Relationship between lifestyle risk factors and development of prediabetes or diabetes, based on BRFSS 2015 questionnaire

Tianyang Jiang, Changhao Jiang, Liuye Huang

May 31, 2023

Abstract

This paper is about diabetes...

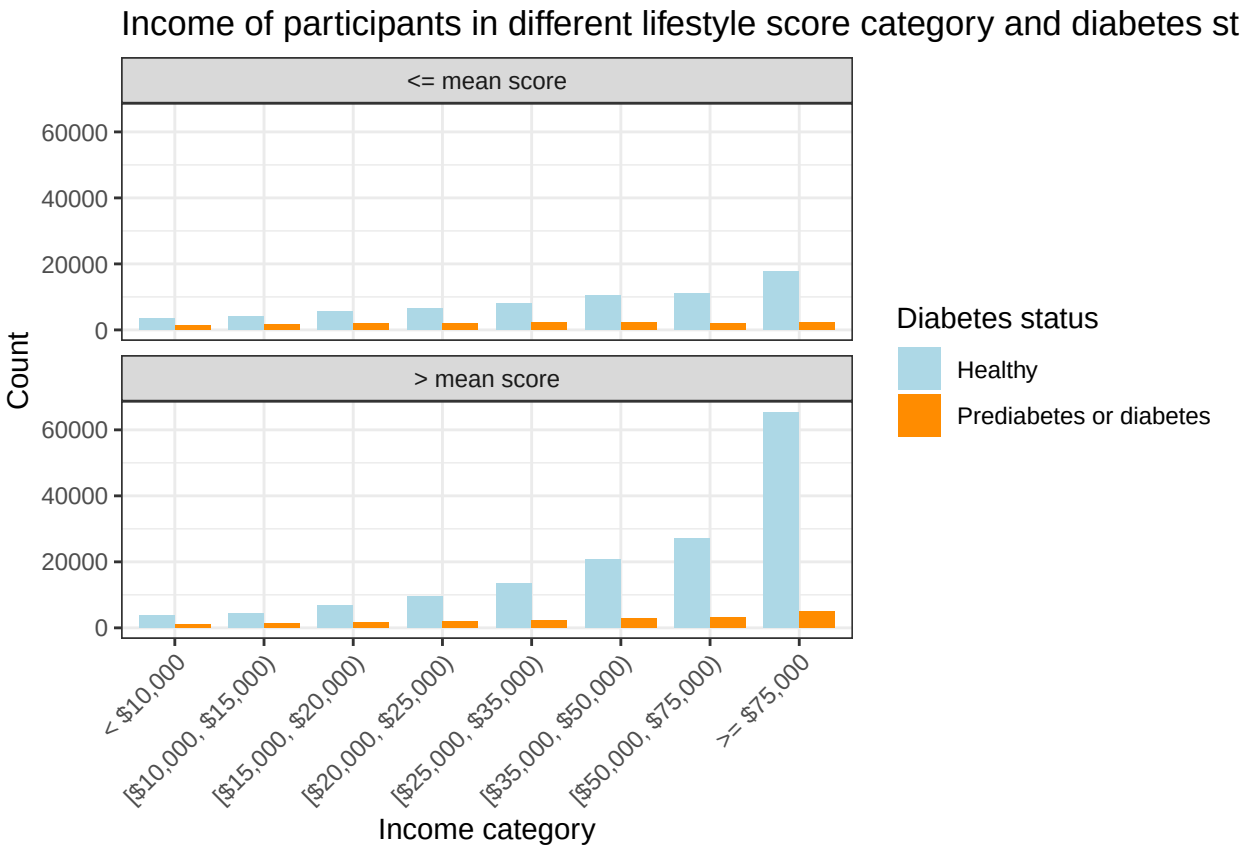
Keywords: Diabetes, Lifestyle intervention, Causal inference, R-learner, PC algorithm, Regression adjustment

