Tobacco shopping and lockdown

02 July 2021

Overall goal: Investigation of tobacco shopping and lockdown

It is well-established that smoking is a risk factor for cardiovascular disease (REF), and recent studies show that smoking worsens the COVID-19 disease course (REF). Some studies from Italy and UK find that smoking decreased during lockdown (REF). No such studies have been found in a Danish context, and furthermore, these existing studies are based on self-reported smoking habits. In this paper, we wish to investigate if lockdown affected tobacco purchases among some groups, based on a time series of credit-card transactions from various Danish supermarkets.

The lockdown period (68 days from 11 March 2020 to 18 May 2020) will be compared to three other control periods of equal length:

- a) 11 March 2019 to 18 May 2019 (68 days)
- b) 02 Jan 2020 to 10 March 2020 (68 days)
- c) 19 May 2020 to 26 July 2020 (68 days)

for the following groups

- 1) Sex.
- 2) Age (18-29, 30-39, 40-49, 50-69).
- 3) Geography (region)
- 4) Education.
- 5) Lifestyle disease (blood lowering/cholesterol lowering drugs at least 6 months before chosen start).
- 6) Work sector.

Supermarket transaction data structure

We have a large amount of supermarket transaction data, where each transaction is uniquely defined by person and time, and contains at least one item. In this framework, one can think of a transaction as a receipt. We let K denote the total number of transactions in the database. Letting M be the total number of items in the database, we define the set of items to be

$$\mathcal{I} = \{I_1, I_2, ..., I_M\}, \tag{1}$$

where item m is denoted I_m. Each item is associated with a positive item price and item quantity. This information is contained in the variables **item**, **itemprice** (in DKK) and **quantity** as seen in table 1 below. Here, the gray and white colors mark the different transactions, which are identified uniquely by a transaction id, **TID**. Thus, in below example in table 1, we have a database consisting of $K = 5$ transactions, and $M = 9$ items:

$$\mathcal{I} = \{I_1, ..., I_9\} = \{bread, dip, dressings, fresh\ eggs, apples, milk, wine, beef, yoghurt\}$$

Note that the total price of the transaction (**transactionprice**) is based on itemprice and quantity.

TID	person	time	item	itemprice	quantity	transactionprice
1	1	17-03-2019 08:03:00	bread	11.95	1	76.95
1	1	17-03-2019 08:03:00	dip	6.00	2	76.95
1	1	17-03-2019 08:03:00	dressings	53.00	1	76.95
2	1	19-03-2019 10:15:53	fresh eggs	27.95	1	78.40
2	1	19-03-2019 10:15:53	apples	15.00	0.700	78.40
2	1	19-03-2019 10:15:53	dip	10.00	2	78.40
2	1	19-03-2019 10:15:53	bread	19.95	1	78.40
3	2	02-02-2020 19:34:01	milk	9.95	1	9.95
4	2	14-02-2020 15:55:04	wine	109.00	3	479.80
4	2	14-02-2020 15:55:04	beef	49.95	2	479.80
4	2	14-02-2020 15:55:04	yoghurt	18.95	1	479.80
4	2	14-02-2020 15:55:04	bread	5.00	5	479.80
4	2	14-02-2020 15:55:04	milk	8.95	1	479.80
5	2	20-02-2020 20:24:10	apples	2.00	2	19.00
5	2	20-02-2020 20:24:10	bread	15.00	1	19.00
...	...	...	...	...	...	

Table 1: Transaction data example with 5 transactions and 9 different items.

- The structure of the transaction data is quite complex: each individual has multiple observations of transactions over time, and the transactions are quite irregular, meaning

that the frequency of transactions differs over the weeks, months and between individuals.
- The number of items for each transaction will also vary between individuals and through time, as seen in above table. Furthermore, we will have new individuals entering the study and people dropping out, or even individuals dropping out and entering again, which creates missing time gaps.
- To describe this complex data structure we will adapt the theory of marked point processes [4] [5].

