Propensity Score Matching for Causal Inference of Chronic Kidney Disease and Obesity

Wenyu Qu

21/12/2020

Abstract

Statistical analysis has been broadly used for different research fields, especially in clinical studies. Chronic kidney disease has become a common disease worldwide, which has caused a series of impacts on people. Obesity has become a common epidemic that could increase the risk of chronic kidney disease. Thus, it is important to use statistical analysis to find causal inference in observational data and to find the causal link between obesity and chronic kidney disease. In this report, variables Age, Blood Pressure, Specific Gravity, Anemia, and Class are selected from the UCI Machine Learning Repository website since they are essential potential driving factors of whether a person has chronic kidney disease. In addition to this, the obesity data will be stimulated and use to find out the causal link between obesity and chronic kidney disease. The logistic regression model is used in the report, as it helps model the association between the response variable (ckd) and predictor variables. The propensity score matching technique is used to find people in both ckd and notckd groups who have the same probability of being in the treatment group. Then a propensity score regression is performed to analyze the significant influence of variables age, bp, sp, ane, and ckd on the average obesity. By conducting a logistic regression on chronic kidney disease and predictor variables, the summary statistics show that whether the person was treated as ckd is influenced by blood pressure and specific gravity. The Huxtable of propensity score regression model indicates that the treatment group: ckd has a significant influence on the average obesity, which is our outcome of interest. Thus, there is a causal link between chronic kidney disease and average obesity.

Keywords

Propensity Score Matching, Logistic Regression, Causal Inference, Chronic Kidney Disease, Obesity, Observational Study

Introduction

Nowadays, chronic kidney disease has become a common disease worldwide, and the percentage of the prevalence of chronic kidney disease has increased by almost 30% since 1990 (Carney, 2020). It has caused a series of complications. Chronic kidney disease could have a wide range of impacts on people, from small symptoms like loss of appetite to severe anemia and even mortality (The National Kidney Foundation, 2017). Statistical analysis has been broadly used for different research fields, especially in clinical studies. And observational data is more appropriate and reliable than experimental data since it observes some variables and tries to find a correlation rather than control some variables and try to find causality. Hence, it is important to use statistical analysis to find causal inference in observational data and to identify some risk factors of diseases, in this case, chronemic kidney disease.

The concept of propensity score matching was first introduced in 1983 by Rosenbaum and Rubin (Austin, 2011). Propensity score matching has become a widely used method for conducting causal inference with

observational data in statistical analysis in recent years (Arbour et al., 2014). The propensity score is the conditional probability of treatment assignment on observed covariates; propensity score matching helps to match treated and untreated subjects that has a similar propensity score (Austin, 2011). This report will perform a propensity score matching to find causal inference of whether a person has obesity and whether a person is confirmed as chronic kidney disease.

Obesity has become a common epidemic in the world. Obesity prevalence has been increased every year, and it has a lot of implications for the risk of chronic kidney disease (Kovesdy et al., 2017). According to Kovesdy et al. (2017), a compensated hyperfiltration happens in order to meet the increased metabolic requirements of high body weight, which will lead to an increase in intraglomerular pressure that damages the kidney and increases the risk of developing chronic kidney disease. Therefore, it is important to find causal obesity and chronic kidney disease.

The purpose of this report is to find the causal link between obesity and chronic kidney disease. In order to make causal inferences of obesity and chronic kidney disease based on the performance of propensity score matching, this report will use the chronic kidney disease dataset. In the Methodology section, the chronic kidney disease data, the logistic model that will be used to conduct the propensity score analysis will be introduced. In the Result section, the results of the logistic model and propensity score matching will be presented. Lastly, in the Discussion section, the causal inferences of the data, the summary.

