Report 4: Add Health

Sally Cochrane

4/26/2021

1 Introduction

Does participation in high school sports increase educational attainment? With increasing pressure on school budgets, especially after the economic downturn caused by the pandemic (Turner, 2020), should schools cut their coaches and sports teams to save money, or is there an educational return on investment from sports? On the one hand, sports could teach discipline, teamwork, and improve overall mood and health, which could in turn improve academic performance. And indeed, some scholars find that participating in school sports has beneficial effects on academic performance and outcomes, from raising GPA (Fox et al., 2009) to increasing college attendance (Marsh, 1993). On the other hand, spending time on sports could cause students to spend less time studying, which could cancel out any beneficial effects of sports. Indeed, some scholars find little to no effect of sports participation on educational outcomes. They argue that any attainment differences between athletes and non-athletes are due to pre-existing individual differences such as in motivation and self-discipline (Daley and Ryan, 2000, Rees and Sabia, 2010). Using data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), we use covariate balancing propensity score (CBPS) matching and linear regression to evaluate how participation in at least one school sport in Wave 1 (1994-95) was associated with educational attainment at Wave 4 (2008-2009). We find evidence consistent with the theory that there is no association between sports participation and educational attainment.

2 Theory

Does participation in high school sports increase educational attainment later in life? On the one hand, participating in sports may build self-discipline, self-esteem, and improve teamwork; features that could help success in school. On the other hand, participating in sports takes away time from studying, and so may have no effect or a negative effect on academic performance. Some scholars have found evidence for the first hypothesis. Marsh (1993), studying high schoolers in 1980 to 1984, found that those who participated in sports during the last two years of high school experienced a range of positive effects: they had increased educational aspirations, reduced absenteeism, and increased subsequent college attendance. Fox et al. (2009) found positive associations between sports participation and higher GPA's in high school for both boys and girls, and girls (but not boys) who did any physical activity also had higher GPA's. Finally, Barron, Ewing, and Waddell (2000) found that high school athletic participation was associated with higher wages and higher educational attainment after high school. However, they remark that it is unclear whether this merely reflected pre-existing differences between individuals' ability or value of leisure time. Other scholars have found modest to negative associations between sports participation and academic outcomes. Rees and Sabia (2010), using data from the National Longitudinal Study of Adolescent Health, found that sports participation was associated with a modest increase in GPA, but that most of this association could be explained by pre-existing individual differences such as motivation, future-orientedness, and self-discipline. Daley and Ryan (2000) found a weakly *negative* correlation between participation in physical activity and English grades for children ages 13 - 16, and no correlation between physical activity and grades in Math and Science. Eide and Ronan (2001) found mixed effects, depending on sex and race: sports participation had a negative effect on educational attainment for white male students; a positive effect for black male students,

and a positive effect on white female students, with no effect on Hispanic males or black and Hispanic female students.

3 Data

Data come from the Add Health longitudinal study of adolescents in grades 7-12 in the United States during the 1994-95 school year (Wave 1), and from follow-up interviews in 2008-2009 when the respondents were 24 to 32 years old (Wave 4). The outcome variable of interest was the highest level of education that the respondent had achieved to date in Wave 4 (2008-2009), ranging from 1 (8th grade or less) to 13 (completed a doctoral degree), with 6 = some college and 7 = completed college (see Appendix for complete coding of education). The treatment variable of interest was whether the student participated in any organized sports at Wave 1. Respondents who reported participating in one or more sports were coded as participating in sports; if they did not mark any sport, they were coded as not participating. Control variables included income in thousands of dollars (reported by a parent in Wave 1); sex; race; self-reported overall health on a scale of 1 (poor) to 5 (excellent); the respondent's desire to attend college from 1 (low) to 5 (high); whether they report drinking alcohol, as some scholars have found that sports participation is associated with increased alcohol use, which may have an effect on academic performance (Mays et al, 2010); and the mean of the parents' educational attainment, from 0 (never went to school) to 8 (professional training beyond 4-year college/university), with 7 = college graduate. The mean was taken because there was no reason from theory to investigate each parents' educational level separately. Other controls included current grades at wave 1, calculated from self-reported English, History, Math, and Science grades, ranging 1 (D or below) to 4 (A). This measure was preferred to the transcripts of cumulative GPAs obtained in wave 3 because if sports do have an effect on educational attainment, they may do so by raising GPA subsequent to sports exposure, and we wanted a measure of academic performance that was as close as possible to a "baseline" as a control variable.

The data contained a significant amount of missingness. 25.8% (1318 entries) was missing for participation in a sport. These observations were removed because this was the treatment of interest, so it was not appropriate to impute the data. Once these missing entries were removed, the variables with the most missingness were income, parent education, and current grades, which were missing 21.9% (831 entries), 11.7% (443 entries), and 1.87% (71 entries) respectively. By substituting the adolescent's report of their resident parents' education levels when the parent's self-report was missing, this missingness was reduced to 0.975% (37 entries). This was deemed acceptable because the distribution of adolescents' reports of average resident parent education was similar to the distribution of parents' self-reported level of education (see Appendix for comparison plots). The rest of the missing data was imputed with Amelia. It was necessary to impute income because exploratory analysis showed that missingness in income was likely not random. Entries with missing income had lower rates of receiving public assistance and were more likely to be able to pay their bills than entries without missing income (see Appendix for analysis), which suggests that participants with missing income data may have had higher incomes than participants without missing income data. Income was logged before imputation, in order to make the distribution closer to normal for linear regression. Other variables with a small amount of missingness (less than 2%) were also imputed with Amelia, as this was unlikely to change the results. Several variables were included only for imputation as they were likely to help predict income and grades. For income, these were: parents' ability to pay all their bills, whether they received public assistance, and whether the parent and the parent's partner were employed full-time; and for grades, these were: whether the student had repeated a grade, been suspended, or had trouble completing homework. After imputation, there was no missing data (See Appendix for summary of imputation procedure).

The data was imbalanced between the sports and non-sports groups in several variables. Out of 3796 students, 2164, or 57%, participated in sports. More males participated in sports than females; health was better in the sports group; sports players had higher grades; and more sports players wanted to go to college. Race, alcohol use, log income, and parent education were roughly similar in treatment and control groups. (See appendix for plots). Four matching methods were examined, and CBPS matching was selected as providing the best balance of the data. **Table 1** shows summary statistics before and after matching. (See appendix

Table 1: Summary Statistics Before and After Matching

	Unmatched			Matched	
	No Sports (N = 1632)	Sports (N=2164)	Total (N=3796)	No Sports	Sports
Log Income	3.41 (0.86)	3.60 (0.89)	3.52 (0.88)	3.604 (0.812)	3.605 (0.887)
Parent Ed.	4.81 (1.79)	5.28 (1.75)	5.08 (1.78)	5.275 (1.734)	5.275 (1.747)
GPA	2.77 (0.75)	2.91 (0.73)	2.85 (0.74)	2.911 (0.733)	2.911 (0.730)
Desire College	4.35 (1.10)	4.61 (0.85)	4.50 (0.97)	4.608 (0.839)	4.609 (0.849)
Health	3.75 (0.92)	4.05 (0.85)	3.92 (0.89)	4.045 (0.855)	4.045 (0.846)
Sex					
Female	1031 (63.2%)	1047 (48.4%)	2078 (54.7%)	0 (48.4%)	0 (48.4%)
Male	601 (36.8%)	1117 (51.6%)	1718 (45.3%)	1 (51.6%)	1 (51.6%)
Race					
White	1004 (61.5%)	1390 (64.2%)	2394 (63.1%)	1 (64.2%)	1 (64.2%)
Black	415 (25.4%)	515 (23.8%)	930 (24.5%)	0 (23.8%)	0 (23.8%)
Hispanic	155 (9.5%)	177 (8.2%)	332 (8.7%)	0 (8.2%)	0 (8.2%)
Other	58 (3.6%)	82 (3.8%)	140 (3.7%)	0 (3.8%)	0 (3.8%)
Alcohol Use					
No	748 (45.8%)	991 (45.8%)	1739 (45.8%)	0 (45.8%)	0 (45.8%)
Yes	884 (54.2%)	1173 (54.2%)	2057 (54.2%)	1 (54.2%)	1 (54.2%)
Ed. Attainment	5.70 (2.41)	6.24 (2.43)	6.01 (2.43)	6.200 (2.462)	6.244 (2.431)

for plots of the balance of covariates with the four matching methods).

