A Comparison of Propensity Score and Logistic Regression Methods

Faith Lembemo (faithlembemo@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Professor Arne Ring
University of the Free State, South Africa

14 May 2020

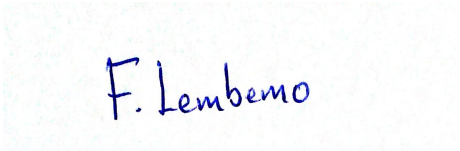
Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

ABSTRACT

When it comes to research in medicine, epidemiology, randomized controlled trials are deemed to be the gold standard of research, but it is not always the case that such studies are feasible. Observational studies are used as alternatives, however, in such studies we face the issue of bias due to uncontrolled covariates. In this essay we look at how the effect of confounding variables is controlled by using propensity scores and logistic regression. We aim to check if the analysis of using both methods leads to similar results. The data set used in this study is from an ongoing cardiovascular cohort study that started in 1948 with 5209 adult participants from the city of Framingham, Massachusetts, from the data set we are going to check whether smoking leads to the development of Coronary Heart Disease (CHD). Missing data were deleted from the data set and we remained with 5039 participants. The backward elimination method was used to select a model to be used in logistic regression. In the propensity score method, we used logistic regression to estimate the propensity scores, the estimated propensity scores were grouped into smokers group and non-smokers group, propensity matching was used to balance the group means so that there are no baseline differences within the groups. The quality of matching was checked by using the significance and standardized difference percentage of the baseline characteristics. We used logistic regression to check on the effect of the exposure variable on the outcome variable. In both methods, the results indicated that smoking leads to the development of CHD. We made all codes used in the analyses available and accessible online.

DECLARATION

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.



Faith Lembemo, 14 May 2020

Contents

1. Introduction

The goal of most statistical tasks is to study variation and find unbiased estimates of the data set, by making sure that all confounding variables are accounted for. This can be achieved by using appropriate methods for data collection and analysis. Randomised controlled trials have been deemed to be the gold standard in research designs, because of their reduction in allocation bias, which comes from baseline characteristics that may influence outcomes (Sullivan, 2011). However, it is not always the case that such experimental designs are possible, practical, or desired in the social and health sciences. Therefore, observational or non-experimental designs are used as alternatives.

In observational studies, randomisation is not used when assigning treatment to study participants, assignments of treatment are based on the characteristics of the study participant. This nonrandomness in the assignment of treatment results in imbalanced baseline characteristics which might be confounded with treatment effects, hence making wrong estimates (Zou et al., 2016). Statisticians have worked on finding different methods to solve such cases, that is by controlling the effect of confounding variables to variables of interest. Many ways are used to control confounding variables in statistics. In this essay we are going to focus on how propensity scores and logistic regression can be used to control the effect of confounders for us to achieve unbiased treatment effect estimates.

1.1 Main research question

- Do adjustments using covariates in logistic regression lead to similar results as propensity scores?

1.2 Specific research questions

1. What are logistic regression and propensity score methods?
2. When do we use logistic regression or propensity score methods?
3. How to analyse propensity score and logistic regression methods using a similar data set?

1.3 Significance of the study

The purpose of this essay is to help researchers and future students in statistics, on how best they can analyse observational studies using the methods discussed in this essay. It is also a well-known fact in theory that propensity scores are the best methods to use when analysing observational studies than logistic regression (Fechtner, 2018), but it is the interest of this essay to find out how much difference do the results have.

2. Literature review

The data set used in this essay was used in the Fechtner (2018) paper, where he defined what is propensity score matching and analysed the impact of smoking on the development of coronary heart disease by controlling covariates. Propensity score matching was used to reduce bias caused by non-randomisation. To check the quality of matching, weighted chi-square and t-test were used. The paper also gave methods on how propensity scores can be implemented in R and SAS and provided codes for both packages. The codes on propensity scores for this essay have been developed from the Fechtner (2018) codes in R. In this chapter, we will look at the handling of missing data, selecting of appropriate models, logistic regression methods and propensity scores method.

2.1 Handling missing data

Missing data is a common problem in epidemiology, medical studies, and social science surveys, where participants of the study fail to provide data due to different reasons such as death, migration, or lack of interest. When a dataset has a lot of missing data, it may result in having biased estimates and making wrong inferences. There are three types of missing data patterns: Missing at Random(MAR), Missing Completely At Random (MCAR), and Missing Not At Random (MNAR). In MCAR pattern missingness is totally at random, and the relationship between missing data and observed data is not seen. In MAR pattern missingness happens due to not including a certain variable, it's pattern is related to the observed values. While MNAR is neither MCAR or MAR, it is related to both observed data and missing data. Due to this, statisticians have looked at strategies of handling missing data by looking at the patterns, before any analysis is done.

According to Rässler et al. (2007), there are four groups of strategies for analyzing missing data. The first group consists of simple methods that are used to deal with missing data, for example, complete case analysis which is also known as listwise deletion, and available case analysis, this removes the units that have incomplete data. He further explains that even though these methods are simple to implement, they lead to inefficient and biased estimates when the percentage of missing data in each variable is high. It is noted that a data set with less than 5% missing data in each variable of interest, list deletion or pair deletion is used (Zwonitzer et al., 2016), in most cases this method is used for MCAR.

The second group consists of weighting procedures, it increases the survey weights for units that gave responses to account for those units that did not respond which will be removed they are removed from further analysis. The third group consists of imputation-based procedures, a standard approach to handling missing data. It fills in all values that are missing and the resulting analysis has no missing values. In the same group, we also have multiple imputation method, which deals with reflecting on the added uncertainty since the values that are entered are not the actual values and it allows the use of complete-data analysis methods. Lastly, the fourth group consists of direct analysis using model-based procedures, the models for the observed data are specified and inferences or conclusions are based on Bayesian analysis or likelihood. According to Rässler et al. (2007), only multiple imputations and direct analysis can result in making valid inferences and are the most useful approaches in epidemiology and medical databases.

2.2 Logistic regression

Multiple regression models are frequently used in observational studies to assess the effect of the variable of interest on the dependent variable while controlling for one or more covariates and its dependent variable is continuous (interval scale). Logistic regression is similar to multiple regression but its dependent variable is dichotomous (Tranmer and Elliot, 2008).

