College Scorecard

December 2017

Introduction

Data on characteristics of US institutions of higher education was collected in an effort to make more transparent issues of cost, debt, completion rates, and post-graduation earning potential. An undertaking of the U.S. Department of Education, the College Scorecard data represent a compilation of institutional reporting, federal financial aid reports, and tax information. The process of gathering and compiling the data is well documented on the College Scorecard website https://collegescorecard.ed.gov/data/documentation/. One caveat is that some of the variables have only been collected on students receiving federal financial aid. Biases inherent to analyses done on data collected from a subgroup should be considered.

Data information & loading data

There are multiple ways of downloading the College Scorecard data. The data are available: for all years (1996-2013) in a .zip file; as the most recent year (as this file is written, the most recent year is 2013) in a .csv file; or as the scorecard only data in a .csv file. https://collegescorecard.ed.gov/data/. For the analysis below, we have used the 2013 most recent data. The original file contains 7804 institutions and 1728 variables.

The dataset is incredibly rich. The variables are broken down by race, family income, first generation status, age of student, etc. It allows for a student to investigate political or personal hypotheses about college education and the costs and benefits within. The variables are described in a data dictionary given at https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary-09-08-2015.csv.

```
college_url <- "https://s3.amazonaws.com/ed-college-choice-public/Most+Recent+Cohorts+(All+Data+Element
college_data <- read_csv(college_url)
dim(college_data)
```

```
## [1] 7804 1728
```

