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

MIKE GRIERSON IS A FEDERAL EMPLOYEE CURRENTLY WORKING ON SAS, DATA WAREHOUSE, AND PROCESS AUTOMATION ISSUES. HE RECEIVED A BSEE DEGREE FROM THE UNIVERSITY OF CENTRAL FLORIDA AND AN MSSE DEGREE (WITH AN OPERATIONS RESEARCH CONCENTRATION) FROM THE GEORGE WASHINGTON UNIVERSITY. AFTER WORKING FOR MANY YEARS ON INFORMATION TECHNOLOGY SYSTEMS, THE OPERATIONS RESEARCH STUDY PROVIDED INSPIRATION FOR HIS FEDERAL SERVICE WORK AS AN OPERATIONS RESEARCHER AND BUSINESS ANALYST (FOR THE DEPARTMENT OF THE ARMY, THE COAST GUARD, THE DEPARTMENT OF VETERANS AFFAIRS AND NOW THE DEPARTMENT OF EDUCATION). HE CURRENTLY LIVES IN THE WASHINGTON DC AREA AND CAN BE CONTACTED AT MCG@MCG-CT.COM.

THIS PRESENTATION IS MIKE'S OWN WORK AND DOES NOT REPRESENT OPINIONS/CONCLUSIONS OF HIS EMPLOYER.

Overview

Confounding
➢ Example article

Methodology:
➢ Get more data (that Includes suspected confounding attributes) and rank that data
➢ Apply Chi Square tests, Correlations, and Logistic Regression

Why this matters
➢ Thinking Fast (Intuitively/Instinctually) vs. Slow (Analytically/Rationally)
➢ Evidence based analytic conclusions

CONFOUNDING

Center for American Progress

New Federal Data Show a Student Loan Crisis for African American Borrowers

References Department of Education data resources

Example article

12/18/2017   New Federal Data Show a Student Loan Crisis for African American Borrowers – Center for American Progress

Center for American Progress

≡

EDUCATION, POSTSECONDARY

New Federal Data Show a Student Loan Crisis for African American Borrowers

By Ben Miller | Posted on October 16, 2017, 9:00 am

Two weeks ago, the U.S. Department of Education provided the first-ever look at long-term outcomes for student loan borrowers, including results by race and ethnicity.

https://www.americanprogress.org/issues/education-postsecondary/news/2017/10/16/440711/new-federal-data-show-student-loan-crisis-african-americ…   1/6

The data show that 12 years after entering college, the typical African American* student who started in the 2003-04 school year and took on debt for their undergraduate education owed more on their federal student loans than they originally borrowed. This holds true even for students who finished a bachelor's degree at a public institution. One reason they might not be paying down their loans? Nearly half of African American borrowers defaulted, including 75 percent of those who dropped out of for-profit colleges.

These results show that the U.S. Department of Education cannot ignore the interaction of race and student loans. Traditionally, the agency has not collected any data on the race of borrowers, except in irregular sample surveys conducted by its quasi-independent statistical arm. Unfortunately, not collecting this information has allowed for the disparate outcomes by race to go unnoticed.

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.

Example follow-on articles

MWSUG2018 AA-030

NCES article and PowerStats reference (cembhag3e)

Repayment of Student Loans as of 2015 Among 1995–96 and 2003–04 First-Time Beginning Students

First Look

← This reference does not mention race. It does mention <u>median income</u> and <u>profit status of college attended</u>.

National Center for Education Statistics PowerStats

Ever defaulted on a federal loan - 12 years by Race/ethnicity, First institution sector (level and control) 2003-04, Cumulative persistence and attainment anywhere 2008-09, Ever received a Pell Grant - 12 years and Income group in 2003-04, for Cumulative amount of undergraduate federal loans - 12 years (X >= 1).

	No (%)	Yes (%)	Total
Estimates			
Total	71.2	28.8	100%
Race/ethnicity			
White	78.5	21.5	100%
Black or African American	50.9	49.1	100%
Hispanic or Latino	63.9	36.1	100%
Asian	87.7	12.3	100%
American Indian or Alaska Native	58.8	41.2	100%
Native Hawaiian / other Pacific Islander	84.5	15.5 !!	100%
Other	74.7	25.3	100%
More than one race	60.3	39.7	100%

PowerStats data is accurately reported.

NCES article and PowerStats reference (cembhag3e)

PowerStats data is accurately reported.

National Center for Education Statistics PowerStats

Ever defaulted on a federal loan - 12 years by Race/ethnicity, First institution sector (level and control) 2003-04, Cumulative persistence and attainment anywhere 2008-09, Ever received a Pell Grant - 12 years and Income group in 2003-04, for Cumulative amount of undergraduate federal loans - 12 years (X >= 1).

