MWSUG 2018 - Paper AA-030

Confounded? This example shows how to use SAS® chi-square tests, correlations and logistic regression to unconfound a result.

By Michael C. Grierson, Washington, DC

ABSTRACT

The purpose of this paper is to describe an example of how to unconfound a confounded statistical result[1] and to present a recipe for unconfounding a result. The confounded result is the conclusion that since African American student loan borrowers are more likely to default on their student loans, that "These results show that the U.S. Department of Education cannot ignore the interaction of race and student loans". This paper shows that student loan defaults are more (by about 5 times) associated with lower median income status than race.

INTRODUCTION

The data supporting the conclusion "that the Department of Education cannot ignore the interaction of race and student loans" is presented in a website posting from the Center for American Progress, that concludes that African American borrowers are defaulting on their student loans at a nearly 50% rate, where other borrowers are defaulting at a rate that is just shy of 30%. The data is accurate and the rates are accurate, but the conclusion (African American default at a greater rate) is not. This data and the confounded conclusions were published by Ben Miller of the Center for American Progress[2] and repeated by the Institute of Higher Education in an article by Paul Fain[3] and in an article for Money magazine by Kaitlin Mulhere at the money part of Time.com[4]. These later two articles are basically a repeat of the first article (as they essentially copied and 'reported' Ben Miller's conclusion). The last article quotes the original article as below "These results show that the U.S. Department of Education cannot ignore the interaction of race and student loans".

The fact that people don't choose and can't change their race, makes us suspect that this result may be confounded by another factor. The prime candidate is greater poverty (people without money cannot pay bills) in the African American community.

The recipe for unconfounding is generally to 1.) get more data, 2.) apply valid statistical processes to that data, 3.) adjust your conclusions based on the resulting facts, and 4.) repeat if necessary.

In this paper, I hope to show that the Department of Education **should** ignore "the interaction of race and student loans" as the real factors involved with people paying back student loans is not significantly associated with race, but it is strongly (by more than 5 times) associated with median income.

---

[1] A description and example of confounding here http://www.statisticshowto.com/experimental-design/confounding-variable/

[2] https://www.americanprogress.org/issues/education-postsecondary/news/2017/10/16/440711/new-federal-data-show-student-loan-crisis-african-american-borrowers/. Specific data used in reference 2 is found by using NCES's PowerStat tool in table id cembhag3e.

[3] https://www.insidehighered.com/news/2017/10/17/half-black-student-loan-borrowers-default-new-federal-data-show

[4] http://time.com/money/4986253/race-gap-student-loan-defaults-debt/

METHODOLOGY

The data source used to study this confounded result is from the mappingstudentdebt.org website[5]. From the Methodology section of this website we use the information below to understand where this data came from:

"This geographic analysis uses two primary datasets: credit reporting data on student debt from Experian and income data from the American Community Survey. The Experian data includes eight key student debt variables (see the header of Table 1) aggregated from household-level microdata to the zip code level. The underlying household data are a snapshot of the entire U.S. population at a single point in time—in this case, the autumn of 2015."

By randomly sampling 60 zip codes (suggest using SAS® rand("Uniform") function and the sashelp.zipcode dataset for convenience) from this website, we can assemble a table as in Table 1 below, with the dn (delinquency number) and albn (average loan balance number) fields being the numeric category assigned in footnote 5.

zip	Delinquency	dn	Average_loan_balance	albn	Median_income	Aamerican	Latino
64854	Extremely High	10	Moderately Low	2	33333	4.2	31.5
48843	Low	3	Slightly High	4	67477	0.7	2.6
85743	Moderately Low	4	Slightly High	4	69577	3.7	19.9
4971	Somewhat Low	5	Moderately Low	2	43393	0.2	
78705	Extremely Low	1	Slightly High	4	12143	4.1	17.3
29056	Moderately High	7	Moderately High	6	23023	80.6	0.2
37871	Very High	9	Average	3	46565	2.4	1.3
85338	Moderately Low	4	Slightly High	4	67132	7.3	34
62959	Moderately High	7	Average	3	45947	6.9	2.4
27807	High	8	Average	3	38532	24.7	31.7
46176	Somewhat High	6	Average	3	45812	1.9	5.5
10965	Extremely Low	1	Somewhat High	5	94271	1	4
45365	Moderately High	7	Average	3	45084	3.4	1.9
4747	Moderately High	7	Moderately Low	2	39048	0.1	0.7
41635	Very High	9	Average	3	25620	0.7	1.1
77008	Extremely Low	1	High	7	70293	4.2	33
45107	Somewhat High	6	Moderately Low	2	52143	0.5	0.4
31308	High	8	Moderately Low	2	47315	8	1.7
28441	Extremely High	10	Moderately Low	2	24514	28.3	14.2
96150	Somewhat Low	5	Slightly High	4	46859	0.9	27.6
22903	Very Low	2	Very High	8	47192	17	4.4
92620	Very Low	2	Very High	8	103385	2	7.8
65101	Somewhat Low	5	Average	3	47111	17.6	1.4
24016	Somewhat Low	5	Somewhat High	4	24044	41.9	5.6
4986	Somewhat Low	5	Moderately Low	2	39612	0.1	0.1
25301	Somewhat High	6	Very High	8	25649	10.4	3

