Ordered Choice Logit Model for Career Satisfaction of Programmers in Eastern Europe

Jorge Bueno Perez - #419034 & Lashari Gochiashvili #425198

Contents

1) Abstract: **1**

2) Introduction: **1**

3) Literature review: **1**

4) Data description: **1**
4.1) Data preparation: 2
4.2) Data representation: 8

5) Application of Econometric models: **10**
5.1) Ordered Choice Models: 10
5.2) Joint significance test: 10
5.3) Goodness of fit test: 10
5.4) Marginal effects: 10

6) Results: **10**

7) Bibliography: **10**

```
library(dplyr)
library(knitr)
library(psych)
library(knitr)
library(dplyr)
```

1) Abstract:

2) Introduction:

3) Literature review:

4) Data description:

`Stack Overflow` is a community forum for programmers. In this forum the users are able to ask questions about their programs, and one is able to reply to them sharing his knowledge.

Being one of the `most known community forums of programmers around the world`, it is a good place to develop a survey in order to know more information about programmers. Every year Stack Overflow develops a `voluntary survey`; however, the main data object of this paper is the survey developed in `January 2018`, which is available in the `website Kaggle`. The survey was answered by almost `98 thousand respondents` and `129 different questions`, rows and columns respectively of the dataset.

[1] "The dataset survey initialy has 98855 rows and 129 columns"

- More details about `Stack Overflow`: https://stackoverflow.com/company
- Next it can be found the link to the data in `kaggle`: https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey

4.1) Data preparation:

During the process of cleaning the data several strategies were applied.

First of all, we `selected several columns`, based on our `intuition`, trying to answer the following question `"Which questions of the survey can explain the best the career satisfaction of the programmers?"`.

`Eight survey questions were selected`, together with the dependent variable of our model – `CareerSatisfaction`. More details about the questions can be found below, linked to the columns of our data:

- `CareerSatisfaction` – Overall, how satisfied are you with your career thus far?
- `Country` - In which country do you currently reside?
- `FormalEducation` - Which of the following best describes the highest level of formal education that you have completed?
- `CompanySize` - Approximately how many people are employed by the company or organization you work for?
- `YearsCoding` - For how many years have you coded professionally (as a part of your work)?
- `JobSatisfaction` – How satisfied are you with your current job? If you work more than one job, please answer regarding the one you spend the most hours on.
- `ConvertedSalary` – What is your current gross salary (before taxes and deductions)? Please enter a whole number in the box below, without any punctuation. If you are paid hourly, please estimate an equivalent weekly, monthly, or yearly salary. If you prefer not to answer, please leave the box empty. *NOTE: (There was an option to select the currency. After, the salary was converted to annual USD salaries using the exchange rate on 2018-01-18, assuming 12 working months and 50 working weeks)
- `Hobby` - Do you code as a hobby?
- `Age` - What is your age? If you prefer not to answer, you may leave this question blank.

```r
survey <- survey %>%
  dplyr::select(CareerSatisfaction, Country, FormalEducation,
                CompanySize, YearsCoding, JobSatisfaction,
                ConvertedSalary, Hobby, Age)
```

In some questions the respondents had the option to not reply, having the option to leave in blank some of them. For this reason, we `excluded the missing value and nulls`, consequently we assume that `programmers are people who are employed and receive salary`. This assumption was motivated by the empty values in the columns `ConvertedSalary` and `CompanySize`. Apart from that, we considered `only programmers that gave their Age`.

```r
survey <- survey %>%
  filter(ConvertedSalary!=0) %>%
  na.omit(survey)
```

We will check if there is any `NA` or `zero` value:

```r
any(is.na(survey))
```

[1] FALSE

```r
any(survey$ConvertedSalary <= 0) %>%
  knitr::knit_print()
```

[1] FALSE

Both of the results were `FALSE`, hence we can continue farther.