Let's only use some of the variables, and also let's make sure that they are all numeric with NA coded appropriately.

```
college_debt = college_data %>%
  dplyr::select(INSTNM,STABBR,PREDDEG, HIGHDEG, region, LOCALE,
         CCUGPROF,HBCU,WOMENONLY, RELAFFIL,ADM_RATE,SATVRMID,
         SATMTMID,SATWRMID,SAT_AVG, UG,NPT4_PUB, NPT4_PRIV,
         COSTT4_A, DEBT_MDN, CUML_DEBT_P90, mn_earn_wne_p10,
         md_earn_wne_p10) %>%
  mutate(ADM_RATE = readr::parse_number(ADM_RATE),
         SATVRMID = readr::parse_number(SATVRMID),
         SATMTMID = readr::parse_number(SATMTMID),
         SATWRMID = readr::parse_number(SATWRMID),
         SAT_AVG = readr::parse_number(SAT_AVG),
         UG = readr::parse_number(UG),
         NPT4_PUB = readr::parse_number(NPT4_PUB),
         NPT4_PRIV = readr::parse_number(NPT4_PRIV),
         COSTT4_A = readr::parse_number(COSTT4_A),
         DEBT_MDN = readr::parse_number(DEBT_MDN),
         CUML_DEBT_P90 = readr::parse_number(CUML_DEBT_P90),
```

```r
        mn_earn_wne_p10 = readr::parse_number(mn_earn_wne_p10),
        md_earn_wne_p10 = readr::parse_number(md_earn_wne_p10)) %>%
  mutate(RELAFFIL = ifelse(RELAFFIL=="NULL", NA, RELAFFIL),
        LOCALE = ifelse(LOCALE =="NULL", NA, LOCALE),
        CCUGPROF = ifelse(CCUGPROF=="NULL", NA, CCUGPROF),
        HBCU = ifelse(HBCU=="NULL", NA, HBCU),
        WOMENONLY = ifelse(WOMENONLY=="NULL", NA, WOMENONLY)) %>%
  mutate(region2 = ifelse(region=="0", "Military",
                ifelse(region=="1", "New England",
                ifelse(region=="2", "Mid East",
                ifelse(region=="3", "Great Lakes",
                ifelse(region=="4", "Plains",
                ifelse(region=="5", "Southeast",
                ifelse(region=="6", "Southwest",
                ifelse(region=="7", "Rocky Mnts",
                ifelse(region=="8", "Far West", "Outlying"))))))))))

str(college_debt)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    7804 obs. of  24 variables:
##  $ INSTNM         : chr  "Alabama A & M University" "University of Alabama at Birmingham" "Amridge Ur
##  $ STABBR         : chr  "AL" "AL" "AL" "AL" ...
##  $ PREDDEG        : int  3 3 3 3 3 3 2 3 3 3 ...
##  $ HIGHDEG        : int  4 4 4 4 4 4 2 3 4 4 ...
##  $ region         : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ LOCALE         : chr  "12" "12" "12" "12" ...
##  $ CCUGPROF       : chr  "9" "8" "6" "8" ...
##  $ HBCU           : chr  "1" "0" "0" "0" ...
##  $ WOMENONLY      : chr  "0" "0" "0" "0" ...
##  $ RELAFFIL       : chr  NA NA "74" NA ...
##  $ ADM_RATE       : atomic  0.899 0.867 NA 0.806 0.512 ...
##   ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  5584 obs. of  4 variables:
##   .. ..$ row     : int  3 7 8 12 13 15 16 17 18 19 ...
##   .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##   .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##   .. ..$ actual  : chr  "NULL" "NULL" "NULL" "NULL" ...
##  $ SATVRMID       : atomic  410 580 NA 575 430 555 NA NA NA 570 ...
##   ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  6503 obs. of  4 variables:
##   .. ..$ row     : int  3 7 8 9 12 13 14 15 16 17 ...
##   .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##   .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##   .. ..$ actual  : chr  "NULL" "NULL" "NULL" "NULL" ...
##  $ SATMTMID       : atomic  400 585 NA 580 425 570 NA NA NA 595 ...
##   ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  6489 obs. of  4 variables:
##   .. ..$ row     : int  3 7 8 9 12 13 14 15 16 17 ...
##   .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##   .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##   .. ..$ actual  : chr  "NULL" "NULL" "NULL" "NULL" ...
##  $ SATWRMID       : atomic  NA NA NA NA NA 540 NA NA NA 565 ...
##   ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  7011 obs. of  4 variables:
##   .. ..$ row     : int  1 2 3 4 5 7 8 9 11 12 ...
##   .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##   .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##   .. ..$ actual  : chr  "NULL" "NULL" "NULL" "NULL" ...
```

```
## $ SAT_AVG        : atomic  823 1146 NA 1180 830 ...
##  ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  6384 obs. of  4 variables:
##  .. ..$ row     : int  3 7 8 12 13 14 15 16 17 18 ...
##  .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##  .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##  .. ..$ actual  : chr  "NULL" "NULL" "NULL" "NULL" ...
## $ UG             : atomic  4380 10331 98 5220 4348 ...
##  ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  2848 obs. of  4 variables:
##  .. ..$ row     : int  19 48 58 59 60 61 62 64 67 79 ...
##  .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##  .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##  .. ..$ actual  : chr  "NULL" "NULL" "NULL" "NULL" ...
## $ NPT4_PUB       : atomic  13415 14805 NA 17520 11936 ...
##  ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  5881 obs. of  4 variables:
##  .. ..$ row     : int  3 8 11 13 14 17 19 23 24 25 ...
##  .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##  .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##  .. ..$ actual  : chr  "NULL" "NULL" "NULL" "NULL" ...
## $ NPT4_PRIV      : atomic  NA NA 7455 NA NA ...
##  ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  3051 obs. of  4 variables:
##  .. ..$ row     : int  1 2 4 5 6 7 8 9 10 12 ...
##  .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##  .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##  .. ..$ actual  : chr  "NULL" "NULL" "NULL" "NULL" ...
## $ COSTT4_A       : atomic  18888 19990 12300 20306 17400 ...
