Marketing Analytics Tutorial 2

Consumer and Customer Analytics

Part A) Binary Choice: Logistic Regression

Exercise 1

We will work with sales data[1] on season passes to an amusement park. Thereby, the sales differ according to two factors: the channel (email, postal mail, in-person) used to deliver the promotion and whether the promotion included the ticket in a bundle with free parking or not. The following tables provide the number of customers that did or did not buy the season pass for each of the possible combinations of promotion channel and bundle offer.

Customers that bought the season pass

	Bundle	No Bundle
Mail	242	359
Park	639	284
Email	38	27

Customers that did not buy the season pass

	Bundle	No Bundle
Mail	449	278
Park	223	49
Email	83	485

a) Store the data tables in R, also including the descriptive marginal labels (Ticket, Channel, Offer).
Hint: First create a single long vector containing the count values and then adjust the dimensions of the object to create a 3x2x2 array.
b) Convert the data into long form to get a data frame of individual customer observations.
Hint: First convert the data object to class "table" and use the package ***vcdExtra****.*
c) Using logistic regression, estimate whether the promotion of the bundle (season pass + free parking) has an effect on the sales of season passes to the amusement park.
d) Manually calculate the association between season pass purchase and the bundle factor by examining the ratio of success to non-success.
Hint: Use the logistic distribution function.
e) How does the model from c) change if channel is added as a predictor?
f) Is there an interaction between the bundle offer and channel in their relationship with the purchase of season passes?

Exercise 2

We will use a simulated data set that contains customer transactions on an e-commerce website together with satisfaction data. The variables are described in the following table:

[1] All data sets in Part A of this Exercise were taken from "R for Marketing Research & Analytics" by Chris Chapman & Elea McDonnell Feit, http://r-marketing.r-forge.r-project.org/data/.

Variable	Description
acctAge	Tenure of the customer (in months)
spendToDate	Total lifetime spending of the customer
spendMonth	Spending in the most recent month
visitsMonth	Number of visits to the website in the most recent month
satSite	Satisfaction rating of the website (1 – 10)
satPrice	Satisfaction rating of the prices (1 – 10)
satQuality	Satisfaction rating of the product quality (1 – 10)
satOverall	Overall satisfaction rating (1 – 10)
region	Geographic region of the customer (US)
coupon	Dummy variable whether a coupon for some promoted product was sent to the customer
purchase	Dummy variable whether the promoted product was purchased

First of all, load the data into R from the web using the following command.

```
salesdata <- read.csv("https://goo.gl/4Akgkt")
```

a) Using logistic regression, estimate the relationship between customers receiving the coupon and their purchasing behavior of the promoted product.
b) Manually calculate the association between purchase and the coupon factor by examining the ratio of success to non-success.
c) How does the model from a) change if region, overall satisfaction, and total spending are added as predictors?
d) Is there an interaction between the coupon variable and overall satisfaction in their relationship with the purchase of the promoted product?

Part B) Product Choice: Multinomial Logit Model

Exercise 3

We use a data set[2] containing the heating system choices for single-family Californian houses. The builders of the houses can choose bewteen five different systems: gas central (gc), gas room (gr), electric central (ec), electric room (er), and heat pump (hp). The variables of the data set are described in the following table:

Variable	Description
idcase	Observation number
depvar	Specifies the chosen alternative (gc, gr, ec, er, or hp)
ic.*alt*	Installation cost ($) for each alternative (ic.gc, ic.gr, ic.ec, ic.er, and ic.hp)
oc.*alt*	Annual operating cost ($) for each alternative (oc.gc, oc.gr, oc.ec, oc.er, and oc.hp)

[2] This data set is part of the mlogit package. Exercise 3 closely follows tasks suggested by Kenneth Train and Yves Croissant.

income	Household annual income
agehed	Age of household lead
rooms	Number of rooms in the house
region	Factor variable for region of the house, either northern coastal (**ncostl**), southern coastal (**scostl**), mountain (**mountn**), or central valley (**valley**)

First of all, load the data into R from the package using the following command.

data("Heating", package = "mlogit")