4 Methods

We ran multiple regression using CBPS weights. CBPS weights approximate random sampling by ensuring that treated and control groups will be close to balanced on observed pre-treatment covariates. Multiple regression was appropriate because we wanted to investigate whether participating in sports was associated with later educational attainment alone and while holding a variety of other variables constant. Some of the assumptions of linear regression may be violated, but the violations were not extreme enough to merit transforming any variable beyond income (which was logged) (see Appendix for diagnostic plots). The residuals vs. fitted values plot shows the linearity assumption is close enough to being met that we proceed as if it is met. A histogram of the residuals shows nearly normal distribution, with a slight skew right, but it is moderate enough that we proceed as if the normality assumption is met. The scale-location plot and histograms of the variables show that the equal variance assumption may be violated, but none of the variables merit transformation, so we proceed as if the equal variance assumption is met. While there are several outliers with influential values, investigation showed these were not due to coding errors so they were retained (all were respondents who achieved PhD's or were PhD candidates in wave 4, but whose parents had less education). Finally, the independence assumption may not be met because the survey included siblings, so there is a possibility that some observations influenced others. However, since the schools where the surveys were conducted were randomly drawn from the entire United States, and from various types of schools (public, private, charter), major clustering problems based on geography are unlikely to be a problem.

A preferred model was selected by adding control variables to the simple bivariate model one at a time and using ANOVA to assess the improvement in explanatory power balanced with parsimony using the F-statistic (see Appendix for ANOVA tables). Using this method, the preferred model is:

$$\widehat{Educational\ Attainment} = \alpha + \beta_1(\text{sports}_{1=\text{yes}}) + \beta_2(\text{gpa}) + \beta_3(\text{parents education}) + \beta_4(\text{college aspirations}) + \beta_5(\text{health}) + \beta_6(\text{log income}) + \beta_7(\text{male}) + \epsilon$$

Where the coefficients are:

$$\widehat{\text{Educational Attainment}} = -2.35 + 0.04(\text{sports}_{1=\text{yes}}) + 1.07(\text{gpa}) + 0.36(\text{parents education}) + 0.42(\text{college aspirations}) + 0.2(\text{health}) + 0.27(\text{log income}) - 0.34(\text{male})$$

5 Results

Table 2: Models for Association between Educational Attainment and Sports Participation

	Dependent variable:		
	Educational Attainment		
	(1)	(2)	(3)
Sports: Yes	0.043 (0.079)	0.042 (0.064)	0.042 (0.064)
Grades (wave 1)		1.067*** (0.048)	1.072*** (0.048)
Parent Education		0.362*** (0.021)	0.369*** (0.021)
Race: Black			0.042 (0.080)
Race: Hispanic			0.267** (0.123)
Race: Other			−0.038 (0.171)
Alcohol: yes			−0.005 (0.066)
College Aspirations		0.418*** (0.040)	0.413*** (0.040)
Health		0.198*** (0.039)	0.196*** (0.039)
Income (log)		0.273*** (0.042)	0.281*** (0.043)
Sex: Male		−0.342*** (0.066)	−0.340*** (0.066)
Constant	6.200*** (0.056)	−2.349*** (0.260)	−2.426*** (0.274)
Observations	3,796	3,796	3,796
R^2	0.0001	0.344	0.345
Adjusted R^2	−0.0002	0.343	0.343
Note:			*p<0.1; **p<0.05; ***p<0.01

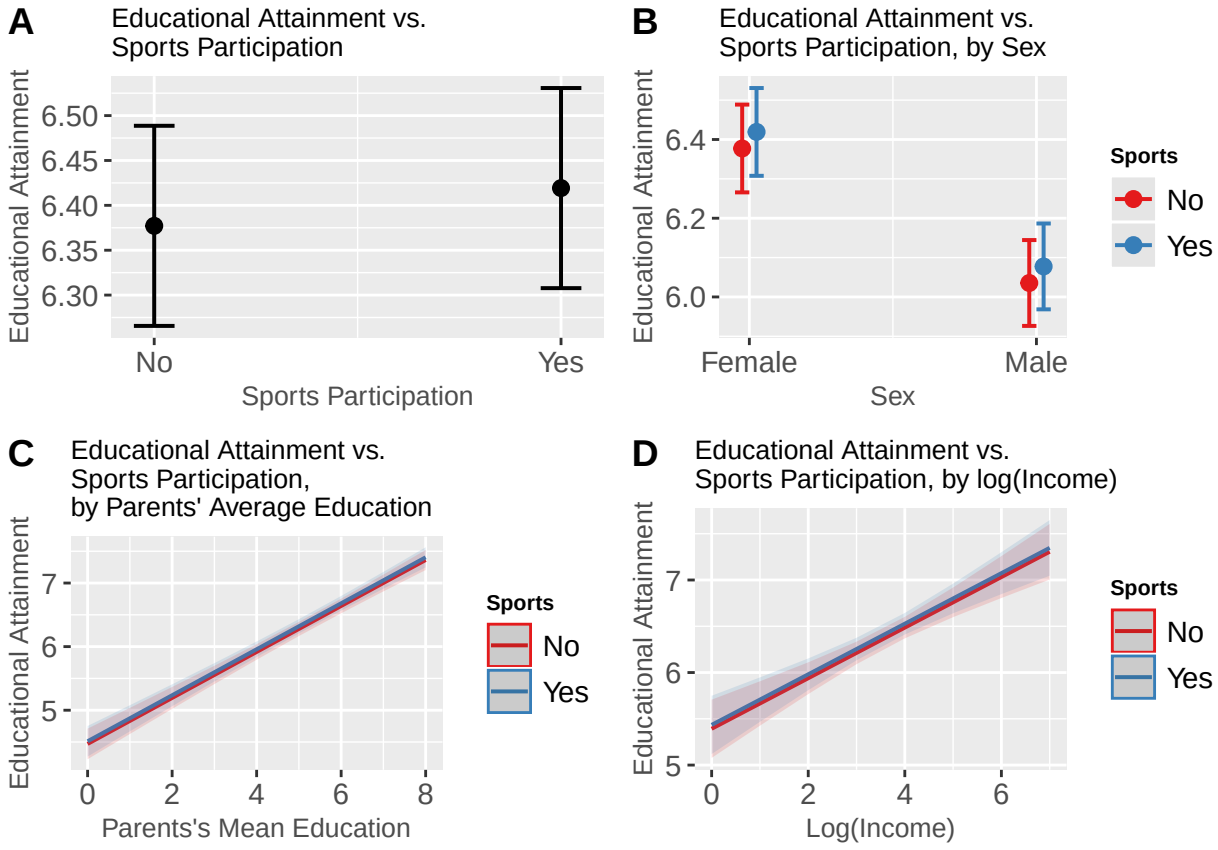
Using linear regression, we find no evidence for an association between participation in sports and educational attainment. As **Table 2** shows, this finding is robust across all three models presented here. It is also robust with all four matching methods presented above (see Appendix, Table 7). When sports participation is the only explanatory variable (model 1), it is associated with 0.043 unit increase in educational attainment as compared to not playing sports. This estimate is near zero, and based on how educational attainment was coded, this represents less than the difference between some schooling at any level and graduation at that level, which would be a one-unit increase. The p-value for this estimate is greater than 0.1, indicating that we cannot reject the null hypothesis that sports participation has no association with educational attainment.

