Decision to Purchase Insurance Based on Default Option

Taarak Shah

Contents

Introduction

In this report, we will examine data from a study about social networks' influence on a farmer's decision to purchase weather insurance based on the default option the farmer is presented with: to purchase or to not purchase. This analysis can help inform whether an "opt-in" structure noticeably influences the individual's decision. The experiment was randomized on farmers in rural China, who were asked to attend information sessions with their peers and presented a default option to buy or not. Information on whether they chose to purchase insurance was then collected.

The study that this report is based on (Cai et al., 2014) looks to determine the influence on social networks and how that changes the farmer's decision to insure. The scope of that essay is much larger than this report, for they are looking to quantify how the farmer's social connections and the decisions of their peers impacts their own decision. In the paper, they showed that the decision to purchase insurance did not quantify the effect of the social networks, rather the fact that farmers shared their knowledge of insurance via these networks, which in turn influenced their final decision. The treatment in their paper was whether the farmer was subjected to an intensive information session or not, where "simple" sessions took 20 minutes, and "intensive" sessions took 45 minutes and covered the benefits of insurance and how it works more thoroughly. In my analysis, the treatment is considered the default option since I was more interested in how this opt-in structure influenced the farmer's decision, regardless of the type of information they received.

Another essay I found interesting when formulating this analysis was a more technical paper that looked to compare estimators of causal treatment effects with propensity score methods (Lunceford et al., 2004). This paper examined popular implementations of propensity score stratification and weighting. It explains how to adjust for confounding and why these methods prove to be useful in practice, and how treatment conditional on observed covariates inform the propensity score, then how individuals are stratified based on the estimated propensity scores. In my analysis, propensity scores were used in three of the four methods, so this paper was helpful in interpretation of my results and how to use these methods appropriately.

Methods

In this section, we will address the data used in the analysis, how we chose what to include in the models, and an overview of the methods used to estimate both the average treatment effect (ATE) and the average treatment effect among the treated (ATT). We first separate the data based on the default option presented: whether to buy or to not buy insurance. The Table One for our data is listed below.

	Level	Not Buy	Buy	p-value	test	SMD
n		727	683			
Purchased Insurance? (Outcome) (%)	No	427 (58.7)	329 (48.2)	<0.001		0.213
	Yes	300 (41.3)	354 (51.8)			
Household Characteristics - Age (mean (SD))		51.93 (12.02)	51.05 (12.29)	0.174		0.072
Household Characteristics - Household Size (mean (SD))		4.97 (2.20)	4.83 (1.94)	0.223		0.065
Area of Rice Production (mu, mu=1/15 hectare) (mean (SD))		13.10 (15.37)	13.76 (27.04)	0.574		0.030
Perceived probability of disaster next year (%) (mean (SD))		33.11 (17.00)	33.09 (16.10)	0.975		0.002
Gender of head of household (%)	F	70 (9.6)	68 (10.0)	0.907		0.011
	M	657 (90.4)	615 (90.0)			
Intensive info session? (%)	Simple	370 (50.9)	347 (50.8)	1.000		0.002
	Intensive	357 (49.1)	336 (49.2)			
Risk aversion (0-1, 0=risk-loving, 1=risk-averse) (mean (SD))		0.17 (0.31)	0.18 (0.31)	0.826		0.012
Literacy (%)	Illiterate	156 (21.5)	136 (19.9)	0.516		0.038
	Literate	571 (78.5)	547 (80.1)			
Takeup rate prior to experiment (mean (SD))		0.36 (0.21)	0.50 (0.25)	<0.001		0.627

The models I fit centered around a few key variables. We wanted to determine the causal effect on the decision to purchase insurance, which has a fairly balanced separation by default option when looking at the Table One. As established, our response variable will be whether the individual purchases insurance or not, and our treatment will be the default option presented to the farmer. Other covariates considered in the model are whether the farmer had insurance prior to the experiment ("Takeup rate prior to experiment"), risk aversion, the area of rice production (i.e. how large their farm is), and their age. These variables were included after fitting a model and utilizing stepwise selection methods. In the binomial logistic regression model fit, the values that minimized the AIC and BIC were the ones listed. Intuitively, these made sense as to why they would influence the decision. From the farmer's perspective, they can rationalize that if

they have a larger farm, it is worth investing in protections for their main source of income. Similarly if the farmer is less risk averse, older, or had already previously purchased insurance, they are more likely to agree it is a necessity and choose to have it.

For the covariates selected in our analysis, there were very few missing observations. There were no missing categorical variables of interest. For the missing continuous values, I utilized multivariate imputation with the `mice` package in R (van Buuren et al., 2011). This method creates a separate model for each incomplete variable and imputes this data based on the result of the model. This is a preferred alternative to using mean or iterative imputation, as it creates the values conditional on what information is contained in the rest of the dataset. It allows us to work with the dataset as a whole and prevents loss of information in our data.