[5] http://mappingstudentdebt.org/#/map-2-race

48309	Low	3	Moderately High	6	84118	6	2.8
97041	Moderately Low	4	Moderately Low	2	64677	0.8	28.8
55362	Very Low	2	Average	3	75079	0.5	4.2
79007	High	8	Average	3	42613	4.2	28.6
80120	Low	3	Low	1	54878	1.2	11.8
57064	Extremely Low	1	Average	3	74359	1.2	1
77320	Moderately High	7	Average	3	44130	27	21.5
35233	Very Low	2	Astronomical	10	49423	43.7	
14608	Somewhat High	6	Somewhat High	4	20796	66.5	10.3
93545	Very High	9	Slightly High	4	32473	1.3	32.3
72384	Extremely High	10	Moderately Low	2	23679	42.1	5.5
53930	Low	3	Moderately Low	2	55078	0.3	7.9
27514	Low	3	Very High	8	56333	9	6.9
44230	Somewhat Low	5	Average	3	51944	1.3	2.3
24084	Moderately High	7	Moderately Low	2	47428	4.8	2
24141	Moderately Low	4	Average	3	39808	5.1	2
56277	Very Low	2	Moderately Low	2	50579	2.4	6.2
33161	Moderately High	7	Slightly High	4	33056	64.9	23.6
63755	Moderately Low	4	Average	3	52144	1.7	1.4
96068	Extremely High	10	Low	1	34375	7.7	19.1
45069	Very Low	2	Slightly High	4	84010	6.3	3.7
91604	Very Low	2	Very High	8	88579	4.5	10.1
61064	Low	3	Average	3	51576	0.4	4.8
32332	High	8	Moderately Low	2	24342	80	16.8
87104	Somewhat High	6	Moderately High	6	43379	1.7	57.8
2458	Extremely Low	1	Extremely High	9	95216	3.7	5.8
53714	Low	3	Average	3	49371	9.9	9.8
36036	Very High	9	Average	3	43977	50.6	
26175	Moderately High	7	Moderately Low	2	42539	0.2	1.2
95207	Very High	9	Average	3	39301	12.2	35.6
16025	Very Low	2	Average	3	58679	0.4	0.2
18017	Low	3	Slightly High	4	62882	5.5	17.3
92648	Moderately Low	4	Moderately High	6	82567	0.7	17.2
48226	Low	3	Extremely High	9	30891	61.2	4.4

Table 1. 60 random samples from the mappingstudentdebt.org website.

This website provides a convenient source of data for median income, race, and student loan default rates by zip code. It's a good source using good methodology (combining default rates from Experian with American Community Survey data and connecting the two by zip code). You are encouraged to create your own random sample from footnote 5.

From this table we can analyze the association between a zip code area's student loan default rates and percentage African American population attributes and compare that to the association between a zip

code area's student loan default rates and levels of median income. This will provide us a data basis for determining which factor is more associated with student loan debt.

SAS® software provides some very convenient procedures for working with the data in table 1. The proc freq procedure provides for chi-square testing, proc corr procedure provides four types of correlation results (we'll look at two here), the proc rank procedure provides a convenient way to rank data into categories (for chi-square tests, spearman correlations, and logistics regressions). We will also use proc logistic (for regression analysis) to put the associations into a context where we can quantify the associations by probability. Proc univariate is also used to describe our sample (the 60 samples shown in table 1).