Appendix

Github Repo link:https://github.com/quwenyu123/final-report

Methodology

Data

Chronic Kidney Disease Dataset is obtained from the UCI Machine Learning Repository website. The dataset collected from a hospital for two months of period, and the initially dataset has 400 observations and 25 variables (Dua & Graff, 2019). In this report, variables Age, Blood Pressure, Specific Gravity, Anemia, and Class are selected, where Age, Blood Pressure, Specific Gravity are numerical variables, and Anemia is categorical variables, and the Class (ckd/ notckd) is the treatment group. According to the National Kidney Foundation (n.d), people over 60 years old are more likely to develop kidney disease since when they are aging, their kidneys are aging at the same time. High blood pressure could constrict the blood vessel, which will damage the kidney's blood vessels, and the kidney will not be able to excrete wastes from a person's body (NIDDK, 2020). Moreover, high blood pressure is more likely to develop when a person is overweight or obese, as the person's heart needs to work harder to pump the blood for the body (Larson, 2019). The urine specific gravity indicates how good a person's kidney concentrates urine and the person's hydration level in order to detect the person's kidney function (Luo, 2018). There are many complications of chronic kidney disease; anemia is one of them; when a person has chronic kidney disease, the kidney cannot filter blood, so the person's lower amount of red blood cells than usual, which leads to Anemia (NIDDK, 2020). And variable Class indicates whether the person has chronic kidney disease or not. Therefore, these variables are essential potential driving factors of whether a person has chronic kidney disease, and also important to find out the causal link between obesity and chronic kidney disease.

The report also needs some measure of a person's average obesity level, so the report simulated data for obesity as there is a lack of variables in the dataset that provide sufficient data for obesity. According to the Centers for Disease Control and Prevention (2020), Body Mass Index (BMI) is between 18.5 to less than 25 is classified as normal weight; when BMI is bigger than 30, then it is classified as obese. Therefore, in this report, if a person's mean obesity BMI is 30, the person has chronic kidney disease. If the person's mean obesity BMI is 22, then the person does not have chronic kidney disease.Thus, chronic kidney disease (Class) is the treatment, and average obesity is the outcome of interest. Here Table 1 is provided to indicate baseline characteristics of the data.

```
table1
```

```
##
##                                Overall
##   n                              336
##   age (mean (SD))              51.32 (16.47)
##   bp (mean (SD))               76.04 (12.20)
##   sg (mean (SD))                1.02 (0.01)
##   ane = yes (%)                   42 (12.5)
##   ckd = 1 (%)                   194 (57.7)
##   average_obesity (mean (SD)) 26.65 (4.15)
```

The table indicates that after the cleaning of the initial Chronic Kidney Disease Dataset by removing all the N/A in variables, there are 336 observations remained. There are 194 out 336 people who have ckd, and 142 out of 336 people are notckd.

–	notckd	ckd
Number of People	142	194

The mean of variable age is 51.32, the standard deviation is 16.47. The mean of variable blood pressure is 76.04, the standard deviation is 12.20. The mean of variable specific gravity is 1.02, the standard deviation is 0.01. The mean of variable anemiayes is 42, the standard deviation is 12.5. The mean of having ckd is 194, the standard deviation is 57.7. And the mean of average obesity is 26.64, and the standard deviation is 4.02.

Model

The model this report is chosen is the logistic regression model, as it helps model the association between response variable Class (ckd) and predictor variables Age, Blood Pressure, Specific Gravity, Anemia. The logistic regression model could include both numerical and categorical predictor variables since Age, Blood Pressure, Specific Gravity are numerical variables, and Anemia is categorical variables. Also, the logistic regression model performs a logistic function to model a binary response variable; in this case, the variable class is selected to indicate whether a person has chronic kidney disease (ckd) or not. Besides, the logistic regression model is appropriated to estimate propensity score as the treatment is an outcome ground on covariate vector (Caetano, 2020). Hence, the logistic regression model is more appropriated for the data. The logistic regression model expression is

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{bp} + \beta_3 x_{sg} + \beta_3 x_{aneyes}$$

Also, the propensity score matching technique assigns some probability to each observation of whether they would be in the treatment group. This report will treat some individuals with chronic kidney disease (ckd) and to see the influence on average obesity. Thus, whether a person has ckd is the treatment, and average obesity is the outcome of interest; and the report wants to find out the propensity of a person who has ckd and then matches based on the propensity. The report is interested in finding out whether a person has ckd or not ckd, so a logistic regression model will be constructed based on different variables, such as age, blood pressure, specific gravity, and anemia, and tries to find out the probability of being either treatment groups (ckd/notckd). Then the report attempts to find people in both ckd and notckd groups who have the same probability of being in the treatment group, and using the matching function in the arm package to find the closest untreated ones and matches to each treated ones. Then a propensity score regression is performed to analyze the significant influence of variables age, bp, sp, ane, and ckd on the average obesity.

