Analysis of Hormone Level Imbalance of Individual Impact on Acne Breakout

Jianzhong You

22/12/2020

Repository of the markdown and simulation code file of this paper can be found in https://github.com/maxxyouu/STA304-FINAL-PROJECT.git

Abstract

Experimental study is the most rigorous method to deriving the casual relationship between the intervention and the outcome as it uses random assignment to get rid of any alternative explanation. However, most studies today are quasi-experiment, which does not involve random assignment and has high probability of contain hidden bias. Furthermore, some studies could not be done simply due to logistic reasons and ethical issue. In this case, we need to make use of some statistical techniques to derive casual relationship from a quasi-experiment. In this paper, I use the propensity score method with nearest neighbor matching technique to draw the casual relationship between the hormone level imbalance variable and the acne breakout response variable from simulated data and ultimately conclude the causal relationship between hormone level imbalance and acne breakout.

Keywords: Propensity Score Matching, Observational Study, Acne Breakout, causal inference, Logistic Regression.

Introduction

Experimental study with random assignment is known as one of the most convincing way to derive the casual relationship between variables of interest. However, in reality, ethical or logistical issues often prevent a rigorous experiment from successfully implemented. In contrast, observational studies from survey are abundant in society and much easier to implement, but at a cost of being bias due to the nature of non-random assignment. To make casual effect conclusion, there are a few requirements that need to be met. First, statistical relationship must exists, often by looking at the correlation. Second, no alternative explanation that able to explain the same effect across the treatment and control group. Third, "there is a reasonable counterfactual"[3]. In observational study, the first requirement is easier to met as we can find the correlation relation use statistical approaches. The second requirement is what defines a experimental and quasi-experiment. In experimental study, researcher often employ random assignment to evenly distribute the treatment and controlled groups, which could lead to the fact that the distribution of covariates between the two groups are the same. If there are any significant difference in outcome between the treatment group and controlled group, researchers are confident that the effect is due to the treatment and no alternative explanation. For counterfactual effect, it is essentially a thought experiment that gives "knowledge of what would have happened to those same people if they simultaneously had not received treatment"[4]. Although derive a cause-and-effect relation between variables in observation study is difficult, by employing some sophisticated statistical techniques, research could at least approximate or imitate the nature of random assignment property.

There are numerous statistical methods that are capable to derive the casual relationship from observational data, such as Regression Discontinuity Design (RDD), Difference in Differences (DID), Propensity Score Matching, and etc. In particular, propensity score matching begins to gain popularity as more casual inferences are done using observational data in various fields of study[5]. By using propensity score matching method, applied researchers able to use the balancing score from the logistic regression (more on this in later section) to make two groups that are of similar distribution of observed covariates [6], which is the key property to make casual inference conclusion. The intuition behind this balancing score number is that researchers do not need to make one-to-one matching (each in different treatment group) based on the long list of observed covariates because as more covariates are observed, the amount of data needed grows exponentially in asymptotic sense. However, when using a single score as a matching metric to divide data into two groups, it is much more flexible and easier to find each pair match base on different matching techniques[3]. In this paper, I will use propensity score matching method as a way to create two groups that are approximately balanced in observed covariates and subsequently draw causal link between the androgen level of individual with acne breakout/inflammation.

Acne problem is the most common skin problem for teenagers or even individuals in early millennial age. Study has shown that more than 85%[1] teenagers suffer from acne breakout to some degree. Although acne is not fetal and it most likely heals by themselves as time go by. However, psychologically, study has shown that severe acne outbreak on the face will cause destruction on the self-confidence of individual. Quotes from a dermatologist, people "with acne can often feel unsupported, socially isolated and become withdrawn"[2]. To successfully combat the acne breakout, individual could change their daily diet to help alleviate this skin disorder early to prevent it from getting worse, or even as a supplemental support that complement the prescription drugs from the doctor to make the recovery more effective and faster (the duration to fully recover acne problem is long).

In this paper, I will first give a description of the simulated data set by examining the predictor and response variables such as age, heredity, protein shake intake, rice consumption level, and bread consumption level, that have impact on the hormone level imbalance. Subsequently, I provide some visualizations as well as some quantitfy measures of the balancing scores, before and after. At the end, I derive the causal link between hormone imbalance of individual and the acne breakout using the propensity score with nearest neighbor matching technique. At the end, I lay out the limitation and further steps desired to make the analysis even more rigorous.