Supermarket transaction data: a marked point process

- Let T_k be the time for the k^{th} supermarket grocery transaction. For each transaction time, T_k, one or more items are purchased.
- For each transaction time T_k, we have information about the items in the transaction, which is described by $(X_1(T_k), ..., X_M(T_k))$. Specifically, $X_m(T_k) = (P_m(T_k), Q_m(T_k))$ denotes a vector that contain information about price and quantity for item m at time T_k.
- Consider the first transaction, $k = 1$, from the example in table 1. We have three different items $I_1 = bread$, $I_2 = dip$ and $I_3 = dressings$, with corresponding price and quantity. For bread we have:

$$X_1(T_1) = (P_1(T_1), Q_1(T_1)) = (11.95, 1) \in \mathcal{X}_1 = \mathbb{R}^+ \times \mathbb{R}^+$$

- Thus, for each transaction we have a positive real price and quantity for each item. We thus have a **mark space** for item m given by $\mathcal{X}_m = \mathbb{R}^+ \times \mathbb{R}^+$.
- With this example in mind, we can now define the marked point process, ϕ for the transaction times, T_k.

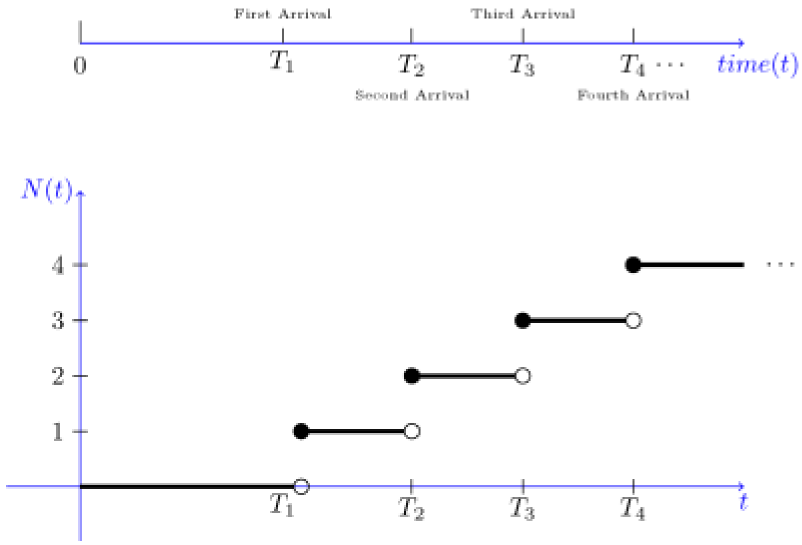
$$\phi = (T_k, (X_1(T_k), ..., X_M(T_k)))_{k \geq 1} \tag{2}$$

where each item has an associated mark space, such that $X_1(T_k) \in \mathcal{X}_1, ..., X_M(T_k) \in \mathcal{X}_M$. Note that the mark spaces do not depend on transaction, but only on item, and the mark space for the entire set of items $\mathcal{I}$ is given by $\mathcal{X}_1 \times ... \times \mathcal{X}_M$.

- Note that in most cases the quantities will be natural numbers, however, in some cases the quantity will be measured in kg or g, and the corresponding price will then be price per kg or price per g. As an example, see transaction two in table 1, where the costumer bought 0.7 kg apples that cost 15.00 DKK per kg.
- The defined marked point process from (2) counts the number of transactions made up to and including time t. When considering the process as a function of t, we have an integer-valued step function with jumps of size +1, which we assume to be right-continous, so that $N(t)$ is the number of events in the time interval $[0, t]$:

$$N(t) = \sum_{k \geq 1} I\{T_k \leq t\}, \tag{3}$$

where we assume $N(0) = 0$. See figure below for an illustration of the marked point process as a counting process [7].