The next step was to select `ten Eastern countries` to study: `Poland`, `Czech Republic`, `Hungary`, `Slovakia`, `Romania`, `Bulgaria`, `Turkey`, `Moldova`, `Belarus` and `Ukraine`.

```r
survey <- survey %>%
  filter(
    Country %in% c(
      "Poland",
      "Czech Republic",
      "Hungary",
      "Slovakia",
      "Romania",
      "Bulgaria",
      "Turkey",
      "Moldova",
      "Belarus",
      "Ukraine"
    )
  )
```

Subsequently, each variable selected was analyzed separately, in order to find some `outliers` in case of `numeric` variables and to `encode the data` in case of `ordinal` features:

- `CareerSatisfaction`: This `characteristic` feature is the `dependent variable`, and has `seven levels` which `can be ordered in a logical way`, from extremely dissatisfied to Extremely satisfied. For the purpoese of this analysis it will be transformed to `three levels`, because it is difficult to define the difference between `extremly disatisfied`, `moderately disatisfied` and `slighty disatisfied`, the same for `satisfied`.

Below more details about the encoding can be found:

- `Disatisfied`: "Extremely dissatisfied", "Moderately dissatisfied" and "Slightly dissatisfied", encoded to 1
- `Neutral`: "Neither satisfied nor dissatisfied", encoded to 2
- `Satisfied`: "Slightly satisfied", "Moderately satisfied" and "Extremely satisfied", encoded to 3

[1] "Slightly satisfied" "Slightly dissatisfied"
[3] "Moderately dissatisfied" "Extremely satisfied"
[5] "Moderately satisfied" "Neither satisfied nor dissatisfied" [7] "Extremely dissatisfied"

```r
survey$CareerSatisfaction <-
  plyr::revalue(
    survey$CareerSatisfaction,
    c(
      "Extremely dissatisfied" = 1,
      "Moderately dissatisfied" = 1,
      "Slightly dissatisfied" = 1,

      "Neither satisfied nor dissatisfied" = 2,

      "Slightly satisfied" = 3,
      "Moderately satisfied" = 3,
      "Extremely satisfied" = 3
    )
  ) %>%
  as.factor()
```

- `Country`: This `characteristic` feature was previously described. However, after removing missing values there is one country less (Moldova), hence we will have `nine countries`. It `cannot be ordered in a logical way`, hence it will be encoded, but it will be transformed to `factor`.

[1] "Poland" "Romania" "Turkey" "Slovakia"
[5] "Bulgaria" "Belarus" "Czech Republic" "Ukraine"
[9] "Hungary"

```
survey$Country <- survey$Country %>%
  as.factor()
```

- `FormalEducation -` This `characteristic` feature has `nine levels` that `can be ordered in a logical way`, from "I never completed any formal education", to "Other doctoral degree (Ph.D, Ed.D., etc.)".

Because this feature can be `ordered in a logical way`, it will be grouped in three categories:

- `Low level of education`: "I never completed any formal education", "Primary/elementary school" and "Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)", encoded to 1
- `Medium level of education`: "Professional degree (JD, MD, etc.)", "Some college/university study without earning a degree" and "Associate degree", encoded to 2
- `High level of education`: "Bachelor's degree (BA, BS, B.Eng., etc.)", "Master's degree (MA, MS, M.Eng., MBA, etc.)" and "Other doctoral degree (Ph.D, Ed.D., etc.)", encoded to 3

Below more details about the encoding can be found:

[1] "Some college/university study without earning a degree"
[2] "Master's degree (MA, MS, M.Eng., MBA, etc.)"
[3] "Bachelor's degree (BA, BS, B.Eng., etc.)"
[4] "Professional degree (JD, MD, etc.)"
[5] "Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)" [6] "Other doctoral degree (Ph.D, Ed.D., etc.)"
[7] "Primary/elementary school"
[8] "Associate degree"
[9] "I never completed any formal education"

```
survey$FormalEducation <- plyr::revalue(
  survey$FormalEducation,
  c(
    "I never completed any formal education" = 1,
    "Primary/elementary school" = 1,
    "Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)" = 1,

    "Professional degree (JD, MD, etc.)" = 2,
    "Some college/university study without earning a degree"  = 2,
    "Associate degree" = 2,

    "Bachelor's degree (BA, BS, B.Eng., etc.)"  = 3,
    "Master's degree (MA, MS, M.Eng., MBA, etc.)" = 3,
    "Other doctoral degree (Ph.D, Ed.D., etc.)" = 3
  )
) %>%
  as.factor()
```

- `CompanySize` - This `characteristic` feature has `eight levels` that `can be ordered in a logical way`, from “Fewer than 10 employees”, to “10,000 or more employees”.