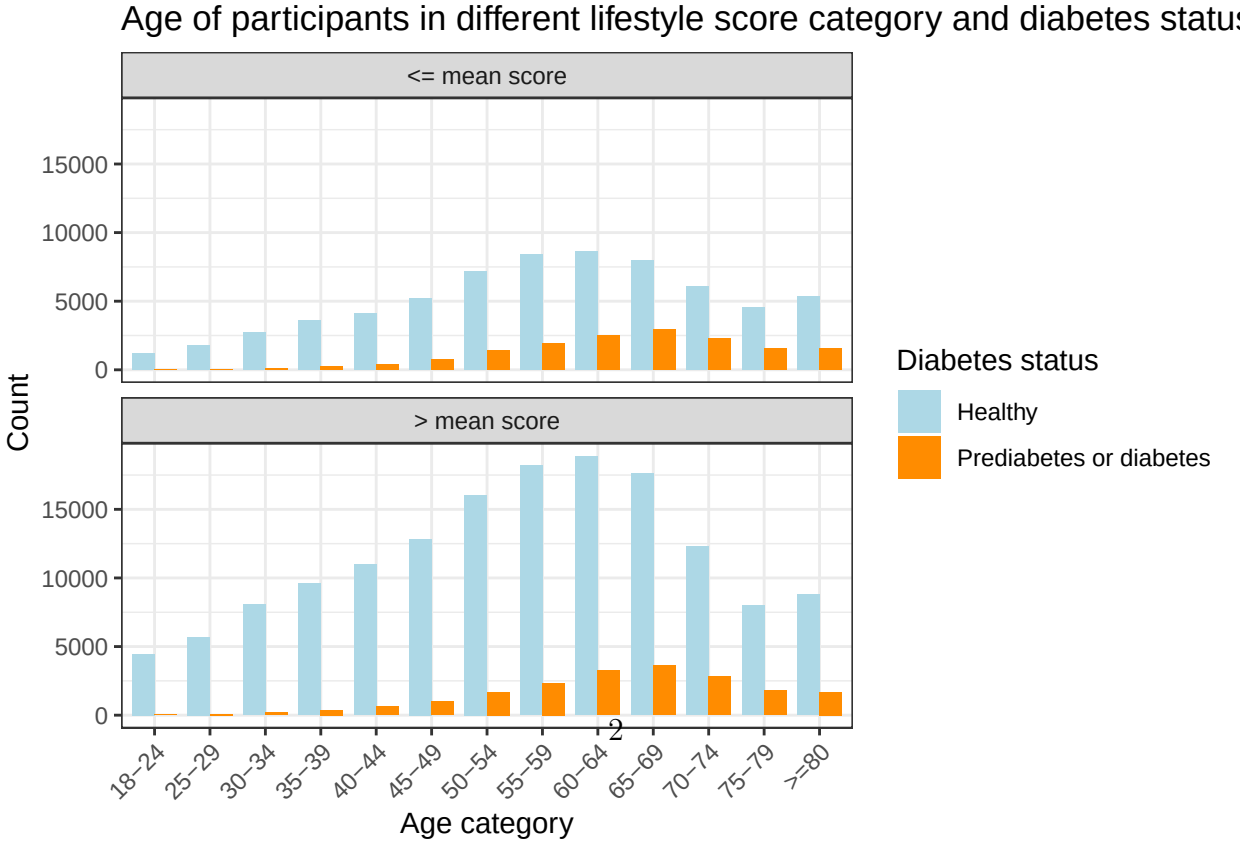
Introduction

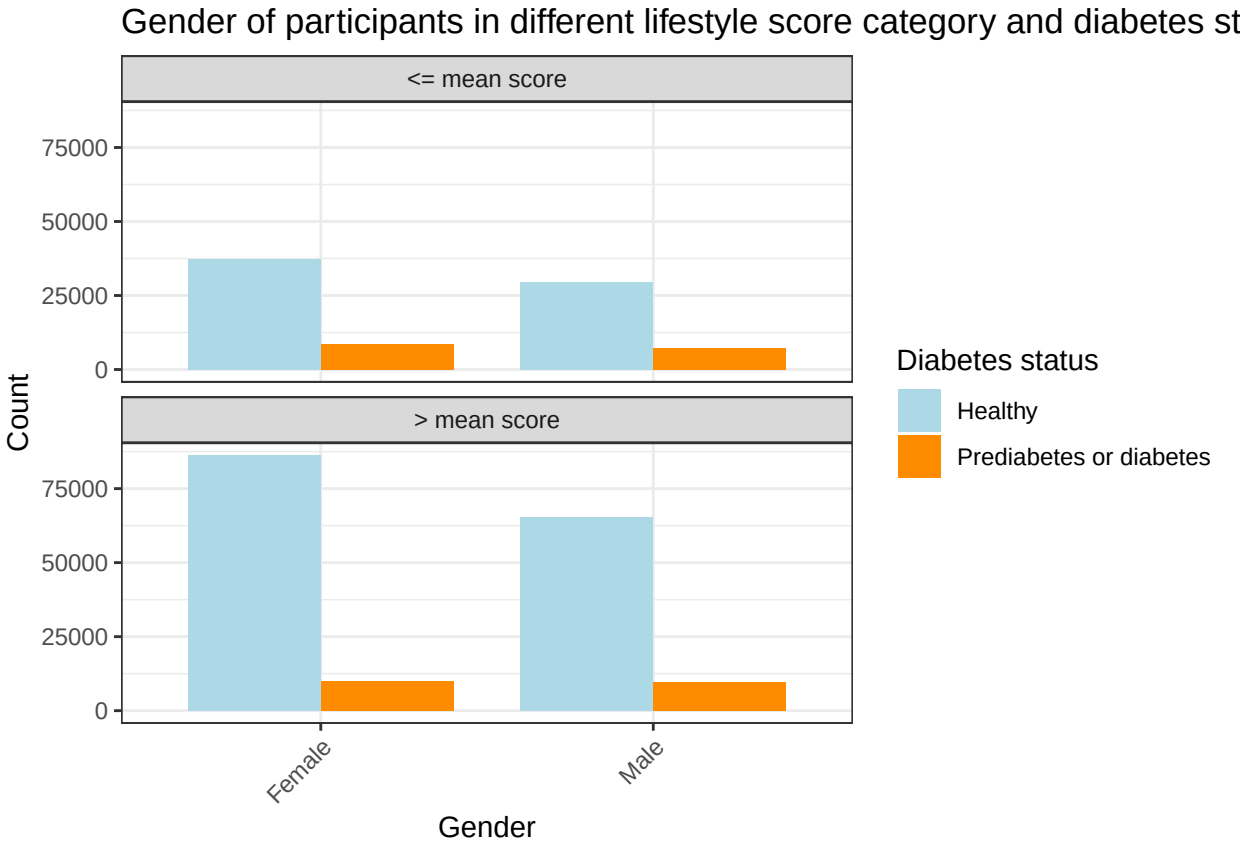
[Provide an introduction to the research topic, including the background, motivation, and research questions/hypotheses.]

Background

Data Description







Statistical Methods

[Explain the statistical methods employed in the analysis, such as linear regression, Bayesian analysis, etc. Provide a brief rationale for their selection.]

Find Scores with Random Forest

Build DAG with PC Algorithm and Literature Review

Logistic Regression with Adjustment

Table 1: Adjustment Results

	Estimate	2.5 %	97.5 %
score	-0.5981	-0.6210	-0.5752
adjusted score	-0.3775	-0.4016	-0.3533

R-learner

In this section, we investigate the heterogeneous effects of lifestyle factors on subgroups defined by age, sex, and income using the R-learner (Nie and Wager, 2017). This approach utilizes machine learning to estimate treatment effects in observational studies.