Methodology

Data

In the simulated data, 100,000 data points are sampled. I only consider seven predictor variables, including the variable of interest–hormone imbalance level of each individual, where the hormone level imbalance serves as the observed intervention. One of the predictor is age, since beyond 30 years, individual is very unlikely to have acne[7] and to prevent further imbalance of the data, the simulated data only consists of individuals that are between 18 and 30 years of age, where each individual is drew from the a uniform distribution. The second covariate I consider is the hereditary (0 or 1) of individual. Study shows that it plays a role on the probability of having acne outbreak during the teenager age due to the genetic component inherited from their parent; that is, the risk of having acne is much higher if parent also had it during their lifetime. Clinical study estimates that there are 50%-90% of acne outbreak cases are attributed to genetic factor[9], I decide to draw the sample from a Bernoulli distribution with the probability of having the inherited genetic as 75% to avoid bias toward either end of the estimated range. The third covariate I consider is the protein shake consumption (0 or 1) of individual. This covariate is simulated by using their workout habit as a proxy variable to determine if a person take protein shake. It is drew from a Bernoulli distribution with probability of taking protein shake as 70% if a person do workout, otherwise 20%. There are numerous studies show that protein shake could "trigger the production of androgens, or hormones that work by overstimulating oil glands."[10], which is the primary cause of acne outbreak. The fourth covariate is the amount of rice consumption annually by an individual. Study shows that the consumption amount of an average person is 26 pounds of rice[11]. Therefore, I simulate this covariate by drawing samples from a normal distribution with the mean as 26 and standard derivation of 4. I allow the spread of the distribution wider as I believe not everyone eats rice as stable food in daily life. The fifth covariate is the annual amount of bread consumption. I draw the sample from a normal distribution with a mean of 37 with standard derivation as 4 [12]. The amount of rice and bread consumption has a huge impact on the acne inflammation. From study, rice and bread are food that have high glycemic index value, which leads to much higher chance of having acne breakout due to change in hormone[13]. The sixth covariate is the categorical variable, college student (assign as 1 if true, 0 otherwise). The value is drew from a Bernouli distribution with the probability of college student 70% if his/her age is less than 25 years old, otherwise, the probability drops to 20%. Finally, the "intervention" variable is the hormone level imbalance of individual, which is determined by the previous mentioned six predictors. since hormone consist of multiple substances and each of them has its unique scale of measurement. For the convenience of the paper, I assign hormone imbalance as 1, 0 otherwise. There is only one dependent variable in the data, which is the acne breakout categorical variable. This variable would be simulated from a Bernouli distribution with the parameter depends on the previous mentioned seven predictors. Figure 1 belows show a brief summary of the raw data await to be further process.

