Exam Applied Mechanism Design and Big Data

Dear student, for this exam you are allowed to consult any reference material, such as lecture notes and exercises. You will be required to both provide answers in the form of text, as through writing short programs. For the programming you are free to use any programming language you want. Your code will not be judged, just the outcome and any additional material that you provide along with the outcome such as plots and explanations in text form. Please make sure that these are separated from the code, or are easily distinguishable from the code (e.g. as in an IPython Notebook). For answers that are not the result of programs, such as mathematical equations and derivations, you are too free to use any means, on paper, via a text processor such as latex, or inline with the coding such as with IPython Notebook, just as long as it is readable!

You are encouraged to provide additional material with your results where possible, such as intermediate results, plots and explanations. These can work in your favour when your result is wrong but (part of) your derivation is correct.

Please read through the sub-exercises first. All information that you need to do any of the sub-exercises is given in the text. You may need an earlier equation for a later sub-exercise. All equations that you need are given, even if you are unable to derive them when asked. If you are unable to answer every question then make sure you answer the ones that you can first: pick your battles.

Gift card analysis

In this exercise you will perform an analysis of gift card usage data. The goal is to model and fit card usage from historical data and subsequently use the model to estimate how many future gift cards will never be used.

An online store sells gift cards for a price of 10 euros each. For each sold gift card the online store has to reserve 10 euros on a dedicated balance account. When the gift card is actually used the 10 euros is deducted again from the balance. However, it is known that some gift cards will never be used, because they are forgotten about, get lost, or are for some other reason never used. For these cards the money is kept on the balance, because the store does not know if cards will be used or not. The online store now asks you to calculate for them how many cards will never be used in the future, and thus how much money they can freely deduct from their balance.

You are asked to model card usage based on historical card usage data and thereby estimate which amount can be freely deducted from the balance. For this purpose you will construct a usage model, fit it to historical usage data using the Maximum Likelihood method, and estimate the deductible amount at 5% confidence level.

0.1 Model construction

Consider card usage as a Poisson process with a usage rate $p(\tau)$. τ is the number of days since the card was purchased. The rate $p(\tau)$ is the probability that a person τ days after the card was purchased still has the card and uses it on that day.

At the moment a card is sold you can calculate the probability that the card will be used τ days later. This probability is:

$$P(\tau) = p(\tau)e^{-\int_0^\tau dsp(s)} \tag{1}$$

It is the combination of $p(\tau)$, the probability that the card is used **on** the day, **and** $P_0(\tau) = e^{-\int_0^\tau dsp(s)}$, the probability that the card was not used **before** that day.

1. Give the Poisson process, i.e. the first order differential equation that results in $P_0(\tau)$, the probability that τ days after card purchase the card has not been used yet.

We will now focus on the usage rate $p(\tau)$. We model it as the result of two influences. On the one hand there is the probability that the card is used. But

on the other hand the card can only be used if it hasn't been lost. So we model the rate as $p(\tau) = pQ_0(\tau)$, a combination of $Q_0(\tau)$ the probability that τ days after purchase the card is not lost **and** a p the probability that the card is used. p is considered constant, the same for every day. $Q_0(\tau)$ is itself the result of a Poisson process, and is modelled as $Q_0(\tau) = e^{-q\tau}$ where q is a constant loss rate.

2. Give the Poisson process, i.e. the first order differential equation that results in $Q_0(\tau)$, the probability that τ days after card purchase the card has not been lost yet.

The usage rate is now given by

$$p(\tau) = pe^{-q\tau} \tag{2}$$

3. Work out $P(\tau)$ to arrive at

$$P(\tau) = pe^{-q\tau - \frac{p}{q}\left(1-e^{-q\tau}\right)} \tag{3}$$

$P(\tau)$ is your model for gift card usage with two parameters p and q. To do the parameter estimation and ultimately estimate the number of cards that will never be used we need two more quantities. The probability that a card purchased at day t_0 has been used somewhere before day t_1 is

$$\text{Norm}(t_0, t_1) = 1 - e^{-\frac{p}{q}\left(1-e^{-q(t_1-t_0)}\right)} \tag{4}$$

This will serve as a normalization in parameter estimation.

