College Scorecard Cohort Analysis

Introduction

This analysis explores the possibility of building on CEW's college ROI work by using all prior College Scorecard datasets. CEW's current ROI methodology only considers the "most recent cohorts" data file, but key variables (such as earnings and debt measures) are also reported for prior cohorts. For the purposes of this analysis, I consider median earnings by cohort 6 years, 8 years, and 10 years after the cohort enters college.

There are three sections to this report:

1. Data Processing: Details the creation of the dataset used for analysis
2. Output Creation: Details the process of creating the graphics used for analysis.
3. Analysis: Provides a qualitative review of the output.

Data Processing

Environment Setup

I use three key sources of data:

1. College Scorecard historical data (available at https://collegescorecard.ed.gov/data/ under the "All Data Files" heading.
2. The BLS CPI "R-CPI-U-RS" (less food and energy) series, available at https://www.bls.gov/cpi/research-series/r-cpi-u-rs-home.htm.
3. A crosswalk matching a given survey to the cohorts whose earnings are being measured (most surveys follow three distinct cohorts between 6 and 10 years after initial college enrollment).[1] The crosswalk file (referred to as "mm_cohort_crosswalk.csv" in the code below) is reproduced below, along with an additional "inflation adjustment" column which shows the terms in which each vintage's earnings are reported:

Source: College Scorecard Data Dictionary, "Institution_Cohort_Map" and "Most_Recent_Inst_Cohort_Map" tabs.

file_name	p6_cohort	p8_cohort	p10_cohort	inflation adjustment
MERGED2003_04_PP.csv	96-97/97-98			2014 dollars
MERGED2005_06_PP.csv	98-99/99-00	96-97/97-98		2014 dollars

file_name	p6_cohort	p8_cohort	p10_cohort	inflation adjustment
MERGED2007_08_PP.csv	00-01/01-02	98-99/99-00	96-97/97-98	2014 dollars
MERGED2009_10_PP.csv	02-03/03-04	00-01/01-02	98-99/99-00	2014 dollars
MERGED2011_12_PP.csv	04-05/05-06	02-03/03-04	00-01/01-02	2014 dollars
MERGED2012_13_PP.csv	05-06/06-07	03-04/04-05	01-02/02-03	2015 dollars
MERGED2013_14_PP.csv	06-07/07-08	04-05/05-06	02-03/03-04	2016 dollars
MERGED2014_15_PP.csv	07-08/08-09	05-06/06-07	03-04/04-05	2017 dollars
MERGED2018_19_PP.csv	11-12/12-13	09-10/10-11	07-08/08-09	2020 dollars
MERGED2019_20_PP.csv	12-13/13-14	10-11/11-12	08-09/09-10	2021 dollars
Most-Recent-Cohorts-Institution.csv	12-13/13-14	10-11/11-12	08-09/09-10	2021 dollars

Below is the R code that sets up the analysis.[2] After identifying which College Scorecard vintages reported earnings data, I created a list of those files ("KEYFILES") and a list of the earnings variables of interest ("KEYVARS"). The "cohorts" data.frame corresponds to the table above.

```
rm(list = ls())
options(scipen = 999)

# Set working directory
#setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
#setwd("..")

# Load packages
library(data.table)
library(tidyverse)
library(magrittr)
library(lubridate)
library(stringr)
library(janitor)
library(openxlsx)

# Define Graph theme
GraphTheme = theme_classic() +
  theme(
    axis.line = element_line(linewidth = 1, colour = "black"),
    panel.grid.major.y = element_line(linewidth = 0.4, color = "grey89"),
    axis.title.x = element_text(size = 14),
    axis.text.x = element_text(size = 12, color = "black"),
    axis.title.y = element_text(size = 13),
    axis.text.y = element_text(size = 12, color = "black"),
    #legend.title = element_blank(),
    legend.key.size = unit(0.8, 'cm'),
    legend.background = element_blank())

# List years of interest
KEYYRS <- c(2003,2005,2007,2009,2011:2014,2018,2019)

# List key files
KEYFILES <- c(paste0("MERGED",KEYYRS,"_",
                     str_sub(as.character(KEYYRS+1),-2),"_PP.csv"),
              "Most-Recent-Cohorts-Institution.csv")

# Key variables
KEYVARS <- c(paste0("md_earn_wne_p",c(6,8,10)))

# Cohort crosswalk
cohorts <- fread("../input/mm_cohort_crosswalk.csv")
```