READ IN THE DATA AND PREPARE THE DATASET FOR ANALYSIS

The data can be read in from a comma separated values (csv) file using the code below in Figure 1.

```
FILENAME REFFILE '/folders/myfolders/zipcodes.csv';
/* import the csv into a dataset */
PROC IMPORT DATAFILE=REFFILE
      DBMS=CSV REPLACE
      OUT=WORK.sd1;
      GETNAMES=YES;
      DATAROW=2;
      GUESSINGROWS=100;
RUN;
/* rank into two groups for chisq and logistic regression calculations */
proc rank groups=2 data=work.sd1 out=work.sd2 ties=low;
  var median_income aamerican latino dn;
  ranks rank_median_income2 rank_aamerican2 rank_latino2 rank_dn2 ;
run;
/* rank into ten groups for spearman correlations */
proc rank groups=10 data=work.sd2 out=work.sd3 ties=low;
  var median_income aamerican latino;
  ranks rank_median_income10 rank_aamerican10 rank_latino10;
run;
/* add formats so that the rankings show as descriptive */
proc format;
  value rank_aamerican2_fmt 0 = "low % African American"
                            1 = "high % African American";
  value rank_median_income2_fmt 0 = "low median income"
                                1 = "high median income";
  value rank_latino2_fmt 0 = "low % latino"
                         1 = "high % latino";
  value rank_dn2_fmt 0 = "low student loan default"
                     1 = "high student loan default";
run;
/* add labels and bump the ranks by 1 */
data work.sd;
  set work.sd3;
  format rank_aamerican2 rank_aamerican2_fmt.
         rank_median_income2 rank_median_income2_fmt.
         rank_latino2 rank_latino2_fmt.
         rank_dn2 rank_dn2_fmt.;
  label dn="Delinquency Category"
        delinquency="Delinquency text"
        albn = "Average Loan Balance category"
        average_loan_balance="Average Loan Balance text"
```

```
        aamerican = "% African American"
        latino="% Latino"
        median_income = "Median Income"
        zip="Zipcode common to Experian and ACS data"
        rank_aamerican10 = "% African American 10 Categories"
        rank_median_income10 = "Median Income 10 Categories"
        rank_latino10 = "% Latino 10 Categories"
        rank_aamerican2 = "African American 2 Categories"
        rank_median_income2 = "Median Income 2 Categories"
        rank_latino2 = "Latino 2 Categories"
        rank_dn2 = "Delinquency 2 Categories";
run;
proc contents data=work.sd order=varnum;
run;
```

Figure 1. SAS code for reading in the zipcodes.csv data

The resulting SAS dataset has columns of interest (median_income, aamerican and latino percentage) in $ values and percentages form. These columns are also ordinally ranked in 2 categories (for chisq analysis) and ordinally ranked in 10 categories (for spearman correlations). The resulting dataset proc contents output is shown in Figure 2 below.

Variables in Creation Order

#	Variable	Type	Len	Format	Informat	Label
1	zip	Num	8	BEST12.	BEST32.	Zipcode common to Experian and ACS data
2	delinquency	Char	15	$15.	$15.	Delinquency text
3	dn	Num	8	BEST12.	BEST32.	Delinquency Category
4	Average_loan_balance	Char	15	$15.	$15.	Average Loan Balance text
5	albn	Num	8	BEST12.	BEST32.	Average Loan Balance category
6	Median_income	Num	8	BEST12.	BEST32.	Median Income
7	Aamerican	Num	8	BEST12.	BEST32.	% African American
8	Latino	Num	8	BEST12.	BEST32.	% Latino
9	rank_median_income2	Num	8	RANK_MEDIAN_INCOME2_FMT.		Median Income 2 Categories
10	rank_aamerican2	Num	8	RANK_AAMERICAN2_FMT.		African American 2 Categories
11	rank_latino2	Num	8	RANK_LATINO2_FMT.		Latino 2 Categories
12	rank_dn2	Num	8	RANK_DN2_FMT.		Delinquency 2 Categories
13	rank_median_income10	Num	8			Median Income 10 Categories
14	rank_aamerican10	Num	8			% African American 10 Categories
15	rank_latino10	Num	8			% Latino 10 Categories

Figure 2. proc contents output

EXAMINE THE DATASET

Next, we test the dataset by running a proc univariate to get descriptive statistics for the sample of 60 shown in Table 1. Remember that we are testing data from 60 random zip codes. This code for the descriptive statistics is shown below in Figure 3.

```
proc univariate data=work.sd;
   var dn median_income aamerican;
run;
```

Figure 3. proc univariate for variables of interest

The output from the proc univariate is partially shown below in Figure 4. This partially validates our sample as we can see that the samples mean median_income variable is about $50K and the samples mean percentage of African Americans is 13.4%. Both numbers are reasonable (within expected ranges).

Variable: median_income (Median Income)

Moments			
N	60	Sum Weights	60
Mean	50155.8833	Sum Observations	3009353
Std Deviation	20258.8013	Variance	410419028
Skewness	0.68105405	Kurtosis	0.14196054
Uncorrected SS	1.75151E11	Corrected SS	2.42147E10
Coeff Variation	40.3916747	Std Error Mean	2615.4

Variable: aamerican (% African American)

Moments			
N	60	Sum Weights	60
Mean	13.3633333	Sum Observations	801.8
Std Deviation	21.1044052	Variance	445.395921
Skewness	2.00025002	Kurtosis	3.01953587
Uncorrected SS	36993.08	Corrected SS	26278.3593
Coeff Variation	157.927702	Std Error Mean	2.724567

