POL346 Take Home Final

Sally Cochrane

2021-05-09

Introduction:

Does work experience in high school affect future earnings? The proportion of teenagers participating in the labor force has continued to decline since its peak in 1979, when nearly 60% of American teenagers worked. In contrast, in 2019 only roughly one third of teens were employed (Dickler, 2019). Are teenagers missing out on valuable work experience? On the one hand, some scholars find that students who work during high school have higher earnings than their non-working peers up to a decade later (Carr et al., 1996), perhaps because working during high school improves responsibility and perceptions of one's own competence–features that may help on the labor market (Cunnien et al. 2009). On the other hand, other scholars find negative effects of working more than a "moderate" number of hours during the school year, including increased delinquent behaviors (Staff and Uggen, 2003), lower GPA's (DeSimone, 2006), and decreased chance of attending college (Lee and Orazem, 2010); effects that could lower future earnings. Using data from the Add Health survey from Wave 1 (1994-1995) and Wave 4 (2008-2009), we use multiple regression with covariate balancing propensity score (CBPS) weights to evaluate the theory that non-summer high school employment in wave 1, defined as 10 hours per week or more, affects earnings reported in wave 4. We find evidence consistent with the theory that high school employment is associated with higher future earnings.

Theory:

Some scholars find beneficial effects of non-summer work during high school. Using data from the Youth Development Study, Cunnien et al. (2009) found that working as an adolescent improved "self-efficacy" (the feeling of self-confidence and ability to meet goals in the face of adversity). If Cunnien et al. are correct, this effect could potentially increase future earnings by improving performance in the labor market. And indeed, some scholars do find that high school workers earn more than their non-working peers later in life. Carr et al. (1996), looking at data from students aged 16-19 in 1979, found that those who worked were less likely to attend or complete college, but that working as a teen positively influenced employment status and income even a decade later, offsetting the negative effects of less education. Similarly, Ruhm (1997) found that low to moderate hours of high school employment was positively correlated with future earnings even though working students had slightly lower educational attainments than their non-working counterparts. On the other hand, some scholars find negative effects from moderate to high levels of employment during the school year. Ruhm's (1997) positive findings, above, only held for students working 10 hours or less per week. Similarly, Lee and Orazem (2010) found that working more hours during high school slightly decreased the likelihood of going to college. And DeSimone (2006), using the Monitoring the Future surveys (1991-2004), found that working up to 15 hours per week increased the GPA's of 12th graders, but working more was associated with declining GPA. Staff and Uggen (2003) found that rates of delinquency (drug use, alcohol use, arrest, and school deviance) were higher among adolescents with jobs that increased autonomy, had higher wages, and more working hours. If these scholars are correct, then the negative effects of moderate to high levels of working hours on GPA, college attendance, and delinquency may decrease future earnings by preventing the students from reaching their full academic and employment potentials.

Data:

Data come from the Add Health survey, waves 1 and 4, conducted in 1994-1995 and 2008-2009. The outcome variable of interest was the respondent's personal earnings before taxes in wave 4. This was logged to create a more normal distribution. The explanatory variable of interest was whether the respondent reported working for pay for 10 or more hours per non-summer week in wave 1. 10 or more hours/week was chosen as the "treatment" because we were interested in testing the effects of employment vs. non-employment, and fewer than 10 hours/week of paid work is likely to reflect incidental work like yard work and babysitting rather than true employment (the type of work was not asked in wave 1). Furthermore, the scholarly literature variously labels 10, 15, and 20 hours per week as "moderate to high" employment, and the level at which negative consequences of employment are theorized to begin, which is the effect we are interested in testing. 10 hours/week was appropriate for this dataset because the mean hours of work per week for the sample was 9.2 and the median was 3. This number thus follows some of the literature while also reflecting that high school employment has declined since most of the literature was published, making fewer hours per week appropriate, and it is more likely to reflect the treatment of interest – employment – rather than incidental paid work.

Age was a control variable, ranging from 14 - 20 at wave 1, and 2147 respondents younger than 14 at wave 1 were removed from the sample as the legal working age in the US is 14. Race was simplified to White, Black, and other, as the number in other categories was small. Other control variables included whether the respondent was enrolled at wave 4, as attending school may limit the number of hours one can work and therefore earnings; family income at wave 1, which was logged to make the distribution more normal; whether the respondent was born a US citizen, as citizenship status may affect job opportunities; self-reported overall health, scaled 1 (poor) to 5 (excellent) at waves 1 and 4, as poor health may impede working ability; highest education level attained by wave 4, scaled from 1 (8th grade or less) to 11 (completed a doctoral degree) (see Appendix for full coding); and whether the respondent worked for pay during the summer at wave 1, to make sure we were isolating the effect of school-year employment, rather than work experience in general.

Missing data for working status were removed because the number missing was small (33, or 0.853%), and because we preferred not to impute the explanatory variable. We also preferred not to impute the control variable of summer work, as it is binary and no other variables in the dataset seemed to predict summer work well. Only 79 entries, or 2.04%, did not report summer work, so removing them was unlikely to bias the results. Missing data for earnings in wave 4 were removed when the respondent also did not provide an estimate for their earnings in the following question. Once these missing entries were removed, the variables with the most missingnesss were wave 1 income (895, or 24.3%) and wave 4 earnings (112, or 3.04%). All other missingess was under 0.1%, so imputation was unlikely to bias the results. We used Amelia for imputation. We included the estimated earnings when using Amelia to help impute earnings, but removed estimated earnings for subsequent analysis. We also included several variables to help predict income which were not part of the final analysis (see Appendix for summary of missingness and imputation procedure).

The working and non-working groups were slightly imbalanced in several variables. Workers were older than non-workers; workers also worked summers more often; and more workers were slightly healthier at both waves. Four matching methods were used, and CBPS weights were selected as providing the best balance (see Appendix for plots of balance using four methods). **Table 1** shows summary statistics before and after weighting with CBPS.

4 Methods

We use multiple regression using covariate balancing propensity score (CBPS) weights. Using CBPS weights allows us to analogize this observational study to an experiment in which students between ages 14 and 20 would be randomly assigned to the treatment condition (working 10 or more hours per week) and the control condition (working less than 10 hours per week), because CBPS weights ensure that the "treated" and "control" groups from the sample are approximately balanced on observed pre-treatment variables. Multiple regression is appropriate because we want to determine whether the "treatment" is associated with earnings in wave 4 on its own and when controlling for other variables. Some of the assumptions of linear regression

Table 1: Summary Statistics by Work Status, Before and After CBPS Matching

	Unmatched			Matched	
	No..N.2343.	Yes..N.1339.	Total..N.3682.	No..N.1.	Yes..N.1.
Log(Earnings)	9.34 (2.89)	9.61 (2.71)	9.44 (2.83)	9.342 (2.970)	9.606 (2.706)
Work Summer					
No	1077 (46.0%)	89 (6.6%)	1166 (31.7%)	0 (7.0%)	0 (6.6%)
Yes	1266 (54.0%)	1250 (93.4%)	2516 (68.3%)	1 (93.0%)	1 (93.4%)
Sex					
Female	1266 (54.0%)	677 (50.6%)	1943 (52.8%)	1 (50.7%)	1 (50.6%)
Male	1077 (46.0%)	662 (49.4%)	1739 (47.2%)	0 (49.3%)	0 (49.4%)
Age	15.34 (1.23)	16.30 (1.17)	15.69 (1.29)	16.286 (1.346)	16.299 (1.170)
Race					
Black	619 (26.4%)	247 (18.4%)	866 (23.5%)	0 (18.5%)	0 (18.4%)
Other	231 (9.9%)	94 (7.0%)	325 (8.8%)	0 (7.0%)	0 (7.0%)
White	1493 (63.7%)	998 (74.5%)	2491 (67.7%)	1 (74.5%)	1 (74.5%)
Health W.1	3.88 (0.92)	3.88 (0.87)	3.88 (0.90)	3.880 (0.903)	3.880 (0.869)
Health W.4	3.68 (0.92)	3.64 (0.91)	3.66 (0.92)	3.644 (0.908)	3.644 (0.910)
Born US Citizen					
No	107 (4.6%)	58 (4.3%)	165 (4.5%)	0 (4.3%)	0 (4.3%)
Yes	2236 (95.4%)	1281 (95.7%)	3517 (95.5%)	1 (95.7%)	1 (95.7%)
Ed. Level	5.66 (2.09)	5.62 (1.95)	5.65 (2.04)	5.620 (2.155)	5.618 (1.950)
Enrolled W.4					
No	1996 (85.2%)	1138 (85.0%)	3134 (85.1%)	1 (85.0%)	1 (85.0%)
Yes	347 (14.8%)	201 (15.0%)	548 (14.9%)	0 (15.0%)	0 (15.0%)
Log(Income)	3.49 (0.94)	3.47 (0.88)	3.49 (0.92)	3.475 (0.949)	3.474 (0.877)