Because this feature can be `ordered in a logical way`, it will be grouped in three categories:

- `Small company size`: “Fewer than 10 employees”, “10 to 19 employees” and “20 to 99 employees”, encoded to 1
- `Medium company size`: “100 to 499 employees” and “500 to 999 employees”, encoded to 2
- `Big comopany size`: “1,000 to 4,999 employees”, “5,000 to 9,999 employees” and “10,000 or more employees”, encoded to 3

Below more details about the encoding can be found:

[1] “20 to 99 employees” “10,000 or more employees” [3] “1,000 to 4,999 employees” “5,000 to 9,999 employees”
[5] “100 to 499 employees” “500 to 999 employees”
[7] “Fewer than 10 employees” “10 to 19 employees”

```
survey$CompanySize <- plyr::revalue(
  survey$CompanySize,
  c(
    "Fewer than 10 employees" = 1,
    "10 to 19 employees" = 1,
    "20 to 99 employees" = 1,

    "100 to 499 employees" = 2,
    "500 to 999 employees" = 2,

    "1,000 to 4,999 employees"  = 3,
    "5,000 to 9,999 employees" = 3,
    "10,000 or more employees"  = 3
  )
) %>%
  as.factor()
```

- `YearsCoding`: This `characteristic` feature has `eleven levels` that `can be ordered in a logical way`, from “0-2 years”, to “30 or more years”.
  Because this feature can be `ordered in a logical way`, it will be grouped in three categories:

- `Short experience coding`: “0-2 years” and “3-5 years”, encoded to 1

- `Medium experience coding`: “6-8 years”, “9-11 years”, “12-14 years” and “15-17 years”, encoded to 2

- `Long experience coding`: “18-20 years”, “21-23 years”, “24-26 years”, “27-29 years”, “30 or more years” and encoded to 3

Below more details about the encoding can be found:

[1] “3-5 years” “6-8 years” “0-2 years” “12-14 years”
[5] “9-11 years” “24-26 years” “18-20 years” “15-17 years”
[9] “30 or more years” “21-23 years” “27-29 years”

```
survey$YearsCoding <-
  plyr::revalue(
    survey$YearsCoding,
    c(
      "0-2 years" = 1,
      "3-5 years" = 1,
      "6-8 years" = 2,
      "9-11 years" = 2,
      "12-14 years" = 2,
      "15-17 years"  = 2,

      "18-20 years" = 3,
      "21-23 years"  = 3,
      "24-26 years"  = 3,
      "27-29 years" = 3,
      "30 or more years" = 3
    )
  ) %>%
  as.factor()
```

- `JobSatisfaction`: This `characteristic` feature has `seven levels` which `can be ordered in a logical way`, from "Extremely dissatisfied", to "Extremely satisfied", the same levels as the `dependent variable`.

Because this feature can be `ordered in a logical way`, it will be grouped in three categories:

- `Disatisfied`: "Extremely dissatisfied", "Moderately dissatisfied" and "Slightly dissatisfied", encoded to 1
- `Neutral`: "Neither satisfied nor dissatisfied", encoded to 2
- `Satisfied`: "Slightly satisfied", "Moderately satisfied" and "Extremely satisfied", encoded to 3

Below more details about the encoding can be found:

[1] "Slightly satisfied" "Moderately satisfied"
[3] "Slightly dissatisfied" "Neither satisfied nor dissatisfied" [5] "Extremely satisfied" "Extremely dissatisfied"
[7] "Moderately dissatisfied"