Target parameter including examples

General target parameter

- We focus on a fixed time period $t \in [a, b]$ which is the same for every person and an item or itemset m^*. Note that m^* can be a specific item (example m^* = tobacco) or an itemset (example m^* = sugary drinks = {soda with sugar, ice tea, alcohol free beer}).
- We now wish to define a target parameter in a general way, such that this can be used to investigate different questions. Following Appendix 4 (example A4.4 in Last and Brandt), we can write the target for m^* in the time period $[a, b]$ as a Lebesgue-Steiltjes integral. Here, we integrate with respect to the counting process defined above, as this process marks the arrivals of the transactions. We define:

$$\begin{aligned}
\mu^{m^*}(a, b) &= E\left(\int_a^b f(X(t)) dN(t)\right) \\
&= \sum_{k:\ a \leq T_k \leq b} E(f(X(T_k))),
\end{aligned}$$

where $f(X(t))$ is a function defining the nature of the target. See below two examples for an understanding of this parameter.

Example 1

- We wish to investigate **the expected number of transactions containing** m^* **for** $t \in [a,b]$. Then we would define $f(X(t)) = I_{\{Q^{m^*}(t)>0,\ P^{m^*}(t)>0\}}$. So, $f(X(t))$ denotes the transactions where m^* was bought, ie. we have a positive quantity and price for m^*. From this, we could also calculate for example the expected number of daily or monthly transactions in the period.

Example 2 (maybe change notation here)

- The idea is to estimate the **expected relative budget spent per transaction on** m^* **for** $t \in [a,b]$.
- We use the same expression for $\mu^{m^*}(a,b)$, however, now defining $f(X(t)) = \frac{P^{m^*}(t) \cdot Q^{m^*}(t)}{\sum_{m=1}^{M} P^m(t) \cdot Q^m(t)}$. We would then need to calculate:

$$\mu_{rel}^{m^*}(a,b) = \frac{\mu^{m^*}(a,b)}{\sum_{k=1}^{K} I_{\{a \le T_k \le b\}}},$$

where the denominater is the number of transactions in the period $[a,b]$.

Including covariates and grocery shopping history

If we were to compare different groups or include grocery shopping history, we could condition on a filtration, $\mathcal{F}^-$, which denotes a set of possibly time varying covariates known just before time t as well as the grocery shopping history up until time t. Note that the shopping history is known at time T_k, but the time varying covariates can change at any time point, s. So, we have:

$$\mathcal{F}_{t^-} = \sigma\{Z(s) : s < t, T_k, X(T_k) : T_k < t\}$$

Using this, the upper example questions could be extended as follows:

Example 1 extension

- Question: “Are there any regions in which tobacco shopping changes during lockdown?” Translated to target: “For the five regions in Denmark, is the expected number of tobacco transactions in lockdown different from the preceding period”?
- Question: “Are there any regional differences in the change in tobacco shopping during lockdown?” Translated to target: “Does the change in the expected number of tobacco transactions between the lockdown period and the preceding period differ between regions?”

Example 2 extension

- Is the expected relative budget spent on sugary drinks from 01.01.2019 until 31.12.2019 different for diabetic and non-diabetic households?

Observed data: time scale and censoring

Choice of time scale

We choose to use calender time scale and not time since storebox start. By doing this, we will take seasonality into account and compare the same transaction times for different subjects. In this way, we can investigate our hypothesis about the impact of lockdown on cigarette shopping, as supposed to truncating the time scale by using time since storebox start, as people enter at different time points.

Censoring

In the observed data, we have different scenarios, which can lead to censored data. Examples are pictured in Figure 1. Here, we have shown transactions for five fictive subjects with and without cigarettes in the period 1 Jan 2020 until 1 Jan 2021.

