Whether Maternal and Paternal Involvement Have Different Effects on Student's Academic Performance

Boyuan Cui

December 22, 2020

Abstract

This report conducts a propensity score matching (PSM) analysis to explore whether mother's involvement and father's involvement have different effects on the student's academic performance. It makes use of an observational dataset offered by the University of Jordan. Two models are constructed in this paper: a logistic model that estimates the propensity score with gender, grade, and place of birth as the baseline characteristics and a linear model that takes both the student's engagement and parent's involvements into account to assess the total marks of the student. Together with these models and the PSM technique, the paper has discovered that the mother's involvement is more likely to help students obtain better academic achievements than the father's involvement, with a p-value of 0.0000828

Github link: https://github.com/cuiboyuan/Parental-Involvement-and-Academic-Performance

Keywords: Parental Involvement, Academic Performance, Observational Study, Propensity Score Matching

Introduction

Parental involvement has always played a critical role in student's education, and schools have been encouraged in many ways to increase parental involvements [**5**]. Extensive researches have been conducted to investigate the effect of parental behaviors on student's school performance, but there has been only a handful of papers interested in the difference between the effect of maternal and paternal involvements on the student's academic achievements. One of the important reasons is that the effect of parental involvement is usually subtle, and it is hard to measure without a long period of time. Randomized experiments that take years are extremely expensive to organize. It is also unethical to interfere with the development of children by asking one of the parents to not involve in their education. Often, only observational data are available for this kind of research. Fortunately, there exist methods that can build causal links through observational studies, and propensity score matching (PSM) is one of those methods. In this report, the author will take advantage of PSM to explore whether there exists a difference between the effect of mother's involvements and father's involvement on the academic performance of the student.

This report will use a dataset on student's academic performance collected on the Kaggle website (www.kaggle.com). Details on how the dataset and PSM are used to explore the causal link between mother and father's involvement on student's total mark will be elaborated in the Methodology section. The result of the PSM analysis will be described in the Result section, and the causal inferences along with the conclusion will be presented in the Discussion section.

Methodology

Data

The dataset is posted on the Kaggle website by professor Ibrahim Aljarah at the University of Jordan (https://www.kaggle.com/aljarah/xAPI-Edu-Data). It consists of 480 students and 16 features including demographic information (e.g. gender, nationality), academic backgrounds (e.g. grade level, total marks), and behavioral features (e.g. raised hands) **[1,2]**.

There are 305 male and 175 females students in the dataset coming from a variety of origins; there are 179 students from Kuwait, 172 from Jordan, 28 from Palestine, 22 from Iraq, 17 from Lebanon, 12 from Tunis, 11 from Saudi Arabia, 9 from Egypt, 7 from Syria, 6 from the USA, 6 from Iran, 6 from Lybia, 4 from Morocco, and 1 from Venezuela. The data is collected by a learning management system called Kalboard 360 through a learner activity tracker tool called experience API (xAPI) which monitors the learning progress and action of students **[1,2]**.

The analysis in this report will only include 9 features from the dataset. Table 1 is a preview of the selected data.

Table 1: Glimpse of Original Data

gender	PlaceofBirth	GradeID	Relation	raisedhands	VisITedResources	AnnouncementsView	Discussion	Class
M	KuwaIT	G-04	Father	15	16	2	20	M
M	KuwaIT	G-04	Father	20	20	3	25	M
M	KuwaIT	G-04	Father	10	7	0	30	L
M	KuwaIT	G-04	Father	30	25	5	35	L
M	KuwaIT	G-04	Father	40	50	12	50	M
F	KuwaIT	G-04	Father	42	30	13	70	M

gender is a binary variable that represents the student's gender; the value is either M (male) or F (female).

PlaceofBirth is the student's place of birth. All possible values are Kuwait, Lebanon, Egypt, SaudiArabia, USA, Jordan, Venezuela, Iran, Tunis, Morocco, Syria, Palestine, Iraq, and Lybia.

GradeID represents which grade the student belongs to. The variable ranges from G-01 to G-12, which indicates grade 1 to grade 12, respectively.