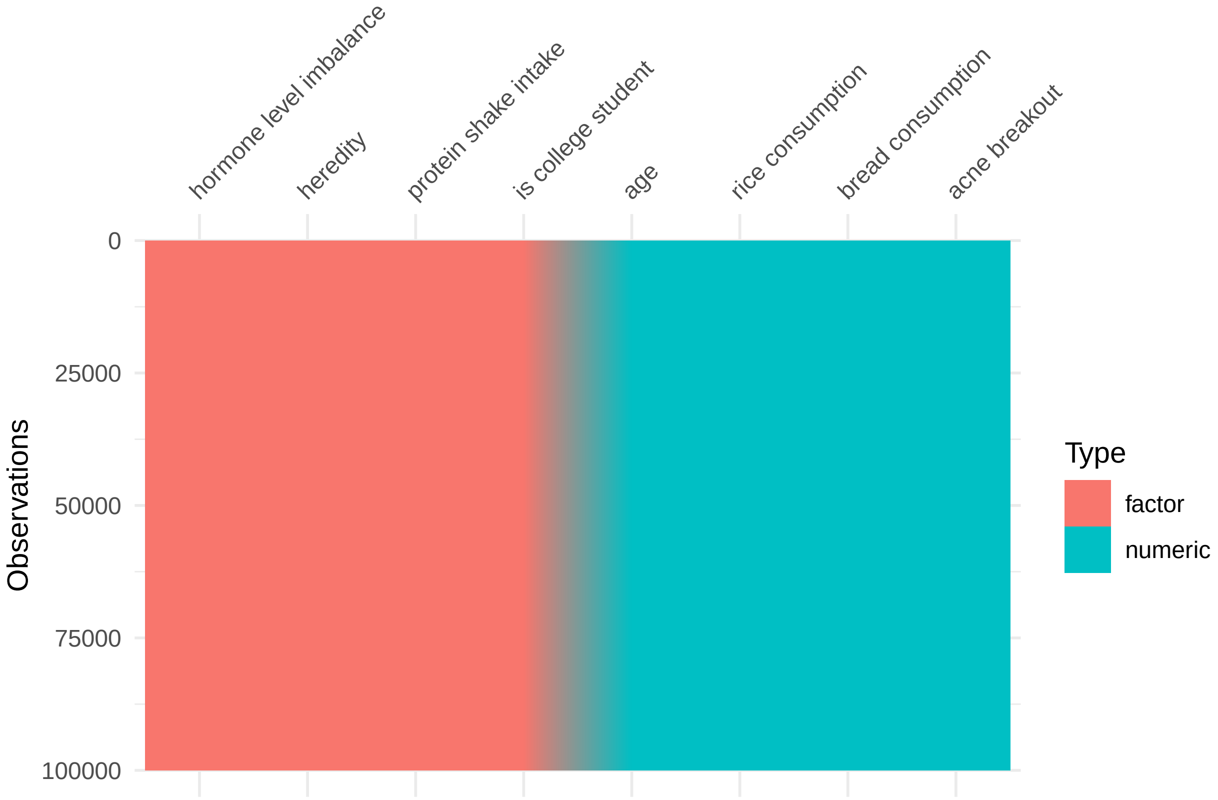

Figure 1: Simulated Raw Data

Model

Before we perform the analysis to derive the casual relationship between the hormone imbalance level and acne breakout, we need to perform a logistic regression on the hormone imbalance level to get the balancing score each data point since the response variable is binary, as follows:

$$\log(\frac{p_h}{1-p_h}) = \beta_0 + \beta_a x_a + \beta_h x_h + \beta_p x_p + \beta_c x_c + \beta_r x_r + \beta_b x_b$$

where p_h is the probability of hormone imbalance given the covariates, but we treat it as the propensity score, also known as the balancing score. x_a is the age, x_h is heredity, x_p is the protein shake intake habit, x_c is the college student categorical variable, x_r is the amount of rice consumption per year, and x_b is the amount of bread consumption per year.

Let $\mathbf{XB} = \beta_0 + \beta_a x_a + \beta_h x_h + \beta_p x_p + \beta_c x_c + \beta_r x_r + \beta_b x_b$, to find the propensity score of each individual, we need to rearrange above equation into the following:

$$p_h = \frac{e^{\mathbf{XB}}}{1 + e^{\mathbf{XB}}}$$

In the matching stage, each pair of data are matched based on the expression above.

After the perform the matching using the *arm* package in R, we use the matched data set to perform a logistic regression modeling to find the casual link between the hormone level imbalance to the acne breakout binary response variable.

$$\log(\frac{p_a}{1 - p_a}) = \beta_0 + \beta_a x_a + \beta_h x_h + \beta_p x_p + \beta_c x_c + \beta_r x_r + \beta_b x_b + \beta_i x_i$$

The predictors are the same as above, but the additional x_i represents the hormone imbalance level variable and p_a this time represents the probability of acne breakout given the observed covariates. The main variable of interest in this study is x_i and I conclude the casual link relationship between x_i and p_a by observing β_i and its corresponding p value significance.

All computations and modelings are performed using the Rstudio IDE and R language.

