Understanding the Value of Higher Education: A Comparison Based on Institutional Demographics

Chase Henley, Harrison Marick, Joe Feldman

4/29/2018

Abstract

Higher education, and specifically, a college education, is known to be an impetus for economic mobility. As the years have progressed, the system of higher education in the United States has become more racially diverse thanks to progressive movements aimed at improving the representation of minorities in America's universities. However, is the economic benefit of a post-secondary degree the same for a white versus a non-white student? Measuring the economic benefit of a college degree as the median income ten years post entry for students coming from a specific institution, this study aims to investigate whether the demographic make up of a university affects the long term benefits of enrollment. After conducting our analysis, there is conclusive evidence that students from colleges with a larger proportion of white students report higher earnings in the intermediate years after enrollment.

Introduction

How much money will I make after attending college? This is an important question that undoubtedly crosses the minds of millions of U.S. college students at some point during their education. It also exerts an especially powerful influence on those who are debating where they want to attend school.

For many, choosing a college is one of the biggest financial decisions in life. College is getting more and more expensive every year, and people want to know which school is best for investing in an education. The return on such an investment is often measured by the economic benefits one receives years after graduation.

As students at Amherst College, an expensive private institution, it would be interesting to know the degree to which our personal investments will pay off. Of course there are benefits to a quality liberal arts education that are unrelated to money, but a primary objective of college is to qualify an individual to earn a comfortable living in his or her profession of choice.

In our country, young adults entering college have a multitude of options when considering where they will pursue their degree, so if it is economic benefit down the road that they seek, what factors or characteristics of an U.S. institution should they consider when choosing a school?

There are a number of variables we could looked at when trying to figure out what makes a college a good investment, but coming from a school as diverse as Amherst College, we decided to focus on racial demographics as our possible explanatory variable for economic success.

It is fairly well-known that white people have had a traditionally stronger representation in the American college system, and white people, on average, have also had higher incomes later in life. However, it seems this trend has recently been changing. In fact, the percentage of white students enrolled in degree-granting post secondary education institutions during the period between 1976 and 2015 has fallen from 84% to 54% (National Center for Education Statistics). Thus, with the influx of minority students into higher education, is there a difference in the economic benefits reaped by white students versus students of color?

Specifically, our project explores differences in students' later-in-life incomes between "predominantly white"" and "non-predominantly white institutions" (This distinction comes later in the paper). We measure the economic benefit derived from an institution (4-year schools only) as students' median earnings 10 years post-enrollment.

<<<<<<< HEAD We begin our analysis by looking at basic uni-variate statistics and then explore variables visually using histograms and density plots— all the while using what we learned throughout the semester about bin width selection methods, kernels, and bandwidth selection methods.

Then, we perform parametric and non-parametric procedures to see if there is a difference in median incomes between students coming from predominantly white versus non-predominantly white institutions. We do a t-test for differences between groups for the parametric procedure and a rank-sum test for the non-parametric procedure. We find concurring results from the non-parametric and parametric tests.

Moreover, we explore the bi-variate relationship between the percentage of white students at a given institution and median incomes 10 years post-enrollment, and we end up building two linear models: one that is constructed parametrically with simple linear regression (SLR) and one that is constructed nonparametrically using a rank-based solution. Our models turn out having quite similar slopes and model utilities.

Without going into too much detail on the statistical finds from these procedures, we will go ahead and say that we found evidence to conclude that students from predominantly white institutions have higher median incomes 10 years post-enrollment compared to students who went to non-predominantly white institutions.

We begin our analysis by looking at basic uni-variate statistics and then explore variables visually using histograms and density plots, all the while using what we learned throughout the semester about bin width selection methods, kernels, and bandwidth selection methods to get a more clear understanding of the distributions of our variables of interest.

Then, we perform parametric and non-parametric procedures to see if there is a difference in median incomes between students coming from predominantly white versus non-predominantly white institutions. We do a t-test for differences between group means for the parametric procedure and a rank-sum test for the non-parametric procedure to look at medians. We find concurring results from the non-parametric and parametric tests.

Moreover, we explore the bi-variate relationship between the percentage of white students at a given institution and median incomes 10 years post-enrollment, and we end up building two linear models: one that is constructed parametrically with simple linear regression (SLR) and one that is constructed nonparametrically using a rank-based solution. Our models turn out having quite similar slopes and model utilities.

Without revealing the specifics of our findings, we have evidence to conclude that students from "predominantly white" institutions have higher median incomes 10 years post-enrollment compared to students who attended non-predominantly white institutions. >>>>>>> 243bf3fc053a4e2e206076c0e7e994e2644fcbdf

However, it is our ethical obligation to consider the limitations to our project and our resulting conclusions. Our scope is constrained by the institutions we looked at in the U.S. and the observations that we were forced to remove due to missing information. With so many factors relating to financial success, there is the possibility of unaddressed confounding variables having an unforeseen effect on our conclusions as well.

Nevertheless, we can safely that we've made significant progress towards fulfilling our initial project goal, and we hope that you continue reading this paper to gain a more in-depth perspective on our findings and how they pertain to the broader question of how racial demographics of an institution relate to post-graduation economic success.

Data

Our data was obtained from the College Scorecard, a database managed by the U.S. Department of Education, and contains 7,312 observations. Each observation is a different 4-year academic institution, and each is located within the United States.

Data was collected by schools themselves via surveys and enrollment information, and then submitted to the U.S. Department of Education.

Analysis

The two variables we looked at in our analysis are the following:

`MD_EARN_WNE_P10`-Median income ($) 10 years after entry of students working and not enrolled in school; quantitative. Used as our measure of economic benefit derived from an institution.

