***PP290: Stata- Final Project\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_***

**Race and Returns to Education**

*By: Inthisar Kamal*

Table of Contents

Summary of Dataset …………………………………………………………………………………………2

Appendix A: Codebook………………………………………………………………………………………………………6

Appendix B: Code…………………………………………………………………………………………………………....8

**Contents of dataset**

This dataset is composed of variables that were merged from three different datasets related to different levels of education. It primarily focuses on both fifth-grade achievement scores and returns to college education in 2004. The survey is composed of 32,042 households. The first set of variables are related to fifth grade reading and math scores in 2004 while variables from the second and third dataset look at college graduation rates (science test scores were dropped as not all states had standardized science tests). Among the variables included are also socio-economic, family background and lastly race indicators. The data from the survey is used to document the relationship between race and education achievement. While the race labels are not all encompassing the data categorizes ethnicity into four dummy variables; White, Black, Hispanic and Other.

**Comparing Average Math and Reading Scores by Race**

|  |  |  |
| --- | --- | --- |
| **5th grade Reading Score Distribution Breakdown** | | |
|  |  |
| Percentiles | | | Smallest |
| 1% 71.68761 | | | 63.0874 |
| 5% 81.75681 | | | 63.30107 |
| 10% 86.41749 | | | 63.63493 | Obs | | 8,105 | |
| 25% 93.66891 | | | 64.03556 | Sum of Wgt. | | 8,105 | |
| 50% 101.5302 | | | Mean | 100 | |  | |
| Largest | | | Std. Dev. | 10 | |  | |
| 75% 107.7444 | | | 117.4619 |  |  | |
| 90% 112.1425 | | | 117.4931 | Variance | | 99.99999 | |
| 95% 113.8118 | | | 117.4931 | Skewness | | -.67794 | |
| 99% 115.8683 | | | 117.6534 | Kurtosis | | 3.172477 | |

|  |  |  |
| --- | --- | --- |
|  | **5th Grade Math Score Distribution Breakdown** | |
|  | |  |  |
| Percentiles | Smallest |  |
| 1% | | 72.52747 | 66.5212 |
| 5% | | 80.14838 | 66.53579 |
| 10% | | 85.25005 | 66.54066 | Obs | | 8,105 |
| 25% | | 94.2862 | 66.56497 | Sum of Wgt. | | 8,105 |
| 50% | | 101.8439 | Mean | 100 | |  |
|  | | Largest | Std. Dev. | 10 | |  |
| 75% | | 107.5292 | 116.8279 |  | |  |
| 90% | | 111.5025 | 117.0857 | Variance | | 99.99999 |
| 95% | | 113.1512 | 117.0857 | Skewness | | -.7970415 |
| 99% | | 115.3981 | 117.0857 | Kurtosis | | 3.179688 |