Key Variable Extraction

Each vintage of College Scorecard's underlying data files contains thousands of variables. In order to efficiently extract only the variables of interest, the following code defines a function ("CLN_YRS") and loops through every relevant data file to standardize and append the data of interest.

```
# Container dataset
allyrs <- data.frame()

# Function for processing individual datasets
CLN_YRS <- function(x) {

  # Download file
  individ_yr <- fread(paste0("../input/",KEYFILES[x])) %>%
    clean_names() %>%
    select(unitid, opeid, opeid6, preddeg, control, st_fips,
           ccugprof, ccsizset, ug, costt4_a, costt4_p, all_of(KEYVARS)) %>%
    # Add years & file name
    mutate(year = paste0(KEYYRS[x],"_",str_sub(as.character(KEYYRS[x]+1),-2)),
           file_name = KEYFILES[x]) %>%
    # Clean earnings variable
    mutate(across(all_of(KEYVARS), as.numeric)) %>%
    # Set all other variables to character
    mutate(across(!all_of(KEYVARS), as.character))

  # Add to overall dataset
  allyrs <- bind_rows(allyrs, individ_yr)

  return(allyrs)
}
```

```
# Apply function
combined_data <- lapply(1:length(KEYFILES), CLN_YRS) %>%
  bind_rows() %>%
  # clean year and add cohorts
  mutate(year = ifelse(year == "NA_NA", "2019_20", year))

rm(allyrs)
```

Finalizing the Dataset

After all of the raw data is collated, the next step is to match the earnings reported in each file to their corresponding cohorts based on the crosswalk. This is done by the following code:

```
full_data <- combined_data %>%
  left_join(cohorts) %>%
  mutate(across(c(p6_cohort, p8_cohort, p10_cohort),
                ~factor(.x,
                        levels = c("96-97/97-98", "98-99/99-00", "00-01/01-02",
                                   "01-02/02-03", "02-03/03-04", "03-04/04-05",
                                   "04-05/05-06", "05-06/06-07", "06-07/07-08",
                                   "07-08/08-09", "08-09/09-10", "09-10/10-11",
                                   "10-11/11-12", "11-12/12-13", "12-13/13-14"))))
```

```
Joining with `by = join_by(file_name)`
```

Finally, I put all earnings in terms of 2023 dollars using the CPI index and the information on the dollar terms used by each vintage of College Scorecard data.

```
# Read in CPS dataset (All items less food & energy)
cps <- read.xlsx("../input/r-cpi-u-rs-alllessfe.xlsx", sheet = 1,
                 startRow = 6) %>%
  clean_names() %>%
  select(year, avg) %>%
  rename(cps_yr = year,
         cps_deflator = avg)

# 2023 CPS avg
Y2023 <- cps[nrow(cps),2]

# Merge into full dataset
full_data_cps <- full_data %>%
  mutate(cps_yr = case_when(
    year == "2012_13" ~ 2015,
    year == "2013_14" ~ 2016,
    year == "2014_15" ~ 2017,
    year == "2018_19" ~ 2020,
    year == "2019_20" ~ 2021,
    T                 ~ 2014)) %>%
  left_join(cps) %>%
  mutate(across(all_of(KEYVARS), ~.x*Y2023/cps_deflator,
                .names = "{.col}_23_dollars"))
```

```
Joining with `by = join_by(cps_yr)`
```

The data is saved in the “intermediate” subfolder.

```
          used (Mb) gc trigger  (Mb) max used  (Mb)
Ncells 1329008 71.0    3059522 163.4  3059522 163.4
Vcells 7046537 53.8   44519895 339.7 46898747 357.9
```

Output Creation

Of primary interest in this analysis is the question of how much value is added to the overall ROI project by creating this time series. This question is multifaceted; there are many ways of trying to answer it using the dataset created above:

- We could check earnings variances by individual colleges’ cohorts. It is possible that some colleges have very stable ROIs and others do not – we might guess that Ivy Leagues would have more stable ROIs than other types of colleges, for example.
- We could check whether certain types of colleges have different trends in earnings (e.g., do 4-year institutions who primarily provide BAs fare differently over time than other types of colleges?).
- We could do a regional analysis to see if there are heterogeneous ROI outcomes by state/region.

