УДК 614.2:004.657
МРНТИ 76.75.33, 14.35.07

# ПОРЯДКОВАЯ РЕГРЕССИЯ В ПРОГРАММНОЙ СРЕДЕ R – КРАТКИЕ РЕКОМЕНДАЦИИ ДЛЯ МАГИСТРАНТОВ И ДОКТОРАНТОВ ПО СПЕЦИАЛЬНОСТИ «МЕДИЦИНА» И «ОБЩЕСТВЕННОЕ ЗДРАВООХРАНЕНИЕ»

**В.Л. ЕГОШИН[1], Н.В. САВВИНА[2], А.М. ГРЖИБОВСКИЙ[1,2,3]**

Северный государственный медицинский университет, Архангельск, Россия
Северо-Восточный федеральный университет, Якутск, Россия
Казахский Национальный университет им. Аль-Фараби, Алматы, Казахстан

Егошин В.Л. – http://orcid.org/0000-0002-8407-3789
Саввина Н.В. – http://orcid.org/0000-0003-2441-6193
Гржибовский А.М. – https://orcid.org/0000-0002-5464-0498

Citation/
Библиографиялық сілтеме/
Библиографическая ссылка:

Egoshin VL, Savvina NV, Grjibovski AM. Multivariable ordinal regression in R: brief guidelines for Master and Doctoral students in medicine and public health. West Kazakhstan Medical Journal 2019;61(3):145–153.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. R бағдарламалық ортадағы реттік регрессия – «Медицина» және «Қоғамдық денсаулық сақтау» мамандықтары бойынша магистранттар мен докторанттарға арналған қысқаша ұсынымдар. West Kazakhstan Medical Journal 2019;61(3):145–153.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. Порядковая регрессия в программной среде R – краткие рекомендации для магистрантов и докторантов по специальности «Медицина» и «Общественное здравоохранение». West Kazakhstan Medical Journal 2019;61(3):145–153.

**Multivariable ordinal regression in R: brief guidelines for Master and Doctoral students in Medicine and Public Health**

V.L. Egoshin[1], N.V. Savvina[2], A.M. Grjibovski[1,2,3]
Northern State Medical University, Arkhangelsk, Russia
North-Eastern Federal University, Yakutsk, Russia
Al-Farabi Kazakh National Medical University, Almaty, Kazakhstan

In this paper we describe basic principles of using R package for multivariable ordinal regression analysis. We present step-by-step guidelines and syntax for creation and evaluation of regression models using practical example with real data from a population-based cross-sectional study in the Almaty region. In addition to the syntax we present R outputs and their interpretation

***Keywords:*** *R, ordinal regression, syntax, listing, modeling, validation.*

**R бағдарламалық ортадағы реттік регрессия – «Медицина» және «Қоғамдық денсаулық сақтау» мамандықтары бойынша магистранттар мен докторанттарға арналған қысқаша ұсынымдар**

В.Л. Егошин[1], Н.В. Саввина[2], А.М. Гржибовский[1,2,3]
[1]Солтүстік мемлекеттік медицина университеті, Архангельск, Ресей
[2]Солтүстік-Шығыс федеральды университеті, Якутск, Ресей
[3]Әл-Фараби атындағы Қазақ Ұлттық университеті, Алматы, Қазақстан

Бұл жұмыста реттік айнымалы жауапқа арналған көптеген регрессиялық талдаулар жасау үшін R бағдарламалық ортасын қолданудың негізгі принциптері ұсынылған. Алматы облысында таңдаулы көлденең зерттеулердің модификацияланған бірнеше нақты мәліметтерін қолдана отырып, практикалық мысал түрінде регрессиялық моделдерді жасау және бағалау үшін R-дің қадамдық алгоритмі және синтаксисі ұсынылыған. Синтаксистен бөлек оларды R, сонымен қатар, олардың интерпритациясы көрсеткендей нәтижелері ұсынылған.

***Негізгі сөздер:*** *R, реттік регрессия, синтаксис, листинг, моделдеу, валидация.*

**Порядковая регрессия в программной среде R – краткие рекомендации для магистрантов и докторантов по специальности «Медицина» и «Общественное здравоохранение»**

В.Л. Егошин[1], Н.В. Саввина[2], А.М. Гржибовский[1,2,3]
[1]Северный государственный медицинский университет, г. Архангельск, Россия
[2]Северо-Восточный федеральный университет, г. Якутск, Россия
[3]Казахский Национальный университет им. Аль-Фараби, г. Алматы, Казахстан

В данной работе представлены основные принципы применения программной среды R для осуществления множественного регрессионного анализа для порядковой (ранговой) переменной отклика. Представлен пошаговый алгоритм и синтаксис в R для создания и оценки регрессионных моделей в

Гржибовский А.М.
*e-mail: andrej.grjibovski@gmail.com*

Received/
Келіп түсті/
Поступила:
06.09.2019

Accepted/
Басылымға қабылданды/
Принята к публикации:
27.09.2019

*ISSN 1814-5620 (Print)*

*Published by West Kazakhstan Marat Ospanov Medical University*

виде практического примера с использованием несколько модифицированных реальных данных выборочного поперечного исследования в Алматинской области. Помимо синтаксиса представлены результаты в том виде, как их выдает R, а также их интерпретация.

