Project submission - Causal Inference 097400

Jonathan Haber Afek , Gefen Pustelnic

link to our code repository:

[CasualInferenceProject/main.ipynb at main · gefen1999/CasualInferenceProject · GitHub]

Introduction

Our problem is exploring the causal effect of engaging in sports on students' GPA final grades. The key focus is on uncovering whether there is a direct causal relationship between doing sports and students' final academic outcomes and whether other factors might mediate this effect.

The underlying question is whether the positive aspects of doing sports—such as improved mental health, enhanced cognitive function, better time management, and reduced stress—translate into better academic performance. Alternatively, participating in sports can also have negative effects, such as reducing the time available for study, which maybe leading to poorer academic results.

This project seeks to estimate the causal effect of sports participation on GPA. The key challenge lies in isolating the influence of sports from other factors that might affect academic performance, such as socioeconomic background, study habit and more. Establishing a clear causal link requires careful control for confounding variables and the use of appropriate methodologies, such as regression models and propensity score matching.

Additionally, this topic is particularly relevant to us as students at the Technion who regularly engage in sports. As students managing a heavy academic workload alongside consistent physical activity, we experience the challenge of balancing study time with time spent on sports. We aim to better understand the potential consequences of our daily choices—not only to optimize our own routines but also to help inform other students facing similar circumstances.

Ultimately, understanding this relationship can provide valuable insights for educational policy and student development programs, allowing institutions to better balance extracurricular and academic activities.

Data

The dataset includes detailed information on 2,392 high school students, covering a range of factors such as demographics, study habits, parental involvement, extracurricular activities, and academic performance. The target variable is **GradeClass**. This comprehensive dataset enables a thorough exploration of the relationships between various student characteristics and academic outcomes, making it a valuable resource for investigating patterns and estimating causal effects in educational contexts.

Tables and Ranges:

StudentID: A unique identifier assigned to each student (1001 to 3392).

Age: The age of the students ranges from 15 to 18 years.

Gender: 0 represents Male and 1 represents Female.

Ethnicity: 0: Caucasian, 1: African American, 2: Asian, 3: Other

ParentalEducation: 0: None, 1: High School, 2: Some College,

3: Bachelor's, 4: Higher

StudyTimeWeekly: Weekly study time in hours, from 0 to 20.

Absences: Number of absences during the school year, from 0 to 30.

Tutoring: Tutoring status, where 0 indicates No and 1 indicates Yes.

ParentalSupport: 0: None, 1: Low, 2: Moderate, 3: High 4: Very High

Extracurricular: Participation in extracurricular activities, where 0 indicates No and 1 indicates Yes.

Sports: Participation in sports, where 0 indicates No and 1 indicates Yes.

Music: Participation in music activities. 0 indicates No and 1 indicates Yes.

Volunteering: Participation in volunteering, where 0 indicates No and 1 indicates Yes.

GPA: Grade Point Average on a scale from 0 to 4.0.

Target Variable: Grade Class

GradeClass: Classification of students' grades based on GPA:

0: 'A' (GPA >= 3.5), 1: 'B' (3.0 <= GPA < 3.5), 2: 'C' (2.5 <= GPA < 3.0),

3: 'D' (2.0 <= GPA < 2.5), 4: 'F' (GPA < 2.0)

Note: The lower the value of GradeClass – the higher the grades.

We assumed an inverse relationship between sport and the GradeClass.

Challenges

1. Missing Variables:

- Challenge: The dataset lacks important lifestyle factors like **sleep quality** and **nutrition**, which are likely correlated with both sports participation and GPA.
- How to Deal: We will acknowledge this as a limitation in our analysis, and we will explore proxy variables such as Absences or StudyTimeWeekly to indirectly capture lifestyle factors. Additionally, we will also suggest future research to include these factors for a more precise causal estimation.

2. Confounding Variables:

- Challenge: Variables such as **ParentalSupport**, **Gender**, **Age**, **Ethnicity**, and **ParentalEducation** could confound the relationship between sports participation and GPA. For example, students with higher levels of parental support or more educated parents may be more likely to participate in sports and also perform better academically. Similarly, demographic factors like gender and ethnicity might influence both sports participation and GPA.
- How to Deal: We will control for these confounders by including them in our model. This will allow us to isolate the effect of sports on GPA while accounting for these background characteristics.