*Table 1: Table 2:*

**Summary of 5th grade Math Score Summary of 5th grade Reading Score**

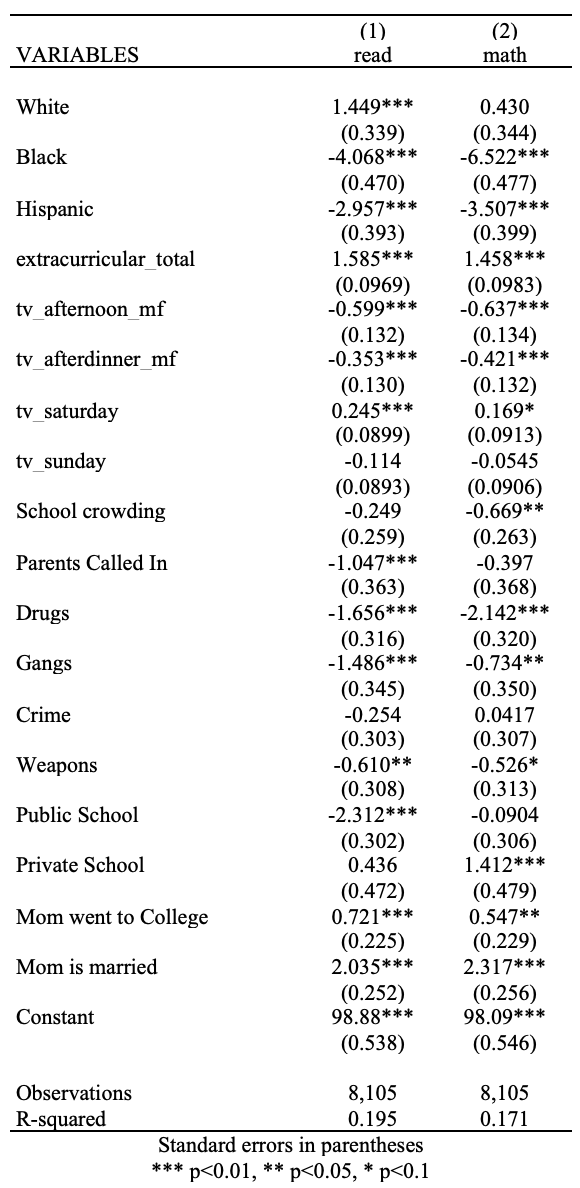
*Table 3: Table 4:*

|  |  |
| --- | --- |
|  | (1) (2) (3) |
| VARIABLES | Mean Std.Dev. Freq |
|  |  |
| White | 102.04356 8.9465693 5,063 |
| Black | 92.463335 10.264275 733 |
| Hispanic | 96.213186 10.31866 1,465 |
| Other | 100.8596 10.206753 844 |
| Total | 99.999999 9.9999995 8,105 |
|  |  |

|  |  |
| --- | --- |
|  | (1) (2) (3) |
| VARIABLES | Mean Std.Dev. Freq |
|  |  |
| White | 102.26931 8.9978369. 5,063 |
| Black | 93.725811 10.274655 733 |
| Hispanic | 95.41872 10.116199 1,465 |
| Other | 99.787927 10.203333 844 |
| Total | 99.999999 9.9999995 8,105 |
|  |  |

***Findings:***

Our data shows that in this sample, white students had higher average scores than Black, Hispanic and other students. The average math and reading score associated with White students lies just above the 50th percentile of the overall math and reading distribution. Hispanic students had a higher average reading and math score than Black students. Their average lay above the 25th percentile in the distribution. Black students had the lowest average standardized test score when compared to the rest with the average score lying below the 25th percentile. The average test score for Hispanic and Black students do not differ significantly but there is a wider gap when compared to the average score associated with white students.

**Regression analysis summary for Reading and Math Scores**

***Findings:***

Key thing to note is that upon regressing school indicators, family background and race on math and reading scores, the variance in these indicators explain only 20% of the variation in the reading and math scores. The ‘other’ race category was omitted to avoid perfect collinearity. The data states:

* Among the variables, mothers’ marital status had the strongest relationship with math and reading scores. This is also statistically significant. Students whose mothers were married scored roughly two points greater than students whose mothers were not married
* Students whose mothers went to college scored 0.7 and 0.5 points higher on average than students whose mothers did not go to college
* White students scored 1.5 points higher on average than non-white students in reading, holding all other factors constant. This is statistically significant (whereas they scored .43 points higher in math but this is not statistically significant)
* Black students scored 4.07 points lower on average than non-black students in reading and 6.5 points lower on average in math, holding all other factors constant. Both of these coefficients are statistically significant
* Hispanic students scored 3 points lower on average than non-Hispanic students in reading and 3.5 points lower on average in math, holding all other factors constant. Both of these coefficients are statistically significant
* Extra-curricular activities played a significant role as well. Each additional part time activity that a child was involved in was associated with a 1.5 point increase in their reading and math score
* Schools that had issues with Drugs, Gangs, Crime and Weapons seemed to score lower on average than schools that did not have these issues
* Public school students scored lower on average while private school students scored higher on average