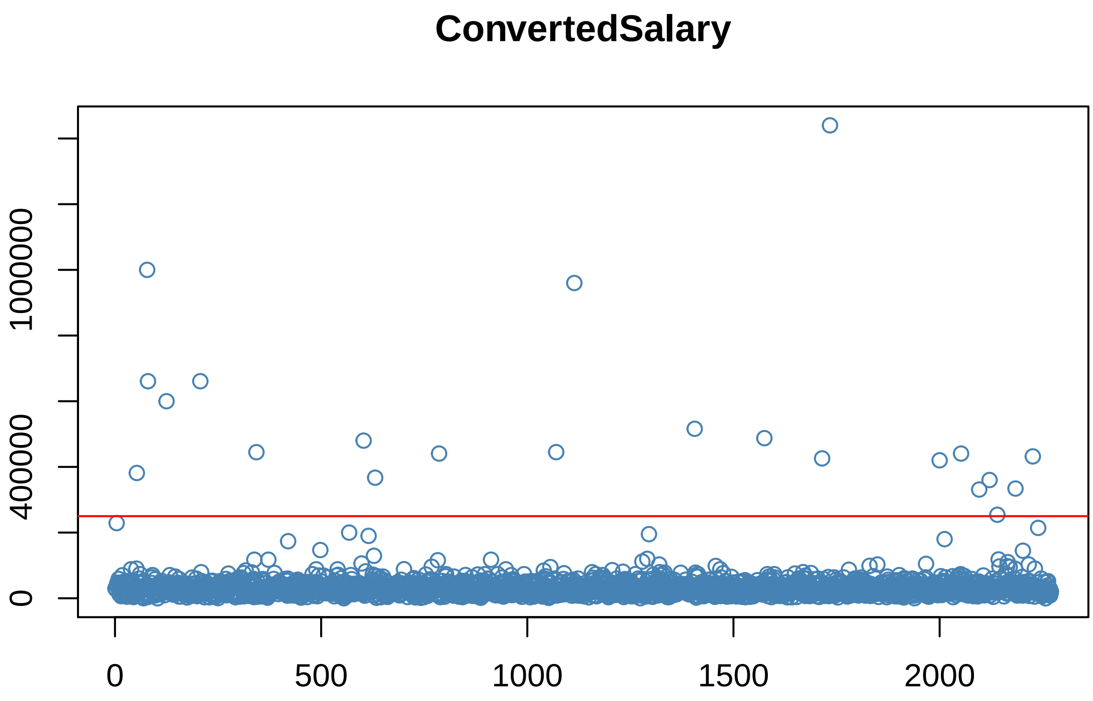
```
survey$JobSatisfaction <-
  plyr::revalue(
    survey$JobSatisfaction,
    c(
      "Extremely dissatisfied" = 1,
      "Moderately dissatisfied" = 1,
      "Slightly dissatisfied" = 1,

      "Neither satisfied nor dissatisfied" = 2,

      "Slightly satisfied" = 3,
      "Moderately satisfied" = 3,
      "Extremely satisfied" = 3
    )
  ) %>% as.factor()
```

- `ConvertedSalary`: This is a `numerical continuous` variable. This feature was analyzed in a graph in order to find some `outliers`. Finally, twenty-two observations were deleted. All of them were greater than the `cutoff point` selected – `250,000 $`. Finally all the values were divided by 1000.

Below we can find a `scatter plot` of this feature, in `red` we can find the `cutoff point`:



We can find below the amount of observation which `ConvertedSalary` is greater than the `cutoff point`:

```r
survey %>%
  filter(survey$ConvertedSalary > 250000) %>%
  nrow() %>%
  knitr::knit_print()
```

[1] 22

As we can we will remove 22 outliers:

```r
survey <- survey %>%
  filter(survey$ConvertedSalary < 250000)
```

Finally as mentioned before, all of the salaries will be divided by 1000:

```r
survey$ConvertedSalary <- survey$ConvertedSalary / 1000
```

- `Hobby` - This `characteristic binomial` feature, "Yes" or "No", was respectively transformed to "1" and "0".

[1] "No" "Yes"

```r
survey$Hobby <-
  plyr::revalue(survey$Hobby,
                c("No" = 0,
                  "Yes" = 1)) %>% as.factor()
```

- `Age` - This `characteristic` feature has `six levels` that can be `ordered in a logical way`, from "Under 18 years old", to "55 - 64 years old".
  Because this feature can be `ordered in a logical way`, it will be grouped in three categories:
- `Early Adulthood` and `Adolescence`: "Under 18 years old" and "18 - 24 years old", encoded to 1
- `Midlife`: "25 - 34 years old" and "35 - 44 years old", encoded to 2
- `Mature Adulthood`: "45 - 54 years old" and "55 - 64 years old", encoded to 3

Below can be found more details about the encoding:

[1] “25 - 34 years old” “45 - 54 years old” “35 - 44 years old” [4] “18 - 24 years old” “Under 18 years old” “55 - 64 years old”

```r
survey$Age <-
  plyr::revalue(
    survey$Age,
    c(
      "Under 18 years old" = 1,
      "18 - 24 years old" = 1,

      "25 - 34 years old" = 2,
      "35 - 44 years old" = 2,

      "45 - 54 years old" = 3,
      "55 - 64 years old" = 3
    )
  ) %>% as.factor()
```

As we saw above, the `characteristics` features were transformed `to factors`.

