Assignment № 3

released: 24.04.2020 at 10:00 **due:** 01.06.2020 at 23.59

Task 1 — 6 points

The data in the file `frequency.cvs` was collected in January-November 2018 with an online survey open to anyone. Respondents were motivated to take the survey by the opportunity to receive personalized results.

The data set contains the following variables:

- Q2: "I think about how my actions affect the environment"
  (5 point rating scale: 1=never, 2=rarely, 3=occasionally, 4=often, 5=always)
- education "How much education have you completed?"
  (1=Less than high school, 2=High school, 3=University degree, 4=Graduate degree)
- urban: "What type of area did you live when you were a child?"
  (1=Rural (country side), 2=Suburban, 3=Urban (town, city))
- gender: "What is your gender?"
  (1=Male, 2=Female, 3=Other)
- age: "How many years old are you?"
- race: "What is your race?"
  (11=Asian, 12=Arab, 13=Black, 14=Indigenous Australian, 15=Native American, 16=White, 17=Other)
- married: "What is your marital status?"
  (1=Never married, 2=Currently married, 3=Previously married)
- familysize: "Including you, how many children did your mother have?"

For all the variables, the value 0 is used to represent missing values.

We would like to test whether there is an association between $Q2$ and the other variables in the data set using the ordered logistic regression model.

(1) Import the data. Recode the variable race so that all the categories that have a frequency lower than 100 are combined into the category "other". Recode missing values in the dataset into NA.

(2) Consider the OLRM intercept model having $Q2$ as the dependent variable (i.e., the OLRM without explanatory variables). Write down the formula of the model and compute the estimate of the intercept parameters without estimating the model.

(3) Use the backward selection procedure to select the best fitting OLRM. The code we used in the practical session on OLRM returns an error. Explain why the error occurs and correct the code so that you can select the best fitting model.

(4) Consider the best fitting OLRM returned by the backward procedure. Test its goodness of fit using an adequate test.

(5) Interpret the model coefficients.

Task 2 **4 points**

The assumption of parallel slopes (or proportional odds) underlying the OLRM implies that the coefficients describing the relationship between the ordinal dependent variable and the explanatory variables are the same between any pair of adjacent categories.

Develop a procedure to informal assess the assumption of parallel slopes.

(1) Describe the procedure and its logic.

(2) Apply the developed procedure to check whether the parallel slopes assumption is matched by the data in Task 1 and the explanatory variables included in the best fitting model (Task 1, point 2).

(Hint: Think about the assumption in terms of logit or odds ratio. Base the procedure on the estimation of binary logistic regression models (one for each category but M) having the form

$$\log\left[\frac{P(Y^m = 1|X)}{P(Y^m = 0|X)}\right] = \beta_0 + \beta_1 X$$

The dependent variable Y^m is a binary variable taking value 1 if the ordinal dependent variable Y is greater than m, and 0 otherwise. Consider one explanatory variable at a time.)

Task 3 **7 points**

The data set `medpar.csv` is an excerpt from US national Medicare inpatient hospital database. It contains 1495 observations on the following variables:

- los: length of hospital stay (in days)
- hmo: patient belongs to a Health Maintenance Organization (1), or private pay (0)
- white: patient identifies themselves as primarily Caucasian (1) in comparison to non-white (0)
- age80: patient age 80 and over (1), or age < 80 (0)
- type: a three-level explanatory variable related to the type of admission (1 = elective, 2 = urgent, and 3 = emergency)

We would like to investigate whether there is an association between the length of hospital stay and the other variables.

(1) Import the data. Based on the descriptive statistics, do you expect that there is a significant relation between `los` and `type`? Justify your answer.

(2) Estimate a Poisson regression model with `los` as dependent variable and `type` as explanatory variable. Name this model Model 1. Interpret the parameters of the model (including the intercept).

(3) Add to the model in (2) the explanatory variables `age80`, `hmo` and `white`. Name the resulting model Model 2. Test whether Model 2 has a better fit than Model 1.

(4) Interpret the parameter related to the variable `age80`.

(5) Test whether the equi-dispersion assumption is matched by the data. If that is not the case:

   i. Estimate an adequate model including all the explanatory variables. Name this model Model 3.

   ii. Compare the results of Model 2 and Model 3. Are they the same?

Task 4 **3 points**

Could you think of a test for equi-dispersed data based on the comparison of two models rather than testing the over-dispersion parameter α of the negative binomial regression model?

(1) Define the null and the alternative hypotheses, the test statistic and the rejection region. Explain the logic of the test.

(2) Apply the test to the data in Task 3.