Assignment $\mathcal{N}^{\underline{o}}$ 2

released: 18.03.2020 at 16:00 **due:** 23.04.2020 at 23:59

Task 1 — 3 points

The following table contains the estimates of a logistic regression model.

						95% C.I. for OR	
	Est.	s.e.	z	p-value	OR	lower	higher
X_1	-0.868		-2.365	0.018		0.205	0.865
X_2	2.404	0.601	4.000	<0.001			
X_3				<0.001		0.010	0.074

Fill in the missing information
(Please report formulas and computation.)

Task 2 — 3 points

During the lecture, we have considered the three systems of hypotheses for the parameters of the MNRM:

- $H_0 : \beta_{jm} = 0 \quad \text{vs.} \quad H_1 : \beta_{jm} \neq 0$
- $H_0 : \beta_{j1} = \ldots = \beta_{j(M-1)} = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_{jm} \neq 0, \ \forall m$
- $H_0 : \beta_{j1} = \ldots = \beta_{j(M-1)} = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_{jm} \quad \forall j, m$

Could you specify another pair of hypotheses H_0 and H_1 for the parameters of the MNRM that we might want to test? Justify the answer.
(You have only to define H_0 and H_1. Defining the statistic and the rejection region of the test is not required!)

Task 3 **8 points**

The website *blablabla.com* sells magazine subscriptions. The related company is going to plan a large e-mail marketing campaign. All of the e-mails that will be sent will go to customers that have previously bought a magazine subscription at *blablabla.com* and who have not opted out of receiving e-mails.

The magazines advertised in each e-mail will be automatically selected for each customer when the e-mail is generated in order to maximize the probability that the customer will buy. The website *blablabla.com* will only include ads for three magazines in each e-mail in a row at the top of the message so that it is likely that the ads will appear in the e-mail preview (and therefore actually be viewed without the receiver actually having to open the e-mail). Moreover, the managers believe that including more ads is ineffective.

To evaluate the efficacy of the campaign, the company run an experiment. They sent 673 e-mails to customers containing the ad for the "Art with you" magazine and recorded whether or not the customer purchased this magazine.

The company has also collected data on the customers by matching the information provided by third party data (which can be purchased from data sources such as the credit scoring agencies) and the recipient of the e-mails when he/she made a purchase at *blablabla.com*.

The data set `magazine.cvs` contains the following information:

- Purchased "Art with you" magazine (Buy = 1 if purchased "Art with you", 0 otherwise)
- Household Income (Income; rounded to the nearest $1,000.00$)
- Gender (IsFemale = 1 if the person is female, 0 otherwise)
- Marital Status (IsMarried = 1 if married, 0 otherwise)
- College Educated (HasCollege = 1 if has one or more years of college education, 0 otherwise)
- Employed in a Profession (IsProfessional = 1 if employed in a profession, 0 otherwise)
- Retired (IsRetired = 1 if retired, 0 otherwise)
- Not employed (Unemployed = 1 if not employed, 0 otherwise)
- Length of Residency in Current City (ResLength; in years)
- Dual Income if Married (Dual = 1 if dual income, 0 otherwise)

- Children (Minors = 1 if children under 18 are in the household, 0 otherwise)
- Home ownership (Own = 1 if own residence, 0 otherwise)
- Resident type (House = 1 if residence is a single family house, 0 otherwise)
- Race (White = 1 if race is white, 0 otherwise)
- Language (English = 1 if language is English, 0 otherwise)
- Previously purchased an art magazine (PrevArt = 1 if previously purchased an art magazine, 0 otherwise).
- Previously purchased a cinema magazine (PrevCin = 1 if previously purchased a cinema magazine)

(1) Estimate a logistic regression model, with Buy as the dependent variable and Gender, Not Employed, Income and Own as explanatory variables. Is there any explanatory variable you would remove from the model? On the basis of which test? Remove the variable(s) and named the resulting model as Model 1.

(2) Add to Model 1 the variables PrevArt and PrevCinema. Name the resulting model as Model2. Compare Model 1 and Model 2 using an appropriate test and comment on the results.

(3) The data set includes many potential explanatory variables. Automatic procedures to select the "best" model are implemented in any software. One of this procedure is called "backward elimination" and performs the following steps:

Step 0 The binary logistic regression model including all the potential explanatory variables is fitted and the likelihood L^0 of the full model is computed.

Step 1 i) All the possible p models obtained by excluding one of the possible p variables are estimated and the model with the lowest AIC (or BIC) is considered.

ii) Let X_j be the variable that is excluded and $L_j^{(1)}$ be the likelihood of the model without X_j. The significance of X_j is tested using the G statistic

$$G_j^{(1)} = -2 \log \left[\frac{L_j^{(1)}}{L_j^{(0)}} \right]$$

If X_j is not significant $\left(\text{i.e., } G_j^{(1)} < \chi^2_{1,1-\alpha}\right)$, the considered model is selected, otherwise the full model is selected and the selection procedure ends.

Step 1 is repeated until the variable that is dropped at the generic step k is significant.

Use the commands:

```
fullmodel<- glm(Buy ~,family='binomial', data=advertisement)
mod.fin <- step(fullmodel, direction = 'backward')
summary(mod.fin)
```

to select the "best" model. Interpret the parameters describing the relation between Buy and all the selected explanatory variables.

Task 4 **6 points**

The dataset `alligator.csv` contains aggregated data concerning the primary food choice of 219 alligators captured in four Florida lakes.
The variables in the dataset are:

- lake: lake of capture (Hancock, Oklawaha, Trafford, and George)
- size: size of the alligator (small= $\leq$ 2.3 meters, large=> 2.3 meters)
- sex: gender of the alligator (0=male, 1=female)
- food: primary food choice (fish, invertebrate, reptile, bird, and other)
- count: number of alligators for each combination of the variables lake, size and sex

A biologist is interested in determining if there is a statistical association between the primary food choice of the alligators and their gender, size and the lake in which they live. Since food is a nominal variable, he would like to estimate a MNRM.

(1) Write down the formulation of the MNRM assuming "fish" as reference category.

(2) Estimate the model using R. Interpret the parameters of the log-odds of reptile vs. fish.

(3) What is the odds ratio of bird versus other for an alligator of small size relative to an alligator of large size? Interpret this odds ratio.

(4) Test the global significance of the variable lake.

(5) Test the fit of the model.