	No (%)	Yes (%)	Total
Estimates			
Total	71.2	28.8	100%
Race/ethnicity			
White	78.5	21.5	100%
Black or African American	50.9	49.1	100%
Hispanic or Latino	63.9	36.1	100%
Asian	87.7	12.3	100%
American Indian or Alaska Native	58.8	41.2	100%
Native Hawaiian / other Pacific Islander	84.5	15.5 !!	100%
Other	74.7	25.3	100%
More than one race	60.3	39.7	100%
First institution sector (level and control) 2003-04			
Public 4-year	81.0	19.0	100%
Private not-for-profit 4-year	81.0	19.0	100%
Public 2-year	73.6	26.4	100%
Private For Profit	46.5	53.5	100%
Other	68.0	32.0	100%
Cumulative persistence and attainment anywhere 2008-09			
Attained bachelor's degree	91.2	8.8	100%
Attained associate's degree	78.5	21.5	100%
Attained certificate	53.7	46.3	100%
No degree, still enrolled	70.7	29.3	100%
No degree, left without return	54.2	45.8	100%
Ever received a Pell Grant - 12 years			
No	87.2	12.8	100%
Yes	64.5	35.5	100%
Income group in 2003-04			
Low	57.5	42.5	100%
Low middle	70.7	29.3	100%
High middle	81.0	19.0	100%
High	86.9	13.1	100%

!! Interpret data with caution. Estimate is unstable because the standard error represents more than 50 percent of the estimate.

The names of the variables used in this table are: FSECTOR, PROUT6, S15FEDCUM1_12Y, RACE, S15EVRDEF_12Y, INCGRP and S15PELLEVR_12Y.

The weight variable used in this table is WTB000.

Source: U.S. Department of Education, National Center for Education Statistics, 2003-04 Beginning Postsecondary Students Longitudinal Study, Second Follow-up (BPS:04/09).