4. Derive equation 4

The probability that a card, that was purchased at day t_0 and has not yet been used by day t_1, will never be used any more is

$$P_{\text{never}}(t_0, t_1) = e^{-\frac{p}{q}\left(e^{-q(t_1-t_0)}\right)} \tag{5}$$

5. Derive equation 5

0.2 Parameter estimation

To estimate the parameters p and q of your gift card usage model you will use the Maximum Likelihood method. You will estimate the parameters on a historical dataset of gift cards that have been used. Then you will use the parameters on a dataset of gift cards that have not been used yet to estimate how many of these that will never be used in the future. This number times 10 euros is the estimate for the amount of money that can be deducted from the balance. The dataset that you received has gift cards purchased between 3-3-2013 and 7-4-2014. Some of these have been used, others have not yet.

The probability $P_r(t_0, t)$ that a card purchased at day t_0 was used at day t is given by

$$P_r(t_0, t) = P(t - t_0)/\text{Norm}(t_0, t_1) \tag{6}$$

with $P(t - t_0)$ from equation 3 and $\text{Norm}(t_0, t_1)$ the normalization factor from equation 4 with t_1=7-4-2014.

If you have a dataset of N gift cards with purchase days t_{0_i} and use days t_i, then the *log-likelihood* of the dataset is given by

$$L = -2 \sum_{i=1}^{N} \text{Log}\left(P_r(t_i, t_{0_i})\right) \tag{7}$$

where $P_r(t_i, t_{0_i})$ is the model probability from equation 6 that card i purchased at day t_{0_i} was redeemed at day t_i. By minimizing the log-likelihood with respect to parameters q and p (and thus maximizing the likelihood) you obtain the values for q and p that are most likely to have produced the dataset.

Read in the provided dataset and split it into two, one of cards that have been used, and one of cards that have not been used yet. You are allowed to express purchase and use days as the number of days since 3-3-2013 to avoid working with datetimes.

6. Explain why the normalization factor is needed in equation 6.
7. For the used cards minimize the log-likelihood of equation 7 to obtain your estimate of q and p.
8. Use your estimates of q and p to estimate how many of the unused cards will never be used in the future, and the amount of money that can be deducted from the balance.

0.3 Bootstrapping

Previously you estimated the parameters q and p from the set of redeemed gift cards. Call these q_0 and p_0. Naturally there is an uncertainty on these parameters, and thus an uncertainty on the amount of money that can be deducted from the balance. To estimate these uncertainties you will use bootstrapping. You will use the parameter estimates q_0 and p_0 to generate new datasets, then fit each dataset to obtain a q and p again. The variation in q and p that you generate in this way is a measure for the uncertainty on the parameters.

Because gift card usage is a probabilistic process the rates that are seen in a (finite) dataset are not necessarily exactly equal to the real underlying usage and loss rates. Only with a very large (infinite) dataset will you see the exact real rates. If you could take another dataset from the one you have then you will generally see slightly different rates. And if you could take various datasets you would see the variation in the rates that is the result of having limited data. You will simulate this process of taking various datasets, by taking your q_0 and p_0 as the 'real' rates and randomly generating new datasets with these values, to uncover the variation/uncertainty on these rates.

- Write a Monte Carlo simulation that randomly generates a use day according to $P(t - t_0)$ of equation 3 for each gift card in the entire dataset using q_0 and p_0 for the model parameters.
- Split this generated dataset in two, one of cards with use days before or at 7-4-2014, and one of cards without use days or with use days after 7-4-2014. Just like the real original dataset.

- For the generated used cards minimize the log-likelihood of equation 7 to obtain estimates of q and p and also use these on the generated unused cards to obtain the amount of money that can be deducted from the balance.

9. Run the Monte Carlo simulation 1000 times, and report the amount for which 95% of the simulations the deductible amount of money is equal or larger. This is at 95% confidence level the amount that can be safely deducted from the balance.

0.4 Dynamical System optimization

The foregoing modelling started with $P_0(\tau)$, the probability that a gift card has not been used yet τ days after purchase. After modelling $p(\tau)$ through a constant loss rate q and a constant usage rate p, $P_0(\tau)$ works out to

$$P_0(\tau) = e^{-\frac{p}{q}\left(1-e^{-q\tau}\right)} \tag{8}$$

10. Show that equation 8 is in fact the optimal solution of the functional

$$S[P_0] = \int_0^\infty d\tau q \log(P_0) + \frac{-\dot{P}_0}{P_0} \log(\frac{-\dot{P}_0}{P_0}) \tag{9}$$

    (Hint: p will appear through an integration constant)

11. Give an interpretation of the functional.