We will now examine an overview of each of the methods used in the analysis. For each method, we estimate the ATE and ATT. The ATE will be compared to the unadjusted ATE, which would overstate the actual impact since it does not adjust for covariates. With these methods, we will hopefully obtain a more reasonable estimate of how much the default option influences the decision to purchase.

The first two methods were regression based methods. One was a standard regression adjustment with covariates and the other was a regression adjustment with propensity score weighting. The binomial logistic regression model was fit with the following covariates:

$$\texttt{takeup.survey} \sim \texttt{1 + defaultBuy + pre.takeup.rate + risk.averse + ricearea.2010 + age} \quad (1)$$

After fitting this model, we then use this model to predict on our data. For each individual, we create two dataframes identical to the original data used to fit the model, then in one dataframe we set the default option to "Buy" and in the other dataframe, we set the default option to "Not Buy". This allows us to isolate the effect of the default option and provide a prediction while accounting for the takeup rate before the experiment, risk aversion, size of the farm, and age. We then subtract the predictions from each other and output an average value over all the predictions, giving us the ATE. For the ATT, we first subset on the data to only include those where the default option was "Buy" already. Then, we create two dataframes identical to this subsetted data and set the default option to "Buy" and "Not Buy" in each dataframe, then predict using our model and average the differences to obtain the ATT.

For regression adjustment with propensity score weighting, we first fit the following model to estimate the propensity score for the data. We build a model with the default option as the response and the same covariates as our original model. The logistic regression model is specified below:

$$\texttt{default} \sim \texttt{1 + pre.takeup.rate + risk.averse + ricearea.2010 + age} \quad (2)$$

We then predict the probabilities of the default option on our data and save these values to be our propensity score. From here, we build a new logistic regression model to predict whether the individual will buy insurance or not, except our only covariate is the treatment and the propensity score. Then, we follow the same procedure as the normal regression adjustment to estimate the ATE and ATT.

Another method used in the analysis was propensity score stratification. The purpose of propensity score stratification in this dataset is to give us an idea of balancing the control and treatment groups in the strata that have similar propensity scores. This provides a comparison to propensity score regression and matching. For this method, we first fit the binomial logistic regression model in Equation (2) and estimate the propensity score. Then, we cut it into 5 quintiles so we can group the individuals based on their propensity scores. We then estimate the ATE and ATT indexed by propensity scores and average over the groups. This allows us to get a better estimate of the ATE since we estimate the treatment groups and use that information to our advantage in determining the decision to purchase.

Finally, we also used propensity score matching. This was implemented with the `MatchIt` package in R (Ho et al., 2011). We only use 1:1 matching since the dataset is fairly small and not every subject is guaranteed a match. This is useful for subjects who are within the same treatment groups but with other covariates that differ. This is dependent on ensuring the propensity score model is correctly specified as well from previous sections, so we hope to build upon the analysis of other methods. In this method, the R package does much of the heavy lifting. It performs pairing and subset selection to create the groups based on covariates and propensity scores. For this method, we used the same covariates as specified in Equation (1). We hope to see individuals matched on these covariates and that they will produce better estimates for the ATE and ATT.

For each of these methods, we obtained the ATE and ATT estimates. We then used a bootstrap method with 100 iterations to estimate standard errors and 95% confidence intervals. Note that the bootstrap method may inflate confidence intervals by nature of variability, and that 100 iterations is limiting in our analysis.

Results

The tables below show the resulting ATE and ATT estimates with corresponding standard errors and 95% confidence intervals for each of our methods.

	Mean (ATE)	SE	95% CI (Lower)	95% CI (Upper)
Unadjusted	0.1063753	0.0297740	0.0505388	0.1757561
Regression adjustment	0.0371483	0.0277923	-0.0123325	0.0941759
Propensity score regression	0.0399585	0.0289668	-0.0160718	0.0998143
Propensity score stratification	0.0463596	0.0289691	-0.0024973	0.1083451
Propensity score matching (k=1)	-0.0015425	0.0242470	-0.0389097	0.0414201

	Mean (ATT)	SE	95% CI (Lower)	95% CI (Upper)
Regression adjustment	0.0373624	0.0251231	-0.0096547	0.0854425
Propensity score regression	0.1054172	0.0248743	0.0593968	0.1493338
Propensity score stratification	0.0461391	0.0292431	-0.0048161	0.1089775
Propensity score matching (k=1)	0.0515636	0.0245333	0.0102914	0.1006489

We will examine each of the results by method.

For standard regression adjustment, the average treatment effect is about 3.7%, indicating a small increase in the decision to purchase insurance if the default option is to purchase. Among those who had the default option as "buy insurance", the ATT is estimated to be about a 3.7% increase as well, which is consistent with the average treatment effect. This helps establish that the treatment assignment was random and that other factors are potentially more useful to distinguish the decision to purchase or not.

For regression adjustment with propensity score weighting, the average treatment effect is about 3.9%, consistent with our output from the regular regression adjustment. However, relative to the regression adjustment without propensity score weighting, the treatment effect among treated is estimated to be about a 10.5% increase, indicating that the treatment group has a large influence on the decision to purchase insurance. This difference can potentially be attributed to the inclusion of the estimated propensity score, allowing us to isolate the effect of the treatment and determine whether it makes a difference on our output.