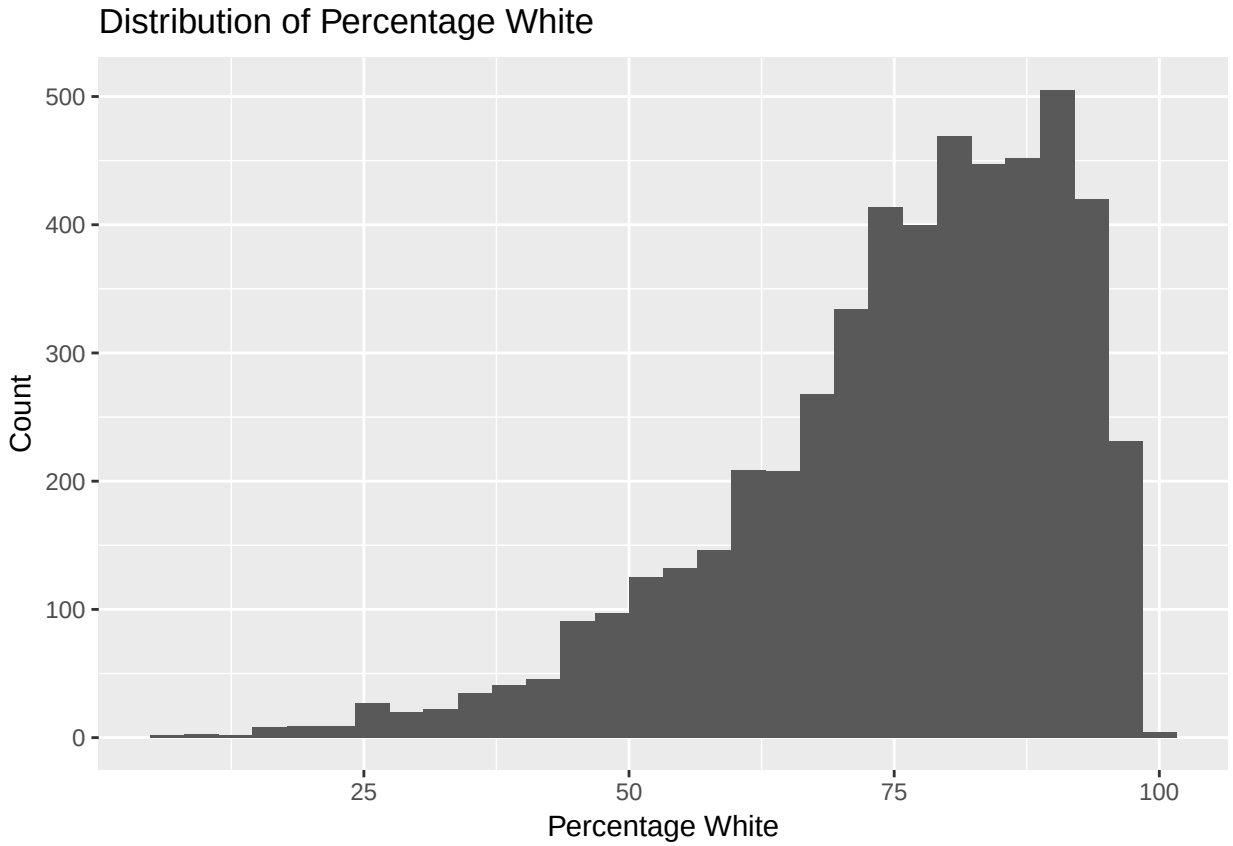
`PCT_WHITE`- Percent of the population from students' zip codes that is White, via Census data; quantitative. Used as a predictor for economic benefit derived from an institution.

<<<<<<< HEAD We also created a new indicator variable, 'Predominantly White', to distinguish schools that contained a study body with a percentage of white students greater than 67.09% as predominantly white from those with a percentage of white students less than or equal to 67.09%, which we declared non-predominantly white. 67.09% is the first quartile for the `PCT_WHITE`, and this cut-off was chosen somewhat arbitrarily, but with respect to the variable's distribution. ======= We selected percent white as an explanatory variable, as it gave a very broad sense of the racial demographics of a school. While it certainly does not paint the entire picture, in a country that is predominantly white, understanding the degree to which a student body is white gives at least an idea of the diversity of the students. >>>>>>> 243bf3fc053a4e2e206076c0e7e994e2644fcbdf

We selected the variable of meadian earnings 10 years after entry as our response variable as it provided a measure of the center of the distribution for students' incomes down the road. We selected median over mean, as it is less sensitive to outliers (i.e. millionaires and billionaires).

We also created a new indicator variable, 'Predominantly White'. Since the distribution of 'PCT_WHITE' had a large left skew, with the vast majority of observations falling above the 50% line, we made the arbitrary distinction of a 'Predominantly White' institution at the first quartile of the 'PCT_WHITE' variable. Specifically, schools with a student body in which 67.09% of those enrolled are white fall into the 'Predominantly White' category, while those with a proportion below this threshold are labeled as 'Non-Predominantly White'.

Before conducting any tests, either parametric or non parametric, it is important to examine the distribution of our variables of interest.