Figure 4. proc univariate output for variables of interest

CHI-SQUARE TEST FOR INDEPENDENCE

Now that we have a dataset that has categories via the rank procedure, lets run some tests of association. A chi-square test for independence is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables[6]. In our case, we have two pairs of variables to test for independence. The pair one test is to see if the categorized delinquency rate (rank_dn2) is independent of the categorized percentage of African Americans (ranked_aamerican2). The pair two test is to test if the categorized delinquency rate (ranked_dn2) is independent of the categorized median income (ranked_median_income2). The SAS code for the chi–square tests is shown below in Figure 5.

```
/* H0 (Null): African American NOT→ Student loan default. */
title "Chisq output for Student Loan default by African American group";
proc freq data=work.sd;
      tables rank_aamerican2*rank_dn2 / cmh chisq expected norow nocol
             nopercent;
run;

/* H0: Median Income NOT → student loan defaults. */
title "Chisq output for Student Loan default group by median income group";
proc freq data=work.sd;
      tables rank_median_income2*rank_dn2 / cmh chisq expected norow nocol
             nopercent;
run;
```

Figure 5. Separate Chi -square SAS code for both proposed independent variables

[6] From website http://stattrek.com/chi-square-test/independence.aspx?Tutorial=AP

The output from these two tests are shown below in Figures 6 and 7 below. Note that the Figure 7 output shows a high Chi-Square value to a high degree of confidence (p-value (Prob.) of < 0.0001) which means that Student Loan Delinquency is dependent on the median income. Figure 6 shows that Student Loan Delinquency is independent of the % African American in these zip codes. This is strong proof that the Department of Education was correct in ignoring "the interaction between race and student loans". Note that the dependent case (in Figure 7) has a contingency table that looks skewed in comparison to the expected.

Chisq output for Student Loan default by African American group

The FREQ Procedure

Frequency
Expected

Table of rank_dn2 by rank_aamerican2			
	rank_aamerican2(African American 2 Categories)		
rank_dn2(Delinquency 2 Categories)	low % African American	high % African American	Total
low student loan default	21 18.133	13 15.867	34
high student loan default	11 13.867	15 12.133	26
Total	32	28	60

Statistics for Table of rank_dn2 by rank_aamerican2

Statistic	DF	Value	Prob
Chi-Square	1	2.2410	0.1344
Likelihood Ratio Chi-Square	1	2.2511	0.1335
Continuity Adj. Chi-Square	1	1.5275	0.2165
Mantel-Haenszel Chi-Square	1	2.2037	0.1377
Phi Coefficient		0.1933	
Contingency Coefficient		0.1898	
Cramer's V		0.1933	

Figure 6. Chi-square output for Student Loan Delinquency by % African American categories

Chisq output for Student Loan default group by median income group

The FREQ Procedure

Frequency
Expected

Table of rank_dn2 by rank_median_income2			
	rank_median_income2(Median Income 2 Categories)		
rank_dn2(Delinquency 2 Categories)	low median income	high median income	Total
low student loan default	7 17	27 17	34
high student loan default	23 13	3 13	26
Total	30	30	60

Statistics for Table of rank_dn2 by rank_median_income2

Statistic	DF	Value	Prob
Chi-Square	1	27.1493	<.0001
Likelihood Ratio Chi-Square	1	30.0065	<.0001
Continuity Adj. Chi-Square	1	24.5023	<.0001
Mantel-Haenszel Chi-Square	1	26.6968	<.0001
Phi Coefficient		-0.6727	
Contingency Coefficient		0.5581	
Cramer's V		-0.6727	

Figure 7. Chi-square output for Student Loan Delinquency by Median Income categories

From the figures above, it is clear that the effect of median income on Student loan default rates is more significant than the effect of % African American, but we can't see how both factors together affect Student loan default rates. The Chi-square test below, shown in Figure 8, will show the interaction of both factors on Student Loan Default rates.

```
title "Chisq output for Student Loan default by Median Income group by %
African American categories";
proc freq data=work.sd;
      tables rank_aamerican2*rank_median_income2*rank_dn2 / cmh chisq
            expected norow nocol nopercent;
run;
```

Figure 8. Combined Chi-square SAS code for both proposed independent variables

Parts of the combined Chi-square outputs are shown below in Figure 9. Note that this is essentially a Chi-square of median income categories by delinquency rates run of the 32 samples with from the low % African American records in the top section of figure 9 and the same run for the 28 samples from the high % African American records. One would expect the high Chi-square result for median income by delinquency rate to be distributed to the low % African American category by 32/60 (or 53%) and the high % African American category would be distributed to 28/60 (46%). The expected percentages are almost matched by the category results 15/27 (or 55%) and 11/27 (or 42%). The Chi-square score of 27 in Figure 7 is distributed nearly as expected in the % African American categories. This result again confirms that the Department of Education was correct in ignoring "the interaction between race and student loans".

