Week 9 Methods: Time

ECON 125: The Science of Population

Setup

Today, we analyze data on manufacturing plants in Indonesia:

- Revenue, value added, capital, materials, and workers
- One observation per plant (called a "firm" in the data) per year
- Manufacturing Survey of Large and Medium-Sized Firms (*Statistik Industri*)

Start by setting up R and loading the data

```r
# Load tidyverse and clear the R environment
library(tidyverse)
rm(list=ls())

# Load dataset
load(url("https://github.com/tomvogl/econ125/raw/main/data/indonesia_fi

# Ask R to not use scientific notation
options(scipen = 999)
```

Glimpse

```r
glimpse(indonesia_firms)
```

```
## Rows: 219,170
## Columns: 14
## $ province        <dbl> 12, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,
## $ year            <dbl> 1994, 1990, 1990, 1990, 1991, 1992, 1993, 19
## $ psid            <dbl> 0, 1750, 1751, 1752, 1752, 1752, 1752, 1752,
## $ product         <fct> "Food, beverages, and tobacco", "Food, bever
## $ workers         <dbl> 145, 138, 125, 129, 135, 141, 118, 133, 129,
## $ materials       <dbl> 5071688.1, 5292000.0, 5401293.0, 4801825.0,
## $ capital         <dbl> 391000.62, 2227000.29, 969999.91, 1791279.96
## $ revenue         <dbl> 8726512.7, 6174643.0, 5962000.0, 6483359.0,
## $ va              <dbl> 2849232.22, 756590.00, 331998.00, 1462863.00
## $ cohort          <dbl> 1994, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA
## $ age             <dbl> 0, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N
## $ years_in_sample <dbl> 1, 1, 1, 11, 11, 11, 11, 11, 11, 11, 11, 11,
## $ enter           <dbl> 1, NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ exit            <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0
```

Years

```
indonesia_firms |> count(year)
```

```
## # A tibble: 11 x 2
##     year     n
##    <dbl> <int>
##  1  1990 16459
##  2  1991 16415
##  3  1992 17564
##  4  1993 18068
##  5  1994 18908
##  6  1995 21424
##  7  1996 22851
##  8  1997 22245
##  9  1998 21285
## 10  1999 21921
## 11  2000 22030
```

Employment

Lots of interesting variables to consider: employemtn, value of materials, value of capital, revenue, value added

We'll focus on the number of workers because it's easy to interpret without worrying about inflation or exchange rates

Can think about this exercise as if we are starting a new manufacturing plant, and we want to know how many workers we should expect to need

The number of workers per plant is very widely dispersed

```r
summary(indonesia_firms$workers)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     11.0     26.0     45.0    188.4    131.0 116052.0
```

Employment over Time

```r
indonesia_firms |>
  group_by(year) |>
  summarise(average = mean(workers),
            p25 = quantile(workers, probs = 0.25),
            p50 = quantile(workers, probs = 0.50),
            p75 = quantile(workers, probs = 0.75))
```

```
## # A tibble: 11 x 5
##     year average   p25   p50   p75
##    <dbl>   <dbl> <dbl> <dbl> <dbl>
##  1  1990    160.    26    42   113
##  2  1991    181.    27    45   133
##  3  1992    187.    26    46   139
##  4  1993    196.    28    49   145
##  5  1994    200.    27    50   148
##  6  1995    193.    26    45   131
##  7  1996    183.    26    43   123
##  8  1997    185.    26    44   125
##  9  1998    192.    26    44   129
## 10  1999    192.    26    45   129
## 11  2000    197.    26    46   132
```

Ages of Plants

The time series of average employment does not give us a good idea of how many workers will be working in our new plant over time

We will need to think about each plant's age: years since opening

Age constructed from observing the same plant ID more than once, so only have it for plants that opened after 1990

```r
summary(indonesia_firms$age)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##      0.0      1.0      2.0      2.6      4.0      9.0   122768
```

Plants present in 1990 have age = `NA`, slightly more than half the observations

```r
nrow(indonesia_firms)
```

```
## [1] 219170
```

For the rest of the analysis, drop plants if we don't know their ages

```r
indonesia_firms <- indonesia_firms |> drop_na(age)
```

Ages of Plants

To better understand the structure of the data, let's look at plant age by year

```
table(indonesia_firms$age, indonesia_firms$year)
```

```
##
##      1991 1992 1993 1994 1995 1996 1997 1998 1999 2000
##   0 2707 2399 1836 2068 3530 3173 1699 1521 1404 1094
##   1    0 2353 2107 1632 1863 3056 2667 1472 1413 1288
##   2    0    0 2133 1926 1515 1641 2515 2258 1388 1323
##   3    0    0    0 1991 1786 1346 1441 2143 2125 1287
##   4    0    0    0    0 1880 1639 1221 1276 2017 1986
##   5    0    0    0    0    0 1732 1510 1075 1205 1895
##   6    0    0    0    0    0    0 1606 1383 1021 1143
##   7    0    0    0    0    0    0    0 1444 1326  962
##   8    0    0    0    0    0    0    0    0 1392 1251
##   9    0    0    0    0    0    0    0    0    0 1338
```

Can you see the consequences of the 1997 Asian financial crisis in the table?