In the preferred model (model 2), when recent grades, parents' mean educational attainment, the respondent's college aspirations, health, income, and sex are held constant, there is also no association between sports participation and educational attainment. In this model, playing sports is associated with an estimated increase in educational attainment of 0.042 units when holding these other variables constant, which, like model 1, is near zero. This estimate has a p-value greater than 0.1, so we cannot reject the null hypothesis that there is no association between sports participation and educational attainment, holding these other variables constant. The full model (model 3), which adds controls for race and alcohol use, has the same estimate with the same p-value for the association between sports and educational attainment as model 2. However, in the preferred model (model 2) and the full model (model 3), many other variables *are* statistically significantly associated with educational attainment, when holding sports participation constant. In model 2, holding sports and the other variables constant, each unit increase in recent grades is associated with a 1.067 increase in educational attainment (p-value less than 0.01); each unit increase in parents' mean education level is associated with a 0.362 unit increase in educational attainment (p-value less than 0.01); each unit increase in desire to go to college is associated with a 0.418 unit increase in educational attainment (p-value less than 0.01); each unit increase in reported health is associated with a 0.198 unit increase in educational attainment (p-value less than 0.01); a doubling of income is associated with a 0.189 ($0.273 * log(2)$) unit increase in educational attainment (p-value less than 0.01); and being male is associated with a 0.342 unit

decrease in educational attainment as compared to being female (p-value less than 0.01). Note that recent grades have the biggest estimated effect on educational attainment, and that all the above control variables are estimated to increase educational attainment except for male sex, holding sports participation constant. In the full model, the estimated associations between these control variables and educational attainment are similar; alcohol use is not statistically significantly associated with educational attainment holding sports and other factors constant; and only Hispanic race is statistically significantly associated with educational attainment, with an estimated 0.267 unit increase in education as compared to the White group (p-value less than 0.05), holding other variables constant.

Taking these results together, the substantive interpretation is that, both in magnitude of the estimated association and in the statistical significance of the estimate, there is no evidence for an association between sports participation and educational attainment; and other factors such as recent grades, parents' education levels, college aspirations, health, income, and sex explain educational attainment better than sports participation.

Plot A shows the association between sports participation and educational attainment. The bars for sports and non-sports overlap substantially, indicating little association. Plot B shows educational attainment by sports participation and sex, showing how males have lower estimated educational attainment than females. Plot C and D show how variables other than sex (here, parents' education and income) are positively associated with educational attainment, though sports participation has little association with education, as the lines for sports and non-sports overlap substantially.



6 Discussion

There are several limitations to this study. First, self-reported answers may not reflect the true answers to the questions. Students may report greater desire to go to college than they feel to please the interviewer, for example, and self-reported participation in sports is less accurate than a team roster would be (which

was not available for this study). Second, missing data, especially for sports participation and income, may have biased the results in unknown ways. Third, estimated average grades at time of interview is not exactly a "baseline" for the child's academic performance before the "treatment" of sports, as they may have been participating in sports before reporting their grades. Future studies could use transcripts to get accurate GPA's before and after students signed up for organized sports. Fourth, this report only looked at the effect of any participation in sports during Wave 1 on educational attainment (the question was not repeated in later waves), which misses subjects who signed up for sports later in high school. Furthermore, it is likely that *more* sports participation would have a different effect from *less* participation, and the simple binary coding here may fail to detect this effect. Future research could look at whether *more hours* spent playing sports and/or *more years* spent involved in sports during high school is associated with educational attainment. Fifth, it may be that colleges preferentially admit students who play sports, so that participation is not directly effecting academic performance, but rather effecting admissions. One way to address this issue would be to study academic performance and higher educational achievement only among those admitted to college.

Finally, we cannot say that the associations (or lack there of) between the explanatory variables and educational attainment were causal. Though the matching procedure mimics random sampling by creating treatment and control groups that are balanced on the observed variables, and therefore gets us closer to being able to infer causality, it is still possible that the two groups were unbalanced on unobserved variables which could act as confounders, predisposing subjects to both their educational attainment level and their status in any of the observed variables in this study. Since the subjects were drawn from a representative sample of middle and high schools around the United States, the results are generalizable to U.S. adolescents between 7th and 12th grades in 1994-5, who were 25-32 in 2008-9.

7 Conclusion

We find no evidence for an association between participating in school sports in Wave 1 with educational attainment in Wave 4. The estimated association between sports participation and educational attainment was near zero and the estimate was not statistically significant, whether sports participation was the only explanatory variable or if other factors were held constant. In the preferred model, however, when sports participation was held constant, recent grades, income, parents' (mean) educational level, the student's college aspirations, health, and sex *were* statistically significantly associated with educational attainment at Wave 4. Higher grades, income, parental education, college aspirations, and health rating were all associated with higher educational attainment, while male sex was associated with lower estimated educational attainment compared to female sex. These results held for the full model as well, where the added variable of alcohol use was not found to have a statistically significant association with educational attainment, and only Hispanic race was found to be statistically significantly associated with increased educational attainment compared to White race. The finding of no association between sports and educational attainment held for models run with all four matching methods. These results undermine the hypothesis that sports participation improves educational attainment and support the hypothesis that sports participation has no effect on educational attainment. While the data were matched to obtain treatment (sports) and control (non-sports) groups that were similar on all the observed variables, which mimics random sampling and brings us closer to inferring causality (or lack thereof) of sports participation on educational attainment, we still cannot say definitively that the relationship was causal, as there may be unobserved characteristics that were not evenly distributed between the treatment and control group which may have driven the results.

8 Bibliography

Data:

- Harris, Kathleen Mullan, and Udry, J. Richard. National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008 [Public Use]. Carolina Population Center, University of North Carolina-Chapel Hill [distributor], Inter-university Consortium for Political and Social Research [distributor], 2018-08-06. https://doi.org/10.3886/ICPSR21600.v21

Software:

- Dahl, David B., David Scott, Charles Roosen, Arni Magnusson and Jonathan Swinton. (2019). xtable: Export Tables to LaTeX or HTML. R package version 1.8-4. https://CRAN.R-project.org/package=xtable
- Fong, Christian, Marc Ratkovic and Kosuke Imai. (2021). CBPS: Covariate Balancing Propensity Score. R package version 0.22. https://CRAN.R-project.org/package=CBPS
- Greifer, Noah. (2020). cobalt: Covariate Balance Tables and Plots. R package version 4.2.4. https://CRAN.R-project.org/package=cobalt
- Hlavac, Marek. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
- Honaker, J., King, G., Blackwell, M. (2011). "Amelia II: A Program for Missing Data." Journal of Statistical Software, 45(7), 1–47. http://www.jstatsoft.org/v45/i07/.
- Lüdecke, D (2021). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.7, <URL: https://CRAN.R-project.org/package=sjPlot>.

Scholarly References:

- Barron, John M., Bradley T. Ewing, and Glen R. Waddell. 2000. "The Effects of High School Athletic Participation on Education and Labor Market Outcomes." *The Review of Economics and Statistics* 82, no. 3:409-421.
- Daley, Amanda J., and Jason Ryan. 2000. "Academic Performance and Participation in Physical Activity by Secondary School Adolescents." *Perceptual and Motor Skills*, 91: 531-534.
- Eide, Eric R., and Nick Ronan. 2001. "Is participation in high school athletics an investmnet or a consumption good?: Evidence from high school and beyond." *Economics of Education Review*, 20, no.5: 431-442.
- Fox, Claudia K., Daheia Barr-Anderson, Dianne Neumark-Sztainer, and Melanie Wall. 2009. "Physical Activity and Sports Team Participation: Associations with Academic Outcomes in Middle School and High School Students." *Journal of School Health* 80, no. 1:31-37.
- Marsh, Herbert W. 1993. "The Effects of Participation in Sport During the Last Two Years of High School." *Sociology of Sport Journal* 10, no. 1:13-43.
- Mays, Darren, Lara DePadilla, Nancy J. Thompson, Howard I. Kushner, and Michael Windle. 2010. "Sports Participation and Problem Alcohol Use: A Multi-Wave National Sample of Adolescents." *Americna Journal of Preventive Medicine* 38, no. 5: 491-498.
- Rees, Daniel I., and Joseph J. Sabia. 2010. "Sports participation and academic performance: Evidence from the National Longitudinal Study of Adolescent Health." *Economics of Education Review* 29, no. 5:751-759.
- Turner, Cory. 2020. "A Looming Financial Meltdown for America's Schools," *National Public Radio*, May 26.