***Ключевые слова:*** *R, порядковая регрессия, синтаксис, листинг, моделирование, валидация.*

**Введение**

В предыдущих статьях нашей серии мы начали знакомить читателей с методами множественного регрессионного анализа, который предназначен для изучения связи между независимыми переменными (предикторами) и зависимой переменной (переменной отклика). Если зависимая переменная представлена непрерывной величиной, то чаще всего используется метод линейной регрессии. При бинарной (имеющей только два значения) переменной отклика используется метод бинарной логистической регрессии при количестве значений больше двух – мультиномиальная логистическая регрессия; если зависимая переменная является порядковой, то используется метод порядковой регрессии (порядковой логистической регрессии).

Метод порядковой логистической регрессии используется в эпидемиологии и общественном здравоохранении при оценке влияния факторов риска на изучаемые заболевания и состояния [1, 2], при прогнозировании течения заболеваний, в визуальной диагностике [3-6], для оценки удовлетворенности пациентов [7] и в других областях здравоохранения.

Теоретическое обоснование использования порядковой логистической регрессии при анализе данных представлено в многочисленных работах [8-11], однако этот вид анализа крайне редко используются на постсоветском пространстве. Современные статистические пакеты (SPSS, SAS, Stata, R и другие) позволяют выполнять логистический регрессионный анализ для порядковых переменных отклика. Достоинства пакета R при этом виде анализа связаны с ее бесплатностью, возможностями в использовании различных моделей, методов трансформации и визуализации данных. Для выполнения порядкового регрессионного анализа в R используются функции разных пакетов: MASS, ordinal, rms и других. Для детального ознакомления с основными принципами выполнения логистического регрессионного анализа для порядковых переменных отклика можно рекомендовать публикации Harrell [12], Шитикова В.К. [13], Christensen [14], Rawat [15] и Sagar [16].

В данной работе обсуждаются основы порядкового регрессионного анализа и использования R для построения упорядоченной логит-модели. После описания и преобразования используемого набора данных (листинг 1) на примере упорядоченной логит-модели с двумя независимыми переменными (листинг 2), демонстрируется, каким образом полученные при её анализе константы регрессии (intercepts) и коэффициенты могут быть использованы для предсказания результатов модели (листинг 3). В листинге 3 также показано использование функции predict, вычисляющей эти предсказываемые значения. В листинге 4 показано, как набор данных разделяется на тренировочный и тестовый сеты, создаётся упорядоченная логит-модель, использующая все независимые переменные набора данных. Полученная модель оценивается с использованием функции car: Anova, создается модель с меньшим количеством предикторов, модели сравниваются между собой. В листинге 5 модель оценивается по уровню ошибок и точности в тренировочном и тестовом сетах. Листинг 6 показывает, как с помощью функций пакета effects можно визуально оценить влияние предикторов на переменную отклика.

В работе используются базовые пакеты R 3.5.3, а также пакеты dplyr, broom, MASS, effects. Работа выполнена в IDE RStudio ver. 1.2.1335. Для создания модели порядковой логистической регрессии использована функция polr из пакета MASS.

*Упорядоченная логит-модель*

Порядковую логистическую регрессию можно рассматривать как расширение простой логистической регрессии для бинарной переменной отклика. В простой логистической регрессии логарифм шансов возникновения события моделируется как линейная комбинация независимых переменных. При выполнении порядкового логистичесого регрессионного анализа наиболее часто выполняется анализ упорядоченной логит модели (ordered logit model, ordered logistic regression, proportional odds model). В этой модели используется накапливание событий для логарифма вычисленных шансов. Графически эта модель представлена на рис.1.

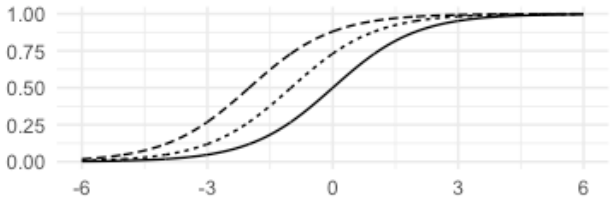


Рис.1 График упорядоченной логистической модели

Модель пропорциональных шансов предполагает, что эффект независимых переменных идентичен для каждого вычисляемого логарифма шансов. При этом значение константы (intercept) имеет различные величины. Модель пропорциональных шансов также предполагает отсутствие мультиколлинеарности (сильной корреляции) между независимыми переменными.