Can you see cohorts in the table?

Ages and Cohorts of Plants

The dataset also has a cohort variable: the year the plant opened (was “born”)

Tabulating age and cohort represents the plant life-cycle in an intuitive way

```r
table(indonesia_firms$age, indonesia_firms$cohort)
```

```
##
##     1991 1992 1993 1994 1995 1996 1997 1998 1999 2000
##   0 2707 2399 1836 2068 3530 3173 1699 1521 1404 1094
##   1 2353 2107 1632 1863 3056 2667 1472 1413 1288    0
##   2 2133 1926 1515 1641 2515 2258 1388 1323    0    0
##   3 1991 1786 1346 1441 2143 2125 1287    0    0    0
##   4 1880 1639 1221 1276 2017 1986    0    0    0    0
##   5 1732 1510 1075 1205 1895    0    0    0    0    0
##   6 1606 1383 1021 1143    0    0    0    0    0    0
##   7 1444 1326  962    0    0    0    0    0    0    0
##   8 1392 1251    0    0    0    0    0    0    0    0
##   9 1338    0    0    0    0    0    0    0    0    0
```

Cross-Sectional Life-Cycles of Plants

Let’s return to thinking about how the number of workers varies with plant age

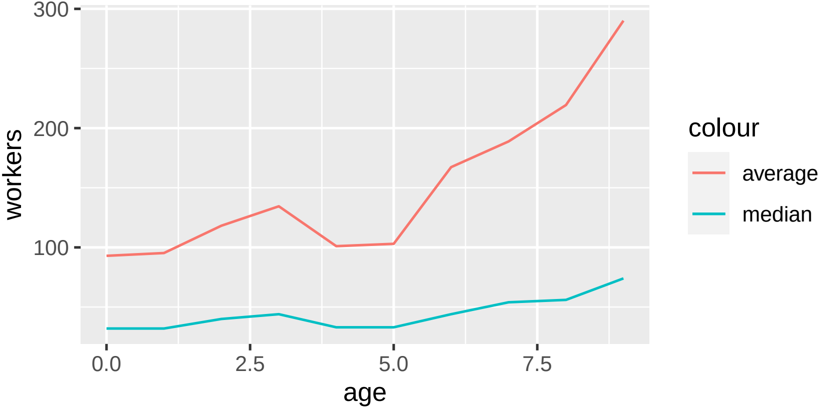
Let’s start naively, by looking at cross-sectional patterns in the final year

```r
tbl <-
  indonesia_firms |>
  filter(year==2000) |>
  group_by(age) |>
  summarise(average = mean(workers),
            median = median(workers))
```

Cross-Sectional Life-Cycle Graph

Both average and median are rising, but slopes are different

```
ggplot(tbl, aes(x=age)) +
  geom_line(aes(y=average, color = "average")) +
  geom_line(aes(y=median, color = "median")) +
  labs(y = "workers")
```

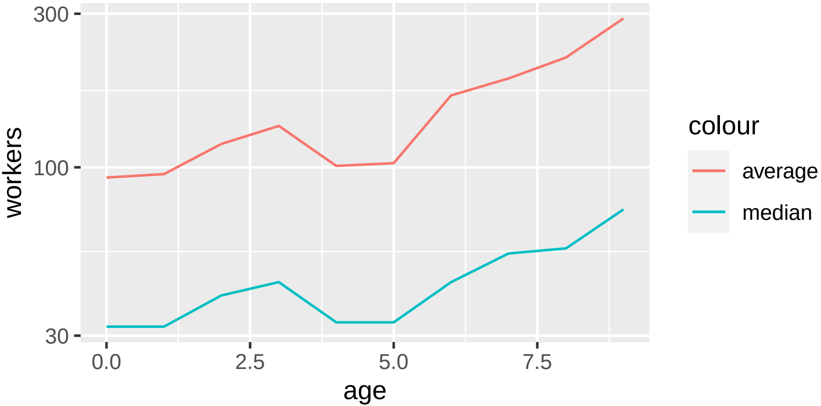


Cross-Sectional Life-Cycle Graph: Log Scale

Differing slopes may be because employment growth is proportional to size

In that case, the plots would be parallel with a log scale

```
ggplot(tbl, aes(x=age)) +
  geom_line(aes(y=average, color = "average")) +
  geom_line(aes(y=median, color = "median")) +
  labs(y = "workers") +
  scale_y_log10()
```