Results

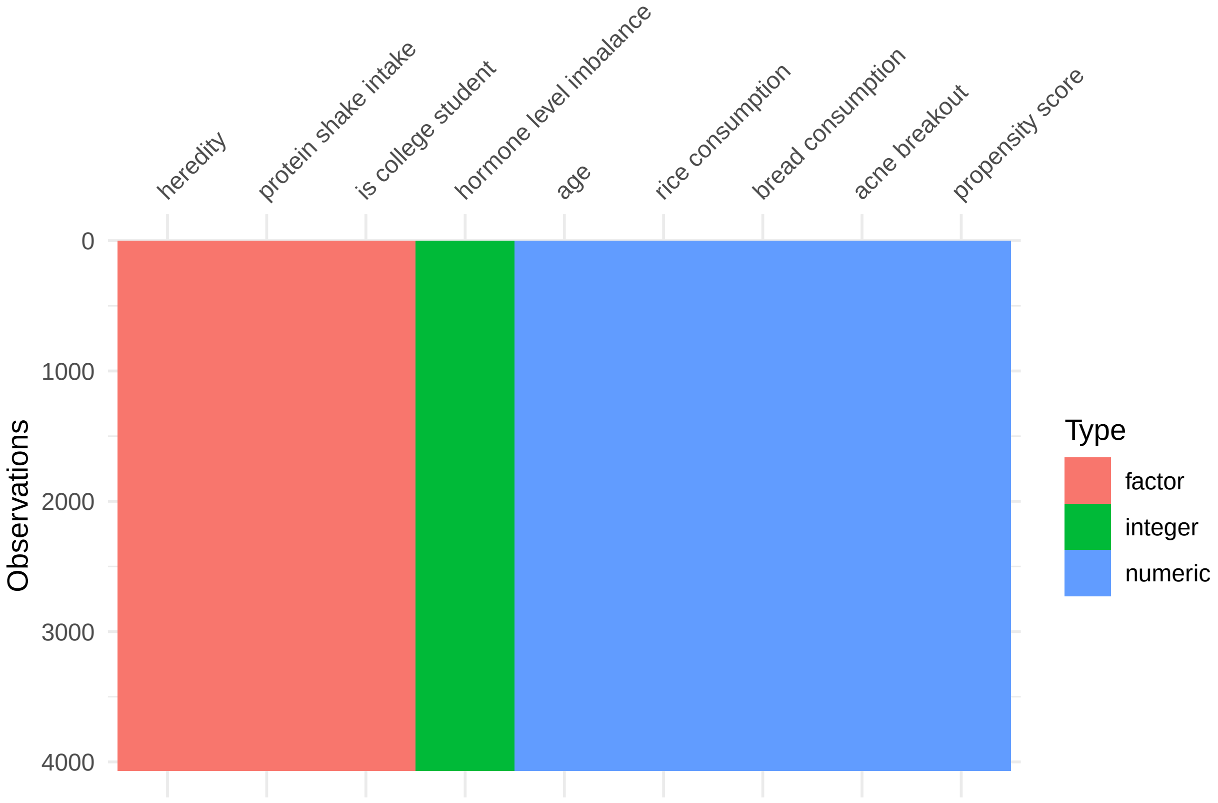

Figure 2: Data after Propensity Score Matching

After propensity score matching, the original data set that contains 100,000 observations is shrank into roughly 4000 observations with equal number of data points in the treatment group (hormone imbalance) and controlled group (hormone balanced). Figure 2 illustrate the property of each covariate as well as the response variable and the conclusion is based on this data set.

Table 1: Unmatched Balance Score

chi-square	df	p-value
5293.215	6	0

Before doing the propensity score matching, I perform a chi-square test to determines the degree of imbalance of the covariates and we have above results. The null hypothesis represents all covariates i the data set are balanced, in other words, it implies that we do not need to strip away too much data when perform the propensity score matching. In contrasts, the alternative hypothesis is that there exists at least one covariate in the data set that is imbalance between the data in treatment group and controlled group. From Table 1, we see that the p value of 0 indicates the null hypothesis is rejected and implies there are imbalance between the covariates.

