Prediction of the body mass index using propensity score matching and multiple linear regression model

Ruike Xu - 1006562550

June 14, 2021

Abstract

Smoking has been one of the most important factors that threaten people's lives, which also potentially influences the weight of smokers. Since smokers and non-smokers are different for many characteristics, propensity score matching would be employed to find matches from treatment and control groups. We would analyze the association of smoking patterns with the body mass index(BMI) for the observations in the Stroke Prediction Dataset. There is no statistical significance for the prediction of BMI based on the smoking pattern of the participants, however, the results indicate a positive tread between smoking and BMI.

Introduction

There is no doubt that smoking has been one of the most crucial factors that risking people's life. Smoking largely increases the probability of getting lung cancer, respiratory and cardiovascular diseases, which has become one of the major preventable causes of death in developed countries. (Doll et al., 2004) Obesity, on the other hand, is the fifth leading cause of death and accounts for 44% of cases of diabetes globally. There is also a study that suggests that the life expectancy of obese smokers is approximately 13 years shorter than non-obese and non-smoking people. (Peeters et al., 2003) The study of the correlation between smoking patterns and obesity of smokers could be complex and not understandable, the published studies have offered conflicting results regarding this topic. While some studies have shown no significant association between smoking patterns and body mass index (BMI), others have suggested that smoking might be associated with lower BMI due to higher energy expenditure and suppressing appetite for smokers. (Dare et al., 2015) Therefore, discovering the relationship between smoking status and BMI of individuals is essential for promoting people's standard of living and life expectancy.

The data set I employed is a Stroke Prediction Dataset (SPD) from Kaggle, in which the author of the data set is Fedesoriano. SPD was created in January 2021 that collects the survey data of participants worldwide. The data set contains information parameters of a patient like gender, age, various disease, smoking status, their corresponding BMI, etc. (Fedesoriano, 2021) SPD provides massive observations for over 5000 patients with various background information so that we could analyze the relationship between the pattern of smoking and the BMI of the participants. In particular, smoking patterns as an important health-related behavior, could be potentially related to the health determination of the participants.

The goal of the study is to determine whether there is a causal relationship between smoking patterns and the BMI of the participant for the world population, in specific, the occurrence of smoking for the participants has a direct influence on a higher level of BMI. We would analyze the SPD data set and employ propensity score matching technique in order to derive a causal relationship for the smoking pattern and BMI of the participants based on this observational data set. The hypothesis I proposed for this study is that there is statistical significance in which the occurrence of smoking for the participants could directly cause a higher level of BMI.

Data

Data Collection Process

My goal of the study is to discover the causal relationship between smoking status and the body mass index for the sampled population. An observational study is required since it's unethical to randomly assign participants to a smoking treatment group. Thus, I deeply searched online for an appropriate observational data set that could provide sufficient observations and various demographic background information. The data set I employed for this study was discovered from an open-source data science community which is the Stroke Prediction dataset created by Fedesoriano in January 2021. The link for the observational survey data set I employed can be found in Appendix. This data was originally used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various conditions of diseases, smoking status, BMI, residence type, etc.

However, some limitations from the chosen data set could potentially affect the further analysis. First, the data set contains an insufficient number of demographic variables for each observation. Some of the important information such as the region that the participant lives in, the ethnicity of the participant is not recorded. That could due to confidential reasons for collecting data. Second, the survey question of this observation data set was not provided by the author. It would be better if we can analyze the purpose of each survey question and perform the data cleaning process more appropriately. Third, the data set was intended for the purpose of predicting stroke for the participants, although it has plenty of demographic information for observations, the data set might have less related information recorded for my goal of study.

Data Summary

The Stroke Prediction Dataset contains 5110 observations, each represents a measurement of a participant. The information that has been recorded for the data set is the ID of the participant, the gender and age of the participant, whether the participant has hypertension and heart disease, whether the participant has married, the type of work, and the residence type of the participant, the average glucose level and BMI of the participant, and the smoking status. The data was collected in January 2021 by Fedesoriano through Kaggle.

In the data cleaning process, we only select the participants that have recorded BMI as non-empty and have a non-zero value, in this case, their responses are useful for the analysis. The observations of smoking status for participants that are recorded unknown are also removed since it has no direct relation with our goal of study. I have also categorized smoking status into a binary variable by gathering participant has smoked and currently smoking into 1. The marriage status of the participants is also transformed into dummy variables, so it would be easier for us to analyze the summary of the data and construct our model. I also mutate and merge two variables that split participants' age into age groups and divide their BMI into different health categories based on the suggestion of the CDC. (About adult bmi, 2020)

Interpretation of Important Variables

Variable name	Data type	Interpretation
ID	Double	Unique identifier for each observation

Variable name	Data type	Interpretation
Gender	Double	Numerical representation of the gender of the participant.(Dummy variable)
Age	Double	Numerical representation of age of the participants
Hypertension	Double	Numerical representation of whether the participant has hypertension.(Dummy variable)
heart disease	Double	Numerical representation of whether the participant has heart disease.(Dummy variable)
Ever married	Double	Numerical representation of whether the participant has married.(Dummy variable)
Work type	Character	Categorical variable for what kind of work the participant does. (Children Government job Never worked Private Self_employed)
Residence type	Character	Categorical variable for the type of living environment of the participant. (Rural / Urban)
Average glucose level	Double	Numerical representation of the average glucose level in blood of the participant
BMI	Double	Numerical variable of the Body Mass Index of the partcipant, BMI is calculated by weight/height^2 in kg/m^2.
Smoking status	Double	Numerical representation of whether the participant smoked.(Dummy variable)
BMI group	Character	Categorical variable of BMI for all observations in the population sample, grouped based on the CDC suggestion of weight status. (Underweight / Normal or Healthy Weight / Overweight / Obese)
Age group	Character	Categorical variable of participants' age in the sampled population. (Under 18 / 18 to 24 / 25 to 34 / 35 to 44 / 45 to 54 / 55 to 64 / Older than 65)

Dummy variable explanation for variables

Gender

Numerical representation	Categories
1	Male
2	Non-Male

Hypertension

Numerical representation	Categories
0	patient doesn't have hypertension
1	patient has hypertension

Heart disease

Numerical representation	Categories
0	patient doesn't have any heart disease
1	patient has heart disease(s)

Ever married

Numerical representation	Categories
0	patient hasn't married before
1	patient has married

Smoking status

Numerical representation	Categories
0	patient hasn't smoked before
1	patient has smoked

Numerical summaries

Variable name	Mean	Variance	Standard deviation	Maximum	Minimum
Gender	1.61	0.24	0.49	2	1
Age	48.65	355.37	18.85	82	10
Average glucose level	108.32	2275.63	47.7	271.74	55.12
BMI	30.29	53.23	7.3	92	11.5
Smoking status	0.46	0.25	0.5	1	0