Relation indicates which parent is responsible for the student. The value is either Mum or Father.

raisedhands is a numeric variable ranging from 0-100 that represents how many times the student has raised his/her hand in the classroom.

VisITedResources is a numeric variable ranging from 0-100 that represents how many times the student visits the course content.

AnnouncementsView is a numeric variable ranging from 0-100 that represents how many times the student checks the new announcements.

Discussion is a numeric variable ranging from 0-100 that represents how many times the student participates in discussion groups.

Class is a categorical variable that indicates which total mark interval the student belongs to. The dataset classifies students into three total mark numerical intervals: L (0-69), M (70-89), and H (90-100).

For the purpose of analysis, **GradeID** variable will be transformed to numeric variable **Grade** where integers 1-12 represent G-01 to G-12, respectively. **Class** variable will be transformed into a numeric variable by using 1 to represent L, 2 to represent M, and 3 to represent H. All other variables will be kept in the original format. Table 2 is a glimpse of the modified dataset.

The reason why these variables are selected will be explained in the Model subsection below.

Table 2: Glimpse of Modified Data

gender	PlaceofBirth	Grade	Relation	raisedhands	VisITedResources	AnnouncementsView	Discussion	Class
M	KuwaIT	4	Father	15	16	2	20	2
M	KuwaIT	4	Father	20	20	3	25	2
M	KuwaIT	4	Father	10	7	0	30	1
M	KuwaIT	4	Father	30	25	5	35	1
M	KuwaIT	4	Father	40	50	12	50	2
F	KuwaIT	4	Father	42	30	13	70	2

Model

Randomized controlled trials (RCT) have always been a gold standard in determining the causal relationship between the intervention and the outcome. RCT ensures that the control and the experiment group are similar by randomly assigning the treatment, so the effectiveness of treatment will not be influenced by confounding variables [**3**]. However, in the real world situation, organizing randomized experiments is often expensive, and sometimes impossible. The observational study is usually a more feasible solution, but it is not resistant to confounders. The observed population could be biased, so researchers cannot make causal inferences through observational data.

That is when the propensity score matching (PSM) becomes useful. A propensity score is defined as the probability of being assigned the treatment conditional on the observed baseline characteristics [**4**]. By this definition, with the same propensity score, the treated subjects will have similar baseline characteristics to the untreated subjects. In the observational study, the true propensity score is unknown and a regression model fit is needed to estimate the propensity score based on the observed baseline characteristic of each student [**3**].

For the topic of this report, the **Relation** variable, i.e. which parent is responsible for the student, will be considered as the treatment, so the propensity score will be the probability that the student's mother/father is responsible for him/her. Because **Relation** variable is binary, this paper will define the propensity score as the probability that a student's mother is responsible for him/her. Hence, untreated students are ones whose father is responsible for him/her, and treated students are ones whose mother is responsible.

After the propensity score regression model is fitted, the propensity score matching will be performed. Based on the propensity score estimation, treated and untreated students who share similar propensity scores will be matched into matching sets. Because they have similar propensity scores, matched students will have common characteristics. Thus, the impact of confounding is reduced when the outcome (the total mark) is assessed, which helps to build the causal link between the treatment and the outcome [**3**].

To summarize, the first step of PSM is to fit a propensity score regression model with the observed characteristics of students; the second step is to match students into treated-untreated pairs based on similar propensity scores; the final step is to compare the outcomes (**Class** variable) within the matched data to assess the effectiveness of the treatment (**Relation** variable).

The analysis will start off with the assumption of the null hypothesis (H_0) that there is no difference between the mother's involvement and the father's involvement in the student's academic performance. The final goal of this report is to determine whether there exists evidence that can reject H_0.

i. Propensity Score Regression Model

As mentioned previously, the probability that **Relation** equals 'Mum' is defined as the propensity score. To fit this model, one would need to find variables that will explain why a student is more likely to have his/her mother as the responsible parents, i.e. the baseline characteristics.