Cross-Sectional Growth Rates

Given the parallel results with a log scale, maybe we should plot growth rates

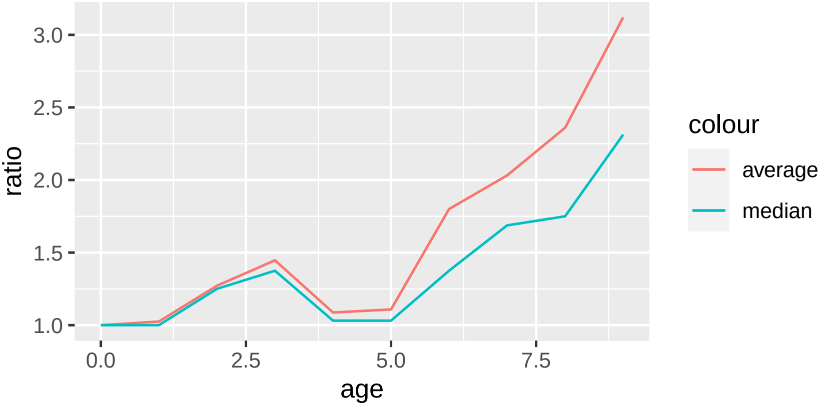
Let’s compute the cross-sectional growth rate of the average and median

```
tbl <-
  tbl |>
  arrange(age) |>
  mutate(average_ratio = average / first(average),
         median_ratio = median / first(median))
```

Cross-Sectional Life-Cycle Graph: Growth Version

Suggests plant size flat for first 5 years, then doubles to triples in next 4

```
ggplot(tbl, aes(x=age)) +
  geom_line(aes(y=average_ratio, color = "average")) +
  geom_line(aes(y=median_ratio, color = "median")) +
  labs(y = "ratio")
```



Cohort Life-Cycles of Plants

The cross-sectional patterns may be very misleading

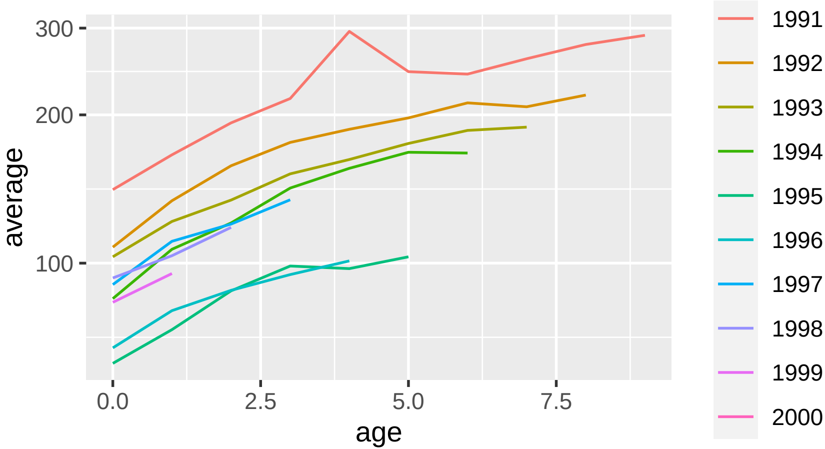
More informative to look at cohorts as they age

For simplicity, we'll focus on averages rather than medians

```r
tbl <-
  indonesia_firms |>
  group_by(age, cohort) |>
  summarise(average = mean(workers))
```

Cohort Life-Cycle Graph

```r
ggplot(tbl, aes(x=age, y=average, color=factor(cohort))) +
  geom_line() +
  scale_y_log10()
```



Steady growth with age, but lots of variation in starting points

Cohort Growth Rates

Growth rates will provide easiest comparison of cohort and cross-section

Let's compute the cohort growth rate of average employment

```
tbl <-
  tbl |>
  group_by(cohort) |>
  arrange(age) |>
  mutate(average_ratio = average / first(average)) |>
  ungroup()
```