Computation by NCES PowerStats on 7/30/2018.

Confounding

Center for American Progress article asserts that race and student loan defaults are associated.

➢ "These results show that the U.S. Department of Education cannot ignore the interaction of race and student loans".

Suspect this result is confounded by median income, because the mechanism of not having money and defaulting on loans is well understood, whereas the mechanism of being a member of a racial group and defaulting on loans is not well understood.

Other examples of possible confounding?
   ➢You will live longer if you drink 4 to 8 cups of coffee a day? (Well off people can afford all that coffee?)
   ➢People with graduate degrees have 11% more brain gliomas? (Graduate degree holders live in cities?)

METHODOLOGY

Find Student Debt Data that Includes Confounders and Apply

Chi Square,

Spearman Correlations and

Logistic Regression

Relevant Data mappingstudentdebt.org

DATA ORIGINS

"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 Figure 2 (@ reference 4)) 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."

SUMMARIZED

At the Zip Code level
➢Experian for student debt data
➢American Community Survey for income data

As of 2015

https://mappingstudentdebt.org/
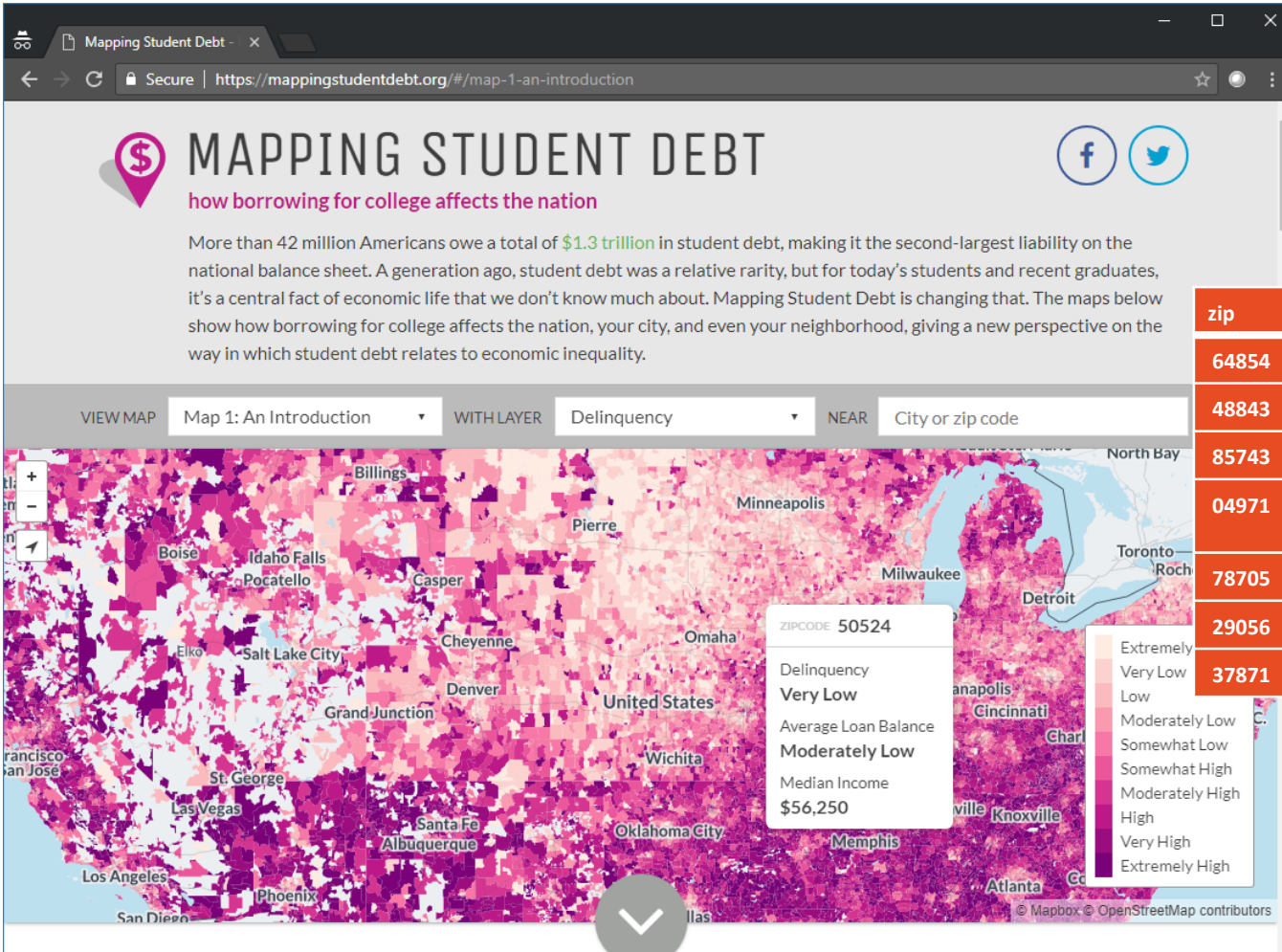
Take a sample of 60 randomly selected zip codes and build data table below