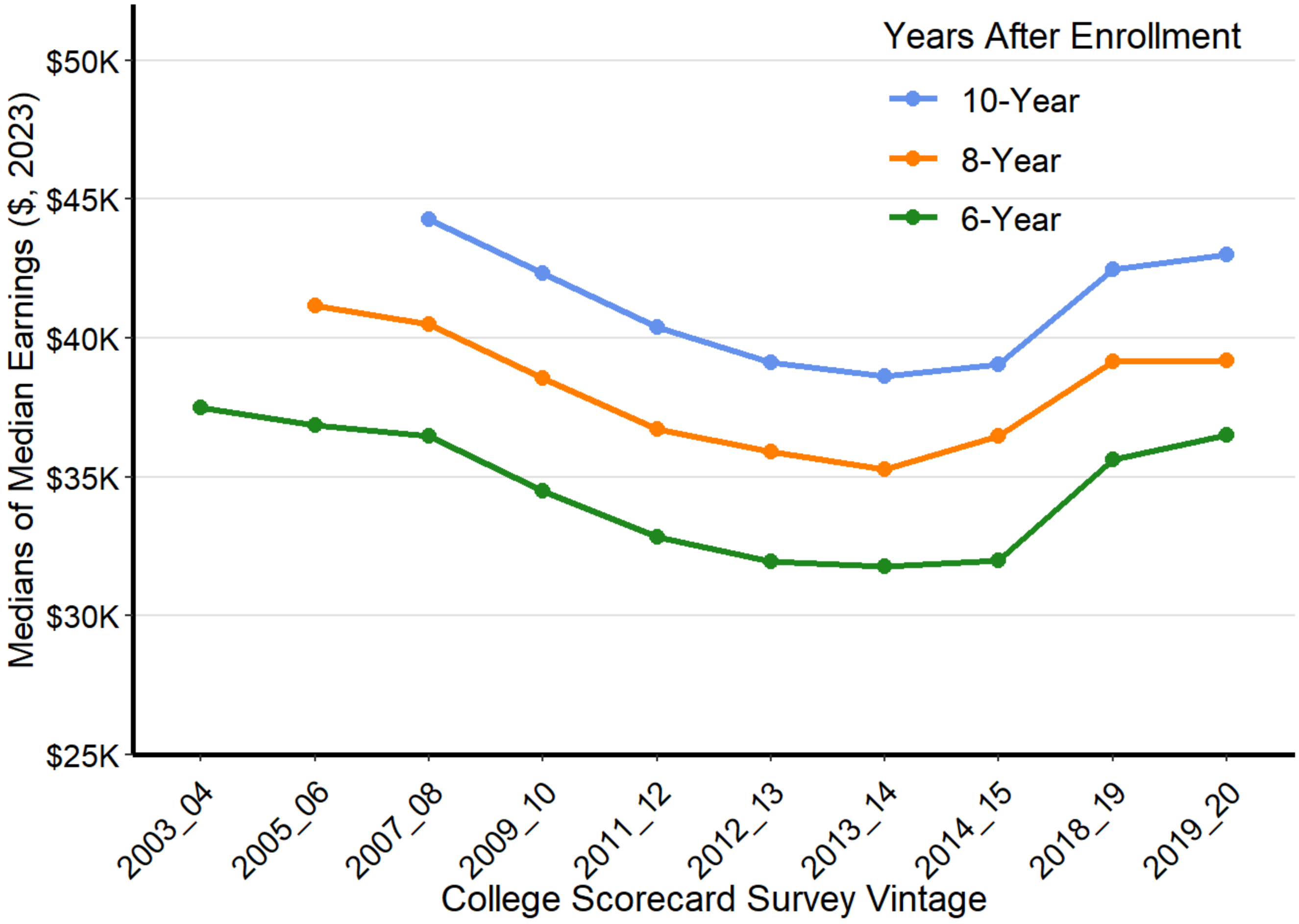
For the purposes of this initial exploratory analysis, we will look at overall time trends across all colleges. Specifically, we will consider **median earnings reported by survey year** and **median earnings by individual cohort**.

