Analyseenheden report for Landsuddeling 2019

Matthew Phelps

24 January, 2020

Table of Contents

# Executive summary

### Problem statement

Landsuddelingen would like to predict which Hjertestarter groups they should target with additional support in order to increase the probability the groups reach the minimum of 15 routes required for a group to receive a free hjertestarter.

### Results

The original question could not be answered due to two reasons:

* The complexity of the relationship between the outcome and the covariates
* Insufficient data quality for key covariates. [[1]](#footnote-23)
* The rarity of groups that were “near-misses” (groups with 10 - 14 routes).

A sub-analysis investigating factors associated with no-shows (people who sign up, but fail to show up on the date) suggests that:

* Collectors connected to a hjertestarter group were **more likely** to show up (OR: 9.3), as compare to those not connected to a group.
* Older age was associated with **higher odds** of showing-up.
* Gender had **no effect** on the probability a collector would show up

### Recommendation

Instead of trying to predict which groups receive assistance, we instead recommend using 2020 to test whether the planned intervention actually improves the likelihood that a Hjertestarter group succeeds in reaching a minimum of 15 routes. Specifically, we recommend that 3 - 4 weeks before the landsuddelingen date, Landsuddelingen randomly split the extant Hjertestarter groups into two groups. The first group will receive no additional assistance (called the “control group”), while the second group will receive the intervention (called the “intervention group”). Then, following completion of the 2020 landsuddeling, test if those in the intervention group had significantly more routes than those in the control group.

# Full Report

## Background

In the 2019 landsuddeling (LU), hjerterstarter groups who were able to get 15 or more routes completed received a free Hjertestarter. The LU organizers noted that some groups nearly reached this goal, but ended up falling short (e.g. achieving between 10 - 14 routes); we call these groups “near misses”. In order to improve these outcomes for subsequent years, the LU organizers wish to provide some assistance to these near-miss groups. However, this requires that the organizers know in advance, which groups are at risk of being near-misses.

LU hypothesized that groups that signed up more collectors (indsamler) over a short time-span were more likely to succeed in reaching 15 routes, and conversely, those groups that were slow to sign up collectors were at risk for being near-misses. For example, a group that signed up the first 5 collectors in 2 weeks might be more likely to succeed than a group that takes 4 weeks to reach the same number. Other potential predictor variables could include the presence/absence of a group coordinator, date of the groups’ creation, and the time lag between the groups creation and the first collector to sign-up.

## Process

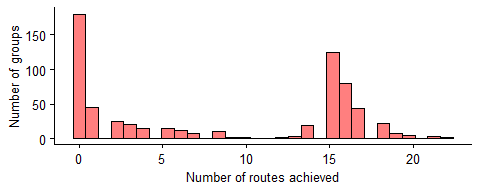
The data were first cleaned and checked for errors/inconsistencies. Problems with the data, and their solutions when possible, were discussed with LU. As the scope and nature of the data problems became apparent, it became clear that addressing the original problem would not be possible, and the scope of the analysis was broadened to explore any insights could be derived from the LU2019 dataset.

## Descriptive statistics

We start with some summary statistics of the attributes of the hjertestarter groups and of the collectors (indsamler).

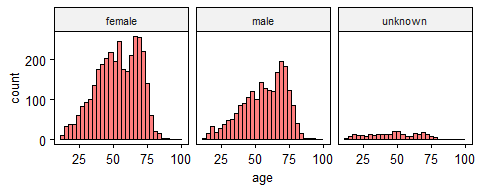
### Hjertestarter groups

* Total active hjertestarter groups: 641
* Number of “near-misses” hjertestarter groups (10-14 routes): 25 (r% of total active groups)
* Number of groups achieving 15 routes: 286 (44.6% of total active groups)
* Distribution of number of routes achieved:

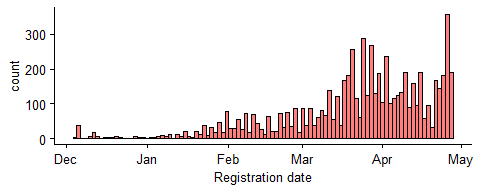


### Collectors (indsamler)

* Total number of collectors who signed up: 6770
* Number of collectors who showed up: 5411 (79.9% of total)
* Number of collectors attached to a hjertestarter group: 5746 (84.9% of total)
* Average age of the collectors: 53.8
* Distribution of ages (by gender):



* Distribution by registration date:



## Challenges

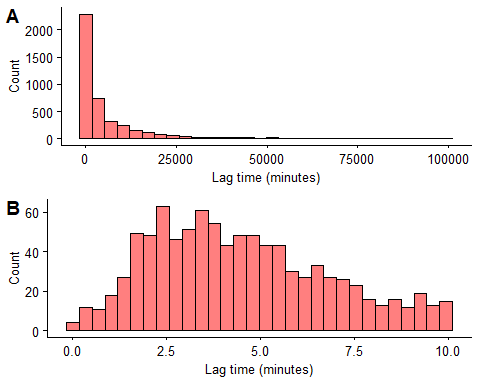
This sections explains the challenges that prevented us from addressing the original problem - predicting which groups would achieve between 10 - 14 routes.

### Few “near-misses”

Only 3.9% of the total active groups ended up being “near-misses”. As such, any model to predict these groups would have to perform very well in order to avoid giving too many false positives (i.e. predicting a group would be a near miss when it wasn’t).

### Lag-time

The main variable of interest, lag-time (time between successive sign-ups within a Hjertestarter group), was of poor quality. Many groups likely organized themselves and signed-up collectors offline before entering all the information into the online platform at once. To illustrate this, we can compare the distribution of all signups under 10 minutes to the distribution of all signups over 10 minutes (below). The distribution of signups with lag times > 10 minutes (plot A) follows the expected exponential distribution, while those with registrations lag times < 10 minutes (plot B) appears to be more normally distributed around the mean. This suggests a different data-generating process for those signups under 10 minutes (i.e.the offline “pre” organization of the hjertestarter group)



There were 939 signups out of 5746 (16.3%) that had a lag-time of < 10 minutes.

### Hjertestarter group creation time

Another covariate, the lag time between the creation of the hjertestarter group and the first collector to sign up for the group, contained invalid data in the form of negative values. A total of 61 groups (11.6%) contained a negative lag time. This likely occurred because some Hjertestarter groups were retroactively assigned collectors that had signed up earlier.

### Many registrations in final weeks

Since LU needs time to implement any intervention, we have to establish a “cut-off” date. This is the date LU will start the intevention, and thus the date on which LU assess to see which groups are at risk of being “near-misses”. Therefore our prediction model can only uses data from before this date.

A total of 2912 (50.7%) collectors attached to a hjertestarter group signed up after the cut-off date we used for this analysis (01-Apr). This is a significant amount of data that we are not able to use. Further, it is likely that the dynamics change in the final weeks, which may limit the utility of a model built on data that does not include these weeks.

## Main analysis: Predicting the number of routes for hjertestarter groups

*Due to the problems described in the “Challenges” section (above), the main analysis is not useful for prediction purposes, however we describe the methods and results below.*

We investigated the relationship between the number of routes each hjertestarter group achieved with a number of explanatory variables. The explanatory variable included:

* Mean lag-time between successive collector signups
* Time until first 5 collectors registered with group
* Lag-time between hjertestarter group creation and first collector signups
* Presence/absence of a group coordinator
* Date of the groups’ creation

The data included all hjertestarter groups registered before 01. Apr that had at least 2 collectors sign up, and the date of group registration < the date of the first collector in the group to be registered (n = 349).

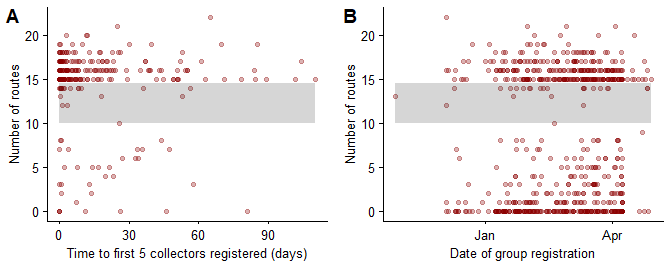
The data was split into training (n = 220) and test (n = 129) datasets. The models were fitted using the training data and applied to make predictions on the test dataset.