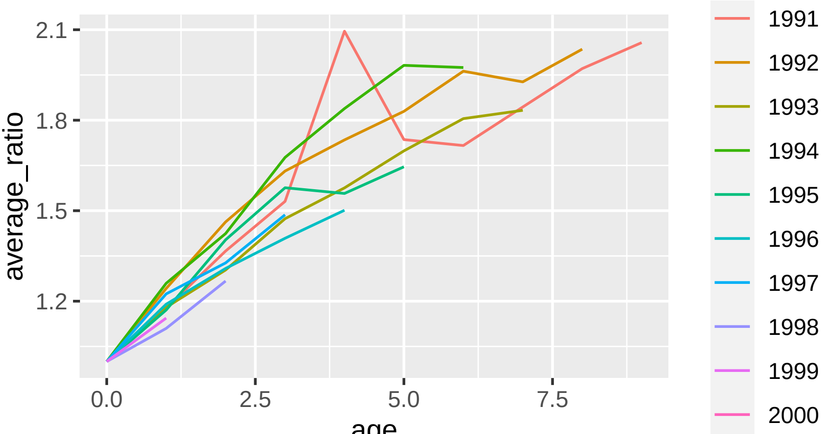
Cohort Life-Cycle Graph: Growth Version

The cross-sectional graph showed a 9-year tripling of the average

The cohort graph shows a doubling

The cross-section overstated employment growth because it ignored cohort effects

```r
ggplot(tbl, aes(x=age, y=average_ratio, color=factor(cohort))) +
  geom_line()
```



Plant Death

Let's look again at our age/cohort table

Do you see plant death?

```r
table(indonesia_firms$age, indonesia_firms$cohort)
```

```
##
##     1991 1992 1993 1994 1995 1996 1997 1998 1999 2000
##   0 2707 2399 1836 2068 3530 3173 1699 1521 1404 1094
##   1 2353 2107 1632 1863 3056 2667 1472 1413 1288    0
##   2 2133 1926 1515 1641 2515 2258 1388 1323    0    0
##   3 1991 1786 1346 1441 2143 2125 1287    0    0    0
##   4 1880 1639 1221 1276 2017 1986    0    0    0    0
##   5 1732 1510 1075 1205 1895    0    0    0    0    0
##   6 1606 1383 1021 1143    0    0    0    0    0    0
##   7 1444 1326  962    0    0    0    0    0    0    0
##   8 1392 1251    0    0    0    0    0    0    0    0
##   9 1338    0    0    0    0    0    0    0    0    0
```

Plant Survival Function

Can compute survival function, just like in week 2

- Count the number of plants in each age/cohort cell
- Group by cohort
- Compute current number as a share of initial number
- Ungroup

```
tbl <-
  indonesia_firms |>
  count(age, cohort) |>
  group_by(cohort) |>
  arrange(age) |>
  mutate(share_surviving = n / first(n)) |>
  ungroup()
```

Plant Survival Function

For intuition, let's look at the part of the table for the oldest cohort

```
tbl |> filter(cohort==1991)
```

```
## # A tibble: 10 x 4
##      age cohort     n share_surviving
##    <dbl>  <dbl> <int>           <dbl>
##  1     0   1991  2707           1
##  2     1   1991  2353           0.869
##  3     2   1991  2133           0.788
##  4     3   1991  1991           0.736
##  5     4   1991  1880           0.694
##  6     5   1991  1732           0.640
##  7     6   1991  1606           0.593
##  8     7   1991  1444           0.533
##  9     8   1991  1392           0.514
## 10     9   1991  1338           0.494
```

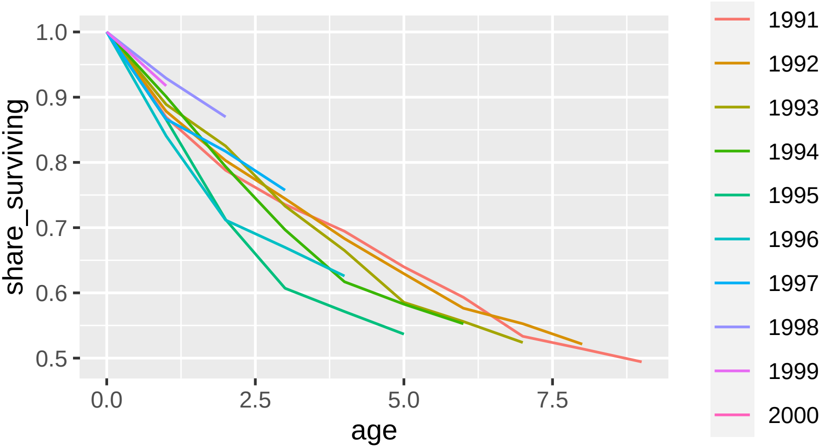
Plant Survival Curves

A lot of plants die!