Chisq output for Student Loan default by Median Income group by % African American categories

The FREQ Procedure

Frequency
Expected

Table 1 of rank_median_income2 by rank_dn2

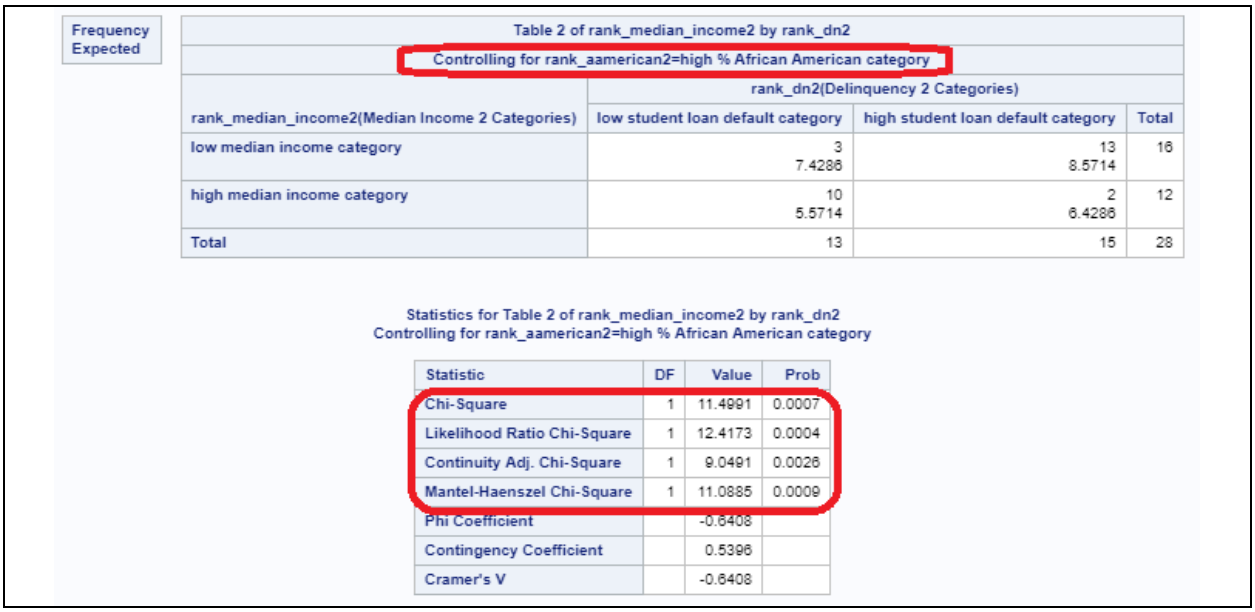
Controlling for rank_aamerican2=low % African American category

rank_median_income2(Median Income 2 Categories)	rank_dn2(Delinquency 2 Categories)		
	low student loan default category	high student loan default category	Total
low median income category	4 9.1875	10 4.8125	14
high median income category	17 11.813	1 6.1875	18
Total	21	11	32

Statistics for Table 1 of rank_median_income2 by rank_dn2
Controlling for rank_aamerican2=low % African American category

Statistic	DF	Value	Prob
Chi-Square	1	15.1479	<.0001
Likelihood Ratio Chi-Square	1	16.7078	<.0001
Continuity Adj. Chi-Square	1	12.3686	0.0004
Mantel-Haenszel Chi-Square	1	14.6746	0.0001
Phi Coefficient		-0.6880	
Contingency Coefficient		0.5668	
Cramer's V		-0.6880	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.



Frequency
Expected

Table 2 of rank_median_income2 by rank_dn2
Controlling for rank_aamerican2=high % African American category

rank_median_income2(Median Income 2 Categories)	rank_dn2(Delinquency 2 Categories)		
	low student loan default category	high student loan default category	Total
low median income category	3 7.4286	13 8.5714	16
high median income category	10 5.5714	2 6.4286	12
Total	13	15	28

Statistics for Table 2 of rank_median_income2 by rank_dn2
Controlling for rank_aamerican2=high % African American category

Statistic	DF	Value	Prob
Chi-Square	1	11.4991	0.0007
Likelihood Ratio Chi-Square	1	12.4173	0.0004
Continuity Adj. Chi-Square	1	9.0491	0.0026
Mantel-Haenszel Chi-Square	1	11.0885	0.0009
Phi Coefficient		-0.6408	
Contingency Coefficient		0.5396	
Cramer's V		-0.6408	

Figure 9. Chi-square output for Student Loan Delinquency by Median Income categories stratified by % African American