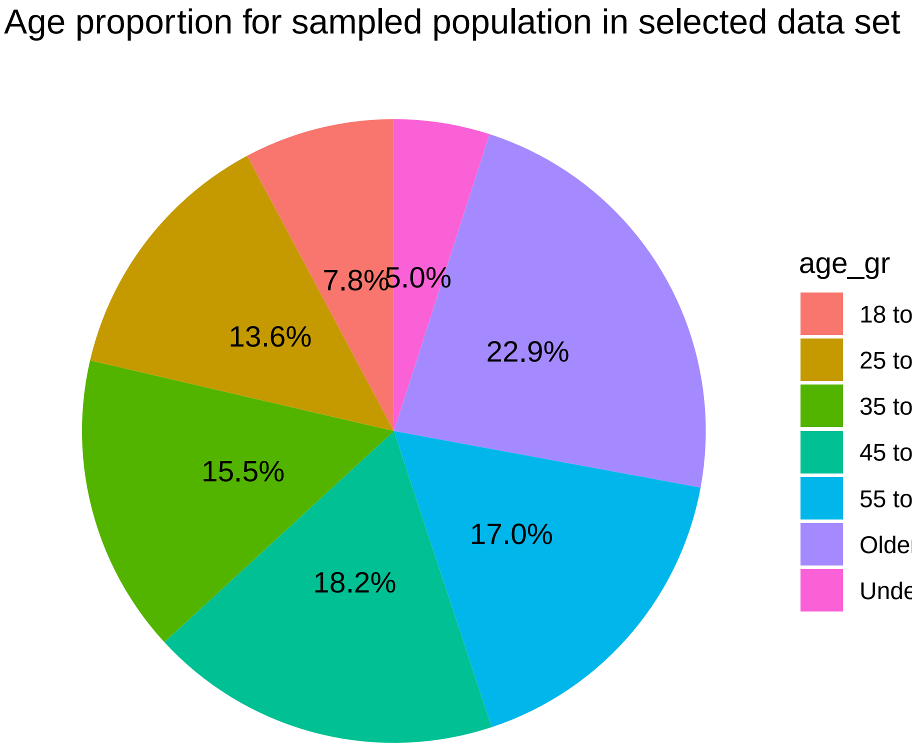


Figure 1

From the above barplot (Figure 1), there is an increasing pattern for the proportions of participants in each age level. Most of the participants concentrated at the ‘older than 65 years old’ level, whereas the age level under 18 has the smallest proportion of the participants. The overall proportion of the participants for age level from 25 to 64 years old is roughly uniformly distributed.

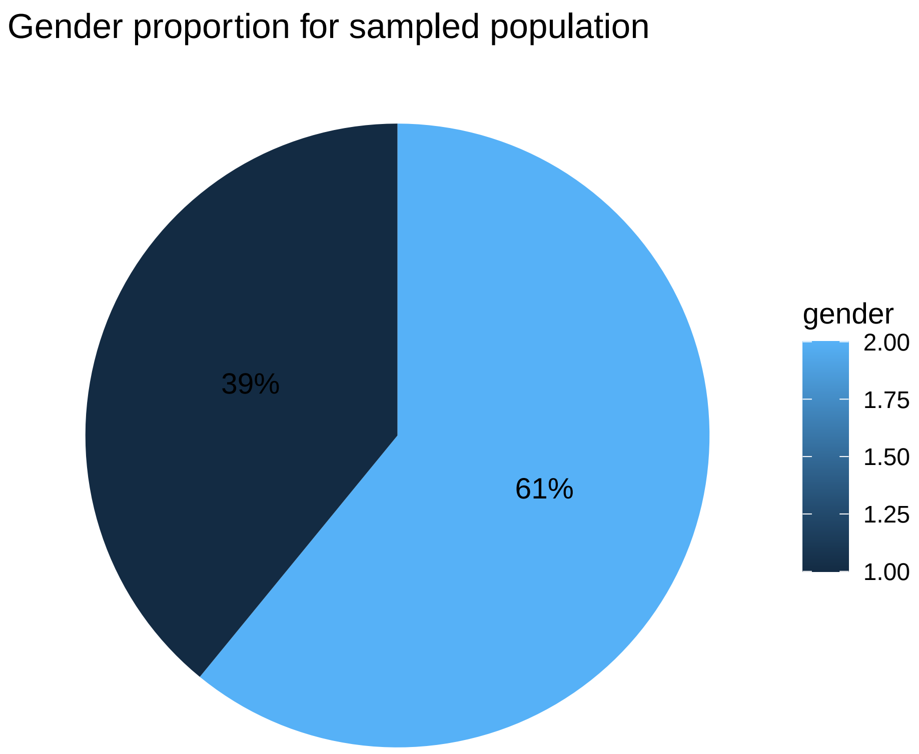


Figure 2

As we can see from the pie chart (Figure 2), the proportion of non-male participants is significantly larger than the male participants. 61% of the participants are non-male, while only 39% of the participants are male.

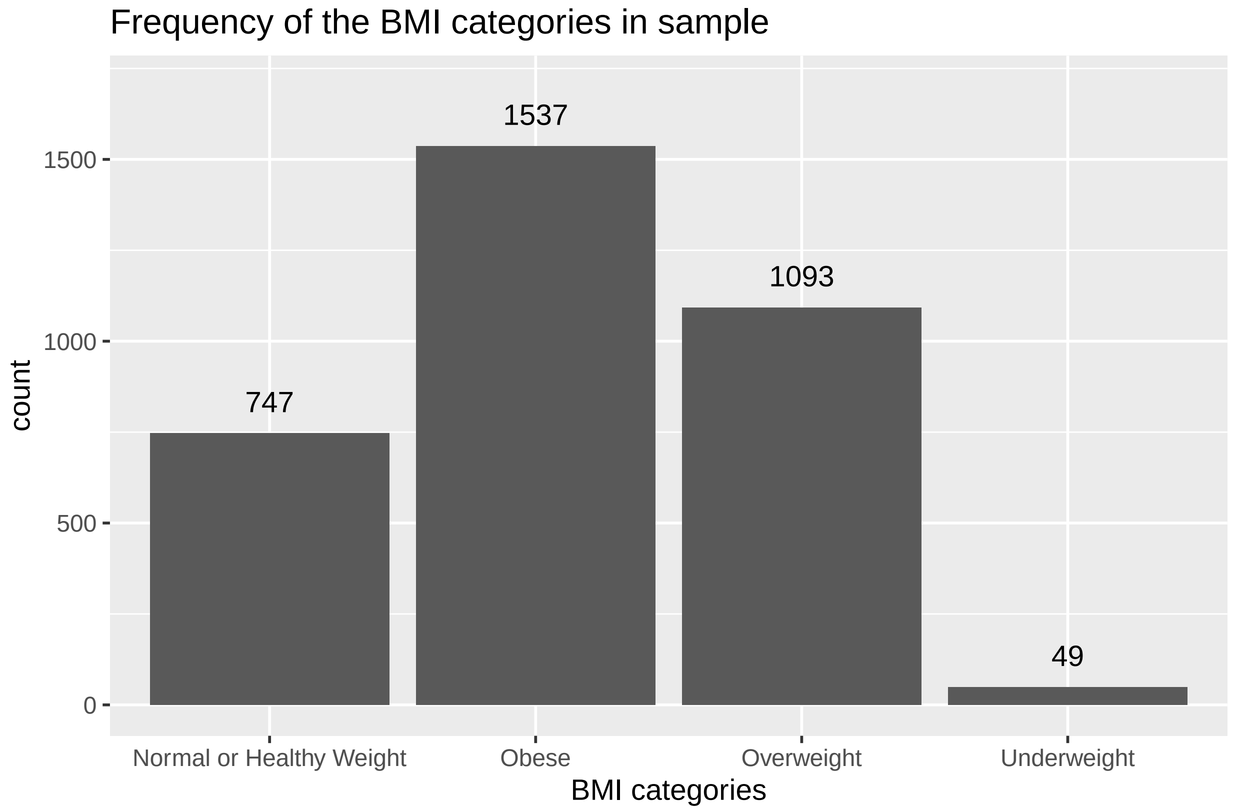


Figure 3

As we can see from the bar plot (Figure 3), the obese category takes most of the percentage of the total sample, which has 1537 participants in that group. The group for normal or healthy weight group has only 747 participants while the overweight group has 1093 participants. The BMI underweight category has the lowest proportion of participants, which only contains 49 participants.