В данной работе используется выборка из реально существующего проекта, реализуемого в

Алматинской области Республики Казахстан [17]. В качестве примера рассматривается оценка влияния различных факторов на самооценку здоровья на случайной выборке из основной базы данных. Переменная SRH (самооценка здоровья) представляет собой порядковую переменную с тремя градациями: Fair or worse (удовлетворительное или хуже), Good (хорошее), Very good (очень хорошее). Возможны следующие варианты логарифма шансов:

$$LogOdds_{\text{Fair or worse}} = \log\frac{P(\text{Fair or worse})}{P(\text{Good}) + P(\text{Very good})}$$

$$LogOdds_{\text{Fair or worse | Good}} = \log\frac{P(\text{Fair or worse}) + P(\text{Good})}{P(\text{Very good})}$$

Формула модели пропорциональных шансов по Rawat (2018):

$$\text{logit}[P(Y \le j)] = \alpha_j - \sum_{1}^{i} \beta_i x_i \ ,$$

где $Y$ – порядковая зависимая переменная; $j$ – число категорий порядковой зависимой переменной минус 1; $\alpha_j$ – intercept, $\beta_i$ коэффициенты независимых переменных $x_i$

**Используемый набор данных**

Начало работы с данными предполагает знакомство с их структурой, выявление пропущенных значений, анализ переменных. В используемом наборе данных все переменные категориальные. После знакомства с данными выполнена их модификация: определены референтные значения и имеющиеся категории, уточнен порядок для зависимой переменной (листинг 1).

Листинг 1

```
# импорт данных
df <- foreign::read.spss(«../data/ordinal_regression.sav», to.data.frame = TRUE)
# структура данных
glimpse(df)
Observations: 1,000
Variables: 7
$ Gender <fct> Male, Male, Male, Male, Male, Female, …
$ Ethnicity <fct> Kazakh, Kazakh, Kazakh, Kazakh, Kazakh…
$ SRH <fct> Fair or worse, Fair or worse, Fair or …
$ MS <fct> Other, Married, Married, Married, Sing…
$ Education <fct> Higher, Vocational, Vocational, Vocati…
$ Occupation <fct> “Part time employed”, “Part time emplo…
$ Smoking <fct> Occasional smoker, Non-smoker, Daily s…
    # определение пропущенных значений
    df %>%
    inspectdf::inspect_na() %>%
    knitr::kable(caption = «Таб.1 Данные о пропущенных значениях»)
```

Таб.1 Данные о пропущенных значениях

| col_name | cnt | pcnt |
|---|---|---|
| Gender | 0 | 0 |
| Ethnicity | 0 | 0 |
| SRH | 0 | 0 |
| MS | 0 | 0 |
| Education | 0 | 0 |
| Occupation | 0 | 0 |
| Smoking | 0 | 0 |

```
    # характеристика переменных, выводится самая частая категория
    df %>%
    inspectdf::inspect_imb() %>%
    knitr::kable(caption = «Таб.2 Частые категории переменных»)
```

Таб.2 Частые категории переменных

| col_name | value | pcnt | cnt |
|---|---|---|---|
| Ethnicity | Kazakh | 80.6 | 806 |
| Smoking | Non-smoker | 73.9 | 739 |
| MS | Married | 68.5 | 685 |
| Occupation | Full time employed | 54.5 | 545 |
| Gender | Female | 51.9 | 519 |
| SRH | Good | 43.5 | 435 |
| Education | Higher | 37.7 | 377 |

```
# функция для представления всех категорий переменной
    tab_apply <- function(variable) {
     janitor::tabyl(variable) %>%
     janitor::adorn_totals(“row”) %>%
 janitor::adorn_pct_formatting()}
# вывод данных о категориях по всем переменным
apply(df, 2, tab_apply)
    $Gender
 variable n percent
 Female 519 51.9%
 Male 481 48.1%
 Total 1000 100.0%

$Ethnicity
 variable n percent
 Kazakh 806 80.6%
 Other 77 7.7%
 Russian 117 11.7%
 Total 1000 100.0%

$SRH
 variable n percent
 Fair or worse 348 34.8%
 Good 435 43.5%
 Very good 217 21.7%
 Total 1000 100.0%

$MS
 variable n percent
```

Married 685 68.5%
Other 100 10.0%
Single 215 21.5%
Total 1000 100.0%

$Education
variable n percent
Basic 59 5.9%
Higher 377 37.7%
Secondary 282 28.2%
Vocational 282 28.2%
Total 1000 100.0%

$Occupation
variable n percent
Full time employed 545 54.5%
Out of work (students, other) 89 8.9%
Part time employed 145 14.5%
Pensioner 75 7.5%
Self-emploed 146 14.6%
Total 1000 100.0%

$Smoking
variable n percent
Daily smoker 145 14.5%
Non-smoker 739 73.9%
Occasional smoker 116 11.6%
Total 1000 100.0%

```
    # модификация набора данных
df <- df %>%
 mutate(
 SRH = factor(SRH, levels = c("Fair or worse", "Good", "Very good"), ordered = TRUE),
 Ethnicity = relevel(Ethnicity, ref = "Kazakh"),
 MS = relevel(MS, ref = "Married"),
 Education = factor(Education, levels = c("Higher", "Basic", "Secondary", "Vocational")),
 Occupation = relevel(Occupation, ref = "Full time employed"),
 Smoking = factor(Smoking, levels = c("Non-smoker", "Daily smoker", "Occasional smoker"))
 )
```

Используемый набор данных содержит 1000 записей в 7 переменных. Пропущенные значения отсутствуют.