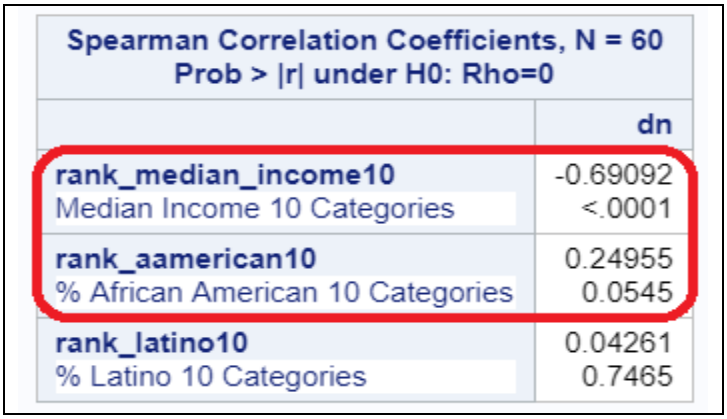
CORRELATIONS

Correlations can be used to verify and additionally to quantify if there is association between student loan default rates and race or median income. Because we have ranked and ordered variables, we'll use Spearman correlations. SAS code for testing correlations is shown below in Figure 8

```
/* spearman correlations between delinquency categories, median_income
   and race */
proc corr data=work.sd spearman plots=scatter;
      var dn;
      with rank_median_income10 rank_aamerican10 rank_latino10;
run;
```

Figure 10. Code for Delinquency Categories vs. Median Income and % African American Categories

The results for the code above are shown below in figure 11.



Spearman Correlation Coefficients, N = 60
Prob > |r| under H0: Rho=0

	dn
rank_median_income10 Median Income 10 Categories	-0.69092 <.0001
rank_aamerican10 % African American 10 Categories	0.24955 0.0545
rank_latino10 % Latino 10 Categories	0.04261 0.7465

Figure 11. Output from the code in Figure 10.

These results show a strong correlation between median income and student loan default rates and a low (and suspect due to the p-value) correlation between being African American and student loan default rates. This is not consistent with the conclusions of the Center for American Progress paper cited in the introduction. This is additional strong proof that the Department of Education was correct in ignoring "the interaction between race and student loans" as recommended by the Center for American Progress.

The low correlation between default rates and being African American is in part due to the correlation between being black and having a lower median income. We can test for a correlation by the same methods using the code below in Figure 12. Note, here we use Pearson correlations (we use continuous variables).

```
/* check correlations between median_income and African Americans */
proc corr data=work.sd spearman plots=scatter;
      var rank_median_income10;
      with rank_aamerican10;
run;
```

Figure 12. proc corr for median_income vs. race

Spearman Correlation Coefficients, N = 60 Prob > |r| under H0: Rho=0	
	rank_median_income10
rank_aamerican10 % African American 10 Categories	-0.38733 0.0022

Figure 13. proc corr output from the code in Figure 10.

The output in Figure 13, shows a medium correlation between median income and % African American, which would explain the low correlation result earlier between student loan default rates and being African American. It further validates that median income has a far greater effect on student loan default rates than being African American. Also, note that if you take the r^2 (R-square) values of the correlations show in Figure 11, you'll get about 0.4039 for median income and about .0731 for being African American. These r^2 values loosely mean that we can model the association with equations that account for 40.49% of the variation for median income, and only about 7.31% of the variation for being African American. If the linear assumptions are valid, this means that the association of median income to student default rates is more than 5 time greater than the association to being African American. This is further evidence that the Department of Education should ignore the interaction of race and student loans and provides a first quantification of the magnitude of the difference between these associations. A better test (does not involve linearity assumptions (which are shaky in this case)) would be logistic regression using the two rankings used in the previous chi-square tests.

QUANTIFY VIA LOGISTIC REGRESSION

To further quantify the relative magnitude of the association between these two pairs (1.) %AfricanAmerican /StudentLoanDefaultRates and 2.) MedianIncome /StudentLoanDefaultRates), the SAS code to run simple logistic regressions is shown below in Figure 14:

```
proc logistic data=work.sd plots(only)=effect;
      class rank_median_income2;
      model rank_dn2=rank_aamerican2 rank_median_income2;
quit;
```

Figure 14. proc logistic for Loan Default Category = %African_American Media_Income

The results of the logistic procedure in Figure 14 are shown in Figure 15 below. The effects plot output shows the probabilities by category for a high student loan default rate (rank_dn2=high student loan default category).