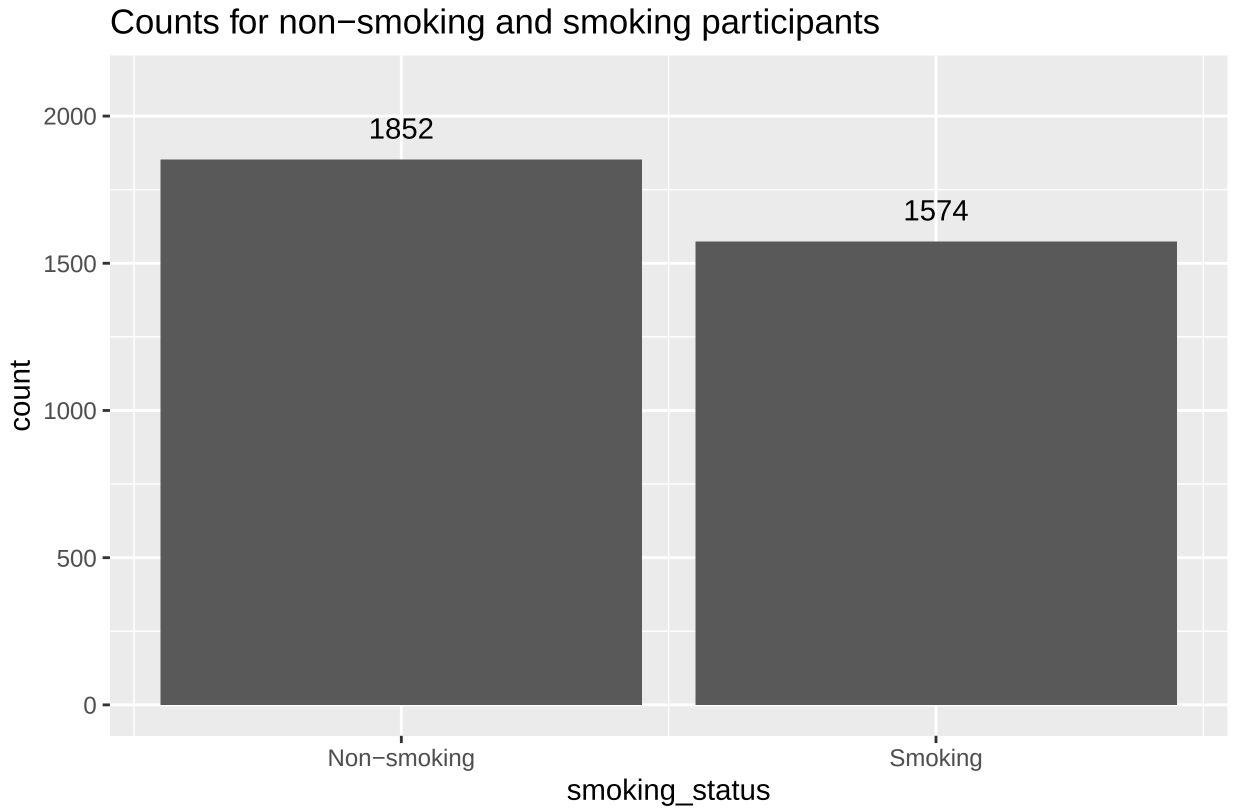


Figure 4

From Figure 4, there are significantly more non-smoking participants among the sampled population from our data set. There are 1852 participants in the non-smoking category while there are 1574 participants in the smoking category among the sampled population from our data.

From Figures 5 and 6 in the Appendix, we can see that most of the participants work for a private firm, there is only a slight proportion of participants who have never worked or are still children. There are approximately the same amount of participants living in rural and urban areas in our data set.

Methods

Constructing causal effects of independent variables on one or more dependent variables are usually processed under the control of several covariates. However, controlling the covariates across independent variables can be challenging, time-consuming, and unethical. For this particular study, we are interested in the causal relationship between smoking status and the BMI of the world population, it would be unethical to randomly assign the participant to a smoking treatment group and employ statistical analysis. Therefore, we would employ the propensity score matching technique to match corresponding observations from treatment groups and control groups in our data. We would then apply a multiple linear regression model to make statistical inferences about the significant level regarding the causation of our study. In the end, we would perform hypothesis testing and likelihood ratio test for the causal effect of smoking status upon the BMI of the participants.

Propensity score matching

Propensity score matching is most commonly used to match treatment groups and control groups according to several covariates. Treatment has been implemented before conducting the study. We can balance inequivalent groups using the propensity score to reduce the biasness of the treatment effect. (Rosenbaum & Rubin, 1983) Propensity scores are most commonly employed using logistic regression, in that case, the dependent variable is the treatment and control group dichotomous variable, and the independent variables are the covariates based on which the groups should be balanced. Logistic regression is used to describe the data between the binary response variable and one or more independent variables. In this case, we use logistic regression to forecast the propensity score of each observation based on their corresponding covariates. The logistic regression model for predicting propensity score is given by

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik},$$

where k is the number of covariates we used to predict the propensity score of each observation and i indicates a particular observation from a total of n observations. (Wu & Thompson, 2020)

It is essential to ensure the result treatment effect estimate is free of confounding and can be effectively interpreted as a causal effect by carefully select covariates. All chosen covariates must be measured before the treatment in order to estimate the total causal effects of the model. (Greifer, 2021) Our choice of covariates is gender, age, whether the participant has hypertension, whether the participant has heart disease, whether the participant has married, the participant's working type and region of living, and the average glucose level of the participant. The predicted probability of each participant falling in treatment or control groups can be calculated based on the logistic model we constructed. Since these propensity scores are the conditional probabilities of participants that would receive particular treatment based on their covariates, they can be applied as a measure of the distance between two observations from the observational data set. There are several assumptions we need to preserve for logistic regression. First, the binary logistic regression requires the response to be a binary outcome. The logistic regression also assumes the linearity between independent variables and the log odds of the event occurrence. (Huber et al., 2016)

The matching technique we would employ is the nearest neighbor matching without replacement. For each observation in the treatment group, the corresponding observation from the control group would be identified based on the propensity score we calculated from their covariates. When the distance of propensity score of two observations from the treatment group and control group is minimal, we would match them as a pair. We can estimate the treatment effects based on the nearest neighbor matches we generated in the matching section using a multiple linear regression model. The average treatment effect in the treated group is the average effect of the treatment for individuals who are actually in the treatment, which can be estimated based on our matches. Multiple linear regression is a statistical technique that employs several expansionary variables to predict the outcome of the response variable. The multiple linear regression model we used for our model is given as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i,$$

where i represents each observations in total n observations and p represents the number of independent variables. The assumptions of multiple linear regression are similar to simple linear regression. We firstly assume the linearity of the regression model, which means that the mean of the response variable is a linear combination of the regression coefficients estimates and the predictors. Since the predictors' values are fixed, the linearity assumption only works on the regression coefficients of the model. The next assumption is homoscedasticity, which means the variances of the random errors do not depend on the values of predictors. Furthermore, the assumption of independence of errors provides the fact that the random errors of the response variables have no relationship with each other. (Wu & Thompson, 2020)

Hypothesis testing

According to my goal of the study of this research, I would like to analyze there is a statistically significant causal relationship between the smoking status of a participant to an increase of BMI of the participant. I would perform a t-test to the treatment group of smoking status (the smoking participants) in the multiple linear regression model constructed for testing the hypothesis. I would like to set $\alpha = 0.05$ as the significant level for this hypothesis test.