##  ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  3667 obs. of  4 variables:
##  .. ..$ row     : int  8 19 58 59 60 61 62 64 71 74 ...
##  .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##  .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##  .. ..$ actual  : chr  "NULL" "NULL" "NULL" "NULL" ...
## $ DEBT_MDN       : atomic  19500 16250 10500 16500 15854 ...
##  ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  1163 obs. of  4 variables:
##  .. ..$ row     : int  20 22 25 26 32 34 43 45 46 49 ...
##  .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##  .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##  .. ..$ actual  : chr  "PrivacySuppressed" "PrivacySuppressed" "PrivacySuppressed" "PrivacySuppress
## $ CUML_DEBT_P90  : atomic  50114 40000 40000 40750 45846 ...
##  ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  1586 obs. of  4 variables:
##  .. ..$ row     : int  25 43 45 49 65 67 81 87 93 117 ...
##  .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##  .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##  .. ..$ actual  : chr  "PrivacySuppressed" "NULL" "NULL" "PrivacySuppressed" ...
## $ mn_earn_wne_p10: atomic  35300 46300 42100 52700 30700 49100 31400 41500 36700 52100 ...
##  ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  2168 obs. of  4 variables:
##  .. ..$ row     : int  19 48 62 64 67 79 80 81 86 105 ...
##  .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##  .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##  .. ..$ actual  : chr  "NULL" "NULL" "NULL" "PrivacySuppressed" ...
## $ md_earn_wne_p10: atomic  31400 40300 38100 46600 27800 42400 27100 39700 34800 45400 ...
##  ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':  2168 obs. of  4 variables:
##  .. ..$ row     : int  19 48 62 64 67 79 80 81 86 105 ...
##  .. ..$ col     : int  NA NA NA NA NA NA NA NA NA NA ...
##  .. ..$ expected: chr  "a number" "a number" "a number" "a number" ...
##  .. ..$ actual  : chr  "NULL" "NULL" "NULL" "PrivacySuppressed" ...
```

```
##  $ region2       : chr  "Southeast" "Southeast" "Southeast" "Southeast" ...
```

```r
summary(college_debt)
```

```
##     INSTNM             STABBR            PREDDEG         HIGHDEG     
##  Length:7804        Length:7804        Min.   :0.000   Min.   :0.000  
##  Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000  
##  Mode  :character   Mode  :character   Median :2.000   Median :2.000  
##                                        Mean   :1.789   Mean   :2.176  
##                                        3rd Qu.:3.000   3rd Qu.:4.000  
##                                        Max.   :4.000   Max.   :4.000  
##                                                                       
##      region         LOCALE            CCUGPROF             HBCU          
##  Min.   :0.000   Length:7804        Length:7804        Length:7804       
##  1st Qu.:3.000   Class :character   Class :character   Class :character  
##  Median :5.000   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :4.621                                                           
##  3rd Qu.:6.000                                                           
##  Max.   :9.000                                                           
##                                                                          
##   WOMENONLY           RELAFFIL            ADM_RATE        SATVRMID    
##  Length:7804        Length:7804        Min.   :0.000   Min.   :290.0  
##  Class :character   Class :character   1st Qu.:0.552   1st Qu.:475.0  
##  Mode  :character   Mode  :character   Median :0.700   Median :515.0  
##                                        Mean   :0.682   Mean   :521.8  
##                                        3rd Qu.:0.834   3rd Qu.:555.0  
##                                        Max.   :1.000   Max.   :760.0  
##                                        NA's   :5584    NA's   :6503   
##     SATMTMID        SATWRMID        SAT_AVG             UG       
##  Min.   :310.0   Min.   :350.0   Min.   : 666.0   Min.   :    0  
##  1st Qu.:483.0   1st Qu.:470.0   1st Qu.: 971.8   1st Qu.:  137  
##  Median :520.0   Median :510.0   Median :1036.5   Median :  754  
##  Mean   :530.8   Mean   :521.2   Mean   :1056.7   Mean   : 2648  
##  3rd Qu.:565.0   3rd Qu.:559.0   3rd Qu.:1117.2   3rd Qu.: 2785  
##  Max.   :785.0   Max.   :755.0   Max.   :1534.0   Max.   :46834  
##  NA's   :6489    NA's   :7011    NA's   :6384     NA's   :2848   
##     NPT4_PUB       NPT4_PRIV       COSTT4_A        DEBT_MDN     
##  Min.   :-1643   Min.   :-1220   Min.   : 4157   Min.   :   333  
##  1st Qu.: 6320   1st Qu.:13132   1st Qu.:14143   1st Qu.:  7710  
##  Median : 8792   Median :18259   Median :22865   Median :  9833  
##  Mean   : 9584   Mean   :18072   Mean   :24354   Mean   : 11830  
##  3rd Qu.:12480   3rd Qu.:22485   3rd Qu.:30383   3rd Qu.: 15462  
##  Max.   :27199   Max.   :87570   Max.   :74473   Max.   :131335  
##  NA's   :5881    NA's   :3051    NA's   :3667    NA's   :1163    
##  CUML_DEBT_P90    mn_earn_wne_p10  md_earn_wne_p10    region2         
##  Min.   :   333   Min.   : 12300   Min.   :  8400   Length:7804       
##  1st Qu.: 14750   1st Qu.: 27300   1st Qu.: 24200   Class :character  
##  Median : 24317   Median : 34500   Median : 31200   Mode  :character  
##  Mean   : 25147   Mean   : 37184   Mean   : 33233                     
##  3rd Qu.: 33798   3rd Qu.: 43300   3rd Qu.: 39200                     
##  Max.   :131335   Max.   :250000   Max.   :250000                     
##  NA's   :1586     NA's   :2168     NA's   :2168                       
```