**Интерпретация результатов упорядоченной логит-модели**

Для получения упорядоченной логит-модели использована функция polr (polr – аббревиатура от proportional odds logistic regression) пакета MASS. В качестве независимых переменных используем переменные семейного статуса и курения. Цель использования модели – предсказать вероятные значения зависимой переменной. Предскажем значение переменной SRH для первой записи набора данных:

***Листинг 2***

```
    # создание модели
model <- polr(SRH ~ MS + Smoking, df, Hess = TRUE)

# вывод данных о модели
summary (model)
    Call:
polr(formula = SRH ~ MS + Smoking, data = df, Hess = TRUE)

Coefficients:
 Value Std. Error t value
MSSingle 0.70481 0.1471 4.7923
MSOther -0.37191 0.2076 -1.7910
SmokingDaily smoker -0.41768 0.1734 -2.4091
SmokingOccasional smoker -0.09109 0.1893 -0.4811

Intercepts:
 Value Std. Error t value
Fair or worse|Good -0.6023 0.0843 -7.1434
Good|Very good 1.3657 0.0946 14.4335

Residual Deviance: 2085.503
AIC: 2097.503
    # данные о независимых переменных модели в первой записи набора данных
df[1, c(«MS», «Smoking»)]
     MS Smoking
1 Other Occasional smoker
```

Полученная модель оценивается по показателям константы (Intercept), коэффициентам и некоторым другим. Константа представляет собой значение изучаемой функции при нулевых значениях x, при работе с категориальными независимыми переменными – при их референтных значениях. Коэффициенты показывают величину влияния независимых переменных на зависимую переменную.

В нашей модели
значения intercepts равны
Fair or worse|Good -0.6023
Good|Very good 1.3657
значения коэффициентов равны
MSSingle 0.70481
MSOther -0.37191
SmokingDaily smoker -0.41768
SmokingOccasional smoker -0.09109

Intercept Fair or worse|Good -0.6023 соответствует $\text{logit}(P(Y = \text{Fair or worse}))$

Intercept Good|Very good 1.3657 соответствует $\text{logit}(P(Y = \text{Fair or worse} \mid Y = \text{Good}))$

Сумма произведений коэффициентов на значение независимых переменных $\sum_{1}^{i} \beta_i x_i$ для данной модели равна

$$0.70481 \cdot \text{MSSingle} + (-0.37191) \cdot \text{MSOther} + (-0.41768) \cdot \text{SmokingDaily smoker} + (-0.09109) \cdot \text{SmokingOccasional smoker}$$

Обращаем Ваше внимание, что участник исследования из первой записи имеет значения переменных MS - Other, Smoking - Occasional smoker.

Поэтому в формуле для расчета суммы произведений коэффициентов на значение независимых переменных значения $x$ для коэффициента MSSingle = 0, MSOther = 1, SmokingDaily smoker = 0, SmokingOccasional smoker = 1.

Вероятность события SRH = Fair or worse может быть рассчитана как $P(\text{Fair or worse}) = \frac{1}{1 + exp^{-(\alpha_1 - \sum \beta_i x_i)}}$, где $\alpha_1$ – intercept Fair or worse|Good -0.6023

Вероятность события SRH = Fair or worse | Good может быть рассчитана как $P(\text{Fair or worse} \mid \text{Good}) = \frac{1}{1 + exp^{-(\alpha_2 - \sum_1^i \beta_i x_i)}}$, где $\alpha_2$ – intercept Good|Very good 1.3657

Вероятность события SRH = Good может быть рассчитана как $P(\text{Good}) = P(\text{Fair or worse} \mid \text{Good}) - P(\text{Fair or worse})$

Вероятность события SRH = Very good может быть рассчитана как $P(\text{Very good}) = 1 - P(\text{Fair or worse} \mid \text{Good})$

Функция predict с аргументом type = «prob» позволяет избегать подобных расчётов.

Листинг 3

```
# сумма произведений коэффициентов на значения x первой записи набора данных
(y1 <- 0.70481 * 0 + (-0.37191) * 1 + (-0.41768) * 0 + (-0.09109) * 1)
[1] -0.463
# вероятность события SRH = Fair or worse
(p1 <- 1/(1 + exp(-(-0.6023 - y1))))
[1] 0.4652312
# вероятность события SRH = Fair or worse | Good
(p1or2 <- 1/(1 + exp(-(1.3657 - y1))))
[1] 0.8616068
# вероятность события SRH = Good
(p2 <- p1or2 - p1)
[1] 0.3963756
# вероятность события SRH = Very good
(p3 <- 1 - p1or2)
[1] 0.1383932
# сравните
round(c(p1, p2, p3),3)
[1] 0.465 0.396 0.138
predict(model, df[1,], type = "prob") %>% round(3)
Fair or worse Good Very good
0.465 0.396 0.138
```

Сравнение результатов расчёта с использованием констант регрессии и коэффициентов модели с результатами выполнения функции predict показывает, что при использовании этой функции отсутствует необходимость выполнения расчётов для получения прогностических значений модели.