After cleaning the data, the dataset contains `2,248 observations` and 9 features.

[1] “The dataset survey after the preparation has 2248 rows and 9 columns”

Finally, we will change the names of the columns, in order to remove the capital letters:

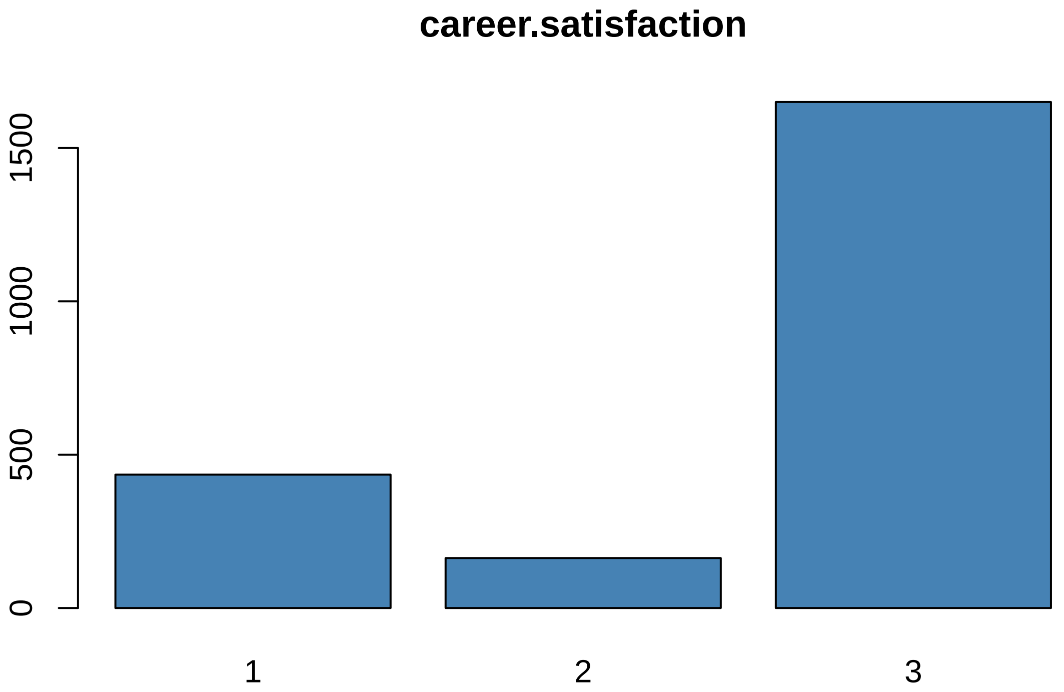
```r
colnames(survey) <- c("career.satisfaction", "country", "education", "company.size",
                      "years.coding", "job.satisfaction", "salary", "hobby", "age")
```

4.2) Data representation:

Below there is a `summary` of the dataset:

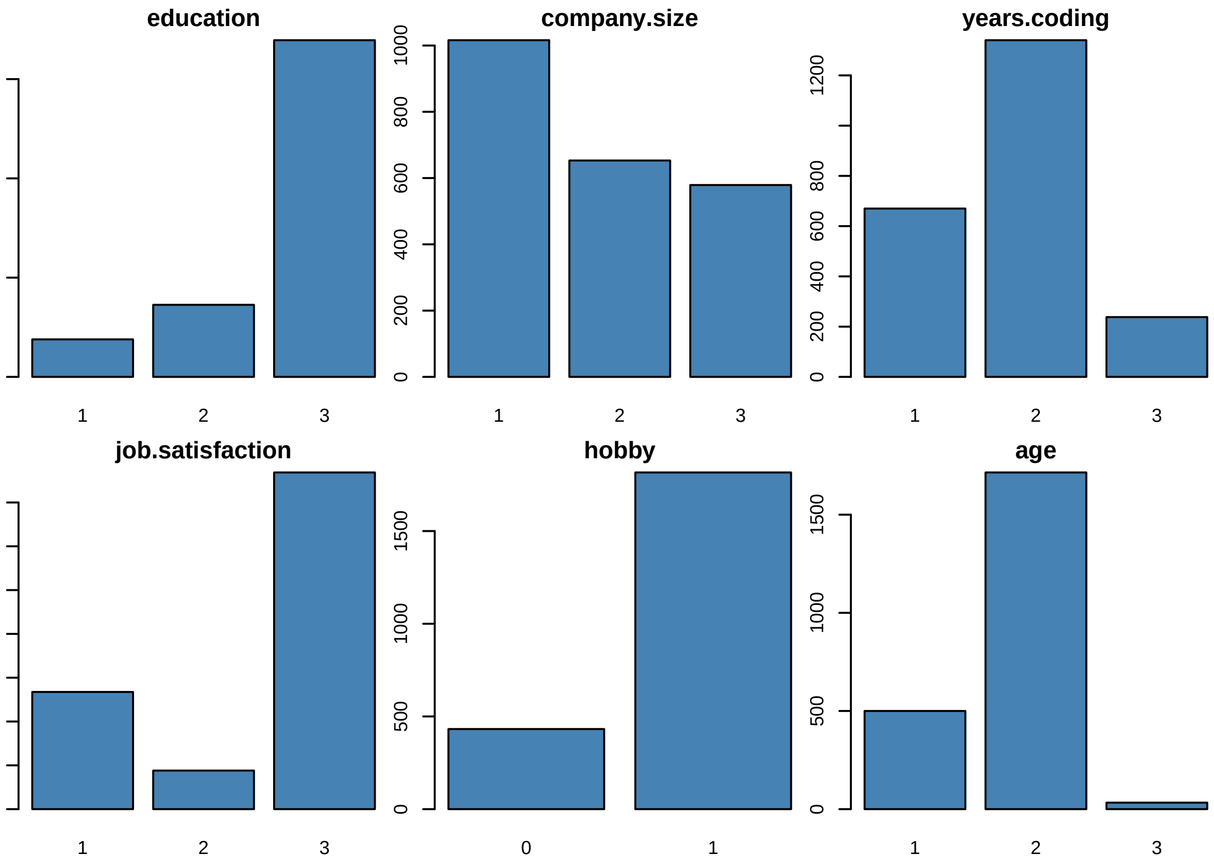
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
career.satisfaction*	1	2248	2.54	0.80	3.00	2.67	0.00	1.00	3.00	2.00	-1.28
country*	2	2248	5.46	2.29	5.00	5.51	2.97	1.00	9.00	8.00	0.00
education*	3	2248	2.67	0.62	3.00	2.82	0.00	1.00	3.00	2.00	-1.70
company.size*	4	2248	1.81	0.82	2.00	1.76	1.48	1.00	3.00	2.00	0.37
years.coding*	5	2248	1.81	0.61	2.00	1.76	0.00	1.00	3.00	2.00	0.12
job.satisfaction*	6	2248	2.45	0.85	3.00	2.56	0.00	1.00	3.00	2.00	-0.99
salary	7	2248	31.27	20.92	28.24	29.03	17.68	0.04	228.89	228.85	2.48
hobby*	8	2248	1.81	0.39	2.00	1.88	0.00	1.00	2.00	1.00	-1.56
age*	9	2248	1.79	0.44	2.00	1.85	0.00	1.00	3.00	2.00	-0.91
NOTE: The features wit	h aster	ik are	‘factors	‘.							

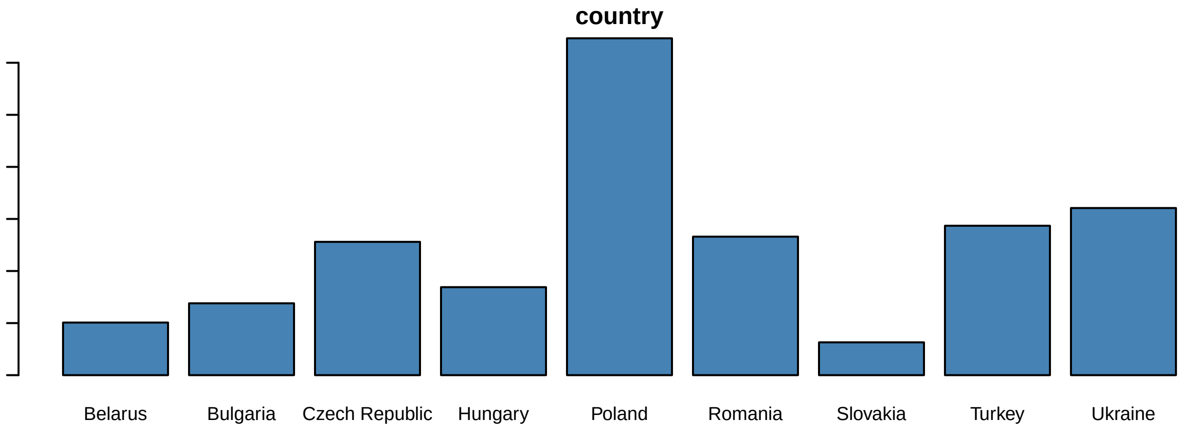
In the next graph we can find the `distribution by levels` of the dependent variable.