Appendix: Education Variables

label	value
8th grade or less	1
Some high school	2
High school graduate	3
Some vocational/techincal training (after high school)	4
Completed vocational/technical training (after high school)	5
Some college	6
Completed college	7
Some post baccalaureate professional education	8
Completed post baccalaureate professional education	9
Some graduate School	10
Completed a master's degree	11
Some graduate training beyond a master's degree	12
Completed a doctoral degree	13

Table 3: Respondent's Educational Attainment at Wave 4

label	value
Never went to school	0
8th grade or less	1
>8th grade/didn't graduate high school	2
Business/trade/voc. school instead high school	3
High school graduate or completed GED	4
Business/trade/voc. school after high school	5
College/didn't graduate	6
Graduated from college/university	7
Prof training beyond 4-year college/univ	8

Table 4: Parent's Educational Attainment

Appendix: Missingness and Imputation.

1. Missingness before any imputation:

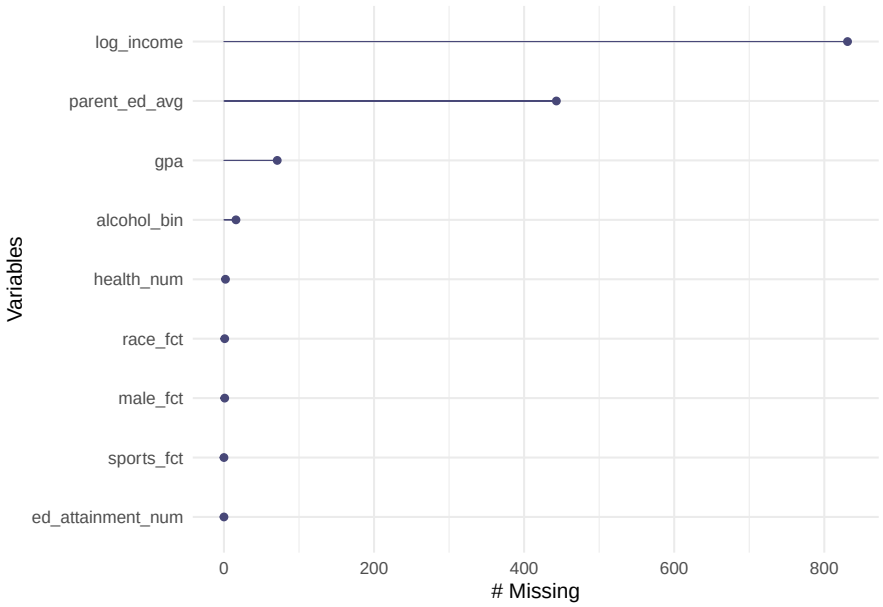


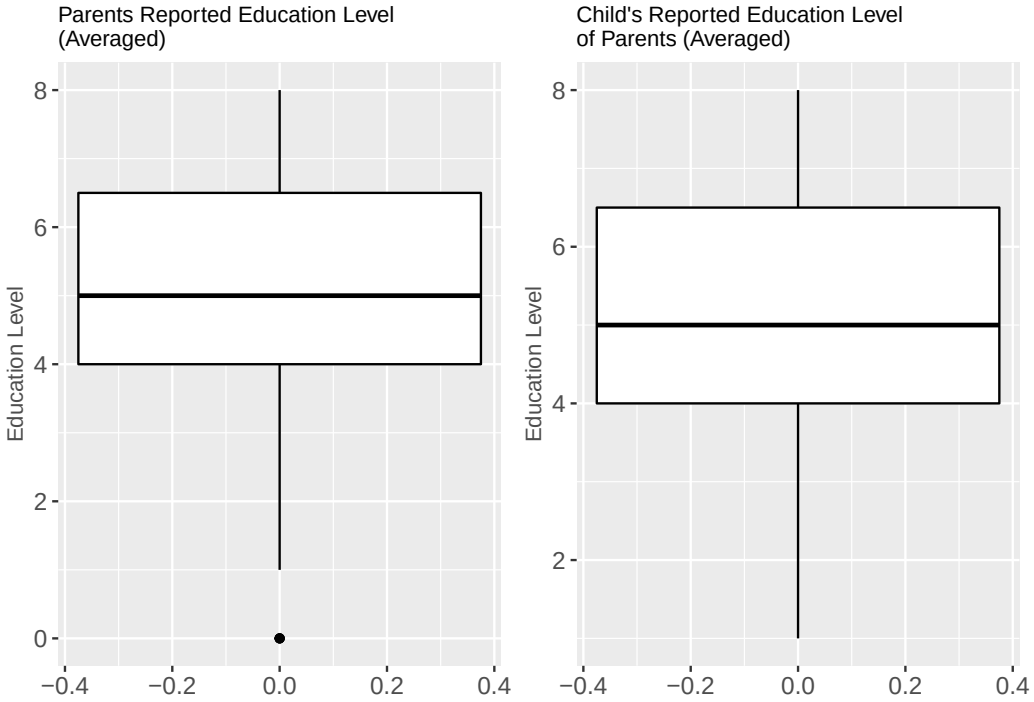
Figure 1: Missingness before imputation.

	variable	n_miss	pct_miss
1	log_income	831	21.89
2	parent_ed_avg	443	11.67
3	gpa	71	1.87
4	alcohol_bin	16	0.42
5	health_num	2	0.05
6	male_fct	1	0.03
7	race_fct	1	0.03
8	sports_fct	0	0.00
9	ed_attainment_num	0	0.00

Table 5: Missingness before imputation

2. Imputing parents' average education:

Childrens' reports of their resident parents' (averaged) education were similar to the parents' self-reports, so the childrens' reported level of parents' education was used to impute missing data in the parents' answers. This method left only 37 entries, or 0.975%, with missing data to be imputed with amelia.



3. Missing income distribution:

After removing missing sports participation data and imputing missing parental education data with the adolescents' responses, the variable with the most missing data was income, with 831 entries, or 21.9% missing; followed by gpa, with 71, or 1.87% missing.

Missing income was not random, as entries with missing income had lower rates of receiving public assistance and were more likely to be able to pay their bills than entries without missing income. Among those with missing income data, 3.9% couldn't afford to pay their bills, while among those without missing income data, 14.0% couldn't afford to pay their bills. Among those with missing income, 3.9% were on public assistance, while among those without missing income, 6.3% were on public assistance. These numbers suggested that missing income may not be random, and so it was imputed.

4. Imputation with amelia:

Income and grades were imputed using the amelia package, and by including several extra variables that were not included as part of the analysis, but that likely predict income and grades. For income, these were: whether the parent responded that they could pay all their bills, whether they received public assistance, and whether the parent and the parent's partner were employed full-time; and for grades, these were: whether the student reported repeating a grade, being suspended, and having trouble completing homework. After imputation, there was no missing data.