In some cultures, the mother might often be expected to take care of young kids, so the place of birth is a reasonable variable to consider. Similarly, in some societies, one might also expect elder male kids to be

more independent of their mother's caring, so age and gender are important as well. The dataset does not contain the age information, but a student's grade can be used as a similar measurement. In conclusion, **Grade**, **gender**, and **PlaceofBirth** will be used as the baseline characteristics.

Since **Relation** variable is binary, the analysis will use the logistic regression to estimate the propensity score. The model can be written as below:

$$logit(p) = \log \frac{p}{1-p} = \beta_0 + \beta_{female} \cdot \mathbb{I}\{female\} + \beta_{BirthPlace} \cdot BirthPlace + \beta_{Grade} \cdot Grade$$

or equivalently:

$$p = \frac{\exp(\beta_0 + \beta_{female} \cdot \mathbb{I}\{female\} + \beta_{BirthPlace} + \beta_{Grade} \cdot Grade)}{1 + \exp(\beta_0 + \beta_{female} \cdot \mathbb{I}\{female\} + \beta_{BirthPlace} + \beta_{Grade} \cdot Grade)}$$

The propensity score is represented as p. β_0 is the intercept term of the logit function $logit(p)$ when all other variables are 0. In the context of this analysis, β_0 has no practical meaning since **Grade** cannot be 0.

β_{female} is the additive coefficient that is added to $logit(p)$ if the student's gender is female; if the student is a male, this coefficient will be 0.

$\beta_{BirthPlace}$ is the additive coefficient that is added to $logit(p)$ if the student is born in $BirthPlace$. Because **PlaceofBirth** is categorical, to fit the model, one country needs to be selected as the reference. In other words, with all other variables the same, $\beta_{BirthPlace}$ can also be interpreted as the difference in $logit(p)$ between the student born in $BirthPlace$ and the student born in the reference country. For this analysis, Kuwait will be selected as the reference country, so $\beta_{Kuwait} = 0$.

β_{Grade} is a slope coefficient that represents the amount that is added to $logit(p)$ when the student's grade (**Grade**) increases by 1.

As an instance, for a grade 6 male student born in Kuwait, the propensity score estimation for him will be:

$$p = \frac{\exp(\beta_0 + 6 \cdot \beta_{Grade})}{1 + \exp(\beta_0 + 6 \cdot \beta_{Grade})}$$

and a grade 12 female student born in Egypt will have a propensity score of:

$$p = \frac{\exp(\beta_0 + \beta_{Egypt} + \beta_{female} + 12 \cdot \beta_{Grade})}{1 + \exp(\beta_0 + \beta_{Egypt} + \beta_{female} + 12 \cdot \beta_{Grade})}$$

The estimates of the fitted model coefficients are shown in the Result section

ii. Propensity Score Matching

Now with the propensity score model fitted, a one-to-one greedy propensity score matching without replacements will be performed. In other words, a treated student (called student A) in the dataset will be first randomly selected, and an untreated student with the propensity score that is closest to student A's propensity score will be matched with student A. Once the match is formed, these two students will no longer be available for the rest of the matching. If there are more than one closest untreated students, a random one will be selected. The process is repeated until a treated-untreated pair cannot be found **[3]**.

Table 3 is a glimpse of the matched pairs. If two students have the same value in the **matched** column, they are matched in a pair. It means that they will have similar baseline characteristics (i.e. similar propensity score) but one is treated (i.e. his/her mother is responsible for him/her) and the other is not.

Table 3: Glimpse of Matched Pairs

gender	PlaceofBirth	Grade	Relation	propensity	matched
F	KuwaIT	8	Mum	0.3056796	1
F	KuwaIT	8	Father	0.3056796	1
F	KuwaIT	7	Father	0.3053464	2
F	KuwaIT	7	Mum	0.3053464	2
M	KuwaIT	7	Father	0.1722119	3
M	KuwaIT	7	Mum	0.1722119	3

iii. Outcome Comparison

Afterward, cases that are not in any matching pairs will be removed from the dataset. Hence, the remaining data can be seen as the RCT experiment participants. Students whose mother is responsible for him/her are in the treatment group, while others belong to the control group.