Using dynamic data within a typical classroom

Using the downloaded data, we start by applying a technique from the introductory curriculum to a research question of interest based on the College Scorecard data. College debt is of particular interest to many college students, but debt can be mediated by post-graduation income. To fully investigate the relationship between the variables, we provide both confidence and prediction intervals for both variables.

After calculating a few intervals, we show the intervals represented graphically and broken down by geographic region. Note that the visual representations are not a summary plot of the data, and we leave it open to the instructor to have the students engage more deeply with the many available variables.

Using the two variables measuring amount of debt of a typical (i.e., median) college graduate and median earning 10 years after matriculation, we create both confidence intervals and prediction intervals – keeping in mind that the observational unit is an academic institution. Note that the calculations below are for both confidence and prediction intervals. The confidence interval agglomerates institutions over the entire dataset; however, the prediction value is for a single *institution* (which is the observational unit). The analysis lends itself nicely to a conversation about confidence vs. prediction intervals as well as observational units as institution vs. as individual student. It is worth pointing out to the students that the prediction intervals likely hold more information related to their individual experiences than the confidence intervals. However, the unit of prediction is for an *institution*, and so the individual student debt and income is likely even more variable than shown here. Additionally, Figure 1 demonstrates the effect of samples size: consider the comparison of the Military intervals (one school) to the intervals for all of the US institutions (about 6000 schools). [Note: the intervals given in Figure 1 were created using an ANOVA model where the within variance is calculated across all regions, which is how the interval for military schools can be calculated. You may or may not want to bring that up with your students.]

The following R code uses the `mosaic` package to directly calculate both prediction and confidence intervals. Note the formula interface given by the tilde is described in detail here: http://rpruim.github.io/eCOTS2014/Workshop/Modeling.html.

```
require(mosaic)
debt_mod <- lm(DEBT_MDN~1, data = college_debt)
debt_fun <- mosaic::makeFun(debt_mod)
debt_fun()
```

```
##        1
## 11829.78
```

```
debt_fun(interval="confidence")
```

```
##        fit      lwr      upr
## 1 11829.78 11692.44 11967.13
```

```
debt_fun(interval="prediction")
```

```
##        fit      lwr      upr
## 1 11829.78 636.2383 23023.33
```

```
earn_mod <- lm(md_earn_wne_p10~1, data = college_debt)
earn_fun <- mosaic::makeFun(earn_mod)
earn_fun()
```

```
##        1
## 33232.59
```

```
earn_fun(interval="confidence")
```

```
##        fit      lwr      upr
## 1 33232.59 32864.78 33600.41
```

```r
earn_fun(interval="prediction")
```

```
##        fit      lwr      upr
## 1 33232.59 5616.893 60848.29
```