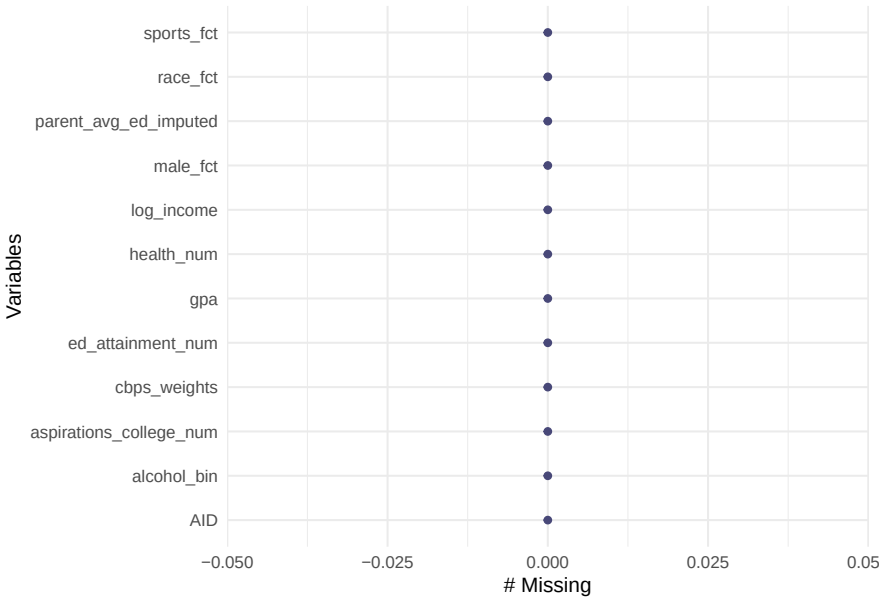


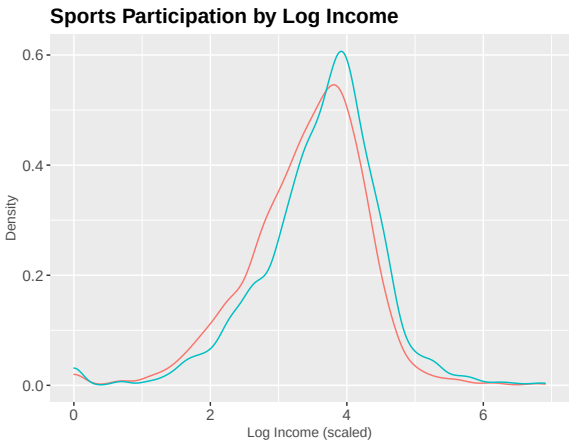
Figure 2: Missingness after imputation

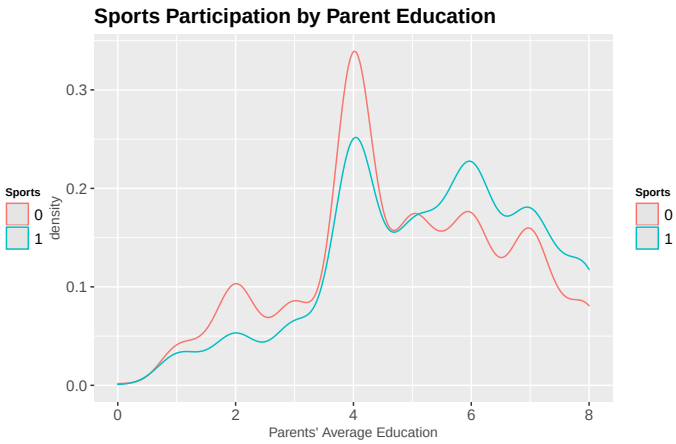
	variable	n_miss	pct_miss
1	AID	0	0.00
2	sports_fct	0	0.00
3	ed_attainment_num	0	0.00
4	parent_avg_ed_imputed	0	0.00
5	log_income	0	0.00
6	male_fct	0	0.00
7	race_fct	0	0.00
8	health_num	0	0.00
9	aspirations_college_num	0	0.00
10	alcohol_bin	0	0.00
11	gpa	0	0.00
12	cbps_weights	0	0.00

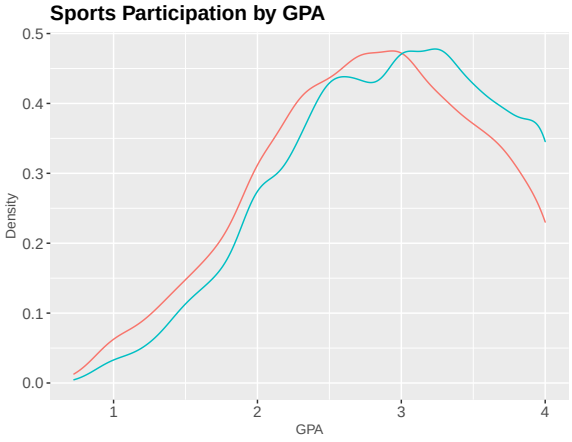
Table 6: Missingness after Imputation with Amelia

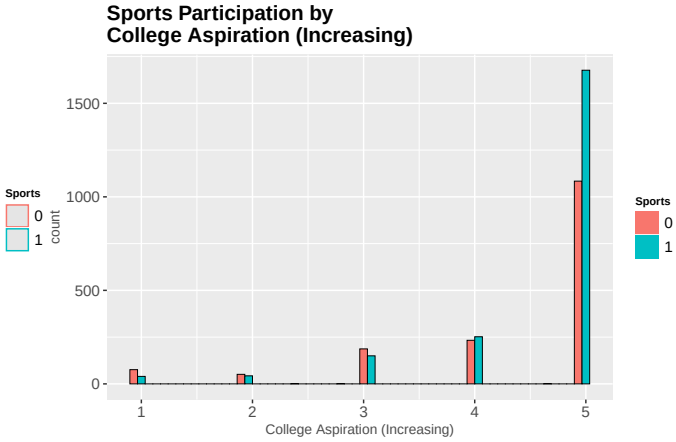
Appendix: Matching Data

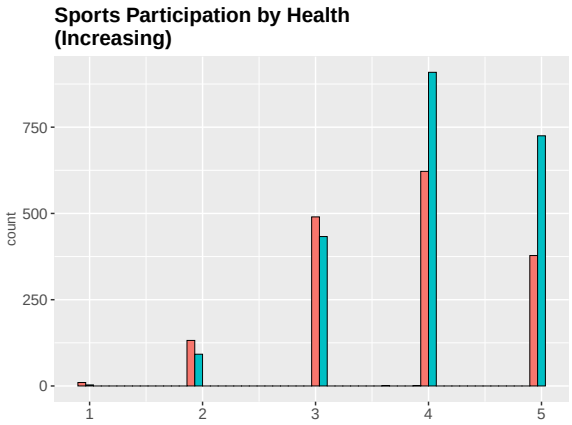
Several variables did not exhibit overlap in the treatment and control group.

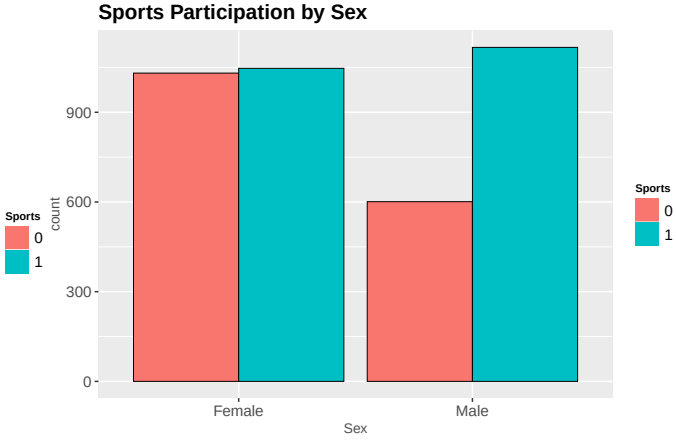

Sports Participation by Log Income
Density
Log Income (scaled)
Sports
0
1

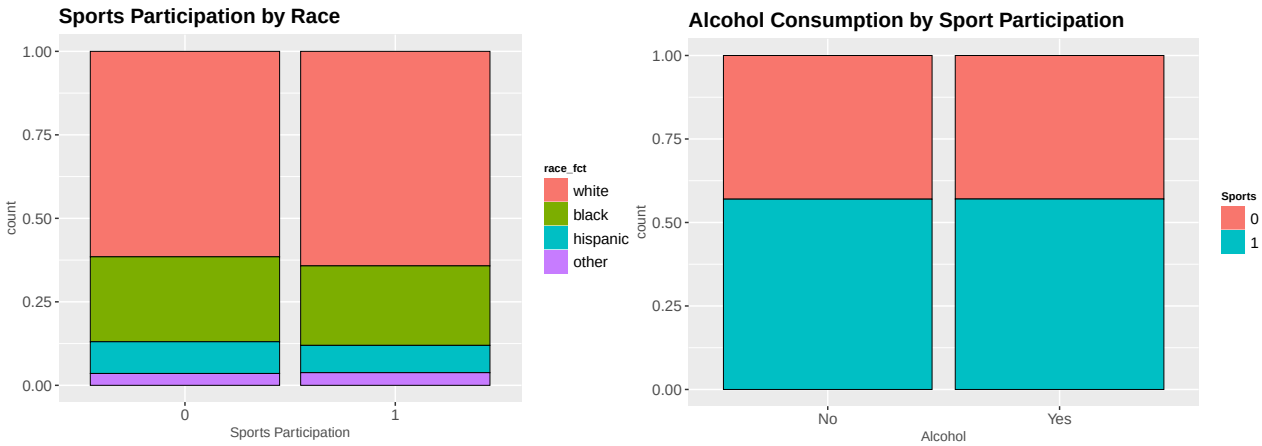

Sports Participation by Parent Education
density
Parents' Average Education
Sports
0
1