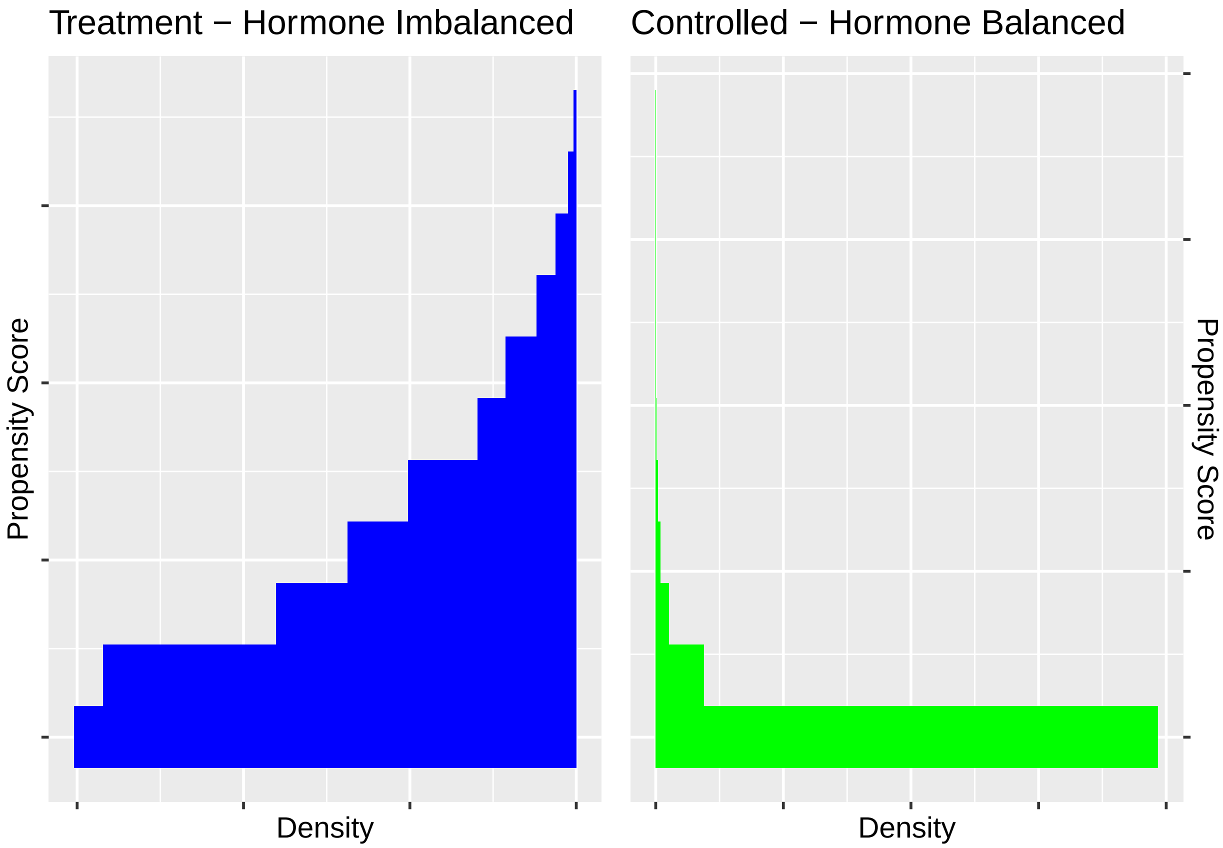

Figure 3: Distribution of Propensity Scores Before Matching

Figure 3 is a visualization used to help illustrate the imbalance covariates situation. The back-to-back histogram shows that the distribution of propensity scores for the treatment (indicates hormone imbalance) and controlled group (indicates no hormone imbalance) are entirely different, implies that requires matching to balance out the covariates between the two groups in order to confidently interpret the causation result. The x-axis is the frequency of each specific propensity score and the y-axis gives each propensity score.

Table 2: Matched Balance Score

chi-square	df	p-value
8.291723	6	0.2175005

Table 2 shows the degree of imbalance in the data set after matching. we see that the chi-square statistics is much smaller and the p value of 21.8% implies that the null hypothesis is accepted. In other words, after matching, the covariates in the data set are imbalanced.