For propensity score stratification, the average treatment effect is about 4.6%, a bit greater than our output from the regression adjustment results. The treatment effect among treated is estimated to be about a 4.6% increase, indicating that the treatment group has some influence on the decision to purchase insurance. The common theme among our results so far is that the propensity score estimation seems to more strongly isolate the effect of the default option being to buy insurance.

However, for propensity score matching, the average treatment effect is about 0%, a stark contrast from the results from all of our other methods to this point. The treatment effect among treated is estimated to

be about a 5.2% increase, which is consistent with our other results. However, we would be more inclined to disregard many of the results from matching. When we perform 1:1 matching, we obtain many errors that not all treated units may have a match, indicating many "NA" values in our data needing to be removed to obtain an estimate of the treatment at all. This method was ultimately not as promising as initially hoped.

Conclusion

Ultimately, it appears the propensity score regression and propensity score stratification were the two more promising methods for this study. Standard regression adjustment accounted for some variation in the data and provided a better baseline ATE and ATT than the unadjusted estimate, but without the propensity score acting as a weight there may have been some confounding in the data that was unaccounted for. Propensity score matching proved to be unreliable as there were many instances where not all subjects were able to obtain a match, so there are a lot of individuals whose characteristics are not approximated well in this method. The other propensity score methods provided weights while also being complete unlike matching.

Overall, our results indicate that there is a small causal effect of the "opt-in" style of presentation when determining if farmers should purchase insurance or not. We obtained a positive ATE and ATT for 3 of our methods. Though the effect is not incredibly pronounced, it is useful to understand how separation under a different treatment than the inital study (Cai et al., 2014) influenced the decision to purchase. The default option to buy or not buy was not explored in depth in the study, and was mainly used as a covariate to quantify an entirely different effect. To analyze the same data but focus on a different treatment allows us to understand the data in more depth and see what other underlying effects may be at play in the decision to purchase insurance or not. Our results imply that presentation of the option to purchase or not purchase insurance is still relatively important.

References

1. Cai, Jing, et al. "Social Networks and the Decision to Insure." American Economic Journal: Applied Economics, vol. 7, no. 2, 2015, pp. 81–108., https://doi.org/10.1257/app.20130442.
2. Daniel E. Ho, Kosuke Imai, Gary King, Elizabeth A. Stuart (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, Vol. 42, No. 8, pp. 1-28. URL https://www.jstatsoft.org/v42/i08/
3. Lunceford, Jared K., and Davidian, M. "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study." Statistics in Medicine, vol. 23, no. 19, 2004, pp. 2937–2960., https://doi.org/10.1002/sim.1903.
4. Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.

Appendix

Data Cleaning

First, we read in the dataset. We see there are some columns that contain missing data, so we used the `mice` package to complete our dataset. Then, we adjust the variable names in preparation to create the Table One.

```
dat <- social_insure


## handle missing data
colSums(is.na(dat))
```

```
##         address         village   takeup_survey             age           agpop
##               0               0               0               4               6
##   ricearea_2010   disaster_prob            male         default       intensive
##               9               0               3               0               0
##     risk_averse        literacy pre_takeup_rate
##               0              21               0
```

```
dat <- complete(mice(dat, maxit=0))
```

```
## Warning: Number of logged events: 2
```

```
sum(is.na(dat)) ## looks good!
```

```
## [1] 0
```

```
dat$takeup_survey <- as.factor(dat$takeup_survey)
levels(dat$takeup_survey) <- c('No', 'Yes')
dat$takeup_survey_ohe <- ifelse(dat$takeup_survey=='Yes', 1, 0)
dat$male <- as.factor(dat$male)
levels(dat$male) <- c('F','M')
dat$default <- as.factor(dat$default)
levels(dat$default) <- c('Not Buy', 'Buy')
dat$intensive <- as.factor(dat$intensive)
```

```
levels(dat$intensive) <- c('Simple', 'Intensive')
dat$literacy <- as.factor(dat$literacy)
levels(dat$literacy) <- c('Illiterate', 'Literate')


## rename data
new.cols <- c('Natural village', 'Admin village', 'Purchased Insurance? (Outcome)',
              'Age', 'Household Size', 'Area of Rice Production (mu, mu=1/15 hectare)',
              'Perceived probability of disaster next year (%)',
              'Gender of head of household', 'Default option',
              'Intensive info session?',
              'Risk aversion (0-1, 0=risk-loving, 1=risk-averse)',
              'Literacy', 'Take-up rate prior to experiment')
orig.names <- names(dat)
names(dat) <- new.cols
```

Table One

Now, we create the table one, stratified by whether the default option when the experiment was presented was either to buy or to not buy. We will see how this affects the outcome in our analysis.