may be violated, but the violations are not extreme enough to merit further transformation of the data (see Appendix for diagnostic plots). The residuals vs. fitted values plot shows that the linearity assumption may be violated, but it is close enough to clustering around a horizontal line that we proceed as if it is met. The normality assumption may be violated: the residuals are nearly normally distributed, but there is a second small peak of very negative residuals, as show in the normal Q-Q plot and histogram of residuals. The scale-location plot does not cluster around a perfectly horizontal line, so the equal variance assumption may be violated. Histograms of the variables show that several are not perfectly normally distributed: age is skewed right, health at waves 1 and 4 are skewed left, education level has a second peak at the low end, and log earnings have a second peak around zero. However, transformation of these variables would not make sense, so we proceed as if the equal variance assumption is met. The residuals vs. leverage plot indicates that there may be several outliers, but investigation showed that none of these were due to obvious coding errors, so they were retained. 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 removing control variables from the full model and using ANOVA to assess the improvement in explanatory power balanced with parsimony using the F-statistic (see Appendix for ANOVA tables). The full model included interaction terms between education level and sex, race, enrollment status, and citizenship, because we suspected that education may have different effects on earnings depending on those features. Using this method, the preferred model is:

$$\begin{aligned}\text{log earnings} = \alpha &+ \beta_1(\text{work}_{\text{yes}}) + \beta_2(\text{education level}) \ + \beta_3(\text{sex}_{\text{male}}) + \beta_4(\text{race}_{\text{other}}) + \beta_5(\text{race}_{\text{white}}) + \\ &\beta_6(\text{enrolled}_{\text{yes}}) + \beta_7(\text{age}) + \beta_8(\text{citizen}_{\text{yes}}) \ + \beta_9(\text{health w. 4}) + \beta_{10}(\text{health w. 1}) + \beta_{11}(\text{ed.} \times \text{sex}_{\text{male}}) + \\ &\beta_{12}(\text{ed.} \times \text{race}_{\text{other}}) + \beta_{13}(\text{ed.} \times \text{race}_{\text{white}}) + \beta_{14}(\text{ed.} \times \text{enrolled}_{\text{yes}}) \ + \beta_{15}(\text{ed.} \times \text{citizen}_{\text{yes}}) + \epsilon\end{aligned}$$

Where the coefficients are:

$$\widehat{\log \text{earnings}} = 5.92 + 0.26(\text{work}_{\text{Yes}}) + 0.32(\text{education level}) + 2.8(\text{sex}_{\text{male}}) + 1.77(\text{race}_{\text{other}}) + 1.44(\text{race}_{\text{white}}) + 1.19(\text{enrolled}_{\text{Yes}}) - 0.08(\text{age}) - 1.05(\text{citizen}_{\text{Yes}}) + 0.22(\text{health4}) + 0.15(\text{health1}) - 0.3(\text{ed.} \times \text{sex}_{\text{male}}) - 0.18(\text{ed.} \times \text{race}_{\text{other}}) - 0.21(\text{ed.} \times \text{race}_{\text{white}}) - 0.23(\text{ed.} \times \text{enrolled}_{\text{Yes}}) + 0.33(\text{ed.} \times \text{citizen}_{\text{Yes}})$$

Results:

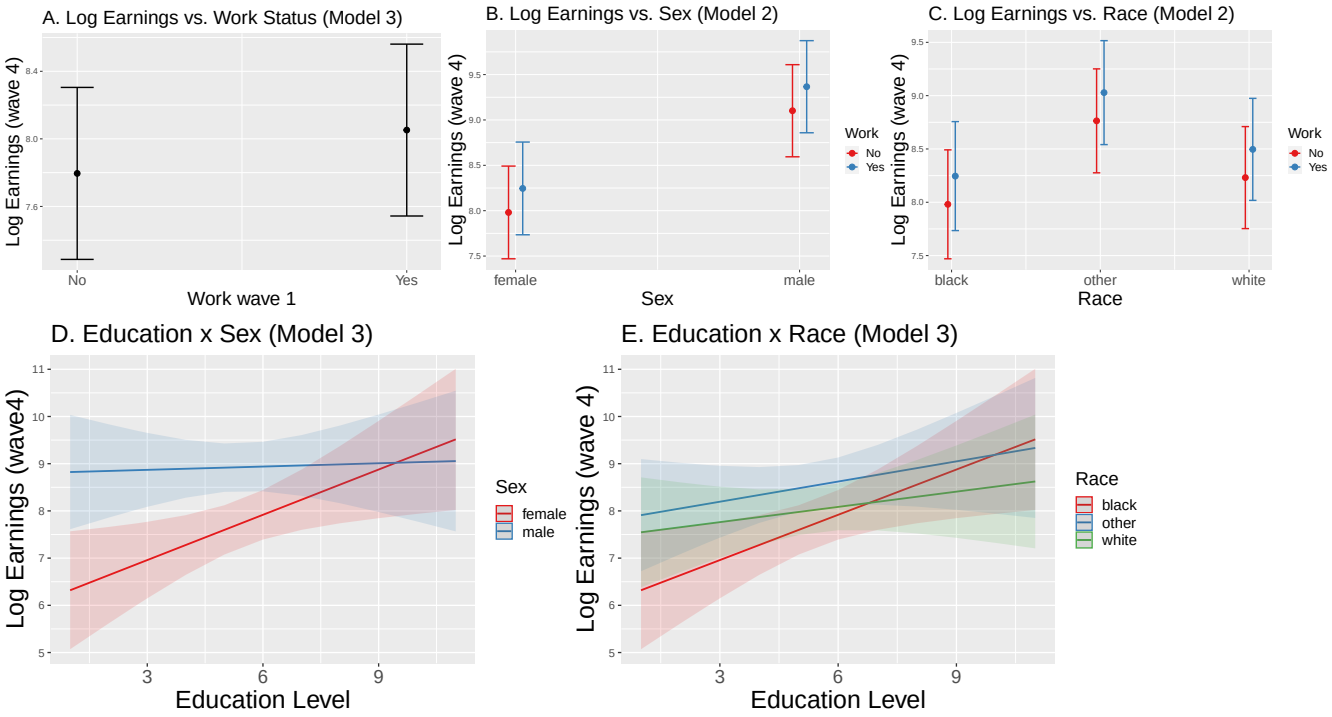
Table 2: Models for Association between Working during School and Later Earnings