In this last step, a model with the outcome as the response variable will be fitted to evaluate whether the treatment is effective. In this paper, the model will include some reasonable explanatory variables along with the treatment variable (**Relation**) to explain the variation in the outcome variable (**Class**). With this model, one can obtain useful statistics to assess the effect of treatment on the outcome.

Although the primary interest of this report is the parental involvement in school performance, the importance of the student's individual engagement should not be neglected. This aspect can be reflected by the student's behavior in the classroom, such as checking resources and announcements frequently, or active participation in group discussions. Therefore, in addition to the **Relation** variable, the model will include the number of raising hands (**raisedhands**), number of visiting class resources (**VisITedResources**), number of viewing announcement (**AnnouncementsView**), and number of participation in discussion groups (**Discussion**) as the explanatory variables.

Since **Class** has more than two possible values, a logistic model would not be appropriate. A linear model will be used in this part of the analysis.

The linear model can also be expressed mathematically as:

$$Class = \beta_0 + \beta_{RaiseHand} \cdot RaiseHand + \beta_{VisitResource} \cdot VisitResource +$$
$$\beta_{ViewAnnouncement} \cdot ViewAnnouncement + \beta_{Discussion} \cdot Discussion + \beta_{mother} \cdot \mathbb{I}\{mother\}$$

β_0 represents the intercept term of the model that represents the predicted **Class** variable value when all other variables are 0.

$\beta_{RaiseHand}, \beta_{VisitResource}, \beta_{ViewAnnouncement}, \beta_{Discussion}$ are the slope coefficient, which represent the predicted amount of increase in total mark level (**Class**) with one unit increment in hand raise, class resource visit, announcement view, and discussion participation, respectively.

β_{mother} indicates that, with all other variables being the same, the increase or decrease in mark level if the student's mother is responsible for him/her. The term $\beta_{mother} \cdot \mathbb{I}\{mother\}$ is 0 if the student's father is responsible for him/her.

As an instance, if a student's mother is responsible, and he/she raised his/her hand 10 times, visited class resource 5 times, view announcement 12 times, and participate in discussion groups 2 times, this model will predict his/her mark level to be:

$$Class = \beta_0 + 10 \cdot \beta_{RaiseHand} + 5 \cdot \beta_{VisitResource} + 12 \cdot \beta_{ViewAnnouncement} + 2 \cdot \beta_{AttendDiscussion} + \beta_{mother}$$

The primary interest of this report is the term β_{mother}. If this coefficient is statistically significant, the null hypothesis that **Relation** has no effect on the student's total mark can be rejected. The estimates of model coefficients are shown in the Result section.

