College Scorecard

June 2016

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
```

It's a really big dataset. 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 %>%
  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 = extract_numeric(ADM_RATE),
         SATVRMID = extract_numeric(SATVRMID),
         SATMTMID = extract_numeric(SATMTMID),
         SATWRMID = extract_numeric(SATWRMID),
         SAT_AVG = extract_numeric(SAT_AVG),
         UG = extract_numeric(UG),
         NPT4_PUB = extract_numeric(NPT4_PUB),
         NPT4_PRIV = extract_numeric(NPT4_PRIV),
         COSTT4_A = extract_numeric(COSTT4_A),
         DEBT_MDN = extract_numeric(DEBT_MDN),
         CUML_DEBT_P90 = extract_numeric(CUML_DEBT_P90),
```

```
        mn_earn_wne_p10 = extract_numeric(mn_earn_wne_p10),
        md_earn_wne_p10 = extract_numeric(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 U
##  $ 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       : num  0.899 0.867 NA 0.806 0.512 ...
##  $ SATVRMID       : num  410 580 NA 575 430 555 NA NA NA 570 ...
##  $ SATMTMID       : num  400 585 NA 580 425 570 NA NA NA 595 ...
##  $ SATWRMID       : num  NA NA NA NA NA 540 NA NA NA 565 ...
##  $ SAT_AVG        : num  823 1146 NA 1180 830 ...
##  $ UG             : num  4380 10331 98 5220 4348 ...
##  $ NPT4_PUB       : num  13415 14805 NA 17520 11936 ...
##  $ NPT4_PRIV      : num  NA NA 7455 NA NA ...
##  $ COSTT4_A       : num  18888 19990 12300 20306 17400 ...
##  $ DEBT_MDN       : num  19500 16250 10500 16500 15854 ...
##  $ CUML_DEBT_P90  : num  50114 40000 40000 40750 45846 ...
##  $ mn_earn_wne_p10: num  35300 46300 42100 52700 30700 49100 31400 41500 36700 52100 ...
##  $ md_earn_wne_p10: num  31400 40300 38100 46600 27800 42400 27100 39700 34800 45400 ...
##  $ region2        : chr  "Southeast" "Southeast" "Southeast" "Southeast" ...
```

```
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

The *mosaic* package formats most data analysis in terms of formulas. The formulas make it clear to the user which variable is the response variable and which is the predictor variable. The formulas also make it straightforward to include additional information to realize further nuances of the underlying relationships.

In this analysis, we will find univariate confidence intervals for amount of debt after graduation, to be compared with earnings 10 years out. Note that the calculations are for both confidence and prediction intervals. However, the prediction value is for an *institution* (which is the observational unit). The analysis below lends itself nicely to a conversation about confidence vs. prediction intervals as well as observational units as institution vs. as individual student. Additionally, the plot below demonstrates the effect of samples

size: consider the comparison of the Military intervals (1 school) to the intervals for all of the US institutions (about 6000 schools).

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

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

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

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

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

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

The prediction intervals are interesting, but 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.

```
#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)
```

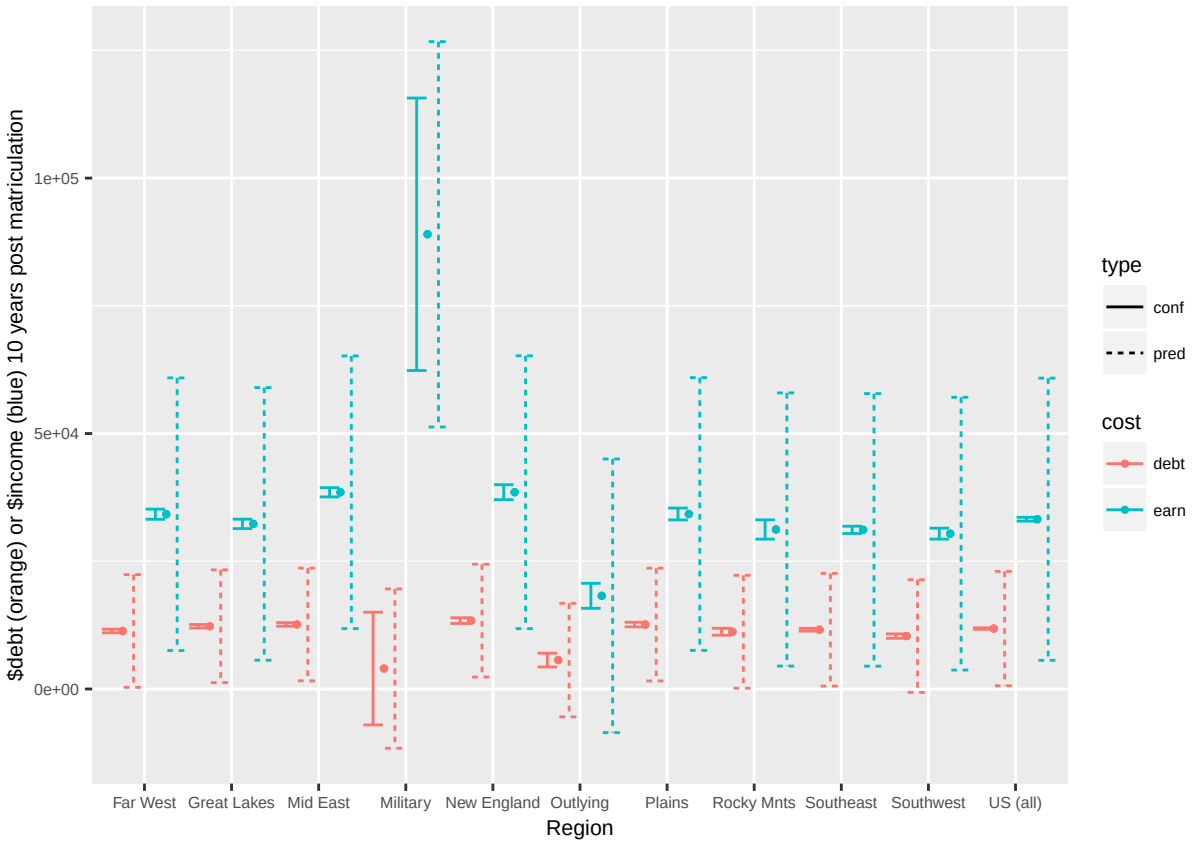
```
worth[1,] <- c(debt_fun(interval="conf"), "debt", "conf", "all", "US (all)")
worth[2,] <- c(debt_fun(interval="pred"), "debt", "pred", "all", "US (all)")
worth[3,] <- c(earn_fun(interval="conf"), "earn", "conf", "all", "US (all)")
worth[4,] <- c(earn_fun(interval="pred"), "earn", "pred", "all", "US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

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,
                    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 dataset is incredibly rich and can be used for a lot of model building: linear, logistic, machine learning. Indeed, thinking about interaction terms could be particularly insightful. Below, we give an example of the variables above with the interaction term as whether or not the institution is one of the Historically Black Colleges and Universities (HBCU).

```
college_debt_nona <- college_debt %>%
  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))
```