	level	Not Buy	Buy	p	test	SMD
n		727	683			
Purchased Insurance? (Outcome) (%)	No	427 (58.7)	329 (48.2)	<0.001		0.213
	Yes	300 (41.3)	354 (51.8)			
Household Characteristics - Age (mean (SD))		51.93 (12.02)	51.05 (12.29)	0.174		0.072
Household Characteristics - Household Size (mean (SD))		4.97 (2.20)	4.83 (1.94)	0.223		0.065
Area of Rice Production (mu, mu=1/15 hectare) (mean (SD))		13.10 (15.37)	13.76 (27.04)	0.574		0.030
Perceived probability of disaster next year (%) (mean (SD))		33.11 (17.00)	33.09 (16.10)	0.975		0.002
Gender of head of household (%)	F	70 (9.6)	68 (10.0)	0.907		0.011
	M	657 (90.4)	615 (90.0)			
Intensive info session? (%)	Simple	370 (50.9)	347 (50.8)	1.000		0.002
	Intensive	357 (49.1)	336 (49.2)			
Risk aversion (0-1, 0=risk-loving, 1=risk-averse) (mean (SD))		0.17 (0.31)	0.18 (0.31)	0.826		0.012
Literacy (%)	Illiterate	156 (21.5)	136 (19.9)	0.516		0.038
	Literate	571 (78.5)	547 (80.1)			
Takeup rate prior to experiment (mean (SD))		0.36 (0.21)	0.50 (0.25)	<0.001	0.627	

Other potential treatment groups

For exploratory purposes, let's see how the other categorical variables in our dataset serve as treatment groups (i.e. is there a separation in the outcome for those variables like there is with the default option as the treatment?)

The following tables are stratified by gender of head of household, whether the subject was given a simple or intensive information session, and whether they are literate or not.

	level	F	M	p	test	SMD
n		137	1273			
Purchased Insurance? (Outcome) (%)	No	77 (56.2)	679 (53.3)	0.583		0.058
	Yes	60 (43.8)	594 (46.7)			

	level	Simple	Intensive	p	test	SMD
n		717	693			
Purchased Insurance? (Outcome) (%)	No	385 (53.7)	371 (53.5)	0.994		0.003
	Yes	332 (46.3)	322 (46.5)			

	level	Illiterate	Literate	p	test	SMD
n		290	1120			
Purchased Insurance? (Outcome) (%)	No	170 (58.6)	586 (52.3)	0.064		0.127
	Yes	120 (41.4)	534 (47.7)			

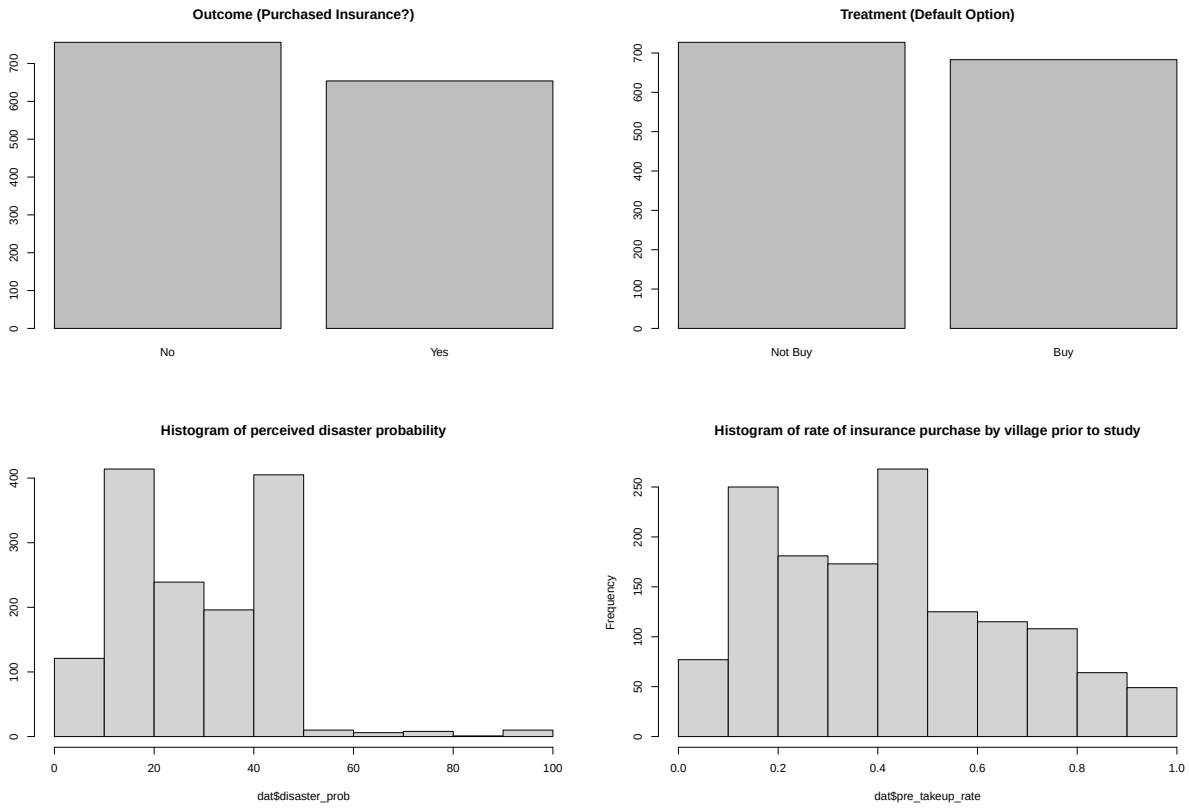
From these three tables, there is not a significant separation in using these variables as the treatment group as none of our p-values are significant. We carry on with our analysis.