Result

Figure 1: Distribution of Student's Gender after Propensity Score Matching

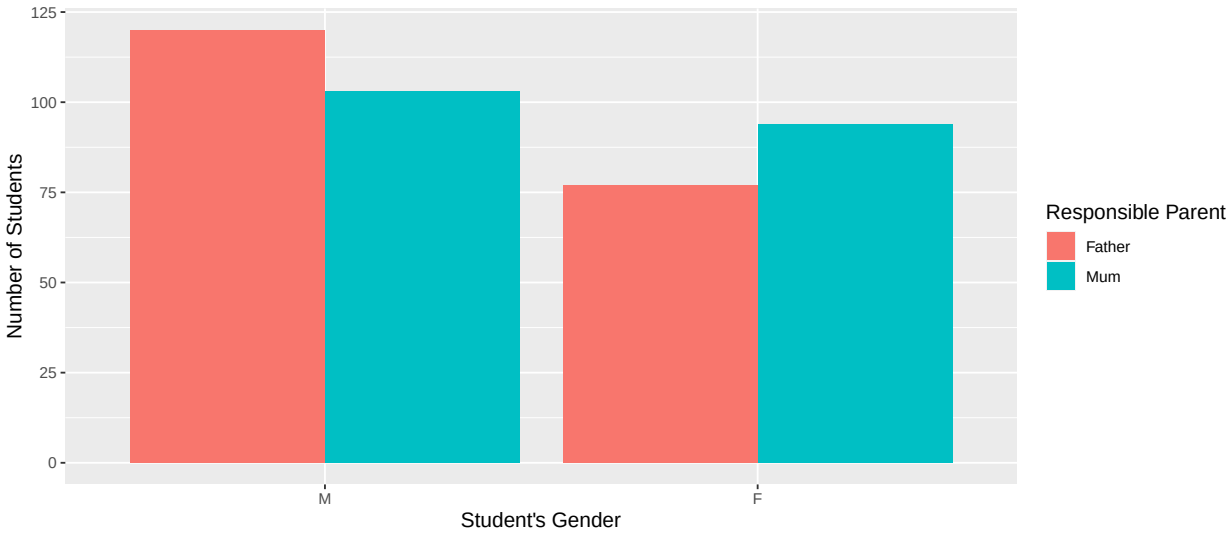

Figure 1 has shown the distribution of student's gender among the treatment and control groups after PSM. It seems that more male students are in the control group (**Relation** variable equals 'Father').

Figure 2: Distribution of Birth Place after Propensity Score Matching

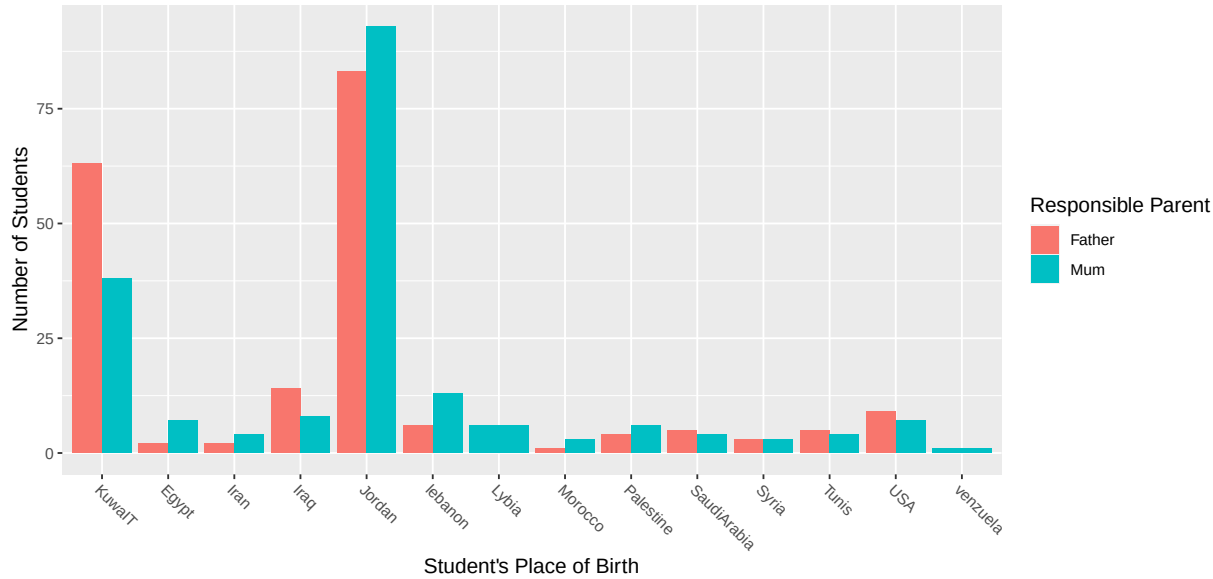

Figure 2 shows the distribution of **PlaceofBirth** after PSM. More students from Kuwait are in the control group, while more students from Jordan are in the treatment group. For other countries, the distribution is roughly similar by briefly inspecting the graph.