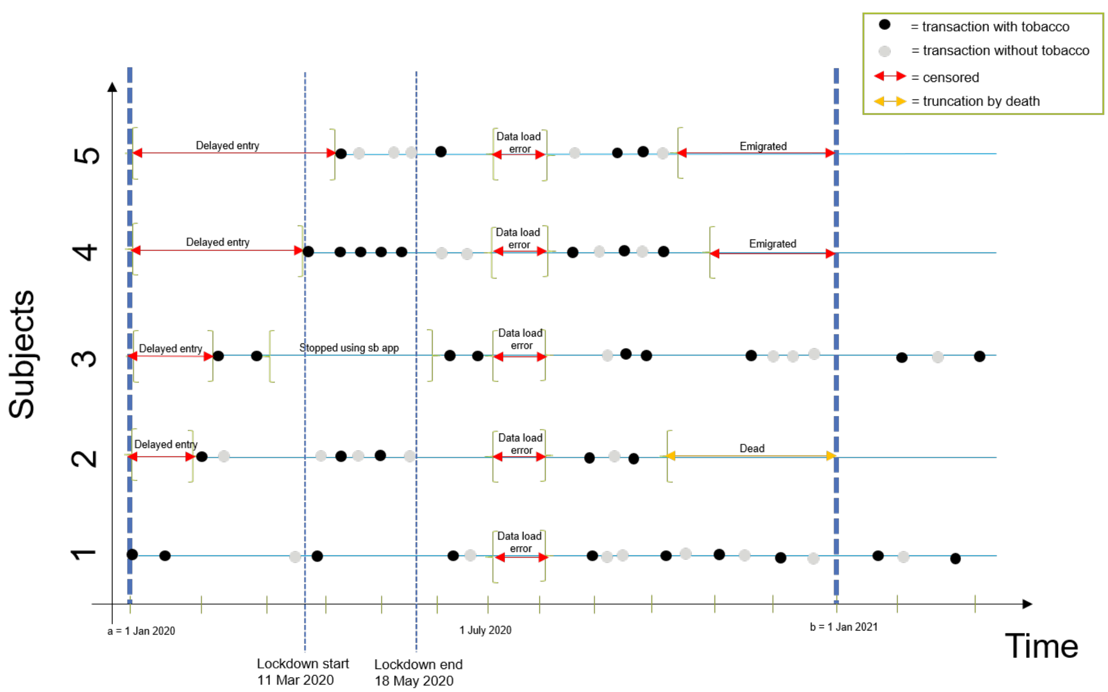


Figure 1: Missing mechanisms

Firstly, note that we have the case of truncation by death as marked by the orange arrow. In this case, the observations are not censored, as we know for sure that a transaction cannot happen from the grave! Secondly, in the periods with large gap times between transactions, we do not know whether the subject stopped using storebox, for example by changing to another supermarket (example subject 3), or if the subject realistically did not do frequent grocery shopping. For now we will not make any assumptions about these transaction gap times. In the periods marked by a red arrow, we know that the subject did not have a possibility of making a transaction using the storebox app. Therefore, the observed data is not complete as these periods are censored. Due to this censoring, $N(t)$ will not be fully observable, but only an incomplete version, $\tilde{N}_i(t)$ will be available for the i^{th} subject:

$$\tilde{N}_i(t) = N_i(t)C_i(t),$$

where the periods with censored data are delayed entry (caused by late entry into storebox or immigration), data load errors and emigration (assuming that the subject does not return to Denmark in the period $[a, b]$). So, we define the censoring process for the i^{th} subject as follows:

$$C_i(t) = \begin{cases} 0 & \text{if} \quad a \leq t \leq \min(T_{i1}, b) & \text{(delayed entry)} \\ 0 & \text{if} \quad e1^{\text{start}} \leq t \leq \min(e1^{\text{slut}}, b) & \text{(data load error)} \\ 0 & \text{if} \quad e2_i^{\text{start}} \leq t \leq \min(e2_i^{\text{slut}}, b) & \text{(emigration)} \\ 1 & \text{otherwise} & \text{(transactions observed)} \end{cases},$$

where $e1_{\text{start}}, e1_{\text{slut}}$ denote the start and end dates for a data load error (same for all subjects) and $e2_i^{\text{start}}, e2_i^{\text{slut}}$ denote start end dates for emigration for subject i. The censoring process is shown in Figure 2 for subject 2 from Figure 1.