a) Currently, the cost attributes are stored in individual variables, or one for each alternative. However, for the remainder of this exercise, a single variable for either cost type is required whose value then differs with the choice alternative. To achieve this format, index the data set on the two variables **idcase** and **depvar**.
Hint: Use the package (and function) ***dfidx*** *and specify the data, choice, and varying arguments accordingly.*
b) Estimate a multinomial logit model with installation and operating costs, but do not include an intercept.
c) How close do the model's estimated probabilities match the shares of houses that actually chose each heating system?
d) Calculate the willingness to pay (WTP) for a $1 reduction in annual operating costs.
e) What is the discount rate r implied by this WTP? Assume a sufficiently long life T of the operating system so that the current value of operating costs cv approaches oc/r for an increasing lifetime.
Hint: The current value of operating costs cv is given by the discounted sum of the operating costs over the life of the system: $cv = \sum_{t=1}^{T} oc/(1+r)^t$
f) Estimate a multinomial logit model that imposes the following restraint on the discount rate: $r = 0.12$. Again, cost is the independent variable and there should be no intercept.
g) Test the assumption made for the discount rate in f) using a likelihood ratio test.
Hint: Use the package lmtest.
h) Estimate a multinomial logit model with installation and operating costs and include alternative-specific constants. Take the heat pump as the base alternative.
i) How close do this model's estimated probabilities now match the shares of houses that actually chose each heating system?
j) Again, compute the corresponding WTP and discount rate.
k) Relate the magnitude of upfront installation costs to household income and add this to the model instead of installation cost by itself. Include operating costs as before.
l) Next, use alternative-specific income effects instead, again specifying the heat pump as the base alternative. Test whether including income effects leads to a better model than one using alternative-specific constants.
m) Finally, we use a multinomial logit model for prediction. The Californian government is considering to offer a 15% rebate on the installation cost of heat pumps. They want to predict the effect of this proposal on the choice of heating systems. Use the estimated coefficients from h) to calculate the new probabilities and predicted shares in the case of the cheaper installation costs for a heat pump.

Part C) Markov Chain Model

Exercise 4

Imagine a popular restaurant is now offering a weekly lunch, where customers can choose between 3 dish options: Pasta, Rice, and Salad. The individual dishes vary weekly, but the base component of each of the 3 alternatives remains the same. Due to the weekly menu, assume that customers will only get lunch at the restaurant once a week.

a) In the first week, 70% of customers ate pasta for lunch, 20% rice, and 10% salad. Create a vector in R with the starting probabilities of the 3 menu options.

b) Since the restaurant just started offering the lunch menu, there is no historical data from which to derive the likelihood of customers switching between the 3 lunch options in the following weeks. Instead, the restaurant used its regular dinner orders to estimate the purchasing behavior of their lunch customers. The resulting transition probabilities are given by the following table. Store the values as a matrix in R.

t +1 t	Pasta	Rice	Salad
Pasta	0.1	0.6	0.3
Rice	0.5	0.4	0.1
Salad	0.2	0.8	0

c) Use the information you have to compute the estimated popularity of each of the 3 lunch alternatives in the second week.
Hint: Think about matrix products.

d) Do the same popularity prediction for the third week in which the restaurant offers the lunch menu.

e) Find the long-term steady state shares of the 3 lunch options.
Hint: Use the package ***expm*** *for matrix exponentiation.*

Exercise 5

We will use a public data set[3] of a web server log that contains requests made to a web server for the US Environmental Protection Agency (EPA). The variables are described in the following table:

Variable	Description
rawhost*	User IP address
host	Unique identifier (observation number)
timestamp*	Time stamp of request
datetime	Date and time information (combined)
request*	Action request, page/file requested and communication protocol
reqtype	Browser action (either **GET**, **POST**, or **HEAD**)
pageytpe	Content (file) requested (**html**, **gif**, **pdf** or **other**)
page	Actual page/file requested

[3] This data set was taken from "R for Marketing Research & Analytics" by Chris Chapman & Elea McDonnell Feit, http://r-marketing.r-forge.r-project.org/data/.

* These variables are given in the raw data set but should not be used for the analysis. Please use the cleaned alternative variables given instead (**rawhost -> host, timestamp -> datetime, request -> reqtype, pagetype & page**).

status	Request status (e.g. 404 Not Found error)
bytes	Number of bytes transmitted

First of all, load the data into R from the web using the following command.

```
epadata <- readRDS(gzcon(url("https://goo.gl/s5vjWz")))
```

a) What are the overall 5 most common page requests? Then, only consider HTML requests and determine the 5 most common requests.
b) When are users most active in making requests? Choose an appropriate visualization method.
c) How many unique users are registered in the data? Are there very active users? How many users account for 80% of the requests?
Hint: Use a cumulative distribution function.
d) To prepare for the sequence analysis, the data set must be adjusted. First, order the requests according to `host` (primary key) and `timestamp` (secondary key). Then, calculate the time gap between requests to determine individual sessions. Take 15 minutes of inactivity as the cutoff value.
Please add 3 new variables to the data set, one for the time difference, one conveying whether a request is part of a new session, and one giving a running total of the overall number of sessions.
Hint: There are three cases that can indicate a new session.
e) How many unique sessions are there? What is the average number of requests per session?
Hint: Use the function `rle()` to get the number of requests made in each session.
f) Create a subset of the EPA data to include only HTML pages that are among the 20 most popular pages.
g) For later calculations, the data should be reformatted so that each session is stored in a single line. To do so, first create a list with separate elements for each session and remove all sessions consisting of a single request. Then, add end states to each session and reformat to have individual lines for each session.
Hint: Use split( ) to first create the list.
h) Estimate the Markov Chain using `fitMarkovChain( )`. In order to do so, import the data from the previous task into a clickstream object.
i) Visualize the transition matrix in a heat map using the package superheat.
j) Use the transition likelihoods to predict the next page request in session 110. Similarly, predict the next two likely pages for session 160.