By 9 years, half close down

Alternatively, then median length of operation for a plant is 9 years

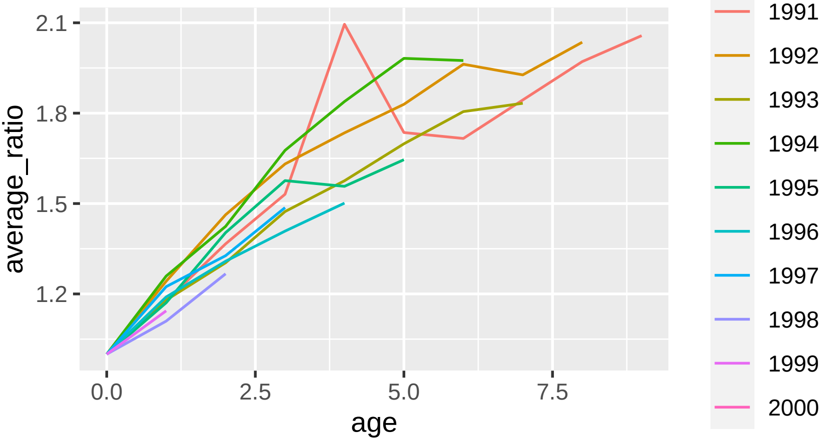
```r
ggplot(tbl, aes(x=age, y=share_surviving, color=factor(cohort))) +
  geom_line()
```



Reconsidering the Cohort Life-Cycle Graph

Now that we have acknowledged that many plants close, we must admit that the cohort graph did **not** track the same plants over the lifecycle

As plants exit the sample, average may change due to selection, not life-cycle



Selection over the Lifecycle

The issue is that cohorts may change in composition over time

Various drivers of this selection

- Here, it is mortality (or plant closing)
- In Abramitzky and Boustan, it is also return migration
- In other studies, it might be variability in who agrees to participate

In repeated cross-section data (as in PS 5), can't do anything about this issue

In panel data (as we have here), can do more, but no perfect solution

Addressing Selection

One fix for the selection issue: focus on plants that were present in all years

This approach is perfect, since we are not sure our plant is going to survive

Can implement it here by using `years_in_sample`

```r
indonesia_firms |> count(years_in_sample)
```

```
## # A tibble: 10 x 2
##    years_in_sample     n
##              <dbl> <int>
##  1               1  3580
##  2               2  6304
##  3               3  7740
##  4               4  8400
##  5               5 13010
##  6               6 13584
##  7               7  9947
##  8               8  8712
##  9               9 11745
## 10              10 13380
```

Let's focus on the firms that were present in every year: "balanced panel"

Employment Growth in the Balanced Panel

Keep the balanced panel, then compute plant-level employment growth

```r
indonesia_firms <-
  indonesia_firms |>
  filter(years_in_sample==10) |>
  group_by(psid) |>
  arrange(age) |>
  mutate(ratio = workers / first(workers)) |>
  ungroup()
```

Now compute average ratio

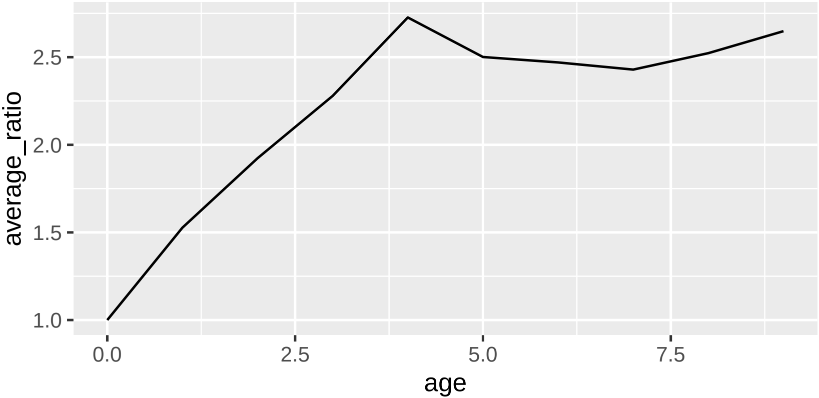
```r
tbl <-
  indonesia_firms |>
  group_by(age) |>
  summarise(average_ratio = mean(ratio))
```

Graph of Employment Growth in the Balanced Panel

The graph shows 2.5x employment growth, between the two sets of earlier results

This number reflects how much employment grew in plants that started in 1991 and continued functioning through 2000

```r
ggplot(tbl, aes(x=age, y=average_ratio)) +
  geom_line()
```


average_ratio
age