3. Post-Treatment Bias:

- Challenge: Variables like **Volunteering**, **Music**, **Extracurricular**, **Tutoring**, **Absences and Study Time Week** could be influenced by sports participation, making them potential post-treatment variables.
- How to Deal: We will create two models: one that excludes these variables to estimate the direct effect of sports on GPA, and another that includes them to estimate the total effect. That will allow us to understand how sports might impact GPA via other factors.

4. Lack of quality Data

- **Challenge**: The dataset lacks detailed information on the quality of important variables. For example, **StudyTimeWeekly** captures only the quantity of study time, not the quality or effectiveness of that time. Similarly, **Sports** is recorded as a simple binary indicator (Yes/No), without details on the type, intensity, or duration of sports participation. Additionally, the **Tutoring** variable does not provide insights into the quality of the tutoring sessions, which could significantly impact academic outcomes.

- **How to Deal**: We will acknowledge these limitations in the analysis and consider them as potential areas for bias. Where possible, we may suggest future research to collect more detailed data on these variables for a more nuanced understanding of their effects on GPA. In the current analysis, we will proceed with caution, interpreting results in the context of these limitations.

Assumptions

Stable Unit Treatment Value Assumption (SUTVA)

1. No Interference Between Units

- **Holds**: In our context, the assumption that there is no interference between units is reasonable. Each student's GPA is likely to be influenced by their own sports participation rather than the sports participation of other students. For example, a student's GPA does not directly depend on whether their peers are involved in sports. Hence, we can safely assume that this condition holds in our analysis, meaning there's no spillover effect between students regarding the treatment (sports participation).

2. No Hidden Variations in Treatment

- **Does Not Hold**: The second part of SUTVA, which requires that all students receiving the treatment (sports participation) experience it in the same way, is more problematic in our case. The dataset only provides a binary indicator for sports participation (Yes/No), without capturing the intensity, type, or duration of the sports activity. This hidden variation in treatment (e.g., some students playing casual sports, while others may engage in competitive, high-commitment sports) violates the second condition of SUTVA. As a result, we cannot assume that the treatment effect is the same across all students who are marked as participating in sports, which introduces potential bias into our analysis.

By recognizing that the second condition of SUTVA does not hold, we are aware of the limitations this introduces in estimating the true causal effect of sports on GPA. This is acceptable for our analysis because we are still able to understand the general influence of sports participation on academic performance, which is the primary goal of our analysis.

Consistency

The Consistency Assumption ensures that the observed outcome for each individual corresponds to the treatment they actually received. In our case, this means that if a student is marked as participating in sports, their GPA reflects the outcome under that condition, and if they are not participating, their GPA reflects the outcome under the "no sports" condition.

This assumption **holds** in our analysis because the dataset consistently categorizes each student as either participating in sports or not, and their GPA is linked to their actual experience. Despite the lack of detailed information on the intensity or type of sports, the binary distinction of participation aligns with the outcomes we observe, allowing us to interpret the relationship between sports participation and GPA correctly.

No Unmeasured Confounders Assumption

The assumption is crucial in causal inference, requiring that all variables influencing both the treatment (sports participation) and the outcome (GPA) are measured and included in the model. In our case, this assumption **holds** reasonably well because key variables like Parental Support, Gender, Age, Ethnicity, and Parental Education—which can influence both sports participation and GPA—are included in the dataset, helping to account for potential confounding factors.

Although certain factors like **sleep** and **diet** are not recorded, they may not violate this assumption as they could be influenced by sports participation (post-treatment variables) rather than confounders affecting both sports and GPA beforehand. Additionally, some of these lifestyle factors might be partially explained by other included variables, such as **StudyTimeWeekly** or **Absences**, which can act as proxies, further reducing the risk of omitted confounders.

Common Support Assumption

The assumption requires that for every level of covariates, there are individuals in both the treatment group (sports participants) and the control group (non-participants). This ensures that we can make valid comparisons between students who participate in sports and those who do not, across a similar range of covariates (e.g., age, gender, parental support).

In our case, the Common Support Assumption **holds** because, as shown in plot number 1, the distribution of propensity scores for both sports participants and non-participants overlaps significantly. This indicates that for most students, regardless of their characteristics, there are comparable individuals in both groups, allowing for a valid estimation of the causal effect of sports on GPA. The graph demonstrates that there is a shared support area where the covariates are well-represented in both groups, affirming that common support is satisfied in our analysis.