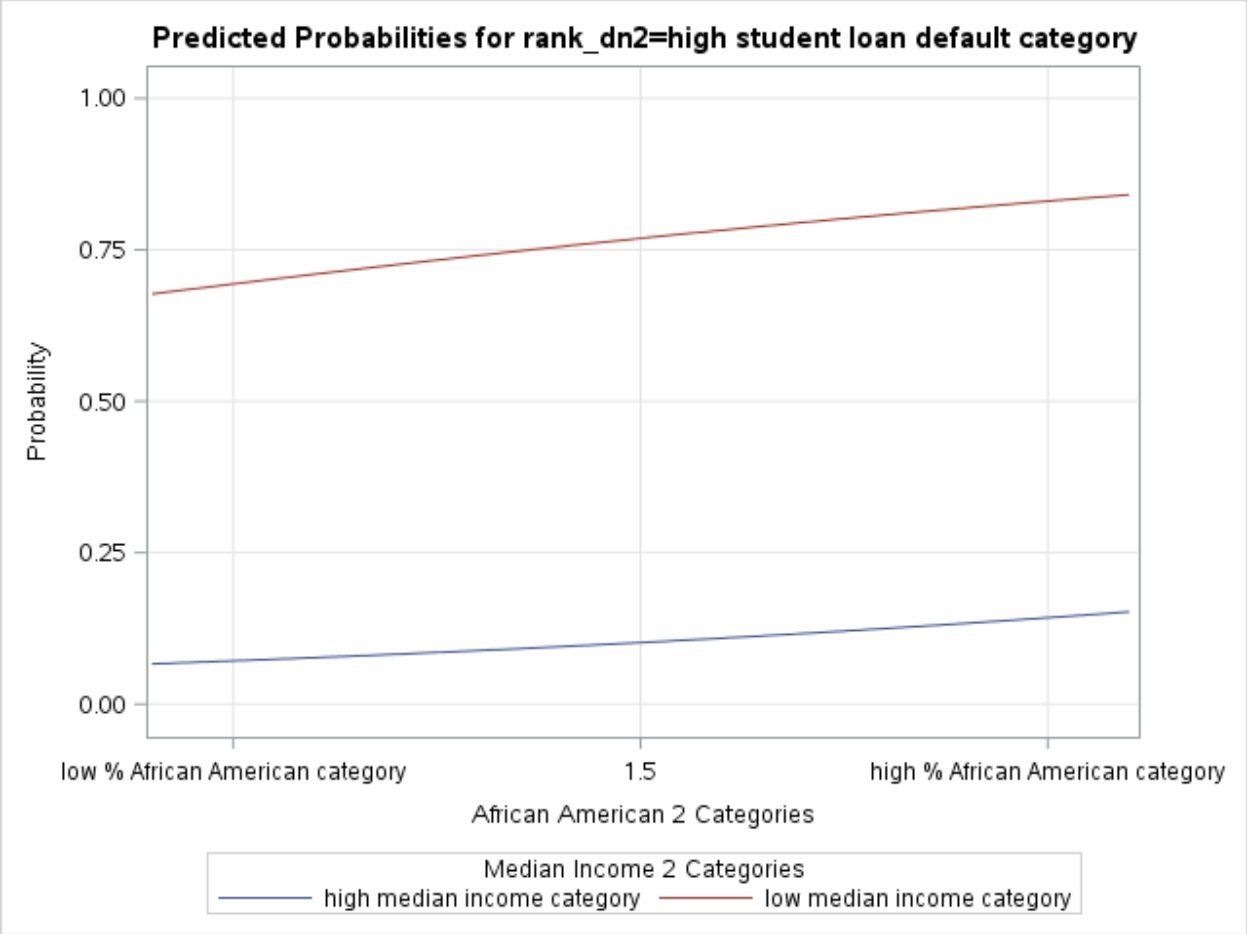


Figure 15. proc logistic effects plot output from the code in Figure 14.

Figure 15 shows that the probability of high student loan default rates is low for borrowers in the zip code areas with a high median income. The figure also shows that the probability of high student loan default rates is high for borrowers in zip codes with low median income.