Figure 3: Distribution of Grade after Propensity Score Matching

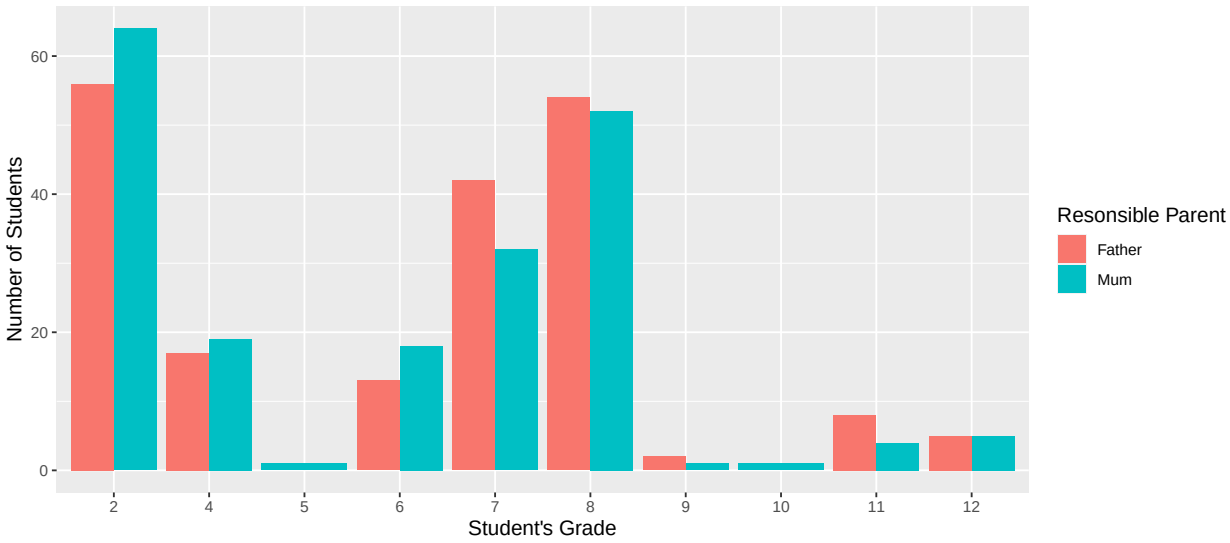

Figure 3 shows the distribution of **Grade** after PSM. By visually inspecting the graph, the distributions of the two groups are somewhat different.

Table 4: Logistic Model for Propensity Score Estimation

	Estimate	Standard Error	z value	p-value
beta_0	-1.5810247	0.2981479	-5.3028196	0.0000001
beta_Grade	0.0015704	0.0372449	0.0421649	0.9663672
beta_female	0.7480653	0.2132194	3.5084301	0.0004508
beta_Egypt	2.6111353	0.8311360	3.1416461	0.0016800
beta_Iran	2.1558197	0.8932975	2.4133278	0.0158076
beta_Iraq	0.7969138	0.4869891	1.6364098	0.1017538
beta_Jordan	1.3644337	0.2416325	5.6467312	0.0000000
beta_Lebanon	1.8120373	0.5370451	3.3740876	0.0007406
beta_Lybia	16.9095131	582.4348860	0.0290325	0.9768387
beta_Morocco	2.5084292	1.1796588	2.1264023	0.0334698
beta_Palestine	1.6931666	0.6809084	2.4866291	0.0128960
beta_SaudiArabia	0.3124146	0.6124069	0.5101422	0.6099518
beta_Syria	1.3272474	0.8512415	1.5591903	0.1189514
beta_Tunis	1.3474106	0.7011603	1.9216870	0.0546452
beta_USA	0.8854636	0.5574683	1.5883660	0.1122036
beta_Venezuela	17.1313887	1455.3975545	0.0117709	0.9906084

Table 4 has shown the estimate of $\beta_0, \beta_{Grade}, \beta_{female}$, and $\beta_{BirthPlace}$ for each place of birth. These are the coefficients of the logistic model that is used to estimate the propensity score in part i under the Model subsection. One noteworthy coefficient in this table is β_{female}. The estimate of β_{female} is positive and the p-value of β_{female} is 0.0004508, which is substantially smaller than 0.05. Both facts imply that a female student is more likely to have her mother as the responsible parent.