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
04971	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

Test the data sample

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

Variable: dn (Delinquency Category)

Moments			
N	60	Sum Weights	60
Mean	5.08333333	Sum Observations	305
Std Deviation	2.75121955	Variance	7.56920904
Skewness	0.22844224	Kurtosis	-1.1262885
Uncorrected SS	1997	Corrected SS	446.583333
Coeff Variation	54.1223519	Std Error Mean	0.35518092

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

Rank the data

```
/* 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;
```
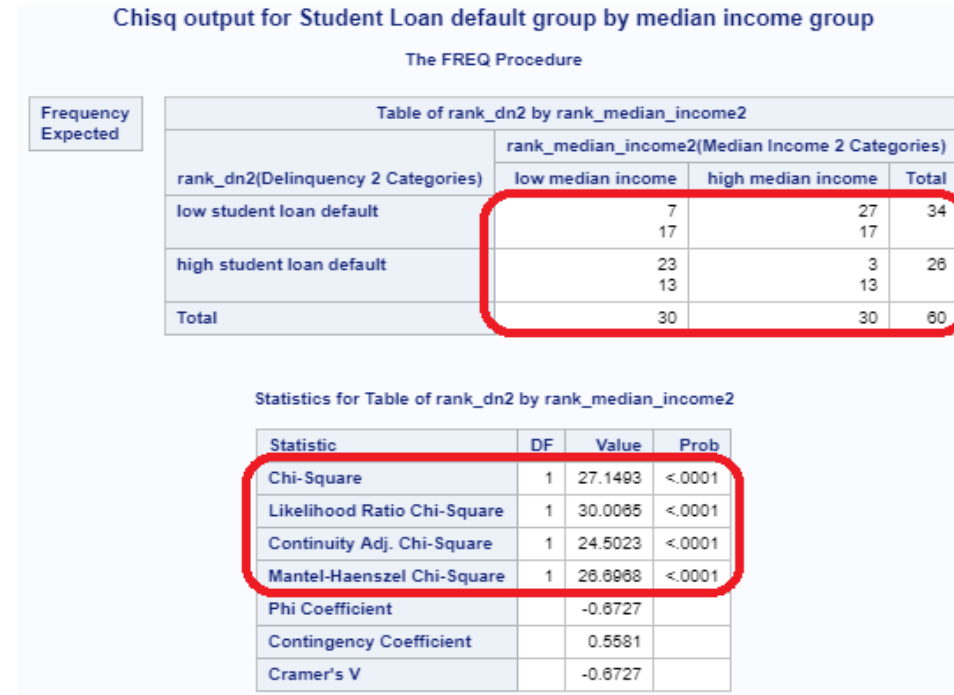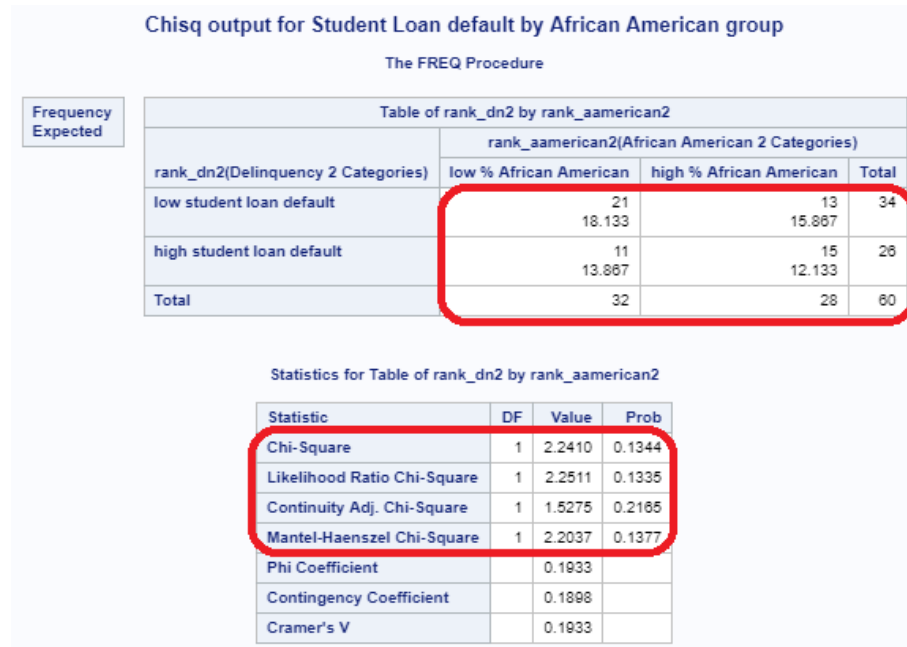
zip	delinquency	dn	Average_loan_balance	albn	Median_income	Aamerican	Latino	rank_median_income2	rank_aamerican2	rank_latino2	rank_dn2	rank_median_income10	rank_aamerican10	rank_latino10
64854	Extremely High	10	Moderately Low	2	33333	4.2	31.5	low median income category	low % African American category	high % latino category	high student loan default category	2	4	8
48843	Low	3	Slightly High	4	67477	0.7	2.6	high median income category	low % African American category	low % latino category	low student loan default category	8	1	3
85743	Moderately Low	4	Slightly High	4	69577	3.7	19.9	high median income category	low % African American category	high % latino category	low student loan default category	8	4	7
4971	Somewhat Low	5	Moderately Low	2	43393	0.2		low median income category	low % African American category		low student loan default category	3	0	
78705	Extremely Low	1	Slightly High	4	12143	4.1	17.3	low median income category	low % African American category	high % latino category	low student loan default category	0	4	7
29056	Moderately High	7	Moderately High	6	23023	80.6	0.2	low median income category	high % African American category	low % latino category	high student loan default category	0	9	0
37871	Very High	9	Average	3	46565	2.4	1.3	low median income category	low % African American category	low % latino category	high student loan default category	4	3	1
85338	Moderately Low	4	Slightly High	4	67132	7.3	34	high median income category	high % African American category	high % latino category	low student loan default category	7	6	9
62959	Moderately High	7	Average	3	45947	6.9	2.4	low median income category	high % African American category	low % latino category	high student loan default category	4	6	2

Chi-Square

```sas
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;


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;
```

Chi-Square

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	

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	

These results show that the U.S. Department of Education SHOULD ignore the interaction of race and student loans because chi-square results show median income levels are far more associated with student loan defaults

Spearman Correlation

```sas
/* 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;
```

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

from http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

.00-.19 "very weak"
.20-.39 "weak"
.40-.59 "moderate"
.60-.79 "strong"
.80-1.0 "very strong"

These results show that the U.S. Department of Education SHOULD ignore the interaction of race and student loans because spearman correlation results show median income levels are far more associated with student loan defaults