**Упорядоченная логит-модель, включающая все переменные**

Упорядоченная логит-модель, включающая все переменные, будет использована для классификации результатов. Для оценки прогностических возможностей модели исходный набор данных будет разделен на тренировочный и тестовый сеты. Модель будет создана на основе тренировочного набора данных. Данные оценки модели будут выведены с использованием функций пакета broom. В таблице 3 будут выведены статистически значимые коэффициенты (не содержащие ноль в доверительном интервале); в таблице 4 - экспоненцированные значения статистически значимых коэффициентов (не содержащие единицу в доверительном интервале); в таблице 5 - экспоненцированные значения intercept. Оценка модели выполнена с использованием функции Anova пакета car. В измененной модели не использована переменная Gender. При сравнении

Таб.3 Коэффициенты модели

| term | estimate | std.error | statistic | conf.low | conf.high | coefficient_type |
|---|---|---|---|---|---|---|
| EducationBasic | 0.826 | 0.350 | 2.363 | 0.141 | 1.516 | coefficient |
| EducationSecondary | -0.176 | 0.186 | -0.945 | -0.542 | 0.189 | coefficient |
| EducationVocational | 0.070 | 0.180 | 0.392 | -0.282 | 0.423 | coefficient |
| EthnicityOther | 0.086 | 0.267 | 0.324 | -0.439 | 0.608 | coefficient |
| EthnicityRussian | 0.789 | 0.236 | 3.343 | 0.328 | 1.255 | coefficient |
| Fair or worse\|Good | -1.068 | 0.174 | -6.152 | NA | NA | zeta |
| GenderFemale | -0.233 | 0.154 | -1.508 | -0.536 | 0.069 | coefficient |
| Good\|Very good | 1.076 | 0.175 | 6.163 | NA | NA | zeta |
| MSOther | -0.444 | 0.259 | -1.715 | -0.957 | 0.060 | coefficient |
| MSSingle | 0.671 | 0.181 | 3.709 | 0.317 | 1.027 | coefficient |
| OccupationOut of work (students, other) | -0.575 | 0.258 | -2.232 | -1.083 | -0.071 | coefficient |
| OccupationPart time employed | -0.392 | 0.225 | -1.744 | -0.835 | 0.048 | coefficient |
| OccupationPensioner | -1.769 | 0.334 | -5.288 | -2.452 | -1.133 | coefficient |
| OccupationSelf-emploed | -0.507 | 0.226 | -2.240 | -0.952 | -0.065 | coefficient |
| SmokingDaily smoker | -0.715 | 0.233 | -3.067 | -1.176 | -0.261 | coefficient |
| SmokingOccasional smoker | -0.331 | 0.234 | -1.414 | -0.791 | 0.127 | coefficient |

моделей функцией anova не выявлено значимых различий между моделями.

***Листинг 4***

```
    # выделение тренировочного и тестового сетов
set.seed(123) # для воспроизведения
# определение номеров записей, входящих в
тренировочный сет
index <- sample(seq_len(nrow(df)), size = .7 *
nrow(df))
# создание тренировочного сета
df_train <- df[index, ]
# создание тестового сета
df_test <- df[-index, ]
    # создание модели
model <- polr(SRH ~ ., data = df_train, Hess = TRUE)
    # параметры модели
glance(model) %>% knitr::kable(digits = 3)
```

| edf | logLik | AIC | BIC | deviance | df.residual |
|---|---|---|---|---|---|
| 16 | -697.912 | 1427.825 | 1500.642 | 1395.825 | 684 |

```
# коэффициенты модели
tidy(model, conf.int = TRUE) %>% knitr::kable(digits =
3, caption = "Таб.3 Коэффициенты модели")
    # значения коэффициентов модели
tidy(model, conf.int = TRUE) %>%
 mutate(ci_estimation = ifelse(conf.low < 0 & conf.high
> 0, FALSE, TRUE)) %>%
 filter(ci_estimation == TRUE) %>%
 dplyr::select(-coefficient_type, -ci_estimation) %>%
 knitr::kable(digits = 3, caption = "Таб.3 Статистически
значимые коэффициенты")
    # экспоненцированные значения коэффициентов
tidy(model, conf.int = TRUE, exponentiate = TRUE)
%>%
 mutate(ci_estimation = ifelse(conf.low < 1 & conf.high
> 1, FALSE, TRUE)) %>%
 filter(ci_estimation == TRUE) %>%
 dplyr::select(-coefficient_type, -ci_estimation) %>%
 knitr::kable(digits = 3, caption = "Таб.4 Статистически
значимые коэффициенты")
    # экспоненцированные значения intercept
tidy(model, conf.int = TRUE, exponentiate = TRUE)
%>%
 filter(coefficient_type == "zeta") %>%
 dplyr::select(-coefficient_type) %>%
 knitr::kable(digits = 3, caption = "Таб.5
Экспоненцированные значения intercept")
    # оценка модели
car::Anova(model)
    Analysis of Deviance Table (Type II tests)

Response: SRH
 LR Chisq Df Pr(>Chisq)
Gender 2.278 1 0.131242
Ethnicity 11.282 2 0.003549 **
```