The intervals are interesting, but they might be even more interesting if broken down by region and shown visually. Note how much smaller the confidence intervals are from the prediction intervals! The difference indicates lots of variability across institutions and large sample sizes.

```r
#creating the models for building confidence and prediction intervals:
debtreg_mod <- lm(DEBT_MDN~as.factor(region), data = college_debt)
debtreg_fun <- makeFun(debtreg_mod)
earnreg_mod <- lm(md_earn_wne_p10~as.factor(region), data=college_debt)
earnreg_fun <- makeFun(earnreg_mod)

# creating a dataframe for holding the information needed to plot

worth <- data.frame(fit = double(),
                    lowerbound = double(),
                    upperbound = double(),
                    cost = character(),
                    type = character(),
                    regNum = character(),
                    regName = character(),
                    stringsAsFactors = FALSE)

worth[1,] <- c(debt_fun(interval="conf"), "debt", "conf", "all", "US (all)")
worth[2,] <- c(debt_fun(interval="pred"), "debt", "pred", "all", "US (all)")
worth[3,] <- c(earn_fun(interval="conf"), "earn", "conf", "all", "US (all)")
worth[4,] <- c(earn_fun(interval="pred"), "earn", "pred", "all", "US (all)")

for(i in 0:9){
  worth <- rbind(worth,
                 c(debtreg_fun(region=i,interval="conf"), "debt","conf",
                   i,college_debt[college_debt$region==i,]$region2[1]))
  worth <- rbind(worth,
                 c(debtreg_fun(region=i,interval="pred"), "debt","pred",
                   i,college_debt[college_debt$region==i,]$region2[1]))

  worth <- rbind(worth,
                 c(earnreg_fun(region=i,interval="conf"), "earn","conf",
                   i,college_debt[college_debt$region==i,]$region2[1]))
  worth <- rbind(worth,
                 c(earnreg_fun(region=i,interval="pred"), "earn","pred",
                   i,college_debt[college_debt$region==i,]$region2[1]))
  }

worth <- worth %>% mutate(fit = readr::parse_number(fit),
                          lowerbound = readr::parse_number(lowerbound),
                          upperbound = readr::parse_number(upperbound))
```

```r
pd <- position_dodge(width = 1)
ggplot(worth, aes(x=regName, y=fit)) +
  geom_point(aes(col=cost), position=pd, size=.8) +
  geom_errorbar(aes(ymin=lowerbound, ymax=upperbound, col=cost,
```