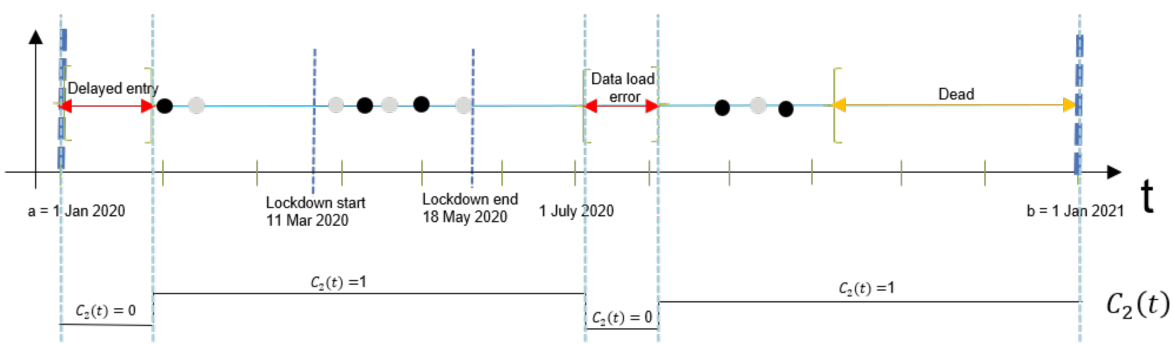


Figure 2: Example of the censoring process $C_2(t)$ for subject 2.

Thus, denoting the total number of transactions for subject i by K_i, we have the following observed data for subjects $i = 1, ..., n$ in the period $[a, b]$:

$$(C_i(t), \tilde{N}_i(t) : a \leq t \leq b, \max(T_{i1}, a) \leq t \leq \min(T_{iK_i}, b))_{i=1}^n$$

Our job is now to estimate a chosen target parameter based on this observed data.

Estimation of the target parameter

First, we write up the target parameter for tobacco for the complete data, conditioning on information up to time t. Here, we use the tower property for conditional expectations. Let $N^{\text{tob}}(t) = N(t)I_{\{Q^{\text{tob}}(t)>0,\ P^{\text{tob}}(t)>0\}}$ be a counting process for tobacco transactions. We get:

$$\begin{aligned}
\mu^{\text{tob}}(a,b) &= E(\int_a^b dN^{\text{tob}}(t)) \\
&= E\left[E(\int_a^b dN^{\text{tob}}(t)|\mathcal{F}_{t-})\right] \\
&= E\left[\int_a^b E(dN^{\text{tob}}(t)|\mathcal{F}_{t-})\right],
\end{aligned}$$

We now with to estimate this target based on the intensity of buying tobacco. From [2] we get the definition of an **intensity process** (here conditioning on $\mathcal{F}_{t-}$):

$$\lambda^{\text{tob}}(t|\mathcal{F}_{t-})dt = P(N^{\text{tob}} \text{ jumps in a time interval of length dt around time t } |\mathcal{F}_{t-}) \tag{4}$$

where $\mathcal{F}_{t-}$ (defined earlier) denotes the past up to the beginning of the small time interval dt (everything that has happened just before time t). In a small time interval, dt, N^{tob} either jumps once or does not jump at all. So the probability of a jump in that interval is close to the expected number of jumps in that interval [2]. From the definition of the intensity, we therefore have:

$$\lambda^{\text{tob}}(t|\mathcal{F}_{t-})dt = E(dN^{\text{tob}}(t)|\mathcal{F}_{t-})$$

Inserting this in the target parameter, we get:

$$\begin{aligned}
\mu^{\text{tob}}(a,b) &= E\left[\int_a^b E(dN^{\text{tob}}(t)|\mathcal{F}_{t-})\right] \\
&= E\left[\int_a^b \lambda^{\text{tob}}(t|\mathcal{F}_{t-})dt\right]
\end{aligned}$$

Assuming for now that expected number of tobacco transactions is independent of tobacco shopping history and covariates, we can remove the expectation, and thus wants to calculate (continuing the expression, for $i = 1, ..., n$ observed subjects for the interval $[a, b]$):

$$
\begin{aligned}
\mu^{\text{tob}}(a,b) &= \int_a^b E(dN^{\text{tob}}(t)) \\
&= \sum_{i=1}^{n} \sum_{j\,:\,a\le t_j\le b} \Delta\tilde{N}_i^{\text{tob}}(t_j),
\end{aligned}
$$

where $j = 1, 2, 3, ...$ and $\Delta\tilde{N}_i^{\text{tob}}(t_j) = \tilde{N}_i^{\text{tob}}(t_j) - \tilde{N}_i^{\text{tob}}(t_{j-1})$ is the observed number of tobacco transactions for subject i between calender day t_{j-1} and t_j.