Таб.3 Статистически значимые коэффициенты

| term | estimate | std.error | statistic | conf.low | conf.high |
|---|---|---|---|---|---|
| EducationBasic | 0.826 | 0.350 | 2.363 | 0.141 | 1.516 |
| EthnicityRussian | 0.789 | 0.236 | 3.343 | 0.328 | 1.255 |
| MSSingle | 0.671 | 0.181 | 3.709 | 0.317 | 1.027 |
| OccupationOut of work (students, other) | -0.575 | 0.258 | -2.232 | -1.083 | -0.071 |
| OccupationPensioner | -1.769 | 0.334 | -5.288 | -2.452 | -1.133 |
| OccupationSelf-emploed | -0.507 | 0.226 | -2.240 | -0.952 | -0.065 |
| SmokingDaily smoker | -0.715 | 0.233 | -3.067 | -1.176 | -0.261 |

Таб.4 Статистически значимые коэффициенты

| term | estimate | std.error | statistic | conf.low | conf.high |
|---|---|---|---|---|---|
| EducationBasic | 2.285 | 0.350 | 2.363 | 1.151 | 4.556 |
| EthnicityRussian | 2.202 | 0.236 | 3.343 | 1.389 | 3.508 |
| MSSingle | 1.956 | 0.181 | 3.709 | 1.374 | 2.792 |
| OccupationOut of work (students, other) | 0.563 | 0.258 | -2.232 | 0.339 | 0.931 |
| OccupationPensioner | 0.171 | 0.334 | -5.288 | 0.086 | 0.322 |
| OccupationSelf-emploed | 0.602 | 0.226 | -2.240 | 0.386 | 0.938 |
| SmokingDaily smoker | 0.489 | 0.233 | -3.067 | 0.308 | 0.770 |

Таб.5 Экспоненцированные значения intercept

| term | estimate | std.error | statistic | conf.low | conf.high |
|---|---|---|---|---|---|
| Fair or worse\|Good | 0.344 | 0.174 | -6.152 | NA | NA |
| Good\|Very good | 2.933 | 0.175 | 6.163 | NA | NA |

```
MS 19.836 2 4.927e-05 ***
Education 8.035 3 0.045288 *
Occupation 36.013 4 2.876e-07 ***
Smoking 10.276 2 0.005870 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
    # модель с использованием всех предикторов кроме Gender
model1 <- polr(SRH ~ . - Gender, data = df_train, Hess = TRUE)
glance(model1)
    # A tibble: 1 x 6
 edf logLik AIC BIC deviance df.residual
 <int> <dbl> <dbl> <dbl> <dbl> <dbl>
1 15 -699. 1428. 1496. 1398. 685
    # сравнение моделей
anova(model, model1)
    Likelihood ratio tests of ordinal regression models

Response: SRH
 Model
1 (Gender + Ethnicity + MS + Education + Occupation + Smoking) - Gender
2 Gender + Ethnicity + MS + Education + Occupation + Smoking
 Resid. df Resid. Dev Test Df LR stat. Pr(Chi)
1 685 1398.102
2 684 1395.825 1 vs 2 1 2.277757 0.1312416
```

**Оценка прогностических возможностей модели**

Чтобы оценить прогностические возможности модели, будет создана матрица несоответствий (confusion matrix) и определена доля ошибок. По диагонали полученной матрицы приводится количество случае совпадения реальных и предсказанных значений. Используя функцию confusionMatrix пакета caret можно оценить точность модели.

***Листинг 5***

```
    # предсказанные значения для тренировочного сета
pred <- predict(model, df_train, type = «class»)

# сравнение реальных и предсказанных значений (первые 6 позиций)
df_train %>% dplyr::select(SRH) %>% cbind(pred) %>% head()
     SRH pred
288 Fair or worse Fair or worse
788 Fair or worse Good
409 Fair or worse Good
881 Good Good
937 Fair or worse Good
46 Good Good
    # матрица несоответствий для тренировочного сета
(tab <- table(pred, df_train$SRH))

pred Fair or worse Good Very good
 Fair or worse 80 39 17
 Good 153 265 118
 Very good 3 8 17
    # определение доли ошибок для тренировочного сета
(m1 <- round(1 - sum(diag(tab)) / sum(tab), 3))
    [1] 0.483
    # оценка точности
(accur1 <- caret::confusionMatrix(tab)$overall[c(1, 3, 4)] %>% round(3))
     Accuracy AccuracyLower AccuracyUpper
 0.517 0.479 0.555
    # предсказанные значения для тестового сета
pred <- predict(model, df_test, type = “class”)

# матрица несоответствий для тестового сета
(tab <- table(pred, df_test$SRH))

pred Fair or worse Good Very good
 Fair or worse 42 21 8
 Good 69 101 54
 Very good 1 1 3
    # определение доли ошибок для тестового сета
(m2 <- round(1 - sum(diag(tab)) / sum(tab), 3))
    [1] 0.513
    # оценка точности
(accur2 <- caret::confusionMatrix(tab)$overall[c(1, 3, 4)] %>% round(3))
     Accuracy AccuracyLower AccuracyUpper
 0.487 0.429 0.545
```