Median Earnings Reported by Survey Year

This analysis is fairly straightforward: I simply take the reported median earnings for each College Scorecard data vintage and plot the results for each of the three cohorts within that data, without identifying any individual cohort across time. The resulting graph (“Figure 1” hereafter) is below:

```
# Set up data
fig1 <- full_data_cps %>%
  group_by(year) %>%
  summarize('6-Year' = median(md_earn_wne_p6_23_dollars, na.rm=T)/1000,
            '8-Year' = median(md_earn_wne_p8_23_dollars, na.rm = T)/1000,
            '10-Year'= median(md_earn_wne_p10_23_dollars, na.rm = T)/1000) %>%
  ungroup() %>%
  pivot_longer(!year, names_to = "cohort", values_to = "earnings") %>%
  mutate(cohort = factor(cohort, levels = c("6-Year", "8-Year", "10-Year"))) %>%
  filter(!is.na(earnings))

# Create plot
ggplot(fig1, aes(x = year, y = earnings, group = cohort, color = cohort)) +
  geom_point(size= 2.5) +
  geom_line(linewidth = 1.2) +
  scale_color_manual(values = c('cornflowerblue', 'darkorange1',
                                'forestgreen'),
                     breaks = c("10-Year", "8-Year", "6-Year")) +
  GraphTheme +
  theme(legend.position = c(0.8,0.83),
        legend.text = element_text(size = 13),
        legend.title= element_text(size=14),
        axis.text.x = element_text(angle = 45, vjust = 0.9, hjust = 1)) +
  labs(color = "Years After Enrollment",
       x = "College Scorecard Survey Vintage",
       y = "Medians of Median Earnings ($, 2023)") +
  scale_y_continuous(expand = c(0,0),
                     limits = c(25, 52),
                     labels = scales::dollar_format(prefix = "$",
                                                    suffix = "K"))
```



```
# Save output
#ggsave("output/earnings_by_survey_yr.png",
#       width = 3500, height = 2750, units = "px")
```

Median Earnings by Individual Cohort

This analysis is a bit more complex, as it requires creating an individual dataset for each cohort and then combining the results. This process is given in the code below:

```
# Create list of unique cohorts
cohort_list_int <- full_data_cps %>%
  select(year, ends_with("_cohort")) %>%
  pivot_longer(!year, names_to = "drop", values_to = "cohort") %>%
  filter(cohort!= "")

cohort_list = unique(cohort_list_int$cohort)

# Create data for each cohort
COHORT <- function(x) {

  indiv_cohort <- full_data_cps %>%
    filter(p6_cohort == x | p8_cohort == x | p10_cohort == x) %>%
    mutate(cohort = x,
      cohort_earnings = case_when(
      p6_cohort == x  ~ md_earn_wne_p6_23_dollars,
      p8_cohort == x  ~ md_earn_wne_p8_23_dollars,
      p10_cohort == x ~ md_earn_wne_p10_23_dollars)) %>%
    mutate(cohort_yr = case_when(
      p6_cohort == cohort  ~ "P6",
      p8_cohort == cohort  ~ "P8",
      p10_cohort == cohort ~ "P10")) %>%
    select(year, cohort, cohort_yr, cohort_earnings) %>%
    filter(!is.na(cohort_earnings))

  return(indiv_cohort)
}
```

Once all the cohorts are identified, we can track any individual cohort across College Scorecard data vintages. Some of the cohorts were tracked across multiple vintages, but several were not. The following graph only considers those cohorts that were considered across more than one vintage.