Spearman Correlation
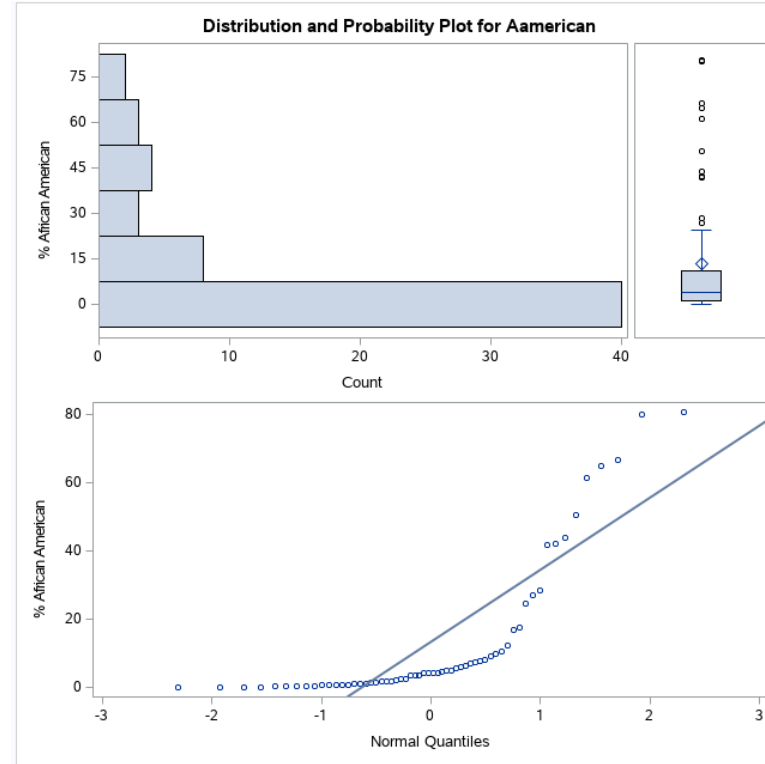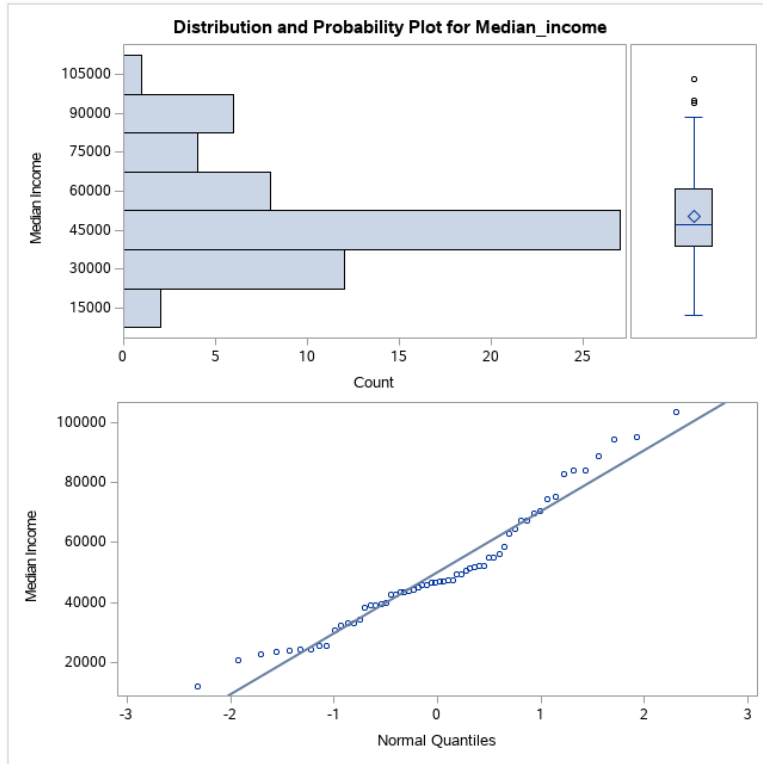
Linearity assumptions for percent African American are suspect

Percent African American data not particularly normally distributed

Logistic Regression

Use Logistic regression effects plots to get some sense of the magnitude of the relative associations (not provided by the earlier tests and not provided by Odds Ratios)