Exploratory data analysis

Our outcome in this dataset is whether the farmers purchase insurance or not, with our treatment group being the default option to buy or not buy. We take a look at the distributions of our variables.

```
# set names back to original
names(dat) <- orig.names


# plot some important variables
# chosen based on context from the paper
# when fitting models I plan to use more than those plotted here
par(mfrow=c(2,2))
plot(dat$takeup_survey, main="Outcome (Purchased Insurance?)")
plot(dat$default, main="Treatment (Default Option)")
hist(dat$disaster_prob, main='Histogram of perceived disaster probability')
hist(dat$pre_takeup_rate, main='Histogram of rate of insurance purchase by village prior to study')
```



From our plots above, we see that some of our key covariates follow a left skewed Normal distribution, which could affect our fit in some of the models we will build due to some key outliers. We will need to keep

this in mind when considering how our covariates in a regression model or our probability weights.

Statistical Analysis Methods

Looking forward, some good model candidates for this data could be to examine the average treatment effect among treated using regression or propensity score stratification. From our Table One, we notice a significant difference between our outcome and the default option assigned. I believe matching could also be useful in this scenario, since we can match among those from similar villages and risk aversion, or perceived probabilities of disaster. This could help explain some of the variation in the outcome and help specify which factors contribute to the decision made to purchase insurance or not.

```
## bootstrap for ATE
boot.ATE <- function(stat.fun, data, match=FALSE, k, B=100){
  n <- nrow(data)

  tab <- matrix(NA, nrow=B, ncol=1)
  for (i in 1:B){
    boot.data <- data[sample(1:n, n, replace = TRUE), ]
    if(match==T){
      tab[i,1] <- get(stat.fun)(boot.data, k)$'ATE'
    }
    else{
      tab[i,1] <- get(stat.fun)(boot.data)$'ATE'
    }
  }
  if(match==T){
    # returns vec with NAs removed
    # needed for k=2 propensity matching
    return(list('Takeup'=c("Mean" = mean(tab[,1], na.rm=T),
                           "SE" = sd(tab[,1], na.rm=T),
                           "CI" = c(quantile(tab[,1], .025, na.rm=T),
                                    quantile(tab[,1], .975, na.rm=T)))))
  }
  return(list('Takeup'=c("Mean" = mean(tab[,1]),
```

```
                          "SE" = sd(tab[,1]),
                          "CI" = c(quantile(tab[,1], .025),
                                   quantile(tab[,1], .975)))))
}


## bootstrap for ATT
boot.ATT <- function(stat.fun, data, match=FALSE, k, B=100){
  n <- nrow(data)


  tab <- matrix(NA, nrow=B, ncol=1)
  for (i in 1:B){
    boot.data <- data[sample(1:n, n, replace = TRUE), ]
    if(match==T){
      tab[i,1] <- get(stat.fun)(boot.data, k)$'ATT'
    }
    else{
      tab[i,1] <- get(stat.fun)(boot.data)$'ATT'
    }
  }
  if(match==T){
    # returns vec with NAs removed
    # needed for k=2 propensity matching
    return(list('Takeup'=c("Mean" = mean(tab[,1], na.rm=T),
                           "SE" = sd(tab[,1], na.rm=T),
                           "CI" = c(quantile(tab[,1], .025, na.rm=T),
                                    quantile(tab[,1], .975, na.rm=T)))))
  }
  return(list('Takeup'=c("Mean" = mean(tab[,1]),
                          "SE" = sd(tab[,1]),
                          "CI" = c(quantile(tab[,1], .025),
                                   quantile(tab[,1], .975)))))
}
```