	Dependent variable:			
	Log Earnings at wave 4			
	(1)	(2)	(3)	(4)
Work: Yes	0.264*** (0.094)	0.265*** (0.089)	0.257*** (0.088)	0.257*** (0.088)
Education Level		0.285*** (0.023)	0.320** (0.130)	0.314** (0.130)
Sex:Male		1.121*** (0.092)	2.800*** (0.264)	2.804*** (0.265)
Race: Other		0.783*** (0.207)	1.768*** (0.619)	1.772*** (0.620)
Race: White		0.251** (0.117)	1.440*** (0.328)	1.418*** (0.329)
Enrolled w.4		−0.306** (0.129)	1.192** (0.473)	1.210** (0.473)
Log Income w.1				0.075 (0.053)
Age w.1		−0.067* (0.036)	−0.079** (0.036)	−0.080** (0.036)
Born Citizen:Yes		0.639*** (0.233)	−1.052 (0.680)	−1.069 (0.681)
Health w.4		0.221*** (0.053)	0.220*** (0.053)	0.219*** (0.053)
Health w.1		0.158*** (0.054)	0.147*** (0.054)	0.145*** (0.054)
Summer Work				−0.139 (0.177)
Ed. Level x Sex:Male			−0.296*** (0.044)	−0.297*** (0.044)
Ed. Level x Race:Other			−0.177* (0.107)	−0.180* (0.107)
Ed. Level x Race:White			−0.212*** (0.055)	−0.212*** (0.055)
Ed. Level x Enrolled:yes			−0.235*** (0.071)	−0.236*** (0.071)
Ed. Level x Citizen			0.327*** (0.119)	0.323*** (0.119)
Constant	9.342*** (0.066)	6.061*** (0.710)	5.921*** (1.022)	5.883*** (1.030)
Observations	3,682	3,682	3,682	3,682
R^2	0.002	0.097	0.116	0.116
Adjusted R^2	0.002	0.095	0.112	0.112
Note:				*p<0.1; **p<0.05; ***p<0.01

Using linear regression with CBPS weights, we find evidence that working in high school is associated with increased earnings at wave 4. As **Table 2** shows, in the preferred model (3), working 10 hours or more per week for students age 14 -20 is associated with a 1.29 ($e^{0.257}$) -fold increase in the median earnings as compared to not working, with a p-value < 0.01, indicating we may reject the null hypothesis that there is no association between working status and earnings in wave 4. Substantively, this multiplicative effect represents, for example, moving from $30,000 (approximately the median earnings in the whole sample), to $38,700. This result is robust across all four models considered, with a similar estimated association between working status and earnings and similar p-value for those estimates. This result is also robust across all matching methods except for nearest neighbor matching (which has a p-value of < 0.1 for this estimate), indicating that the results are likely not being driven by the balance of the data (see Appendix). **Plot A** shows the relationship between working status and log earnings in the preferred model.

It is also notable that in model 2, model 3 (the preferred model), and model 4 (the full model), several other variables are also statistically significantly associated with earnings. In the preferred model (3), each unit increase in health at wave 1 and health at wave 4 are associated with a 1.16 ($e^{0.147}$) and 1.25 ($e^{0.220}$) -fold increase in median earnings, respectively (p-value < 0.01 for both estimates). Also in the preferred model, the interactions between education level and sex, race, enrollment status, and citizenship were statistically significantly associated with estimated earnings. For males, each unit increase in education level is associated with a multiplicative change of 0.74 ($e^{-0.296}$) in median earnings (p-value < 0.01). As **Plot D** shows, This means that females earn more with each added unit of education than males. However, at lower education

levels, males are estimated to earn more than females. With a high school education (ed. level = 3), males have a 6.77 ($e^{(2.800+3*-0.296)}$) -fold greater median earnings compared to women, and with a college education (= 7), men have a 2.07 ($e^{(2.800+7*-0.296)}$) -fold greater median earnings as compared to women. It is only at education level 9.46, which is beyond having completed a master's degree (ed. level = 9), that women's median earnings are as much or more than men's with each unit increase in education. This finding also holds in model 2, without interaction terms, where males have an estimated median earnings 3.07 ($e^{1.121}$) times higher than females (p-value < 0.01) (see **Plot B**). There is also an interactive effect between education level and race. For the White group, each unit increase in education is associated with a 0.81 ($e^{-0.212}$) multiplicative change in median earnings as compared to the Black group (p-value < 0.01). As **Plot E** shows, this means that people in the Black group earn more with each unit increase in education than those in the White group. However, it is only at education = 6.79, or almost having completed college (=7), that those in the Black group earn more than those in the White group. This finding also holds in model 2, without interaction terms, where those in the White group have an estimated median earnings 1.29 ($e^{0.251}$) times higher than those in the Black group ($p < 0.5$). However, the interaction between education and "other" race is not statistically significantly associated with earnings in model 3 ($p > .05$), though "other" race *is* statistically significantly associated with an increase in median earnings as compared to the Black group *without* interaction terms (by 2.19 times in model 2, and by 5.86 times in model 3, p-value < 0.01 for both estimates) (see **Plot C** for the association between race and earnings from model 2). For those enrolled at wave 4, median earnings are estimated to change by a multiplicative factor of 0.79 ($e^{-0.235}$) for each unit increase in education as compared to the unenrolled (p-value < 0.01). This means that the enrolled group's estimated median earnings are roughly 20% less with each unit increase in education level as compared to the unenrolled. For those who report an education level at wave 4 of 5.07 or higher (the equivalent of completing vocational training after high school), being unenrolled predicts higher earnings than being enrolled. This may indicate that the enrolled group is sacrificing current earnings for education. This result also holds in model 2, where being enrolled is associated with a multiplicative change in median earnings of 0.74 ($e^{-0.306}$) as compared to not being enrolled. Finally, for each unit increase in education, being born a US citizen is associated with a 1.39 ($e^{0.327}$) -fold increase in estimated median earnings as compared to not being born a US citizen. For a college graduate (education = 7), this translates into median earnings 3.45 times greater for a born US citizen than a non-citizen. (See Appendix for additional plots). The estimates from the full model are similar to the estimates from the preferred model, with similar statistical significance.



Discussion:

There are several limitations to this study. First, self-reported answers may not be accurate, and may bias the results in unknown ways. Students may under- or over- report working hours or earnings, and more objective measures would be preferable for further research than self-report. Second, missing data, especially for working hours and earnings, may have biased the results in unknown ways, though the chance is small as the number of missing entries was small. Third, this report only looked at the "treatment" of working 10 + hours per week vs. the "control" of working less than 10 hours per week as a way to evaluate the effects of employment on future earnings. However, it may be that hours worked has a continuous effect on future earnings, and further studies could determine whether fewer or more hours of work per week have different effects. Fourth, future studies could determine if different types of employment have different effects on future earnings (a topic that was not possible with the Add Health survey here, which only asked about paid work in general). Fifth, we only looked at reported work status at wave 1, but this misses students who may have begun working later in high school. Future studies could consider the total amount of work done in high school. Finally, we cannot say that the association between the explanatory variables and earnings at wave 4 were causal. Although the CBPS weighting 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 work during high school and have higher earnings after high school. Since the subjects were drawn from a representative sample of high schools around the United States, the results are generalizable to U.S. adolescents ages 14 - 20 in 1994-1995, who were 25-32 when earnings were measured in 2008-2009.

Conclusion

We find evidence that working 10 or more hours per week in high school is associated with higher earnings roughly 14 years later at wave 4. In all four models considered here, working during high school was associated with median earnings roughly 1.3 times higher at wave 4 than the earnings of those who did not work at wave 1. This result was robust whether working status was the only explanatory variable and when a number of other variables such as sex, race, family income, citizenship, education level, enrollment status at wave 4, and health were controlled for. This finding was also robust across three of four matching methods, indicating it is likely not driven by the balance of the data. These results undermine the hypothesis that working moderate to large numbers of hours during high school is associated with negative effects on future earnings. In the models which included control variables, earnings was also statistically significantly associated with education level, sex, race, enrollment status, family income at wave 1, and health at waves 1 and 4, though it was not statistically significantly associated with whether the respondent worked during the summer or the respondent's age. In the preferred model, there were also statistically significant interaction effects between education level at wave 4 and sex, race, enrollment status, and citizenship status, indicating that education has different effects on earnings depending on those variables. While the data were matched using CBPS to obtain treatment and control groups that were similar on all observed covariates, which mimics random sampling and brings us closer to inferring a causal relationship between working and higher future earnings, 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.