Null hypothesis: There is no statistical significance for the sampled participants that smoking would have a direct impact on increasing body mass index.

Alternative hypothesis: There is statistical significance for the sampled participants that smoking would have a direct influence on increasing body mass index.

Results

General results

We employed propensity score matching to estimate the average treatment effect between the smoking treatment and the non-smoking controlled groups on the sampled participants based on the covariates we selected from the participants. We first constructed a no-matching model and employ the same covariates to construct the nearest neighbor matching based on propensity score in order to check the imbalance of the original data set. Both matching methods are implemented based on propensity score distance of observations by using logistic generalized linear regression to predict. There is a significant improvement in covariate balance from the nearest neighbor matching approach without replacement, which we can observe from the summary of the matching table. There are 1574 participants in the treated group and all of them are matched with observations from the controlled group. Only 278 observations from the controlled group are unmatched. The quality of the match is also assessed using the distribution of propensity scores plot, eQQ plots, and Love plots. (Greifer, 2021) The hypothesis of this study is tested by using a t-test in the multiple linear regression model constructed based on matching.

The R code for propensity score matching is partially cited from *Matchit: Getting started.* (Greifer, 2021)

Sample sizes

	Control observations	Treated observations
All	1852	1574
Matched	1574	1574
Unmatched	278	0
Discarded	0	0

Analyzing the imbalance of the data by comparing no-match and nearest-neighbor matching without replacement

(graph here)

We would use the summary table of the nearest neighbor matching approach (without replacement) to analyze the improvement of balance from the matching technique. As we can see from the table above, there is a significant improvement in the balance for our data set, which we can observe from standardized mean differences (Std. Mean Diff), variance ratios (Var. Ratio), and empirical cumulative density function (eCDF).

The standardized mean difference is one of the most widely used for assessing the balance of data set after propensity score match, it is the difference in the means of each covariate between treatment groups under the same scale. (Greifer, 2021) A Std. Mean Diff closes to zero suggests a good balance of the match. We could observe that Std. Mean Diffs of covariates for the matched data are all around or lower than 0.1, the maximum of that value is from gender, indicating that match is slightly off for that covariate.

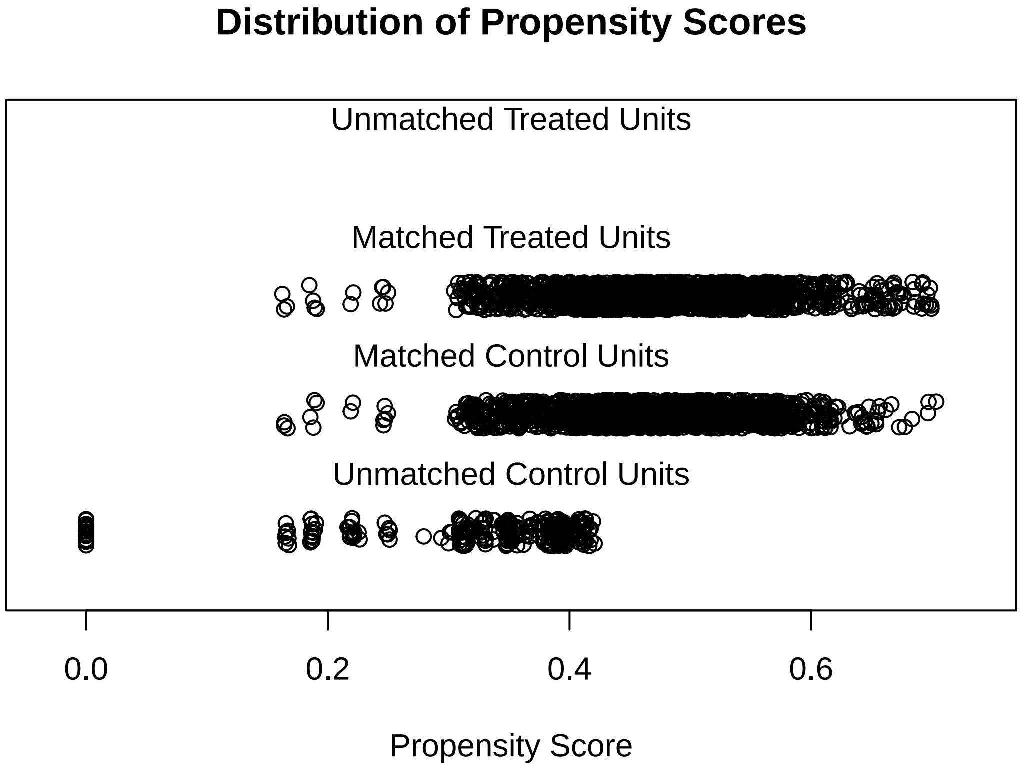
Variance ratios measure the ratio of the variance of a covariate in one group to the variance in the other group. A Var. Ratio that is close to 1 suggests a good balance because the variances of matched pairs are similar to each other. (Greifer, 2021) Thus, from the summary of balance for matched data, we could observe a pretty balanced match of the data as the variance ratios of covariates are mostly centered around 1.

The empirical cumulative density function indicates the difference in the overall distribution of the covariates between treated groups, which ranges from 0 to 1. A value that is closer to 0 indicates a good balance. (Greifer, 2021) Although there is no specific benchmark for us to examine the balance based on this parameter, a noticeably large number needs to be addressed. The result from the summary table seems appropriate, only eCDF means of covariates age, gender, and heart disease are slightly off.

(graph here)

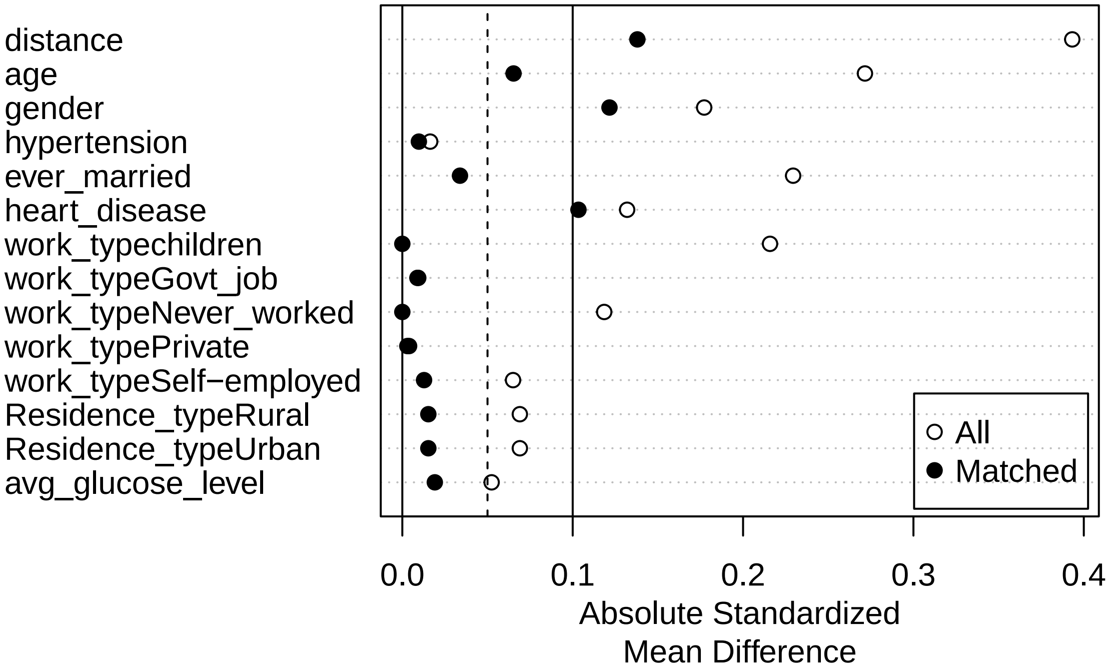
The above percent balance improvement table compares statistical parameters of summaries of matched data from non-match and nearest-neighbor matching without replacement. As we can observe that there is more or less improvement of balance after matching our data. The covariate categories work type as children and work type as never worked have 100% improvement in the balance measurement. Only covariate work type as private jobs has decreased the balance of the match.