Figure 4: Responsible Parents and Student's Predicted Mark Levels

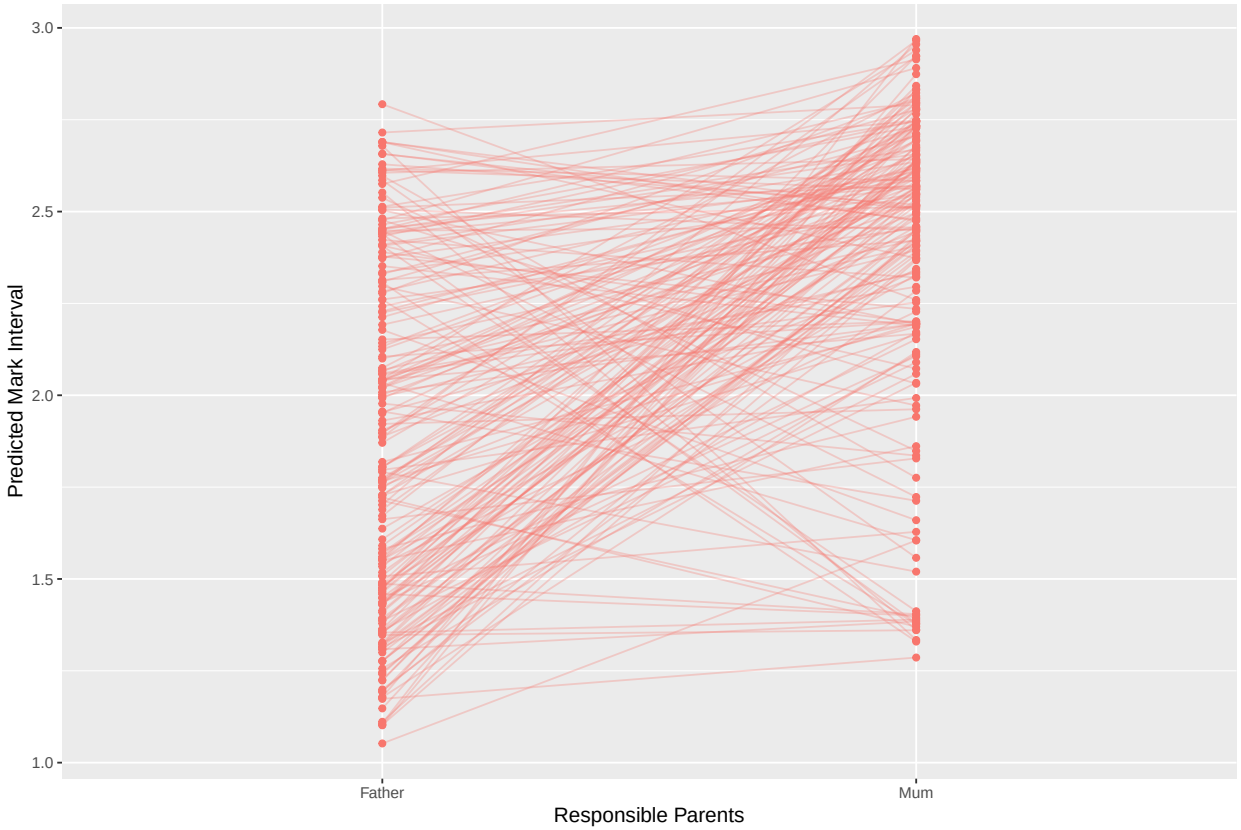

Figure 4 shows the different effects of father's and mother's involvement on the predicted marks of the student, produced by the fitted linear model in part iii under Model subsection. Two points are connected by the line if they are matched in a pair during the propensity score matching. By visually inspecting the graph, most students whose mother is the responsible parents seem to have higher predicted marks, which may imply that if a student's mother is responsible for him/her in school, he/she is predicted to have better academic performances.