We now wish to investigate this target parameter in lockdown as compared to the preceding period (control period). Therefore, we define the following for the lockdown and control period:

$$
L = [11.03.2020, 18.05.2020] \quad \text{and} \quad \overline{L} = [02.01.2020, 10.03.2020]
$$

So, the null hypothesis that we wish to investigate is:

$$
H_0 : \mu^{\text{tob}}(L) = \mu^{\text{tob}}(\overline{L})
$$

that the expected number of tobacco transactions in lockdown and the control period is the same. In the lockdown and control period we get the following expected number of tobacco transactions:

$$
\begin{aligned}
\hat{\mu}^{\text{tob}}(L) &= 1 + 2 + 5 + 1 = 9 \\
\hat{\mu}^{\text{tob}}(\overline{L}) &= 2 + 1 + 2 = 5
\end{aligned}
$$

Note that in order to compare these without bias, we need to make an assumption about which subjects to include in the analysis, such that we are comparing the same subjects in the two periods. For example, we could choose to include only subjects who had at least one tobacco purchase before 01-01-2020, who did not emigrate or die during any of the two periods in order to get an unbiased, interpretable estimate.

Another way of taking the total observation time per person into account is to model $\hat{\lambda}^{\text{tob}}(t)$ as the tobacco transaction rate per person-day. To estimate this based on the observed data, we start by plugging in an estimator for the intensity, as defined in [6] p. 77 (again assuming constant for $\mathcal{F}_{t-}$):

$$
\hat{\lambda}^{\text{tob}}(t) = \frac{\text{total number of tobacco transactions at time t}}{\text{total observation time at time t}}
$$

So, $\hat{\lambda}^{\text{tob}}(t)$ can be interpreted as the tobacco transaction rate per person-day, if we consider a time unit of days. Plugging this in and continuing the expression, for $i = 1, ..., n$ observed subjects for the interval $[a, b]$, we get ([6] p. 83):

$$
\begin{aligned}
\hat{\mu}^{\text{tob}}(a,b) &= \int_a^b \hat{\lambda}^{\text{tob}}(t)dt \\
&= \frac{\sum_{i=1}^n \sum_{j\,:\,a\le t_j\le b} \Delta \tilde{N}_i^{\text{tob}}(t_j)}{\sum_{i=1}^n \sum_{j\,:\,a\le t_j\le b} C_i(t_j)},
\end{aligned}
$$

where $j = 1, 2, 3, ...$ and $\Delta\tilde{N}_i^{\text{tob}}(t_j) = \tilde{N}_i^{\text{tob}}(t_j) - \tilde{N}_i^{\text{tob}}(t_{j-1})$ is the observed number of tobacco transactions for subject i between calender day t_{j-1} and t_j.

During lockdown for the fictive data, we get:

$$
\begin{aligned}
\hat{\mu}^{\text{tob}}(L) &= \int_L \hat{\lambda}^{\text{tob}}(t)dt \\
&= \frac{\sum_{i=1}^n \sum_{j\,:\,11.03.20\le t_j\le 18.05.20} \Delta \tilde{N}_i^{\text{tob}}(t_j)}{\sum_{i=1}^n \sum_{j\,:\,11.03.20\le t_j\le 18.05.20} C_i(t_j)} \\
&= \frac{\sum_{i=1}^n \sum_{k\ge 1} I\{11.03.20 \le T_{ik}^{tob} \le 18.05.20\}}{\sum_{i=1}^n \sum_{j\,:\,11.03.20\le t_j\le 18.05.20} C_i(t_j)} \\
&= \frac{1+2+0+5+1}{68+68+68+68+(68-13))} \\
&= \frac{9}{327} = 0.028 \ \text{ tobacco trans / pers-day}
\end{aligned}
$$

This gives $0.028 \cdot 30.44 = 0.85$ tobacco transactions per person-month in the lockdown period.