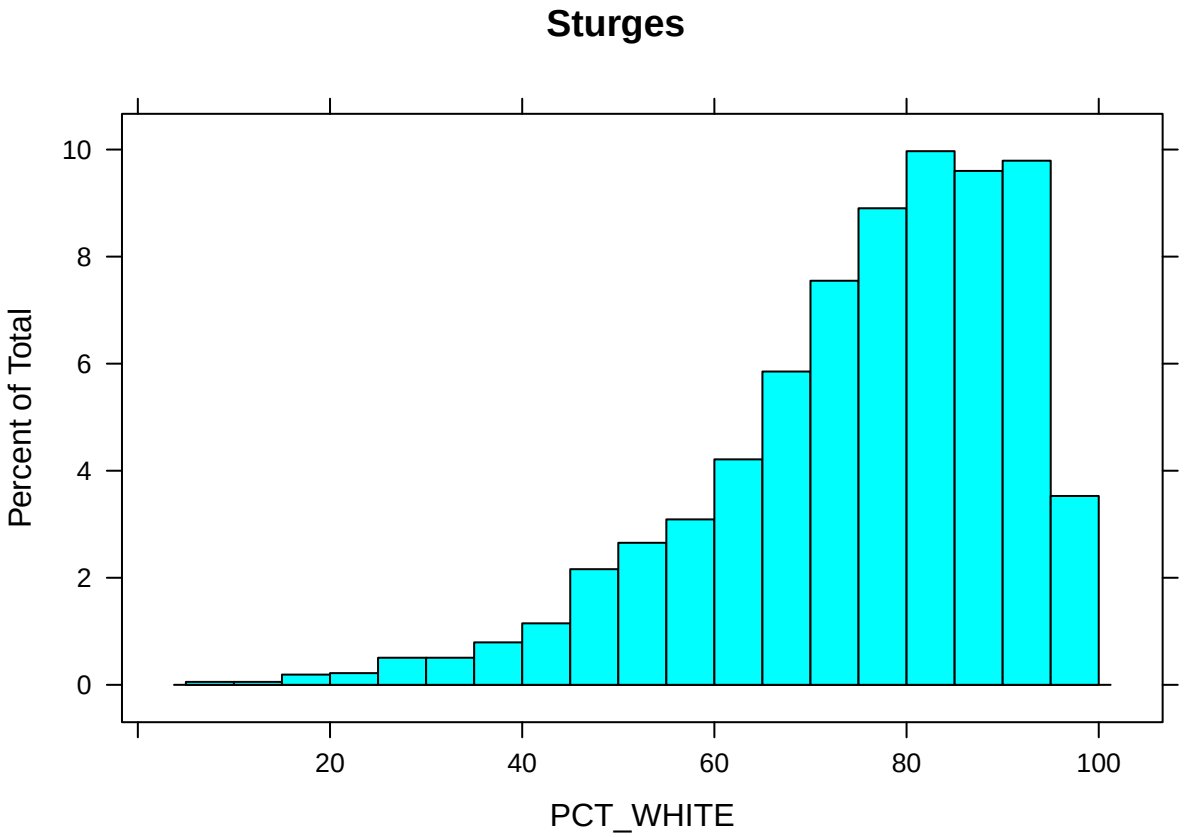
<<<<<<< HEAD

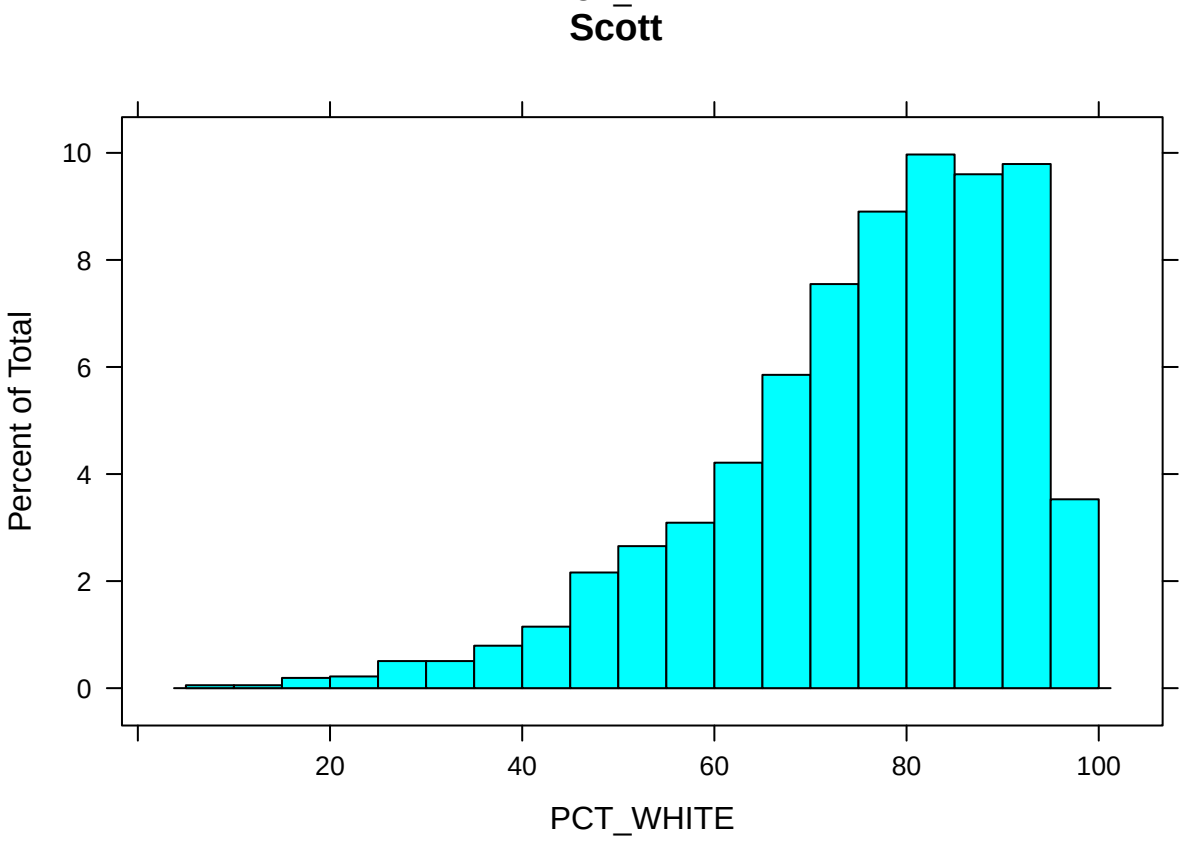
Conducting Uni-variate Analysis on the distribution of the percentage of students that are white, it is clear that this distribution is quite skewed left. The mean percentage is 75.509 while the median is 78.685. The data ranges from from 5.34 to 98.98 with a standard deviation of 15.879.