We investigated 9 models. The model with the best trade-off of complexity and performance was:

* number\_routes ~ rcs(time\_first\_5\_collectors) + date\_group\_created

This model shows that the lag-time between successive collector signups does not have a linear relationship with the outcome. In other words, the relationship between the covariate (lag-time) and the outcome changes depending on the value of the lag-time. The model also shows that the earlier the group was created, the greater the number of routes the group achieved.

However, as the plots below show, there are clearly two clusters of hjertestarter groups when plotted against either explanitory variable (plot A: time to first 5 collectors, plot B: date of group registration). Thus the relationship between the explanitory variables and the number of routes as described by the model is likely not usable.



## Sub-analysis: Predicting no-shows

We investigate the risk-factors for a collector not showing-up on the collection day. The potential risk-factors examined included:

* gender (binary)
* age (continuous)
* member/non-member of hjertestarter group (binary)

The data included all collectors who signed up before 01. Apr (n = 3858). The data was split into training (n = 2438) and test ( n = 1420) datasets. We investigated 4 logistic models that were fitted using the training dataset:

* model 1: no\_show ~ hs\_group + age + gender
* model 2: no\_show ~ hs\_group + rcs(age) + gender
* model 3: no\_show ~ hs\_group + age
* model 4: no\_show ~ hs\_group

Model 3 provided the best trade-off of complexity and performance, therefore our inferences are drawn from his model.

This models shows that those collectors who are not connected to a hjertestarter group have 9.3 (95% CI: 11.8 - 7.3) greater odds of not showing up as compared to collectors who are connected to a hjertestarter group. Increasing age was also associated with higher odds of showing up, with each year of age increasing the odds by 1.7 (95% CI: 2 - 1.4).

## Recommendations

For the 2020 LU, we recommend testing to see if the planned intervention is effective at improving the likelihood that a group will achieve 15 routes (or if the intervention increases the number of routes a group achieves on average). This can be done in the following steps in 2020:

1. Select a date when the intervention should start (e.g. 3 weeks before collection day).
2. On that date, randomly divide the existing active hjertestarter groups into two - the intervention group and the control group - and ensure that the data on which group each hjertestarter group belongs to is tracked and stored. The size of the intervention group will be determined by a combination of how many resource for the intevention are available, and how many resource each individual group requires for the intervention.
3. Apply the intervention to the those in the intervention group, and apply the standard practices to those in the control group.
4. After LU 2020 is finished, test if the intervention worked. This can be done in several ways. The simplest ways are:
   * Build a logistic model, where the outcome is the binary variable of whether the hjertestarter group achieved 15 routes (yes/no), and the explanitory variable is which intervention group the hjerterstarter group belonged to (intervention or control)
   * Build a linear model where the outcome is the continuous variable of how many routes each hjertestarter group achieved, and the explanitory variable is which intervention group the hjerterstarter group belonged to (intervention or control)

Additionally, it may be helpful to try and collect some additional variables from the collectors that might help predict no-shows ahead of time. These could include:

* Previous experience as a collector (yes/no)
* Whether they are personally affected by hjerte-kar sygdomme, e.g. if they, or a family or friend has a CVD (note - this may have GDPR implications since it is probably personfølsomme oplysninger).
* Gender is not recommend to be included. This can be inferred from names, and furthermore, it seems to be unrelated to the probability of no-shows.

These data would only be worthwhile to collect if there is a reasonable probability that a high percentage of people will provide the data. Otherwise the data will end up too biased to be useful.

1. This is not a criticism of the data collection process, rather it’s a result of the fact that the data-infrastructure for LU is designed to facilitate an efficient implementation of the activity, and not designed for data-collection for research purposes. [↑](#footnote-ref-23)