Cohort Analysis

Josep Curto Díaz, Adjunct Professor[a]

[a]IE Business School, Madrid, 28006, Spain

This version was compiled on March 17, 2019

This technical note introduces what is Cohort Analysis and its benefits and limitations.

cohort analysis | clustering | customer analytics | r

The problem.

- **Problem**: we don't know how customers' behaviour changes over time
- **Goals**:
- Understanding customers' behaviour evolution
- Understanding behaviour using groups that evolve over time, not individually
- Use groups based on business criteria instead of relaying on a simple metric or a segmentation technique
- **Why?** Perform specific actions on groups, detect similar temporal patterns on group

We will use what is called **cohort analysis**.

Definition.

> **Cohort** is a group of people used in a study who have something (such as age, social class, when they become our customers) in common.

We can find the origin in ancient times. A cohort was a **military unit**, one of ten divisions in a Roman legion.

Nowadays, we have a formal definition:

> **Cohort analysis** is a observational, analytical and longitudinal study. It is a comparison of the evolution of a particular aspect (KPI). Individuals comprising study groups are selected based on the presence of a particular characteristic.

Cohort analysis is a traditional tool in epidemiology. When we applied this technique in other industries most of the times, we use metrics that are easier to capture and analyse. They can be:

- Direct: number of customers, revenue, cost,...
- Derived: retention,...

Types of Age-Period-Cohort Analysis.

Age-period-cohort analysis is one of the methods used in an effort to separate the effects of age, period, and cohort.

Keyes et al. (2010) distinguish between three different kind of models, based on their distinction between 1st order and 2nd order effects:

- Models in which 1st order effects are estimated and interpreted;
- Models in which 2nd order effects are estimated and interpreted; and
- Hybrid models in which 1st order effects are estimated but 2nd order effects interpreted.

1st order effects are determined based on the assumption that age, period, and cohort can exist independently of each other and have a linear relationship with the outcome of interest. Each linear slope is estimated by controlling for the additive effect of the other two effects. These linear relationships are what Keyes et al. term 1st order effects. 2nd order effects are those which have a non-linear relationship with the outcome of interest.

How to apply the analysis.

- **[BU]** Determine business questions/needs, measure to study and cohorts of interest
- **[DU]** Data Sourcing, Cleaning & Exploration
- **[DP]** Create cohorts, extract data according to cohorts
- **[M]** Calculate the measure
- **[E]** Analyze results and adjust parameters
- **[D]** Present and explain the results

Example (I)

First we load the required packages: ggplot2 and dplyr (in case we don't have them).

Let's imagine that we have a startup and we have data from the weekly evolution of several cohorts (defined as cohort X refers to the customers that subscribe to the service in week X). We are going to create a data frame with four fields: cohort, week, value (number of active customers) and action (the company did a marketing action to adquire customers).

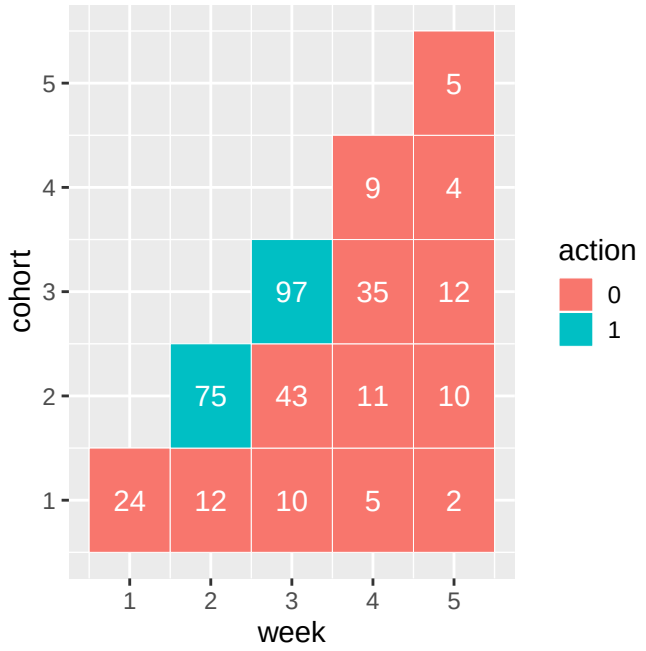
```
# Dataframe creation
df <- structure(list(
  cohort = c(1L, 1L, 1L, 1L, 1L, 2L,
             2L, 2L, 2L, 3L, 3L, 3L, 4L, 4L, 5L),
  wk = c(1L, 2L, 3L, 4L, 5L, 2L, 3L, 4L, 5L, 3L,
         4L, 5L, 4L, 5L, 5L),
  value = c(24L, 12L, 10L, 5L, 2L, 75L, 43L, 11L,
            10L, 97L, 35L, 12L, 9L, 4L, 5L),
  flag = c(0L, 0L, 0L, 0L, 0L, 1L, 0L, 0L, 0L,
           1L, 0L, 0L, 0L, 0L, 0L)),
  .Names = c("cohort", "week", "value", "action"),
  class = "data.frame",
  row.names = c(NA, -15L))

# Coarse action to factor
df$action <- as.factor(df$action)

# View
df
```

```
#    cohort week value action
#  1      1    1    24      0
#  2      1    2    12      0
#  3      1    3    10      0
#  4      1    4     5      0
#  5      1    5     2      0
```

```
#  6       2    2    75       1
#  7       2    3    43       0
#  8       2    4    11       0
#  9       2    5    10       0
#  10      3    3    97       1
#  11      3    4    35       0
#  12      3    5    12       0
#  13      4    4     9       0
#  14      4    5     4       0
#  15      5    5     5       0
```