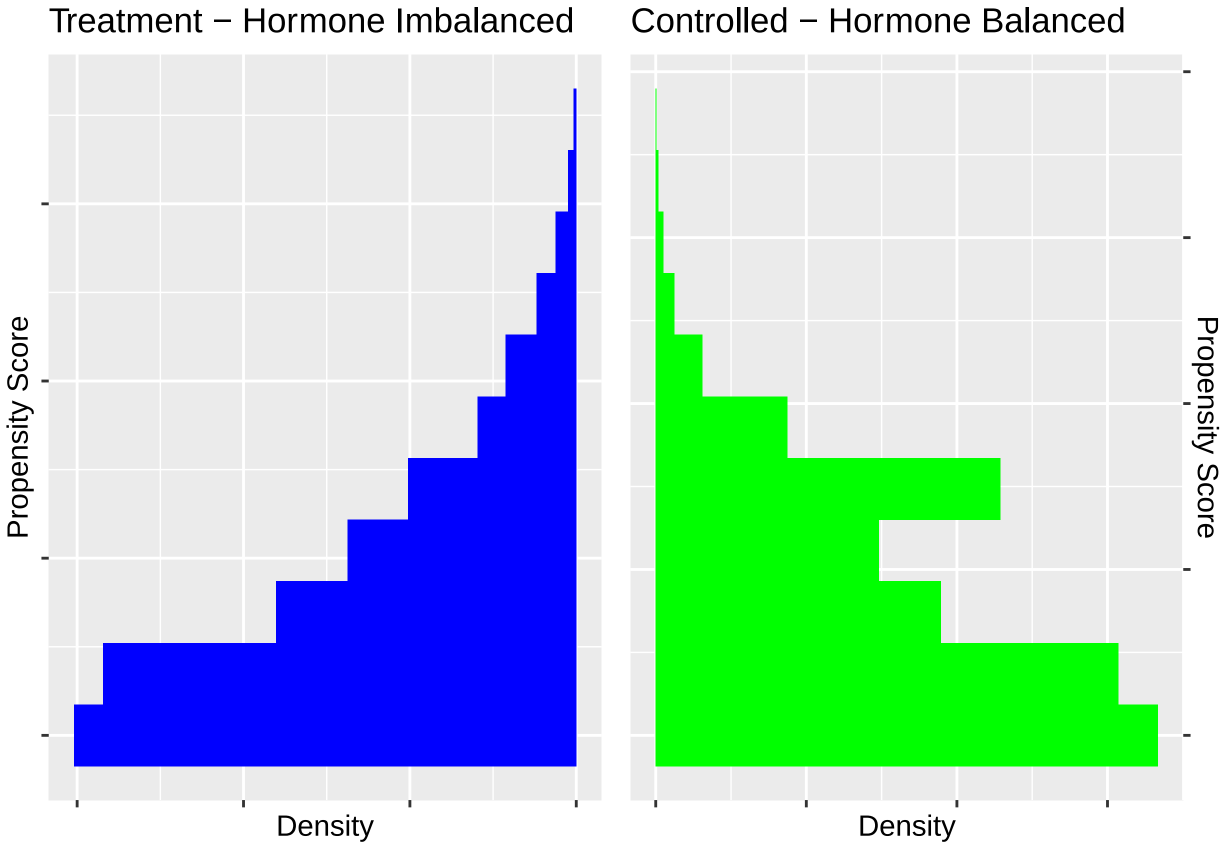

Figure 4: Distribution of Propensity Scores After Matching

After matching, figure 4 shows that the distribution of propensity scores between the two groups are roughly similar. Therefore, we are more confident, at this stage, that any difference in outcome between the two groups can be attributed to the "intervention" on the treatment group, the hormone level imbalance. The x-axis is the frequency of each specific propensity score and the y-axis gives each propensity score.

Table 3: Causal Inference Statistics Table

Predictor of Interest	Logit Effect
Intercept	-2.1837313478926 **
Hormone Imbalance	2.89268322627829 ***

Table 3 above shows only the predictors of interest as well as the reference level of all predictors (the intercept) in this study. In technical term, logit number represents the effect of hormone level imbalance on the odds of getting acne breakout. In other words, it is the same as saying $\log \frac{p_a}{1-p_a} = 2.89268322627829$, where p_a is the probability of having acne breakout. With some rearrangement, we can conclude that $p_a \approx 94$. That is, if a person has hormone level imbalance, the probability of having acne breakout is roughly 94%. The "***" indicates there are strong evidences in the matched data that the effect is statistically significant – with p value less than 5%. In contrast, for the reference level (the intercept), where the person has a balanced hormone, not drink protein shake, not consume excess amount of rice and bread, does not have acne genetic from parents, and not a stressed college student, the probability of having acne breakout immediately drops to roughly 10%, by evaluate $\frac{exp(-2.184)}{1+exp(-2.184)}$.