Graphical interpretation



The distribution of propensity scores presents treated and control units after matching with the corresponding propensity score. The matching left 278 unmatched, which are mostly concentrated around propensity score 0.3 and some are away center of the data set (around propensity score 0). We can see that there is a high concentration of controlled observation around the propensity score of 0.3 in the matched control units. Since all treatment observations are matched with a controlled observation, it might just due to a high concentration of data that causes unmatched control units.

For the eQQ plots in the Appendix, the y-axis indicates values of the covariates for the treatment observations, the x-axis displays the value of the covariate of the corresponding quantile in the controlled group. The concentration of data points falling onto the solid diagonal line indicates the proportion of balance between two groups. (Greifer, 2021) I selected four essential covariates (age, average glucose level, type of work) in the matching and plot the above eQQ plots. As we can observe, there is a significant improvement in age and average glucose level as the concentration of the data points for the two plots are mostly on the diagonal line. The matching has a moderate improvement of balance between two groups for work types as children, never worked, and self-employed.



The Love plots are a straightforward way to summarize the balance of matching visually. It's the balance is fairly improved after the nearest-neighbor matching, more than half of the covariates have absolute standardized mean difference less than the threshold 0.1. The huge improvement can be seen by the distances between absolute standardized mean differences from all and matched for each covariate.

Interpretation of multiple linear regression model and performing hypothesis testing

```
##
## Call:
## lm(formula = bmi ~ as.factor(smoking_status) + age + as.factor(gender) +
##     as.factor(hypertension) + as.factor(ever_married) + as.factor(heart_disease) +
##     as.factor(work_type) + as.factor(Residence_type) + avg_glucose_level,
##     data = survey_near_match_data, weights = weights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.821  -4.875  -0.973   3.606  59.617
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       20.729266   1.374174  15.085  < 2e-16 ***
## as.factor(smoking_status)1         0.280401   0.251030   1.117    0.264    
## age                               -0.038924   0.008801  -4.423 1.01e-05 ***
## as.factor(gender)2                -0.018209   0.256662  -0.071    0.943    
## as.factor(hypertension)1           2.286542   0.397094   5.758 9.32e-09 ***
## as.factor(ever_married)1           2.180308   0.355964   6.125 1.02e-09 ***
## as.factor(heart_disease)1         -0.747978   0.529910  -1.412    0.158    
## as.factor(work_type)Govt_job       7.395624   1.406264   5.259 1.55e-07 ***
## as.factor(work_type)Private        7.383693   1.370805   5.386 7.72e-08 ***
## as.factor(work_type)Self-employed  7.154459   1.413054   5.063 4.36e-07 ***
## as.factor(Residence_type)Urban    -0.155200   0.250610  -0.619    0.536    
## avg_glucose_level                  0.022539   0.002671   8.438  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.017 on 3136 degrees of freedom
## Multiple R-squared:  0.05864,    Adjusted R-squared:  0.05534 
## F-statistic: 17.76 on 11 and 3136 DF,  p-value: < 2.2e-16
```