Conducting Univariate Analysis on the distribution of the percentage of students that are white, it is clear that this distribution is skewed left. The mean percentage is 75.509 while the median is 78.685. The data ranges from from 5.34 to 98.98 with a standard deviation of 15.879. >>>>>>> 243bf3fc053a4e2e206076c0e7e994e2644fcbdf

Intuitively, this distribution makes sense. White people have a traditionally stronger representation in the American college system, and this distribution supports that claim as there is more density around colleges with a proportionately high white student body.

Naturally, we explored different bin-widths with our histograms in order to produce the most accurate picture of the distribution. While we explored various different bin-widths, we found that Scott's and Sturge's methods of bin-width selection produced accurate histograms. In fact, R produced the same exact plot using both methods.



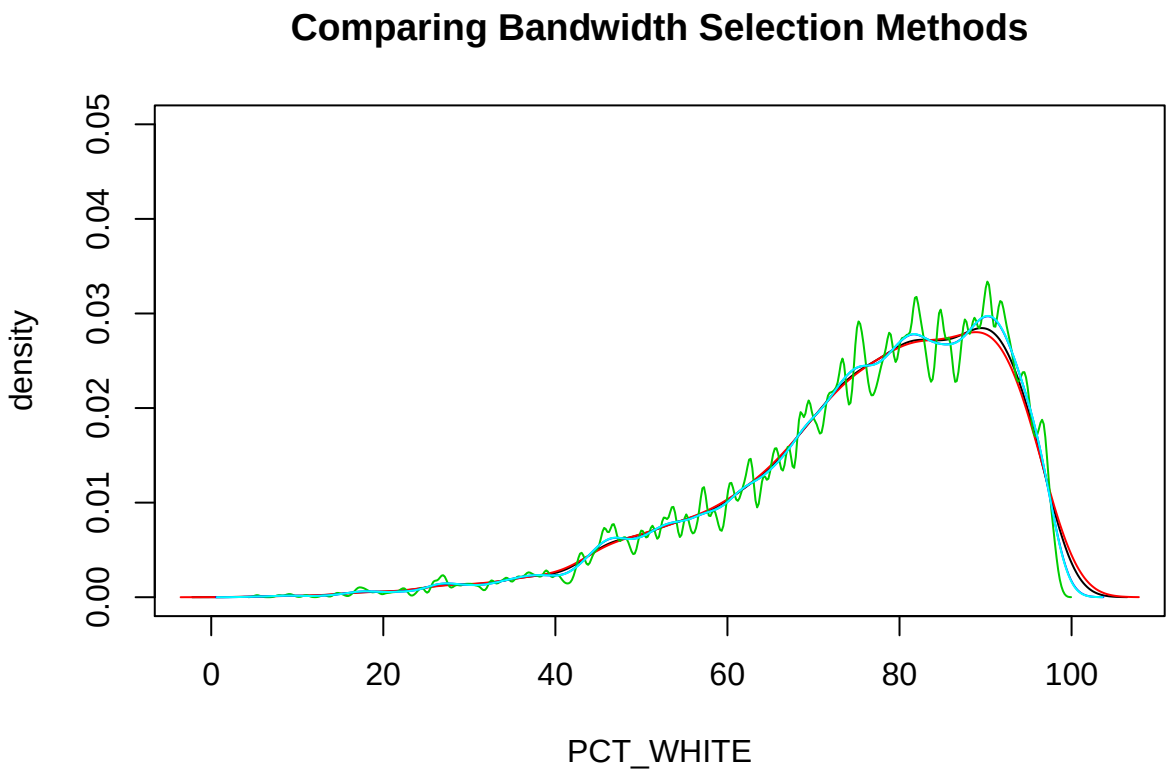


<<<<<<< HEAD In the interest of being as thorough as possible, we elected to also look at kernel density

plots. We varied the kernel function and the bandwidth selection in order to find the ideal combination. =======

In the interest of being as thorough as possible, we elected to also look at kernal density plots. We varied the kernal function and the bandwidth selection in order to find the ideal combination. >>>>>>> 243bf3fc053a4e2e206076c0e7e994e2644fcbdf

After comparing different kernel functions with the default bin-width, we found the normal kernel to be the most smooth, producing what we felt was the best representation of the data. Below, we have compared different bandwidth selection methods.



We see that the green and light blue lines are the least smooth—these were produced using the ucv and SJ bandwidth selection methods, respectively. These are both data-driven cross-validation methods.

The black, red, and darker blue lines are all fairly similar in terms of smoothness, and these were produced with the nrd0, nrd, and ucv bandwidth selection methods. The nrd0 and nrd methods are both normal-based and work ideally with data sets that are normally distributed. The ucv method is a cross-validation method as its name implies. Thus, considering our data is not normally distributed, we prefer the ucv method in this setting. Having said that, the ucv method produces a density plot that is not very smooth. As a result, we feel any of the other lines suffice to accurately depict our data.

With an accurate picture of the distribution of `PCT_WHITE`, we can now begin to conduct tests to determine if there is a significant difference in future income based on the "whiteness" of a school.