Table 5: Linear Model for Outcome

	Estimate	Standard Error	t value	p-value
beta_0	1.0086139	0.0650046	15.5160270	0.0000000
beta_RaiseHand	0.0063217	0.0012262	5.1556356	0.0000004
beta_VisitResource	0.0001131	0.0013748	0.0822418	0.9344968
beta_ViewAnnouncement	0.0092407	0.0010885	8.4891196	0.0000000
beta_AttendDiscussion	0.0031026	0.0010399	2.9834263	0.0030306
beta_mother	0.2209342	0.0555337	3.9783831	0.0000828

Table 5 has shown the estimate of linear model coefficients $\beta_0, \beta_{RaiseHand}, \beta_{VisitResource}$, $\beta_{ViewAnnouncement}, \beta_{AttendDiscussion}, \beta_{mother}$ in part iii of the Model subsection.

By Table 5, the estimate of $\beta_{mother} \approx 0.221$, and it has a p-value of 0.0000828. This is calculated from the linear model that is fitted to predict the **Class** variable, which accounted for the number of raising hands (**raisedhands**), number of visiting class resources (**VisITedResources**), number of viewing announcement (**AnnouncementsView**), number of participation in discussion groups (**Discussion**), and which parent is responsible for the student (**Relation**).

Discussion

Summary

The analysis makes use of the PSM technique to investigate the effect of the mother's involvement on the student's academic performance. First, a logistic model is fitted to estimate the propensity score of each student (details of the fitted model are described in Table 4). Then these estimations are used to match students in pairs so that control and treatment groups are formed. Lastly, a linear model is constructed to assess the effect of the mother's involvement (details of the fitted model are described in Table 5).

Conclusion

Table 5 shows that a student whose mother is responsible for him/her is predicted to have a 0.22 higher total mark level (**Class** variable) than a student whose father is responsible. The p-value of β_{mother} is substantially smaller than 0.05, so the effect of the mother's involvement is statistically significant. Hence, there exists strong evidence that rejects the null hypothesis. In other words, the analysis has shown that the mother's involvement has a better effect on the student's academic performance than the father's involvement.

Weakness & Next Steps

One potential flaw of the analysis is that a linear model may not be appropriate to fit the **Class** variable. The mark interval represented by **Class** is discrete and each level consists of different ranges of marks, so a linear model may not be valid. In the next steps, I could use diagnostic plots to check whether the linear model is appropriate. If it is not, I could try to apply other models that are commonly used to fit discrete variables, such as a multinomial logistic regression, in further analysis.

Another drawback is that PSM does not give similar distributions among the two groups. As shown in Figure 1, 2, 3, the distribution of baseline characteristics is slightly different in treatment and control groups after PSM. This could imply that the data after matching does not approximate RCT well enough, which may lead to bias results. One possible reason for this weakness is that I use the one-to-one greedy propensity score matching in the analysis, which does not guarantee to yield optimal matches. Thus, in future improvements, I could try alternative PSM techniques, such as an optimal matching or match with replacements [**3**], to see which one gives a more balanced matched data.

Additionally, propensity score matching is not always an ideal approach to construct causal links. King and Nielsen have shown that propensity score matching might suffer from *PSM paradox*: as the propensity score matching reduces the imbalance of observational data, the inference becomes less useful [**6**]. One possible solution is, as suggested by King and Nielsen, to use other matching methods instead of PSM [**6**]. To further extend this analysis, I could evaluate whether the matching in this report is appropriate, and learn to apply other matching approaches to reduce the influence of *PSM paradox*.

References

[**1**] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). *Mining Educational Data to Predict Student's academic Performance using Ensemble Methods.* International Journal of Database Theory and Application, 9(8), 119-136.

[**2**] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). *Preprocessing and analyzing educational data set using X-API for improving student's performance.* In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

[**3**] Austin P. C. (2011). *An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies*. Multivariate behavioral research, 46(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

[**4**] Rosenbaum P.R., Rubin D.B. (1983). *The central role of the propensity score in observational studies for causal effects*. Biometrika. 1983a;70:41–55.

[**5**] Chen, G. (2020, October). *Parental Involvement is Key to Student Success*. Public School Review. https://www.publicschoolreview.com/blog/parental-involvement-is-key-to-student-success

[**6**] King, G., Nielsen R. (2019). *Why Propensity Scores Should Not Be Used for Matching*. Political Analysis.