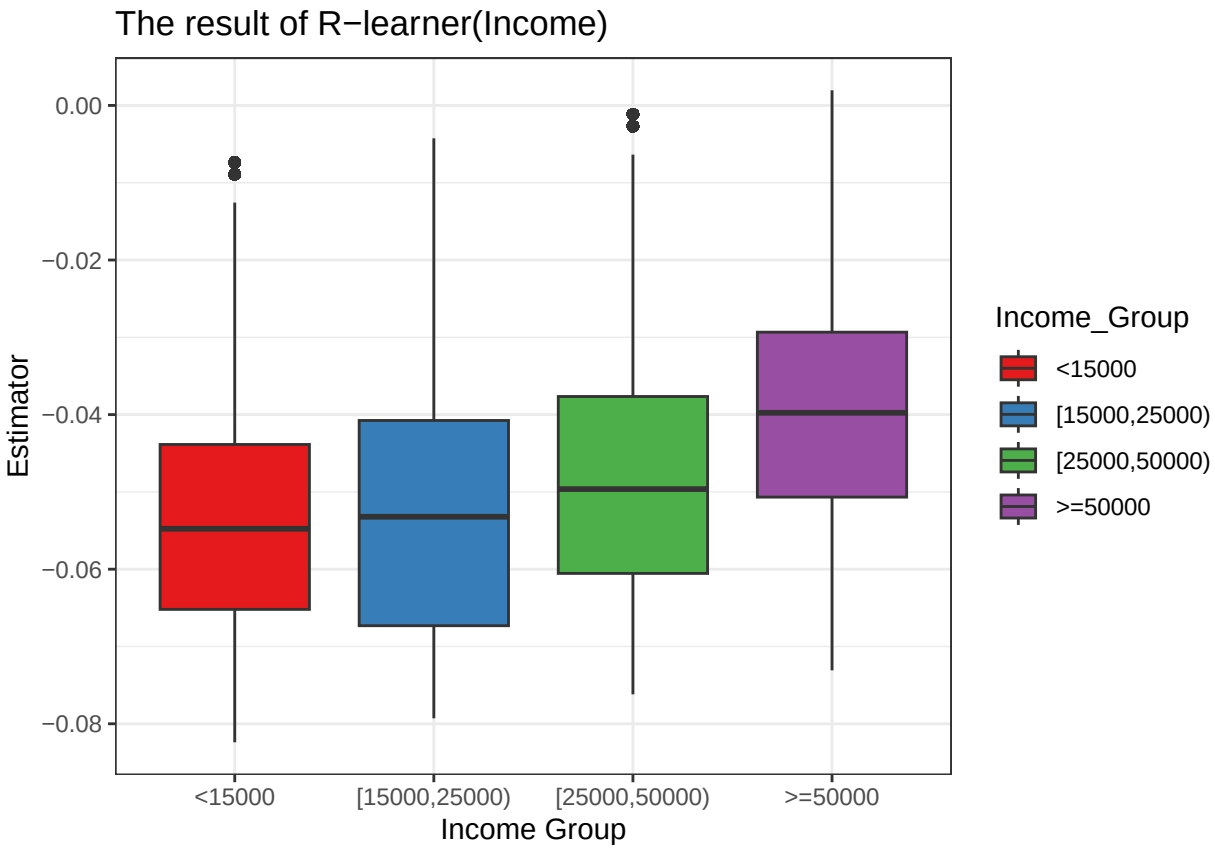
The R-learner is used to estimate the Conditional Average Treatment Effect (CATE), $\tau^*(z_1, z_2, z_3) = E(Y(1) - Y(0)|Z_1 = z_1, Z_2 = z_2, Z_3 = z_3)$, with Z_1, Z_2, and Z_3 denoting individual features (age, income, and gender) and Y_i the observed outcome.