When we have a proportion as a response, we use logistic or logit models to link the dependent variable to the set of independent variables. The logit link has the form:

$$\text{Logit } (\mathrm{P}(X_i)) = \ln \left(\frac{\mathrm{P}(X_i)}{1 - \mathrm{P}(X_i)} \right), \tag{2.1}$$

where, $\mathrm{P}(X_i)$ denotes the odds of an event occurring at value X_i. The odds of failure are the probability of an event not happening divided by the probability of an event happening(Agresti, 2018).

The logistic regression model can be written as:

$$\ln \left(\frac{\mathrm{P}(X_i)}{1 - \mathrm{P}(X_i)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m, \tag{2.2}$$

where, $X_1, X_2, \ldots, X_m$ are the predictor variables, m is the number of variables being considered, the parameter β_0 gives the log odds, $\beta_1, \beta_2 \ldots, \beta_m$ are constant terms representing unknown parameters. The logistic model is designed to describe a probability that is always within 0 and 1.

In logistic regression, we don't use the β coefficients to check the relationship between the independent variable and the dependent, instead, we use odds ratios (OR), which is the result of dividing the odds of success by odds of failure or vise-versa. From an R output we find the odds ratio by finding the exponent of the coefficients i.e., e^{β} and interpret the odds ratio as follows;

when the $OR < 1$ it is less in favor and when the $OR > 1$ it is more in favor.

2.3 Model selection

When conducting a logistic regression or any type of regression analysis, the main focus is on the selection of appropriate variables to be included in the model and used for controlling confounding. However the selection process becomes a problem when the number of explanatory variables is large (Heinze et al., 2018). In such case, there are many methods that are used to select an appropriate model to explain the variability in the outcome variable. For example, basing on previous work in similar studies or using statistical methods such as stepwise variable selection algorithm and the Akaike Information Criterion (AIC) method.

2.3.1 Stepwise variable selection algorithm.

The stepwise variable selection algorithm is the combination of backward elimination and forward elimination methods. The backward elimination method starts with a complex model and sequentially drops the variables (Agresti, 2018). In doing so, all independent variables are included in the regression model and at each step, variables that are not significant are dropped. The best model is detected when all variables in the regression output are significant. For a dataset that has a continuous dependent variable, linear regression model is used, but when the dependent variable is binary, the logistic regression model is used to select the variables.

Agresti (2018) further explains that including a variable in a model should not sorely depend on the statistical significance of a variable. Some variables can be included in the model when they are considered important for the purpose of the study. When such variables are kept in the model they help to reduce bias in estimating the effect of other explanatory variables and can also be used to compare results from other studies where the effect is significant.

Forward elimination is the reverse of backward elimination method. Instead of dropping a variable at each step, we begin with a model that only has the one explanatory variable. The other variables are added sequentially to the model until we can not find any variable that provides strong evidence of their importance in the model (Agresti, 2018).

2.3.2 Akaike information criterion (AIC) method.

The AIC model is known to be the best in model selection, it selects the best model by comparing the values of the simple model (known as fitted values) with the complex model values (known as expected values), the latter contain all the covariates. The best model is the one that has its fitted values close to the true outcome probabilities (Agresti, 2018) and the model reduces bias and variance at the same time. The following equation is used to calculate the value of AIC.

$$\text{AIC} = -2(\text{log L- K}) \tag{2.3}$$

Where K is the number of estimated parameters in the model and L is the maximized likelihood function for the estimated parameters.

When then the value of AIC is small it indicates a good model.In this essay, we are going to use the backward regression method to select the covariates due to time constraits for the essay period.

2.4 Propensity score method

Propensity scores method is a tool proposed by Rosenbaum and Rubin in 1983(Fechtner, 2018), it has been used widely to adjust for confounding variables in the statistical analysis of observational or nonexperimental data, where all confounding variables are observed and included in the propensity score model. According to Elze et al. (2017) in other models like regression where all relevant participant characteristics are included, there is a common citation concern: when there are large numbers of covariates, it results in overfitting of the model (covariates are characteristics of study participants). However when using the propensity score model, it summarizes all participant's characteristics into one covariate, hence, reducing the possibility of overfitting.

A propensity score E_i is defined as the conditional probability of a study participant being assigned to receive specific treatment Z_i, given a vector of observed covariates $X_i,\ i = 1, ..., n$

$$E_i = \text{P}(Z_i|X_i = x_i) = \frac{e^{x_i\zeta_i}}{1 + e^{x_i\zeta_i}} = \frac{1}{1 + e^{-x_i\zeta_i}} \implies log\left(\frac{E_i}{1 - E_i}\right) = x_i\zeta_i, \tag{2.4}$$

where, Z_i denotes the binary treatment condition ($Z_i = 1$, if a case is in the treatment group, and $Z_i = 0$, if the case is in control group), ζ_i denotes the vector of regression parameters and n the number of subjects in the study. The covariates X_i are variables which are associated as potential confounders regarding the main analysis. The probability E_i can be estimated by using the maximum likelihood

estimator of ζ_i resulting from the logistic regression model and considering the predefined potential confounders X_i.

A pair of treated and control groups having the same propensity score are essentially viewed as comparable, even though they might have different values of specific covariates. When propensity score methods are applied appropriately they can help to solve problems of selection bias and provide reasonable estimates of average treatment effect (Guo and Fraser, 2014). There are four major Propensity score approaches namely; propensity score matching, propensity score stratification, covariate adjustment by propensity scores and propensity score based inverse probability weighting (IPW). In this essay will focus only on propensity score matching, for the other methods the, interested reader can refer to Guo and Fraser (2014).

2.4.1 Propensity score matching.

This method is used to find matches between the treated and control groups with balance baseline covariates (Fechtner, 2018). We have three steps in propensity score matching:

1. Estimate the propensity score

   Using equation 2.4 propensity scores can be estimated manually, however in cases where a large data set is used the propensity scores are estimated using statistical packages like R. The estimated propensity scores of the participants are grouped into the control and treatment groups.

   To make sure that the created groups have equal distributions of the measured baseline covariates, a sufficient overlapping of the propensity score is required. Figure 2.1 shows an example of two different samples of propensity scores where the propensity matching makes sense for one case and where it does not for the other case.