Note that the x-axis refers to the survey vintage, while each individual line corresponds to a given cohort. The labels ("P6", "P8", and "P10") refer to the type of earnings measured (e.g., "P6" refers to earnings measured 6 years after cohort entry into college).

As an example, consider the leftmost, red line. This line corresponds to the median earnings of the 96-97/97-98 cohort. In the 2003/2004 College Scorecard survey, the median earnings (6 years after entering college) were $37.5K. Two years later, in the 2005/2006 College Scorecard Survey, the median earnings (now 8 years after entering college) were $41.2K.

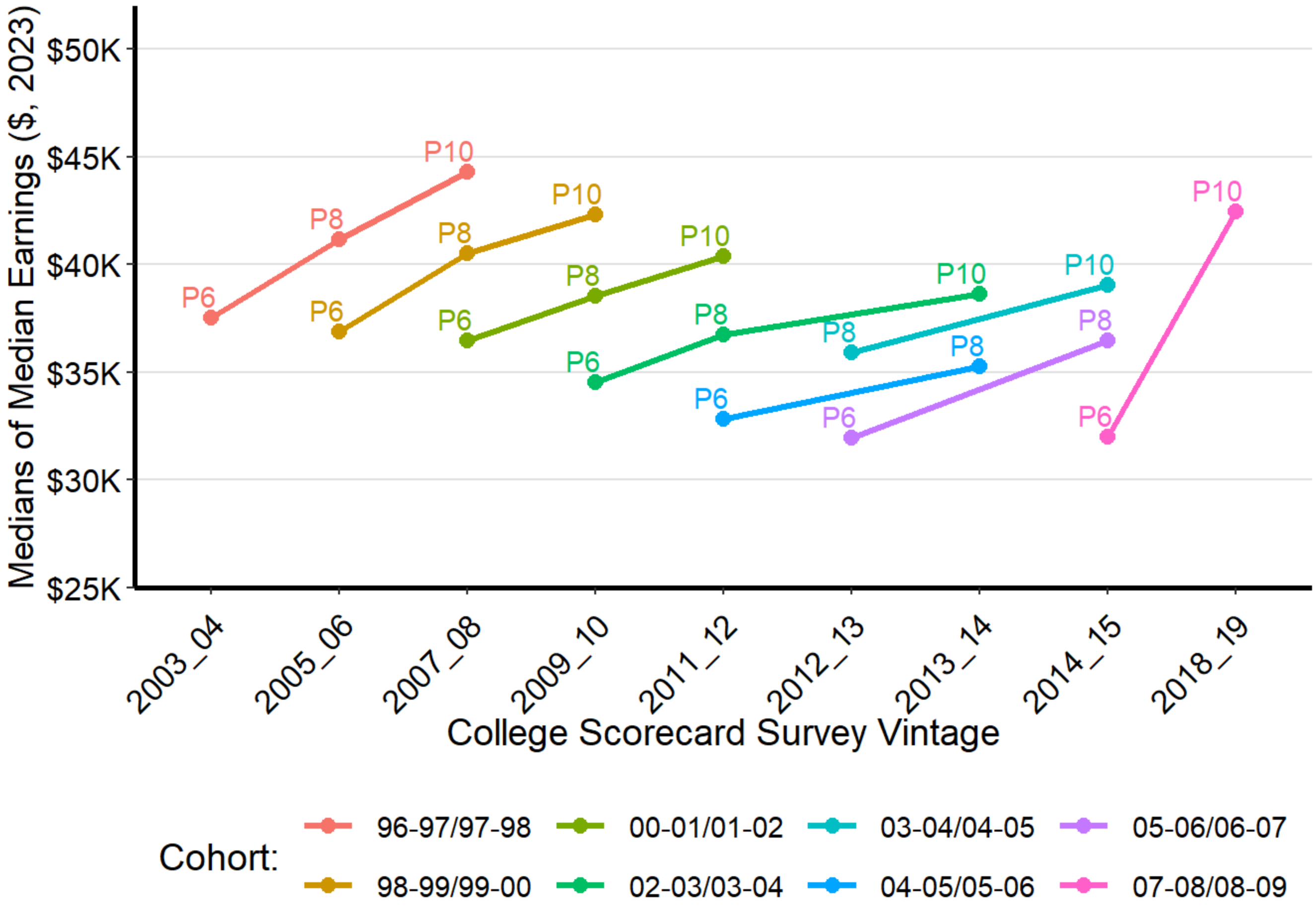
The full graph ("Figure 2" hereafter) is below:

```
# Create cohort data
fig2 <- lapply(cohort_list, COHORT) %>% bind_rows() %>%
  group_by(year, cohort, cohort_yr) %>%
```

```r
  summarize(earnings = median(cohort_earnings)/1000) %>%
  ungroup() %>%
  mutate(cohort =
           factor(cohort,
                  levels = c("96-97/97-98", "98-99/99-00", "00-01/01-02",
                             "01-02/02-03", "02-03/03-04", "03-04/04-05",
                             "04-05/05-06", "05-06/06-07", "06-07/07-08",
                             "07-08/08-09", "08-09/09-10", "09-10/10-11",
                             "10-11/11-12", "11-12/12-13", "12-13/13-14"))) %>%
  group_by(cohort) %>%
  mutate(count = n()) %>%
  ungroup()
```

```
`summarise()` has grouped output by 'year', 'cohort'. You can override using
the `.groups` argument.
```