Discussion

The casual inference analysis is done based on the data (Figure 1) after propensity score matching. The number of observations is significantly reduced, specifically, 60% of reduction. However, we gain the advantage of having two balanced groups in terms of observed covariates and thus imitate the nature of random assignment in experimental study by removing the initial selection bias[3]. It is this characteristics that give researchers the higher confidence to conclude the causation between two variables of interest, in this case, the hormone level imbalance and acne breakout.

Summary

We have done a simulated data set (Figure 1), where each row represents a set of features of an individual as well as the “treatment”, the hormone level imbalance categorical variable. From there, I used a logistic regression to find the propensity score for each individual, which represents the probability of assign the person to the treatment group (hormone imbalance) given the observed features/covariates. At the subsequent stage, I perform the matching by the nearest neighbors approach, which is a pair-wise matching technique that for each propensity score, find two individuals in two different groups that has the difference less than some small threshold. After the matching stage, the number of observations are shrank to only around 4,000 (Figure 2) with equal split between the treatment and control groups, with the average distribution of the observed covariates in each group approximately the same, except the people in the treatment group have imbalanced hormone level. Thus, I am more confident that any effect between the two groups could attributes to the only intervention between the two groups, hormone imbalance. However, the final data for casual inference analysis is much smaller to work with than the original one. Subsequently, I perform an another logistic regression on the acne breakout variable with the rest of the predictors including the categorical hormone imbalance variable (the variable of interest). At the end, the statistical significant estimation on the hormone imbalance variable indicates that there is indeed a causal link between the acne breakout and hormone imbalance. That is, if a person has hormone imbalance, it would cause the individual having higher chance of getting acne breakout.

Conclusion

The propensity score analysis with nearest neighbors matching indicates that the hormone level imbalance of an individual is one of the casual factor that leads to higher risk of having acne breakout. In particular, the result section shows that the odds of having acne breakout is 2.83 higher if a person has hormone imbalance, in other words, there is approximately 94% of chance having acne breakout if hormone level imbalance happens. Furthermore, the statistically significant p value indicates that the result from this analysis rejects the null hypothesis — the hormone imbalance has no effect on the acne breakout, in other words, there are strong evidence from the data that the hormone imbalance does have effect on acne breakout.

In conclusion, from the analysis of the paper, we already know that hormone imbalance does have causal effect on the acne breakout. From study, diet is one of the associated factor that has impact on the hormone level, either restore it back to balanced level or make it imbalance[14]. The implication is that to avoid having acne breakout or inflammation, we could make use of a proper diet to prevent acne breakout or at least alleviate the severity and prevent it from getting worse. From the introductory section, I have demonstrated varies studies on the effect of different food type impact on the hormone level, including protein shake, high glycemic index food such as white rice and bread. Therefore, if readers of this paper suffer from acne breakout or inflammation, beside consult to the dermatologists, we could be on guard about the daily diet intake to complement the existing treatment such as avoid the food mentioned in this paper that mess up the hormone level, and ultimately make the healing period faster.

Weakness

Although when simulate the data set, I tried to use relevant finding from literatures to make educated guess on each distribution parameter. However, when I could not find any relevant information, such as the proportion of people who drink protein shake when they workout, I only able to make the parameter predictions based on my personal experience and surrounding. Therefore, some hidden bias of the simulated data are inevitable.

There are limitations about the propensity score method. Firstly, when we are deriving the casual inference, we assume that the distribution of observed covariates between the two groups are similar. However, the unobserved covariates of each individual also plays a factor in real setting. That is, there is a possibility that the confounding variable is one of the unobserved covariates and thus our analysis only able to conclude the association instead of causation of the intervention. Secondly, the propensity score matching method requires huge amount of data so that the overlap region of propensity scores is large enough to perform a better analysis. Relevant studies show that collect 3-4 times more sample data for the control group than the treatment group is recommended to ensure large enough overlap of propensity scores[3] based on the intuition that the probability of matching each propensity score in the treatment group to the controlled group is higher.

Next Step

A sensible next step would be perform a sensitivity analysis of the result. In other words, when we perform the matching using the propensity score, we have to remove the unmatched data and thus the final result is subject to hidden bias that are not identified by us. Therefore, some sensitivity analysis could be used to measure how robust the result is. For example, the output from Wilcoxon Signed Rank Test by Rosenbaum represents "how much the odds need to be change before the statistical significance of the outcome shifts"[3].