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

Logistic Regression
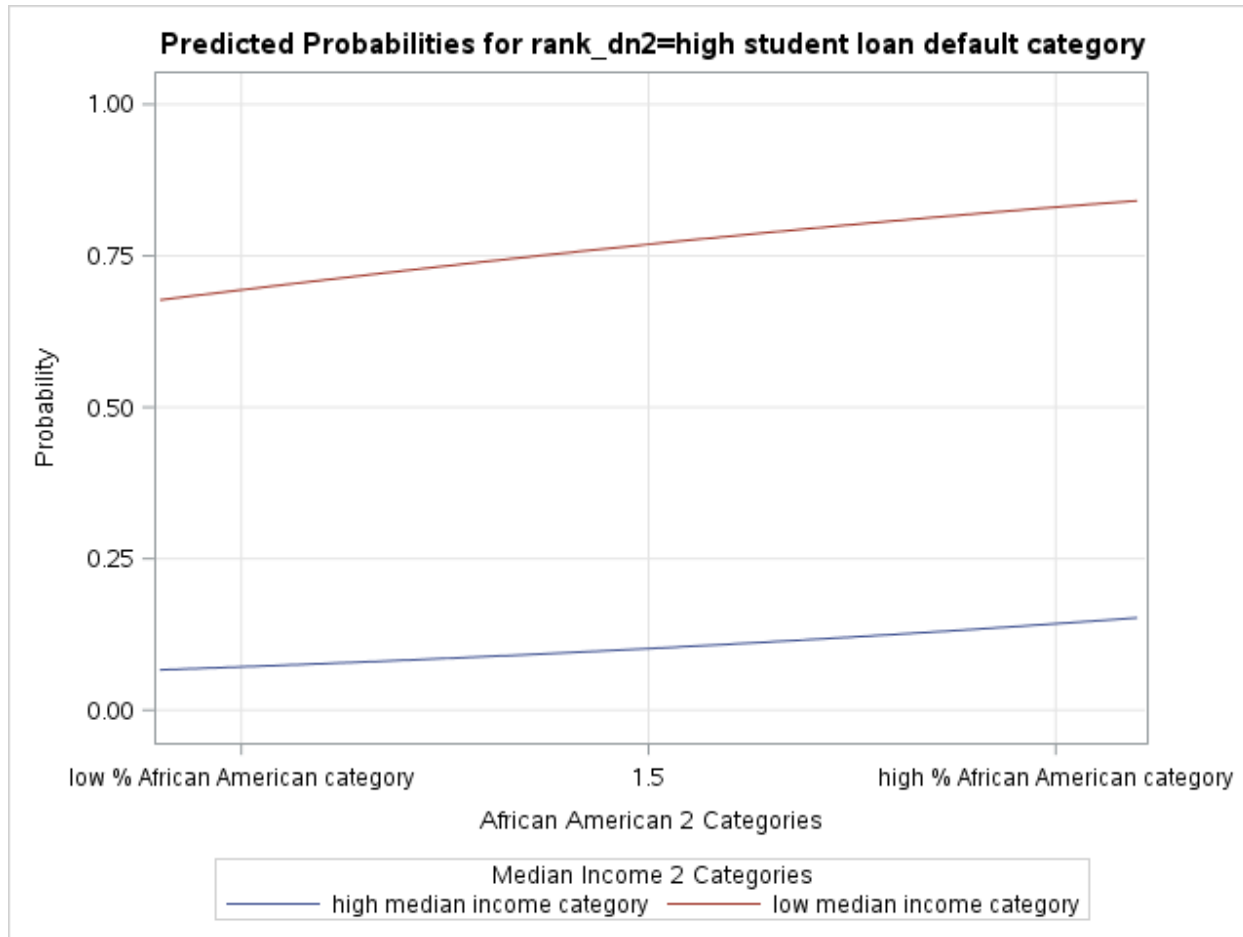
Predicted Probabilities for rank_dn2=high student loan default category

Effects plots use Maximum Likelihood Estimators (MLE) to provide probabilities for all 60 of the sample cases in the sample.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.8178	0.5357	2.3303	0.1269
rank_aamerican2		1	0.7710	0.7178	1.1536	0.2828
rank_median_income2	high median income category	1	-3.3815	0.7578	19.9140	<.0001

Logistic Regression
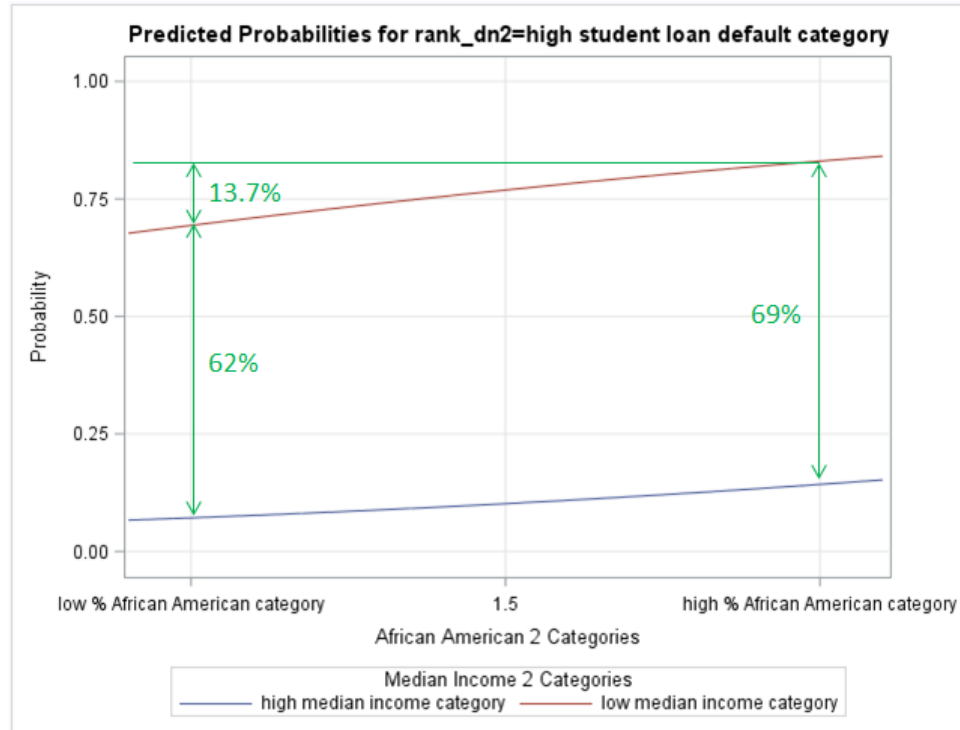
Changes in median income category change the probability estimate of student loan default by greater than (>) 62%