Уровень ошибок при оценке тренировочного сета составил 48.3%, при оценке тестового сета - 51.3%. Точность модели при оценке тренировочного сета: 0.517, доверительный интервал 0.479-0.555; при оценке тестового сета 0.487, доверительный интервал 0.429-0.545.

**Графическое представление результатов**

Интерпретация порядковой логистической регрессии в терминах логарифма шансов достаточно сложна. Визуализация с использованием пакета effects упрощает эту задачу. Показатели на шкалах выводятся в значениях вероятности (листинг 6). Например, на рисунке 2, оценивающем эффект фактора образования, можно увидеть, что в случае базового образования вероятность оценки собственного здоровья как Fair or worse будет достоверно ниже, чем вероятность оценки Good; в случае высшего образования вероятность оценки Fair or worse выше вероятности оценки Very good. Функции пакета effects позволяют также оценить влияние двух факторов (рисунок 7).

***Листинг 6***

```
plot(Effect(focal.predictors = “Education”, model))
```

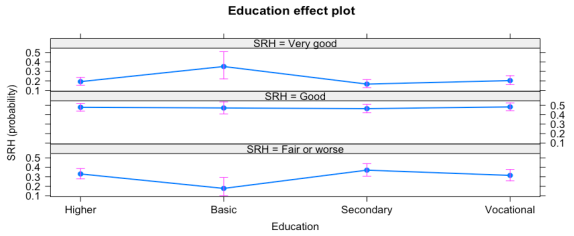


Рис.2 Графическая оценка эффекта фактора образования

**plot(Effect**(focal.predictors = "MS", model))

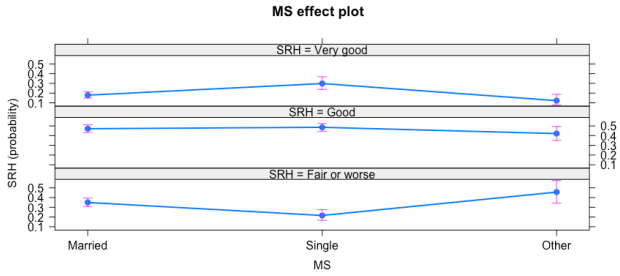


Рис.3 Графическая оценка эффекта фактора семейного положения

**plot(Effect**(focal.predictors = "Ethnicity", model))

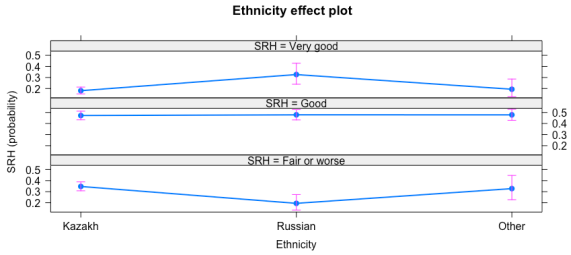


Рис.4 Графическая оценка эффекта этнического фактора

**plot(Effect**(focal.predictors = "Smoking", model))

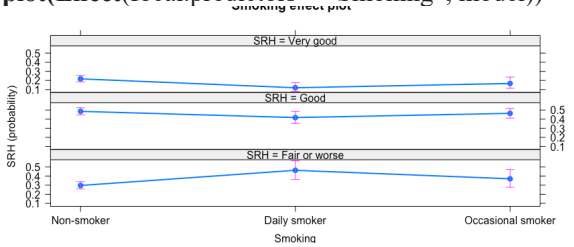


Рис.5 Графическая оценка эффекта фактора курения

**plot(Effect**(focal.predictors = "Occupation", model))

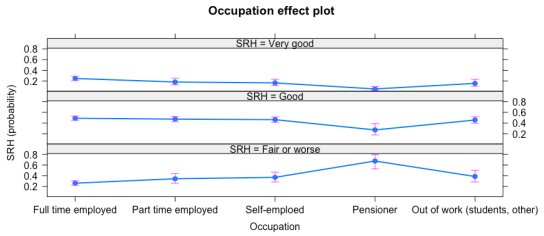


Рис.6 Графическая оценка эффекта фактора трудовой деятельности

plot(Effect(focal.predictors = c("MS", "Gender"), model))