Model Selection

Before any specific methods, I plan to address model selection to be carried through to the other methods. We would use all-subset selection and both backward/forward stepwise selection with BIC to select our optimal model while penalizing for including too many covariates. Our response variable will consist of whether the farmer decided to purchase insurance or not (`takeup.survey`). Once the optimal model is selected, we can generalize that model with the chosen covariates to the different methods we plan to address in the analysis.

```
mod_c <- glm(takeup_survey ~ default +  age + agpop + ricearea_2010 +
                                        disaster_prob + male + intensive +
                                        risk_averse + literacy + pre_takeup_rate,
             family='binomial', data=dat)

mod_s <- glm(takeup_survey ~ default, family='binomial', data=dat)

n <- nrow(dat)
step(mod_c, scope=list(upper=mod_c, lower=mod_s), direction='backward',
     data=dat, k=log(n), trace=0)
```

```
##
## Call:  glm(formula = takeup_survey ~ default + age + ricearea_2010 +
##     risk_averse + pre_takeup_rate, family = "binomial", data = dat)
##
## Coefficients:
##     (Intercept)       defaultBuy              age    ricearea_2010
##        -3.12536          0.16735          0.02677          0.02865
##     risk_averse  pre_takeup_rate
##         1.06378          2.22890
##
## Degrees of Freedom: 1409 Total (i.e. Null);  1404 Residual
## Null Deviance:       1947
## Residual Deviance: 1780  AIC: 1792
```

```
step(mod_s, scope=list(upper=mod_c, lower=mod_s), direction='forward',
     data=dat, k=log(n), trace=0)
```

```
##
## Call:  glm(formula = takeup_survey ~ default + pre_takeup_rate + risk_averse +
##     ricearea_2010 + age, family = "binomial", data = dat)
##
## Coefficients:
##     (Intercept)       defaultBuy  pre_takeup_rate      risk_averse
##         -3.12536          0.16735          2.22890          1.06378
##   ricearea_2010              age
##          0.02865          0.02677
##
## Degrees of Freedom: 1409 Total (i.e. Null);  1404 Residual
## Null Deviance:       1947
## Residual Deviance: 1780  AIC: 1792
```

From the above result, our model that minimizes BIC is given by:

`takeup.survey ~ 1 + defaultBuy + pre.takeup.rate + risk.averse + ricearea.2010 + age`

Before accepting this as the final model, we consider if including interactions between these covariates is beneficial.

```
mod_int <- glm(takeup_survey ~ default + (pre_takeup_rate + risk_averse + ricearea_2010 + age)^2,
               family='binomial', data=dat)
summary(mod_int)
```

```
##
## Call:
## glm(formula = takeup_survey ~ default + (pre_takeup_rate + risk_averse +
##     ricearea_2010 + age)^2, family = "binomial", data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2873  -1.0314  -0.6692   1.1159   2.1674
```

```
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.9629279  0.7243239  -5.471 4.47e-08 ***
## defaultBuy                      0.1623265  0.1186267   1.368 0.171193
## pre_takeup_rate                 3.5592706  1.2593128   2.826 0.004708 **
## risk_averse                     0.0109740  1.1030643   0.010 0.992062
## ricearea_2010                   0.0762264  0.0284820   2.676 0.007444 **
## age                             0.0428042  0.0122748   3.487 0.000488 ***
## pre_takeup_rate:risk_averse    -0.1357026  0.8414957  -0.161 0.871886
## pre_takeup_rate:ricearea_2010 -0.0160261  0.0279732  -0.573 0.566707
## pre_takeup_rate:age           -0.0215311  0.0209099  -1.030 0.303147
## risk_averse:ricearea_2010       0.0265487  0.0219454   1.210 0.226371
## risk_averse:age                 0.0154744  0.0170678   0.907 0.364595
## ricearea_2010:age              -0.0009902  0.0005265  -1.881 0.060015 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1947.3  on 1409  degrees of freedom
## Residual deviance: 1773.8  on 1398  degrees of freedom
## AIC: 1797.8
##
## Number of Fisher Scoring iterations: 5
```

From our summary, we see none of the covariates are significant, so we accept the following for our covariates.

`takeup.survey ~ 1 + defaultBuy + pre.takeup.rate + risk.averse + ricearea.2010 + age`

We proceed to estimate the ATE and ATT for each of our methods outlined below.

```
## unadjusted ATE
unadjusted <- function(data){
  ## ATE estimate
```

```
  ate.obj <- table(data$takeup_survey, data$default)
  ate.val <- ate.obj[2,2] / (ate.obj[2,2] + ate.obj[1,2]) -
             ate.obj[2,1] / (ate.obj[2,1] + ate.obj[1,1])

  return(list("ATE" = ate.val))
}

set.seed(7485)
unadj.ATE <- boot.ATE("unadjusted", data=dat, B=100)
unadj.ATE
```

```
## $Takeup
##       Mean         SE    CI.2.5%   CI.97.5%
## 0.10637526 0.02977399 0.05053877 0.17575606
```