No Reverse Causality Assumption

The assumption ensures that the treatment (sports participation) affects the outcome (GPA), but not the other way around. In our case, this assumption **holds** because the GPA is recorded after the sports participation has taken place. This temporal sequence guarantees that the GPA reflects academic performance following the period of sports involvement, eliminating the possibility that a student's GPA influenced their decision to participate in sports. Thus, we can confidently interpret sports participation as influencing GPA, rather than GPA determining sports participation.

The assumptions we rely on allow us to estimate the causal effect of sports participation on GPA. They ensure that the relationships we observe between sports and GPA are not driven by external factors, unmeasured variables, or reverse causality. While some limitations exist, such as the lack of detailed information on the intensity of sports, we can still draw meaningful conclusions about the overall effect of sports on academic performance. By accounting for key confounders, ensuring consistency, and confirming the temporal order of variables, these assumptions provide a solid foundation for causal inference in this analysis.

Methods

ATE (Average Treatment Effect): The average difference in outcomes between a treatment group and a control group (Those who participate in sport and those who do not), assessing the overall impact of a treatment across the population.

ATT (Average Treatment Effect on the Treated): The average effect of a treatment specifically on those who received it, highlighting the treatment's impact on that subgroup

Propensity Score Matching:

Method: Propensity score matching calculates the probability (propensity score) of each individual receiving the treatment (sports participation) based on covariates. Individuals from the treated group are matched with individuals from the control group with similar propensity scores.

Purpose: This method reduces selection bias by creating comparable groups, allowing for the estimation of **ATE** and **ATT** based on matched samples.

Calibration:

Method: Calibration ensures that the predicted probabilities from the propensity score model are well-calibrated to reflect the actual treatment assignment probabilities. This improves the reliability of subsequent matching and IPW estimations.

Purpose: Properly calibrated propensity scores are crucial for accurate estimation of both **ATE** and **ATT** by reducing bias in matching and weighting techniques.

Bootstrap Sampling:

Method: Bootstrap sampling involves repeatedly resampling the data with replacement and recalculating treatment effects. This provides a distribution of the estimated **ATE** and **ATT**, enabling confidence interval (CI) calculation.

Purpose: Bootstrapping improves robustness by estimating the variability of **ATE** and **ATT** and producing confidence intervals for these estimates.

Root Mean Squared Error (RMSE):

Method: RMSE evaluates the predictive performance of the models, indicating how well the models fit the training and validation data.

Purpose: By ensuring the models used for causal effect estimation are well-calibrated and accurate, RMSE helps validate the quality of the ATE and ATT estimates.

S-Learner:

Method: The S-Learner trains a single machine learning model (random forest) on the covariates along with the treatment indicator (sports participation). The model predicts outcomes under both the treatment and control conditions by altering the treatment indicator.

Purpose: Estimates both **ATE** and **ATT** by comparing predicted outcomes for treated and untreated groups.

T-Learner:

Method: The T-Learner involves training two separate models—one for the treated group and one for the control group. Each model is then used to predict outcomes for the respective group.

Purpose: The T-Learner allows for more flexibility in estimating both **ATE** and **ATT** by comparing predictions from the two models trained separately on treated and untreated individuals.

Matching:

Method: Matching techniques pair treated and control individuals based on similar covariates, ensuring comparability between the two groups.

Purpose: Matching ensures comparability between treated and control individuals, leading to more accurate estimates of both **ATE** and **ATT**.

Inverse Probability Weighting (IPW):

Method: IPW assigns weights to individuals based on the inverse of their propensity score. This creates a weighted dataset that mimics a randomized experiment, where the weights balance the treatment and control groups.

Purpose: By reweighting the sample, IPW allows for estimation of both **ATE** and **ATT**, providing a more balanced comparison between treated and control groups.

Models

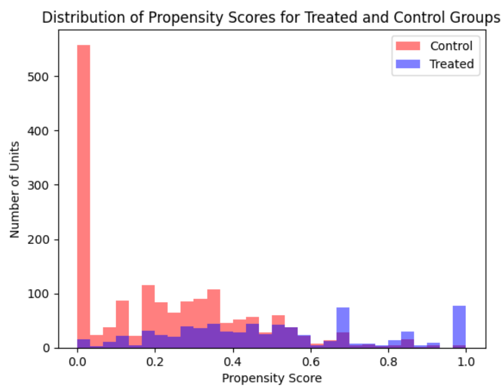
In this project, we selected the Random Forest Regressor and compared its performance to the XGB Regressor across various approaches.

Random Forest Regressor: This model is an ensemble learning method that creates multiple decision trees and averages their predictions, which helps reduce overfitting and provides a robust solution for many regression problems.