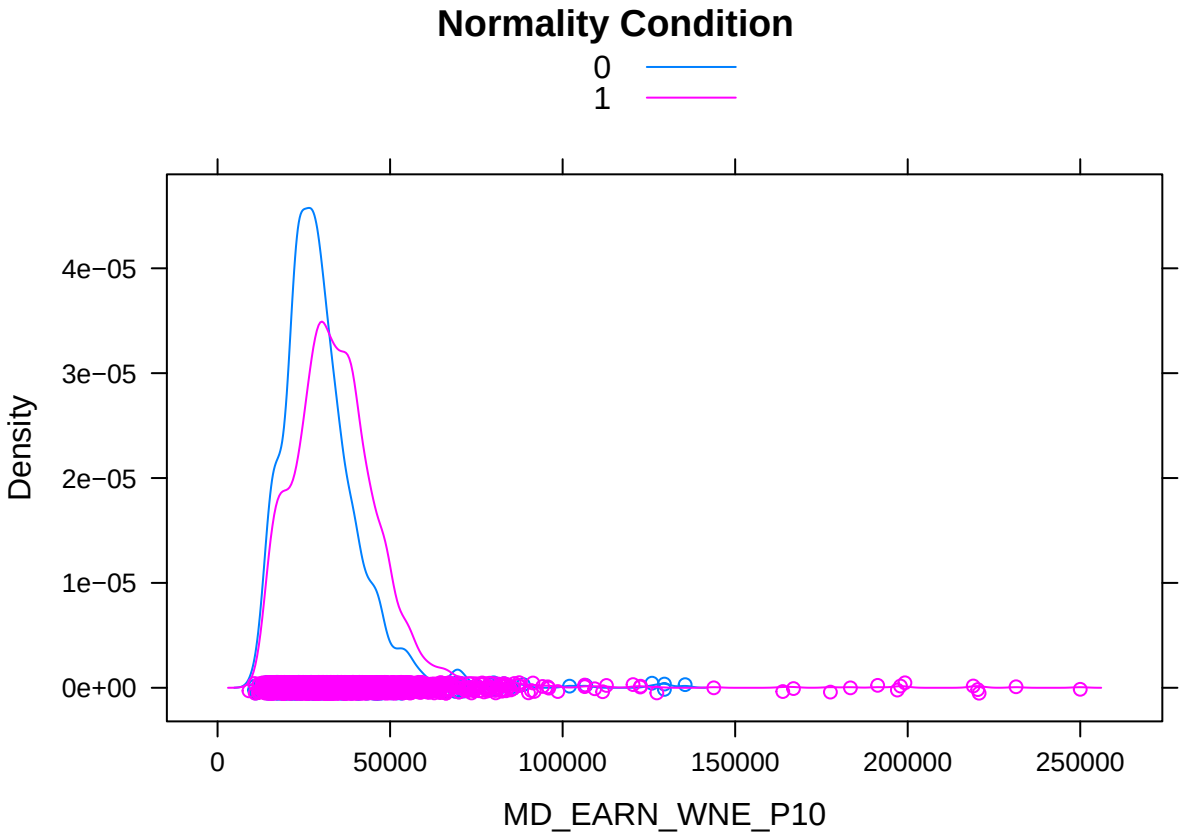
We elected to conduct a Two Sample T-Test to test whether or not there is a difference in average median income 10 years post entry between schools that are predominantly white versus schools with a lower percentage of students that are white. Before actually conducting the test, we needed to split the data into two groups. As referenced above, we user a `PredominantlyWhite` variable to distinguish between schools. This value corresponds to the first quartile of the `PCT_WHITE` variable.

After creating two groups, we conducted a parametric two sample t-test.

```
##
##  Welch Two Sample t-test
##
## data:  filter(data2, PredominantlyWhite == 1)$MD_EARN_WNE_P10 and filter(data2, PredominantlyWhite ==
## t = 10.377, df = 2623.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4008.179 5875.991
## sample estimates:
## mean of x mean of y
##  35487.93  30545.85
```

<<<<<<< HEAD We have a p-value for our t-statistic that is virtually 0, which is statistically significant at any reasonable alpha level. Additionally, we have a 95% confidence interval for the different in means between the two groups to be (-5875.99, -4008.18), which is completely below 0. We have sufficient evidence to reject the null in favor of the alternative that the predominantly white schools have a higher mean value for `MD_EARN_WNE_P10`. In other words, the mean of the median earning 10 years after entry is higher for predominantly white schools. ======= We have a p-value for our t-statistic that is virtually 0, which is statistially significant at any reasonable alpha level. Additionally, we have a 95% confidence interval for the difference in average median income between the two groups that lies completely above 0. We have sufficient evidence to reject the null in favor of the alternative that the predominantly white schools have a higher mean value for `MD_EARN_WNE_P10`. In other words, the mean of the median earning 10 years after entry is higher for predominantly white schools. >>>>>>> 243bf3fc053a4e2e206076c0e7e994e2644fcbdf

Of course, this parametric t-test depends on the normality of the data, which is in question.



<<<<<<< HEAD Notice that for both groups, there is a distinct right skew, which means that the normality

condition for the parametric t-test do not hold. Unfortunately, the conclusions above are now in question. We conducted a Rank-Sum Procedure in order to verify our conclusions using a non-parametric procedure.