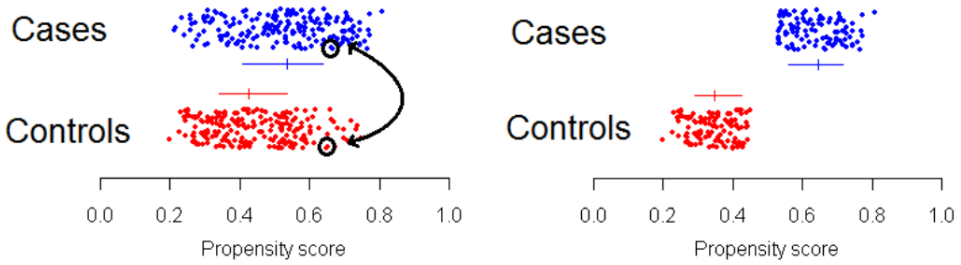


Figure 2.1: Overlapping of the propensity scores, from(Fechtner, 2018).

   From Figure 2.1 the left diagram indicates that there is sufficient overlapping of propensity scores, thus matching makes sense. The right diagram indicates that there is no overlapping of the propensity scores, thus matching does not make sense.

2. Match

   In cases where propensity matching makes sense ,ie., when we have overlapping of the groups, we now create groups by matching up individuals that have similar propensity scores or likelihoods. There are many methods that we can use to match, for example:

   (a) Nearest neighbor matching

$E(x_i)$ is the propensity score for the participant in the control group and $E(x_j)$ the propensity score for a participant in the treatment group. Let I_0 and I_1 be the set containing control participants and treated participants, respectively. A neighborhood $C(E_j)$ contains a control participant i (i.e.,$i \in I_0$) as a match for treated participant j (i.e.,$j \in I_1$), if the absolute difference between $E(x_j)$ and $E(x_i)$ is the smallest among all possible pairs of propensity scores between i and j (Guo and Fraser, 2014) as:

$$C(E_j) = \min_i\{|E(x_j) - E(x_i)|\}, \quad i \in I_0 \tag{2.5}$$

(b) Caliper matching

If the smallest absolute difference between $E(x_j)$ and $E(x_i)$ is within a predetermined caliper band b; that is,

$$C(E_j) = \min_i\{|E(x_j) - E(x_i)| < b\}. \tag{2.6}$$

Austin (2009) conducted empirical investigations and Monte Carlo simulations to investigate different sizes of calipers which are defined by the standard deviation of the logit of the propensity score. From his findings he suggested that a caliper band width of 0.2 should be used.

(c) Radius matching

It is a one-to-many matching, whereby it matches each unit i in the control group with multiple units in the treatment group within a predetermined band b; that is,

$$d(i, j) = \{|E(x_i) - E(x_i)|\} < b \tag{2.7}$$

Greedy matching or Optimal matching algorithms are used to implement the propensity matching methods above. In greedy matching algorithm, a unit i from the control group is matched to a unit j from the treated group, once this match is made, the matched units can not be replaced no matter how better the other match is.

While optimal matching you can be able to change the previous matches before you make the current match so that you can achieve the overall minimum or optimal distance. However, when using optimal matching it results in finding smaller overall distances within matched units. Therefore, to find a well-matched group, greedy matching is said to be the most preferred one (Stuart, 2010).

3. Evaluate the quality of matching

   The goal of matching techniques is to achieve a balance between the treatment and control group on observable traits. To check the quality of matching we use different approaches, such as, comparing means using t-test or chi-square, calculating the standardized difference percentage, or using graphical presentations such as histograms and box-plots.

4. The final step will use logistic regression to check the effect of the exposure variable on the outcome variable, together with the other variables that we believe affect the outcome, using the matched sample.

2.4.2 Properties of using propensity score matching.

Propensity scores help to reduce the effect of dimensionalization, when there are more confounding variables than outcome events this is done by compiling all the confounding variables into a single score which can be used for adjustment (Yang et al., 2014). It is also known that the results that are obtained using propensity score methods are more robust, precise, and have less bias as compared to other methods like regression (Fechtner, 2018). However, the use of propensity score methods requires a large sample size and adjustments can only be done for observed covariates.

2.4.3 Why and when do we use logistic regression and propensity score matching.

Propensity score methods are used to analyze the effect of an exposure variable on the outcome variable in cases where randomization was not used and there are also used in making causal inferences (Guo and Fraser, 2014). Logistic regression is used when the dependent variable for the study is a binary coded variable and when you want to control the effects of confounders.

3. Methodology

3.1 DataSet

The dataset used in this essay is from an ongoing cardiovascular cohort study that started in 1948 with 5209 adult participants from the city of Framingham, Massachusetts. The study is now on its fourth generation participants. With the aid of str() and summary() function in R, we explored the variable names and types and the first few values of the data frame, as shown in Table 6.1 in the Appendix. To summarise the properties of the datsaset, Table 3.1 shows a summary of the results.

VARIABLES IN CREATION ORDER					
No:	**Variable**	**Type**	**Label**	**Categories**	**Units**
1	Status	chr		Alive Dead	
2	DeathCause	chr	Cause of Death	Cancer Cerebral Vascular disease Coronary Heart disease Other Unknown	
3	AgeCHDdiag	num	Age CHD Diagnoised		year
4	Sex	chr		Female Male	
5	AgeAtStart	num	Age at Start		year
6	Height	num			in
7	Weight	num			lbs
8	Diastolic	num			mmHg
9	Systolic	num			mmHg
10	MRW	num	Metropolitan Relative Weight (reference weight for a given height)		
11	Smoking	num			cigarettes
12	AgeAtDeath	num	Age At Death		year
13	Cholesterol	num			mg/dL
14	Chol_Status	chr	Cholesterol Status	Borderline Desirable High	
15	BP_Status	chr	Blood Pressure Status	High Normal Optimal	
16	Weight_Status	chr	Weight Status	Normal Overweight Underweight	
17	Smoking_Status	chr	Smoking Status	Heavy (16-25) Light (1-5) Moderate (6-15) Non-smoker Very Heavy (> 25)	

Table 3.1: Variables in the dataset.

3.2 Data analysis

The main analysis of this study will be based on checking whether smoking has an impact on the development of CHD by using Propensity scores and logistic regression methods.