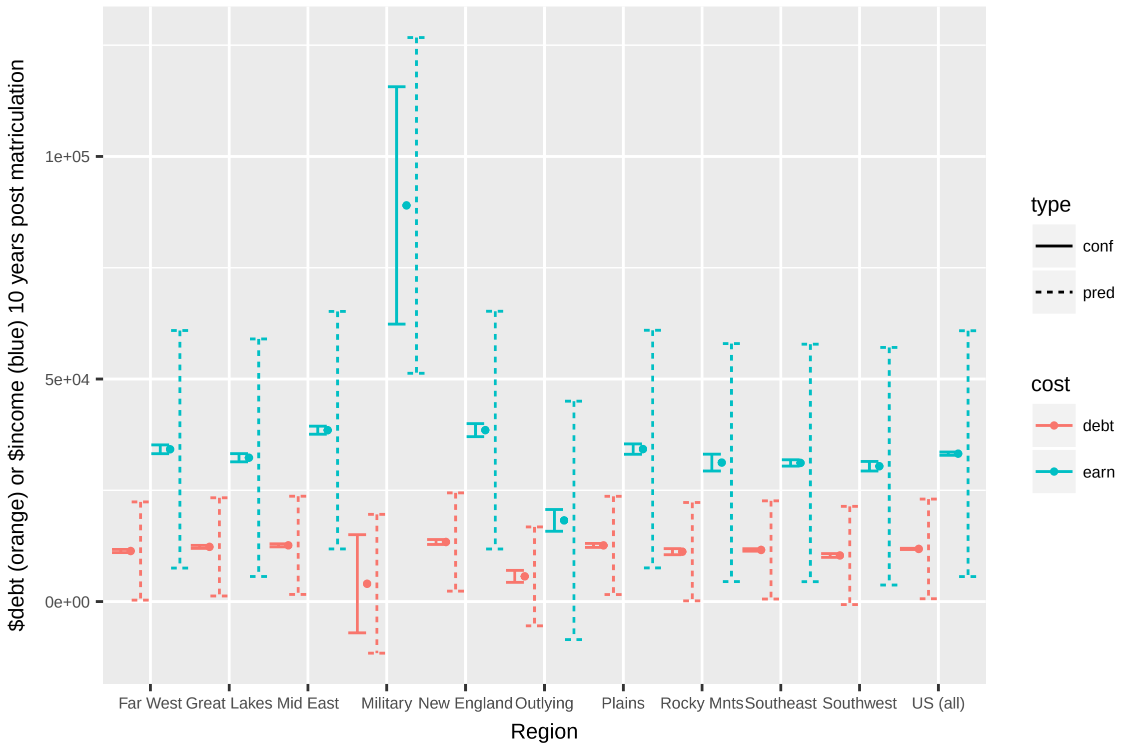

Figure 1: The x-axis represents the region of the institution. The y-axis represents either the amount of debt 10 years after matriculation (orange) or the amount of income 10 years after matriculation (blue). Confidence intervals for the average values (within region) are given by the solid lines. Prediction intervals for individual institutions are given by the dashed lines. The solid dot represents the center of both types of intervals (broken down by debt and income).

```
                  lty=type), position=pd) +
  xlab("Region") + ylab("$debt (orange) or $income (blue) 10 years post matriculation") +
  theme(text = element_text(size=8))
```

Thinking outside the box

The College Scorecard dataset is incredibly rich and can be used for many different types of model building: linear, logistic, machine learning. Indeed, thinking about interaction terms could be particularly insightful. Here, we give an example of regressing earnings on debt with the interaction term as whether or not the institution is one of the Historically Black Colleges and Universities (HBCU). Figure 2 displays the separate regression lines for the two distinct types of institutions.

```
college_debt_nona <- college_debt %>%
  dplyr::select(md_earn_wne_p10, DEBT_MDN, HBCU)
college_debt_nona <- college_debt_nona[complete.cases(college_debt_nona),]

earn_lm <- lm(md_earn_wne_p10 ~ DEBT_MDN*HBCU, data=college_debt_nona)
summary(earn_lm)
```

```
##
## Call:
## lm(formula = md_earn_wne_p10 ~ DEBT_MDN * HBCU, data = college_debt_nona)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32083  -6755   -413   4779 177289
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    18869.2571   371.0353  50.856  < 2e-16 ***
## DEBT_MDN           1.2119     0.0275  44.071  < 2e-16 ***
## HBCU1           1632.2247  3894.8953   0.419  0.67518
## DEBT_MDN:HBCU1    -0.6329     0.2239  -2.826  0.00473 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10880 on 4963 degrees of freedom
## Multiple R-squared:  0.2829, Adjusted R-squared:  0.2825
## F-statistic: 652.8 on 3 and 4963 DF,  p-value: < 2.2e-16
```

```
ggplot(college_debt_nona, aes(x=DEBT_MDN, y=md_earn_wne_p10, color=HBCU)) +
  geom_text(aes(DEBT_MDN,md_earn_wne_p10, label=toString(equation_end[1,-1])),
            data=data.frame(DEBT_MDN=25000, md_earn_wne_p10=180000, HBCU="0"))+
  geom_text(aes(DEBT_MDN,md_earn_wne_p10, label=toString(equation_end[2,-1])),
            data=data.frame(DEBT_MDN=25000, md_earn_wne_p10=160000, HBCU="1"))+
  geom_point(alpha=.25, size=.25) +
  geom_smooth(method="lm", fill=NA, lwd=.5) +
  xlab("Debt at Graduation") +
  ylab("Median income 10 years post matriculation")+
  theme(text = element_text(size=10))
```