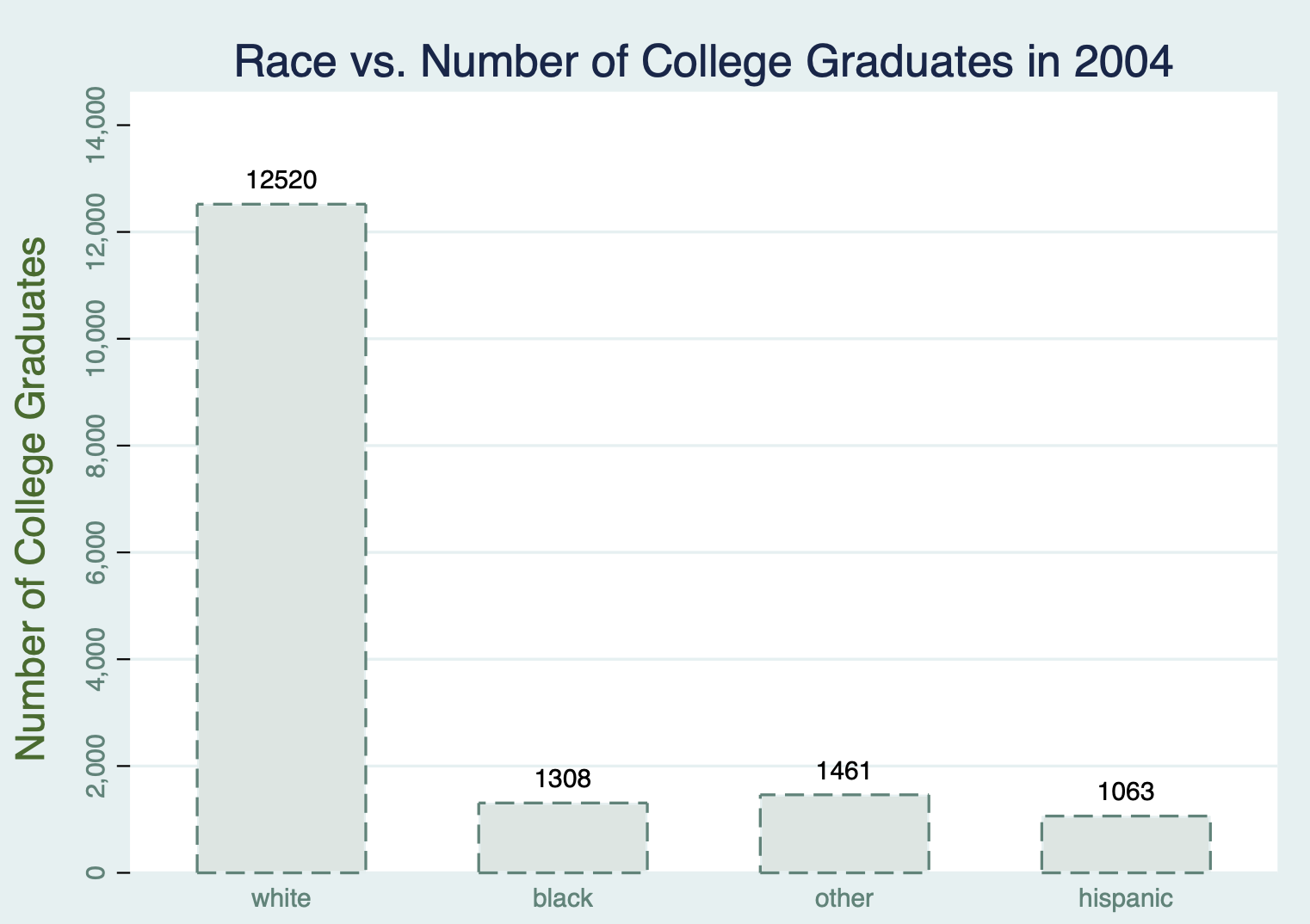
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| (1) family\_income | 1.000 |
| (2) total\_testscore | 0.302 | 1.000 |
| (3) extra-curricular | 0.232 | 0.279 | 1.000 |
| (4) public\_school | -0.207 | -0.166 | -0.172 | 1.000 |
| (5) private\_school | 0.148 | 0.105 | 0.077 | -0.525 | 1.000 |
| (6) white | 0.223 | 0.298 | 0.182 | -0.141 | 0.083 | 1.000 |
| (7) black | -0.145 | -0.234 | -0.096 | 0.082 | -0.024 | -0.407 | 1.000 |
| (8) Hispanic | -0.172 | -0.211 | -0.163 | 0.101 | -0.076 | -0.606 | -0.148 | 1.000 |
| (9) mom married | 0.212 | 0.217 | 0.159 | -0.107 | 0.041 | 0.185 | -0.273 | -0.040 | 1.000 |
| (10) mom\_education | 0.391 | 0.444 | 0.353 | -0.247 | 0.132 | 0.312 | -0.098 | -0.331 | 0.184 | 1.000 |
|  | | | | | | | | | | |

**Correlation between race and other variables that have an association with reading and math scores**

***Findings***

After running the regression and seeing how the variables were associated with the scores it was striking how much larger the coefficients associated with race were compared to other indicators. Therefore, this analysis compared the correlation between race and the other variables. None of the correlations were particularly strong in either direction. What was insightful was that White students who scored higher scores on average compared to non-White students, were positively correlated with the variables that had a positive association with the scores. For example, White students were positively correlated with family income, extracurricular activities and mothers’ education level, variables that Hispanic and Black students shared a negative correlation with.