Missing data will be removed before any analysis is done and variables that have "character" data type will be changed to a "Factor" data type so that frequencies of the categorical variables are shown. The codes for all the analyses can be accessed here.

3.3 Creating new variables

From the dataset, the goal is to check how smoking leads to the development of CHD, given the other covariates. However, we do not have a CHD variable and some variables like Chol_Status and Smoking_Status have more than 2 categories, therefore we will create new variables that only have two categories, for easy interpretation of the results. BMI variable will be created in place of height and weight.

1. CHD

   The variable CHD will be created from the variable " AgeCHDdiag" which is a continuous variable. All participants that gave their age when CHD was diagnosed, were regarded as those that have CHD and coded with a binary variable 1. All missing values for variable " AgeCHDdiag" were regarded as the participants that did not have CHD and were coded with a binary variable 0.

2. Smoker

   The variable Smoker will be created from the Smoking variable which is a continuous variable. The smoker variable will be a binary coded i.e., 0=non-smoker's and 1 = smokers. The non-smoker's group will consist of participants that do not smoke or that only smoke $<= 5$ cigarettes a day, while the smoker's group will consist of participants who smoke > 5 cigarettes a day.

3. Cholesterol_Status

   The variable Chol_Status is categorical and it has three levels i.e., borderline, desirable, and high. We are going to form a new variable that will have two categories, for those that have a Cholesterol value of < 240 will be categorised as those that have low cholesterol status and those that have Cholesterol value of $>= 240$ will be categorised as those that have high status.

4. Body Mass Index(BMI)

   Body Mass Index is a person's weight measured in kilograms divided by the square of height measured in meters (Calle et al., 1999). High levels of BMI indicate body fatness.

$$\text{BMI} = \frac{\text{weight}(Kg)}{[\text{height}(m)]^2}, \tag{3.1}$$

   when weight is measured in pounds (lbs) and height is measured in inches (in) then we can find BMI by converting Kg to lbs and meters to inches, i.e., 1kg= 2.20462 lbs and 1 m= 39.3701 in.

   We will use equation 3.1 to calculate the BMI of the participants. Instead of using weight status, which has three levels, we will use BMI which is a continuous variable.

After the creation of the new variables will check the summary of the variables and will create a subset of the dataset that will only contain the variables of interest i.e., CHD, Smoker, Sex, AgeAtStart, Diastolic, Cholestrol_Status, BP_Status, Systolic and BMI.

3.4 Model selection

Backward regression variable selection method will be used to find an appropriate model which will be fitted in the logistic regression. The first step will be to put all the covariates in the regression model. After analysing we will remove one variable at a time, the one that has the highest p-value when compared with the others, until all variables are significant.

3.5 Logistic regression

Using the variables that were obtained during variable selection, we are going to fit the variables in the following model.

Let Y denote the binary dependent variable, i.e., $Y = 1$ if the participant has CHD and $Y = 0$ if the participant does not have CHD. $\mathsf{P}(X_i)$ denotes the odds of an event occurring at value X_i.

$$\ln \left(\frac{\mathsf{P}(X_i)}{1 - \mathsf{P}(X_i)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m, \tag{3.2}$$

X_i's are the explanatory variables, we can write equation 3.2 in terms of probability of having CHD as:

$$\mathsf{P}(X_i) = \frac{\mathsf{exp}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m)}{1 - \mathsf{exp}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m)}, \tag{3.3}$$

The logistic model is designed to describe a probability, which is always some number between 0 and 1. After fitting the model in R, we will find the parameter estimates β_i's which will be used in the interpretation of the results.

Given the following logistic regression:

$$\ln \left(\frac{\mathsf{P}(C_i)}{1 - \mathsf{P}(C_i)} \right) = \beta_0 + \beta_1 C_1 + \beta_2 C_2, \tag{3.4}$$

Lets suppose that C_1 is a binary coded explanatory variable and C_2 is a continuous independent variable,

When $C_1 = 0$:

$$\ln \left(\frac{\mathsf{P}(C_1 = 0)}{1 - \mathsf{P}(C_1 = 0)} \right) = \beta_0 + \beta_1(0) + \beta_2 C_2 = \beta_0 + \beta_2 C_2, \tag{3.5}$$

and when $C_1 = 1$:

$$\ln \left(\frac{\mathsf{P}(C_1 = 1)}{1 - \mathsf{P}(C_1 = 1)} \right) = \beta_0 + \beta_1(1) + \beta_2 C_2 = \beta_0 + \beta_1 + \beta_2 C_2, \tag{3.6}$$

Note that when equation 3.5 is subtracted from equation 3.6 it yields to:

$$\ln \left(\frac{\mathsf{P}(C_1 = 1)}{1 - \mathsf{P}(C_1 = 1)} \right) - \ln \left(\frac{\mathsf{P}(C_1 = 0)}{1 - \mathsf{P}(C_1 = 0)} \right) = \beta_0 + \beta_1 + \beta_2 C_2 - (\beta_0 + \beta_2 C_2) = \beta_1, \tag{3.7}$$

which implies that:

$$\ln\left(\frac{\mathrm{P}(C_1=1)}{1-\mathrm{P}(C_1=1)}\times\frac{1-\mathrm{P}(C_1=0)}{\mathrm{P}(C_1=0)}\right)=\beta_1, \tag{3.8}$$

by taking exponents of both sides we get:

$$\frac{\mathrm{P}(C_1=1)}{1-\mathrm{P}(C_1=1)}\times\frac{1-\mathrm{P}(C_1=0)}{\mathrm{P}(C_1=0)}=e^{\beta_1}, \tag{3.9}$$

which means that the exponents of the estimated coefficients β_i's in the logistic regression are the odds ratio's.

3.6 Propensity score analysis

A propensity score is a conditional probability of a study participant being assigned to a treatment group given a vector of observed covariates. Logistic regression is used to estimate the probabilities of the fitted values. We are going to use all the variables that are believed to have an effect on the development of CHD. According to Greenland et al. (2003) study, he identified the major risk factors of CHD as, diabetes, blood pressure, cigarette smoking, cholesterol status, and adverse dietary habits. From the dataset we are going to use all the characteristics of the participants as covariates i.e., Sex, AgeAtStart, Smoker, Cholesterol_Status, BP_Status, Diastolic, Systolic, and BMI. The following steps will be used for the analysis.