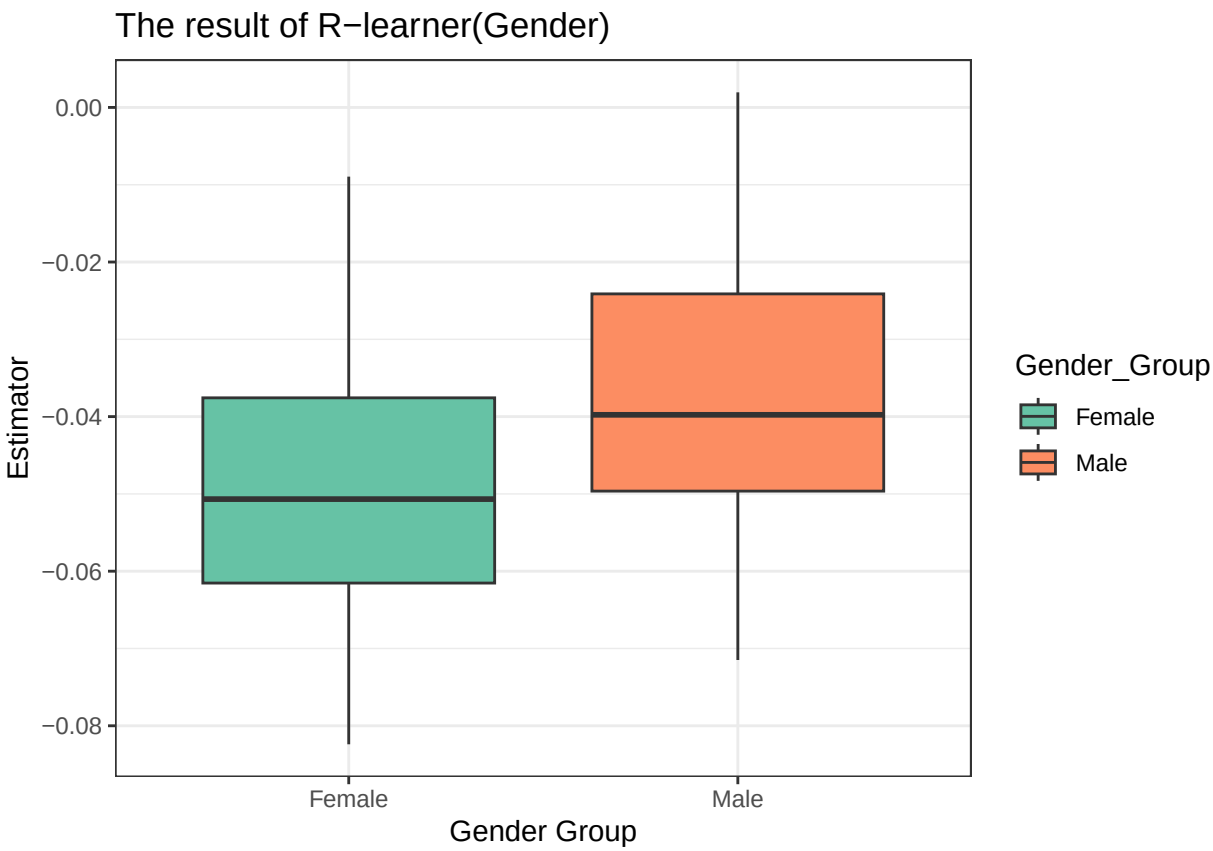
Using the `rlasso` method in `rlearner`, we input age, gender, and income as features, and a categorical score as the treatment variable, with diabetes presence indicating observed response. The algorithm of R-learner also includes cross-validation, we set it with ten folds. Finally, we visualize CATE CATE across income levels, gender, and age groups.