**College Graduation**

****Tabulation of Race v. College Graduation**

|  |  |  |  |
| --- | --- | --- | --- |
| race | =1 if college grad, =0 if high school | | |
|  | 0 | 1 | Total |
| White | 10536 | 12520 | 23056 |
|  | *45.70* | *54.30* | **100.00** |
| Black | 1955 | 1308 | 3263 |
|  | *59.91* | *40.09* | **100.00** |
| Other | 850 | 1461 | 2311 |
|  | *36.78* | *63.22* | 100.00 |
| Hispanic | 2349 | 1063 | 3412 |
|  | *68.85* | *31.15* | **100.00** |
| Total | 15690 | 16352 | 32042 |
|  | 48.97 | 51.03 | 100.00 |
| First row has *frequencies* and second row has *row percentages* | | | |

***Findings***

We merged this dataset with two datasets showing returns to college and race in 2004 to see if the same trends are apparent. Among White students it is almost even split with 45% having not graduated from college. Among Black students the proportion of non-college graduates is higher at 60%. However, the highest proportion of non-college graduates is within the Hispanic student group with approximately 69% not having a college degree.

**Regression analysis summary on Weekly Earnings**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| weekly\_earn | Coef. | | St. Err. | t-value | | p-value | [95% Conf | | Interval] | Sig |
| white | 64.316 | | 16.478 | 3.90 | | 0.000 | 32.019 | | 96.613 | \*\*\* |
| black | -139.745 | | 20.605 | -6.78 | | 0.000 | -180.132 | | -99.358 | \*\*\* |
| Hispanic | -122.871 | | 20.508 | -5.99 | | 0.000 | -163.067 | | -82.675 | \*\*\* |
| college\_grad | 530.042 | | 8.556 | 61.95 | | 0.000 | 513.272 | | 546.811 | \*\*\* |
| Constant | 643.804 | | 16.598 | 38.79 | | 0.000 | 611.271 | | 676.338 | \*\*\* |
|  | | | | | | | | | | |
| Mean dependent var | | 933.265 | | | SD dependent var | | | 807.126 | |
| R-squared | | 0.127 | | | Number of obs | | | 32042.000 | |
| F-test | | 1160.312 | | | Prob > F | | | 0.000 | |
| Akaike crit. (AIC) | | 515550.181 | | | Bayesian crit. (BIC) | | | 515592.055 | |
|  | | | | | | | | | | |
| *\*\*\* p<0.01, \*\* p<0.05, \* p<0.1* | | | | | | | | | |

**Matrix of correlations**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| (1) college\_grad | 1.000 |
| (2) white | 0.105 | 1.000 |
| (3) black | -0.074 | -0.538 | 1.000 |
| (4) Hispanic | -0.138 | -0.554 | -0.116 | 1.000 |
| (5) parents\_married | 0.083 | 0.149 | -0.152 | -0.061 | 1.000 |
| (6) weekly\_earn | 0.342 | 0.124 | -0.090 | -0.106 | 0.165 | 1.000 |
| (7) hourly\_wage | 0.362 | 0.121 | -0.088 | -0.109 | 0.169 | 0.960 | 1.000 |
|  | | | | | | | |

***Findings***

College graduation rates and race are statistically significant in the regression on Weekly Earnings. Those who were white were associated with earnings that were $64.00 higher on average than non-white individuals. Black and Hispanic individuals earned more than $100.00 less on average than non-Black and non-Hispanic students. When evaluating the correlation between these variables after merging once again Black and Hispanic students were negatively correlated variables that had a positive association with weekly earnings.