References

1: Gardner, S. (2020, November 12). Teen Acne: Causes, Symptoms, Treatments, & More. Retrieved December 09, 2020, from https://www.webmd.com/skin-problems-and-treatments/acne/what-is-acne

2: Cashin-Garbutt, I. (2018, August 23). How does acne affect self-confidence? Retrieved December 09, 2020, from https://www.news-medical.net/news/20170508/How-does-acne-affect-self-confidence.aspx

3: Olmos, A., & Govindasamy, P. (2015). Propensity Scores: A Practical Introduction Using R. Retrieved December 17, 2020, from https://journals.sfu.ca/jmde/index.php/jmde_1/article/view/431/414

4: Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental design for generalized causal inference. Boston: Houghton Mifflin Company.

5: Caliendo, Marco, and Sabine Kopeinig. "SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PROPENSITY SCORE MATCHING." Wiley Online Library, John Wiley & Sons, Ltd, 31 Jan. 2008, onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6419.2007.00527.x?casa_token=vJSIHkwqM6oAAAAA%3AoZSbl
P1Su8N6vQ.

6: Austin, Peter C. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." Multivariate Behavioral Research, Taylor & Francis, May 2011, www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/.

7: Jaliman, D. (2019, May 17). Acne Causes: What Is Acne and Why Do I Have It? Retrieved December 18, 2020, from https://www.webmd.com/skin-problems-and-treatments/acne/understanding-acne-basics

8: Whelan, C. (2020, May 18). Is Acne Genetic? Learn About Hereditary Risks Factors. Retrieved December 18, 2020, from https://www.healthline.com/health/is-acne-genetic

9: Thomas, D. (2018, November 16). Genetics of Acne. Retrieved December 18, 2020, from https://www.news-medical.net/health/Genetics-of-Acne.aspx

10: Shiffer, E. (2019, February 25). Your Protein Shake Might Be Giving You Acne. Retrieved December 18, 2020, from https://www.menshealth.com/health/a19546380/protein-shakes-and-acne/

11: Major U.S. Crops: Rice. (n.d.). Retrieved December 18, 2020, from http://www.crosscurrents.hawaii.edu/content.aspx?lang=eng

12: Three bread trends shaping American diets. (n.d.). Retrieved December 19, 2020, from https://www.world-grain.com/articles/8702-three-bread-trends-shaping-american-diets

13: University of Wisconsin Hospitals and Clinics Authority. (n.d.). Acne and Your Diet: How the Glycemic Index Affects Your Skin. Retrieved December 19, 2020, from https://www.uwhealth.org/news/acne-diet-glycemic-index/46499

14: Boyers, L. (2019, April 01). How Your Diet Affects Your Hormones. Retrieved December 21, 2020, from https://www.healthline.com/health/menopause/diet-hormones

Package Reference

Wickham, H. (2019, November 21). Easily Install and Load the ‘Tidyverse’ [R package tidyverse version 1.3.0]. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/tidyverse/index.html

Wickham, H. (2020, August 18). A Grammar of Data Manipulation [R package dplyr version 1.0.2]. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/dplyr/index.html

Hayes, A. (2020, December 16). Convert Statistical Objects into Tidy Tibbles [R package broom version 0.7.3]. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/broom/index.html

Gelman, A., & Su, Y. (2020, July 27). Data Analysis Using Regression and Multilevel/Hierarchical Models [R package arm version 1.11-2]. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/arm/index.html

Wilke, C. (2020, September 08). Streamlined Plot Theme and Plot Annotations for ‘ggplot2’ [R package cowplot version 1.1.0]. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/cowplot/index.html

Bowers, J., Fredrickson, M., & Hansen, B. (n.d.). Package RItools. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/RItools/index.html

Tierney, N. (2019, February 15). Preliminary Visualisation of Data [R package visdat version 0.5.3]. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/visdat/index.html

Dowle, M. (2020, December 08). Extension of ‘data.frame’ [R package data.table version 1.13.4]. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/data.table/index.html

Hugh-Jones, D. (2020, October 27). Easily Create and Style Tables for LaTeX, HTML and Other Formats [R package huxtable version 5.1.1]. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/huxtable/index.html

Xie, Y. (2020, September 22). A General-Purpose Package for Dynamic Report Generation in R [R package knitr version 1.30]. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/knitr/index.html