Sports Participation by GPA
Density
GPA


Sports Participation by
College Aspiration (Increasing)
count
College Aspiration (Increasing)
Sports
0
1


Sports Participation by Health
(Increasing)
count


Sports Participation by Sex
count
Female
Male
Sex
Sports
0
1



Plots of covariates with four methods of matching:

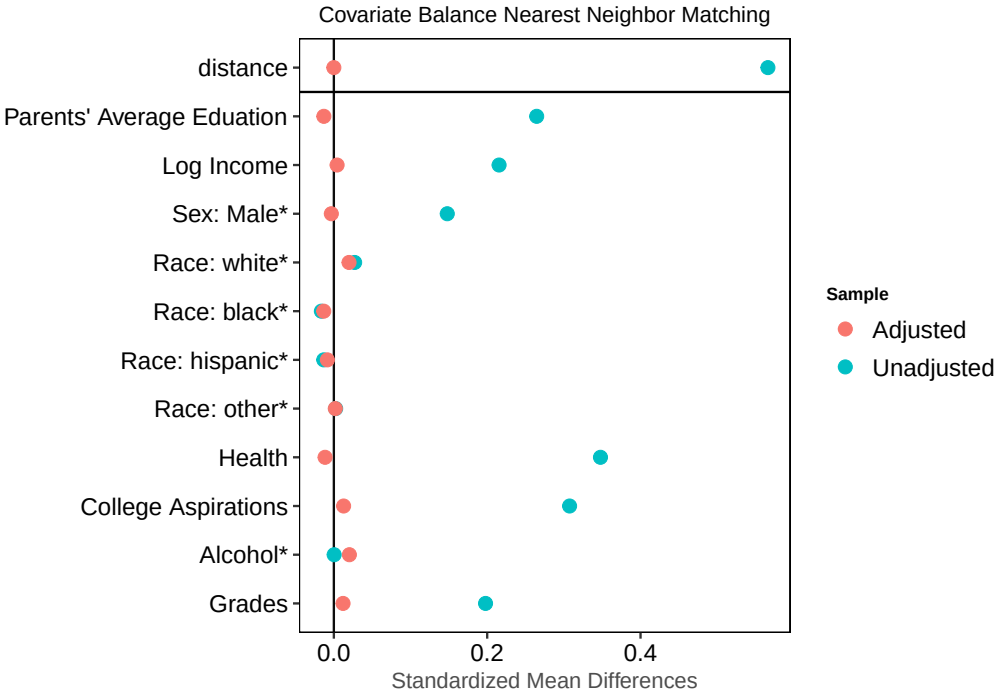


Figure 3: * indicates variables for which the displayed value is the raw (unstandardized) difference in means.

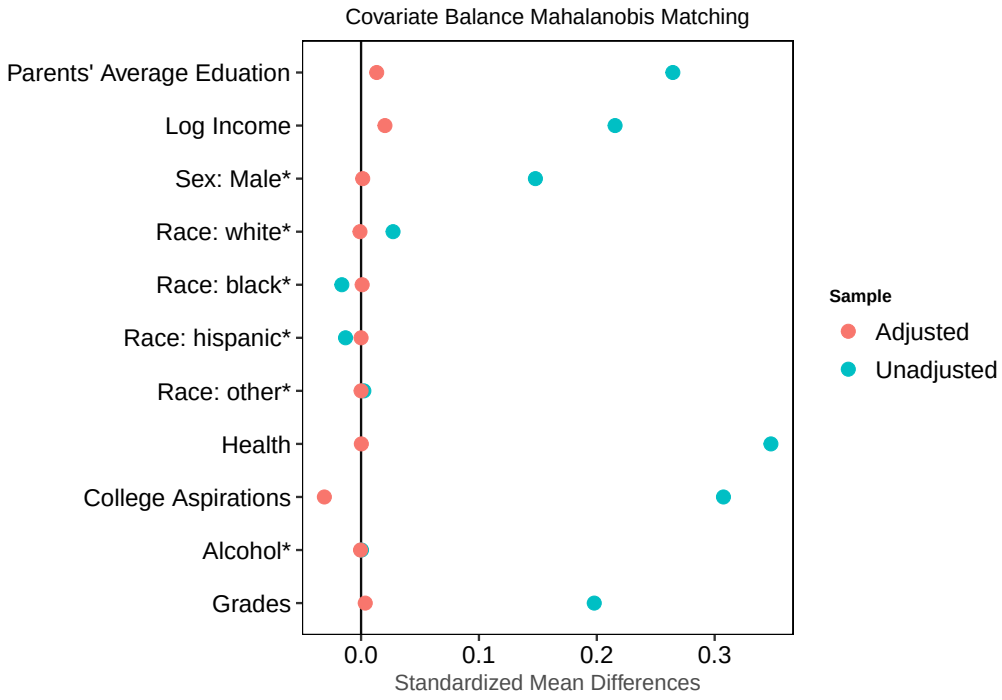


Figure 4: * indicates variables for which the displayed value is the raw (unstandardized) difference in means.

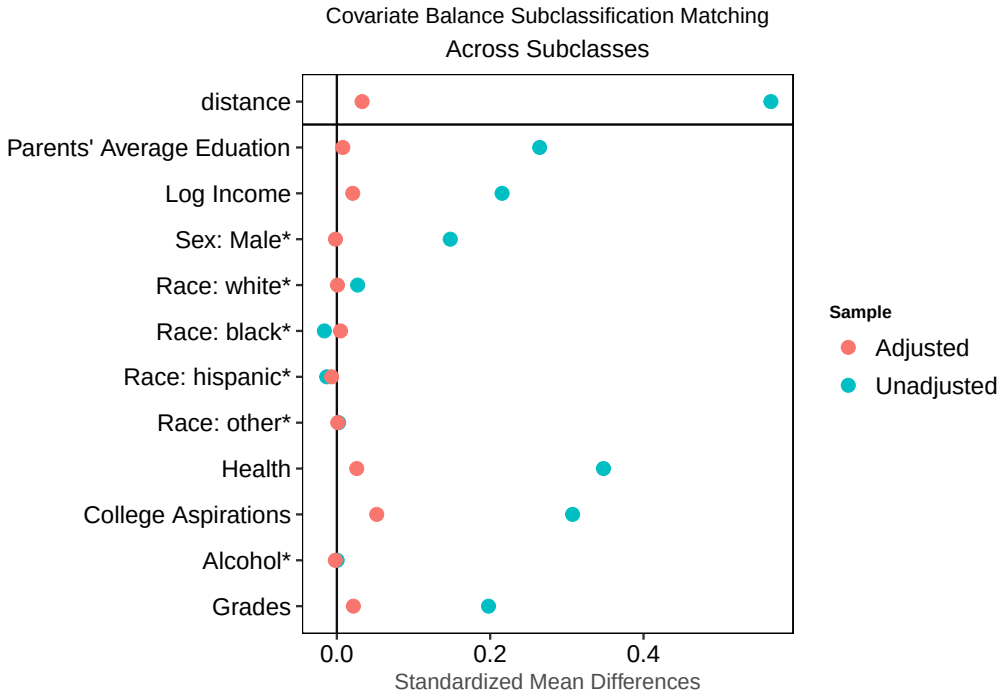


Figure 5: * indicates variables for which the displayed value is the raw (unstandardized) difference in means.