3.6.1 Estimating propensity scores.

Using the logistic equation below we can estimate the propensity score from the data.

$$\ln\left(\frac{E_i}{1-E_i}\right)=\zeta_0+\zeta_1x_1+\zeta_2x_2+\cdots+\zeta_mx_m, \tag{3.10}$$

The results that are obtained from the logistic regression will be used to calculate the propensity scores, where $x_1, x_2 \ldots, x_m$ are characteristics of the participant and $\zeta_1, \zeta_2, \ldots, \zeta_m$ are the regression coefficients. Equation 3.10 can be written in terms of propensity scores;

$$E_i=\frac{\exp(\zeta_0+\zeta_1x_1+\zeta_2x_2+\cdots+\zeta_mx_m)}{1+\exp(\zeta_0+\zeta_1x_1+\zeta_2x_2+\cdots+\zeta_mx_m)}, \tag{3.11}$$

$E_i's$ are the probabilities that predict whether the participant received treatment or not. After the propensity scores are estimated for each participant, we will create a variable named "pscore.treated.all" that consist of propensity scores for all smokers, and another variable named "pscore.control.all" for propensity scores for all non smokers.

3.6.2 Checking the distributions of propensity scores before matching.

In order to match the estimated propensity scores, the distributions of the treatment groups should overlap, otherwise, matching can not be done. Therefore, we will create a mirror histogram that will be used to check if the distribution of propensity scores in the smoker's group overlaps with the non-smoker's group.

3.6.3 Matching.

Nearest neighbour matching without replacement will be used, with a caliper band of 0.1 and a ratio of 1:1. In nearest neighbour matching: a participant (i) from the control group with propensity score E(x_i), is matched with a participant (j) from the treatment group with propensity score E(x_j), if and only if the absolute difference $C(E_j)$ between the scores is the smallest.

i.e., $C(E_j) = \underset{i}{min}\{|E(x_j) - E(x_i)|\}, \quad i \in I_0$

The ratio of the matching will be 1:1, where a participant from the control group will be matched to a participant in the treatment group, without replacement i.e., once a match is found it is removed from the set .

3.6.4 Checking baseline characteristics.

To assess the quality of matching, we will check if there exist baseline differences between the smoker's group and non-smoker's group. The following hypothesis will be tested by using p-value and the standardised difference percentage.

H_0 : The means between groups are equal.

H_1 : The means between groups are different.

The p-value and the standardised difference percentages will be taken from the R output. Having a small data set we can manually calculate the standardised difference percentage as follows.

For a continuous covariate: let d be the standardised difference percentage, where $\bar{x}_T$ is the mean for the treatment group and $\bar{x}_C$ is the mean for the control group, s_C is the standard deviation of the covariates for all the units in the control group and s_T in the treatment group.

$$d = \frac{(\bar{x}_\mathsf{T} - \bar{x}_\mathsf{C}) \times 100\%}{\sqrt{\frac{s^2_\mathsf{T} + s^2_\mathsf{C}}{2}}} \tag{3.12}$$

For a binary covariate: where P_C is the proportion of the control group and P_T the proportion for the treatment group.

$$d = \frac{(P_T - P_C) \times 100\%}{\sqrt{\frac{P_T(1 - P_T) + P_C(1 - P_C)}{2}}} \tag{3.13}$$

If the standardized differences d is greater than 10% and the p-value is less than 5%, then it indicates the means are different within the groups, hence, will support the claim, otherwise we fail to reject the null hypothesis that the mean between groups are equal.

In most cases, P-values are used however, they are not fully trusted when the sample size is large. Therefore, in addition to P-values we also use the absolute standardised difference.

3.6.5 Using logistic regression for the output matched dataset.

The aim of matching is to mimic randomization in the assignment of treatment groups. Therefore, having an appropriate matched sample will conduct a logistic regression to see the effect of the exposure variable on the outcome variables.

4. Results

4.1 Descriptive analysis

Table 4.1 is a summary of all the variables before creating new variables, it shows the frequencies of the categorical variables, the units for the numerical variables and the number of missing data for each variable.

VARIABLES IN CREATION ORDER							
No:	**Variable**	**Type**	**Label**	**Categories**	**Frequencies**	**Missing data(NA's)**	**Units**
1	Status	Factor		Alive Dead	3218 1991	None	
2	DeathCause	Factor	Cause of Death	Cancer Cerebral Vascular disease Coronary Heart disease Other Unknown	539 378 605 357 112	3218	
3	AgeCHDdiag	num	Age CHD Diagnoised			3760	year
4	Sex	Factor		Female Male	2873 2336	None	
5	AgeAtStart	num	Age at Start			None	year
6	Height	num				6	in
7	Weight	num				6	lbs
8	Diastolic	num				None	mmHg
9	Systolic	num				None	mmHg
10	MRW	num	Metropolitan Relative Weight			6	lbs
11	Smoking	num				36	cigarettes
12	AgeAtDeath	num	Age At Death				year
13	Cholesterol	num				152	mg/dL
14	Chol_Status	Factor	Cholesterol Status	Borderline Desirable High	1861 1405 1791	152	
15	BP_Status	Factor	Blood Pressure Status	High Normal Optimal	2267 2143 799	None	
16	Weight_Status	Factor	Weight Status	Normal Overweight Underweight	1472 3550 181	6	
17	Smoking_Status	Factor	Smoking Status	Heavy (16-25) Light (1-5) Moderate (6-15) Non-smoker Very Heavy (> 25)	1046 579 576 2501 471	36	

Table 4.1: Variables in the dataset.

Table 4.2 show the old and new created variables, that will be used for the analysis, the frequencies for all categorical variables are included and the units, after removing all missing values we remain with 5039 participants.

VARIABLES IN CREATION ORDER						
No:	**Variable**	**Type**	**Label**	**Categories**	**Frequencies**	**units**
1	Sex	Factor		Female Male	2764 2275	
2	AgeAtStart	num	Age at Start			year
3	CHD	Factor	Coronary Heart Disease	No CHD(0) CHD(1)	3256 1783	
4	Smoker	Factor		Non smokers (0) Smokers(1)	2994 2045	
5	Cholesterol_Status	Factor	Cholesterol status	Low(0) High(1)	3256 1783	
6	BP_Status	Factor	Blood Pressure Status	High Normal Optimal	2198 2075 766	
7	Diastolic	num				mmHg
8	Systolic	num				mmHg
9	BMI	num	Body Mass Index			lbs/in

Table 4.2: Variables in the subset.