Bibliography:

Data:

- Harris, Kathleen Mullan, and Richard J. Udry. 2018. "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 Sources:

- Carr, Rhoda V., James D. Wright, and Charles J. Brody. 1996. "Effects of High School Work Experience a Decade Later: Evidence from the National Longitudinal Survey." *Sociology of Education* 69: 66-81.

- Cunnien, Keith A., Nicole MartinRogers adn Jeylan T. Mortimer. 2009. "Adolescent work experience and self-efficacy." *International Journal of Sociology and Social Policy* 29, no 3/4: 164-175.

- DeSimone, Jeff. 2006. "Academic performance and part-time employment among high school seniors." *The B.E. Journal of Economic Analysis & Policy* 6, no. 10:1466

- Dickler, Jessica. 2019. "Why so few teenagers have jobs anymore." *CNBC* Oct. 6, 2019.

- Lee, Chanyoung and Peter F. Orazem. 2010. "High school employment, school performance and college entry." *Economics of Education Review* 29, no. 1: 29-39.

- Ruhm, Christopher J. 1997. "Is high school employment consumption or investment?" *Journal of Labor Economics* 15, no. 4:735–776

- Staff, Jeremy, and Christopher Uggen. 2003. "The Fruits of Good Work: Early Work Experiences and Adolescent Deviance." *Journal of Research in Crime and Delinquency* 40, no. 3: 263-290.

Appendix:

Coding for Education Level at wave 4:

Table 3 Shows the coding for the education level variable at wave 4.

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

Table 3: Respondent's Education Level at Wave 4

Missingness:

The following show missingess before removing NA's in explanatory and outcome variables:

	variable	n_miss	pct_miss
1	log_income	950	24.55
2	log_earnings	221	5.71
3	work_summer	79	2.04
4	hrs_nonsummer	33	0.85
5	work_nonsummer	33	0.85
6	race_fct	3	0.08
7	health1	3	0.08
8	sex_fct	1	0.03
9	ed_level	1	0.03
10	enrolled	1	0.03
11	AID	0	0.00
12	age_full	0	0.00
13	health4	0	0.00
14	citizen_fct	0	0.00

Table 4: Missingness before removing missing outcome and explanatory variables

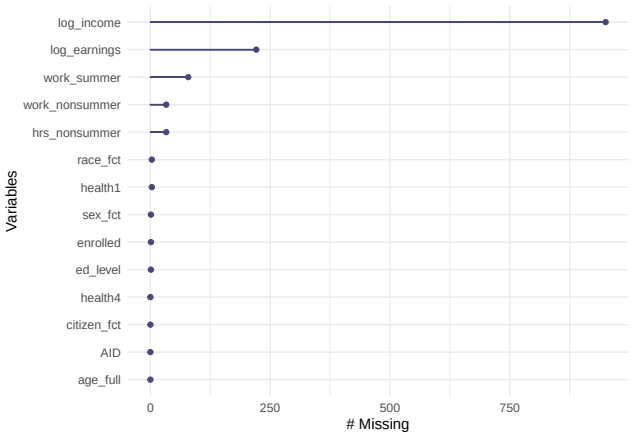


Figure 1: Missingness before removing missing outcome and explanatory variables.

The following show missingness after removing NA's in explanatory and outcome variables:

	variable	n_miss	pct_miss
1	log_income	895	24.31
2	log_earnings	112	3.04
3	race_fct	2	0.05
4	health1	1	0.03
5	ed_level	1	0.03
6	AID	0	0.00
7	age_full	0	0.00
8	hrs_nonsummer	0	0.00
9	work_nonsummer	0	0.00
10	work_summer	0	0.00
11	sex_fct	0	0.00
12	health4	0	0.00
13	citizen_fct	0	0.00
14	enrolled	0	0.00

Table 5: Missingness after removing missing outcome and explanatory variables

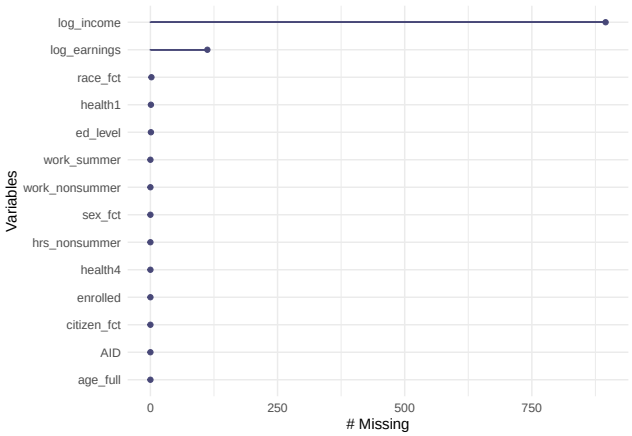


Figure 2: Missingness after removing missing outcome and explanatory variables, before imputing with Amelia.

The following show missingness after imputation with Amelia:

	variable	n_miss	pct_miss
1	AID	0	0.00
2	age_full	0	0.00
3	work_nonsummer	0	0.00
4	hrs_nonsummer	0	0.00
5	work_summer	0	0.00
6	log_earnings	0	0.00
7	sex_fct	0	0.00
8	race_fct	0	0.00
9	health1	0	0.00
10	health4	0	0.00
11	citizen_fct	0	0.00
12	ed_level	0	0.00
13	enrolled	0	0.00
14	log_income	0	0.00
15	cbps_weights	0	0.00

Table 6: Missingness after imputation with Amelia

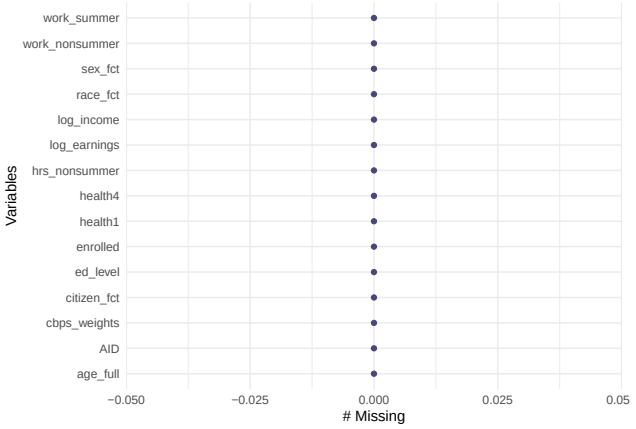
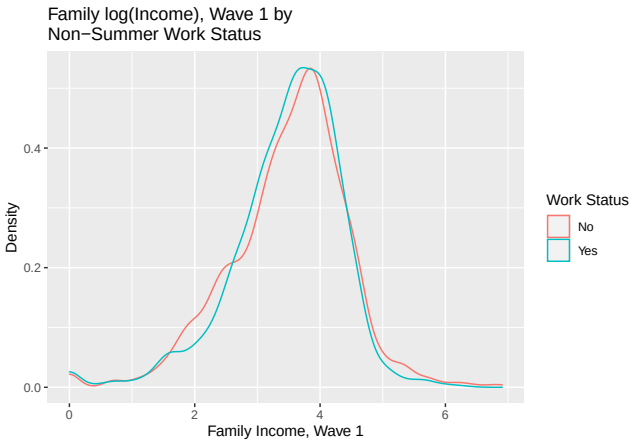
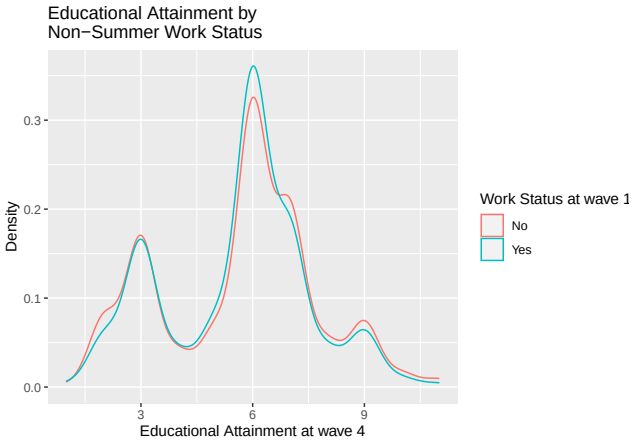