Changes in African American category change the probability estimate of student loan default by less than (<) 14%

These results show that the U.S. Department of Education SHOULD ignore the interaction of race and student loans because logistic regression effects plot results show median income levels are far more associated with student loan defaults

WHY THIS MATTERS

Instinctual/Intuitive thinking vs Analytic/Rational thinking

See: Kahneman, D. (2011): Thinking, Fast and Slow
          Penguin, 496 pp., ISBN 978-0141033570

          Lewis, M. (2016): The Undoing Project
          W. W. Norton & Company, 368 pp., ISBN   978-0393254594

Consequences of incorrect analytic conclusions

Analytic conclusions

Consequences of incorrect associations

➢Approximately 40 million people in poverty in the US (2016 ACS)

> Approximately 8 or 9 million of those in poverty are African American.

> African Americans are over represented in poverty roles (by more than 2 times), but most impoverished people are NOT African American.

➢A 'solution' to student loan defaults that has a focus on race (in particular African Americans) would leave out most impoverished people (perhaps 75% of the folks needing help due to poverty).

Name: Michael Grierson

Organization: Federal Employee

E-Mail: mcg@mcg-ct.com

Web: https://www.mcg-ct.com/saspaper4/

Logistic Regression MLE

$$\text{logit}(p_i)=\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\dots \hat{\beta}_k x_{ki}$$

From SAS® online course "Predictive Modeling Using Logistic Regression (v14.2)"

Beta hats are MLE coefficients that provide the maximum probability (likelihood) of the equation above using all 60 data points as constraints on an optimization to maximize likelihood

Intuitive explanation of maximum likelihood estimation

Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed.

From https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

Logistic Regression effects plot

Logistic Regression effects plot

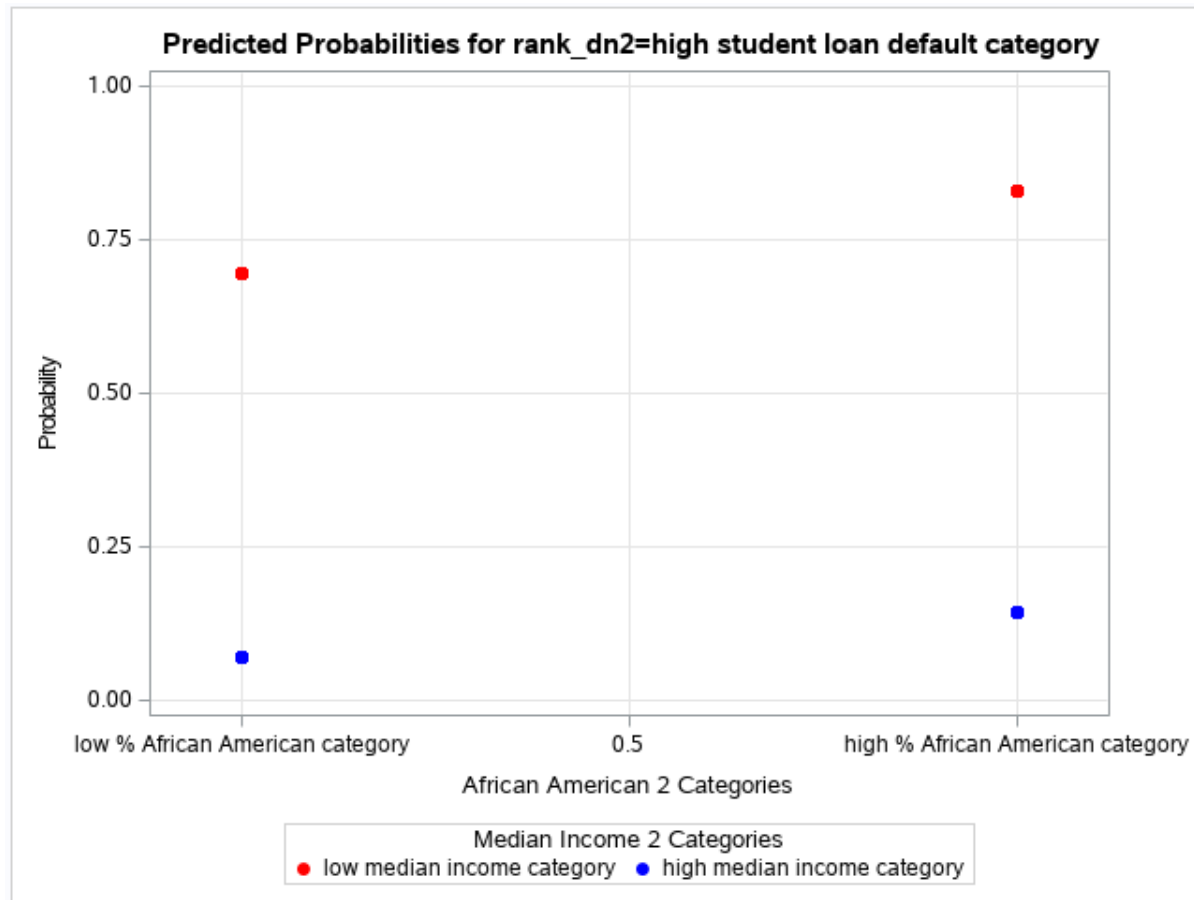# Logistic Regression effects plot

```
proc logistic data=work.sd plots(only)=effect;
  class rank_median_income2 / param=ref;
  model rank_dn2=rank_aamerican2 rank_median_income2 ;
  code file='/folders/myfolders/effectscode.sas';
quit;
title ' ';
data predprob;
  set work.sd;
  %include '/folders/myfolders/effectscode.sas';
run;
```