4.2 Analyses using logistic regression

4.2.1 Model selection.

Backward variable regression method was used to select the best model to be fitted in logistic regression analysis. The models were entered sequentially as follows:

model 1: CHD$\sim$ Sex + AgeAtStart + Diastolic + Systolic + BP_Status + BMI + Cholesterol_Status + Smoker

model 2: CHD$\sim$ Sex + AgeAtStart + Diastolic + Systloic + BMI + Cholesterol_Status + Smoker

model 3: CHD$\sim$ Sex + AgeAtStart + Systolic + BMI + Cholesterol_Status + Smokers

The variables that were not significant and dropped from the models at each step have been summarised in the following table:

Model	**Not Significant variables**	**Dropped variable**	**P-value of dropped variable**
model 1	Diastolic, BP_Status	BP_Status	0.939019
model 2	Diastolic	Diastolic	0.323474
model 3	none	none	

Table 4.3: Summary of the variable selection steps.

In **model 3** all variables were significant.

Fitting **model 3** as a logistic regression model in R we get the output in Table 4.4.

Coefficients:					
	Estimate(β)	**Std.Error**	**z value**	**exp(β)**	**Pr(>|z|)**
(Intercept)	-6.079848	0.295159	-20.599	0.002289	<2e-16
Sex Male	0.747450	0.071300	10.483	2.111608	<2e-16
AgeAtStart	0.040061	0.004263	9.398	1.040874	<2e-16
Systolic	0.010175	0.001518	6.705	1.010227	2.01e-11
Smoker 1	0.482231	0.068645	7.025	1.200941	2.14e-12
Cholesterol_Status 1	0.049322	0.008317	5.930	1.619683	3.02e-09
BMI	0.183105	0.073993	2.475	1.050558	0.0133

Table 4.4: Logistic regression summary.

From Table 4.4, looking at the P-value for the variable Smoker, it indicates that indeed smoking has an effect on the development of CHD, but how sure can we be with the effect by only concluding using the P-value? As a result we further look at the odds ratio $(e^{\beta}) = 1.20$ which indicates that: for variable "Smoker", the odds for smokers developing CHD are 20% higher than the odds for non smokers.

4.3 Analyses using propensity scores

Table 4.5 show the baseline characteristics of the Non smoker and Smokers before matching was done, which indicates that that the means between the groups are different.

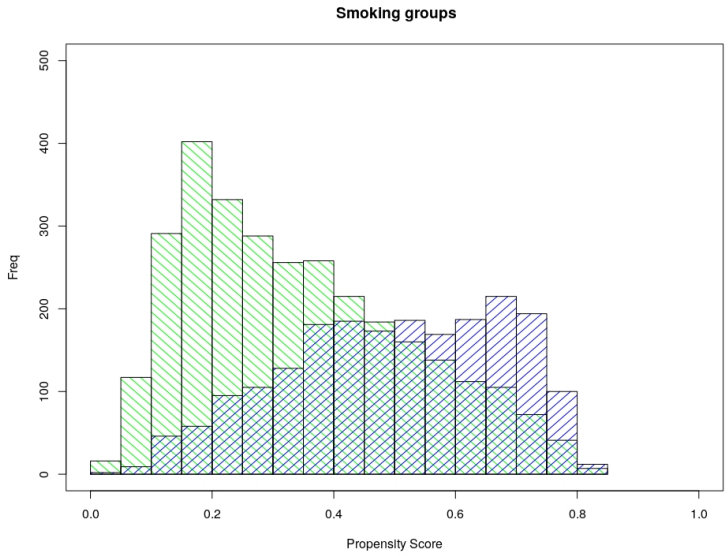
	STRATIFIED BY SMOKER			
	Non Smoker (n=2994)	**Smoker (n=2045)**	**P-value**	**SMD**
Sex=Male(%)	958(32.0)	1317 (64.4)	< 0.001	0.686
AgeAtStart(mean(SD))	45.42 (8.68)	42.10 (8.01)	< 0.001	0.398
Diastolic(mean(SD))	86.39 (13.31)	83.98 (12.35)	< 0.001	0.187
Cholesterol_Status=High(%)	1089(36.4)	694 (33.9)	0.081	0.051
BP_Status(%)			< 0.001	0.166
High	1402 (46.8)	796(38.9)		
Normal	1180 (39.4)	895 (43.8)		
Optimal	412 (13.8)	354 (17.3)		
Systolic(mean(SD))	139.30 (25.22)	133.61 (20.99)	< 0.001	0.245
BMI(mean(SD))	26.08 (4.39)	24.82 (3.85)	< 0.001	0.305

Table 4.5: Baseline characteristics of smokers.

Therefore we are going to use propensity scores to create samples that have the same baseline values, with the aim to reduce the potential effect of the baseline measurements on the main analysis.

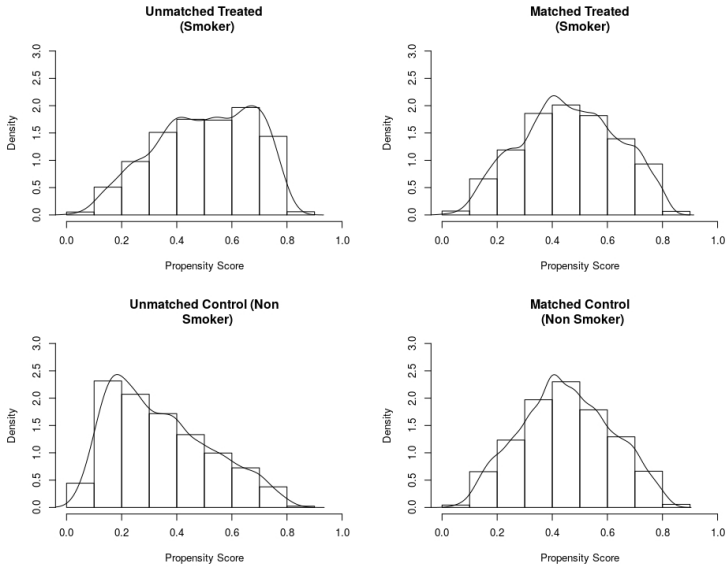
Note that the analysis using propensity scores can not be done, when the covariates contain missing

values. After estimating the propensity scores, we now check the overlapping of the groups, so that matching should be implemented. From Figure4.1, it indicates that there is overlapping between the groups.