Regression

The regressions performed would be consistent with the covariates chosen in the "Model Selection" section. I would want to compare the ATT based on the "default option" assigned to the subjects, i.e. whether the experiment is framed that the default option is to buy insurance, and the subject must opt-out if they do not want insurance, or if the default option is to not buy insurance, and the subject must opt-in if they want insurance. The first regression will be a standard regression adjustment with covariates and the second will be a regression adjustment with propensity score weighting.

```
adj.reg <- function(data){
  mod <- glm(takeup_survey ~ 1 + default + pre_takeup_rate + risk_averse + ricearea_2010 + age,
             family='binomial', data=data)

  ## ATE estimate
  data_trt <- data_ctr <- data
  data_trt$default <- "Buy"
  data_ctr$default <- "Not Buy"
```

```
  pred1 <- predict(mod, newdata = data_trt, type = "response")
  pred0 <- predict(mod, newdata = data_ctr, type = "response")
  ate.val <- mean(pred1 - pred0)

  ## ATT estimate
  data_trt <- data_ctr <- data[data$default == "Buy", ]
  data_trt$default <- "Buy"
  data_ctr$default <- "Not Buy"

  pred1 <- predict(mod, newdata = data_trt, type = "response")
  pred0 <- predict(mod, newdata = data_ctr, type = "response")
  att.val <- mean(pred1 - pred0)

  return(list("ATE" = ate.val, "ATT" = att.val))
}

set.seed(7485)
adj.reg.ATE.out <- boot.ATE("adj.reg", data = dat, B=100)
adj.reg.ATT.out <- boot.ATT("adj.reg", data = dat, B=100)
adj.reg.ATE.out
```

Regression adjustment with covariates

```
## $Takeup
##        Mean          SE     CI.2.5%    CI.97.5%
##  0.03697057  0.02775021 -0.01267159  0.09392068
```

```
adj.reg.ATT.out
```

```
## $Takeup
##        Mean          SE     CI.2.5%    CI.97.5%
##  0.03760726  0.02516304 -0.00991900  0.08539772
```

The average treatment effect is about 3.5%, indicating a small increase in the decision to purchase insurance if the default option is to purchase. Among those who had the default option as "buy insurance",

the effect is estimated to be about a 3.7% increase as well, which is consistent with the average treatment effect. This helps establish that the treatment assignment was random and that other factors are potentially more useful to distinguish the decision to purchase or not. However, our 95% confidence interval contains 0, indicating that there may not be a significant effect.

```
ps.reg <- function(data){
  mod <- glm(default ~ 1 + pre_takeup_rate + risk_averse + ricearea_2010 + age,
             family='binomial', data=data)
  data$ps <- predict(mod, type='response')

  data_trt <- data_ctr <- data
  data_trt$default <- "Buy"
  data_ctr$default <- "Not Buy"

  mod.ps <- glm(takeup_survey ~ default*rcs(ps, 5), data=data, family='binomial')
  pred1 <- predict(mod.ps, newdata = data_trt, type = "response")
  pred0 <- predict(mod.ps, newdata = data_ctr, type = "response")
  ## ATE estimate
  ate.val <- mean(pred1 - pred0)
  ## ATT estimate
  att.val <- mean(pred1[data$default == 'Buy'] - pred0[data$default == 'Not Buy'])

  return(list("ATE" = ate.val, "ATT" = att.val))
}

set.seed(7485)
ps.reg.ATE.out <- boot.ATE("ps.reg", data = dat, B=100)
ps.reg.ATT.out <- boot.ATT("ps.reg", data = dat, B=100)
ps.reg.ATE.out
```

Regression adjustment with propensity score weighting

```
## $Takeup
```

```
##        Mean          SE     CI.2.5%    CI.97.5%
##  0.03982978  0.02891239 -0.01588241  0.09953879
```

```
ps.reg.ATT.out
```

```
## $Takeup
##       Mean         SE    CI.2.5%   CI.97.5%
## 0.10541669 0.02487218 0.05940634 0.14931526
```

The average treatment effect is about 3.9%, consistent with our output from the regular regression adjustment. However, relative to the regression adjustment without propensity score weighting, the treatment effect among treated is estimated to be about a 10.5% increase, indicating that the treatment group has a large influence on the decision to purchase insurance.

Propensity Score Stratification

The purpose of propensity score stratification in this dataset is to give us an idea of balancing the control and treatment groups in the strata that have similar propensity scores. This provides a comparison to propensity score matching that I believe will be the most promising method to estimate the treatment effect.

```
pss <- function(data){
  mod <- glm(default ~ 1 + pre_takeup_rate + risk_averse + ricearea_2010 + age,
             family='binomial', data=data)

  pred.pss <- predict(mod, type = "response")
  ps.quint <- cut(pred.pss,
                  breaks = c(0, quantile(pred.pss, p = c(0.2, 0.4, 0.6, 0.8)), 1),
                  labels = 1:5)

  ## ATE estimate
  ate.nA <- nrow(data)
  ate.nAj <- table(ps.quint)

  ate.quint <- tapply(data$takeup_survey_ohe[data$default == "Buy"], ps.quint[data$default == "Buy"], m
                tapply(data$takeup_survey_ohe[data$default == "Not Buy"], ps.quint[data$default == "Not
```

```
  ate.val <- sum(ate.quint * ate.nAj/ate.nA)

  ## ATT estimate
  att.nA <- nrow(data[data$default == "Buy", ])
  att.nAj <- table(ps.quint[data$default == "Buy"])


  att.quint <- tapply(data$takeup_survey_ohe[data$default == "Buy"], ps.quint[data$default == "Buy"], m
                tapply(data$takeup_survey_ohe[data$default == "Not Buy"], ps.quint[data$default == "Not
  att.val <- sum(att.quint * att.nAj/att.nA)

  return(list("ATE" = ate.val, "ATT" = att.val))
}

set.seed(7485)
pss.ATE.out <- boot.ATE("pss", data = dat, B=100)
pss.ATT.out <- boot.ATT("pss", data = dat, B=100)
pss.ATE.out
```

```
## $Takeup
##         Mean           SE      CI.2.5%     CI.97.5%
##  0.046083729  0.029304449 -0.005088994  0.108386703
```

```
pss.ATT.out
```

```
## $Takeup
##         Mean           SE      CI.2.5%     CI.97.5%
##  0.038069101  0.026936929 -0.007605104  0.089457236
```