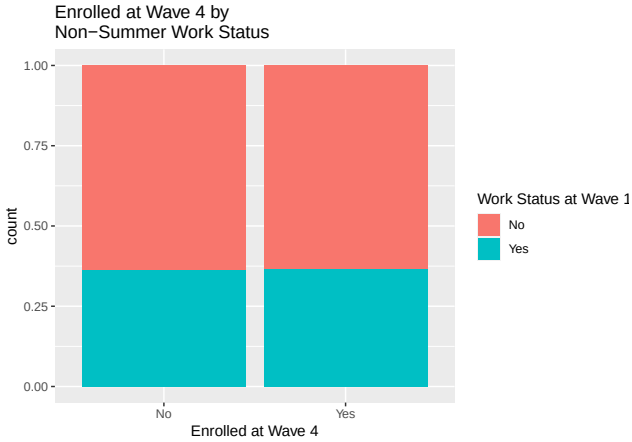
Figure 3: Missingness after imputation with Amelia

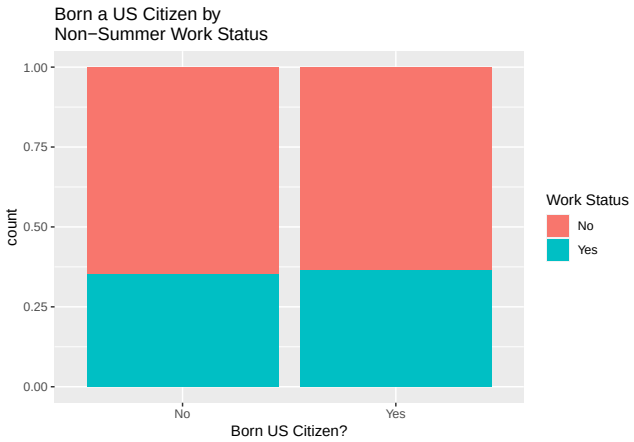
Matching methods:

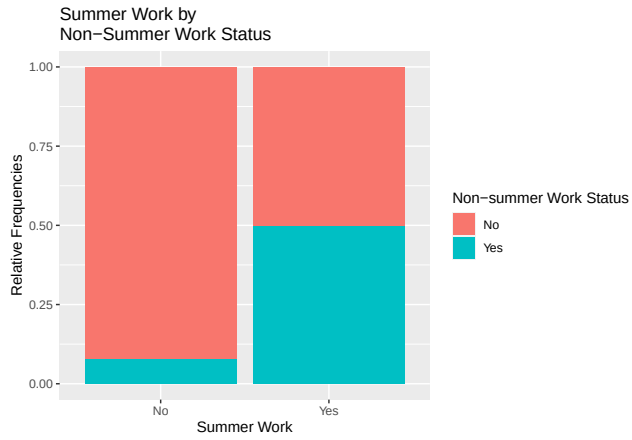
As the following plots show, the working and non-working groups were imbalanced in several variables.

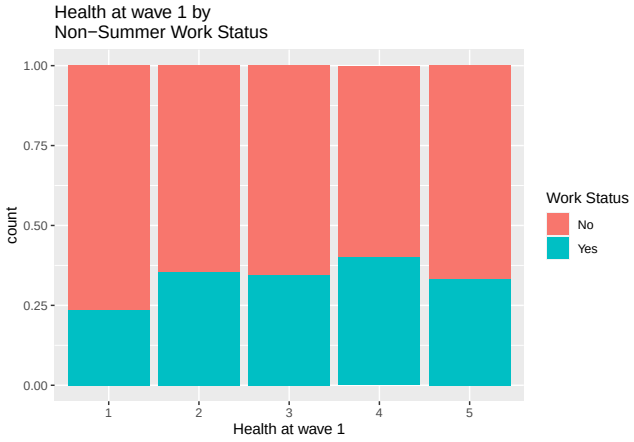

Family log(Income), Wave 1 by
Non−Summer Work Status
Density
0.0
0.2
0.4
0
2
4
6
Family Income, Wave 1
Work Status
No
Yes

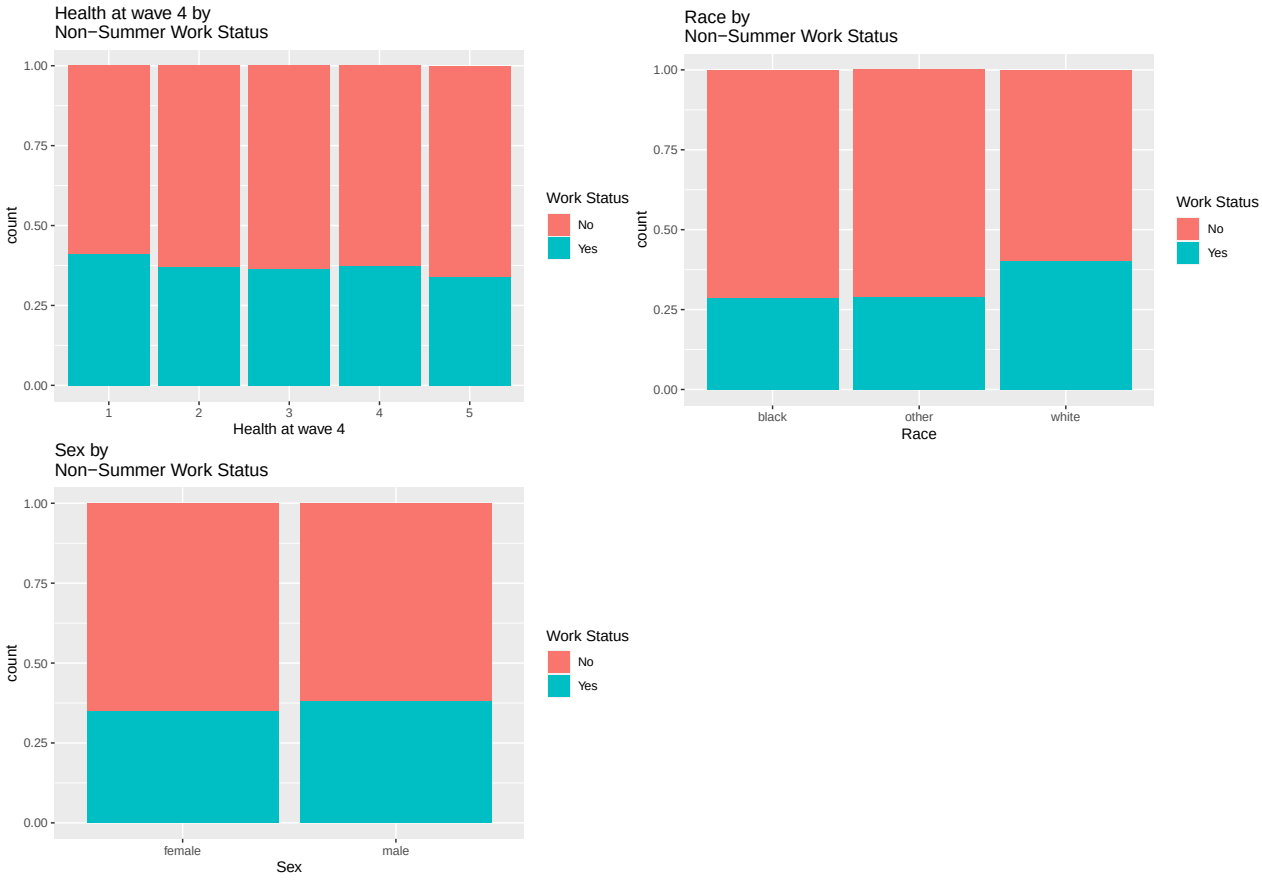

Educational Attainment by
Non−Summer Work Status
Density
0.0
0.1
0.2
0.3
3
6
9
Educational Attainment at wave 4
Work Status at wave 1
No
Yes