.

Figure 4.1: Overlapping of the treatment groups, blue shaded region is for the control and green is for the treatment



.

Figure 4.2: Distributions of the propensity scores before and after the matching

From Figure 4.2 the distributions of the propensity scores of smokers and non smokers shows that they were different before matching, but after matching they look quite similar.

Having these matching distributions we now look at the significance of the baseline characteristics after matching, , i.e., if there are any significant differences of the baseline characteristics.

	STRATIFIED BY SMOKER			
	Non Smoker (n=1670)	**Smoker(n=1670)**	**P-value**	**SMD**
Sex = Male (%)	911 (54.6)	943 (56.5)	0.280	0.039
AgeAtStart (mean (SD))	43.03 (8.65)	42.89 (8.25)	0.625	0.017
Diastolic (mean (SD))	84.61 (12.73)	84.24 (12.57)	0.404	0.029
Cholestrol_Status = 1 (%)	535 (32.0)	550 (32.9)	0.605	0.019
BP_Status (%)			0.654	0.032
High	667 (39.9)	652 (39.0)		
Normal	709 (42.5)	735 (44.0)		
Optimal	294 (17.6)	283 (16.9)		
Systolic (mean (SD))	134.20 (21.34)	134.00 (21.77)	0.786	0.009
BMI (mean (SD))	25.06 (3.78)	24.95 (3.99)	0.376	0.031

Table 4.6: Baseline characteristics of smokers after adjustments.

Table 4.6 show the baseline characteristics p-values and standardized differences (SMD) after matching, we notice that all variables are statistically not significant and the standardized differences are less than 10% which indicates that we do not have differences between the two groups in the baseline values.

We can now check the effect of smoking on the development of CHD using methods that are used when we have a randomised sample (matched data set). Fitting the variables that were used to create propensity scores in a logistic regression model, we get the following output.

Coefficients:					
	Estimate(β)	**Std.Error**	**z value**	**exp(β)**	**Pr(>|z|)**
(Intercept)	-6.495048	0.589221	-11.023	0.001510903	$<$ 2e-16
SexMale	0.836837	0.092689	9.028	2.309052948	$<$ 2e-16
AgeAtStart	0.044007	0.005268	8.354	1.044989337	$<$ 2e-16
BMI	0.046573	0.011754	3.962	1.047674957	7.43e-05
Diastolic	0.008476	0.005881	1.441	1.008511940	0.14952
Systolic	0.006087	0.003275	1.859	1.006105808	0.06310
BP_StatusNormal	0.032610	0.123922	0.263	1.033147596	0.79244
BP_StatusOptimal	0.071244	0.193555	0.368	1.073843685	0.71281
Cholestrol_Status1	0.469546	0.086615	5.421	1.599268707	5.92e-08
Smoker1	0.232301	0.083263	2.790	1.261499318	0.00527

Table 4.7: Logistic regression summary of the matched sample.

From Table 4.7, Looking at the P-value for variable Smoker it also indicates that smoking has an effect on the development of CHD, when we further look at the odds ratio $(e^{\beta}) = 1.261$ which indicates that: for variable "Smoker",the odds for smokers developing CHD is 26.1% higher than the odds for non smokers.

4.4 Comparison of methods

Table 4.8 shows the output of a logistic regression that was obtained without any covariate adjustments and logistic regression output that is obtained after propensity score matching.

Coefficients:	Logistic regression					Propensity Score				
	Estimate	**Std.Error**	**z value**	**exp(β)**	**Pr(>|z|)**	**Estimate**	**Std.Error**	**z value**	**exp(β)**	**Pr(>|z|)**
(Intercept)	-6.080	0.295	-20.599	0.002	<2e-16	-6.495	0.589	-11.023	0.002	< 2e-16
Sex Male	0.747	0.071	10.483	2.112	<2e-16	0.837	0.093	9.028	2.309	< 2e-16
AgeAtStart	0.040	0.004	9.398	1.041	<2e-16	0.044	0.005	8.354	1.045	< 2e-16
BMI	0.183	0.074	2.475	1.051	0.0133	0.047	0.012	3.962	1.048	7.43e-05
Diastolic						0.008	0.006	1.441	1.009	0.150
Systolic	0.010	0.002	6.705	1.010	2.01e-11	0.006	0.003	1.859	1.006	0.063
BP_StatusNormal						0.033	0.124	0.263	1.033	0.792
BP_StatusOptimal						0.071	0.194	0.368	1.074	0.713
Cholesterol_Status 1	0.049	0.008	5.930	1.620	3.02e-09	0.470	0.087	5.421	1.599	5.92e-08
Smoker 1	0.482	0.069	7.025	1.201	2.14e-12	0.232	0.083	2.790	1.261	0.005

Table 4.8: A comparisons between logistic regression and propensity scores results.

Note that Diastolic and BP_Status were not included in the logistic regression without adjustments, hence the spaces. For the Smoker variable in Table 4.8 we can see that the odds for developing CHD due to being classified as a smoker using propensity scores is 1.261, which is higher than the odds for developing CHD from the logistic regression without adjustments, which is 1.201 and the difference between the odds is very small and they all lead to the same interpretation. Therefore, in both cases it is indeed, true that smoking leads to the development of CHD.

5. Discussion

In the dataset, it was discovered that before creating a new variable, missing data should be dealt with, using any appropriate method. For example when the "Smoker" variable was created from the "Smoking" variable before removing missing data of the original variable. The new Smoker variable did not have NA's, instead the NA spaces were given 0 values which would result in making wrong estimates.

On the same issue of missing values, we noticed that AgeCHDdiag had the highest number of missing values, this is because most participants were not diagnosed with having CHD, therefore the spaces were left blank. When deleting the missing variables we only targeted the other variables except for AgeCHDdiag, the same applies to the DeathCause variable.