In the control period for the fictive data, we get (what about death?):

$$
\begin{aligned}
\hat{\mu}^{\text{tob}}(\overline{L}) &= \int_{\overline{L}} \hat{\lambda}^{\text{tob}}(t)dt \\
&= \frac{\sum_{i=1}^n \sum_{j\,:\,02.01.20\le t_j\le 10.03.20} \Delta \tilde{N}_i^{\text{tob}}(t_j)}{\sum_{i=1}^n \sum_{j\,:\,02.01.20\le t_j\le 10.03.20} C_i(t_j)} \\
&= \frac{\sum_{i=1}^n \sum_{k\ge 1} I\{02.01.20 \le T_{ik}^{tob} \le 10.03.20\}}{\sum_{i=1}^n \sum_{j\,:\,02.01.20\le t_j\le 10.03.20} C_i(t_j)} \\
&= \frac{2+1+2+0+0}{68+(68-30)+(68-40)+0+0} \\
&= \frac{5}{134} = 0.037 \ \text{ tobacco trans / pers-day}
\end{aligned}
$$

See Figure 1 to see the censored periods subtracted from the person-days in the denominator.

This gives $0.023 \cdot 30.44 = 1.13$ tobacco transactions per person-month in the control period.

Additional notes: data dictionary and definition of kovariates

The following data dictionary gives the variables that we will include to model the expected number of soda transactions:

Name	Label	Levels	Exp effect
hyp	Hypertension for sb users	0 = No hypertension prior to storebox 1 = Hypertension prior to storebox defined by combination of two medicines 180 days before start date	Higher risk Lower risk
sex	Sex at sb start	Female Male	Lower risk Higher risk
agegr	Age group at sb start	< 18 18-29 30-39 40-49 50-69 > 69	Lower risk Higher risk Higher risk Higher risk Lower risk Lower risk
edu	Education at sb start	Basic Vocational training Upper secondary (videregående) Bachelor Higher education	Higher risk Higher risk Lower risk Lower risk Lower risk
householdn	Number in household at sb start	Continous	Higher risk
hotdogpast	Whether hotdogs was bought during the month	1 = hotdogs bought in [t-30,t] 0 otherwise	 Lower risk

Table 2: Data dictionary for variables to include in hotdog model

- Maybe define this way
- We have the following time invariant covariates:

$$Z = \begin{bmatrix} Z^{\text{hypertension}} \\ Z^{\text{sex}} \\ Z^{\text{agegroup}} \\ Z^{\text{education}} \\ Z^{\text{number in household}} \end{bmatrix},$$

and one time dependent covariate, given by

$$\bar{Z}^{\text{past}}(t) = \begin{cases} 1 \text{ if hotdogs were bought during the month [t-30 days, t]} \\ 0 \text{ otherwise} \end{cases}$$

References

[1] Gill, D. R., Andersen, P. K. (1982). *Understanding Cox's Regression Model: A Martingale Approach*, The Annals of Statistics, 10 (4), p. 1100-1120.

[2] Gill, D. R. (1984). *Understanding Cox's Regression Model: A Martingale Approach*, Journal of the American Statistical Association, 79 (386), p. 441-447.

[3] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning* (Second ed.). Springer.

[4] Karr, A. F. (1986). *Point Processes and Their Statistical Inference* (First Ed.). Marcel Dekker.

[5] Last, G., Brandt, A. (1995). *Marked Point Processes on the Real Line* (First Ed.). Springer.

[6] Marubini, E. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies* (First Ed.). John Wiley and Sons.

[7] Pishro-Nik, H. (2014). *Introduction to Probability, Statistics, and Random Processes* (First ed.). Kappa Research.

[8] Tan, P. N., Steinbach, M., Karpatne, A., Kumar, V. (2019). *Introduction to Data Mining* (First ed.). Pearson.

[9] Walley, Rosalind et al. (2016). *Using Bayesian analysis in repeated preclinical in vivo studies for a more effective use of animals*, Pharmaceutical Statistics, No. 15, 2016, p. 277-285.