The average treatment effect is about 4.6%, a bit greater than our output from the regression adjustment results. The treatment effect among treated is estimated to be about a 3.8% increase, indicating that the treatment group has some influence on the decision to purchase insurance.

Matching

Finally, I also plan to explore 1:1 matching for subjects who are within the same treatment groups but with other covariates that differ, such as their risk aversion, the village they are from, their perceived probability of disasters, or others. This could help provide insight into how much these other factors influence their decision to buy insurance or not. This is dependent on ensuring the propensity score model is correctly specified as well from previous sections, so we hope to build upon the analysis of other methods.

```
psm.k <- function(data, k){
  mod.match <- matchit(default ~ 1 + pre_takeup_rate + risk_averse + ricearea_2010 + age,
                       distance = "logit", method = "nearest", ratio = k, data=data)

  ## ATE estimate
  p1 <- mean(data$takeup_survey_ohe)
  p0 <- mean(data$takeup_survey_ohe[as.numeric(mod.match$match.matrix)])
  n1 <- table(data$default)[2]
  n0 <- length(as.numeric(mod.match$match.matrix))

  ate.val <- p1 - p0

  ## ATT estimate
  p1 <- mean(data$takeup_survey_ohe[data$default == "Buy"])
  p0 <- mean(data$takeup_survey_ohe[as.numeric(mod.match$match.matrix)])
  n1 <- table(data$default)[2]
  n0 <- length(as.numeric(mod.match$match.matrix))

  att.val <- p1 - p0

  return(list("ATE" = ate.val, "ATT" = att.val))
}

set.seed(7485)
# k = 1
psm.k1.ATE.out <- boot.ATE("psm.k", data=dat, match=T, k=1, B=100)
```

```
psm.k1.ATT.out <- boot.ATT("psm.k", data=dat, match=T, k=1, B=100)
psm.k1.ATE.out
```

```
## $Takeup
##          Mean           SE      CI.2.5%     CI.97.5%
## -0.001574138  0.024416166 -0.039666098  0.043115907
```

```
psm.k1.ATT.out
```

```
## $Takeup
##        Mean          SE     CI.2.5%    CI.97.5%
## 0.051696743 0.024748336 0.009751132 0.109210047
```

The average treatment effect is about 0%, a stark contrast from the results from all of our other methods to this point. The treatment effect among treated is estimated to be about a 5.2% increase, indicating that the treatment group has some influence on the decision to purchase insurance, which is consistent with our other results.

The results tables collected by ATE and ATT are below.

```
## ATE
ate.res <- rbind(unadj.ATE$Takeup,
                 adj.reg.ATE.out$Takeup,
                 ps.reg.ATE.out$Takeup,
                 pss.ATE.out$Takeup,
                 psm.k1.ATE.out$Takeup)
rn <- c("Unadjusted",
        "Regression adjustment",
        "Propensity score regression",
        "Propensity score stratification",
        "Propensity score matching (k=1)")
rownames(ate.res) <- rn
ate.res
```

```
##                                      Mean         SE      CI.2.5%   CI.97.5%
## Unadjusted                     0.106375264 0.02977399  0.050538767 0.17575606
```

```
## Regression adjustment            0.036970568 0.02775021 -0.012671589 0.09392068
## Propensity score regression      0.039829775 0.02891239 -0.015882413 0.09953879
## Propensity score stratification  0.046083729 0.02930445 -0.005088994 0.10838670
## Propensity score matching (k=1) -0.001574138 0.02441617 -0.039666098 0.04311591
```

```
## ATT
att.res <- rbind(adj.reg.ATT.out$Takeup,
                 ps.reg.ATT.out$Takeup,
                 pss.ATE.out$Takeup,
                 psm.k1.ATT.out$Takeup)
rownames(att.res) <- rn[-1]
att.res
```

```
##                                       Mean         SE      CI.2.5%   CI.97.5%
## Regression adjustment           0.03760726 0.02516304 -0.009919000 0.08539772
## Propensity score regression     0.10541669 0.02487218  0.059406335 0.14931526
## Propensity score stratification 0.04608373 0.02930445 -0.005088994 0.10838670
## Propensity score matching (k=1) 0.05169674 0.02474834  0.009751132 0.10921005
```