411	642
Matched number of observations (unweighted).	1721	2215

Table II: Impact of microfinancing on amount of land irrigated

	Model I	Model II
Estimate...	-0.52392	0.80932
AI SE......	0.58769	0.72512
T-stat.....	-0.89148	1.1161
p.val......	0.37267	0.26437
Original number of observations.............	4126	4888
Original number of treated obs..............	411	642
Matched number of observations..............	411	642
Matched number of observations (unweighted).	1721	2195

Lam Wun Ming 1155049536
Asgn5

Table III: Match balance (per capita expenditure)

Model I:

```
Before Matching Minimum p.value: < 2.22e-16
Variable Name(s): age agehead HGC  Number(s): 2 12 15

After Matching Minimum p.value: < 2.22e-16
Variable Name(s): age HGChead  Number(s): 2 10
```

Model II:

```
Before Matching Minimum p.value: < 2.22e-16
Variable Name(s): savings hhsize fed med mar age2 age3 age4 age agehead  Number(s): 2 7 10 11 12 13 14 15 18 19

After Matching Minimum p.value: < 2.22e-16
Variable Name(s): savings livevalue fed agehead  Number(s): 2 4 10 19
```

Table IV: Match balance (amount of land irrigated)

Model I:

```
Before Matching Minimum p.value: < 2.22e-16
Variable Name(s): age agehead HGC  Number(s): 2 12 15

After Matching Minimum p.value: < 2.22e-16
Variable Name(s): age HGChead  Number(s): 2 10
```

Model II:

```
Before Matching Minimum p.value: < 2.22e-16
Variable Name(s): savings hhsize fed med mar age2 age3 age4 age agehead HHlandirrigated  Number(s): 2 7 10 11 12 13 14 15 18 19 93

After Matching Minimum p.value: < 2.22e-16
Variable Name(s): savings livevalue agrincome hhsize fed  Number(s): 2 4 5 7 10
```

Lam Wun Ming 1155049536
Asgn5

Appendix II (Source Code)

```
#libraries
library(Matching)
library(foreign)
#reading data
df <- read.dta("~/Downloads/asgn5.DTA")

#Model I
tmp1 <- df[,c(95,7,82,8,84:94,98,10:81,97)]
##delete na value
tmp1 <- tmp1[complete.cases(tmp1), ]
Y1 <- tmp1$percaexpenditure
Tr1 <- tmp1$part
#rule out the dependent variable from the regression dataset
tmp1 = subset(tmp1, select = -c(percaexpenditure))
logit1 <- glm(part~.,family="binomial",data=tmp1)
##propensity score
ps1 <- logit1$fitted
##Matching
rr1 <- Match(Y=Y1, Tr=Tr1, X=ps1, M=1)
##Matching Balance
mb1 <- MatchBalance(part~., data=tmp1, match.out=rr1, nboots=500)

#Model II
##create dummy variables
df$HGCdummy <- ifelse(df$HGC > 0, c(1), c(0))
tmp2 <- df[,c(95,108,107,106,105,104,99,83,1,2,4,6,3,100:102,5,7,82,92,93,10:81,97)]
tmp2 <- tmp2[complete.cases(tmp2), ]
Y2 <- tmp2$percaexpenditure
Tr2 <- tmp2$part
tmp2 = subset(tmp2, select = -c(percaexpenditure))
logit2 <- glm(part~.,family="binomial",data=tmp2)
ps2 <- logit2$fitted
rr2 <- Match(Y=Y2, Tr=Tr2, X=ps2, M=1)
mb2 <- MatchBalance(part~., data=tmp2, match.out=rr2, nboots=500)

### If we want to see the effect on HHlandirrigated, replace 97 to 103
```