Problem Set 1. Visualizing Life Course Inequality

Info 3370. Studying Social Inequality with Data Science. Spring 2023

Due: 5pm on 10 Feb 2023.

Welcome to the first problem set!

- Use this .Rmd template to complete the problem set
- If you want to print the assignment, here is a pdf
- In Canvas, you will upload the PDF produced by your .Rmd file
- Don't put your name on the problem set. We want anonymous grading to be possible
- We're here to help! Reach out using Ed Discussion or Office Hours

This problem set involves both reading and data analysis.

Reading for this the problem set

Everyone will read p.~1–7 of following paper. Stop before the section "Analytic Framework for Decomposing Inequality."

> Cheng, Siwei. 2021. The Shifting Life Course Patterns of Wage Inequality.. *Social Forces* 100(1):1–28. https://doi.org/10.1093/sf/soab003

Graduate students will also read this paper for the bonus question

> DiPrete, Thomas A., & Eirich, Gregory M. 2006. Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology* 32:271-297. https://doi.org/10.1146/annurev.soc.32.061604.123127

Data analysis

This problem set uses the data `lifeCourse.csv`.

```
library(tidyverse)
library(scales)
lifeCourse <- read_csv("https://info3370.github.io/assets/data/lifeCourse.csv")
```

The data contains information on the life course earnings profiles for four cohorts of American workers: those born in 1940, 1950, 1960, and 1970. Each row contains a summary of the annual earnings distribution for a particular birth cohort at a particular age, among the subgroup with a particular level of education. To prepare these data, we aggregated microdata from the Current Population Survey, provided through the Integrated Public Use Microdata Series.

The data contain five variables.

1. `quantity` is the metric by which the earnings distribution is summarized: 10th, 50th, or 90th percentile
2. `education` is the educational subgroup being summarized: College Degree, Less than College
3. `cohort` is the cohort (people with a given birth year) to which these data apply: 1940, 1950, 1960, 1970
4. `age` is the age at which earnings were measured: 30–45
5. `income` is the value for the given earnings percentile in the given subgroup. Income values are provided in 2022 dollars

1. Visualize the data (20 points)

Use `ggplot` to visualize these data. To denote the different trajectories,

- use `color` for `quantity`
- use `facet_grid` to make a panel of facets where each row is an education value and each column is a cohort value

- Hint: See the class website, Ed Discussion, and R4DS 3 for help on getting started on the graph.

Modify the axis titles and labels as appropriate to make the visualization easy to read.

```
# your code goes here
```

2. Interpret what you see

Write 2-3 sentences summarizing the trends that you see in the data.

2.1 (5 points). How does the distribution of incomes change over the life course for those with a college degree? Explain your observations for at least 2 percentiles and compare the percentiles overall.

Type your answer here.

2.2 (5 points). How does the pattern differ for those with less than a college degree? Explain your observations for at least 1 percentile and compare the percentiles overall.

Type your answer here.

2.3 (5 points). How do these patterns (such as inequality) change across cohorts?

Type your answer here.

3. Connect to theories (15 points)

This section involves p. 1–7 of the Cheng (2021) paper referenced at the top of the assignment. Our data are not the same as Cheng's. But our analysis is able to reproduce many of her findings. Answer each question in two sentences or less.

3.1 (3 points) Cheng discusses period trends, cohort trends, and age trends. Which two of these is visually apparent in your graph, and where in the graph do you look to see them?

Type your answer here.

3.2 (3 points) Define the intragenerational process of stratification, for someone who has never heard of it.

Type your answer here.

3.3 (3 points) Cheng discusses education-based cumulative advantage. Describe how you see this in your graph.

Type your answer here.

3.4 (3 points) Cheng discusses within-education trajectory heterogeneity. Describe how you see this in your graph.

Type your answer here.

3.5 (3 points) Cheng discusses wage volatility. Our data doesn't speak to this concept. What kind of data would we need to study this concept?

Type your answer here.

Bonus question

- For graduate students, this question is worth 20 points.
- For undergraduate students, this question is optional and worth 0 points.

The growth of inequality over the life course is a descriptive pattern that corresponds to a widespread theory of inequality: cumulative advantage. By 'theory', I mean an abstraction that applies to many social processes.

To complete this section, read DiPrete and Eirich (2006), referenced at the top of the assignment.

B.1 (5 points) Give a canonical example of the strict form of cumulative advantage in the Mertonian sense.

Type your answer here.

B.2 (5 points) The visualization above is descriptively consistent with cumulative advantage, but the data provided to you cannot provide a direct test of the strict form of the theory. Why not?

Type your answer here.

B.3 (10 points) Propose a research question about cumulative advantage. Is your proposal cumulative advantage in the strict sense or the descriptive sense? What data would you want in order to empirically investigate this question? Answer this question in fewer than 200 words.

Type your answer here.

Computing environment

Leave this at the bottom of your file, and it will record information such as your operating system, R version, and package versions. This is helpful for resolving any differences in results across people.

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] scales_1.2.1    forcats_1.0.0   stringr_1.5.0   dplyr_1.1.0
##  [5] purrr_1.0.1     readr_2.1.3     tidyr_1.3.0     tibble_3.1.8
##  [9] ggplot2_3.4.0   tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.0    xfun_0.36           haven_2.5.1
##  [4] gargle_1.3.0        colorspace_2.1-0    vctrs_0.5.2
##  [7] generics_0.1.3      htmltools_0.5.4     yaml_2.3.7
## [10] utf8_1.2.3          rlang_1.0.6         pillar_1.8.1
## [13] glue_1.6.2          withr_2.5.0         DBI_1.1.3
## [16] bit64_4.0.5         dbplyr_2.3.0        modelr_0.1.10
## [19] readxl_1.4.1        lifecycle_1.0.3     munsell_0.5.0
## [22] gtable_0.3.1        cellranger_1.1.0    rvest_1.0.3
## [25] evaluate_0.20       knitr_1.42          tzdb_0.3.0
## [28] fastmap_1.1.0       curl_5.0.0          parallel_4.2.2
## [31] fansi_1.0.4         broom_1.0.3         backports_1.4.1
## [34] googlesheets4_1.0.1 vroom_1.6.1         jsonlite_1.8.4
## [37] bit_4.0.5           fs_1.6.0            hms_1.1.2
```

```
## [40] digest_0.6.31       stringi_1.7.12      grid_4.2.2
## [43] cli_3.6.0           tools_4.2.2         magrittr_2.0.3
## [46] crayon_1.5.2        pkgconfig_2.0.3     ellipsis_0.3.2
## [49] xml2_1.3.3          reprex_2.0.2        googledrive_2.0.0
## [52] lubridate_1.9.1     timechange_0.2.0    assertthat_0.2.1
## [55] rmarkdown_2.20      httr_1.4.4          rstudioapi_0.14
## [58] R6_2.5.1            compiler_4.2.2
```