Results

```
huxtable::huxreg(propensity_score_regression)
```

	(1)
(Intercept)	7.875
	(14.605)
age	0.004
	(0.004)
bp	0.005
	(0.005)
sg	13.226
	(14.278)
aneyes	0.221
	(0.180)
ckd1	8.122 ***
	(0.175)
N	336
R2	0.940
logLik	-481.555
AIC	977.110

*** p < 0.001; ** p < 0.01; * p < 0.05.

According to summary statistics, the formula of estimated logistic regression model is:

$$log(\frac{p}{1-p}) = 588.178 + 0.018x_{age} + 0.052x_{bp} – 581.833x_{sg} + 19.955x_{aneyes}$$

Where $\beta_0 = 588.178$, $\beta_1 = 0.018$, $\beta_3 = -581.833$, and $\beta_4 = 19.955$ (3 decimal points). The β_0 represents the intercept of the model and is the probability of ckd when age equals to 0, blood pressure equals to 0, specific gravity equals to 0, and anemia is no. The β_1 represents for every one unit increase in age, there is an expectation that β_1 increase in the probability of ckd. Similarly, β_2 represents for every one unit increase in the blood pressure, there is an expectation that β_2 increase in the probability of ckd. The β_3 represents for every one unit increase specific gravity, here is an expectation that β_3 increase in the probability ckd. The β_4 represents for every one unit increase anemia, there is an expectation that β_4 increase in the probability of ckd.

There is a weak positive correlation between log odds of ckd and both age and blood pressure, a strong positive correlation between ckd and have anemia and a strong negative correlation between ckd and specific gravity. The intercept H_0: $\beta_0 = 0$, H_a: β_0 0, in the summary statistics, p-value of the intercept coefficient equals to 2.98e-14, which is smaller than 0.05. Therefore, reject H_0 and support H_a 0. Similarly, for the age

H_0: $\beta_1 = 0$, H_a: β_1 0, p-value of age equals 0.135, which is bigger than 0.05, hence, fail to reject H_0 and indicates that there is not much correlation between ckd and age. For the blood pressure, H_0: $\beta_2 = 0$, H_a: β_2 0, p-value of bp equals 0.009, which is smaller than 0.05, hence, reject H_0 and support H_a 0, indicates that there is correlation between ckd and blood pressure. For the specific gravity, H_0: $\beta_3 = 0$, H_a: β_3 0, p-value of sp equals 2.00e-14, which is smaller than 0.05, hence, reject H_0 and support H_a 0, indicates that there is correlation between ckd and specific gravity. Lastly, for the age H_0: $\beta_4 = 0$, H_a: β_4 0, p-value of aneyes equals 0.987, which is bigger than 0.05, hence, fail to reject H_0 and indicates that there is not much correlation between ckd and anemia.

By using the matching function in the arm package to find the closest untreated ones and matches to each treated ones, the report could find out how ckd would affect people's average obesity. The propensity score regression of the outcome of interest, which is average obesity based on age, blood pressure, specific gravity, anemia, as well as the treatment, whether a person has ckd or not based on the data. The Huxtable indicates that the p-value of intercept of propensity score regression is bigger than 0.05, which indicates intercept is not significant in influencing on average obesity. The p-value of age is bigger than 0.05, which indicates age is not significant in influencing on average obesity. The p-value of bp is bigger than 0.05, which indicates blood pressure is not significant in influencing on average obesity. The p-value of sp is bigger than 0.05, which indicates specific gravity is not significant in influencing on average obesity. The p-value of aneyes is bigger than 0.05, which indicates that having anemia is not significant in influencing on average obesity. The p-value of ckd1 is smaller than 0.001, which indicates chronic kidney disease is very significant in influencing on average obesity. Therefore, only the ckd have significant influences on average obesity.

Discussion

Summary

In order to find out the causal link between obesity and chronic kidney disease, the report used some potential driving factors, such as age, blood pressure, specific gravity, and anemia to find their relationship with chronic kidney disease. This report first conducted a logistic regression model to find out whether the person was treated as a function of the variables. Then a forecast was created to add on the ckd dataset to create matches, then found people who have treated as ckd to match with the closest untreated ones based on the propensity score. The propensity score matching technique is used to assigns some probability to each observation of whether they would be in the treatment group, where ckd is the treatment, and average obesity is the outcome of interest. To find a person who has the same probability of being in the treatment group and the propensity to get both ckd and notckd, the matching function in the arm package was used to perform propensity score matching. A variable treated was created, then using the matching function to match whether a person has ckd or not. After that, the report combined the matches data to the ckd data, then conducted a propensity score regression to determine what factors significantly influence the outcome of interest: average obesity.