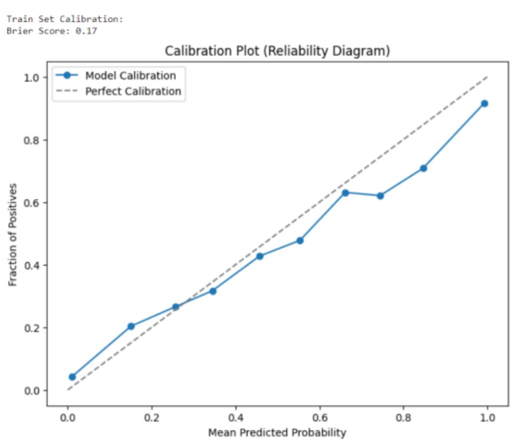
XGBRegressor: This model is based on gradient boosting, where trees are built sequentially, with each new tree correcting the errors of the previous ones, leading to high accuracy, especially in complex datasets.

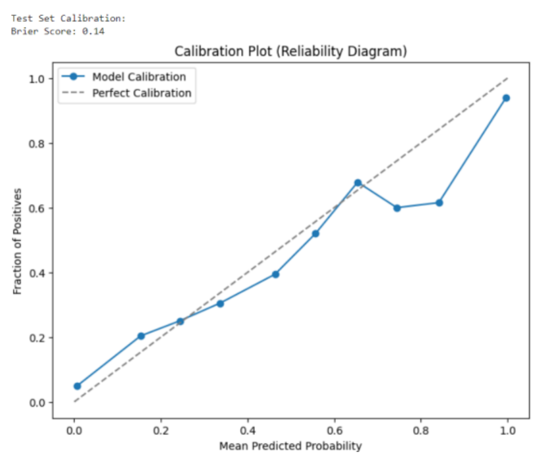
Results

First, we calculated the distribution of the propensity scores because understanding the overlap between the treated and control groups is crucial for reliable causal inference.



there is good overlap between the groups, primarily in the **0.1** to **0.8** range, which supports some comparability. However, we observe a large concentration of control units with very low propensity scores, and a few treated units clustered around higher scores, indicating limited overlap at the extremes.





Then, we examined the calibration of the propensity score model to assess how well the predicted probabilities align with actual outcomes. In the plot on the left, representing the train set, we achieved a Brier score of **0.17**. The test set, shown on the right, has a slightly lower (better) Brier score of **0.14**, suggesting that the model's predictions generalize well to new data. In both plots, the blue line (representing model calibration) closely follows the diagonal line of perfect calibration, indicating that the model's predicted probabilities are generally well-calibrated to the observed outcomes. This supports the reliability of the propensity scores in estimating treatment probabilities

Next, we evaluated treatment effect estimates using S-learner and T-learner models with both RandomForest and XGBRegressor. In the S-learner, both models achieved similar Mean Training and Test RMSE values. The ATE and ATT values are close to zero, with confidence intervals indicating no significant treatment effect. Adding sample weights had minimal impact.

For the T-learner, RMSE values were also similar across models. RandomForest slightly outperformed XGBRegressor in terms of RMSE on the test set. Overall, both learners and models suggest a negligible treatment effect.

S Learner

Score / Model	Mean Training RMSE	Mean Test RMSE	ATE	ATE - CI	ATT	ATT - CI
RandomForest *sample weights*	1.19	1.22	-0.07	[-0.226, 0.066]	-0.06	[-0.223, 0.066]
RandomForest	1.18	1.21	-0.01	[-0.161, 0.043]	-0.01	[-0.162, 0.04]
XGBRegressor *sample weights*	1.12	1.22	-0.08	[-0.24, 0.055]	-0.09	[-0.241, 0.06]
XGBRegressor	1.09	1.20	-0.07	[-0.212, 0.057]	-0.08	[-0.207, 0.055]

T Learner

Score / Model	Mean Training RMSE (Treated)	Mean Training RMSE (Control)	Mean Test RMSE (Treated)	Mean Test RMSE (Control)	ATE	ATE - CI	ATT	ATT - CI
RandomForest	1.11	1.10	1.23	1.17	-0.12	[-0.205, 0.041]	-0.12	[-0.192, 0.053]
XGBRegressor	1.10	1.13	1.24	1.20	-0.02	[-0.222, 0.05]	-0.03	[-0.212, 0.067]

The next table compares the Matching and Inverse Probability Weighting (IPW) methods for estimating treatment effects. Matching shows a Mean RMSE of **1.74**, with ATE and ATT values of **-0.08** and **-0.06**, respectively, and relatively narrow confidence intervals that include zero, indicating no significant treatment effect. IPW, on the other hand, has a lower Mean RMSE of **1.42**, suggesting a better fit than Matching, with ATE and ATT values of **-0.08** and **-0.11**. However, IPW's confidence intervals are wider, still including zero, which implies no significant treatment effect but potentially less stability in the estimates. Overall, all methods suggest a negligible treatment effect, with IPW achieving a slightly better fit than Matching but worse than the S learner and T learner.