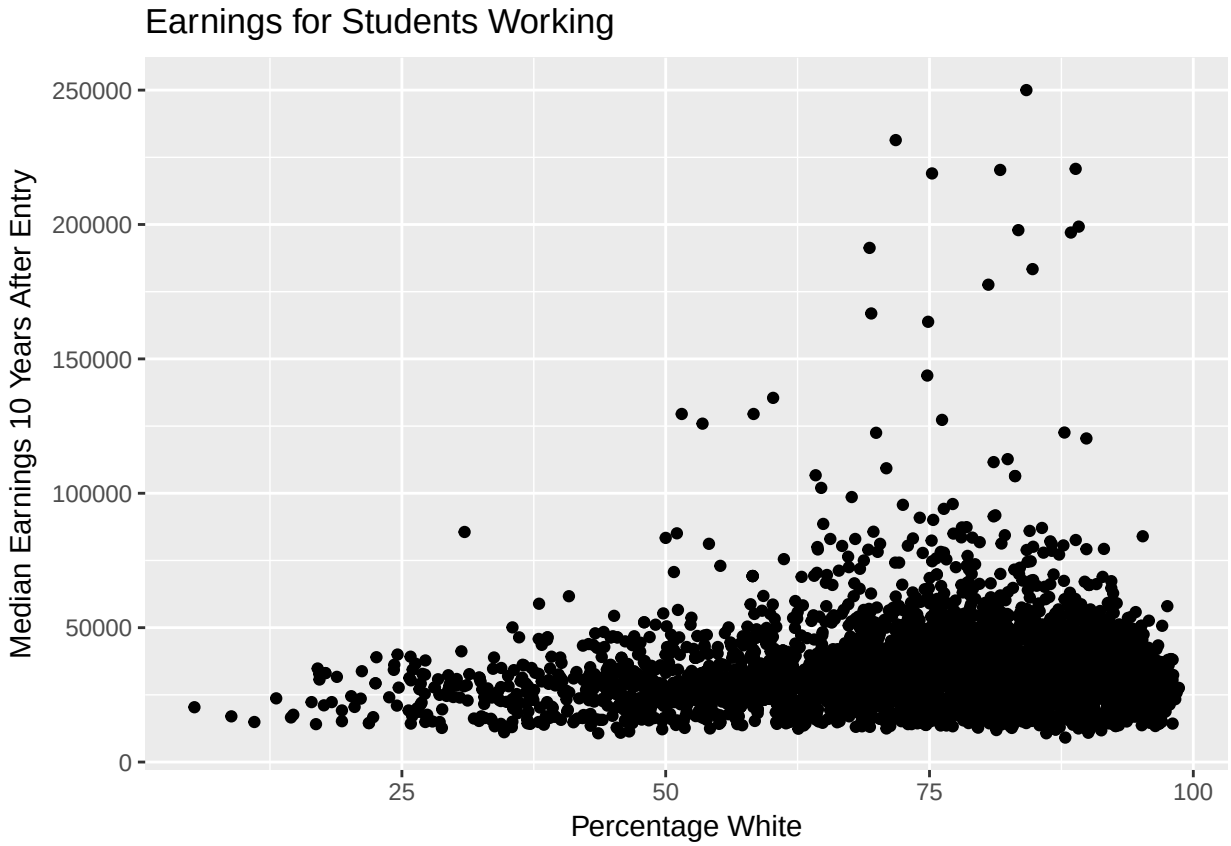
Looking at the density plot above, it seems reasonable to utilize a shift model, since the distributions appear to have the same shape but for potentially different centers.

Notice that for both groups, there is a distinct right skew, which means that the normality condition for the parametric t-test does not hold. Unfortunately, the conclusions above are now in question. We conducted a Rank-Sum Procedure in order to verify our conclusions using a nonparametric procedure. Looking at the density plot above, it seems reasonable to utilize a shift model, since the distributions appear to have the same shape but for potentially different centers. >>>>>>> 243bf3fc053a4e2e206076c0e7e994e2644fcbdf