**APPENDIX A: CODEBOOK**

|  |  |
| --- | --- |
| Variable Name | Definition |
| year | 2004 |
| age | Age |
| hrsworked\_per\_week | Weeks worked last year |
| weeks\_worked\_2003 | Usual hours worked per week (last year) |
| college\_grad | =1 if college grad, =0 if high school |
| parents\_married | =1 married, = 0 if unmarried |
| weekly\_earn | weekly earnings |
| hourly\_wage | hourly wage |
|  |  |
|  |  |
| race |  |
| white | =1 if ethnicity is white, =0 if otherwise |
| black | =1 if ethnicity is black, =0 if otherwise |
| Hispanic | =1 if ethnicity is Hispanic, =0 if otherwise |
| other | =1 if ethnicity is non-white, non-black and non-Hispanic, =0 if otherwise |
| mom\_college | =1 if mother has a college degree, =0 if otherwise |
| mom\_married\_at\_birth | 1=yes, 0=no |
| family\_income | annual family income |
| mom\_work\_status | 1= mom works >=35 hours/week, =2 is < 35 hours, =3 if not |
| siblings | number of siblings in home |
| hhsize | household size |
| dance | =1 if participate in dance lessons after school, =0 otherwise |
| athletics | =1 if participate in athletics after school, =0 otherwise |
| club | =1 if participate in some club activities after school, =0 otherwise |
| music | =1 if participate in music lessons after school, =0 otherwise |
| art | =1 if participate in art lessons after school, =0 otherwise |
| tv\_afternoon\_mf | hours of tv watching per day after school & before dinner, monday-friday |
| tv\_afterdinner\_mf | hours of tv watching per day after dinner, monday-friday |
| tv\_saturday | hours of tv watching on an average saturday |
| tv\_sunday | hours of tv watching on an average sunday |
| dinner\_as\_family | nights per week eat dinner as a family |
| home\_language\_nonenglish | 1=yes 0=no |
| both\_parents | 1=living w/ both parents, 0=not |
| crowding | =1 if principal reports school has problem with overcrowding, 0 otherwise |
| parents | =1 if principal reports school has problem w/parental involvement, 0 otherwise |
| drugs | =1 if principal reports school has problem with drugs, 0 otherwise |
| gangs | =1 if principal reports school has problem with gangs, 0 otherwise |
| crime | =1 if principal reports school has problem with crime, 0 otherwise |
| weapons | =1 if principal reports school has problem with weapons at school, 0 otherwise |
| has\_library\_card | =1 if kid has library card, =0 otherwise |
| Variable Name | Definition |
| school\_has\_security | =1 if school hires security guards, =0 otherwise |
| Read | reading test score |
| Math | math test score |
| mom\_curr\_married | =1 if mom is currently married, =0 if otherwise |
| family\_type | =1 2par/sibs, =2 2par/no sib, =3 1par/sib, =4 1par/no sib, 5=other |
| total\_testscore | Summation of reading and math score |
| Public\_school | =1 if student attends a public school, =0 if otherwise |
|  |  |
| Private\_school | =1 if student attends a private school, =0 if otherwise |
| Total\_extracurricular. | Sum of number of extracurriculars a student participates in |
|  |  |

**APPENDIX B: CODE**

\*\*\*\*Begin with opening the 2004 5th grade test scores dataset\*\*\*\*

use "/Users/inthisarkamal/Downloads/ecls\_5th\_grade.dta"

drop pct\_minority

drop bmi

drop problem\_turnover

drop problem\_attacks

drop has\_library\_card

ssc install outreg2

ssc install asdoc

gen read = reading\_test

gen math = math\_test

gen crowding = problem\_crowding

gen drugs = problem\_drugs

gen gangs = problem\_gangs

gen crime = problem\_crime

gen weapons = problem\_weapons

drop science\_test

label var read "5th grade Reading Score"

label var math "5th grade Math Score"

gen extracurricular\_total = part\_dance + part\_athletics + part\_club + part\_music + part\_art

su extracurricular\_total

gen any\_extracurricular = .

replace any\_extracurricular = 1 if part\_dance == 1 | part\_athletics == 1 | part\_club == 1 | part\_music == 1 | part\_art == 1

su any\_extracurricular

tab extracurricular\_total

tab any\_extracurricular

gen public\_school =.

replace public\_school = 1 if school\_type ==1

replace public\_school = 0 if school\_type == 2 | school\_type == 3

gen private\_school =.

replace private\_school=1 if school\_type == 3

replace private\_school = 0 if school\_type ==1 | school\_type ==2

gen mommarried = .

replace mommarried = 1 if mom\_curr\_married ==1

replace mommarried = 0 if mom\_curr\_married ==0

gen total\_testscore = read + math

label define race\_vals 1 "White" 2 "Black" 3 "Hispanic" 4 "Other"

label values race race\_vals

gen White = race == 1

gen Black = race == 2

gen Hispanic = race == 3

gen Other = race == 4

label define school\_type\_vals 1 "Public School" 2 "Catholic School" 3 "Private School"

label values school\_type school\_type\_vals

gen mom\_college = mom\_educ == 4

label define mom\_educ\_vals 4 "Mom went to College"

label values mom\_educ mom\_educ\_vals

gen dance = part\_dance

gen athletics = part\_athletics

gen club = part\_club

gen music = part\_music

gen art = part\_art

\*\*\*Running descriptive statistics\*\*\*

asdoc summarize family\_income siblings hhsize read math

tabstat read math, statistics(mean, sd, min max)

tabstat math, stat(mean, sd, min max)

tabstat read math, stat(mean, p25, p50, p75, min max)

asdoc sum math, detail

asdoc sum read, detail

tab race, summarize (math)

tab race, summarize (read)