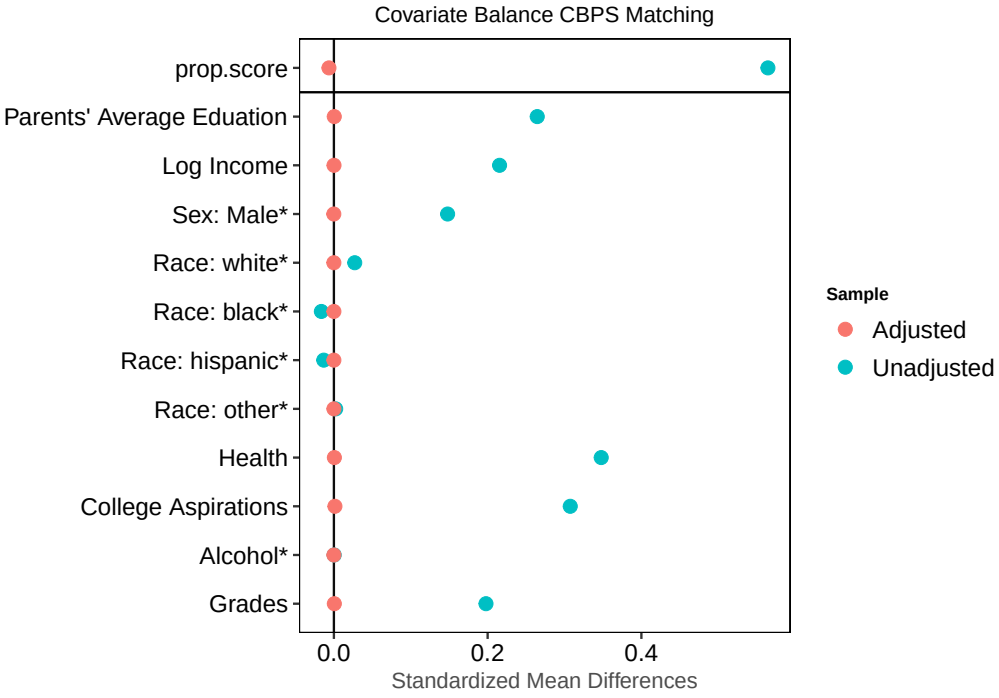


Figure 6: * indicates variables for which the displayed value is the raw (unstandardized) difference in means.

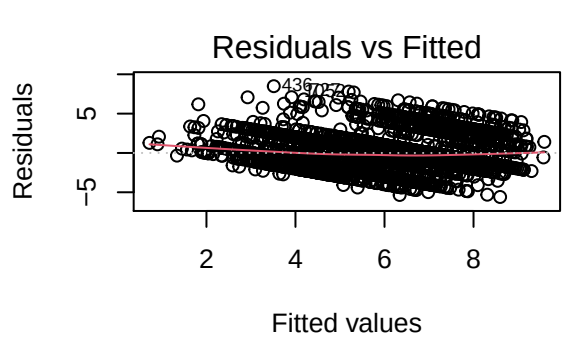
Linear regression models with four methods of matching:

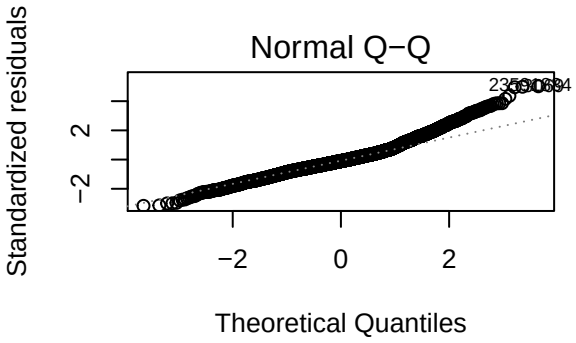
All four methods of matching gave similar results in linear regression models, indicating that the results are robust across matching methods.

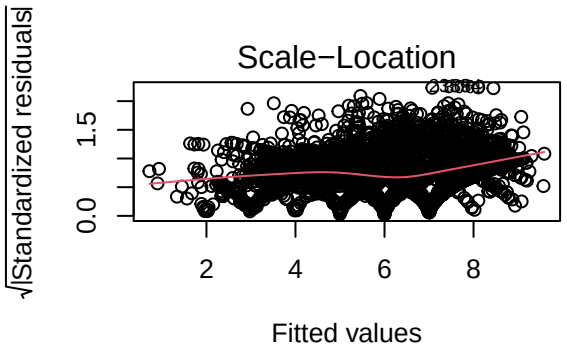
Table 7: Estimates of Sports Participation on Educational Attainment using Different Matching Methods

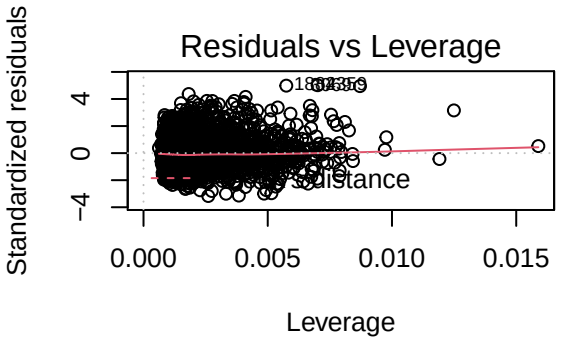
	Dependent variable:			
	Educational Attainment			
	Nearest Neighbor	Mahalanobis	Subclassification	CBPS
	(1)	(2)	(3)	(4)
Sports: Yes	0.093	0.109	0.030	0.042
	(0.078)	(0.076)	(0.065)	(0.064)
Parent Education	0.373***	0.375***	0.368***	0.369***
	(0.023)	(0.023)	(0.021)	(0.021)
Race: Black	−0.003	0.047	0.040	0.042
	(0.088)	(0.087)	(0.080)	(0.080)
Race: Hispanic	0.337**	0.361***	0.280**	0.267**
	(0.132)	(0.132)	(0.121)	(0.123)
Race: Other	0.036	−0.158	−0.057	−0.038
	(0.191)	(0.184)	(0.172)	(0.171)
Alcohol: yes	0.027	0.033	−0.007	−0.005
	(0.073)	(0.071)	(0.066)	(0.066)
Income (log)	0.240***	0.272***	0.275***	0.281***
	(0.045)	(0.046)	(0.042)	(0.043)
Health	0.150***	0.184***	0.179***	0.196***
	(0.043)	(0.043)	(0.039)	(0.039)
Grades	1.009***	1.020***	1.055***	1.072***
	(0.053)	(0.054)	(0.049)	(0.048)
Sex: Male	−0.427***	−0.473***	−0.385***	−0.340***
	(0.074)	(0.073)	(0.066)	(0.066)
College Aspirations	0.431***	0.401***	0.399***	0.413***
	(0.044)	(0.042)	(0.039)	(0.040)
Constant	−2.042***	−2.197***	−2.184***	−2.426***
	(0.297)	(0.285)	(0.262)	(0.274)
Observations	3,064	3,161	3,796	3,796
R^2	0.341	0.350	0.345	0.345
Adjusted R^2	0.339	0.347	0.343	0.343
Note:				*p<0.1; **p<0.05; ***p<0.01

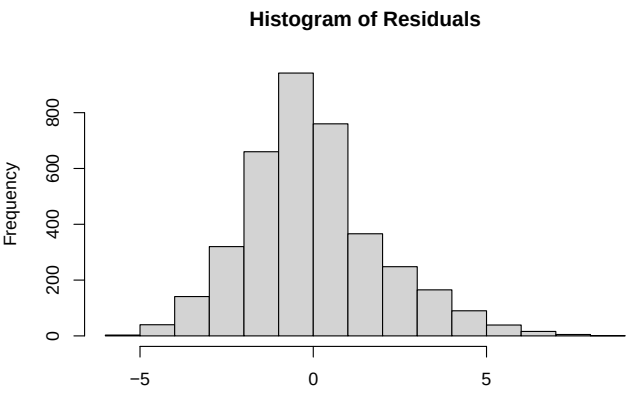
Appendix: Assumptions of Linear Regression:

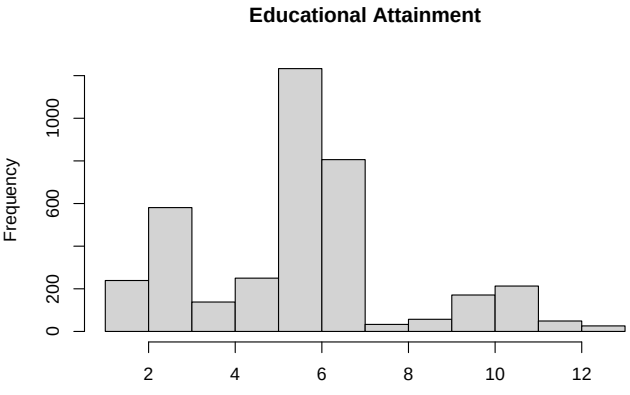


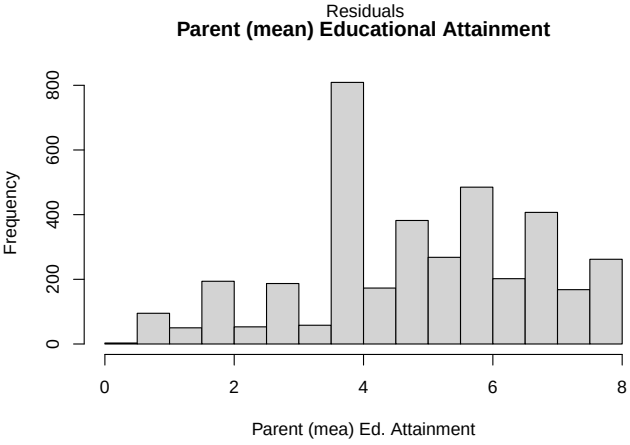


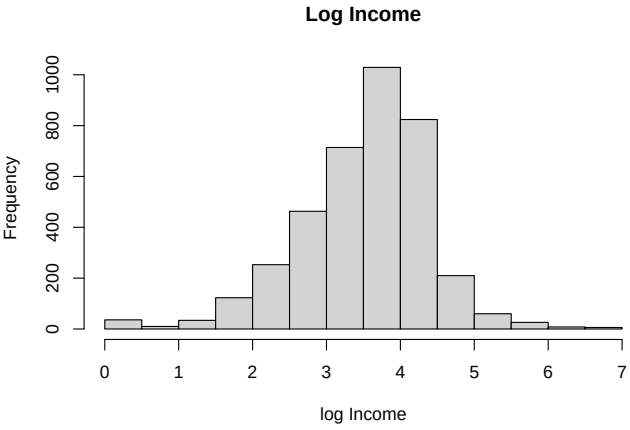


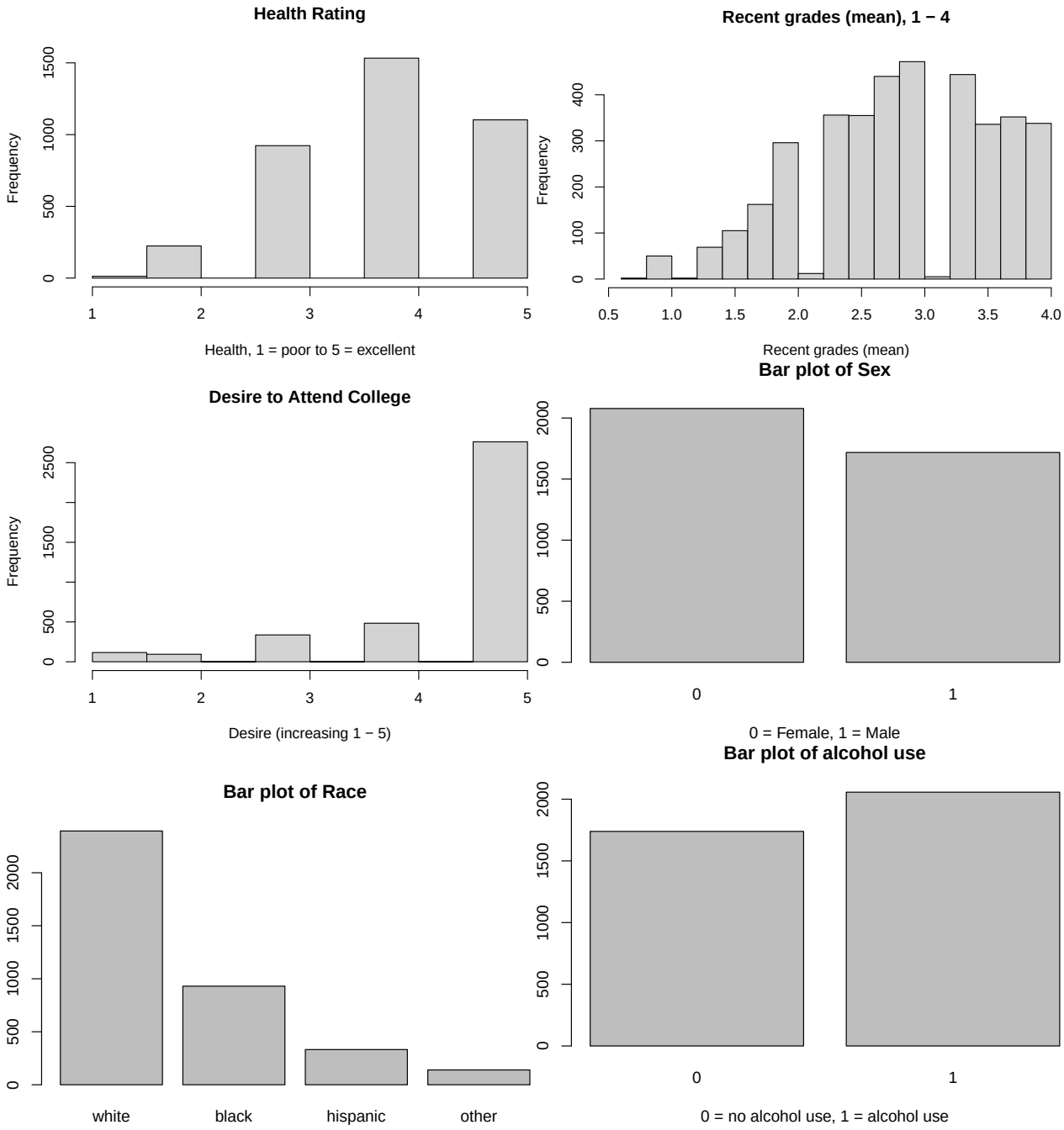












Appendix: ANOVA tables to select best model

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3794	11.96				
2	3793	9.47	1	2.49	998.75	0.0000

Table 8: Simple Bivariate Model (1) vs. Model Controlling gpa (2)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3793	9.47				
2	3792	8.31	1	1.16	530.28	0.0000

Table 9: Model Controlling gpa (2) vs. gpa and parents' education (3)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3792	8.31				
2	3791	8.04	1	0.27	127.06	0.0000

Table 10: gpa + parents' education (3) vs. + college aspirations (4)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3791	8.04				
2	3790	7.99	1	0.05	22.30	0.0000

Table 11: gpa + parents' ed + college aspirations (4) vs. + health

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3790	7.99				
2	3789	7.90	1	0.09	41.55	0.0000

Table 12: gpa + parents' ed + college aspirations + health (5) vs. + log income

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3789	7.90				
2	3788	7.85	1	0.06	26.72	0.0000

Table 13: gpa + parents' ed + college aspirations + health + log(income) (6) vs. + sex

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3788	7.85				
2	3785	7.84	3	0.01	1.64	0.1787

Table 14: Preferred model (7) vs. (7) + race

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3788	7.85				
2	3787	7.85	1	0.00	0.00	0.9455

Table 15: Preferred model (7) vs. (7) + alcohol

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3788	7.85				
2	3784	7.84	4	0.01	1.23	0.2964

Table 16: Preferred model (7) vs. full model

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3789	7.85				
2	3788	7.85	1	0.00	0.43	0.5135

Table 17: Preferred model (7) vs. (7) without sports