CA - S9: Cohort Analysis

Josep Curto

June 28, 2018

Contents

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
- **Why?** Perform specific actions on groups, detect similar temporal patterns on group

But attention, we decide the groups instead of relaying on a simple metric or a segmentation technique.

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.

Then,

> **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:

- Metrics are easier to capture and analyse
- 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.

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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).

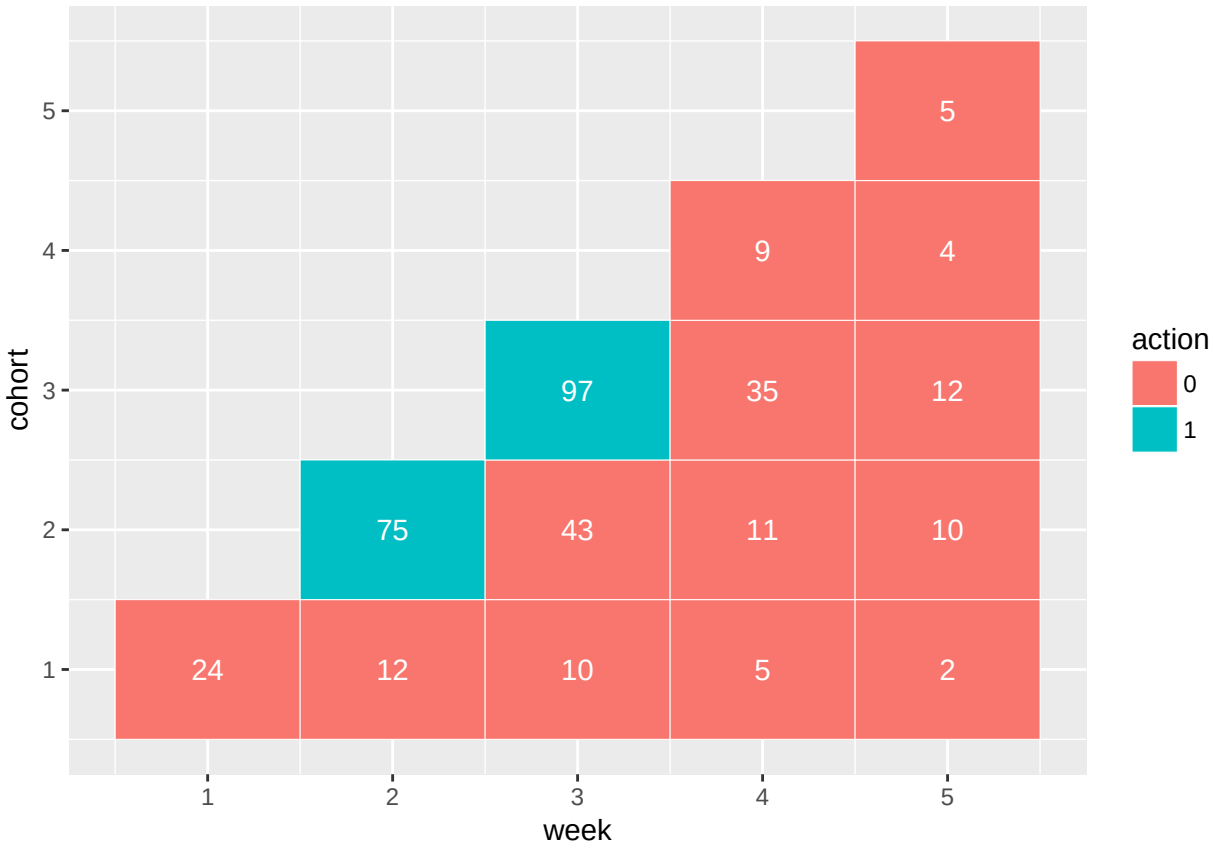
```
df <- structure(list(cohort = c(1L, 1L, 1L, 1L, 1L, 2L, 2L, 2L, 2L, 3L, 3L, 3L, 4L, 4L, 5L), wk = c(1L,
df$action <- as.factor(df$action)
df
```

```
##    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
```

In this data set we have four fields: cohort, week, value (number of active customers) and action (the company did a marketing action to adquire customers).

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. From the dataset we can obtain relevant information as well. For example, the total amount of customers alive per week:

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

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

We can know as well the amount of new customers per week just reviewing the diagonal.

Example (II)

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

cohort	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
C01	270000	85000	72000	52000	50000	51000	51000	46000	38000	39000	38000	35000
C02	0	275000	63000	42000	45000	52000	69000	85000	42000	38000	42000	35000
C03	0	0	277000	76000	60000	55000	48000	77000	72000	45000	31000	38000
C4	0	0	0	361000	80000	51000	45000	41000	41000	33000	32000	45000
C05	0	0	0	0	288000	58000	42000	38000	31000	34000	26000	35000
C06	0	0	0	0	0	253000	54000	37000	30000	34000	28000	32000
C07	0	0	0	0	0	0	272000	74000	49000	46000	43000	48000
C08	0	0	0	0	0	0	0	352000	107000	83000	82000	44000
C09	0	0	0	0	0	0	0	0	285000	69000	51000	47000
C10	0	0	0	0	0	0	0	0	0	279000	87000	52000
C11	0	0	0	0	0	0	0	0	0	0	282000	92000
C12	0	0	0	0	0	0	0	0	0	0	0	500000

This table provide diffent types of information:

- **Diagonal**: Revenue generated by the the new customers per month
- **Row**: Total revenue per cohort during the period of analysis.
- **Columns**: Total revenue generated per month.

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