\*\*\*\*Run summary statistics to determine any relationships between test scores and extra-curricular activities\*\*\*\*

corr total\_testscore extracurricular\_total

corr total\_testscore athletics

corr total\_testscore club

corr total\_testscore dance

corr total\_testscore music

corr total\_testscore art

\*\*\*\*Run summary statistics to determine relationships between income and extra-curricular activities\*\*\*\*

corr family\_income any\_extracurricular

corr family\_income athletics

corr family\_income club

corr family\_income dance

corr family\_income music

corr family\_income art

\*\*\*\*Run summary statistics to determine relationship between family background/school type and math and reading test scores\*\*\*\*

corr total\_testscore race

corr total\_testscore school\_type

corr total\_testscore mom\_educ

asdoc corr family\_income total\_testscore extracurricular\_total public\_school private\_school White Black Hispanic mommarried mom\_educ, replace

\*\*\*\*Run a regression analysis on reading and math test scores and import to word\*\*\*\*

reg read White Black Hispanic extracurricular\_total tv\_\* problem\_\* public\_school private\_school mom\_college mommarried

outreg2 using regression\_results, replace word

reg math White Black Hispanic extracurricular\_total tv\_\* problem\_\* public\_school private\_school mom\_college mommarried

outreg2 using regression\_results, append word

\*\*\*\*Save the new 5th grade dataset\*\*\*\*

save "/Users/inthisarkamal/Downloads/ecls\_5th\_grade.dta", replace

clear

\*\*\*\*Bring in new dataset, "Returns to College (Female)"\*\*\*\*\*

use "/Users/inthisarkamal/Downloads/march\_cps\_females.dta"

sort year

\*\*\*\*Append with dataset "Returns to College (Male)"\*\*\*\*

append using "/Users/inthisarkamal/Downloads/march\_cps\_males.dta"

sort year

keep if year == 2004

\*\*\*\*Clean dataset\*\*\*\*

drop veteran

gen hrsworked\_per\_week = uhrswork

gen weeks\_worked\_2003 = wkswork1

gen race = race4

label define race\_vals 1 "white" 2 "black" 3 "other" 4 "hispanic"

label values race race\_vals

tab race

label define marital\_status\_vals 1 "married" 2 "divorced/separated" 3 "widowed" 4 "unmarried"

label values marital\_status marital\_status\_vals

tab marital\_status

gen parents\_married =.

replace parents\_married = 1 if marital\_status == 1

replace parents\_married = 0 if marital\_status == 2| marital\_status == 4

gen white = race == 1

gen black = race == 2

gen hispanic = race == 4

gen other = race == 3

\*\*\*Create table to tabulate number of college graduates by race in 2004\*\*\*

asdoc tabulate race college\_grad, column

asdoc tabulate race college\_grad, row

\*\*\*Merge with new dataset of 5th grade test scores\*\*\*

merge m:m race using "/Users/inthisarkamal/Downloads/ecls\_5th\_grade.dta"

\*\*\*\*See correlations with college\_grad\*\*\*\*

corr college\_grad white black hispanic parents\_married weekly\_earn hourly\_wage, mom\_college

\*\*\*\*Run regression on Weekly Earnings\*\*\*\*

reg weekly\_earn white black hispanic college\_grad mom\_college

\*\*\*Running a loop\*\*\*

foreach y in "reg" "logit" "probit" {

`y' college\_grad mom\_educ

}

foreach x of varlist part\_\* {

summ `x'

}

\*\*\*\*Determine relationship between graduating from college and 5th grade reading and math scores\*\*\*\*

corr college\_grad math

corr college\_grad read

\*\*\*\*Determine relationship between probability of graduating from college and attending a private school in the fifth grade\*\*\*\*

asdoc corr college\_grad private\_school

\*\*\*Summary statistics with the merged data\*\*\*

summarize hrsworked\_per\_week weekly\_earn family\_income

graph bar (count) college\_grad, over(race)

graph save "Graph" "/Users/inthisarkamal/Downloads/Data Visualization.gph"

asdoc describe, replace

drop race

drop race4

drop part\_\*

drop problem\_\*

save "/Users/inthisarkamal/Downloads/Final final dataset.dta"