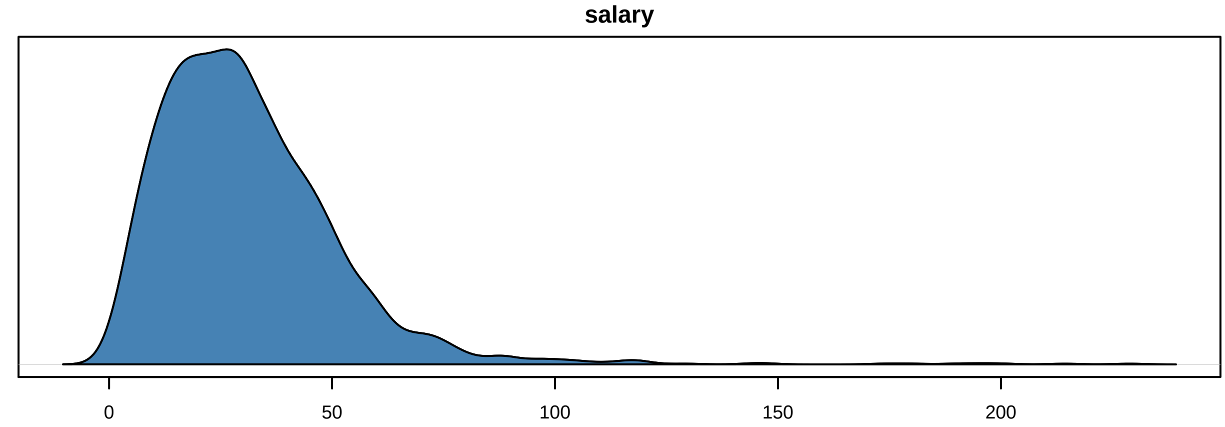


As we can observe the data is `imbalanced`. There are much more people with the level 3 `satisfied` than the other levels, 1 `disatisfied` and 2 `neutral`.

Below we can find a representation of the `covariates`:







Analyzing the table and graph, one can extract several characteristics of the respondents: + Most of them are from Poland. + Their level of education is high. + The majority works in small/medium size companies. + Most of them are amateur, because in general they have not long experience. + Mostly, they are satisfied with their jobs. + They are young. + Coding is a hobby for most of them. + The most common average salary is around 31,000 $ per year.

5) Application of Econometric models:

5.1) Ordered Choice Models:

5.2) Joint significance test:

5.3) Goodness of fit test:

5.4) Marginal effects:

6) Results:

7) Bibliography:

- Joo, B. and McLean, G.N. (2006), “Best employer studies: a conceptual model from a literature review and a case study”, Human Resource Development Review, Vol. 5 No. 2, pp. 228-57.
- Joo, B. and Park, S. (2009), “Career satisfaction, organizational commitment, and turnover intention”, Leadership & Organizational Development Journal, Vol. 21 No. 6, pp. 482-486.
- Judge, T.A., Cable, D.M., Boudreau, J.W. and Bretz, R.D. (1995), “An empirical investigation of the predictors of executive career success”, Personnel Psychology, Vol. 48 No. 3, pp. 485-519.

- Kozlowski, S.W.J., Gully, S.M., Brown, K.G., Salas, E., Smith, E.M. and Nason, E.R. (2001), “Effects of training goals and goal orientation traits on multidimensional training outcomes and performance adaptability”, Organizational Behavior and Human Decision Processes, Vol. 85 No. 1, pp. 1-31.
- Fried, Y., and Ferris, G.R. The validity of the job characteristics model. Personnel Psychology, 40, 2 (1987), 287-322.
- Hu, Nan; Poon, Simon; Zhong, Jiangfan; and Wan, Yun, “Job Satisfaction of Information Technology Professionals” (2004). AMCIS 2004 pp. 3616-3623. http://aisel.aisnet.org/amcis2004/456