The regression coefficient estimates of the multiple linear regression model we chose indicate the correlation between the predictors (age, smoking condition, gender, hypertension, marriage, heart disease, types of work, region of living, average glucose level) in our data set and participant's body mass index. According to the table, participant's age, average glucose level, the occurrence of hypertension and marriage, and working types as a government job, private job, self-employed job have statistical significance for predicting the BMI of the participants. The overall statistical significance of the model for all the predictors is extremely small (2.2e-16) while the intercept of the multiple linear regression model is 20.7292. The regression coefficient estimate for the participant has smoking pattern is 0.2804.

```
## 
## t test of coefficients:
## 
##                                     Estimate Std. Error t value  Pr(>|t|)    
## (Intercept)                       20.7292657  0.8685148 23.8675 < 2.2e-16 ***
## as.factor(smoking_status)1         0.2804011  0.2526321  1.1099   0.26712    
## age                               -0.0389241  0.0087414 -4.4529 8.769e-06 ***
## as.factor(gender)2                -0.0182092  0.2532503 -0.0719   0.94268    
## as.factor(hypertension)1           2.2865417  0.4469602  5.1158 3.313e-07 ***
## as.factor(ever_married)1           2.1803078  0.3827410  5.6966 1.336e-08 ***
## as.factor(heart_disease)1         -0.7479777  0.4197250 -1.7821   0.07484 .  
## as.factor(work_type)Govt_job       7.3956244  0.9221722  8.0198 1.481e-15 ***
## as.factor(work_type)Private        7.3836934  0.8697640  8.4893 < 2.2e-16 ***
## as.factor(work_type)Self-employed  7.1544592  0.9265930  7.7213 1.538e-14 ***
## as.factor(Residence_type)Urban    -0.1552003  0.2566892 -0.6046   0.54547    
## avg_glucose_level                  0.0225388  0.0028988  7.7753 1.013e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis states that there is no statistical significance for the sampled participants that smoking would have a direct impact on increasing body mass index. Based on the result of the t-test for the participant who currently has a smoking pattern, the smoking variable as a predictor has no significant impact on predicting the BMI of the participant. The P_value for the t-test of smoking category in smoking status is 0.264, which is a bit higher than the 0.05 α level of significance. Thus, we would have not enough evidence to reject the null hypothesis via the t-test result based on the multiple linear regression model. Therefore, we can't conclude that smoking has a direct impact to an increase of BMI of the participant.

Conclusions

The overall goal of this study is to determine whether there is a causal relationship between smoking status and the BMI of the participants. Smoking, one of the most critical dangers that challenge people's life, should be studied thoroughly in order to determine what could be the potential influences under different smoking conditions. The study could also suggest the government enforce proper restrictive policies to adjust the consumption of tobacco products in the market. In this study, I proposed the hypothesis that whether there is a statistical significance to the sampled participants in which smoking would directly increase the body mass index of the participant. I employed the SPD data set and used the demographic information, various disease information to construct a nearest-neighbor propensity score matching without replacement for the smoking treatment group and non-smoking controlled group. The propensity score matching offers an opportunity to derive a causal relationship based on observational data set, in this case, our study would be unethical to implement for the treatment group. It also results a well-balanced and less biased data for comparison. The prediction model is constructed through multiple linear regression based on the matched pairs and various background information as predictors. Participant's age, average glucose level, the occurrence of hypertension and marriage, and working types as a government job, private job, self-employed job have statistical significance for predicting the BMI of the participants. Although the regression coefficient estimate of smoking (0.2804) has indicated a positive trend that smoking could potentially increase the BMI of the participant, the t-test result does not provide a sufficient condition for us to reject the null hypothesis. As far as I'm concerned, the result is reasonable since our data set does not offer more demographic information for matching and the categories of the smoking status of the participants are insufficient. Some studies indicate conflicting results for the relationship between smoking status and BMI of the participants. While some studies have shown no significant association between smoking patterns and body mass index (BMI), others have suggested that smoking might be associated with lower BMI due to higher energy expenditure and suppressing appetite for smokers. (Dare et al., 2015)

Weaknesses

In this study, the Stroke Prediction Dataset has few variables for each observation. Some of the important information of the participants such as region of living, education level, race, etc are not measured in this data set. Thus, the observations matched from treatment and control groups based on propensity score matching could potentially be less precise. The number of observations from our data set is also insufficient for causal analysis. There are 278 controlled group observations unmatched since there are significantly more smokers in the cleaned data set compared to non-smokers. This is also a drawback of propensity score matching as it seriously reduces the observations we can analyze. On the other hand, the unobserved information that has a direct impact on the treatment is not considered in the propensity score matching process, which could cause the hidden bias to remain in the data set after the matching procedure. (Garrido et al., 2014) Also, during the data cleaning process, I mutate the 'has smoked' 'smoking' categories from the smoking status to 'has smoking pattern' and mutate only 'never smoked' as non-smoking since I need to construct binary categories for calculating propensity score using logistic regression. The reduction of categories could lead to a reduction of data and the ambiguous meaning of the variable.

Next Steps

For further study, I would suggest collecting more demographic information about the participants to get a more accurate match. The balance in numbers of smokers and non-smokers is also important to address so that the number of unmatched observations would decrease, in other words, the data is applied more efficiently. The result of no significant association between smoking patterns and BMI might due to unclear categories of the smoking pattern. For instance, if we can specifically record how much tobacco the participant consumes each day and collect more observations, there could be a more distinct result for the association.

Discussion

In conclusion, the multiple linear regression model I constructed based on the Stroke Prediction Dataset indicates a positive trend for the BMI of the participant is the person has a smoking pattern. The study is essential for people that have a smoking pattern to adjust their lifestyle. It also suggests governments propose appropriate restrictive policies to monitor and solve the health issues of their citizens.

Bibliography

Fedesoriano. (2021, January 26). Stroke prediction dataset. Retrieved June 14, 2021, from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Huber, S., Dietrich, J. F., Nagengast, B., & Moeller, K. (2016). Using propensity score matching to construct experimental stimuli. *Behavior Research Methods, 49*(3), 1107-1119. doi:10.3758/s13428-016-0771-8

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,*41–55. doi:10.2307/2335942

Doll, R., Peto, R., Boreham, J., & Sutherland, I. (2004). Mortality in relation to smoking: 50 Years' observations on male British doctors. *BMJ, 328*(7455). doi:10.1136/bmj.38142.554479.ae

Dare, S., Mackay, D. F., & Pell, J. P. (2015). Relationship between smoking and Obesity: A cross-sectional study of 499,504 middle-aged adults in the UK general population. *PLOS ONE, 10*(4). doi:10.1371/journal.pone.0123579

Peeters, A., Barendregt, J. J., Willekens, F., Mackenbach, J. P., Mamun, A. A., & Bonneux, L. (2003). Obesity in adulthood and its consequences for life EXPECTANCY: A Life-Table Analysis. *Annals of Internal Medicine, 138*(1), 24. doi:10.7326/0003-4819-138-1-200301070-00008

About adult bmi. (2020, September 17). Retrieved June 14, 2021, from https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Greifer, N. (2021, May 26). Matchit: Getting started. Retrieved June 16, 2021, from https://cran.r-project.org/web/packages/MatchIt/vignettes/MatchIt.html

Wu, C., & Thompson, M. E. (2020). *Sampling Theory and Practice.* Springer International Publishing. (Last Accessed: June 16, 2021)

Greifer, N. (2021, May 26). Assessing balance. Retrieved June 18, 2021, from https://cran.r-project.org/web/packages/MatchIt/vignettes/assessing-balance.html

Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., & Aldridge, M. D. (2014). Methods for constructing and assessing propensity scores. *Health services research, 49*(5), 1701–1720. https://doi.org/10.1111/1475-6773.12182

Appendix

Supplementary Data

A glimpse of the original observational data set

```
## Rows: 5,110
## Columns: 12
## $ id                <dbl> 9046, 51676, 31112, 60182, 1665, 56669, 53882, 10434~
## $ gender            <chr> "Male", "Female", "Male", "Female", "Female", "Male"~
## $ age               <dbl> 67, 61, 80, 49, 79, 81, 74, 69, 59, 78, 81, 61, 54, ~
## $ hypertension      <dbl> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1~
## $ heart_disease     <dbl> 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0~
## $ ever_married      <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No~
## $ work_type         <chr> "Private", "Self-employed", "Private", "Private", "S~
## $ Residence_type    <chr> "Urban", "Rural", "Rural", "Urban", "Rural", "Urban"~
## $ avg_glucose_level <dbl> 228.69, 202.21, 105.92, 171.23, 174.12, 186.21, 70.0~
## $ bmi               <chr> "36.6", "N/A", "32.5", "34.4", "24", "29", "27.4", "~
## $ smoking_status    <chr> "formerly smoked", "never smoked", "never smoked", "~
## $ stroke            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

Link for the Stroke Prediction Data that I employed for the study:

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

Supplementary Plots

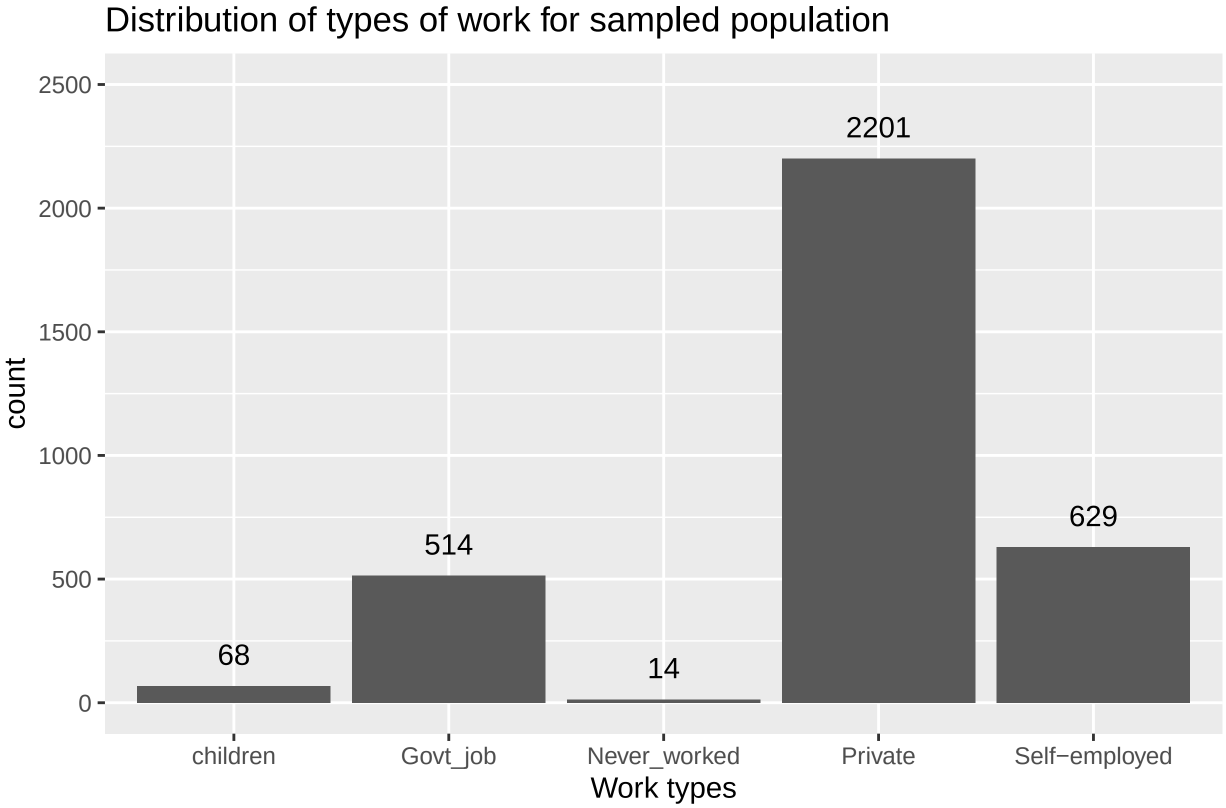


Figure 5

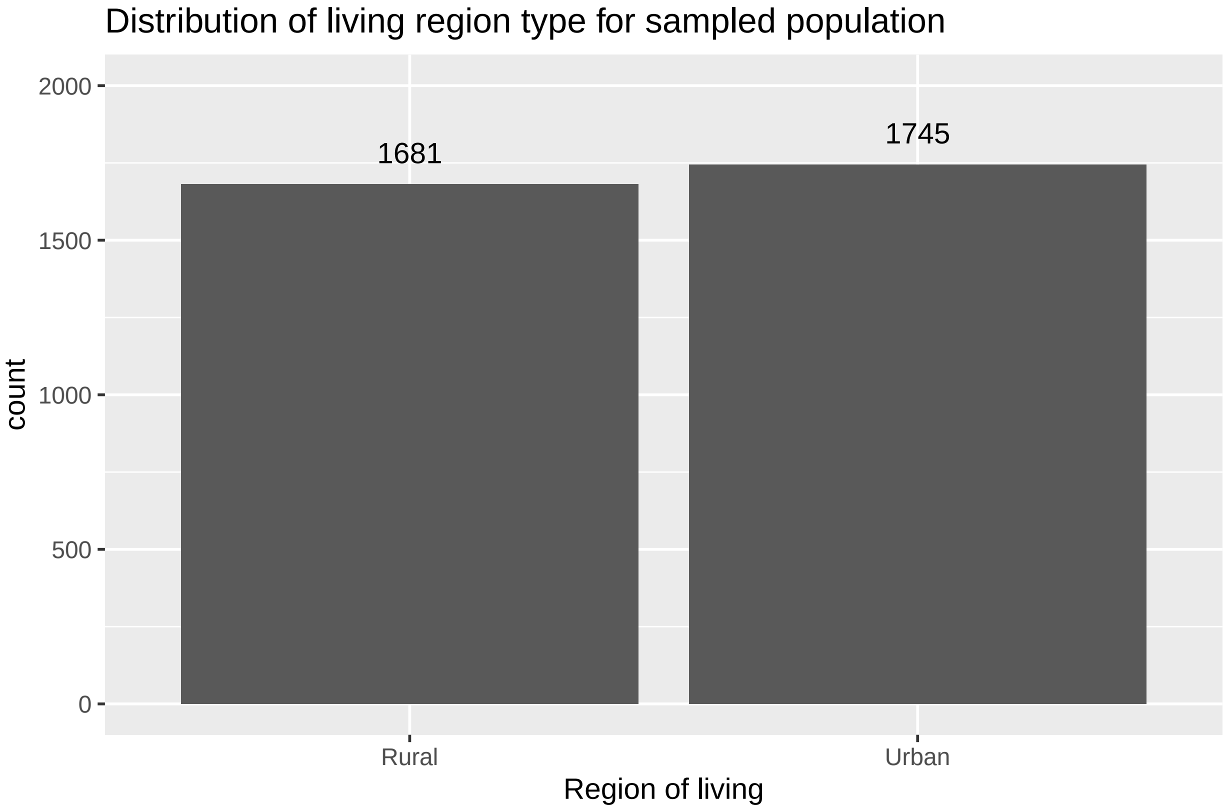
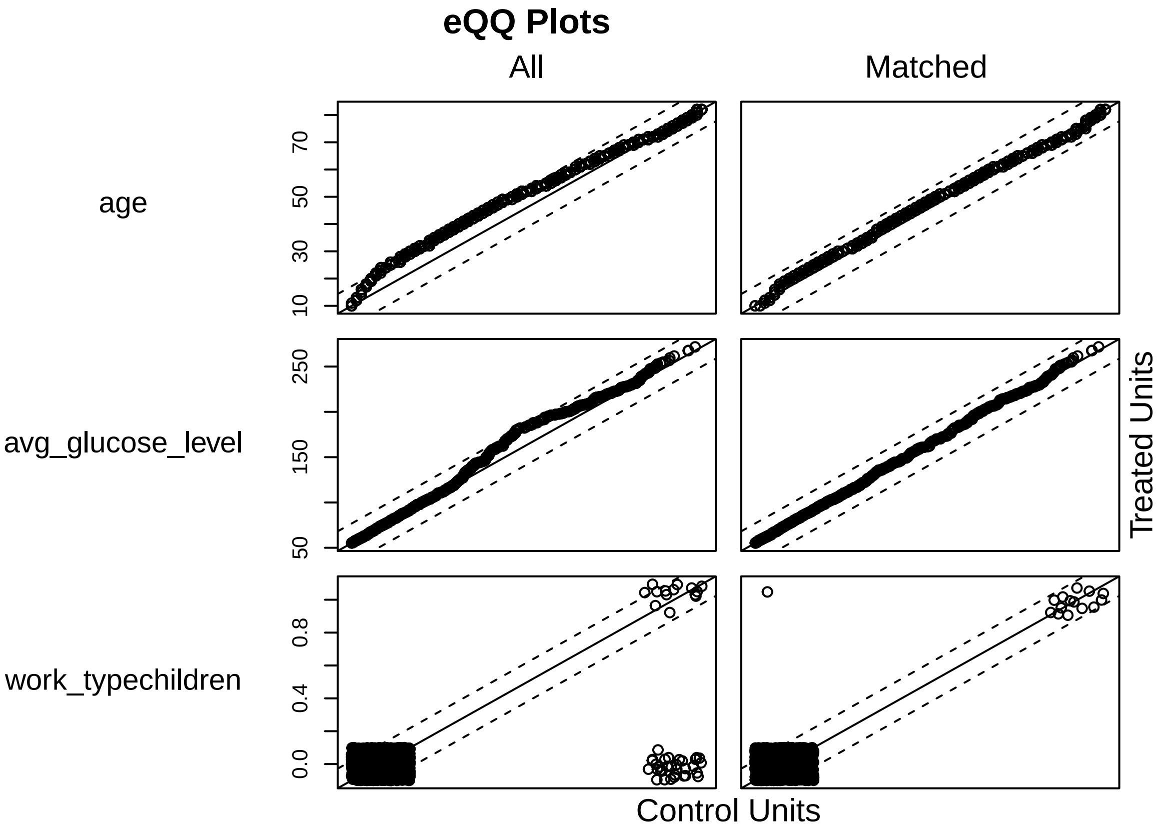
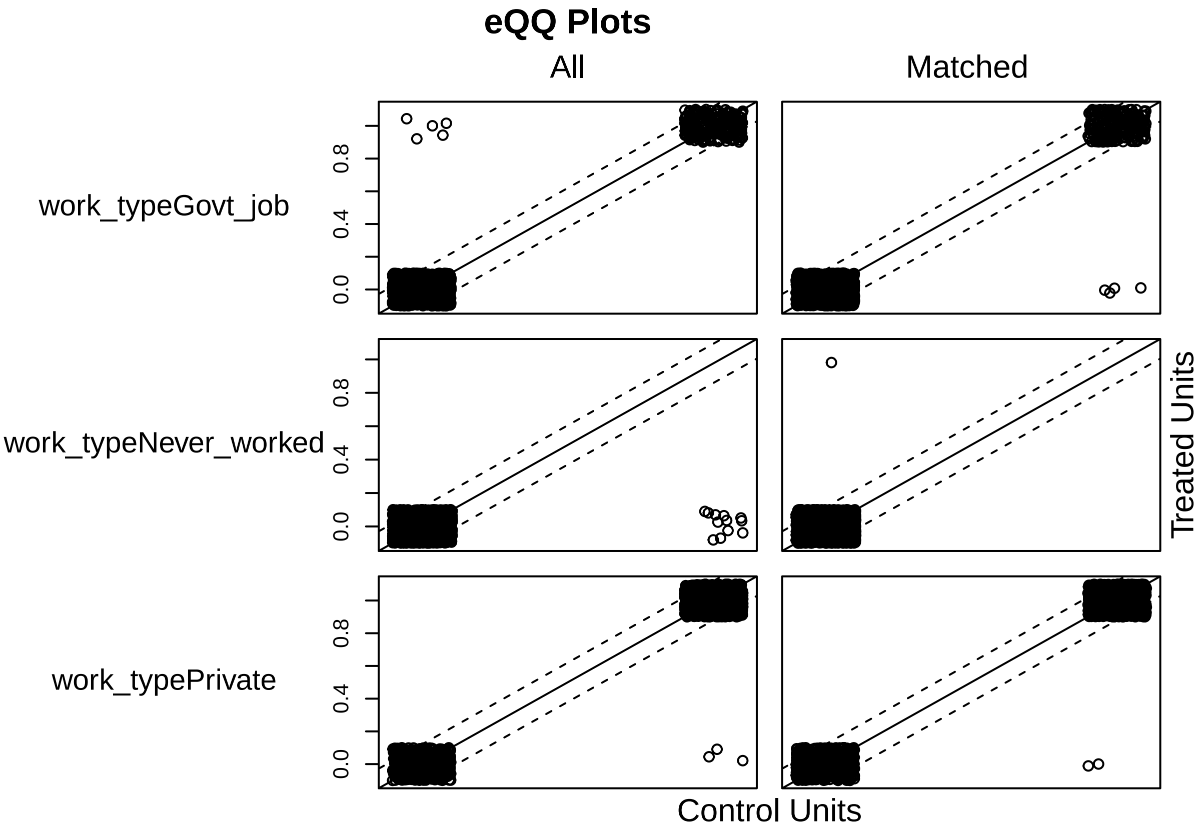
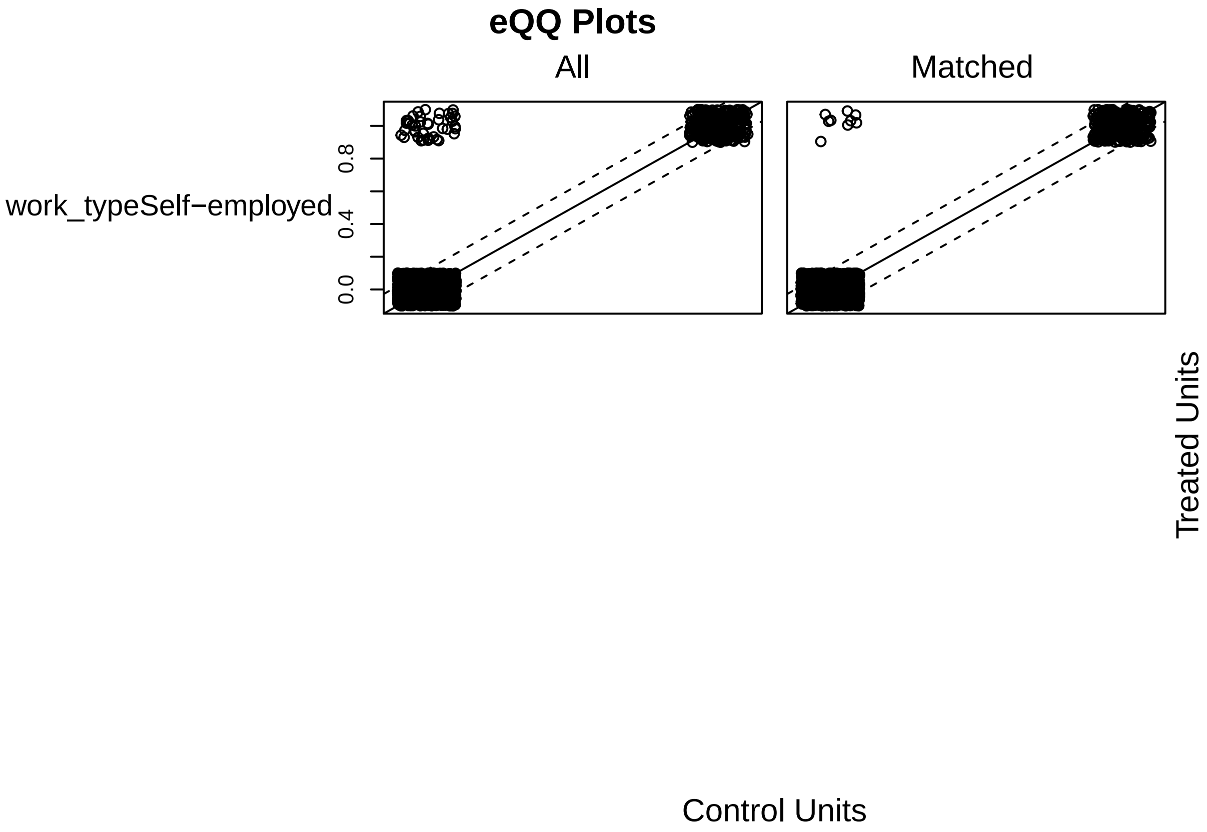


Figure 6


eQQ Plots
All
Matched
age
10
30
50
70
avg_glucose_level
50
150
250
work_typechildren
0.0
0.4
0.8
Treated Units
Control Units


eQQ Plots
All
Matched
work_typeGovt_job
work_typeNever_worked
work_typePrivate
0.0
0.4
0.8
Treated Units
Control Units


eQQ Plots
All
Matched
work_typeSelf−employed
0.0
0.4
0.8
Treated Units
Control Units