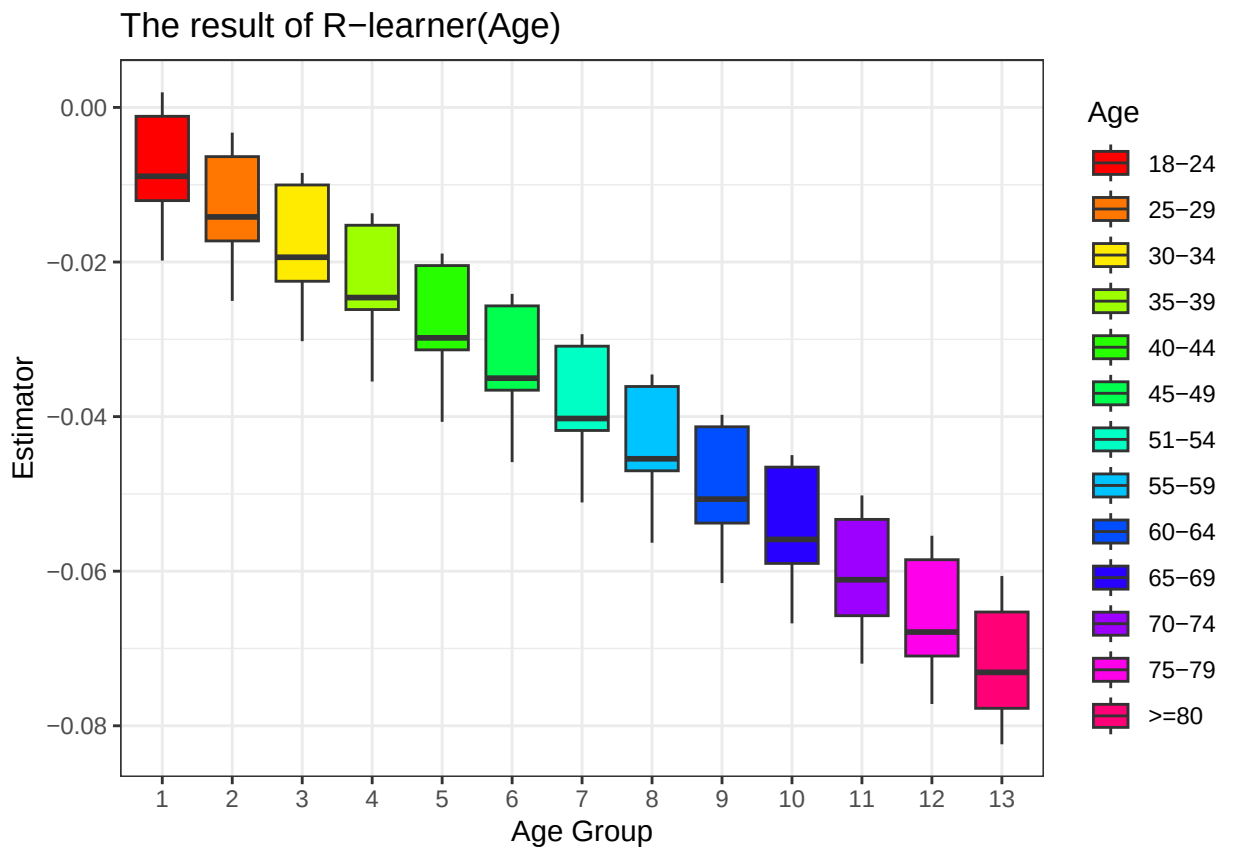
In the visualization of CATE across different income groups, we observe a correlation between higher income and a larger average treatment effect within the same income bracket. This trend is particularly pronounced among individuals earning over $50,000 annually, where the mean effect significantly surpasses that of other groups.

When comparing CATE across genders, it is noteworthy that the average treatment effect for females is lower than that for males.

The CATE across age groups shows a distinct decline as age increases. For individuals younger than 24, there is little discernible difference in the outcome between those maintaining good or poor lifestyles.


The result of R−learner(Income)
0.00
−0.02
−0.04
−0.06
−0.08
Estimator
<15000
[15000,25000)
[25000,50000)
>=50000
Income Group
Income_Group
<15000
[15000,25000)
[25000,50000)
>=50000


The result of R−learner(Gender)
0.00
−0.02
−0.04
−0.06
−0.08
Estimator
Female
Male
Gender Group
Gender_Group
Female
Male



Results

From the R-learner plots, we discern that higher income individuals are less likely to develop diabetes, even have poor lifestyle choices. This pattern is particularly noticeable for those earning over $50,000 annually. Regarding gender, the difference in diabetes risk between good and poor lifestyle habits is more pronounced in females than males, underscoring the importance of healthy habits for women. In terms of age, the substantial variation in CATE across different age groups suggests that maintaining a healthy lifestyle becomes increasingly vital as one ages. This is because poor lifestyle choices are more likely to induce diabetes as age increases.

Discussion

[Discuss the implications and significance of the findings, relate them to the research questions/hypotheses, and address any limitations or potential sources of bias.]

Conclusion

[Summarize the key takeaways from the study and suggest possible avenues for future research.]

References

[Include a list of cited references using a suitable citation style (e.g., APA, MLA, IEEE).]

Appendix