However it is still not clear when it comes to selecting variables or an appropriate model to use in the propensity score method. In some research papers like Brookhart et al. (2006), they have recommended that we should include all outcome variables, but this causes debates and confusion when you look at the possibility of having a lot of confounding variables, which is a common problem in epidemiology or medical surveys. Some researchers have thought of only including variables that affect either the outcome or exposure variable.

In applying the propensity score, we compared the methods that were used in the Fechtner paper and the methods that have been used in this essay. The Fechtner paper used a matching ratio of 2:1, the matching method was the nearest neighbor with replacement and the caliper band of 0.2 (which is the standard), to check the baseline characteristics they used the weighted chi-square and weighted t-test, which showed that the groups were well balanced. However using the same methods on the dataset and checking the balance of the baseline characteristics using "tableone" package, it indicated that the groups were not matched perfectly as some of the variables like "Sex" and "AgeAtStart" indicated that significant differences within the groups still existed.

After several trials on changing the caliper band value, the matching ratio, and replacement value, the baseline characteristic's p-value and absolute standardised differences were all less than 5% and 10%, respectively. From the Logistic regression results using or without using any adjustment on the covariates, smokers have a higher possibility of developing coronary heart disease than those that do not smoke.

In conclusion, it is a well-known concept in theory that propensity scores are the best methods when it comes to analysing observational data, but from our results we have seen that both methods lead to the same interpretation, with a very minor difference in the actual values.

Acknowledgements

First and foremost, I thank God for making things possible and beautiful for me in His own time. I want to thank AIMS and it's funders for supporting this work. Very special gratitude to my supervisor Professor Arne Ring for his support and guidance throughout the development of this essay. I would also like to thank Samar Elsheikh our AIMS tutor for her continuous support and encouragement. Lastly I would like to thank my family for always believing in me.

6. Appendix

6.1 Descriptive statistics table

tibble [5,209 × 17] (S3: tbl_df/tbl/data.frame)		
$ Status	chr [1:5209]	"Dead" "Dead" "Alive" "Alive" ...
$ DeathCause	chr [1:5209]	"Other" "Cancer" NA NA ...
$ AgeCHDdiag	num [1:5209]	NA NA NA NA NA NA NA NA NA NA ...
$ Sex	chr [1:5209]	"Female" "Female" "Female" "Female" ...
$ AgeAtStart	num [1:5209]	29 41 57 39 42 58 36 53 35 52 ...
$ Height	num [1:5209]	62.5 59.8 62.2 65.8 66 ...
$ Weight	num [1:5209]	140 194 132 158 156 131 136 130 194 129 ...
$ Diastolic	num [1:5209]	78 92 90 80 76 92 80 80 68 78 ...
$ Systolic	num [1:5209]	124 144 170 128 110 176 112 114 132 124 ...
$ MRW	num [1:5209]	121 183 114 123 116 117 110 99 124 106...
$ Smoking	num [1:5209]	0 0 10 0 20 0 15 0 0 5 ...
$ AgeAtDeath	num [1:5209]	55 57 NA NA NA NA NA 77 NA 82 ...
$ Cholesterol :	num [1:5209]	NA 181 250 242 281 196 196 276 211 284 ...
$ Chol_Status :	chr [1:5209]	NA "Desirable" "High" "High" ...
$ BP_Status :	chr [1:5209]	"Normal" "High" "High" "Normal" ...
$ Weight_Status :	chr [1:5209]	"Overweight" "Overweight" "Overweight" "Overweight" ...
$ Smoking_Status:	chr [1:5209]	"Non-smoker" "Non-smoker" "Moderate (6-15)" "Non-smoker" ...

Table 6.1: summary of the variable selection steps

6.2 Codes

The codes for all the analyses can be accessed here.

References

Agresti, A. *An introduction to categorical data analysis*. John Wiley & Sons, 2018.

Austin, P. C. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and monte carlo simulations. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(1):171–184, 2009.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

Calle, E. E., Thun, M. J., Petrelli, J. M., Rodriguez, C., and Heath Jr, C. W. Body-mass index and mortality in a prospective cohort of us adults. *New England Journal of Medicine*, 341(15):1097–1105, 1999.

Elze, M. C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., Nichols, M., Stone, G. W., and Pocock, S. J. Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *Journal of the American College of Cardiology*, 69(3):345–357, 2017.

Fechtner, S. *Propensity score matching*, 2018. URL https://www.lexjansen.com/phuse/2018/rw/RW03.pdf.

Greenland, P., Knoll, M. D., Stamler, J., Neaton, J. D., Dyer, A. R., Garside, D. B., and Wilson, P. W. Major risk factors as antecedents of fatal and nonfatal coronary heart disease events. *Jama*, 290(7): 891–897, 2003.

Guo, S. and Fraser, M. W. *Propensity score analysis: Statistical methods and applications*, volume 11. SAGE publications, 2014.

Heinze, G., Wallisch, C., and Dunkler, D. Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449, 2018.

Rässler, S., Rubin, D. B., and Zell, E. R. 19 incomplete data in epidemiology and medical statistics. *Handbook of statistics*, 27:569–601, 2007.

Rosenbaum, P. R. and Rubin, D. B. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

Stuart, E. A. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Sullivan, G. M. Getting off the "gold standard": randomized controlled trials and education research. *Journal of graduate medical education*, 3(3):285–289, 2011.

Tian, Y., Schuemie, M. J., and Suchard, M. A. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *International journal of epidemiology*, 47(6): 2005–2014, 2018.

Tranmer, M. and Elliot, M. Binary logistic regression. *Cathie Marsh for census and survey research, paper*, 20, 2008.

Yang, W., Joffe, M. M., Hennessy, S., and Feldman, H. I. Covariance adjustment on propensity parameters for continuous treatment in linear models. *Statistics in medicine*, 33(26):4577–4589, 2014.

Zou, B., Zou, F., Shuster, J. J., Tighe, P. J., Koch, G. G., and Zhou, H. On variance estimate for covariate adjustment by propensity score analysis. *Statistics in medicine*, 35(20):3537–3548, 2016.

Zwonitzer, M. R., Soupir, M. L., Jarboe, L. R., and Smith, D. R. Quantifying attachment and antibiotic resistance of escherichia coli from conventional and organic swine manure. *Journal of environmental quality*, 45(2):609–617, 2016.