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  filter(data2, PredominantlyWhite == 1)$MD_EARN_WNE_P10 and filter(data2, PredominantlyWhite ==
## W = 2542200, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  3900 5300
## sample estimates:
## difference in location
##                   4600
```

<<<<<<< HEAD With a test statistic of 0 and corresponding p-value of virtually 0, we have sufficient evidence to reject our null hypothesis in favor of the alternative that the difference in distribution centers is not zero. With a confidence interval that lies completely below 0, we have sufficient evidence to conclude that the predominantly white group has a median median income 10 years post entry that is higher than that of the non predominantly white group. This conclusion is consistent with the results of the parametric test ======= With a test statistic of 2542200 and corresponding p-value of virtually 0, we have sufficient evidence to reject our null hypothesis in favor of the alternatie that the difference in distribution centers is not zero. With a confidence interval that lies completely above, we have sufficient evidence to conclude that the predominantly white group has a median median income 10 years post entry that is higher than that of the non predominantly white group. This conclusion is consistent with the results of the parametric test. >>>>>>> 243bf3fc053a4e2e206076c0e7e994e2644fcbdf

After confidently finding a difference between the two group means, we looked for a linear relationship between `PCT_WHITE` and `MD_EARN_WNE_P10`.

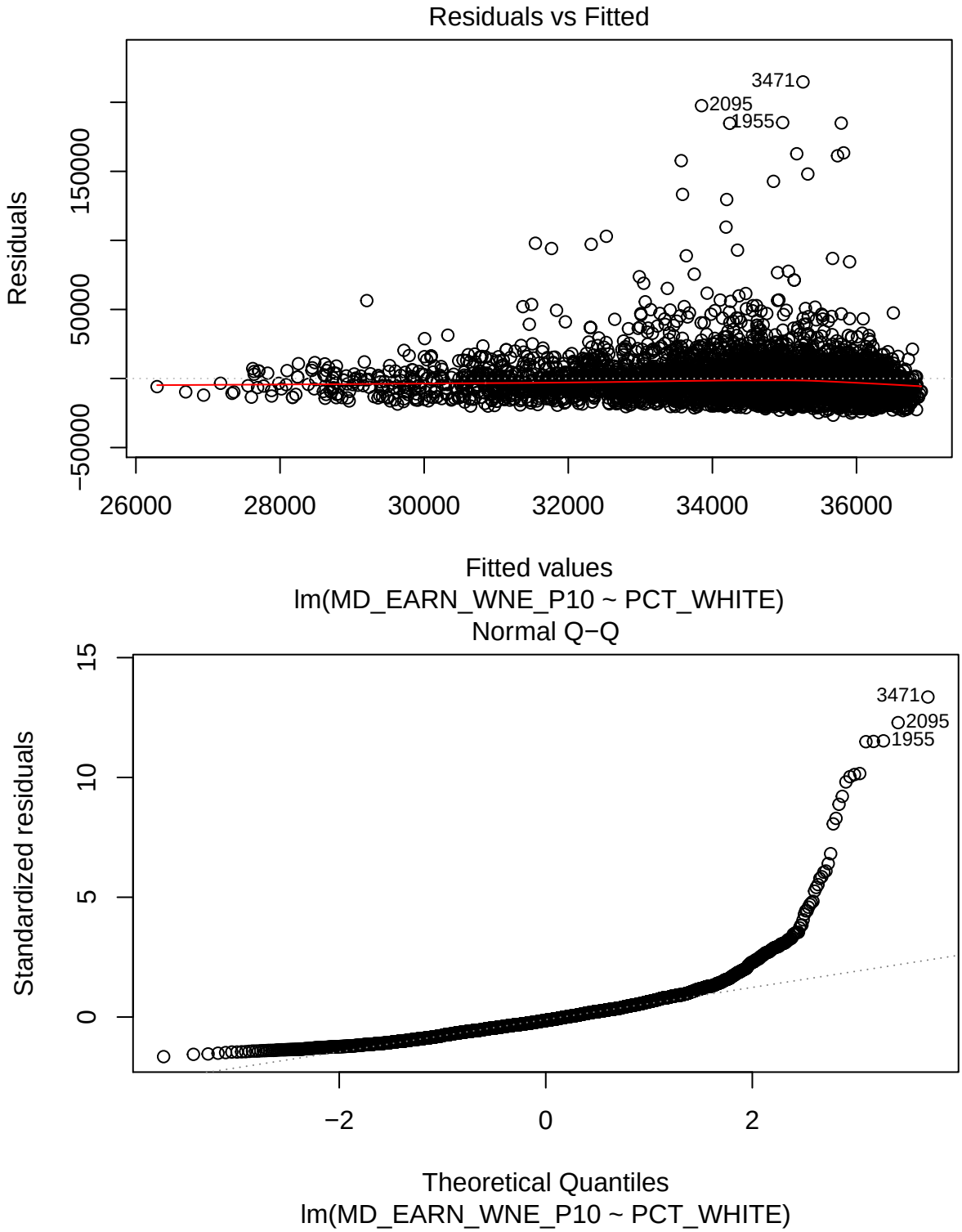


There appears to be a weak positive linear relationship between the two variables. It seems reasonable to conduct a simple linear model.

```
##
## Call:
## lm(formula = MD_EARN_WNE_P10 ~ PCT_WHITE, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -26576  -9009  -2248   5558 214743
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25686.44    1143.28  22.467  < 2e-16 ***
## PCT_WHITE     113.69      14.88   7.641  2.6e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16080 on 4650 degrees of freedom
## Multiple R-squared:  0.0124, Adjusted R-squared:  0.01219
## F-statistic: 58.38 on 1 and 4650 DF,  p-value: 2.603e-14
```

Beginning with the parametric model we have a slope coefficient of 113.69, which has a statistically significant p-value at any reasonable alpha level. The coefficient of 113.69 means that for every additional percentage "whiter" a school is, the median student earnings 10 years post entry is expected to increase by $113.69. Additionally, the overall model utility test is statistically significant. As a result, the parametric model revealed there is a relationship between `PCT_WHITE` and `MD_EARN_WNE_P10`.

Naturally, the conclusions from the parametric model are only valid if we can verify the conditions.



Examining both the residuals vs. fitted plot and the qqplot, both the constant variance and normality of the errors conditions are not met. There is a clear increase in the bandwidth of the errors as the fitted values increase and the upper tail of the plot moves well off the normal line. In this instance, a non-parametric

model appears to be a better choice.