Many interesting conversations can ensue based on the regression of income on debt. Reminding the students that each observation is an institution is an important starting point. Additionally, students should be able to volunteer the dangers of using a model like this to suggest causality. Last, there might be room to discuss an inferential analysis of whether HBCUs are statistically different from non-HBCUs (noting the substantial differences in sample sizes).

It is not hard to come up with additional questions to investigate with the College Scorecard data. Indeed, because the data relate directly to college students, they should be able to find many ways to engage with the data. We recommend continued conversations about how the data are valuable to the larger community, but that the information is not always complete (e.g., many variables are collected only on students who fill out financial aid forms) and not causative.

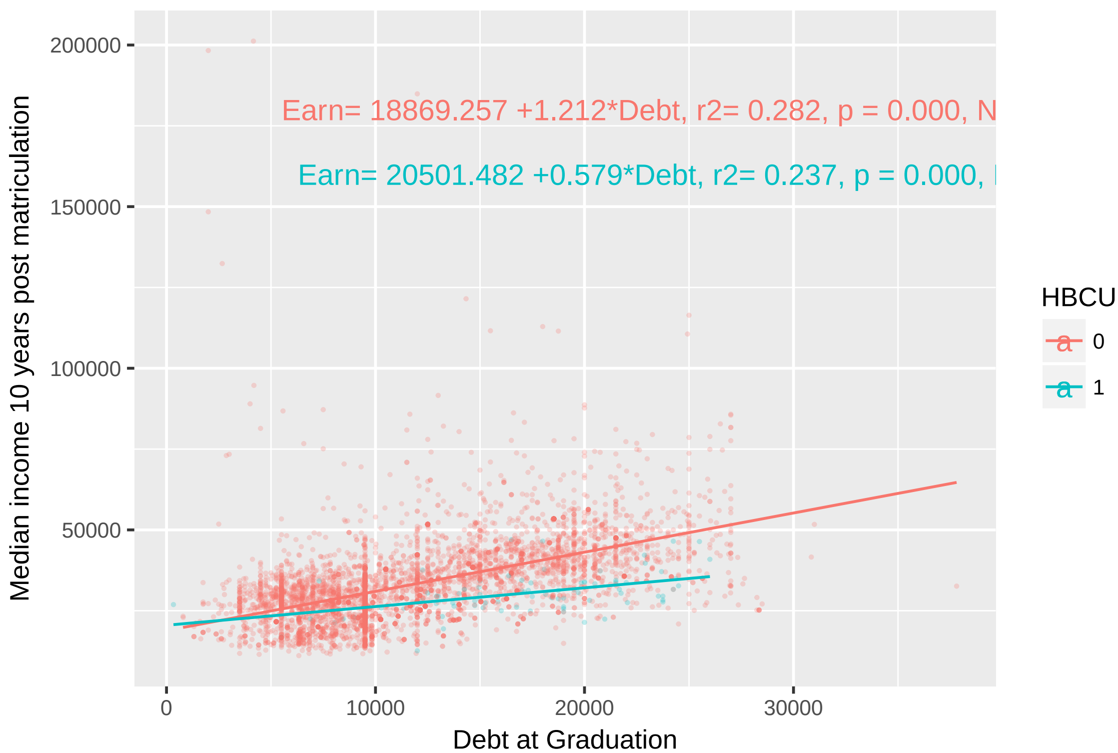

Figure 2: Median income regressed on debt. For the analysis, HBCU is interacted with debt to provide two distinct (and not parallel) regression lines. HBCU institutions are given in blue, and non-HBCU institutions are given in orange.