Conclusions

The initial research indicates that factors age, blood pressure, specific gravity, and anemia all have an impact on whether a person has chronic kidney disease or not. By conducting a logistic regression on chronic kidney disease and predictor variables, the summary statistics show that age, blood pressure, and whether having anemia all have a positive correlation with chronic kidney disease based on the dataset. But specific gravity has a strong negative correlation with chronic kidney disease. According to the p-values of these variables, whether the person was treated as ckd is influenced by blood pressure and specific gravity.

Then a propensity score matching was performed to match those who were treated as ckd and to those who were untreated based on their propensity score. And propensity score regression was conducted, and the Huxtable of propensity score regression model indicates that the treatment group: ckd has a significant influence on the average obesity, which is our outcome of interest. Yet, other variables, age, blood pressure, specific gravity, and anemia, do not have a significant influence on average obesity. Therefore, the Huxtable indicates that there is a causal link between chronic kidney disease and average obesity.

Weakness

Chronic Kidney Disease Dataset is a relatively small dataset that has 400 observations and 25 variables, and the data was collected from one hospital in two months (Dua & Graff, 2019). Therefore, the dataset may not be very representative of the actual findings. There are more significant variables that are not considered in the report, such as red blood cells, bacteria, and hypertension, etc. The average obesity was simulated for the report; it could not include every single scenario. Also, the propensity score matching has some flaws. According to Gary King and Richard Nielsen (2019), propensity score matching tries to conduct a fully randomized experiment and ignore the large percentage of imbalance that could be removed by other matching methods, so propensity score matching increases imbalance, it is inefficient and has more bias. Moreover, there are other more sophisticated models and codes that more appropriate and accurate for this report.

Next Steps

Researchers need to find datasets collected from different hospitals in different cities or even countries over a longer period of time. And try to conduct the causal inference of obesity and chronic kidney disease by considering other driving factors of chronic kidney disease. Instead of stimulating data on obesity, try to find or ask for access to better and more comprehensive datasets that provide more factors and observations and contain obesity variables for statistical analysis. To avoid the flaws of propensity score matching, use different matching methods to reconduct the research again. In addition, try to use different and more sophisticated models and codes to perform the research again and try to find the different results of the research.

References

- Alexander, Rohan (2020), "Running Through a Propensity Score Matching Example", Created 11 November 2020. University of Toronto
- Austin, P. (2011, May). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Retrieved December 22, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/
- Caetano, S. (2020). Week 10 Lecture: Causality in Observational Studies. STA304 Surveys, Sampling and Observational Data. University of Toronto
- Carney, E. (2020). The impact of chronic kidney disease on global health. Retrieved December 22, 2020, from https://www.nature.com/articles/s41581-020-0268-7
- Centers for Disease Control and Prevention. (2020, September 17). Defining Adult Overweight and Obesity. Retrieved December 22, 2020, from https://www.cdc.gov/obesity/adult/defining.html
- Introduction to tableone, an R package to facilitate creation of Table 1. (n.d.). Retrieved December 22, 2020, from http://rstudio-pubs-static.s3.amazonaws.com/13321_da314633db924dc78986a850813a50d5.html
- King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. Political Analysis, 27(4), 435-454. doi:10.1017/pan.2019.11
- Larson, J. (2020, February 13). The Obesity and High Blood Pressure Connection. Retrieved December 22, 2020, from https://www.healthgrades.com/right-care/weight-control-and-obesity/the-obesity-and-high-blood-pressure-connection
- Luo, E. K. (2018). Urine specific gravity test: Procedure and results. Retrieved December 22, 2020, from https://www.medicalnewstoday.com/articles/322125
- National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). (2020, March 01). High Blood Pressure & Kidney Disease. Retrieved December 22, 2020, from https://www.niddk.nih.gov/health-information/kidney-disease/high-blood-pressure

- National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). (2020, September 01). Anemia in Chronic Kidney Disease. Retrieved December 22, 2020, from https://www.niddk.nih.gov/health-information/kidney-disease/anemia
- National Kidney Foundation. (2020, March 12). Aging and Kidney Disease. Retrieved December 22, 2020, from https://www.kidney.org/news/monthly/wkd_aging