We can present this information as a cohort table to understand what is happening.

```
ggplot(df, aes(x = week,
               y = cohort, fill = action)) +
    geom_tile(color = "white") +
    geom_text(aes(label = value), color = "white")
```



The diagonal gives information about the number of new customers per week. Every row is giving as information about how many customers in every cohort remain alive. Similar happens if we analyze the information considering the diagonal (the amount of new customers per week) or the columns. In addition, we can obtain relevant information from the dataset as well. For example, the total amount of customers alive per week:

```
weeklyCustomers <- aggregate(value ~ week, df, sum)
weeklyCustomers
```

```
#    week value
#  1    1    24
#  2    2    87
#  3    3   150
#  4    4    60
#  5    5    33
```

Example (II)

Let's imagine that we have the following cohort table:

```
cohort.sum <- data.frame(
  cohort=c('C01', 'C02', 'C03', 'C04', 'C05',
           'C06', 'C07', 'C08', 'C09', 'C10',
           'C11', 'C12'),
  M1=c(270000,0,0,0,0,0,0,0,0,0,0,0),
  M2=c(85000,275000,0,0,0,0,0,0,0,0,0,0),
  M3=c(72000,63000,277000,0,0,0,0,0,0,0,0,0),
  M4=c(52000,42000,76000,361000,0,0,0,0,0,0,
       0,0),
  M5=c(50000,45000,60000,80000,288000,0,0,0,
       0,0,0,0),
  M6=c(51000,52000,55000,51000,58000,253000,
       0,0,0,0,0,0),
  M7=c(51000,69000,48000,45000,42000,54000,
       272000,0,0,0,0,0),
  M8=c(46000,85000,77000,41000,38000,37000,
       74000,352000,0,0,0,0),
  M9=c(38000,42000,72000,41000,31000,30000,
       49000,107000,285000,0,0,0),
  M10=c(39000,38000,45000,33000,34000,34000,
        46000,83000,69000,279000,0,0),
  M11=c(38000,42000,31000,32000,26000,28000,
        43000,82000,51000,87000,282000,0),
  M12=c(35000,35000,38000,45000,35000,32000,
        48000,44000,47000,52000,92000,500000)
  )
```

This table provide diffent types of information that are relevant to undestand the customer behaviour as whole and based on cohorts:

- **Diagonal**: it represents the revenue generated by the new customers per month. Let's check the value in the third month.

```
d <- as.vector(diag(as.matrix(cohort.sum[, -1])))
d[3]
```

```
#  [1] 277000
```

- **Row**: it represents the total revenue per cohort during the period of analysis. We can calculate the total amount for a particular cohort (in this case C04).

```
cohort.sum %>% filter(cohort == "C04") %>% select(-cohort) %>%
```

```
#  [1] 729000
```

- **Columns**: it represents the total revenue generated per month (by all the active customers). We can calculate the total amount for a particular month.

```
cohort.sum %>% select(M4) %>% sum()
```

```
#  [1] 531000
```

Benefits.

- Understand Customer Lifecycle/Journey: length, value, situation,...
- Identify patterns
- Behavioral/Psychographic analysis

Use Cases.

- Examine where cashflow is coming from and understand the health of your business
- Easily see how much monthly or quarterly revenue is driven from newer and older cohorts
- Study customer retention patterns to see if they are getting better or worse
- Compare cohorts of users from different segments

References.

- Bent Nielsen. apc: An R Package for Age-Period-Cohort Analysis. The R Journal, 7(2):52-64, Dec. 2015
- Startup Metrics for Pirates
- Lean Analytics
- Cohort Analysis Cheat Sheet
- Data Analytics for Startups - Tetuan Valley Startup School Fall 2015