Logistic Regression effects plot

I_rank_dn2	U_rank_dn2	P_rank_dn20	P_rank_dn21	_LMR_BAD	_st12	_0_0	_LP0	_TEMP	_P0	_P1	_MAXP	_IY
1	1	0.306225779	0.693774221	0	0	1	-3.381537272	0.441391119	0.306225779	0.693774221	0.693774221	2
0	0	0.928489406	0.071510594	0	1	0	0	2.563713366	0.928489406	0.071510594	0.928489406	1
0	0	0.928489406	0.071510594	0	1	0	0	2.563713366	0.928489406	0.071510594	0.928489406	1
1	1	0.306225779	0.693774221	0	0	1	-3.381537272	0.441391119	0.306225779	0.693774221	0.693774221	2
1	1	0.306225779	0.693774221	0	0	1	-3.381537272	0.441391119	0.306225779	0.693774221	0.693774221	2
1	1	0.169552694	0.830447306	0	0	1	-4.152514091	0.204170321	0.169552694	0.830447306	0.830447306	2
1	1	0.306225779	0.693774221	0	0	1	-3.381537272	0.441391119	0.306225779	0.693774221	0.693774221	2
0	0	0.857262458	0.142737542	0	1	0	-0.770976819	1.792736547	0.857262458	0.142737542	0.857262458	1
1	1	0.169552694	0.830447306	0	0	1	-4.152514091	0.204170321	0.169552694	0.830447306	0.830447306	2
1	1	0.169552694	0.830447306	0	0	1	-4.152514091	0.204170321	0.169552694	0.830447306	0.830447306	2
1	1	0.306225779	0.693774221	0	0	1	-3.381537272	0.441391119	0.306225779	0.693774221	0.693774221	2
0	0	0.928489406	0.071510594	0	1	0	0	2.563713366	0.928489406	0.071510594	0.928489406	1
1	1	0.306225779	0.693774221	0	0	1	-3.381537272	0.441391119	0.306225779	0.693774221	0.693774221	2
1	1	0.306225779	0.693774221	0	0	1	-3.381537272	0.441391119	0.306225779	0.693774221	0.693774221	2
1	1	0.306225779	0.693774221	0	0	1	-3.381537272	0.441391119	0.306225779	0.693774221	0.693774221	2
0	0	0.928489406	0.071510594	0	1	0	0	2.563713366	0.928489406	0.071510594	0.928489406	1
0	0	0.928489406	0.071510594	0	1	0	0	2.563713366	0.928489406	0.071510594	0.928489406	1
0	0	0.857262458	0.142737542	0	1	0	-0.770976819	1.792736547	0.857262458	0.142737542	0.857262458	1
1	1	0.169552694	0.830447306	0	0	1	-4.152514091	0.204170321	0.169552694	0.830447306	0.830447306	2
1	1	0.306225779	0.693774221	0	0	1	-3.381537272	0.441391119	0.306225779	0.693774221	0.693774221	2
0	0	0.857262458	0.142737542	0	1	0	-0.770976819	1.792736547	0.857262458	0.142737542	0.857262458	1

Logistic Regression

Predicted Probabilities for rank_dn2=high student loan default category

(Probability plotted on y-axis from 0.00 to 1.00; African American 2 Categories on x-axis: low % African American category, 0.5, high % African American category)

Median Income 2 Categories
● low median income category   ● high median income category

Effects plots use Maximum Likelihood Estimators (MLE) to provide probabilities for all 60 of the sample cases in the sample.

The effects plot just connects the dots