```
## Call:
## rfit.default(formula = MD_EARN_WNE_P10 ~ PCT_WHITE, data = data)
##
## Coefficients:
##              Estimate Std. Error t.value   p.value
## (Intercept) 23257.840    795.038  29.254 < 2.2e-16 ***
## PCT_WHITE     116.008     10.319  11.242 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared (Robust): 0.02728063
## Reduction in Dispersion Test: 130.4127 p-value: 0
```

<<<<<<< HEAD In the interest of brevity, we will spare the formal interpretation of the non-parametric results. In short, the conclusions are the same, as the p-value for the slope coefficient is virtually zero. With a coefficient of 116.01, the slope for this model is very similar to that of the parametric model. Since there are virtually no conditions to account for here, other than the independence and randomization of our data, we feel confident in the results from the non parametric model. ======= In the interest of brevity, we will spare the formal interpretation of the nonparametric results. In short, the conclusions are the same, as the p-value for the slope coefficient is virtually zero. With a coefficient of 116.01, the slope for this model is very similar to that of the parametric model. Since there are virtually no conditions to account for here, other than the independence and randomization, which are both inherently present in our dataset, we feel confident in the results from the non parametric model. >>>>>>> 243bf3fc053a4e2e206076c0e7e994e2644fcbdf

Results

In summary, we found that there is a clear difference in the median median income 10 years post entry between predominantly white schools and non predominantly white schools. Additionally, we found a positive linear relationship between percentage white and median income 10 years post entry. For both tests, we relied heavily upon the non parametric test procedures, as assumptions for parametric tests could not be met.

Based on our analysis, there appears to be an economic benefit later in life to attending a predominantly white college or university. It is important to note that this does not mean that the reason for this relationship is because the college or university is "whiter". Our analysis is merely based on correlation, which says nothing of causal relationships. The primary third variable that we have not accounted for is quality of the school. It is reasonable to expect more prestigious schools to produce a higher median income 10 years post entry. Either way, there appears to be a correlative economic benefit to attending a more predominantly white college or university.

<<<<<<< HEAD ###Conclusion =======

Conclusion

Our analysis was aimed at investigating whether there are discrepancies in the monetary gain of attending a certain university based on its demographic make up. Specifically, we sought to highlight a potential difference in the earnings of students who attended what we labeled as "Predominantly White" Universities versus students who attended "Non-Predominantly White" Universities. After conducting both parametric and non parametric analyses for the differences in group means or medians (depending on the statistical framework - parametric vs. non-parametric) of ten year post entry median earnings of students, the results were significant. Students from predominantly white universities reported median incomes higher than students coming from non predominantly white institutions. The tests demonstrated that students from predominantly white institutions reported median incomes between $$$4,000 and $$$5,000 higher than students from non predominantly white institutions. In fitting a model in both parametric and non-parametric settings, the

goal was to highlight the relationship between the percentage of the student body that is white and median income post entry 10 years post entry. In each model, all coefficients were significant and positive.

With this relationship established, several potential avenues for research are opened up. Namely, the "Non Predominantly White" institutions could be filtered from the data set, and investigated for potentially similar characteristics. For instance, one could look at the breakdown of majors from these universities and conduct similar statistical tests to see whether the percentage of STEM majors at these institutions is significantly different from the proportion at "Predominantly White" universities, since STEM degrees traditionally afford more lucrative opportunities upon graduation. There could also be analysis on the structure of these universities: What is the average or median student faculty ratio? What are the retention rates? What are the four year graduation rates? All of these statistics could then be compared to those from "Predominantly White" universities.

Of course, any conclusions drawn from our research should not be made without acknowledgement of the flaws of our data set. For one, nearly a third of the observations of in our whole set were missing, limiting the scope of inference in our study. In addition, it is necessary to note that the observations in our set come at the institution level. With that said, there is no concept of the income distribution by race within the university, only the mean and median ten years post entry of the entire student body. This makes the conclusion of our study slightly more difficult to interpret. We cannot say that enrolling in a university that is proportionately more white is more lucrative down the line for an individual student. For instance, it may be the case that minorities at such a school earn way less, but that the white demographic, since it represents the majority of the student body, outweighs this effect in calculation of the median and mean earnings. Therefore, the scope of our research extends only to the institution level. Specifically, universities that have a higher proportion of white students report higher median incomes for their students in the intermediate future post graduation.
>>>>>>> 243bf3fc053a4e2e206076c0e7e994e2644fcbdf

Our analysis was aimed at investigating whether there are discrepancies in the monetary gain of attending a certain university based on its demographic make up. Specifically, we sought to highlight a potential difference in the earnings of students who attended what we labeled as "Predominantly White" Universities versus students who attended "Non-Predominantly White" Universities. After conducting both parametric and non parametric analyses for the differences in group means or medians (depending on the statistical framework - parametric vs. non-parametric) of ten year post entry median earnings of students, the results were significant. Students from predominantly white universities reported median incomes higher than students coming from non predominantly white institutions, and the tests demonstrated that students from predominantly white institutions reported median incomes between $4,000 and $5,000. In fitting a model in both parametric and non-parametric settings, the goal was to highlight the relationship between the percentage of the student body that is white and median income post entry. In each model, all coefficients were significant. The interpretation is that for every increase in percentage of the student body that is white, potential income down the road is increased by nearly $120.

With this relationship established, several potential avenues for research are opened up. Namely, the "Non Predominantly White" institutions could be filtered from the data set, and investigated for potentially similar characteristics. For instance, one could look at the breakdown of majors from these universities and conduct similar statistical tests to see whether the percentage of STEM majors, majors that afford students opportunities in particularly lucrative fields post-graduation, is significantly different from students at "Predominantly White" universities. There could also be analysis on the structure of these universities: What is the average or median student faculty ratio. What are the retention rates? What are the four year graduation rates? All of these statistics could then be compared to the "Predominantly White" universities.

<<<<<<< HEAD Of course, any conclusions drawn from our research should not be made without acknowledgement of the flaws of our data set. For one, nearly a third of the observations of in our whole set were missing, limiting the scope of inference in our study. In addition, it is necessary to note that the observations in our set come at the institution level. With that said, there is no concept of the income distribution by race within the university, only the mean and median ten years post entry of the entire student body. This makes the conclusion of our study slightly more difficult to interpret. We cannot say that enrolling in a university that is proportionately more white is more lucrative down the line for an

individual student. For instance, it may be the case that minorities at such a school earn way less, but that the white demographic, since it represents the majority of the student body, outweighs this effect in calculation of the median and mean earnings. Therefore, the scope of our research extends only to the institution level. Specifically, universities that have a higher proportion of white students report higher median incomes for their students in the intermediate future post graduation. ======= >>>>>>>

243bf3fc053a4e2e206076c0e7e994e2644fcbdf

Works Cited

"Status and Trends in the Education of Racial and Ethnic Minorities." Revenues and Expenditures for Public Elementary and Secondary Education: School Year 2001-2002, E.D. Tab, nces.ed.gov/pubs2010/2010015/indicator6_24.asp.