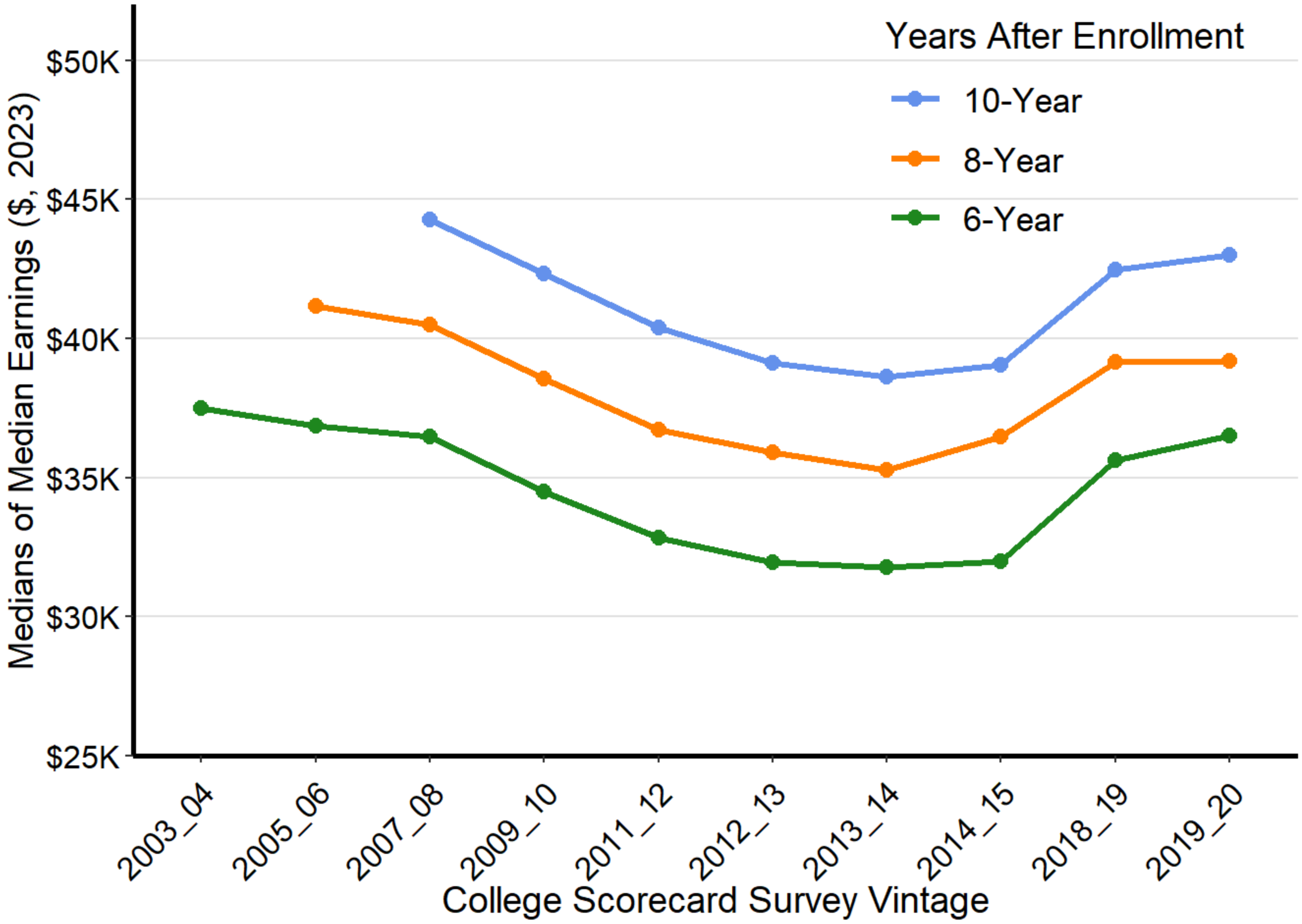
```r
# Graph data
ggplot(fig2 %>% filter(count>1), aes(x = year, y = earnings,
                 group = cohort, color = cohort)) +
  geom_point(size= 2.5) +
  geom_text(aes(x = year, y = earnings, label = cohort_yr),
            vjust =-0.5, hjust = 0.85, show.legend = F) +
  guides(color = guide_legend(nrow =2, title = "Cohort:"))+
  geom_line(linewidth = 1.2) +
  GraphTheme +
  theme(legend.position = "bottom",
        legend.direction = "horizontal",
        legend.text = element_text(size = 11),
        legend.title=element_text(size=14),
        axis.text.x = element_text(angle = 45, vjust = 0.9, hjust = 1)) +
  labs(color = "Cohort: ",
       x = "College Scorecard Survey Vintage",
       y = "Medians of Median Earnings ($, 2023)") +
  scale_y_continuous(expand = c(0,0),
                     limits = c(25, 52),
                     labels = scales::dollar_format(prefix = "$",
                                                    suffix = "K"))
```



```
# Save output
#ggsave("output/earnings_by_cohort.png",
#       width = 3750, height = 2750, units = "px")
```