Enrolled at Wave 4 by
Non−Summer Work Status
count
0.00
0.25
0.50
0.75
1.00
No
Yes
Enrolled at Wave 4
Work Status at Wave 1
No
Yes

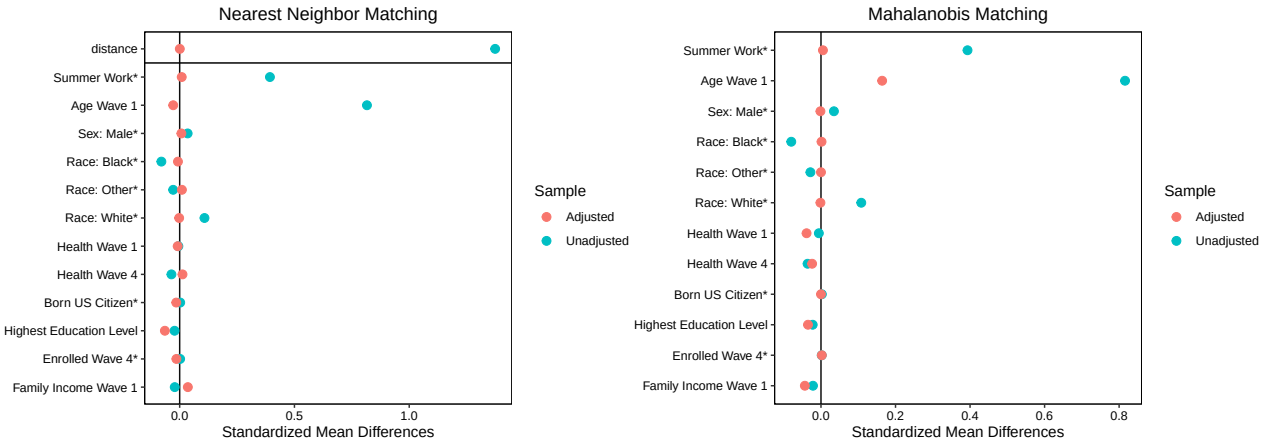

Born a US Citizen by
Non−Summer Work Status
count
0.00
0.25
0.50
0.75
1.00
No
Yes
Born US Citizen?
Work Status
No
Yes

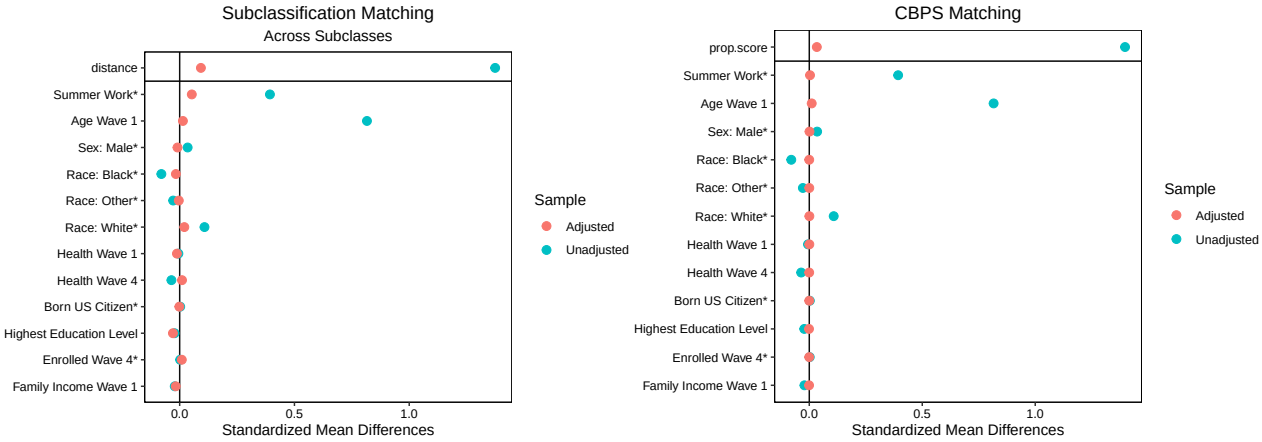

Summer Work by
Non−Summer Work Status
Relative Frequencies
0.00
0.25
0.50
0.75
1.00
No
Yes
Summer Work
Non−summer Work Status
No
Yes


Health at wave 1 by
Non−Summer Work Status
count
0.00
0.25
0.50
0.75
1.00
1
2
3
4
5
Health at wave 1
Work Status
No
Yes



The following show the balance of the covariates using four matcihng methods:



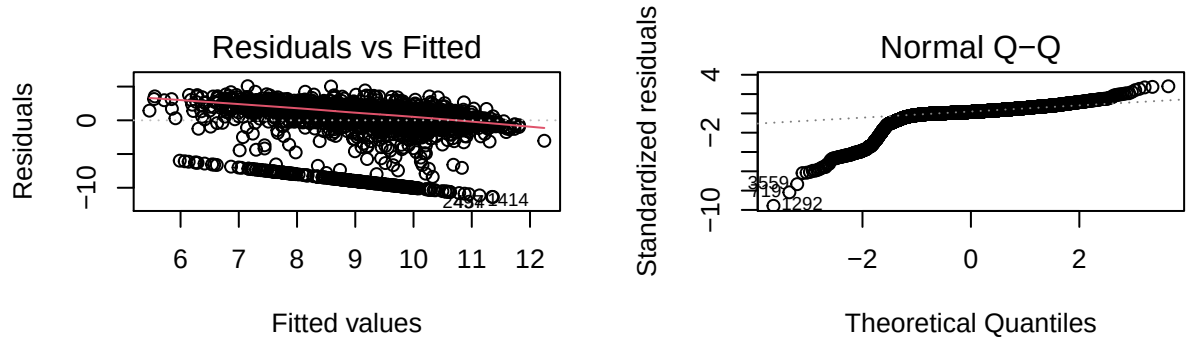


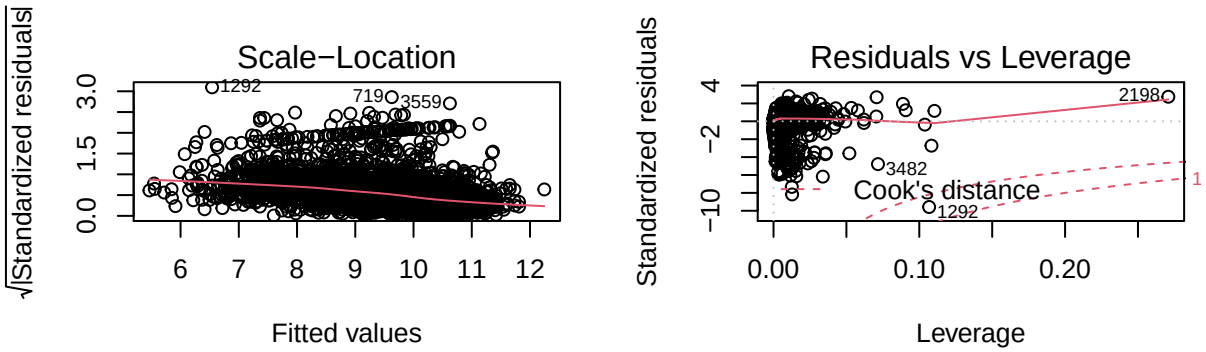
The following table shows linear regression using the four matching methods:

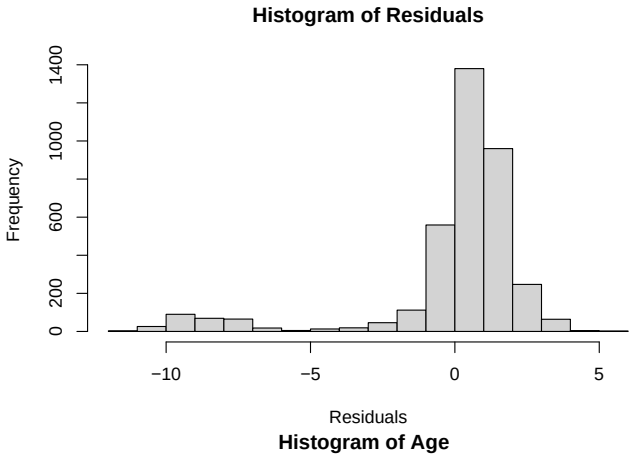
Table 7: Log Earnings wave 4 vs. Work Status wave 1 using Different Matching Methods

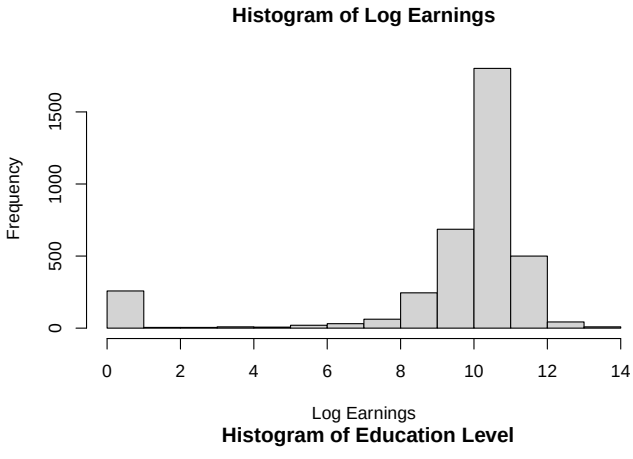
	Dependent variable:			
	Log Earnings at wave 4			
	Nearest Neighbor	Mahalanobis	Subclassification	CBPS
	(1)	(2)	(3)	(4)
Work:yes	0.245*	0.272**	0.251***	0.265***
	(0.126)	(0.127)	(0.093)	(0.089)
Sex:Male	0.249***	0.231***	0.295***	0.285***
	(0.031)	(0.032)	(0.023)	(0.023)
Race:other	1.166***	1.105***	1.101***	1.121***
	(0.121)	(0.122)	(0.092)	(0.092)
Race:white	0.669**	0.445*	0.508**	0.783***
	(0.270)	(0.265)	(0.202)	(0.207)
Enrolled:yes	0.221	−0.036	0.112	0.251**
	(0.152)	(0.154)	(0.116)	(0.117)
Age	−0.553***	−0.498***	−0.278**	−0.306**
	(0.168)	(0.170)	(0.131)	(0.129)
Health w.4	−0.043	−0.020	−0.039	−0.067*
	(0.050)	(0.051)	(0.038)	(0.036)
Health w.1	1.300***	0.817***	0.602**	0.639***
	(0.309)	(0.305)	(0.234)	(0.233)
Constant	0.230***	0.279***	0.249***	0.221***
	(0.070)	(0.070)	(0.053)	(0.053)
health1	0.133*	0.097	0.175***	0.158***
	(0.073)	(0.075)	(0.054)	(0.054)
Constant	5.362***	5.719***	5.553***	6.061***
	(0.957)	(0.963)	(0.732)	(0.710)
Observations	1,994	2,049	3,682	3,682
R^2	0.100	0.083	0.099	0.097
Adjusted R^2	0.095	0.079	0.096	0.095
Note:				*p<0.1; **p<0.05; ***p<0.01

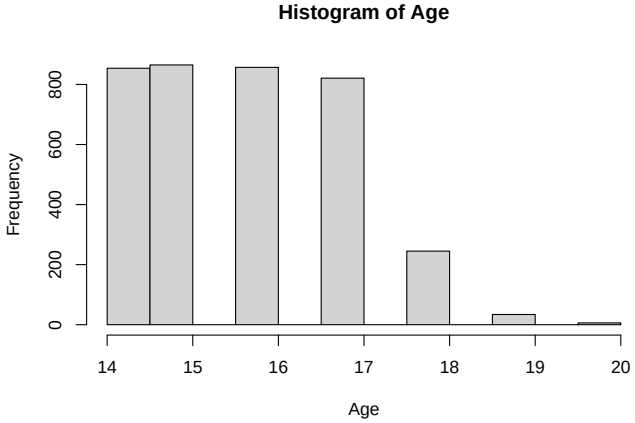
Diagnostic Plots for Assumptions of Linear Regression:

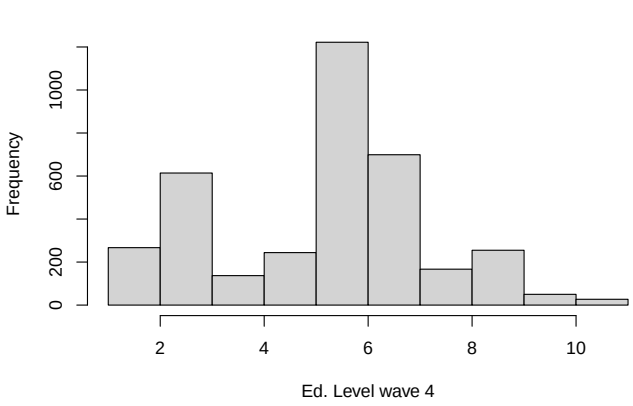


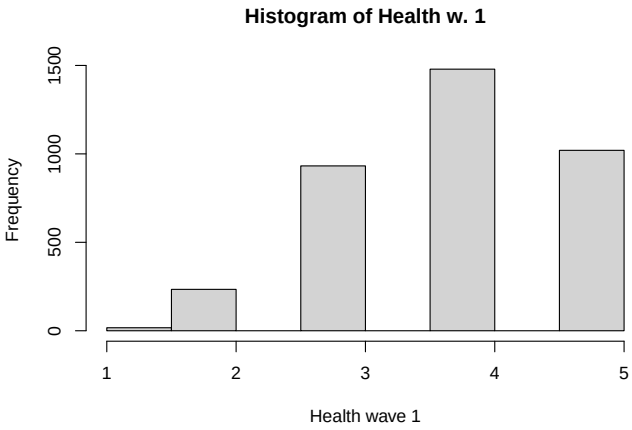


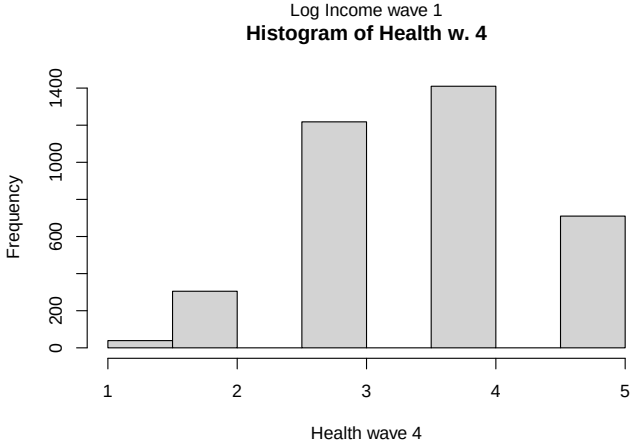






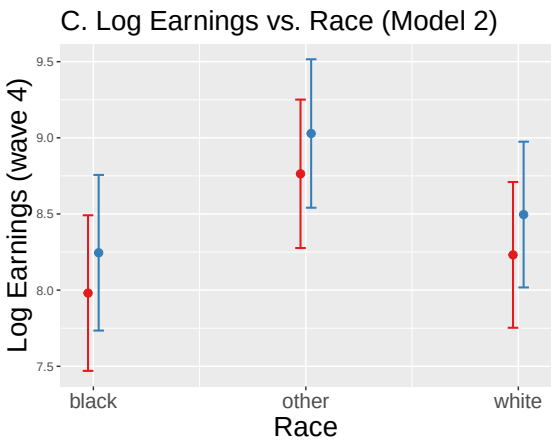


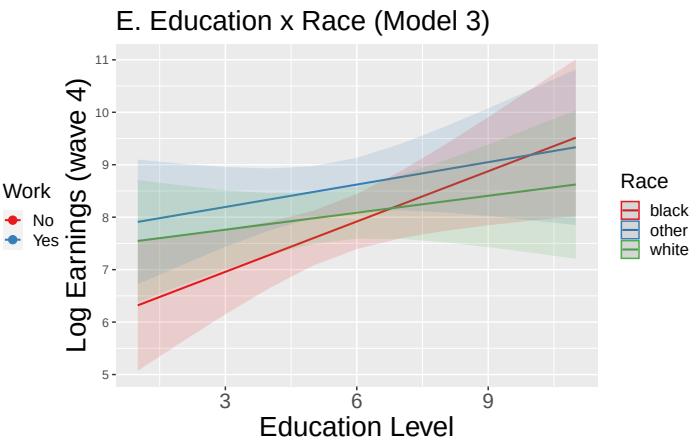


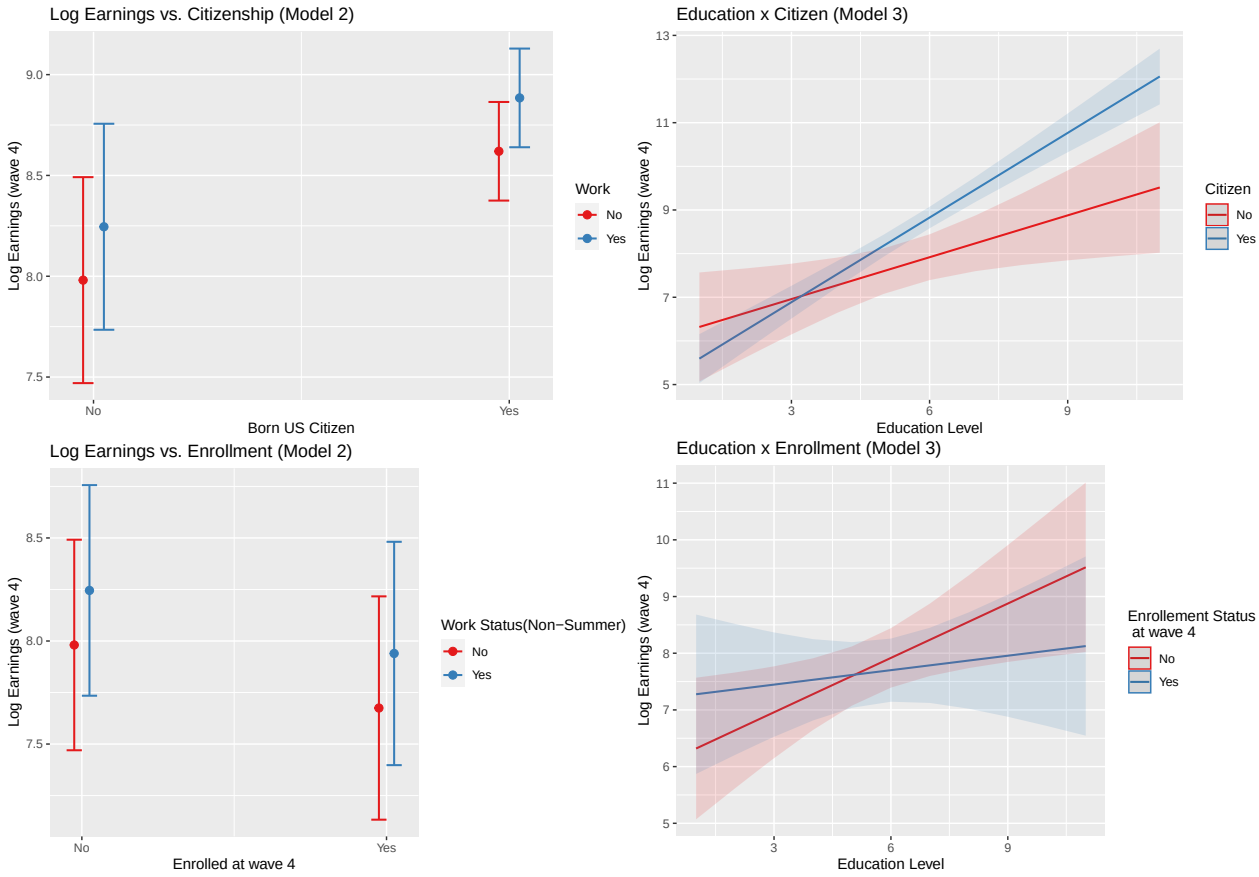


Additional Plots of Model Variables:

The following are additional plots of control variables in model 2 and interaction terms in model 3.







ANOVA to select best models:

The following tables show how the preferred model (model 3 in **Table 2**) is better than the preferred model without interaction terms (model 2), which is better than the simple model (model 1), but that the full model (model 4) is no better than the preferred model (model 3).

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3680	16.13				
2	3671	14.60	9	1.53	42.86	0.0000

Table 8: Model 1 vs. Model 2

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3671	14.60				
2	3666	14.30	5	0.30	15.48	0.0000

Table 9: Model 2 vs. Model 3

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3666	14.30				
2	3664	14.28	2	0.01	1.32	0.2661

Table 10: Model 3 vs. Model 4

The following tables show how the preferred model was selected by removing variables one at a time from the full model until removing more variables decreased the explanatory power of the model.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3664	14.28				
2	3665	14.29	-1	-0.00	0.61	0.4342

Table 11: Full model vs. Remove Summer Work

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3666	14.30				
2	3665	14.29	1	0.01	2.04	0.1536

Table 12: Full model - summer work vs. - income

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3667	14.31				
2	3666	14.30	1	0.02	4.97	0.0258

Table 13: Preferred model vs. - Age, Can't take out age

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3667	14.32				
2	3666	14.30	1	0.03	7.43	0.0065

Table 14: Preferred model vs. - Health 1, Can't take out health1

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3667	14.36				
2	3666	14.30	1	0.07	17.53	0.0000

Table 15: Preferred model vs. - Health 4, Can't take out health4

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3667	14.32				
2	3666	14.30	1	0.03	7.59	0.0059

Table 16: Preferred model vs. - Ed x Citizen, Can't take out Ed x Citizen

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3667	14.34				
2	3666	14.30	1	0.04	11.03	0.0009

Table 17: Preferred model vs. - Ed x Enrolled, Can't take out Ed x Enrolled

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3668	14.35				
2	3666	14.30	2	0.06	7.37	0.0006

Table 18: Preferred model vs. - Ed x Race, Can't take out Ed X Race

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3667	14.47				
2	3666	14.30	1	0.18	45.49	0.0000

Table 19: Preferred model vs. - Ed x Sex, Can't take out Ed X Sex