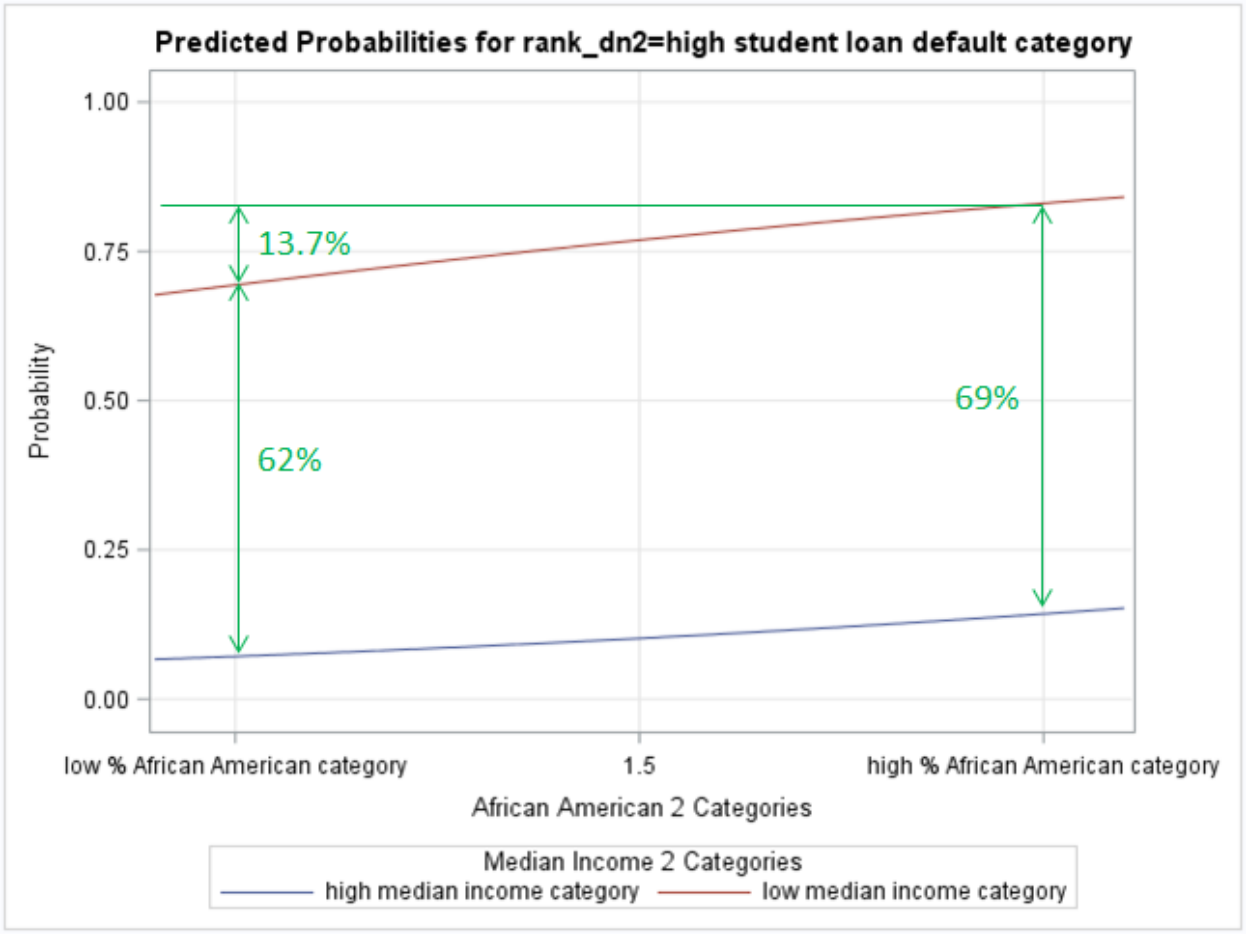


Figure 16. proc logistic output highlighting (in green) quantified differences.

Figure 16 is the same output that shows quantified differences between the circumstances we are comparing. The probability between median income categories is more than 62% whereas the probability differences between the % African American categories is less than 13.7%. Clearly the median income category is the more significant factor in student loan defaults (in this case about 5 times larger probability).

CONCLUSION

The recommendation about "the interaction of race and student loans" ("These results show that the U.S. Department of Education cannot ignore the interaction of race and student loans".) should be ignored because the effect of the borrower's median income category confounds what appears to be the effect of the borrower's race on student loan default rates.

More generally, when you suspect a confounded result, the process of unconfounding that result begins with getting more data, applying statistical processes (SAS procedures make this easy), arriving at potentially adjusted conclusions. Then repeat that process until you understand the conclusions as reasonable and you understand the 'mechanism' of the result (in this case, people without money have a more difficult time paying bills).

Lastly, the opinions and conclusions expressed in this paper are mine and are not necessarily those of my employer

REFERENCES

1. The web article titled “Confounding Variable: Simple Definition and Example” http://www.statisticshowto.com/experimental-design/confounding-variable/
2. https://www.americanprogress.org/issues/education-postsecondary/news/2017/10/16/440711/new-federal-data-show-student-loan-crisis-african-american-borrowers/. Specific data used in reference 2 is found by using NCES’s PowerStat tool in table id cembhag3e.
3. https://www.insidehighered.com/news/2017/10/17/half-black-student-loan-borrowers-default-new-federal-data-show
4. http://time.com/money/4986253/race-gap-student-loan-defaults-debt/
5. The web site http://mappingstudentdebt.org/#/map-2-race uses Experian data and the American Community Survey to map student debt to zip code areas.
6. From website http://stattrek.com/chi-square-test/independence.aspx?Tutorial=AP

RECOMMENDED READING

- The SAS online class (EXPMLR41) titled: Predictive Modeling Using Logistic Regression (v14.2) available via https://vle.sas.com and a SAS® Learning Subscription.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Michael C. Grierson
mgrierson@gmail.com