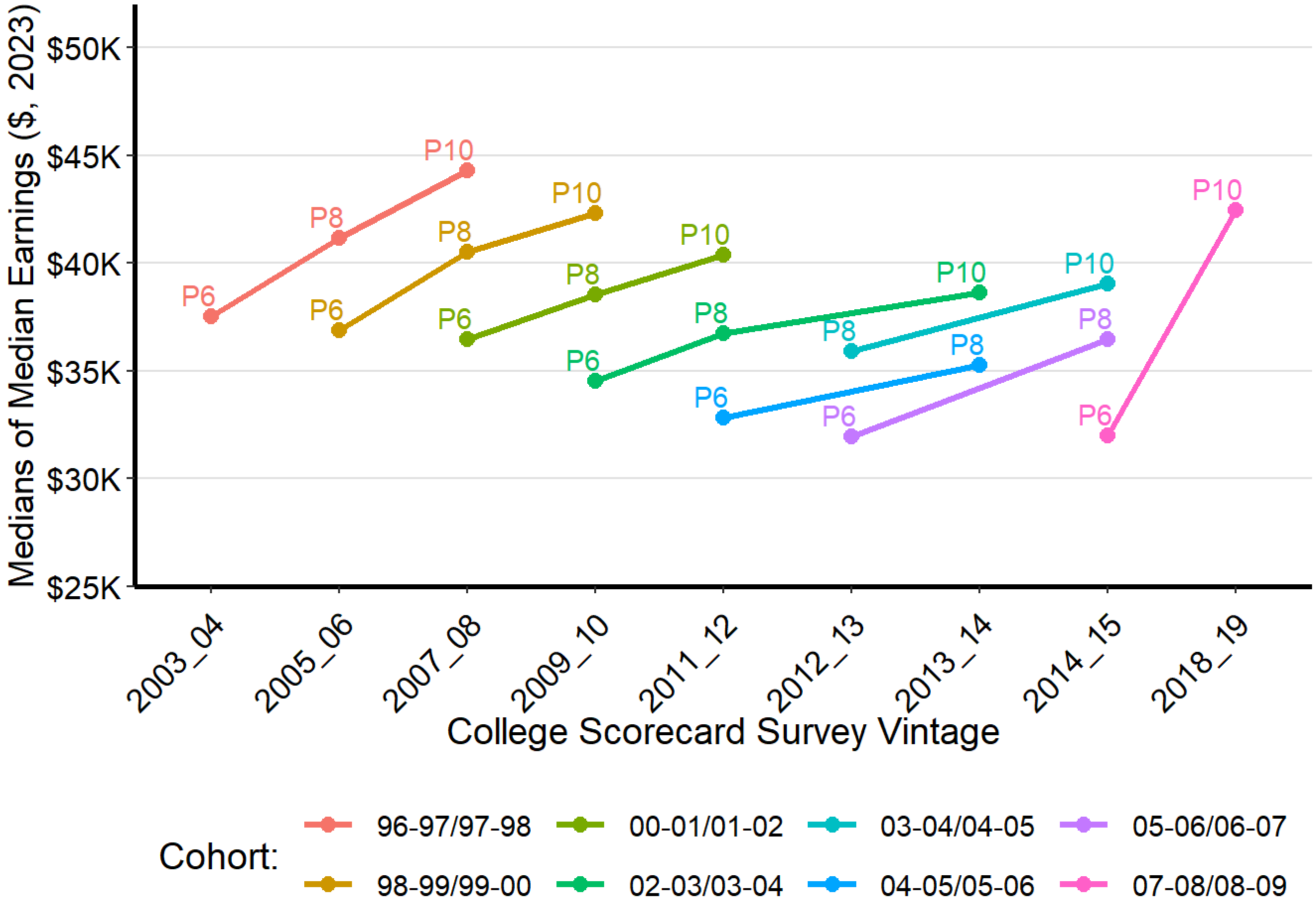
Analysis

Figure 1 (reproduced below) shows that earnings for all cohorts fell during the Great Recession, reaching a nadir in the 2013/2014 College Scorecard vintage. After beginning a slow rise the year after, the next available data (2018/2019 vintage) shows a marked rise, especially in the 6-Year category. This particular cohort entered college in the 2012/13 and 2013/2014 academic years (referred to in the data as the “12-13/13-14” cohort). However, by 2019/2020, the reported median earnings were still below pre-Recession levels.



The overall trend holds regardless of cohort, which suggests that the effect of general, negative labor market conditions had a stronger impact on earnings than the effect of tenure post-recession, even for those cohorts that graduated and established careers before the recession.

Figure 2 (below) considers earnings by individual cohort:



One striking trend is a steady decline in the positive effect of labor market tenure, measured by the slope of each line, with the only marked difference coming from the 07-08/08-09 cohort. It appears that the most severely-affect cohort was the 04-05/05-06 cohort, which saw median earnings rise less than $2,500 in two years.

The positive relationship between tenure and earnings is strongest for the 07-08/08-09 cohort, but there was also a small increase for the 05-06/06-07 cohort (most of whom were graduating during the tail end of the Great Recession), suggesting that the decline in the positive effect of labor market tenure was temporary.

Footnotes

1. For example, the 2019-2020 version of the College Scorecard data measures the earnings of three separate cohorts: (1) the cohort who entered college in AY2008/09-AY2009/10, (2) the cohort who entered college in AY2010/11-AY2011/12, and the cohort who entered college in AY2012/13-AY2013/14.↩
2. Within the main analysis folder, I have four key subfolders: (1) “code” contains the R code files, (2) “input” contains all the raw college scorecard data, as well as the CPI series and the crosswalk file, and (3) “intermediate” contains intermediate files (e.g., the cleaned dataset), and (4) “output” is where final graphs/tables/results are stored.↩