Score **Model**	**Mean RMSE**	**ATE**	**ATE - CI**	**ATT**	**ATT - CI**
Matching	1.74	-0.08	[-0.237, 0.069]	-0.06	[-0.253, 0.141]
IPW	1.42	-0.08	[-0.433, 0.263]	-0.11	[-0.418, 0.169]

In the next analysis, we used Random Forest to examine the potential impact of various possibly post-treatment features, on treatment effects. The results indicate that Volunteering, Music, Extracurricular activities and Weekly Study Time have minimal treatment effects, with ATE and ATT values close to zero and confidence intervals near zero, suggesting no significant impact of these features. For **Tutoring**, however, we observed a slightly larger negative effect, with an ATE and ATT of **-0.09** that slightly favor negative values, hinting at a minor adverse impact. **Absences** stood out with a lowest Mean Test RMSE of **0.79**, suggesting better fit and more precise model predictions for GradeClass. The ATE and ATT estimates for Absences are **-0.05**, with confidence intervals also slightly leaning negative. In summary, while most possibly post-treatment features do not show a strong treatment effect, Tutoring and Absences might have a slight negative impact. Overall, the Random Forest model performed consistently across features, with **Absences** providing the best model fit based on the lowest Test RMSE.

Score / features	Mean Training RMSE	Mean Test RMSE	ATE	ATE - CI	ATT	ATT - CI
Volunteering	1.18	1.21	-0.03	[-0.155, 0.033]	-0.03	[-0.156, 0.033]
Music	1.17	1.21	0.02	[-0.167, 0.041]	0.03	[-0.166, 0.04]
Extracurricular	1.17	1.20	0.00	[-0.164, 0.034]	0.00	[-0.161, 0.035]
Tutoring	1.17	1.20	-0.09	[-0.142, 0.035]	-0.09	[-0.14, -0.036]
Absences	0.76	0.79	-0.05	[-0.134, -0.019]	-0.05	[-0.137, - 0.02]
StudyTimeWeekly	1.14	1.19	-0.02	[-0.138, 0.027]	-0.01	[-0.139, 0.027]

Note: we didn't saw a reason to try double robust estimator because we tried a lot of methods (including the ones that are used in the double robust estimator) and we got a very similar results

possible weaknesses

1. **Missing Variables**: The dataset lacks important lifestyle factors like sleep quality and nutrition, which are likely correlated with both sports participation and GPA. Without these variables, there may be unaccounted influences on academic performance.
2. **Post-Treatment Bias**: Some variables (e.g., Volunteering, Music, Extracurricular, Tutoring, Absences, and StudyTimeWeekly) are potentially influenced by sports participation and thus could introduce post-treatment bias. This could confound the causal relationship between sports and GPA.
3. **SUTVA Violation**: The Stable Unit Treatment Value Assumption (SUTVA) does not fully hold because the treatment (sports) is only recorded as a binary variable, without accounting for differences in sports intensity or type. This introduces hidden variations in treatment, which could lead to biased estimates.
4. **Reliance on Propensity Score Methods**: The project relies heavily on propensity score matching and weighting, which might not fully eliminate bias from unobserved confounders.

These weaknesses suggest that while the project provides insights into the relationship between sports participation and academic performance, the results should be interpreted cautiously, considering the limitations in data and methodological assumptions.

Discussion

1. **Conclusions**:

   - The results indicate that sports participation has a negligible impact on GPA, as shown by the low ATE and ATT values across multiple methods, with confidence intervals often including zero. This suggests that, for this dataset, participating in sports does not have a strong, consistent effect on academic performance. This finding may imply that other factors, potentially unmeasured, play a more significant role in influencing GPA.

2. **Future Research:**

   Future studies could address the limitations identified in this work by including more detailed variables. Another direction could involve testing different model architectures. Additionally, continuous data that tracks changes over time would enable a deeper exploration of how sports participation may impact academic trajectories rather than single-point GPA measures.