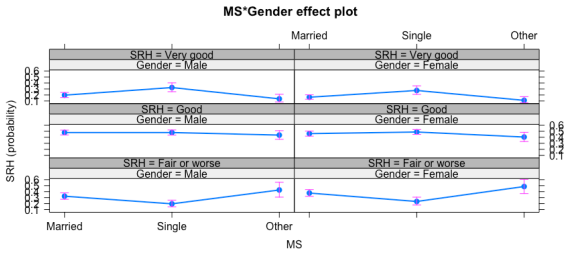


Рис.7. Графическая оценка эффекта факторов семейного положения и курения

Работа с R

Программная среда R является свободно распространяемым кросс-платформенным программным средством, использующимся для статистических вычислений и визуализации данных. Дистрибутивы R доступны на сайтах The Comprehensive R Archive Network, https://cran.r-project.org, Microsoft R Application Network, https://mran.microsoft.com/download. Удобным IDE (integrated development environment, интегрированная среда разработчика) для программы R является программа RStudio, свободно распространяемый дистрибутив может быть загружен на сайте RStudio IDE, https://www.rstudio.com/products/rstudio/. В наших более ранних публикациях (18) мы уже касались вопросов применения программной среды R в биомедицинских исследованиях. Использованный в работе файл с набором данных и скрипт с кодом доступны на сайте https://github.com/valegoshin/Paper_Scripts.

***Конфликт интересов.*** *Авторы заявляют об отсутствии конфликта интересов. Работа выполнена без внешнего финансирования.*

*Список литературы / References:*

1. Abikulova AK, Tulebaev KA, Akanov AA, Turdalieva BS, Kalmahanov SB, Kumar AB, Izekenova, Mussaeva BA, Grjibovski AM. Inequalities in self-rated health among 45 + year-olds in Almaty, Kazakhstan: a cross-sectional study. BMC Public Health. 2013;13:654.
2. Siqueira AL, Clareci CS. Ordinal logistic regression models: application in quality of life studies. Cad. Saúde Pública. 2008;24:581–591.
3. Doyle OM, Ashburner J, Zelaya FO, Williams SCR, Mehta MA, Marquand AF. Multivariate Decoding of Brain Images Using Ordinal Regression. Neuroimage. 2013;81:347–357.
4. Doyle OM, Westman E, Marquand AF, Mecocci P, Vellas B, Tsolaki M, Kłoszewska I. Predicting Progression of Alzheimer's Disease Using Ordinal Regression. PLoS One. 2014;9.
5. Keeble C, Baxter PD, Gislason-Lee AJ, Treadgold LA, Davies AG. Methods for the Analysis of Ordinal Response Data in Medical Image Quality Assessment. Br J Radiol. 2016;89(1063):20160094.
6. Satake E, Majima K, Aoki SC, Kamitani Y Sparse Ordinal Logistic Regression and Its Application to Brain Decoding. Front Neuroinform. 2018;12:51.
7. Koletsi B, Pandis N. Ordinal Logistic Regression. Am J Orthod Dentofacial Orthop. 2018;153(1):157–158.
8. Agresti A. Categorical Data Analysis. Hoboken, New Jersey: Wiley; 2002.
9. McCullagh P. Regression models for ordinal data. Journal of the Royal Statistical Society. 1980;42(2):109–142.
10. Winship C, Mare R. Regression models with ordinal variables. American Sociological Review. 1984;49:512–525.

11. Preisser JS, Phillips C, Perin J, Schwartz TA. Regression models for patient-reported measures having ordered categories recorded on multiple occasions. Community Dent Oral Epidemiol. 2011;39 (2):154–163.
12. Harrell F. Modeling Strategies. Switzerland: Springer International Publishing; 2015.
13. Шитиков ВК, Мастицкий СЭ. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. https://ranalytics.github.io/data-mining/index.html.
*Shitikov VK, Mastitskiy SE. Claccification, regression and other Data Mining algorithms using R. https://ranalytics.github.io/data-mining/index.html. [In Russian]*
14. Christensen R. Cumulative Link Models for Ordinal Regression with the R Package Ordinal. https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf
15. Sagar C. How to Perform Ordinal Logistic Regression in R. https://r-posts.com/how-to-perform-ordinal-logistic-regression-in-r/.
16. Rawat A. Ordinal Logistic Regression. An Overview and Implementation in R. https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5.
17. Индершиева ЕВ, Турдалиева БС, Усатаев ММ, Аимбетова ГЕ. Изучение распространенности избыточной массы тела среди взрослого населения Алматинской области и факторов, влияющих на ее развитие. Вестник КазНМУ. 2019;3:302–304.
*Indershiyeva EV, Turdaliyeva BS, Usatayev MM, Aimbetova GE. To study the prevalence of overweight among the population of Almaty region and factors influencing its development. Vestnik KazNMU. 2019;3:302–304.[In Russian]*
18. Егошин ВЛ, Саввина НВ, Иванов СВ, Капанова ГЖ, Гржибовский АМ. Основы работы в программной среде R при анализе биомедицинских данных. Экология человека. 2018;7:55–64.
*Egoshin VL, Savvina NV, Ivanov SV, Kapanova GZ, Grjibovski AM. Basic Principles of Biomedical Data Analysis in R. Ekologiya cheloveka [Human Ecology] 2017;7:55–64. [In Russian]*